7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

Few-Shot Learning for Intent Classification Using Pre-Trained BERT Model

Mohammad Amin Kanani[1], Mahdi Aminian[2]

[1] BSc. in Computer Engineering, Computer Engineering Department, Guilan University, Rasht, Iran
mohammadaminkanani@gmail.com

[2] Assistant Professor, Computer Engineering Department, Guilan University, Rasht, Iran
mahdi.aminian@guilan.ac.ir

Abstract

Intent classification is one of the important tasks in natural language understanding which aims to classify queries based on their intent, goal, or purpose which is implied in the content. However, the problem in this task is the lack of data labeled by a human agent. This problem leads to weakness in the generalization of models, especially when the models face rare words. Using pre-trained models can be useful in the generalization of language representation. BERT pre-trained language model which has been currently published has made a major impact in the natural language processing field. This model which is trained on unlabeled large-scale corpora, with fine-tuning could achieve state-of-art results in various natural language processing tasks e.g. question answering and sentiment analysis. In this paper, we compared the BERT model with conventional machine learning models and shown that the BERT model has a better performance than conventional machine learning models in few-shot learning.

Keywords: Natural Language Processing, Natural Language Understanding, Intent Classification, Conversational Agents, Deep Learning, Machine Learning

7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

یادگیری مجموعه داده محدود برای طبقه‌بندی مقصود با استفاده از مدل پیش‌آموزش داده شده BERT

محمد امین کنعانی[۱]، مهدی امینیان[۲]

[۱] دانشجوی کارشناسی، گروه مهندسی کامپیوتر، دانشگاه گیلان، رشت
mohammadaminkanani@gmail.com

[۲] استادیار و عضو هیئت علمی، گروه مهندسی کامپیوتر، دانشگاه گیلان، رشت
mahdi.aminian@guilan.ac.ir

چکیده

طبقه‌بندی مقصود یکی از مسائل مهم در فهم زبان طبیعی است که هدف آن طبقه‌بندی پرسش‌ها بر اساس مقصود، هدف، یا منظوری است که در محتوا بیان شده است. اما مشکلی که در این نوع مسائل وجود دارد کمبود داده‌هایی است که توسط عامل انسانی برچسب‌گذاری شده باشند. این مشکل باعث ضعف در جامع‌سازی مدل‌ها می‌شود، مخصوصا وقتی که مدل‌ها با کلمات نادر مواجه می‌شوند. استفاده از مدل‌های از پیش آموزش داده شده می‌تواند در ارائه‌ای جامع از زبان مفید واقع شوند. مدل زبانی از پیش آموزش داده شده BERT که اخیراً منتشر شده است اثر مهمی در حوزه پردازش زبان طبیعی گذاشته است. این مدل زبانی که با استفاده از یک پیکره زبانی بسیار بزرگ بدون برچسب پیش آموزش داده شده است، با تنظیم دقیق توانسته است در بسیاری از مسائل پردازش زبان طبیعی مانند سیستم‌های پرسش و پاسخ و تحلیل احساس نتایج بسیار خوبی کسب کند. در این مقاله سعی شده است که مدل زبانی BERT با مدل‌های رایج یادگیری ماشین برای طبقه‌بندی مقصود مقایسه شود و نشان داده شده است که برای یادگیری مجموعه داده محدود مدل BERT عملکرد بهتری نسبت به مدل‌های رایج یادگیری ماشین دارد.

کلمات کلیدی

پردازش زبان طبیعی، فهم زبان طبیعی، طبقه‌بندی مقصود، عوامل مکالمه‌ای، یادگیری عمیق، یادگیری ماشین، انتقال یادگیری

۱- مقدمه

در سال‌های اخیر، استفاده از سیستم‌های مکالمه گفتاری یا عوامل گفتاری در بلندگوهای هوشمند خانگی یا در بخش‌های مختلف تجارت‌های الکترونیک مانند دستیارهای هوشمند خرید یا پشتیبانی مشتریان رواج پیدا کرده است و انتظار می‌رود که استفاده از این تکنولوژی در آینده بیشتر شود. یکی از بخش‌های حیاتی عوامل گفتاری، بخش فهم زبان طبیعی است که پرسش افراد را به یک ساختار یا ارائه‌ای انتزاعی که معمولا مقصود یا عمل گفتگو نامیده می‌شود نگاشت می‌کند [۱]. طبقه‌بندی مقصود برای پیش‌بینی منظور کاربر استفاده می‌شود و هنگامی که مدل برای پرسش کاربر یک برچسب پیش‌بینی کرد، سیستم از این برچسب برای تولید یک جواب یا انجام دادن یک کار استفاده می‌کند. همچنین فهم مقصود کاربر در عوامل گفتاری برای پردازش وظایف دیگر مانند مدیریت گفتگو[۱] بسیار حیاتی و مهم است. همچنین در مکالمات، دامنه باز اطلاعات زمینه‌ای (یک یا چند سخن قبلی) برای فهم زبان بسیار مهم است. جدول (۱) یک نمونه از طبقه‌بندی مقصود برای پرسش "وقتی سریال کارت هدیه رو می‌زنم سیستم پیغام می‌دهد کارت هدیه من نامعتبر است. چکار کنم؟" است.

طبقه‌بندی مقصود یک مسئله از نوع طبقه‌بندی متن است که هدف آن پیش‌بینی برچسب مقصود y^i است که i نشان دهنده داده iام است و y

پرسش	وقتی سریال کارت هدیه را می‌زنم سیستم پیغام می‌دهد کارت هدیه من نامعتبر است. چکار کنم؟
مقصود	wrong_serial_number_for_gift_card

جدول (۱) : یک مثال از نگاشت پرسش کاربر به مقصود

برچسب این داده است. دست آوردهایی بر اساس شبکه‌های عصبی بازگشتی[۲]، مخصوصا شبکه‌های حافظه‌ی طولانی کوتاه مدت[۳] و واحدهای درگاهی مکرر[۴]، نتایج بسیار خوبی در طبقه‌بندی مقصود کسب کرده‌اند. همچنین نشان داده شده است که سازوکار توجه[۵] [۲] به شبکه‌های عصبی بازگشتی کمک می‌کند که بتوانند درک بهتری از وابستگی‌های با طول بلند در متن داشته باشد.

یکی از مشکلات وظیفه طبقه‌بندی مقصود، وجود مجموعه داده‌های محدود برای آموزش مدل‌های یادگیری عمیق است. مجموعه داده‌های محدود برای هر کلاس تعداد کمی داده برچسب شده دارند که باعث بیش برازش در این نوع از مدل‌ها می‌شود. برای حل این مشکل روش‌های مبتنی بر انتقال یادگیری که با حجم زیادی از داده‌های بدون برچسب با منظور عمومی آموزش داده می‌شوند پیشنهاد شده است. روش‌های اولیه مانند کلمه به بردار[۶] [۳] یک ارائه‌ای از تعبیه کلمه فارغ از زمینه‌ای که کلمه در آن آمده است یاد می‌گیرد و رهیافت‌های دیگری مانند ELMo [۴] ارائه‌ای از تعبیه کلمه را با توجه به زمینه‌ای که کلمه در آن آمده است یاد می‌گیرد. این بردارهای تعبیه کلمات معمولا به صورت ویژگی‌های افزوده به وظیفه اصلی اضافه می‌شوند. با این وجود، به صورت تجربی اثبات شده است که پیش-آموزش بر اساس شکل‌های مختلف مدل زبانی عملکرد بهتری دارند [۵]. تنظیم دقیق مدل زبانی فراگیر [۶] که مبتنی بر شبکه‌های حافظه‌ی طولانی کوتاه مدت است یک روش برای تنظیم دقیق مدل زبانی از پیش آموزش داده شده است که در شش مجموعه داده طبقه‌بندی متن نتایج بسیار خوبی کسب کرده است. اخیرا مدل‌های زبانی از پیش آموزش داده شده مانند OpenAI GPT [۷] و BERT [۸] نشان داده‌اند که با بکار گرفتن حجم زیادی از داده‌های بدون برچسب در یادگیری ارائه‌های رایج زبانی بسیار مفید هستند. مدل BERT بر اساس مبدل چند لایه دوطرفه [۹] است و بر روی مجموعه دادگان بدون برچسب آموزش داده است.

اگرچه استفاده از مدل‌های زبانی باعث بهبود در نتایج می‌شود، اما برای این که این مدل‌ها بتوانند بردارهایی جامع از متن ارائه کنند نیازمند پیکره‌های زبانی بزرگ هستند و هزینه‌های محاسباتی بالایی دارند. همچنین با توجه به تفاوت‌های زبان‌ها این مدل‌های زبانی غالبا بر روی زبان انگلیسی پیش آموزش داده می‌شوند. اما وجه تمایزی که مدل BERT دارد این است که این مدل چند زبانه است و می‌تواند ۱۰۴ زبان را پردازش کند و به همین دلیل است که از این مدل برای روش پیشنهادی استفاده شده است.

۲- مفاهیم اولیه

داده‌های بدون ساختار مانند متن همه جا هستند، مانند ایمیل‌ها، صفحات وب، شبکه‌های اجتماعی، چت بات‌ها و غیره. متن می‌تواند منبع غنی از اطلاعات باشد، اما استخراج اطلاعات مفید از آن به دلیل ماهیت غیر ساختاری که دارد می‌تواند سخت و زمان‌بر باشد.

طبقه‌بندی متن یکی از راه‌های استخراج اطلاعات از متن است که به دو نوع متفاوت روش‌های مبتنی بر قاعده و روش‌های مبتنی بر یادگیری ماشین دسته‌بندی می‌شود.

۲-۱- روش‌های مبتنی بر قاعده

رهیافت‌های مبتنی بر قاعده متن‌ها را بر اساس مجموعه‌ای از قوانین زبان شناختی در دسته‌های مختلف طبقه‌بندی می‌کند. این قواعد با استفاده عناصر معنایی که در محتوای متن استفاده شده دسته مربوطه‌ی متن را پیدا می‌کند.

۲-۲- روش‌های مبتنی بر یادگیری ماشین

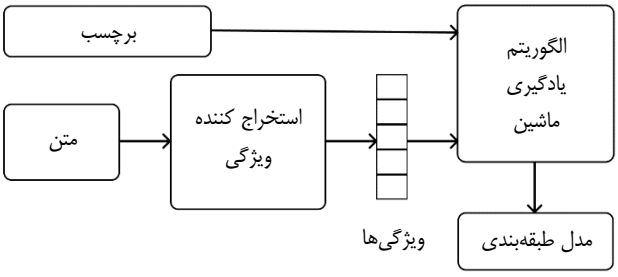
بر خلاف تکیه بر قواعد دستی، طبقه‌بندی متن با یادگیری ماشین یاد می‌گیرد که بر اساس مشاهدات قبلی وظیفه طبقه‌بندی را انجام دهد. با استفاده از نمونه‌های از قبل برچسب‌گذاری شده به عنوان داده‌های آموزش، الگوریتم یادگیری ماشین می‌تواند پیوستگی‌های متفاوت بین تکه‌های متن و یک خروجی خاص که برای یک ورودی مشخص مورد انتظار است را یاد بگیرد.

اولین قدم برای آموزش یک طبقه‌بند استخراج ویژگی است، که روشی است برای تبدیل متن به یک بردار نمایش دهنده عددی. یکی از پرکاربردترین روش‌ها برای تبدیل متن به بردار عددی روش کوله کلمات[۷] است که یک بردار نشان دهنده تعداد هر کلمه در یک دیکشنری تعریف شده از کلمات است.

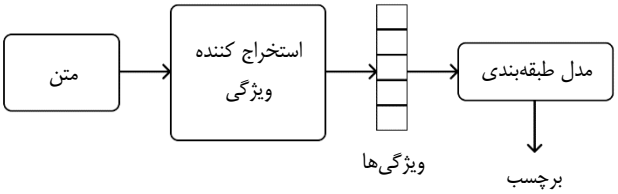
بعد از استخراج ویژگی، الگوریتم یادگیری ماشین با استفاده از داده‌های آموزشی که یک جفت از بردار ویژگی برای هر نمونه و برچسب متناظرش است تغذیه می‌شود که در نهایت یک مدل طبقه‌بندی ساخته می‌شود. شکل (۱) نمونه‌ای از مراحل آموزش یک طبقه‌بند است.

هنگامی که این مدل با نمونه‌های کافی آموزش یافت، مدل یادگیری ماشین می‌تواند شروع به پیش‌بینی کند. همان روند استخراج ویژگی نیز برای متن دیده نشده انجام می‌شود، و سپس ویژگی‌های استخراج شده از متن به مدل داده می‌شود تا برچسب آن متن را پیش‌بینی کند. در شکل (۲) مراحل استفاده از مدل آموزش داده شده برای پیش‌بینی متن دیده نشده نشان داده شده است.

طبقه‌بندی با استفاده از یادگیری ماشین معمولا دقیق‌تر از روش‌های مبتنی بر قاعده است. طبقه‌بندها با یادگیری ماشین برای نگهداری و توسعه ساده‌تر هستند. الگوریتم‌های بیز ساده[۸]، ماشین بردار پشتیبان[۹] و یادگیری عمیق چند نمونه از الگوریتم‌های یادگیری ماشین هستند.



شکل (۱) : مراحل آموزش یک مدل طبقه‌بند متن



شکل (۲) : استفاده از مدل آموزش داده شده برای پیش‌بینی برچسب

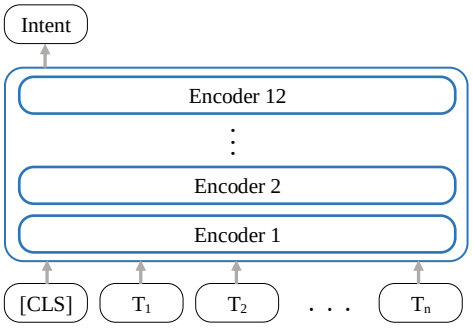
۲-۳- مدل زبانی BERT

معماری مدل BERT (Bidirectional Encoder Representations from Transformers) بر اساس قسمت رمزگذار معماری مبدل [۹] است که یک مدل چندلایه دو طرفه است. ورودی این مدل می‌تواند یک جمله یا دو جمله (به عنوان مثال برای وظیفه پرسش و پاسخ) باشد و از تعبیه‌های WordPiece [۱۹] استفاده شده است. اولین توکن ورودی هر دنباله همیشه یک توکن مخصوص طبقه‌بندی ([CLS]) است. اخرین حالت نهفته مربوط به این توکن به عنوان ارائه‌ای از کل جمله در مسائل طبقه‌بندی استفاده می‌شود. در این مدل نیز می‌توان دو جمله را بهم چسباند و یک جفت جمله تشکیل داد. در این مدل جمله‌ها به دو صورت قابل تمایز هستند. اول جملات با یک توکن خاص ([SEP]) از هم جدا می‌شوند. دوم به هر توکن یک نهفته اضافه می‌شود که نشان می‌دهد توکن متعلق به جمله اول است یا جمله دوم. با فرض یک دنباله از توکن‌ها $\boldsymbol{x} = (x_1, \dots, x_T)$، خروجی مدل $\mathbf{H} = (h_1, \dots, h_T)$ است که مدل BERT برای طبقه‌بندی بردار h_1 که مربوط به توکن مخصوص طبقه‌بندی ([CLS]) است را می‌گیرد.

برخلاف روش‌های رایج چپ به راست و راست به چپ، مدل BERT به صورت بدون نظارت با استفاده از دو وظیفه بر روی یک مجموعه عظیم پیکره متنی پیش آموزش داده شده است. شکل (۳) یک نمای کلی از معماری مدل BERT است.

۲-۳-۱- مدل زبانی ماسک شده

برای آموزش یک ارائه دوطرفه عمیق، یک درصدی از توکن‌های ورودی به



شکل (۳) : یک نمای کلی از مدل پیشنهادی با فرض دنباله‌ای از توکن‌های $\mathbf{T} = ([\boldsymbol{CLS}], \boldsymbol{T_1}, \dots, \boldsymbol{T_N})$ و خروجی مدل که مقصود پرسش است.

صورت تصادفی با یک توکن مخصوص ([MASK]) پوشانده می‌شوند و مدل باید پیش‌بینی بکند این لغت پوشانده شده چه لغتی است. مشکلی که در اینجا وجود دارد این است که بین پیش‌آموزش و تنظیم دقیق عدم تطابق وجود دارد، زیرا توکن مخصوص نقاب در حین تنظیم دقیق آورده نمی‌شود. برای کم کردن این مشکل ۱۵٪ از کلمات دنباله ورودی به صورت تصادفی انتخاب می‌شوند. اگر یک توکن انتخاب شود به احتمال ۸۰٪ با توکن مخصوص نقاب جایگذاری می‌شود و با احتمال ۱۰٪ با یک توکن تصادفی دیگر جا به جا می‌شود و در بقیه موارد توکن تغییری نمی‌کند.

۲-۳-۲- پیش‌بینی جمله بعدی

بسیاری از وظایف در پردازش زبان طبیعی بر اساس درکی از رابطه بین دو جمله است، که مستقیما در مدل‌های زبانی کسب نمی‌شود [۸]. برای این که مدلی آموزش داده شود که بتواند رابطه بین دو جمله را درک بکند، مدل BERT برای یک وظیفه طبقه‌بندی متن دودویی که از هر مجموعه داده‌ی متن تک زبانه‌ای می‌تواند تولید شود، پیش آموزش داده شده است.

با فرض یک جفت جمله A و B، در نصف موارد جمله B دقیقا همان جمله‌ای است که بعد از جمله A آمده است (که برچسب بعدی–است می‌گیرد) و در بقیه موارد یک جمله تصادفی از بقیه پیکره متنی انتخاب می‌شود (که برچسب بعدی–نیست می‌گیرد).

۳- پیشینه و کارهای مرتبط

مدل‌های یادگیری عمیق به صورت گسترده‌ای در فهم زبان طبیعی بررسی شده‌اند. به دلیل شباهت‌های بین طبقه‌بندی مقصود و طبقه‌بندی متن، کارهای مرتبط را به دو دسته طبقه‌بندی مقصود و طبقه‌بندی متن دسته‌بندی می‌کنیم. ضمنا به دلیل کارهای بسیار زیادی که در این زمینه انجام شده است فقط به صورت مختصر به معرفی دست آوردهایی که به صورت گسترده استفاده شده‌اند اشاره می‌کنیم.

۳-۱- طبقه‌بندی مقصود

نشان داده شده است که مدل‌های بر اساس شبکه‌های عصبی بازگشتی و شبکه‌های حافظه‌ی طولانی کوتاه مدت عملکرد بهتری نسبت به روش‌های مبدأ مبتنی بر مدل زبانی n-gram، مدل زبانی مبتنی بر شبکه‌های عصبی پیشخور[۱۰] و طبقه‌بندهای تقویت کننده در طبقه‌بندی مقصود دارند [۱۰]. در روش حافظه طولانی کوتاه مدت سلسله مراتبی[۱۱] هر مکالمه یک دنباله‌ی سلسله مراتبی از داده در نظر گرفته می‌شود که هر جمله یک دنباله‌ای از کلمات و هر نشست یک فهرستی از جملات است. با فرض یک مکالمه‌ای از n جمله در ابتدا با استفاده از یک حافظه طولانی کوتاه مدت تمام جملات به طور مستقل مدل می‌شوند. حالت‌های نهفته هر جمله که در این مرحله بدست آمده است با استفاده از یک حافظه طولانی کوتاه مدت دیگر به یک بردار جمله برای هر جمله در مکالمه تبدیل می‌شود. همچنین برای ارتقا بیشتر مدل کردن زمینه مکالمات پیچیده، یک مؤلفه حافظه نیز اضافه شده است. این مؤلفه در خروجی حافظه طولانی کوتاه مدت دوم قرار داده می‌شود که اطلاعات زمینه‌ای مورد نیاز را موقع محاسبه بردار جمله فراهم و به یاد می‌سپارد. برای استخراج ویژگی از متن معمولا از روش‌های رایجی مانند ویژگی‌های n-gram استفاده می‌شود، اما در یادگیری عمیق ویژگی‌ها به صورت پیوسته و خودکار یاد گرفته می‌شوند. در روش شبکه عصبی پیچشی بازگشتی [۱۲] از شبکه عصبی پیچشی[۱۱] برای استخراج ویژگی از دنباله کلمات ورودی استفاده شده است. هر کلمه در جمله به یک بردار متوالی ثابت رمز می‌شود و کل جمله الحاقی از بردارهای کلمات است. در نتیجه لایه نهفته ویژگی h شامل هر بردار ویژگی استخراج شده از پنجره پیچشی مشترک است که در موقعیت هر کلمه قرار دارد و این سازوکار استخراج ویژگی اطلاعات پیچیده‌تری نسبت به تعداد n-gram کسب می‌کند. این ویژگی‌های بدست آمده هم برای وظیفه طبقه‌بندی مقصود هم برای وظیفه برچسب گذاری دنباله در وظایف فهم زبان گفتاری استفاده می‌شوند. همچنین با اضافه کردن لایه‌های مختص هر وظیفه بر روی لایه ویژگی این شبکه عصبی پیچشی می‌توان به طور همزمان وظیفه طبقه‌بندی و برچسب گذاری را انجام داد. برای وظیفه طبقه‌بندی مقصود یک لایه max-pooling که با یک لایه خروجی softmax همراه شده است اضافه می‌شود. همچنین برای وظیفه برچسب‌گذاری از مدل‌های زمینه‌های تصادفی شرطی[۱۲] استفاده می‌شود. برای آوردن دانش زمینه‌ای ارتباطات بازگشتی از خروجی مقصود و شکاف قبلی به مدل شبکه عصبی پیچشی غیر زمینه‌ای اضافه می‌شود.

اگرچه شبکه‌های عصبی رمزکننده رمزگشا[۱۳] مبتنی بر توجه نتایج امیدوار کننده‌ای در وظایف ترجمه ماشین و تشخیص گفتار کسب کرده‌اند، نشان داده شده است که این نوع از شبکه نیز در وظیفه تشخیص مقصود و پرکردن شکاف هم قابل استفاده هستند [۱۳]. در این مدل یک شبکه عصبی بازگشتی دو طرفه دنباله مبدأ را از جلو و عقب می‌خواند. در اینجا از مدل حافظه طولانی کوتاه مدت برای واحد شبکه عصبی بازگشتی استفاده شده است. در هر قدم حالت نهفته h_i ترکیبی از حالت نهفته رو به جلو fh_i و حالت نهفته رو به عقب bh_i است ($h_i = [fh_i, bh_i]$). هر حالت نهفته h_i شامل اطلاعاتی از کل کلمه دنباله ورودی است. این حالت نهفته سپس با بردار زمینه c_i، که به میانگینی وزن‌دار از حالت‌های نهفته ($\mathbf{H} = (h_1, \dots, h_T)$) شبکه عصبی بازگشتی است ترکیب می‌شود و در نهایت این بردار ترکیب شده برای تولید یک توزیع بین برچسب‌ها استفاده می‌شود. برای تشخیص مقصود اگر از ساز و کار توجه استفاده نشود، بر روی بردارهای حالت نهفته mean-pooling همراه با رگرسیون لجستیک[۱۴] اعمال می‌شود؛ و اگر از ساز و کار توجه استفاده شود میانگین وزن‌دار بردارهای حالت نهفته استفاده می‌شود. همچنین از مدل BERT برای وظایف پرکردن شکاف نیز استفاده شده است [۱۴]. در این روش تعبیه هر کلمه با یک لایه softmax همراه می‌شود که یک توزیع بین برچسب‌ها بدست می‌آورد.

۳-۲- طبقه‌بندی متن

اگر چه شبکه‌های عصبی پیچشی معمولا برای وظایف بینایی کامپیوتر استفاده می‌شوند، از آن‌ها برای طبقه‌بندی متن نیز استفاده شده است [۱۵]. در این روش هر کلمه تبدیل به یک بردار ثابت می‌شود و جمله نیز به صورت دنباله‌ای از بردار کلمات پشت سر هم در نظر گرفته می‌شود. هر فیلتر به یک پنجره‌ای از کلمات اعمال می‌شود و بعد از این که بر روی کل جمله اعمال شود یک برداری از ویژگی‌ها می‌سازد. بر روی این بردار نگاشت ویژگی عملیات max-overtime pooling اعمال می‌شود و مقدار بدست آمده ویژگی مرتبط به آن فیلتر در نظر گرفته می‌شود. این عمل pooling باعث می‌شود که مدل بتواند جمله‌هایی با طول متغیر را بپذیرد. همچنین شبکه‌های عصبی پیچشی می‌توانند عمل استخراج ویژگی را بر روی کاراکترها انجام دهند [۱۶]، به این صورت که جمله به جای این که تبدیل به دنباله‌ای از بردارهای کلمات شود تبدیل به دنباله‌ای از بردارهای کاراکترها می‌شود و هر بردار کاراکتر یک بردار one-hot است و اندازه این بردار به تعداد کاراکترها است. همچنین در این روش نمی‌توان ورودی‌هایی با سایز متغیر داشت و طول ویژگی‌های ورودی باید ثابت باشد. توجه داشته باشید که اگر طول ورودی کمتر از ورودی مدل باشد جمله مورد نظر با بردارهایی به مقدار صفر پر می‌شود.

بدلیل این که شبکه‌های عصبی پیچشی بر روی پنجره‌ای از کلمات اعمال می‌شوند، توانایی یادگیری اطلاعات زمینه‌ای بلند مدت و رابطه‌ی بین کلماتی که پشت هم نیستند را ندارند. اما با اضافه کردن یک لایه توجه بین لایه ورودی، دنباله‌ای به طول ثابت از بردارهای کلمه، و شبکه عصبی پیچشی می‌توان یک بردار زمینه برای هر کلمه بدست آورد. این بردار زمینه‌ای به بردار کلمه ملحق می‌شود تا یک بردار ارائه جدیدی از کلمه بسازد و در نهایت این بردارهای جدید کلمات هستند که به شبکه عصبی تغذیه می‌شوند. ایده ساز و کار توجه این است که یاد بگیرد که موقع استخراج بردار زمینه برای هر کلمه بر روی کلمات خاصی تمرکز کند [۱۷].

یادگیری انتقال استنتاجی اثر بسیار زیادی در حوزه بینایی کامپیوتر گذاشته است، اما رهیافت‌های پردازش زبان طبیعی که تا اینجا معرفی شدند نیازمند تغییرات مختص به هر وظیفه دارند و باید از ابتدا آموزش داده شوند. تنظیم دقیق مدل زبانی فراگیر [۶] یک روش مؤثر انتقال یادگیری است که برای بسیاری از وظایف پردازش زبان طبیعی از جمله طبقه‌بندی متن استفاده می‌شود. این روش شامل سه مرحله است که در مرحله اول مدل بر روی یک پیکره زبانی با دامنه عمومی آموزش داده می‌شود تا بتواند ویژگی‌های عمومی زبان را در لایه‌های مختلف بدست آورد. بعد از این مرحله کل مدل زبانی بر روی داده‌های وظیفه هدف با استفاده از تنظیم دقیق افتراقی و نرخ یادگیری مثلثی شیب‌دار تنظیم دقیق می‌شود. در نهایت یک طبقه‌بند با استفاده از آزاد کردن تدریجی لایه‌ها و نرخ یادگیری مثلثی شیب‌دار بر روی داده‌های هدف تنظیم دقیق می‌شود. این عمل باعث می‌شود که ارائه‌های زبانی سطح پایین حفظ شود و فقط ارائه‌های سطح بالا که مربوط به وظیفه هدف است آموزش داده شود. مدل از پیش آموزش داده شده BERT نیز با تنظیم دقیق بر روی داده‌های هدف برای وظیفه طبقه‌بندی نیز استفاده می‌شود [۵].

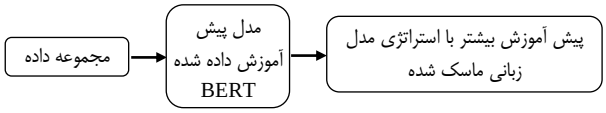
٤- روش پیشنهادی

مدل‌های یادگیری عمیق برای این که بتوانند بردارهایی جامع از داده ارائه کنند نیازمند مقدار بسیار زیادی داده هستند. به همین دلیل استفاده از مدل-های یادگیری عمیق برای مجموعه داده‌های محدود باعث بیش برازش می-شود. اما مدل BERT که به صورت خود نظارت شده بر روی پیکره زبانی با مقیاس بزرگ آموزش داده شده است، توانایی درک کلی از زبان را دارد و می-تواند ویژگی‌های زبانی جامعی را از متن استخراج بکند. به همین دلیل این مدل بدون این که بیش برازش شود، توانایی یادگیری مجموعه داده‌های محدود را دارد.

برای وظیفه طبقه‌بندی مقصود در این روش، ابتدا مدل بر روی مجموعه داده پیش آموزش بیشتر داده می‌شود و بعد از پیش آموزش بیشتر، بر روی مجموعه داده تنظیم دقیق می‌شود.

٤-١- پیش آموزش بیشتر مدل BERT

از آنجایی که مدل BERT بر روی پیکره زبانی با منظوری عمومی پیش آموزش داده شده است، برای طبقه‌بندی در یک دامنه معین، به عنوان مثال داده‌هایی برای تحلیل احساس، توزیع داده‌ها ممکن است با داده‌های مدل BERT متفاوت باشد. برای همین بهتر است با یکی از استراتژی‌هایی که در مقاله اصلی ذکر شده است مدل یک بار دیگر پیش آموزش داده شود که در این مقاله بدلیل این که مسئله از نوع طبقه‌بندی است از استراتژی مدل زبانی نقاب شده برای پیش آموزش بیشتر استفاده می‌کنیم. روند این فرآیند در شکل (۴) نشان داده شده است.



شکل (٤) : پیش آموزش بیشتر مدل BERT

٤-٢- تنظیم دقیق مدل BERT

برای طبقه‌بندی مقصود، مدل BERT آخرین حالت نهفته بردار h_1 را به عنوان ارائه‌ای از کل جمله می‌گیرد که بردار متناظر توکن [CLS] است. یک طبقه

بند ساده softmax به انتهای مدل اضافه می‌کنیم که احتمال مقصود c را پیش‌بینی می‌کند:

$$p(c|h_1) = softmax(Wh_1) \qquad (۱)$$

که در اینجا W ماتریس پارامترهای مخصوص وظیفه است. تمام پارامترهای مدل BERT همراه با ماتریس W با بیشینه کردن لگاریتم احتمال برچسب-های درست تنظیم دقیق می‌شوند.

٥- تحلیل نتایج

در این بخش مدل پیشنهادی را بر روی مجموعه داده پرسش‌های پرتکرار فروشگاه اینترنتی دیجیکالا[۱۵] ارزیابی کردیم.

٥-١- مجموعه داده

مجموعه داده‌های که در این مقاله استفاده شده است با توجه به پرسش‌های پرتکرار فروشگاه اینترنتی دیجیکالا تهیه شده است. این مجموعه داده شامل ۱۶۲۱ پرسش با ۱۲۹ مقصود است. از این مجموعه داده مقدار ۱۲۹۶ به مجموعه داده آموزش و مقدار ۳۲۵ به مجموعه داده تست اختصاص داده شده است. پ

٥-٢- جزییات آموزش

در این مقاله از مدل چند زبانه $\text{BERT}_{\text{base}}$ استفاده کردیم که بردار نهفته‌ای با اندازه ۷۶۸ دارد. این مدل بر روی ۱۰۴ زبان پیش آموزش داده شده است که زبان فارسی را نیز شامل می‌شود. همچنین این مدل دارای ۱۲ بلوک مبدل [۹و۱۲] سر self-attention است. برای پیش آموزش بیشتر این مدل، اندازه دسته‌ها را ۶۴، بیشینه طول دنباله را ۱۲۸و تعداد گام‌ها را ۱۰۰ در نظر گرفتیم. برای نرخ یادگیری یکی از مقادیر ۵e−۵، ۴e−۵، ۳e−۵، و ۲e−۵ پیشنهاد می‌شود [۸]، که به صورت تجربی مقدار ۵e−۵ عملکرد بهتری نسبت به بقیه مقادیر داشت.

برای تنظیم دقیق این مدل برای وظیفه طبقه‌بندی مقصود، اندازه دسته‌ها را ۱۶ گرفتیم. همچنین از بهینه ساز Adam [۱۹] با $\beta_1 = 0.9$ و

F1	Recall	Precision	Accuracy	
۸۸.۱۲	۹۰.۰۲	۸۸.۹۱	۹۲.۳	Logistic Regreesion (BoW)
۸۳.۸۶	۸۵.۱۵	۸۵.۳۳	۸۹.۸۴	Logistic Regression (tf-idf)
۸۷.۸۱	۸۹.۵۱	۸۸.۷۲	۹۲.۶۱	SVM (BoW)
۸۷.۷۴	۸۹.۴۲	۸۸.۲۸	۹۲.۳	SVM (tf-idf)
۹۱.۳۹	۹۳.۰۴	۹۲.۷۴	۹۳.۵۳	BERT (Without Further Fine-Tuning)
۹۴.۷۶	۹۵.۴۷	۹۶	۹۵.۰۷	BERT (With Further Fine-Tuning)

جدول (۲) : نتایج بدست آمده از مدل‌های مختلف (برحسب درصد)

$\beta_2 = 0.999$. استفاده کردیم و نرخ آموزش را به صورت تجربی ۵e−۵ در نظر گرفتیم. تعداد گام‌ها را نیز به صورت تجربی ۴ در نظر گرفتیم

۵-۳- نتایج

برای درک بهتر از نتایج این مدل، ما از مدل‌های رگرسیون لجستیک و ماشین بردار پشتیبانی برای مقایسه نتایج استفاده کردیم. برای این دو مدل یادگیری ماشین، در مرحله پیش پردازش علامت‌های نگارشی و ایست واژه‌ها را از پرسش‌ها حذف کردیم. در این دو مدل هر کدام از پرسش‌ها را با استفاده از دو روش کوله کلمات و tf-idf[۱۶] تبدیل به بردار عددی کردیم.

مدل BERT را نیز یک بار بدون این که بر روی مجموعه داده‌ها بیشتر پیش آموزش کنیم تنظیم دقیق کردیم که بتوانیم مقایسه بهتری بین روش‌های رایج تنظیم دقیق این مدل داشته باشیم. برای این مدل به صورت تجربی تعداد گام‌ها را ۵ گرفتیم و بقیه هایپرپارامترهای این مدل با مدلی که بیشتر پیش آموزش داده شده است برابر است. نتایج بدست آمده از این مدل‌ها در جدول (۲) آمده است. با توجه به نتایج می‌توان گفت که مدل بیشتر آموزش داده شده BERT نتایج بسیار بهتری نسبت به الگوریتم‌های دیگر یادگیری ماشین کسب کرده است که نشان از قدرت این مدل در وظیفه طبقه‌بندی مقصود دارد. همچنین پیش آموزش بیشتر مدل BERT با مجموعه داده دقت این مدل را نسبت به حالتی که پیش آموزش بیشتر داده نشود بالا می‌برد.

٦- جمع بندی

در این مقاله یک مدل طبقه‌بندی مقصود معرفی شده است که بر اساس مدل چند زبانه بر پایه BERT تلاش می‌کند مشکل قابلیت جامع سازی مدل‌های رایج فهم زبان طبیعی را پوشش دهد. در طبقه‌بندی مقصود، کمبود داده‌های برچسب گذاری شده باعث بیش برازش در مدل‌های یادگیری عمیق می‌شود. اما نتایج تجربی نشان می‌دهند که مدل پیشنهادی BERT عملکرد بهتری نسبت به مدل‌های رایج یادگیری ماشین دارد و همچنین مدل‌های یادگیری عمیق که از پیش آموزش داده شده‌اند عملکرد بسیار بهتری در وظیفه طبقه‌بندی مقصود دارند. همچنین می‌توان نتیجه گرفت که پیش‌آموزش بیشتر مدل BERT در بهبود عملکرد این مدل برای وظیفه طبقه‌بندی مقصود بسیار مؤثر است. در آخر، افزایش قابلیت جامع سازی این مدل در وظیفه‌های مشخص و تقویت بردارهای ارائه برای یادگیری مجموعه داده محدود از جمله کارهایی است که می‌تواند در آینده مورد بررسی و تحقیق قرار بگیرد.

سپاسگزاری

در تهیه این سند از مجموعه داده سوالات پرتکرار فروشگاه اینترنتی دیجیکالا که توسط تیم هوش مصنوعی کوونتا[۱۷] تولید و برچسب گذاری شده است استفاده شده، که از زحمات آنان سپاسگزاری می‌شود.

مراجع

[1] Liu, X., Eshghi, A., Sweitojanski, P., Rieser, V., "Benchmarking Natural Language Understanding Services for Building Conversational Agents", 10th International Workshop on Spoken Dialogue Systems Technology, Siracusa, Sicily, Italy, 2019.

[2] Bahdanau, D., Cho, K., Bengio, Y., "Neural Machine Translation by Jointly Learning to Align and Translate", arXiv preprint arXiv: 1409.0473, 2016.

[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., "Distributed Representations of Words and Phrases and their Compositionality", The 26th International Conference on Neural Information Processing Systems, Vol. 2, pp. 3111–3119, 2013.

[4] Peters, Matthew E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., "Deep contextualized word representations", Proceedings of NAACL, pp. 2227–2237, 2018.

[5] Sun, C., Qiu, X., Xu, Y., Huang, X., "How to Fine-Tune BERT for Text Classification", arXiv preprint arXiv: 1905.05583, 2020.

[6] Howard, J., Ruder, S., "Universal Language Model Fine-Tuning for Text Classification", Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp 328–339, Melbourne, Australia, 2018.

[7] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., "Improving Language Understanding by Generative Pre-training", arXiv preprint arxiv: 1607.06450, 2016.
[8] Devlin, J., Chang, M., Lee, K., Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 4171–4186, 2018.
[9] Vaswani, A., Shazeer, N., Parmer, N., Uszkoreit, J., Jones, L., Gomez, Aidean N., Kaiser, L., Polosukhin, I., "Attention Is All You Need", arXiv preprint arXiv: 1706.03762, 2017.
[10] Ravuri, S., Stolcke, A., "Recurrent Neural Network and LSTM Models for Lexical Utterance Classification", 16th Annual Conference of the International Speech Communication Association, 2015.
[11] Meng, L., Huang, M., "Dialogue Intent Classification with Long Short-Term Memory Networks", Natural Language Processing and Chinese Computing, pp. 42-50, 2018.
[12] Liu, C., Xu, P., Sarikaya, R., "Deep Contextual Language Understanding in Spoken Dialogue Systems", International Speech Communication Association, 2015.
[13] Liu, B., Lane, I., "Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling", arXiv preprint arXiv: 1609.01454, 2016.
[14] Chen, Q., Zhuo, Z., Wang, W., "BERT for Joint Intent Classification and Slot Filling", arXiv preprint arXiv: 1902.10909, 2019.
[15] Kim, Y., "Convolutional Neural Networks for Sentence Classification", arXiv preprint arXiv: 1408.5882, 2014.
[16] Zhang, X., Zhao, J., LeCun, Y., "Character-level Convolutional Networks for Text Classification", Proceedings of the 28th International Conference on Neural Information Processing Systems, Vol. 1, pp. 649–657, 2015.
[17] Zhao, Z., Wu, Y., "Attention-Based Convolutional Neural Networks for Sentence Classification", INTERSPEECH, pp. 705-709, 2016.
[18] Wu, Y., Schuster, M., Chen, Z., Vle, Q., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al, "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation", arXiv preprint arXiv: 1609.08144, 2016.
[19] Kingma, Diederik P., Ba, J., "Adam: A Method for Stochastic Optimization", arXiv preprint arXiv: 1412.6980, 2017.

زیرنویس‌ها

1 Dialogue Management
2 Recurrent Neural Networks (RNN)
3 Long Short Term Memory (LSTM)
4 Gated Recurrent Unit (GRU)
5 Attention Mechanism
6 Word2Vec
7 Bag of Words
8 Naïve Bays
9 Support Vector Machine (SVM)
10 Feedforward Neural Networks
11 Convolutional Neural Networks (CNN)
12 Conditional Random Field (CRF)
13 Encoder-Decoder
14 Logistic Regression
15 www.digikala.com
16 Term Frequency–Inverse Document Frequency
17 www.kouventa.ir