



دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی
مهندسی کامپیوتر

عنوان:

تحلیل و تشخیص احساسات نظرات دیجی کالا

نگارش:

مصطفی قدیمی

استاد راهنما:

دکتر حمید بیگی

بهمن ۱۴۰۰

سلام الغفران

چکیده

در سال‌های اخیر رشد چشم‌گیر و قابل توجهی در مورد تحقیق درباره‌ی تحلیل احساسات از روی متن مشاهده شده است. تحلیل احساسات هم از جنبه‌ی کسب‌وکار و هم از نظر علمی حائز اهمیت است. با روی کار آمدن مدل‌های زبان از پیش‌آموزش‌دیده و هم‌چنین با ایجاد مدل‌های زبانی قدرتمند، عصر جدیدی را در زمینه پردازش زبان طبیعی (NLP) آغاز شده است. در میان این مدل‌ها، مدل‌های مبتنی بر مبدل‌ها مانند BERT به دلیل عملکرد فوق‌العاده خود، به طور افزایشی محبوب شده‌اند. با این حال، این مدل‌ها معمولاً بر روی زبان انگلیسی متمرکز هستند و زبان‌های دیگر را به مدل‌های چند زبانه با منابع محدود واگذار می‌کنند. برای زبان فارسی نیز تحقیقات در زمینه‌ی مدل تک‌زبان‌ی فارسی انجام شده است. پردازش زبان طبیعی به ما در انجام وظایف مختلفی کمک می‌کند. در این میان تجزیه و تحلیل نظرات متنی مردم به‌خصوص در شبکه‌های اجتماعی، کسب‌وکارهای اینترنتی و... از اهمیت و ارزش بالایی برخوردار است. در این پایان‌نامه با محوریت تحلیل عواطف و احساسات نظرات شرکت دیجی‌کالا، ابتدا چالش‌های پیش‌روی در مدیریت حجم بسیار زیاد داده مطرح شده و سپس روش‌هایی برای تایید خودکار نظرات به کمک هوش مصنوعی ارائه شده است. برای استخراج احساسات با کمک مدل BERT از یک روش خلاقانه با تعداد شاخص‌های مشخص استفاده شده است که عمل کرد و خروجی شبکه را تا حد بسیار زیادی بهبود می‌دهد. تاثیرگذاری این مدل در شرکت دیجی‌کالا به نحوی بوده که هوش مصنوعی سهم ۹۰ درصدی از رد یا تایید نظرات را به خود اختصاص داده است.

کلیدواژه‌ها: یادگیری ماشین، پردازش زبان طبیعی، دسته‌بندی متن، تجزیه و تحلیل احساسات، کلان‌داده، دیجی‌کالا

فهرست مطالب

۱	مقدمه	۱
۱-۱	تعریف مسئله	۱
۲-۱	اهمیت موضوع	۲
۳-۱	اهداف تحقیق	۵
۴-۱	ساختار پایان نامه	۵
۲	مفاهیم اولیه و کارهای پیشین	۶
۱-۲	هوش مصنوعی	۶
۲-۲	پردازش زبان طبیعی	۷
۳-۲	وظایف پردازش زبان طبیعی	۸
۴-۲	پیش پردازش متن	۹
۱-۴-۲	یافتن ریشه کلمات ^۱	۱۰
۲-۴-۲	ایست واژه ^۲	۱۰
۵-۲	مبدل ها	۱۰
۱-۵-۲	آشنایی با شبکه های عصبی بازگشتی ^۳	۱۱

^۱Lemmatization or Stemming

^۲Stopword

^۳Recurrent Neural Networks

۱۱	۲-۵-۲ مدل توجه ^۴
۱۳	۶-۲ Google BERT
۱۳	۲-۶-۱ توصیف مدل
۱۴	۲-۶-۲ شهرت نام تجاری ^۵
۱۵	۷-۲ ParsBert
۱۶	۸-۲ DeepSentiPers
۱۶	۹-۲ جمع‌بندی
۱۷	۳ روش پیشنهادی
۱۷	۳-۱ مرحله اول: انتخاب ویژگی
۱۹	۳-۲ مرحله دوم: پیش پردازش
۱۹	۳-۳ مرحله سوم: معماری مدل
۲۱	۳-۴ مرحله چهارم: پس پردازش
۲۱	۳-۵ مرحله پنجم: شناسایی تولیدکنندگان نظرات نامناسب
۲۲	۳-۶ مرحله ششم: مدل BERT خلاقانه
۲۵	۳-۷ جمع‌بندی
۲۶	۴ نتایج تجربی
۲۶	۴-۱ مجموعه دادگان
۲۶	۴-۲ معیارهای اندازه‌گیری
۲۷	۴-۳ دقت مدل بر اساس معیارهای متفاوت روی شاخص کیفیت
۲۸	۴-۴ دقت مدل بر اساس معیارهای متفاوت روی شاخص ارزش نسبت به قیمت
۲۸	۴-۵ دقت مدل بر اساس معیارهای متفاوت روی شاخص اصالت

^۴Attention^۵Net Brand Reputation

۶-۴	دقت مدل بر اساس معیارهای متفاوت روی شاخص گارانتی	۲۸
۷-۴	دقت مدل بر اساس معیارهای متفاوت روی شاخص ابعاد	۲۹
۸-۴	دقت مدل بر اساس معیارهای متفاوت روی شاخص مغایرت کالا با کالای خریداری شده	۲۹
۹-۴	دقت مدل بر اساس معیارهای متفاوت روی شاخص رایحه/بو	۲۹
۱۰-۴	دقت مدل بر اساس معیارهای متفاوت روی شاخص تاریخ انقضا	۳۰
۱۱-۴	جمع بندی	۳۰

۵ نتیجه گیری

۱-۵	مسیرهای احتمالی برای کارهای آتی	۳۱
-----	---------------------------------	----

فصل ۱

مقدمه

۱-۱ تعریف مسئله

پردازش زبان طبیعی^۱ و تجزیه و تحلیل متن شامل استفاده از الگوریتم‌های یادگیری ماشین و هوش مصنوعی برای درک معنی اسناد متنی است. این اسناد می‌تواند تقریباً شامل هر چیزی باشد که شامل متن است: نظرات رسانه‌های اجتماعی، بررسی‌های آنلاین، پاسخ‌های نظرسنجی، حتی اسناد مالی، پزشکی، قانونی و نظارتی. در اصل، نقش یادگیری ماشینی و هوش مصنوعی در پردازش زبان طبیعی و تجزیه و تحلیل متن، بهبود، تسریع و خودکارسازی عملکردهای تحلیل متن و ویژگی‌های پردازش زبان طبیعی است که این متن بدون ساختار را به داده‌ها و بینش‌های قابل استفاده تبدیل می‌کند. مدل تجزیه و تحلیل احساسات^۲ برای تجزیه و تحلیل متن ترکیبی از تکنیک‌های پردازش زبان طبیعی و یادگیری ماشینی است تا نمرات احساس شده را به موجودات، موضوعات و دسته‌بندی‌ها در یک جمله یا عبارت اختصاص دهد.

تجزیه و تحلیل احساسات به تحلیلگران داده‌ها در شرکت‌های بزرگ کمک می‌کند تا افکار عمومی را بسنجند، تحقیقات ظریف بازار را انجام دهند، برند و شهرت محصول را زیر نظر بگیرند و تجربیات مشتری را بفهمند. علاوه بر این، شرکت‌های تجزیه و تحلیل داده‌ها اغلب با نظارت بر رسانه‌های اجتماعی با تجزیه و تحلیل احساسات این داده‌ها بینش مفیدی را به مشتریان خود ارائه دهند.

¹Natural Language Processing (NLP)

²Sentiment Analysis

یکی از بزرگ‌ترین چالش‌های دیجی‌کالا در مسیر پیشرفت، خودکار کردن فرایندهایی است که به‌صورت دستی و توسط کارشناس‌ها انجام می‌شود. برای مثال مدیریت نظرات کاربران یکی از آن فعالیت‌هایی است که در دیجی‌کالا توسط کارشناس‌ها تیم محتوای شرکت انجام می‌شود. روزانه صدها نظر در پلتفرم‌های مختلف دیجی‌کالا ثبت می‌شود که پیش از انتشار، باید آن‌ها را بررسی کنیم. اما این بررسی‌ها در یک فرایند کاملاً دستی و به کمک نیروهای انسانی انجام می‌شود. به همین دلیل، با رشد سریع دیجی‌کالا و افزایش تعداد کاربران آن، محدودیت‌هایی برای واحدهای عملیات انسانی به ویژه واحد بررسی نظرات کاربران ایجاد شده است. محدودیت‌های زمانی، هزینه‌های بالا و افزایش حجم کاری کارشناس‌ها این واحد، فقط بخشی از چالش‌هایی است که با آن مواجه شدیم. تا قبل از مهر ۱۳۹۹، تنها یک درصد از حجم نظرات کاربران توسط هوش مصنوعی بررسی می‌شد و بررسی ۹۹ درصد دیگر به عهده نیروهای انسانی بود که کاری سخت و زمان‌بر است. در حال حاضر دیگر نمی‌توانیم حجم بالایی از نظرات را فقط به کمک کارشناس‌ها، بررسی و منتشر کنیم. به همین دلیل در تیم AI دیجی‌کالا، تصمیم گرفتیم تا قابلیت‌های هوش مصنوعی را بیشتر از قبل در فرایندهای انسانی وارد کرده و از قابلیت خودکارسازی فرایندها برای رفع محدودیت‌ها و برطرف کردن چالش‌ها استفاده کنیم.

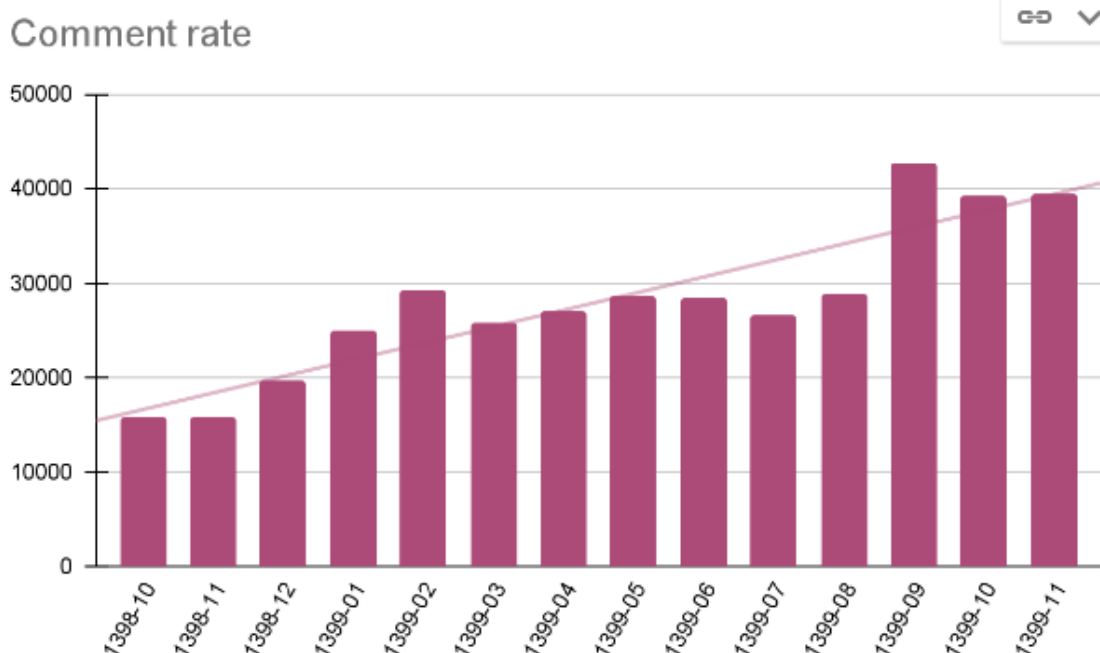
۲-۱ اهمیت موضوع

تجزیه و تحلیل احساسات از اهمیت بالایی برخوردار است زیرا به مشاغل اجازه می‌دهد تا احساسات مشتریان خود را نسبت به نام تجاری خود درک کنند. با بررسی احساسات پشت گفتگوها، بررسی‌ها^۳ و موارد دیگر، مشاغل می‌توانند تصمیمات بهتر و آگاهانه‌تری بگیرند. این ابزار مدیران محصول را قادر می‌سازد تا احساسات مشتریان را در کمپین‌های بازاریابی خود درک کنند. این یک عامل مهم در مورد شناخت محصول و نام تجاری، وفاداری مشتری، رضایت مشتری، موفقیت تبلیغات و تبلیغات و پذیرش محصول است. درک روانشناسی مصرف‌کنندگان می‌تواند به مدیران محصول کمک کند تا نقشه راه محصول خود را با دقت بیشتری تغییر دهند. [۱] [۲]

با رشد دیجی‌کالا و افزایش تعداد کاربران، میزان نظرات هم با افزایش قابل توجهی روبه‌رو شد. در نمودار زیر شاهد رشد چشمگیر تعداد نظرات از بهمن ماه ۱۳۹۸ تا دی ماه ۱۳۹۹ هستیم که رشدی ۲ تا ۳ برابری را نشان می‌دهد.

³Review

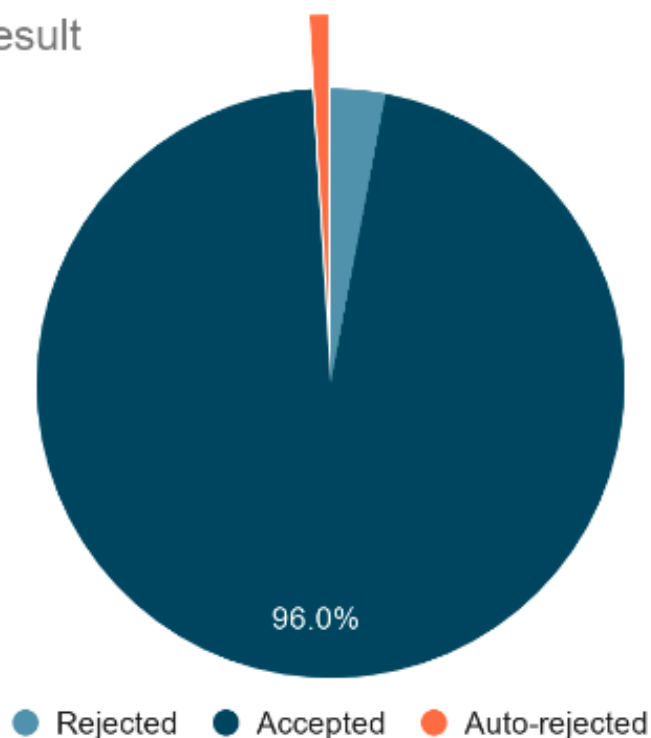
شکل ۱-۱: تعداد نظرات در ماه‌های مختلف



در این ماه‌ها، نظرات ثبت شده در دیجی کالا از ۱۵ هزار عدد نظر ثبت شده به ۴۰ هزار عدد رسیده است! باید بگوییم که در حال حاضر یعنی در سال ۱۴۰۰، حتی این عدد به ۵۰ هزار هم نزدیک شده است. اینجا بود که فهمیدیم، کارشناس‌ها واحد تیم محتوا دیگر نمی‌توانند به‌صورت دستی، این حجم از نظرات را مدیریت کنند. چون با همان سیستم قبلی باید شدت حدود ۳ برابر کار می‌کردند که این عملاً غیر ممکن است! از نظر ما، تنها راهکار برطرف کردن این چالش، خودکارسازی فرایندها به کمک هوش مصنوعی بود. اما قبل از اینکه، از خودکارسازی و استفاده از قابلیت‌های هوش مصنوعی صحبت کنیم؛ بهتر است که اول نگاهی به سیستم قبلی داشته باشیم

در مدل قبلی، ۹۹ درصد از نظرات توسط کارشناس‌ها مدیریت می‌شد. به این ترتیب که ۹۶ درصد آن‌ها تایید و ۳ درصد آن‌ها رد می‌شد. سهم هوش مصنوعی در مدیریت این نظرات، فقط یک درصد بود! یعنی فقط یک درصد از کل نظرات ثبت شده در دیجی کالا به‌صورت خودکار رد می‌شدند. برای درک بهتر این موضوع، لازم است تا نگاهی به نمودار زیر بیاندازیم:

Performance result



شکل ۱-۲: سهم مدل قبلی از نظرات

این مدل، یک شبکه عصبی عمیق بر پایه بلندواژه‌ی کوتاه حافظه^۴ است. شاید در ظاهر مدل ناکارآمدی نباشد، اما در کاهش چالش‌ها برای واحد مدیریت نظرات، خیلی هم موثر نبود. چون این مدل فقط می‌توانست بخشی از نظرات را «رد» کند. یعنی قابلیت بررسی نظرات برای «تایید» آن‌ها را نداشت! جالب است بدانید که باتوجه‌به به استانداردها و سیاست‌های تیم محتوای دیجی‌کالا، فقط ۴ درصد از کل نظرات رد می‌شوند. پس این مدل فقط می‌توانست روی ۴ درصد تاثیرگذار باشد که از این مقدار هم، طبق نمودار بالا، فقط یک درصد از نظرات را تشخیص داده و رد می‌کرد. به همین دلیل سهم چشمگیری در کاهش حجم کاری کارشناسان ایفا نمی‌کرد. پس به مدلی نیاز داشتیم که بتواند علاوه بر رد کردن نظرات، به تایید آن‌ها هم بپردازد.

⁴LSTM

۳-۱ اهداف تحقیق

در این پایان‌نامه هدف پیاده‌سازی مدل تجزیه و تحلیل احساسات نظرات دیجی‌کالا است. به همین منظور ابتدا روش‌هایی برای مدیریت تعداد زیاد نظرات به کمک هوش مصنوعی می‌پردازیم و سپس روش خلاقانه‌ای برای تحلیل احساسات نظرات ارائه خواهیم نمود.

۴-۱ ساختار پایان‌نامه

در فصل دوم به بررسی مفاهیم اولیه مرتبط با پایان‌نامه پرداخته و گریزی به روش‌های پیشین خواهیم داشت؛ سپس در فصل سوم به تشریح روش تکمیلی و خلاقانه‌ای که از آن در شرکت دیجی‌کالا بهره برده‌ایم، خواهیم پرداخت. در فصل چهارم به بررسی نتایج به دست آمده حاصل از روش خلاقانه پرداخته‌ایم و در نهایت در فصل پنجم کارها و مسیرهای احتمالی پیش رو را تبیین می‌نماییم.

فصل ۲

مفاهیم اولیه و کارهای پیشین

در این بخش از پایان نامه به معرفی اجمالی مفاهیم استفاده شده می پردازیم.

۱-۲ هوش مصنوعی

هوش مصنوعی یکی از جدیدترین رشته‌ها در علوم و مهندسی است و در حال حاضر طیف وسیعی از زیرشاخه‌ها (از حوزه‌های عمومی مانند یادگیری و ادراک^۱ تا حوزه‌های خاص مانند بازی کردن شطرنج، اثبات قضایای ریاضی و ...) را دربرمی گیرد. تعاریف مختلف و متنوعی از هوش مصنوعی وجود دارد که به طور کلی می توان آن‌ها را به چهار دسته‌ی زیر تقسیم بندی کرد.

۱. فکر کردن مانند انسان:

- خودکارسازی فعالیت‌هایی که ما با تفکر انسان مرتبط می کنیم. فعالیت‌هایی مانند تصمیم گیری، حل مسئله، یادگیری و

۲. فکر کردن عقلانی:

- مطالعه‌ی قوای ذهنی از طریق مدل‌های محاسباتی
- مطالعه‌ی محاسباتی که ادراک، استدلال و عمل را امکان پذیر می کند.

¹Perception

۳. عمل کردن مانند انسان:

- هنر ایجاد ماشین‌هایی که کارهایی را انجام می‌دهند که وقتی توسط انسان‌ها انجام می‌شود، نیازمند هوش‌مندی است.
- مطالعه‌ی چگونگی انجام دادن کارها با کامپیوتر که در حال حاضر (سال ۱۹۹۱) انسان‌ها آن کارها را بهتر انجام می‌دهند.

۴. عمل کردن عقلانی:

- هوش محاسباتی مطالعه‌ی طراحی نماینده^۲ های هوش‌مند است.

۲-۲ پردازش زبان طبیعی

پردازش زبان طبیعی به شاخه‌ای از علوم کامپیوتر، به‌طور مشخص‌تر به شاخه‌ی عمل کردن مانند انسان هوش مصنوعی که پیش‌تر به آن اشاره شد، مربوط می‌شود. این حوزه به کامپیوترها این توانایی را می‌دهد تا متن و کلمات محاوره‌ای را همان‌طور که انسان‌ها متوجه می‌شوند، بفهمند. پردازش زبان طبیعی، زبان‌شناسی محاسباتی (مدل‌سازی مبتنی بر قواعد زبان انسانی^۳) را با مدل‌های آماری، یادگیری ماشین و یادگیری عمیق ترکیب می‌کند. این فن‌آوری‌ها در کنار هم، کامپیوترها را قادر می‌سازند تا زبان انسان را به صورت متنی یا داده‌های صوتی پردازش کنند و معنای کامل و دقیق آن را با توجه به هدف و احساسات گوینده یا نویسنده درک کنند. پردازش طبیعی زبان نیروی پیش‌ران‌های برنامه‌های کامپیوتری‌ای هستند که متن‌ها را از یک زبان به زبان دیگر ترجمه می‌کنند (مترجم گوگل^۴)، به دستورات گفتاری پاسخ می‌دهند (مانند دستیار صوتی شرکت اپل^۵) و حجم زیادی از متن را به سرعت بی‌درنگ^۶ خلاصه می‌کنند. در زندگی معمولی با احتمال خوبی با کاربردهای پردازش زبان طبیعی روبه‌رو شده‌ایم. برخی از آن‌ها عبارتند از:

- سیستم‌های مکان‌یاب^۷

^۲agent

^۳rule-based modeling of human language

^۴Google Translate

^۵Siri

^۶Real-Time

^۷GPS

- دستیارهای دیجیتال
- نرم افزارهای تبدیل گفتار به متن
- ربات های گفت گو^۸
- خدمات مشتری
- و ...

۲-۳ وظایف پردازش زبان طبیعی

زبان انسان مملو از ابهاماتی است که نوشتن نرم افزاری که به طور دقیق معنای متن یا داده های صوتی را تعیین می کنند، سخت می کند:

- هم نام ها^۹
- هم آواها^{۱۰}
- کنایه ها
- اصطلاحات
- استعاره ها
- قواعد، استثناها و تنوع در ساختار جمله

موارد بالا فقط تعداد محدودی از بی نظمی ها^{۱۱}ی زبان انسان است که یادگیری آن برای انسان ها چندین سال طول می کشد اما برنامه نویس باید برنامه های کاربردی مبتنی بر زبان طبیعی را آموزش دهند تا از ابتدا به طور دقیق مفاهیم را شناسایی و درک کنند. بعضی از وظایف پردازش زبان طبیعی این است که متن و داده های صوتی انسان را به طوری تقسیم بندی کند تا به کامپیوترها کمک کند تا آنچه را که قرار است تجزیه کنند را بفهمند. برخی از این وظایف عبارتند از:

⁸Chat Bots

⁹Homonyms

¹⁰Homophones

¹¹Irregularities

• **تشخیص گفتار:** در این کاربرد، به کمک پردازش زبان طبیعی گفتار به نوشتار تبدیل می‌شود. تشخیص گفتار برای هر برنامه‌ای که از دستورات صوتی استفاده می‌کند یا به سوالات گفتاری پاسخ می‌دهد، مورد نیاز است. چیزی که تشخیص گفتار را به صورت ویژه‌ای چالش برانگیز می‌کند، نحوه‌ی صحبت افراد است مانند سرعت تکلم، قرینه‌های لفظی و معنوی، تاکید و لحن‌های مختلف، استفاده نادرست از قواعد زبان و ...

• **برچسب‌گذاری گفتار:** فرآیند تشخیص بخشی از گفتار از یک کلمه یا قسمتی از متن با استفاده از نحوه‌ی استفاده از آن و زمینه است. مثلاً در جمله‌ی « من این کتاب را خریدم»، خریدم فعل جمله است.

• **شناسایی موجودیت‌های نام‌گذاری‌شده:** کلمات یا عبارات را به عنوان موجودیت‌های مفید شناسایی می‌کند. مثلاً شناسایی این که «ناجا» یک سازمان و یا «علی» نام یک مرد است را برعهده دارد.

• **تجزیه و تحلیل احساسات:** پردازش طبیعی زبان در این کاربرد سعی می‌کند تا کیفیت‌های ذهنی^{۱۲} را از متن استخراج کند. کیفیت‌هایی ذهنی نظیر:

– نگرش‌ها

– احساسات

– کنایه

– سوءظن

– ابهامات و سردرگمی‌ها^{۱۳}

۲-۴ پیش پردازش متن

پیش پردازش متن یک مرحله مهم برای کارهای پردازش زبان طبیعی است. این کار متن را به شکل قابل هضم‌تری تبدیل می‌کند، به طوری که الگوریتم‌های یادگیری ماشین می‌توانند عملکرد بهتری داشته باشند. به طور کلی، ۳ پیش‌پردازش مهم وجود دارد:

¹²Subjective Qualities

¹³Confusions

- توکن‌سازی: به طور خلاصه، توکن‌سازی در مورد تقسیم رشته‌های متن به قطعات کوچکتر یا "توکن" است. پاراگراف‌ها را می‌توان به جملات و جملات را می‌توان به کلمات تبدیل کرد.
- نرمال‌سازی: هدف نرمال‌سازی این است که همه متن را در یک سطح بازی برابر قرار دهد، به عنوان مثال، تبدیل همه کاراکترها به حروف کوچک.
- حذف نویز: حذف نویز متن را پاک می‌کند، به عنوان مثال، فضاهای سفید اضافی را حذف می‌کند.

در زیر به چند مفهوم معروف پیش پردازش اشاره می‌کنیم.

۲-۴-۱ یافتن ریشه کلمات^{۱۴}

در پیش‌پردازش کلمات را به ریشه اصلی خود برمی‌گردانیم. یا حروف اضافه را از آن‌ها حذف می‌کنیم ریشه‌یابی یک ابتکاری است که انتها (یا بخشی از) کلمات را بریده و بنابراین، ممکن است کلمات خوب یا واقعی نباشند.

۲-۴-۲ ایست‌واژه^{۱۵}

کلمات کلیدی کلمات بسیار متداول در هر زبان هستند. برای مثال در فارسی حروف ربط، افعال پرتکرار مانند "است" و بسیاری دیگر از کلمات جزو ایست‌واژه‌ها محسوب می‌شوند که ما در پیش‌پردازش آن‌ها را حذف می‌کنیم.

۲-۵ مبدل‌ها

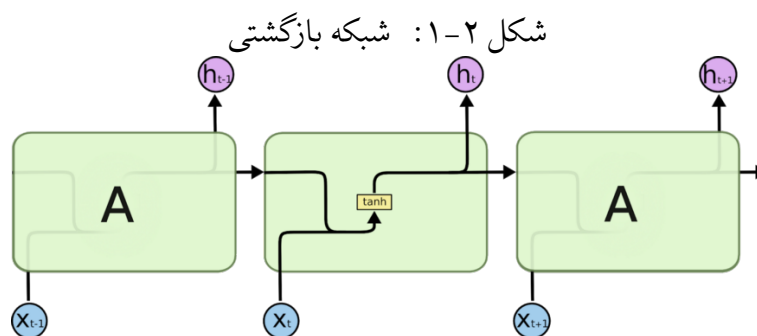
شبکه عصبی مبدل یک معماری جدید است که هدف آن حل مسائل دنباله به دنباله است که وابستگی‌های با فاصله‌های زیاد را به راحتی مدیریت می‌کند که اولین بار در مقاله [۳] در سال ۲۰۱۷ پیشنهاد شد.

¹⁴Lemmatization or Stemming

¹⁵Stopword

۲-۵-۱ آشنایی با شبکه‌های عصبی بازگشتی^{۱۶}

شبکه‌های عصبی بازگشتی جزو شبکه‌های پیشخور^{۱۷} هستند، که به مرور زمان راه‌اندازی می‌شوند. برخلاف شبکه‌های عصبی معمولی، شبکه‌های بازگشتی طوری طراحی شده‌اند که مجموعه‌ای از ورودی‌ها را بدون محدودیت اندازه از پیش تعیین شده دریافت کنند. "سری"ها در هر ورودی از آن دنباله رابطه‌ای با همسایه خود دارد یا تاثیری بر آنها دارد.



شبکه‌های عادی پیشرو چیزهایی را که در طول آموزش آموخته‌اند به خاطر می‌آورند. در حالی که شبکه‌های بازگشتی در حین آموزش به طور مشابه یاد می‌گیرند، علاوه بر این، هنگام تولید خروجی (ها) چیزهایی را که از ورودی‌های قبلی آموخته‌اند به خاطر می‌آورند. در واقع شبکه‌های بازگشتی حافظه از ورودی قبلی دارند و دنباله ورودی‌ها را متوجه می‌شوند.

۲-۵-۲ مدل توجه^{۱۸}

این مدل توجه از دو جهت با مدل کلاسیک متفاوت است. در مقایسه با مدل ساده مدل کلاسیک، در اینجا واحد رمزگذار^{۱۹} داده‌های بیشتری را به واحد رمزگشا^{۲۰} منتقل می‌کند. قبلاً تنها، آخرین حالت مخفی^{۲۱} قسمت رمزگذاری به رمزگشایی ارسال می‌شد، اما اکنون رمزگذار تمام حالت‌های پنهان (حتی حالت‌های میانی) را به رمزگشایی منتقل می‌کند. قسمت رمزگشایی قبل از تولید خروجی یک مرحله اضافی (در زیر توضیح داده شده) انجام می‌دهد. آخرین مرحله رمزگشایی به شرح زیر است:

¹⁶Recurrent Neural Networks

¹⁷Feed-Forward

¹⁸Attention

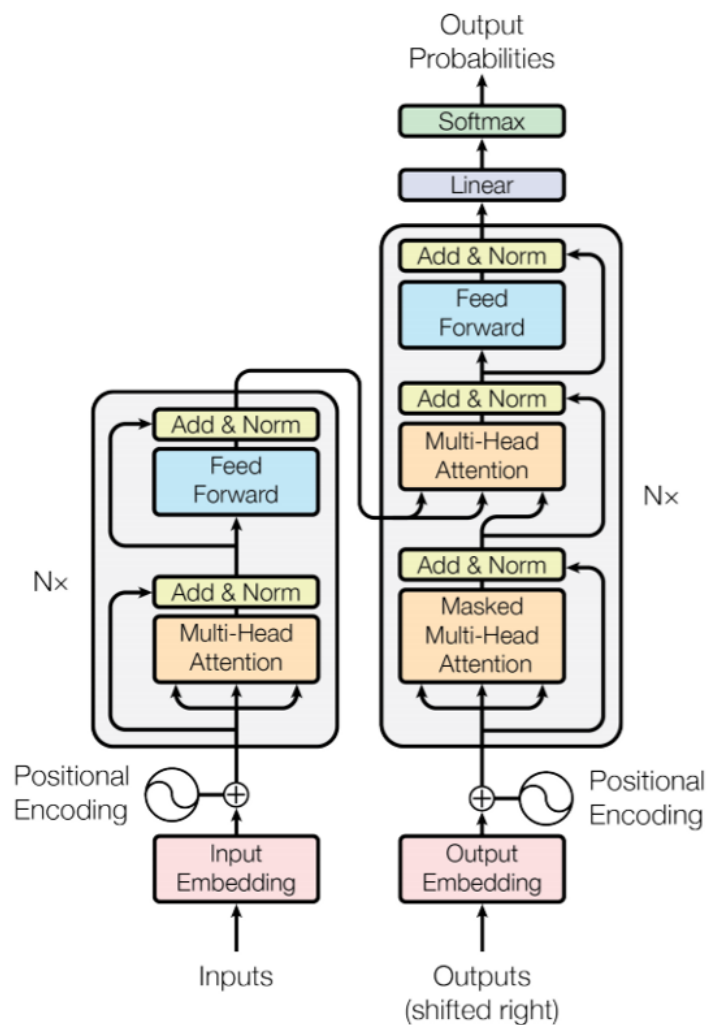
¹⁹Encoder

²⁰Decoder

²¹Hidden State

- هر حالت پنهانی را که دریافت کرده است بررسی می‌کند زیرا هر حالت پنهان رمزگذار بیشتر با کلمه خاصی از جمله ورودی مرتبط است.
- به هر حالت پنهانی امتیاز می‌دهد.
- در نهایت هر نمره در نمره softmax مربوط ضرب می‌شود، بنابراین حالت‌های پنهان را با امتیاز بالا تقویت می‌کند و حالت‌های پنهان را با امتیاز پایین از بین می‌برد.

شکل ۲-۲: معماری مدل مبدل [۳]



۶-۲ Google BERT

Bert^{۲۲} [۴] [۵] مقاله ای است که اخیراً توسط محققان Google AI Language منتشر شده است، که در بسیاری از مسائل پردازش زبان طبیعی پیشرفته‌ترین^{۲۳} نتایج را داشت. مهمترین نوآوری فنی BERT استفاده از آموزش دو طرفه مبدل، در مدل سازی زبان است. نتایج مقاله نشان می‌دهد که یک مدل زبانی که به صورت دو طرفه آموزش دیده است، می‌تواند نسبت به مدل‌های زبانی تک جهت، حس عمیق تری از موضوع متن و جریان آن زبان داشته باشد.

BERT از مبدل استفاده می‌کند و بر اساس سازوکار توجه روابط متنی بین کلمات (یا کلمات فرعی) را در یک متن می‌آموزد. در شکل وانیلی، مبدل شامل دو مکانیسم مجزا است - یک رمزگذار که ورودی متن را می‌خواند و یک رمزگشایی که یک پیش‌بینی برای ورودی تولید می‌کند. از آن‌جا که هدف BERT ایجاد یک مدل زبانی است، فقط مکانیزم رمزگذار لازم است. Bert از موفقیت بی نظیر در پردازش زبان طبیعی به کمک آموزش^{۲۴} به کمک مدل سازی ماسک زبان (MLM)^{۲۵} و پیش‌بینی جمله بعدی^{۲۶} (NSP) از موفقیت بی نظیری برخوردار است. مدل BERT به ما این امکان را می‌دهد که با استفاده از مدل ازپیش‌آموزش داده‌شده^{۲۷} آن را برای مساله‌ی مورد نظر خود تنظیم دقیق^{۲۸} کنیم.

۱-۶-۲ توصیف مدل

Bert یک مدل مبدل است که بر روی یک مجموعه‌نوشته‌ها^{۲۹} ی بزرگی از داده‌های چند زبانه به صورت خودنظارتی^{۳۰} پیش‌آموزش می‌شود. این به این معنی است که فقط بر روی متون خام پیش‌آموزش انجام می‌گیرد، بدون این که توسط انسان برچسب گذاری شوند و با یک فرایند خودکار تولید ورودی‌ها و برچسب‌ها انجام می‌شود. در نگاهی دقیق تر، این کار با دو هدف مورد توجه قرار می‌گیرد.

• مدل سازی زبان ماسک (MLM): با دریافت یک جمله، مدل به طور تصادفی ۱۵٪ از کلمات

²²Bidirectional Encoder Representations from Transformers

²³state-of-the-art

²⁴Training

²⁵Masked-Language Modeling

²⁶Next Sentence Prediction

²⁷Pre-trained

²⁸Fine-tune

²⁹Corpus

³⁰Self-supervised

را در ورودی ماسک می‌کند و سپس کل جمله ماسک را از طریق مدل اجرا کرده و باید کلمات ماسک را پیش بینی کند. تفاوت این مدل با شبکه‌های عصبی بازگشتی که معمولاً کلمات را یکی پس از دیگری می‌بینند، این اجازه را می‌دهد تا مدل برای یادگیری، یک نمایش دو طرفه از جمله را بیاموزد.

- پیش بینی بعدی جمله (NSP): مدل‌ها دو جمله ماسک را به عنوان ورودی‌ها در طول پیش‌آموزش به هم پیوند می‌زنند. آموزش بر این اساس انجام می‌شود که آیا دو جمله در متن اصلی در کنار یک‌دیگر قرار داشته‌اند یا خیر. داده‌های آموزش نیز بر همین اساس ساخته شده و به عنوان ورودی مدل در نظر گرفته می‌شوند.

۲-۶-۲ شهرت نام تجاری^{۳۱}

رسانه‌های اجتماعی^{۳۲}، علم و فناوری سطح زندگی مردم و جامعه را افزایش داده است. امروزه رسانه‌های اجتماعی نقش بسیار مهمی را با پیشرفت صنایع و شرکت‌ها ایفا می‌کنند. این امر به عنوان یک عامل اساسی توسط شرکت‌ها و همچنین مردم در نظر گرفته شده است. به نحوی که همه مردم به هر طریقی به شبکه‌های اجتماعی متصل هستند. ترکیب فناوری و روابط اجتماعی باعث شده است که افراد بتوانند اطلاعات را با یک‌دیگر به اشتراک بگذارند. رسانه‌های اجتماعی در ده سال گذشته به عنوان یک بستر غالب برای به اشتراک گذاری اطلاعات ظاهر شده است. تجزیه و تحلیل احساسات به کاربر این امکان را می‌دهد تا احساسات، اعتقادات و دیدگاه‌ها را در جهان بزرگ‌تر نشان دهد. علم و فناوری کامپیوتر به اما این امکان را می‌دهد تا بتوانیم با دقت نسبتاً خوبی تمایل احساسات و نظرات مردم در شبکه‌های اجتماعی را محاسبه کنیم به شکلی که بتوانیم هر متن به اشتراک گذاشته شده را از نظر مثبت یا منفی بودن ارزیابی کنیم. حال برای بدست آوردن شهرت نام تجاری یک نشان تجاری سه مرحله داریم.

- استخراج پیام‌های مربوط به برند مربوط

در این مرحله با توجه به نوع تحقیق و پژوهشی که در حال انجام است، اولین نیاز برای محاسبه شهرت نام تجاری یک برند استخراج پیام‌هایی است که مردم در یک یا چند رسانه اجتماعی به اشتراک گذاشته‌اند. در اینجا تمامی تحلیل‌های ما بر روی متن پیام‌هاست.

³¹Net Brand Reputation

³²Social Media

• محاسبه تمایل هر پیام مربوط

در این مرحله به کمک مدل‌های شبکه عصبی آموزش داده شده می‌توانیم تمایل هر پیام استخراج شده مربوط به نشان تجاری مورد نظر را بدست آوریم. به طوری که به هر پیام بر اساس بار معنایی به هر پیام یک برچسب پیش‌بینی مثبت یا منفی می‌زنیم.

• محاسبه شهرت نام تجاری یک نشان تجاری

در این مرحله بعد از بدست آوردن پیام‌های مربوط به برند مورد نظر به همراه برچسب‌های مثبت یا منفی به معنای بار احساسات آن‌ها، معیاری تعریف می‌کنیم که عددی بین -۱ و +۱ به ما بدهد (+۱ به معنای شهرت مثبت بالا و -۱ به معنای شهرت منفی بالا) و آن را عدد شهرت نام تجاری نشان تجاری مورد نظرمینامیم. این رابطه به شکل زیر است.

$$NBR = \frac{\#Positive - \#Negative}{\#Positive + \#Negative} \quad (۱-۲)$$

همانطور که در رابطه ۱-۲ می‌بینیم تعداد پیام‌های با بار مثبت منهای تعداد پیام‌های با بار منفی شده و سپس بر جمع آن‌ها تقسیم شده است. معیار بدست آمده را شهرت نام تجاری برندی در نظر می‌گیریم که پیام‌های استخراج شده مربوط به آن اند.

۷-۲ ParsBert

ParsBERT [۶] مدلی تک زبانه برای Bert در زبان فارسی پیشنهاد می‌کند که عملکرد برتر را در مقایسه با معماری‌ها و مدل‌های چند زبانه دارد. همچنین، از آنجا که میزان داده‌های موجود برای وظایف پردازش زبان طبیعی به زبان فارسی بسیار محدود است، یک مجموعه داده عظیم برای کارهای مختلف پردازش زبان طبیعی و همچنین پیش آموزش این مدل جمع‌آوری شده است. ParsBERT نتایج بهتری را در همه مجموعه‌داده‌گان نسبت به BERT چند زبانه و سایر کارهای قبلی در تجزیه و تحلیل احساسات و دسته بندی متن بدست آورده است.

۸-۲ DeepSentiPers

در این مدل [۷] ترکیبی از شبکه LSTM [۸] [۹] و شبکه CNN [۱۰] [۱۱] برای دسته‌بندی جملات ارائه می‌شود.

دست‌آورد اصلی مجموعه‌های یادگیری عمیق پیشنهادی هر دو نمرات $f1$ مناسب را کسب کرده‌اند. اما، B-LSTM به دلیل واحدهای حافظه داخلی که قادر به کنترل وابستگی‌های طولانی مدت هستند، برای طبقه‌بندی احساسات بهتر عمل می‌کند. تحقیقات همچنین این ایده را تأیید می‌کند که معماری CNN برای محدوده پردازش تصویر مناسب است نه طبقه‌بندی متن. بنابراین، تحقیقات بیشتری می‌توان در LSTM دو طرفه در ارتباط با تجزیه و تحلیل احساسات انجام داد.

۹-۲ جمع‌بندی

با توجه به دو بخش آخر این فصل و بیان مزایا و معایب هر کدام از مدل‌های DeepSentiPers و ParsBert و همچنین تحقیق در مورد آن‌ها ما از مدل ParsBert استفاده کرده و آن را تا حد بسیار زیادی بهبود داده‌ایم.

فصل ۳

روش پیشنهادی

همیشه یکی از چالش‌های مهم در پیاده‌سازی مدل‌های مختلف به خصوص روی کلان‌داده^۱ انتخاب مدلی با کمترین پیچیدگی است. به همین دلیل در طراحی سیستم جدید، به سراغ مدل‌های سنگین‌تر مانند BERT نرفتیم و تلاش کردیم تا همان مدل قبلی بر پایه LSTM را بهبود داده و قابلیت‌های جدیدی مثل فرایند تایید خودکار را به آن اضافه کنیم تا به یک مدل به اصطلاح بی‌درنگ برسیم.

اما اضافه کردن فرایند تایید خودکار نظرات کار ساده‌ای نیست. چون می‌تواند هزینه‌های مستقیم و غیرمستقیم زیادی را به شرکت تحمیل کند. پس باید یک پس‌پردازش^۲ هم برای کاهش خطاهای احتمالی در نظر می‌گرفتیم که در ادامه، بیشتر از آن صحبت می‌کنیم.

حال در این قسمت نیاز است تا نگاهی دقیق به ساختار کلی مدل بیاندازیم:

۳-۱ مرحله اول: انتخاب ویژگی

انتخاب ویژگی^۳ همان فرایند انتخاب زیرمجموعه‌ای از ویژگی‌های مرتبط برای ساخت مدل است. ما در این فرایند، چند ویژگی مشخص را به عنوان ورودی مدل انتخاب کردیم که عبارت‌اند از:

• محتوای نظر

¹Bigdata

²Post-Processing

³Feature Selection

- نقاط ضعف و قوت نظر

- عنوان نظر

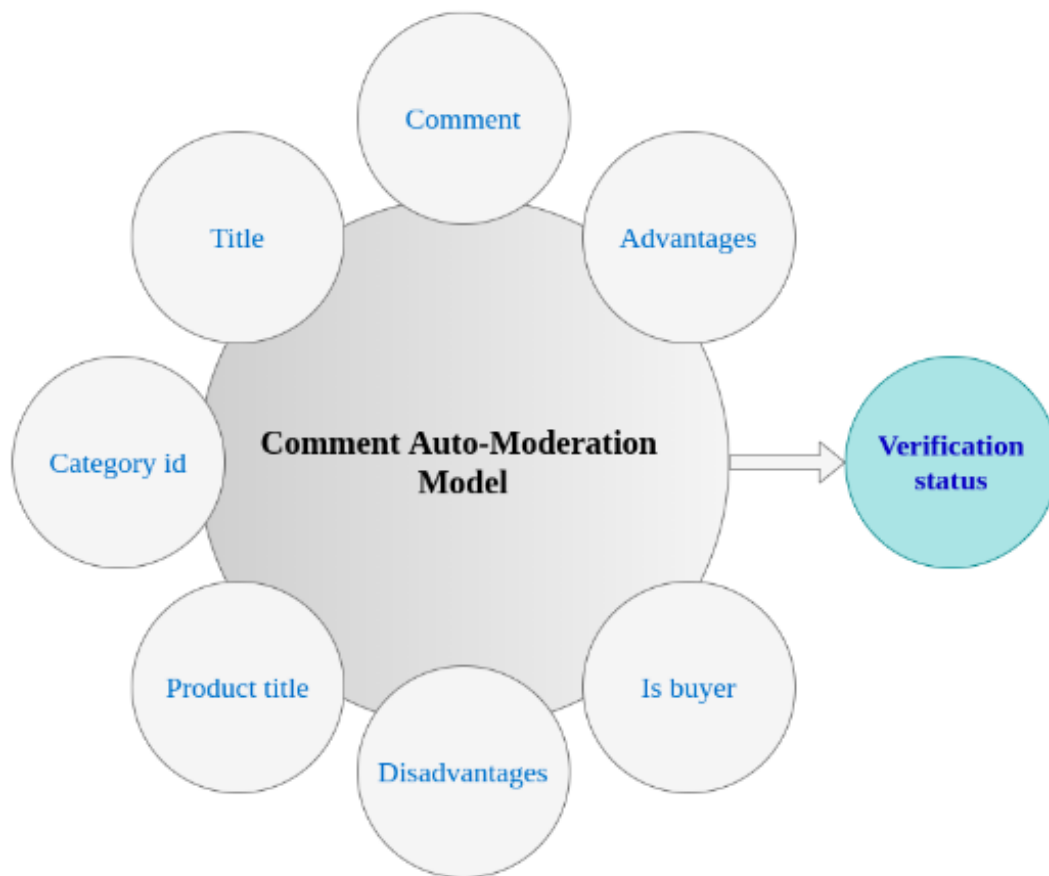
- گروه کالایی کالا

- عنوان کالا

- خریدار بودن یا نبودن شخصی که نظر را ثبت کرده

در نهایت خروجی، میزان احتمال تایید شدن نظر ورودی را تخمین می‌زند.

شکل ۳-۱: مدل تایید خودکار



با پشت سر گذاشتن مرحله اول یعنی انتخاب ویژگی، به مرحله دوم می‌رسیم. مرحله‌ای که باید از پیش‌پردازش بیش‌تر بگوییم..

۲-۳ مرحله دوم: پیش پردازش

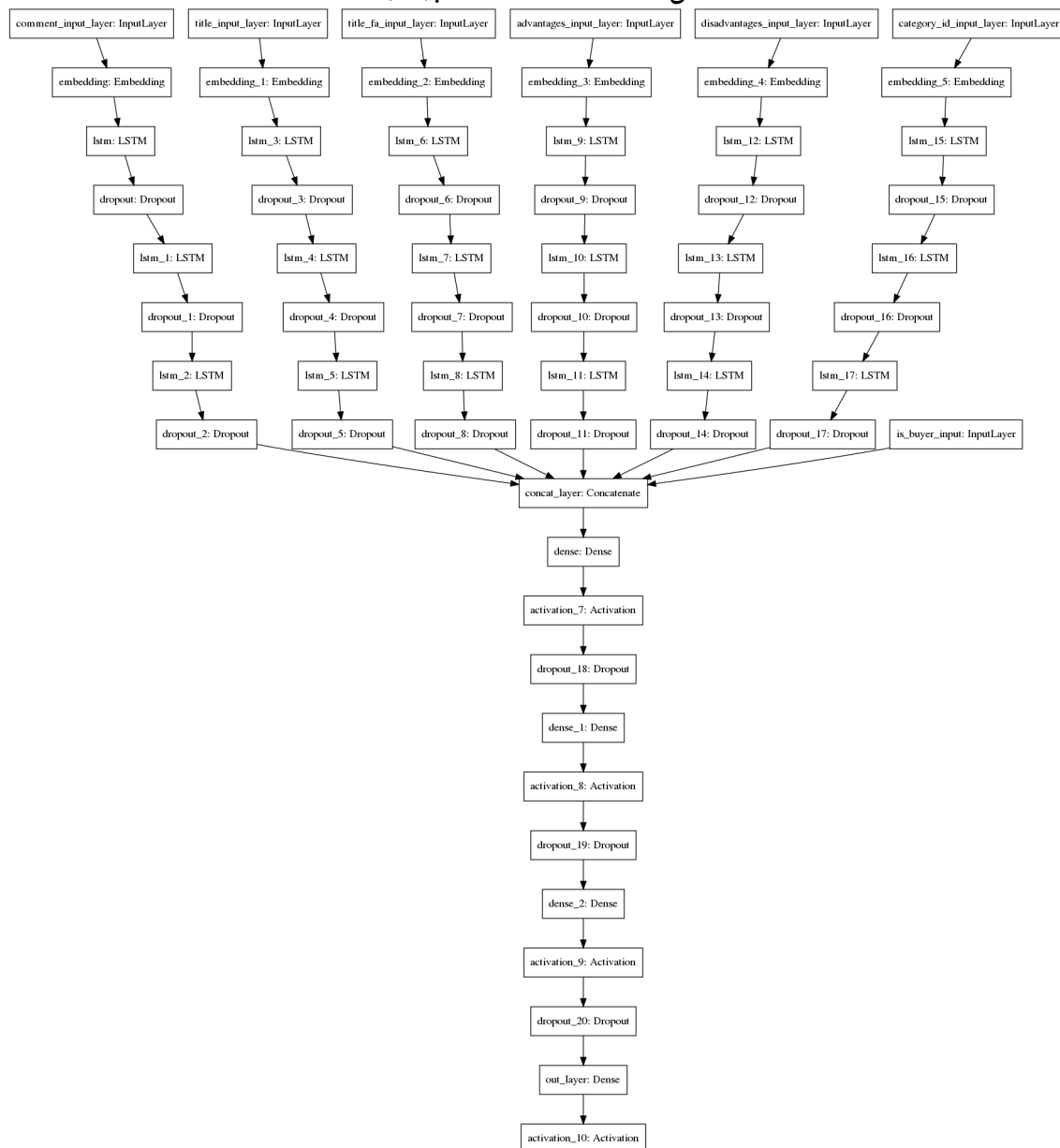
در مدل‌های متنی معمولاً یک مرحله پیش‌پردازش وجود دارد. پیش‌پردازی که ما روی متن‌های نظرات انجام می‌دهیم از دو بخش تشکیل شده است. در بخش اول که به اصطلاح به آن Normalizer گفته می‌شود، سعی می‌کنیم از پیچیدگی‌های اضافی ورودی کم کنیم. برای مثال اعداد انگلیسی به فارسی تبدیل می‌شود، علائم نگارشی از کلمات جدا می‌شود، «ی» عربی به فارسی تبدیل می‌شود و... اما بخش دوم یا Tokenizer، کمی مفصل‌تر است. در نظر بگیرید که هر ویژگی متنی توزیع جداگانه‌ای دارد و یکی از ویژگی‌های ما مثل «گروه کالایی محصول» هم متنی نیست. برای این مسئله به طور خلاصه، هر کدام از ویژگی‌های متنی به یک Tokenizer و سپس Indexer وصل می‌شود که در فرایند Tokenizing، بخش Indexing درست می‌شود که این بخش هم به Embedding مدل وصل می‌شود. اما سوالی که به وجود می‌آید، این است که ویژگی عنوان کالایی محصول یا همان Category ID را چگونه به مدل دهیم؟ همان‌طور که می‌دانید اولین راه‌حل برای ویژگی‌های چند مقداری یا Categorical، استفاده از الگوریتم One-Hot Encoder است؛ اما در دیجی‌کالا تعداد همه‌ی عنوان‌های کالا بسیار زیاد است و این کار خیلی منطقی نیست. راه‌حلی که ما استفاده کردیم این بود که به ویژگی «عنوان کالا» به عنوان یک متن نگاه کرده و به یک بردار ۶۴ تایی نگاشت^۴ می‌کنیم؛ این بردار در طول فرایند یادگیری آموزش می‌بیند.

۳-۳ مرحله سوم: معماری مدل

ساختار مدل پیشنهادی را می‌توانید در شکل زیر ببینید:

^۴mapping

شکل ۳-۲: معماری مدل پیشنهادی



همان‌طور که در بلوک دیاگرام مدل قابل مشاهده است، هرکدام از بخش‌های مبتنی بر متن ورودی، پس از گذر از لایه Embedding، وارد یک زیر شبکه‌ی شامل سه لایه LSTM می‌شود. پس خروجی هر زیر شبکه LSTM در کنار هم قرار می‌گیرند و آخرین ورودی که خریدار بودن یا نبودن کاربر را تعیین می‌کند و تک نوروں است هم در کنار سایر خروجی‌های مختلف قرار می‌گیرد.

خروجی شبکه مقدار احتمال تایید شدن نظر ورودی است. برای تعیین وضعیت نظرات از روی خروجی شبکه، دو مقدار Threshold برای تایید و رد، در نظر گرفته شده است که نظراتی که احتمال

تایید آن‌ها بین این دو مقدار قرار می‌گیرد، به معنای عدم اطمینان مدل در تصمیم‌گیری هستند و برای بررسی بیشتر برای کارشناس‌ها ارسال می‌شوند.

۴-۳ مرحله چهارم: پس پردازش

در مرحله پس‌پردازش به دنبال کاهش مثبت‌های کاذب یا False Positive هستیم. در واقع در این مرحله، می‌خواهیم نظرات نامناسبی که به هیچ وجه نباید تایید شوند را مشخص کنیم. برای این کار از Tokenizer یک مدل BERT استفاده کردیم تا ریشه عبارت‌های نامناسب را پیدا کنیم. سپس این نظرات در چند فهرست قرار می‌گیرند که این فهرست‌ها شامل کلمات نامناسب و سایر عبارت‌های غیرقابل انتشار هستند. سپس این کلمات بررسی شده و اگر تایید شوند، مدل تایید نهایی را ارایه می‌دهد. برای مثال در کلمه «نشستندی»، توکنایزر این کلمه را به صورت «نشست + ندی» تقسیم می‌کند و اگر کلمه «نشست» در فهرست سیاه باشد، تمام مشتقات این کلمه را تا حد خوبی شناسایی کرده و نظراتی که دارای این مشتقات هستند را مستقیماً برای کارشناس ارسال می‌کند.

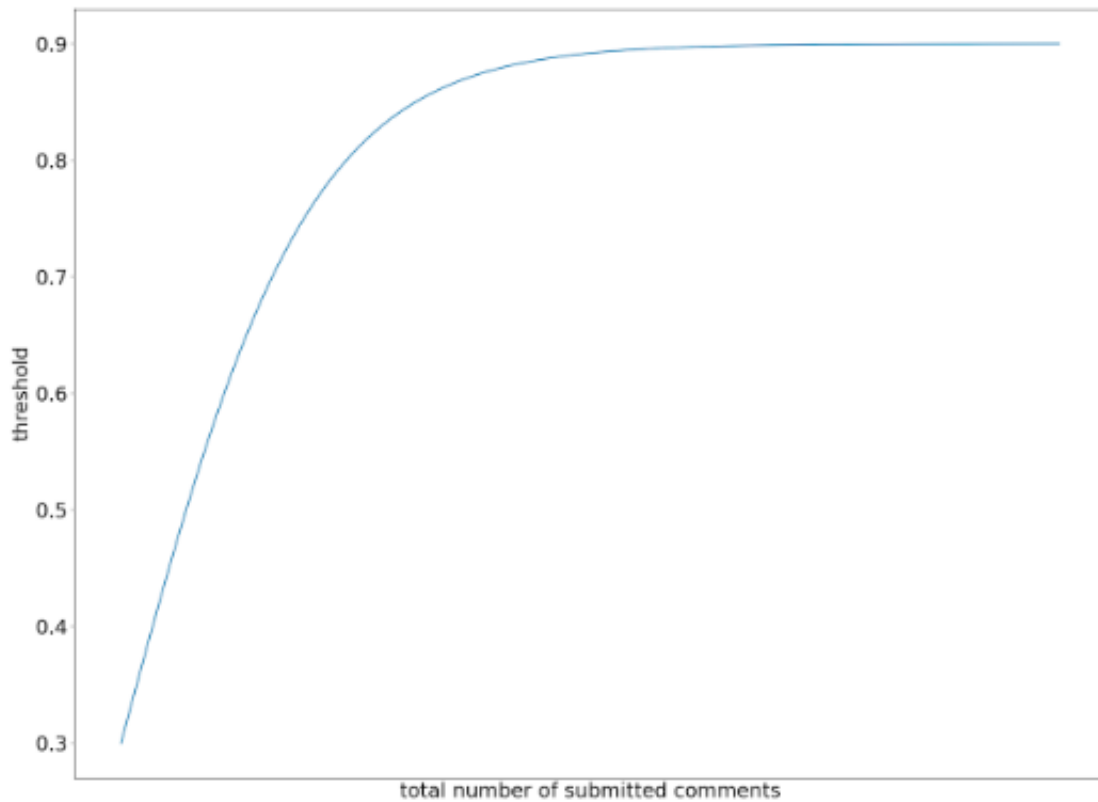
۵-۳ مرحله پنجم: شناسایی تولیدکنندگان نظرات نامناسب

یکی از مشکلاتی که ما در بخش نظرات با آن روبه‌رو هستیم، این است که بعضی کاربران به دلایل مختلف مثل دریافت امتیاز دیجی‌کلاب، تخریب کالای رقیب و... به تولید تعداد زیادی نظر می‌پردازند. برای حل این مشکل معیاری به عنوان Spammer بودن تعریف می‌کنیم. معیاری که توسط آن، احتمال Spammer بودن هر کاربر مشخص می‌شود، به صورت زیر تعریف شده است:

$$user\ performance = \frac{number\ of\ verified\ comments}{total\ number\ of\ submitted\ comments} \quad (۱-۳)$$

پس مقدار بازدهی هر کاربر در بازه زمانی مشخصی محاسبه شده و باتوجه‌به نمودار زیر، وضعیت Spammer بودن آن تعیین می‌شود.

شکل ۳-۳: نمودار نسبت تعداد کل نظرات به نظرات تایید شده



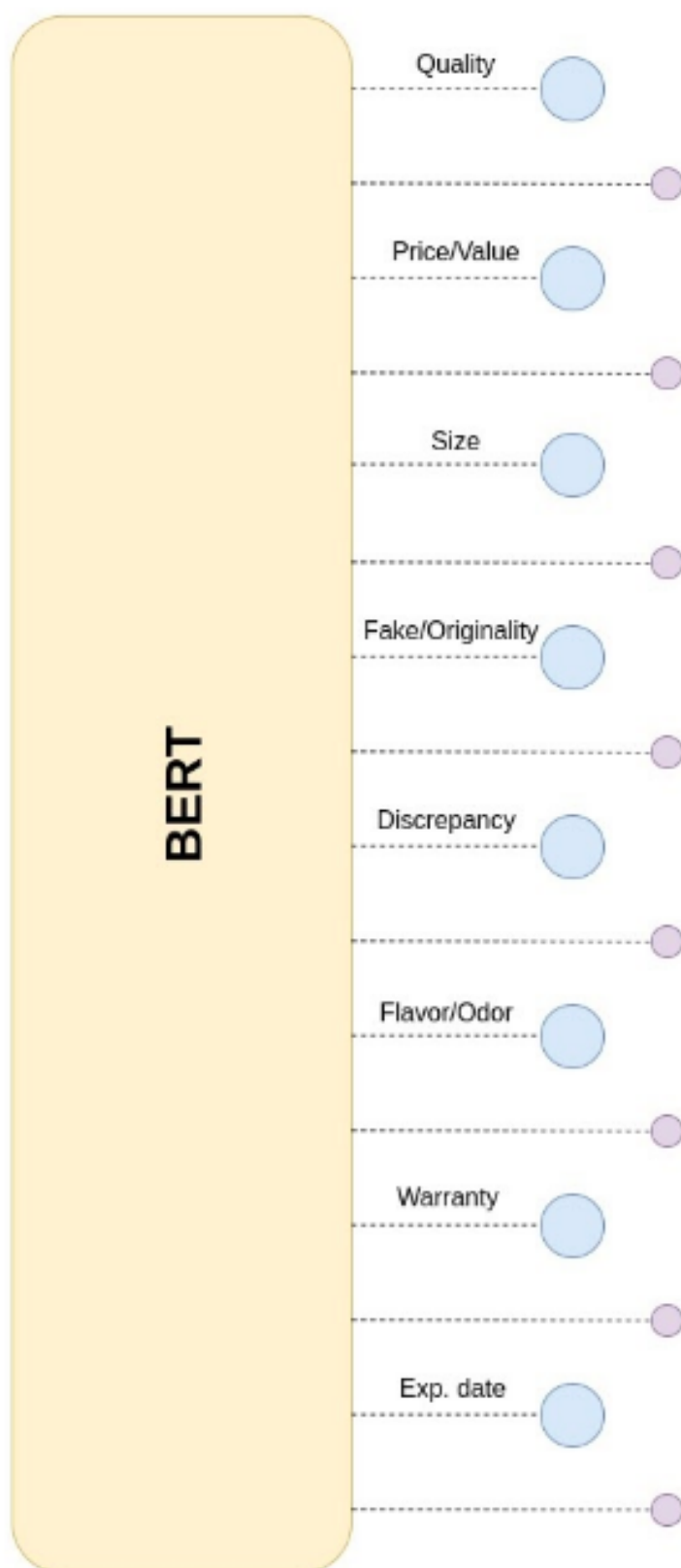
نمودار بالا مقدار Threshold برای تصمیم‌گیری در مورد Spammer بودن هر کاربر را تعیین می‌کند. مقدار این Threshold به صورت داینامیک و با توجه به تعداد کل نظرات ثبت شده توسط هر کاربر در بازه زمانی مورد نظر به دست می‌آید. طبق نمودار، هرچه تعداد کل نظرات ثبت شده توسط کاربر بیشتر باشد، بازدهی مورد انتظار از آن کاربر هم بیشتر می‌شود. در نهایت اگر مقدار بازدهی کاربر مورد نظر کمتر از Threshold آن کاربر باشد، به عنوان Spammer شناسایی شده و برای کارشناس به منظور بررسی دقیق‌تر فرستاده می‌شود. غیر خطی بودن این نمودار به این دلیل است که با تعداد نظرات کم، Threshold بالاتری در نظر گرفته شود و شناسایی کمتری انجام شود.

۳-۶ مرحله ششم: مدل BERT خلاقانه

برای بررسی احساسات و معنا^۵ ی نظرات از یک سری شاخص استفاده می‌کنیم:

⁵Intent sentiment analysis

شکل ۳-۴: مدل پیشنهادی BERT در حالت جدید



این intentها عبارتند از:

۱. کیفیت
۲. ارزش خرید نسبت به قیمت
۳. ابعاد
۴. اصالت کالا
۵. مغایرت کالا با کالای خریداری شده
۶. طعم/بو
۷. گارانتی
۸. تاریخ انقضا

در حالت کلی برای هر یک از intentها سه sentiment داریم که عبارتند از حالت‌های مثبت، منفی و خنثی. با توجه به ۸ intentای که در بالا بدان اشاره شد، اگر بخواهیم هر کدام را به صورت مجزا در نظر بگیریم، ۲۴ خروجی خواهیم داشت (زیرا ۸ شاخص داریم که هر کدام ۳ حالت مثبت، منفی و یا خنثی را می‌توانند داشته باشند) که اولاً خروجی مدل شبکه را بسیار سنگین و پراکنده^۶ می‌کند و ثانياً شبکه به درستی آموزش داده نمی‌شود. از همین رو، روش خلاقانه‌ای که در این مدل از آن استفاده کردیم، پیش‌بینی intentها از روی متن و پس از آن تشخیص احساسات از روی آن شاخص بود. به بیان دیگر، اول بررسی می‌شود که کدام شاخص‌ها در متن به کار رفته‌اند؛ برای مثال، ابتدا در نظر بررسی می‌شود که آیا کلمه‌ای به کار رفته است که به مفهوم کیفی یک کالا اشاره کند یا خیر و سپس در صورتی که جواب مثبت باشد، بررسی می‌شود که آیا آن کلمه مثبت، منفی و یا خنثی بوده است. استفاده از این روش ساده و خلاقانه، باعث بهبود خروجی شبکه که باعث افزایش سرعت و دقت مدل نهایی می‌شود.

با توجه به مطالبی که در بند قبل به آن اشاره کردیم، پروژه‌ی حال حاضر یادگیری ماشین چند وظیفه‌ای^۷ است؛ زیرا دو تا کار تحلیل احساسات و تشخیص شاخص‌ها به صورت هم‌زمان در حال اجرا هستند. یادگیری چندوظیفه‌ای زیر مجموعه‌ای از یادگیری ماشین است که در آن چندین کار یادگیری هم‌زمان حل

^۶sparse

^۷multi task

می‌شود، در حالی که از نقاط اشتراک و تفاوت بین وظایف استفاده می‌شود. این می‌تواند باعث بهبود کارایی یادگیری و دقت پیش‌بینی برای مدل‌های خاص وظیفه شود، در مقایسه با آموزش مدل‌ها به‌طور جداگانه. نسخه‌های اولیه یادگیری چندوظیفه‌ای «اشاره» نامیده می‌شدند. [۱۲] [۱۳] [۱۴] به کارگیری تحلیل احساسات و تشخیص شاخص‌ها در کنار یک‌دیگر به ما این امکان را می‌دهد تا این شبکه‌ها به یک‌دیگر کمک کنند تا یادگیری بهتری در کل داشته باشیم.

برای اندازه‌گیری دقت مدل باید از یک تابع هزینه^۸ استفاده کنیم. تابع هزینه‌ای که برای مدل در نظر گرفته‌ایم به شکل زیر است:

$$loss = intent\ loss + optimized\ sentiment\ loss \quad (۲-۳)$$

$$optimized\ sentiment\ loss = intent\ label \times sentiment\ loss \quad (۳-۳)$$

در فصل بعدی به بررسی نتایج به دست‌آمده با استفاده از این روش خواهیم پرداخت.

۷-۳ جمع‌بندی

همان‌طور که در بالاتر توضیح داده شد، ما با استفاده از تعداد شاخص‌های محدود برای تشخیص بودن یا نبودن یک موضوع خاص در نظرات، توانستیم از مقدار پراکندگی و تنک بودن داده‌ها بکاهیم و با توجه به حجم زیاد داده‌ها، فرآیندها را با کارایی بسیار بالاتری انجام دهیم و هم‌چنین با استفاده از مدل خودکار رد یا تایید نظرات، توانسته‌ایم بر کیفیت داده‌های خود بیفزایی.

^۸loss function

فصل ۴

نتایج تجربی

در این فصل به بررسی نتایج به دست آمده حاصل از روش پیشنهادی که در فصل گذشته به طور مفصل راجع به آن صحبت کردیم، می پردازیم.

۴-۱ مجموعه دادگان

در این قسمت مجموعه دادگان آموزش و آزمون مدل های پیشنهادی معرفی می شوند و پیش پردازش های انجام شده روی هریک آورده شده است. همان طور که در قسمت قبل یعنی مدل پیشنهادی صحبت شد، در آن مدل ما تعدادی شاخص داریم مانند کیفیت کالا. حال در مجموعه دادگان آموزش اپراتورهای دیجی کالا حدود ۲۵ میلیون کامنت را برچسب گذاری کرده اند و از آن در مدل خود استفاده نموده ایم. اندازهی مجموعه دادگان نقش بسیار مهمی در دقت مدل به دست آمده دارد.

۴-۲ معیارهای اندازه گیری

برای اندازه گیری دقت مدل، روش های گوناگونی وجود دارد که ابتدا سه تا از مهم ترین معیارها را معرفی کرده و در قسمت بعدی به مقایسه ی نتایج به دست آمده از دقت هر کدام از روش ها برای همه ی شاخص ها خواهیم پرداخت.

- معیار Precision برابر است با تقسیم تعداد مواردی که توسط مدل درست تشخیص داده شده است بر تعداد مواردی که واقعاً درست هستند، درست تشخیص داده شده‌اند. از همین رو فرمول این معیار به شیوه‌ی زیر محاسبه می‌شود.

$$recall = \frac{true\ positive}{true\ positive + false\ positive} \quad (۱-۴)$$

- معیار Recall برابر است با تقسیم تعداد مواردی که توسط مدل درست تشخیص داده‌اند شده بر تعداد کل مواردی که توسط مدل ایجاد شده‌اند. به بیان دیگر از تعداد کل موارد صحیح، چه تعداد را توانسته‌ایم درست حدس بزنیم. از همین رو فرمول این معیار به شیوه‌ی زیر محاسبه می‌شود.

$$precision = \frac{true\ positive}{true\ positive + false\ negative} \quad (۲-۴)$$

- معیار F1 در دو مورد قبل، Precision و Recall را معرفی و به صورت دقیق تعریف کردیم. حال معیار F1 Measure یک نوع میانگین بین این دو است که به شکل زیر محاسبه می‌شود.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (۳-۴)$$

۳-۴ دقت مدل بر اساس معیارهای متفاوت روی شاخص کیفیت

شکل ۴-۱: مدل بر اساس معیارهای متفاوت روی شاخص کیفیت

(Macro) Avg	Body	Adv/disadv	Positive	Negative	Neutral
Precision	0.88	0.97	0.95	0.81	0.63
Recall	0.85	0.96	0.95	0.87	0.57
F1	0.86	0.97	0.95	0.84	0.60

۴-۴ دقت مدل بر اساس معیارهای متفاوت روی شاخص ارزش نسبت به قیمت

شکل ۴-۲: مدل بر اساس معیارهای متفاوت روی شاخص ارزش نسبت به قیمت

(Macro) Avg	Body	Adv/disadv	Positive	Negative	Neutral
Precision	0.90	0.95	0.96	0.87	0.72
Recall	0.94	0.97	0.93	0.88	0.80
F1	0.92	0.96	0.94	0.87	0.76

۵-۴ دقت مدل بر اساس معیارهای متفاوت روی شاخص اصالت

شکل ۴-۳: مدل بر اساس معیارهای متفاوت روی شاخص اصالت

(Macro) Avg	Body	Adv/disadv	Positive	Negative	Neutral
Precision	0.87	0.90	0.85	0.87	0.54
Recall	0.75	0.72	0.84	0.85	0.57
F1	0.80	0.78	0.84	0.86	0.55

۶-۴ دقت مدل بر اساس معیارهای متفاوت روی شاخص گارانتی

شکل ۴-۴: مدل بر اساس معیارهای متفاوت روی شاخص گارانتی

(Macro) Avg	Body	Adv/disadv	Positive	Negative	Neutral
Precision	0.89	0.89	0.87	0.80	0.39
Recall	0.85	0.92	0.82	0.88	0.34
F1	0.87	0.90	0.84	0.83	0.36

۷-۴ دقت مدل بر اساس معیارهای متفاوت روی شاخص ابعاد

شکل ۴-۵: مدل بر اساس معیارهای متفاوت روی شاخص ابعاد

(Macro) Avg	Body	Adv/disadv	Positive	Negative	Neutral
Precision	0.87	0.89	0.90	0.82	0.72
Recall	0.92	0.91	0.81	0.71	0.84
F1	0.89	0.90	0.85	0.76	0.78

۸-۴ دقت مدل بر اساس معیارهای متفاوت روی شاخص مغایرت کالا با کالای خریداری شده

شکل ۴-۶: مدل بر اساس معیارهای متفاوت روی شاخص مغایرت کالا با کالای خریداری شده

(Macro) Avg	Body	Adv/disadv	Positive	Negative	Neutral
Precision	0.86	0.76	0.88	0.90	0.77
Recall	0.86	0.88	0.02	0.98	0.70
F1	0.86	0.81	0.04	0.94	0.73

۹-۴ دقت مدل بر اساس معیارهای متفاوت روی شاخص رایحه/بو

شکل ۴-۷: مدل بر اساس معیارهای متفاوت روی شاخص رایحه/بو

(Macro) Avg	Body	Adv/disadv	Positive	Negative	Neutral
Precision	0.94	0.92	0.96	0.76	0.65
Recall	0.97	0.97	0.97	0.79	0.61
F1	0.95	0.94	0.97	0.77	0.63

۴-۱۰ دقت مدل بر اساس معیارهای متفاوت روی شاخص تاریخ انقضا

شکل ۴-۸: مدل بر اساس معیارهای متفاوت روی شاخص تاریخ انقضا

(Macro) Avg	Body	Adv/disadv	Positive	Negative	Neutral
Precision	0.95	0.87	0.82	0.80	0.45
Recall	0.87	0.93	0.76	0.80	0.49
F1	0.91	0.90	0.79	0.80	0.47

۴-۱۱ جمع‌بندی

با توجه به نتایج به دست آمده که آن‌ها را به‌طور دقیق در شکل‌ها بررسی نمودیم و با توجه به بالا بودن مقدار دقت با استفاده از تابع‌های هزینه‌ی Precision ، Recall و F1 متوجه می‌شویم که مدل خروجی بسیار خوبی و مورد انتظاری دارد. لذا تنها کار برای پیش‌رفت مدل، جامع و مانع کردن آن می‌باشد که در بخش نتیجه‌گیری به‌طور کامل بدان می‌پردازیم.

فصل ۵

نتیجه‌گیری

۵-۱ مسیرهای احتمالی برای کارهای آتی

همان‌طور که در مدل بررسی کردیم، شاخص‌هایی که انتخاب شده، براساس تجربه‌ی مهندس یادگیری ماشین بوده است. انتخاب شاخص در ابعادی و معیارهایی که بحث کلان داده مطرح می‌باشد، کار به شدت حیاتی و مهم است. بنابراین بهتر است این معیارها به صورت نظارت نشده^۱ و بدون کمک برچسب‌های آن‌ها صورت پذیرد. فلذا یکی از بهترین کارهایی که می‌تواند در ادامه‌ی این مسیر انجام شود، اولاً خوشه‌بندی^۲ است و سپس مشخص کردن این شاخص‌ها به صورت پویا است.

^۱unsupervised

^۲clustering

مراجع

- [1] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.
- [2] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.
- [3] N. S. Ashish Vaswani. Attention is all you need. *arXiv:1706.03762v5*, 2017.
- [4] M.-W. C. Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805v2*, 2019.
- [5] Y. Liu. Roberta: A robustly optimized bert pretraining approach. 2019.
- [6] M. G. Mehrdad Farahani. Parsbert: Transformer-based model for persian language understanding. *arXiv:2005.12515v2*, 2020.
- [7] M. G. Javad PourMostafa Roshan Sharami. Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus. *arXiv:2004.05328v1*, 2020.
- [8] S. Hochreiter. Long short-term memory. 1997.
- [9] J. Wang. Dimensional sentiment analysis using a regional cnn-lstm model.
- [10] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.
- [11] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

- [12] J. Baxter. A model of inductive bias learning journal of artificial intelligence research. pages 149–198, 2000.
- [13] S. Thrun. Is learning the n-th thing any easier than learning the first?. in advances in neural information processing systems. 1996.
- [14] R. Caruana. Multi-task learning. 1997.

واژه‌نامه

الف

آموزش : Training

ایست‌واژه : Stopword

ازپیش‌آموزش داده‌شده : Pre-trained

ب

یافتن ریشه کلمات : Lemmatization/Stemming

بررسی : Review

پ

پراکنده : Sparse

پردازش زبان طبیعی : Natural Language

Processing

پس‌پردازش : Post-processing

پیش‌خور : Feed-Forward

پیشرفته‌ترین : State-of-the-art

ت

تابع هزینه : loss function

تجزیه و تحلیل احساسات : Sentiment Analysis

تنظیم دقیق : Fine-tune

چ

چند وظیفه‌ای : Multi task

ح

حالت مخفی : Hidden State

خ

خودنظارتی : Self-supervised

خوشه‌بندی : Clustering

ر

رابط : Interface

رسانه‌های اجتماعی : Social Media

رمزگذار : Encoder

رمزگشا : Decoder

ش

شبکه‌های عصبی بازگشتی : Recurrent Neural

Networks

شهرت نام تجاری : Net Brand Reputation

ک

کلان‌داده: Bigdata

م

مجموعه‌نوشته‌ها: Corpus

مدل‌سازی ماسک زبان: Masked-Language

Modeling

ن

نظارت نشده: Unsupervised

نگاشت: Mapping

ه

همبستگی: Correlation

Abstract

The growth of pre-trained language models has opened a new era in Natural Language Processing (NLP), allowing us to build powerful language models. Transformer-based models, such as BERT, have grown in popularity as a result of their cutting-edge performance. These models, however, are often centred on English, leaving other languages to multilingual models with minimal resources. In this thesis, the preceding models have been reviewed in details and proposed a new creative model using BERT and machine learning algorithms for two major purposes. The latter is to handle automatic handling (reject/accept) the comments and the former is the main topic of the thesis. This proposal is a real world example of research in industry, which helps Digikala company to automatically handle the procedures.

Keywords: Machine Learning, Natural Language Processing, Text Classification, Intent-Sentiment Analysis, Bigdata, Digikala



Sharif University of Technology
Department of Computer Engineering

Bachelor Thesis
Computer Engineering

Sentiment Analysis of Digikala Comments

By:

Mostafa Ghadimi

Supervisor:

Dr. Hamid Beigy

January 2022