DOR: 20.1001.1.2538161.1403.9.1.6.5

DOI: 10.22034/csj.2024.197230

بهار ۱۴۰۳، دوره ۹، شماره ۱، پیاپی ۳۲

دریافت مقاله: ۱۴۰۲/۱۲/۰۴

پذیرش مقاله: ۱۴۰۳/۰۱/۳۰

نوع مقاله: پژوهشی

ایجاد مدل هوش مصنوعی بر پایه مبدل به‌منظور پاسخ به سؤالات از داده‌های بدون ساختار

فریبرز دلیری*

دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشکده و پژوهشکده رایانه شبکه و ارتباطات، دانشگاه جامع امام حسین(ع)، تهران، ایران

پست الکترونیکی: fariborz.daliri@ihu.ac.ir

محسن نوروزی

پژوهشگر گروه مهندسی کامپیوتر، دانشکده و پژوهشکده رایانه شبکه و ارتباطات، دانشگاه جامع امام حسین(ع)، تهران، ایران

پست الکترونیکی: m.norouzi@ihu.ac.ir

علی ناصری

استاد گروه مهندسی کامپیوتر، دانشکده و پژوهشکده رایانه شبکه و ارتباطات، دانشگاه جامع امام حسین(ع)، تهران، ایران

پست الکترونیکی: a.naseri@ihu.ac.ir

چکیده

به دنبال موفقیت‌های چشم‌گیر مدل‌های زبانی از پیش‌آموزش‌دیده و یادگیری انتقالی، اخیراً موج قابل‌توجهی از تحقیقات هوش مصنوعی در این حوزه انجام شده است. مدل‌های زبانی یکی از مؤلفه‌های اساسی در پردازش زبان طبیعی هستند. این مدل‌ها تنها یک‌بار روی پیکره‌های بزرگ متنی به‌صورت خود ناظر آموزش‌دیده و سپس در حل مسایل مختلف فهم زبانی مانند پاسخ به پرسش، بسط متن و طبقه‌بندی متون به‌کارگیری می‌شوند. به‌کارگیری مبدل‌ها برای حل مسایل گوناگون مستلزم فراهم نمودن سه پیش‌نیاز (مدل، مجموعه داده مناسب، استانداردهایی به‌منظور توصیف نحوه ارزیابی مدل) است. بررسی حجم بالای اسناد بدون ساختار که بیشتر در قالب متن هستند برای متخصصان امری دشوار است. از طرفی پاسخ به بسیاری از سؤالات تحقیق نیازمند بررسی حجم بالایی از متون است. در این پژوهش به‌منظور پاسخ به سؤالات از روی متون بدون ساختار در زمینه تحلیل امنیت شبکه‌های کامپیوتری سیستمی بر پایه هوش مصنوعی ارایه شده است. سیستم مذکور بر پایه مبدل‌ها پیاده‌سازی شده و با زبان پایتون و در بن‌سازهٔ گوگل کولب‌پرو کدنویسی و پیاده‌سازی شده است. این سیستم پس از آموزش روی پرسش و پاسخ‌های حوزه امنیت شبکه‌های کامپیوتری قادر به استخراج پاسخ سؤالات از متن بدون ساختار است. به‌منظور آموزش مدل یک مجموعه داده سفارشی به نام هوشفا توسعه داده شده است که شامل ۱۲,۱۵۷ رکورد است. سیستم مذکور پس از ارزیابی دقت ارایه پاسخ در معیار تطبیق دقیق را بسته به نوع مدل پایه به میزان ۲۰ تا ۴۳ درصد و در معیار اف۱ به میزان ۸ تا ۱۸ درصد بهبود بخشیده است. همچنین برای تعیین داده‌های مرتبط آموزشی برای آموزش مدل‌ها، سیستم پیشنهادی از یک

* نویسندهٔ مسئول

مدل طبقه‌بندی داده اســتفاده می‌کند. نمرات ارزیابی مدل طبقه‌بندی داده نتایج بالای ۸۰٪ در معیارهای اف۱، پوشش و صحت را نشان داده است. همچنین نمرات ارزیابی تعیین داده‌های غیر مرتبــط نتایج بالای ۹۷٪ در معیارهای اف۱، پوشش و صحت را نشان داده است.

واژه‌های کلیــدی: مبدل‌هــا، تحلیل داده‌هــای بدون ساختار، پردازش زبان طبیعی، امنیت شبکه‌های کامپیوتری، مدل‌های زبانی از پیش‌آموزش دیده

۱. مقدمه

اخیراً مدل‌های زبانــی[1] از پیش آموزش‌دیده در حوزه پردازش زبــان طبیعی به‌طور قابل‌توجهــی افزایش یافته است. سیستم‌های پاسخ به سؤال نوعی از مدل‌های زبانی هستند که در سال‌های اخیر موردتوجه زیادی قرار گرفته است[۳۲]. بیشتر آثار مربوط به پاسخ به سؤال در ابتدا برای زبان انگلیسی پیشنهاد شده اســت، اما برخی از مطالعات تحقیقاتی اخیراً بر روی زبان‌های غیرانگلیسی انجام شده اســت[۱۰]. فهم زبان طبیعی زیرشاخه‌ای از پردازش زبان طبیعی و در رده ســخت‌ترین مســایل هــوش مصنوعی قرار می‌گیرد[۱]. با توجه به کاربردهای عملی این فناوری تمایل زیادی به آن وجــود دارد و می‌توان از کاربردهای آن، اســتدلال خودکار، ترجمه ماشینی، دسته‌بندی اسناد، جمع‌آوری خبر، پاســخ به پرســش‌ها، تشــخیص گفتار، پردازش تصویر(مانند نویسه‌خوان نوری)، فهرست بندی و تحلیل محتوا در ابعاد بزرگ اشاره کرد.

از مهم‌ترین کاربرد پــردازش زبان طبیعی می‌توان به استخراج دانش نهفته از متون بدون ساختار (نوشته متنی ســاده) اشاره نمود. متون بدون ســاختار به جهت این که مدل داده‌ای از پیش تعیین شده ندارند، در استخراج دانش با چالش‌های زیادی همراه هستند. طبق منابع علمی مختلف داده‌های بدون ســاختار بیــش از ۸۰ درصد کل داده‌ها را تشــکیل می‌دهند[۱-٤] و با این که ســایر پژوهشگران در ایــن زمینه تحقیقاتــی انجام داده‌اند اما هنــوز نیاز به کار بیشتر در این حوزه (به خصوص در زبان فارسی) وجود دارد. مبدل‌ها[2] از ســازوکار خودتوجه استفاده می‌کنند و این امکان را به ماشــین می‌دهد تــا در محدوده‌های محلی و عمومی متمرکز شود و عمل یادگیری را علاوه بر توجه به مکان فعلی کلمه، با در نظر داشــتن کل جمله انجام دهد. تحقیقات اخیر در حوزه مدل‌های زبانی از پیش‌آموزش داده شده نشان داده‌اند این مدل‌ها در زمان پیش‌آموزش به‌طور ضمنی و بدون نظارت، دانش‌های زبانی را در بازنمایی‌های خود کد می‌کنند. به‌عنوان مثال برت[3] این دانش‌ها را به‌طور سلسله‌مراتبی فراگرفته است؛ به‌طوری‌که لایه‌های اولیه در مدل، دانش‌های زبانی سطح پایین مانند اطلاعات مربوط به موقعیت مکانی کلمات را کــد می‌کنند و لایه‌های میانی در اطلاعات نحوی مانند بازسازی درخت تجزیه جمله غنی‌تر هستند.

۱-۱. بیان مسئله پژوهش

در این پژوهش بــا به‌کارگیری تکنیک‌هــای یادگیری عمیق و پــردازش زبان طبیعی، مدلی آموزش داده شــده اســت تا با اســتفاده از دریافت ورودی در قالب داده‌های بدون ساختار، قادر به بررسی و ارایه پاسخ به موضوعات مرتبط با امنیت شبکه‌های کامپیوتری باشد. مدل مذکور از مقادیر زیادی داده‌های متنی جمع‌آوری شده برای آموزش اســتفاده نموده و با ایجــاد درک از نقش کلمه در جملات (مشــابه مغز انســان)، قادر به یادگیری، بسط و پاسخ به سؤالات است. محصول تولید شــده قابلیت تنظیم دقیق و به‌کارگیــری در همــه حوزه‌ها را دارد؛ امــا در این تحقیق بر روی مفاهیم امنیت شــبکه‌های کامپیوتری به‌خصوص حملات ســایبری کــه منجر بــه نقض امنیــت و خارج از سرویس شدن شــبکه‌های کامپیوتری می‌شــوند، تمرکز شده اســت. همچنین برای به‌روز بودن مدل، سابقه برخی موارد مرتبط با امنیت شبکه‌های سایبری مانند گزارش‌های تحلیلی در وب‌گاه ها از جمله مراکز آپا نیز به مدل آموزش داده می‌شــود. انتظار می‌رود مدل مذکور پس از آموزش،

1-Language Model

2-Transformers

3-Bert

توانایی‌های تحلیلی و اســتدلالی مناســبی در مورد تحلیل امنیت شبکه‌های کامپیوتری ارایه نماید.

ارایه خروجی به‌صورت قراردادی از نیازهای اطلاعاتی انجام می‌شود. به‌عنوان مثال یک روش این است که کاربر اعلان‌هایی را به ســامانه ارسال می‌کند و هوش مصنوعی ســامانه که در قالب مدل آموزش داده شــده، با توجه به مطالبی که از پیش یاد گرفتــه، در هر اعلان یک پاراگراف بسط ارایه می‌کند. روش دیگر ارایه نتایج به‌صورت پرسش و پاسخ اســت و برای این منظور نیاز است مجموعه داده مرتبط ایجاد و به مدل آموزش داده شود.

۲-۱. ضرورت و اهمیت پژوهش

در ایــن پژوهش بــا به‌کارگیــری مبدل‌هــا از ترکیب شــبکه‌های یادگیری عمیــق و تکنیک‌های پــردازش زبان طبیعی اســتفاده شــده تا یک مدل هــوش مصنوعی برای ارزیابــی امنیت شــبکه‌های کامپیوتری ایجاد شــود. مدل با اســتفاده از مجموعه داده ایجاد شــده در حوزه امنیت شــبکه‌های کامپیوتری، آموزش‌دیده است و از سرعت و دقت بالاتری نســبت به سایر روش‌های مشابه برخوردار است. با توسعه این مدل محققان حوزه سایبری می‌توانند پاســخ ســؤالات تحلیلی خود را از انبوهــی از متن بدون ساختار استخراج کنند.

۳-۱. انگیزه‌ها و دستاوردهای پژوهش

اصلی‌تریــن انگیزه این پژوهش ایجــاد یک مدل هوش مصنوعی به‌منظور ارزیابی متون بدون ساختار در حوزه امنیت شبکه‌های کامپیوتری است. برای این منظور نیاز به یک مجموعه داده آموزشــی وجود دارد. با ایجاد مجموعه داده، سایر محققان می‌توانند باهدف تحلیل امنیت شبکه‌های کامپیوتری مدل‌هایی طراحی و پیاده‌سازی نموده و همچنین تصمیم به بهبود و گسترده‌تر کردن مدل و یا مجموعه داده نمایند.

در این پژوهش با ارایه سیستمی به جستجوی پاسخ‌ها در زمینه امنیت شبکه‌های کامپیوتری از متن بدون ساختار پرداخته شــده است. همچنین با تولید یک مجموعه داده به نام هوشــفا میزان دقت مدل‌ها در ارایه پاســخ را بسته به مدل‌های مختلف؛ در معیار تطبیق دقیق به میزان ۲۰ تا ۴۳ درصد و در معیار اف۱ به میزان ۸ تا ۱۸ درصد نسبت به پژوهش‌های قبلی بهبودیافته است.

۴-۱. تعاریف و مفاهیم پایه

یک مــدل مبدل جهانی کامل و کارآمد برای پاســخ به سؤال وجود ندارد [۵]. بهترین مدل برای یک مسئله، مدلی است که بهترین خروجی را برای یک مجموعه داده با وظیفه خاص تولید کند. از ســال ۲۰۱۷ تا ۲۰۲۳ تعداد پارامترها از 65M به 1/7T در مدل جی‌پی‌تی-٤ افزایش یافته اســت. روند رشد پارامترها در جدول زیر قابل مشاهده است:

جدول (۱). تکامل تعداد پارامترهای مبدل

مدل مبدل	مقاله	پارامتر
Transformer Base	Vaswani et al. (2017)	65M
Transformer Big	Vaswani et al. (2017)	213M
BERT-Base	Devlin et al. (2019)	110M
BERT-Large	Devlin et al. (2019)	340M
GPT-2	Radford et al. (2019)	117M
GPT-2	Radford et al. (2019)	345M
GPT-2	Radford et al. (2019)	1.5B
GPT-2	Brown et al. (2020)	175B
GPT-3	Brown et al. (2020)	175B
GPT-3.5	OpenAI.(2022)	175B
GPT-4	OpenAI.(2023)	1/7T

با توجه به ایــن که مدل‌ها در اندازه‌های مختلفی ایجاد می‌شــوند ممکن است برخی از نســخه‌های مدل‌های قبلی نسبت به مدل‌های جدید دارای پارامترهای بیشتری باشند. به‌عنوان‌مثــال، مدل کوچــک جی‌پی‌تی-۳ شــامل 125M پارامتر اســت که از یک مدل جی‌پی‌تی-۲ که شامل 345M پارامتر است کوچک‌تر است.

۱-۴-۱. معرفی بن‌سازهٔ اشتراک مدل‌های مبدل

هاگینگ فیس به یک بن‌ســازهٔ[4] مرکزی اطلاق می‌شود

4-Platform

که فرصتِ کشف، استفاده و کمک به ارتقای پیشرفته‌ترین مدل‌ها و مجموعه داده‌ها را به کاربران می‌دهد. این بن‌سازه از طیف گسترده‌ای از مدل‌ها میزبانی می‌کند و بیش از ده هزار موردِ آن در دسترس محققان قرار دارد. مدل‌های موجود در هاگینگ فیس‌هاب[5] به مبدل‌ها یا حتی پردازش زبان طبیعی محدود نیستند. به‌عنوان نمونه، مدل‌های مربوط به گفتار، بینایی ماشین و غیره نیز وجود دارند. هرکدام از این مدل‌ها در قالب یک منبع گیت[6] میزبانی می‌شوند. اشتراک‌گذاری مدل در هاب به معنای این است که عموم کاربران می‌توانند از آن استفاده کنند. دیگر کاربران مجبور نیستند مدل را آموزش دهند و کار اشتراک و استفاده بسیار آسان شده است. اشتراک‌گذاری و استفاده از مدل‌ها در هاب کاملاً رایگان است. البته اگر قصد اشتراک‌گذاری مدل‌های خصوصی مطرح باشد، می‌توان از طرح‌های پولی استفاده نمود. خط لوله[7] اصلی‌ترین شیء در کتابخانه مبدل‌ها است. خط لوله مدل را به مراحل پیش‌پردازش متصل کرده و امکان می‌دهد هر متنی را به‌طور مستقیم وارد و پاسخی قابل‌قبول دریافت نمود. خط‌لوله به‌صورت پیش‌فرض یک مدل خاص و آموزش دیده را که برای تحلیل احساسات به زبان انگلیسی تنظیم شده است را انتخاب می‌کند.

۱-۴-۲. سوگیری و محدودیت‌ها در مدل‌ها

مدل‌ها علی‌رغم این‌که ابزارهای قدرتمندی هستند دارای سوگیری[8] و محدودیت هم هستند. بزرگ‌ترین محدودیت این است که برای انجام پیش‌آموزش روی حجم زیادی از داده‌ها، اغلب باید تمام محتوای ممکن شامل بهترین و بدترین موارد موجود در اینترنت جستجو و به کار گرفته شوند. اگرچه به‌ندرت می‌توان مدلی مثل برت را در میان مدل‌های مبدل یافت اما حتی چنین مدلی نیز می‌تواند پیش‌بینی‌های دارای سوگیری ارایه دهد. هنگام استفاده از این مدل باید توجه کرد که مدل اولیه به‌راحتی می‌تواند متنی تولید کند که دارای سوگیری یا محدودیت است و ممکن است حتی حاوی محتوای نژادپرستانه باشد. همچنین ممکن است تنظیم دقیق مدل بر روی‌داده‌های جدید نیز نتواند این سوگیری ذاتی را از بین ببرد.

۱-۴-۳. مجموعه داده

در دسترس بودن مجموعه داده‌های ارزیابی، باکیفیت بالا یک ضرورت برای ارزیابی قابل اعتماد از پیشرفت در وظایف و دامنه‌های مختلف فهم زبان طبیعی است[٦]. مجموعه داده به مجموعه‌ای از داده‌های آماری یا رایانه‌ای مربوط به یک پایگاه داده گفته می‌شود، که با هدف یکپارچه نمودن داده‌ها، محتویات آن در قالب یک جدول پایگاه داده یا یک ماتریس داده‌ای، تنظیم و مرتب شده و در آن ستون پایگاه داده، نشان‌دهنده یک متغیر خاص و هر ردیف به یکی از اعضای مجموعه داده موردنظر مرتبط می‌باشد.

۱-۴-۴. مطالعه موردی

یک آزمایشگاه تحقیقاتی هوش مصنوعی به نام اوپن‌ای‌آی که توسط ایلان ماسک تأسیس‌شده، آخرین نسخه سیستم پردازش زبان طبیعی مبتنی بر هوش مصنوعی به نام جی‌پی‌تی-۳ را راه‌اندازی کرده که می‌تواند زبان انسان را تقلید کند. جی‌پی‌تی-۳ می‌تواند برای تفسیر متن پیچیده، ایجاد هشدار، شناسایی متن تکراری، خودکارسازی مواردی که ممکن است برای کاربران انسانی آشکار نباشد، استفاده شود. همچنین در سایر حوزه‌ها می‌تواند از تحلیل گزارش‌های پزشکی تا تولید شعر نیز استفاده کرده و متنی تولید کند که اغلب از متن انسانی قابل تشخیص نیست. به‌عنوان مثال، یک وب نویس از جی‌پی‌تی-۳ برای نوشتن پست‌های وب‌نوشتی استفاده کرده و عملکرد فوق‌العاده‌ای را در هکرنیوز[9]، رقم‌زده است.

با وجود موارد ذکر شده خروجی‌ها از عجیب تا عاقلانه

5-Huging Face Hub
6-Git
7-Pipeline
8-Bias
9-Hacker News

متفاوت هستند و وجود تناقض در برخی موارد دور از انتظار نیست و نیاز به کارهای بیشتر وجود دارد. این مدل در هسته خود از مبدل استفاده می‌کند و برای توابع تولید متن، پاسخگویی به سؤال، تحلیل احساسات، شناسایی موجودیت‌های نامدار، ترجمه ماشینی و خلاصه کردن متن طراحی شده است. مدل جی‌پی‌تی-۳ شامل یادگیری عمیق با ۱۷۵ میلیارد پارامتر است و بر روی مجموعه داده‌های متنی بزرگ شامل صدها میلیارد کلمه آموزش داده شده است. برای آموزش محتوای متنی صفحات وب از سال ۲۰۱۶ تا ۲۰۱۹ مصرف شده که دارای ۴۵ ترابایت داده متنی فشرده است. مدل‌های قبلی جی‌پی‌تی-۱ از پایگاه داده ۵ گیگابایتی متشکل از محتوای متنی بیش از ۷۰۰۰ کتاب توسط دانشگاه تورنتو و ام‌آی‌تی آموزش دید و تعداد کلمات حدود یک‌میلیون کلمه بود. همچنین مدل جی‌پی‌تی-۲ برای آموزش محتوای متنی یک مجموعه داده بومی، متشکل از هشت میلیون صفحه وب با ۴۰ گیگابایت داده را دریافت کرد. همان‌طور که ذکر شد جی‌پی‌تی-۳ مهارت فوق‌العاده‌ای دارد و می‌تواند با دریافت یک جمله یا حتی چند کلمه، پنج پاراگراف کامل را نگارش کند. در حال حاضر این مدل انحصاراً در اختیار مالکان(اوپن‌ای‌آی[10] و مایکروسافت[11]) است و به سایرین استفاده محدود داده می‌شود تا از پی‌آمدهای ناخواسته مانند جعل عمیق جلوگیری شود. مدل جی‌پی‌تی-۴ در سال ۲۰۲۳ با تعداد پارامترهای بیشتر منتشر شده است. توجه به‌سرعت به‌عنوان عنصری حیاتی در مدل‌های پردازش زبان طبیعی موجب شد گوگل با استفاده از مبدل مدل برت را ایجاد کند که یکی دیگر از مدل‌های زبانی بسیار موفق است.

این مدل‌ها در همه حوزه‌ها خلاقیت زیادی را ایجاد می‌کنند و می‌توان از آن‌ها در همه زمینه‌ها استفاده نمود. ممکن است برای استفاده از مدل‌ها در حوزه صنعتی نیاز به امضای توافق‌نامه عدم افشای اطلاعات باشد. در منشورهای برخی شرکت‌های هوش مصنوعی بیان شده وظیفه امانت‌داری اولیه در قبال بشریت است و شرکت برای برنده شدن در مسابقه هوش مصنوعی، ایمنی را به خطر نمی‌اندازد.

10-Open AI
11-Microsoft

۲. سوابق پژوهش

طی بیش از صدسال گذشته بسیاری از محققان روی انتقال توالی و مدل‌سازی زبانی کارکرده‌اند و ماشین‌ها به‌تدریج یاد گرفتند که چگونه توالی‌های احتمالی کلمات را پیش‌بینی کنند. برای استناد به همه مواردی که این اتفاق را انجام داده است، به‌اندازه یک کتاب کامل زمان لازم است؛ لذا در این بخش مواردی بررسی می‌شوند که زمینه ورود به مباحث مبدل‌ها در پردازش زبان طبیعی فراهم شود. در ادامه، پیشینه پردازش زبان طبیعی که منجر به ظهور مبدل‌ها شد را مورد بررسی قرار داده و مشاهده می‌کنیم که چگونه مدل مبدل اختراع شده توسط تیم تحقیقاتی گوگل چندین دهه تحقیق، توسعه و پیاده‌سازی‌های پردازش زبان طبیعی را متحول کرده است.

در اوایل قرن بیستم، آندری مارکوف[12] مفهوم مقادیر تصادفی را معرفی و نظریه‌ای از فرآیندهای تصادفی ایجاد کرد. این نظریه در هوش مصنوعی با عناوین فرآیندهای تصمیم‌گیری مارکوف[13]، زنجیره‌های مارکوف[14] و فرآیندهای مارکوف[15] شناخته می‌شوند. در سال ۱۹۰۲، مارکوف نشان داد که می‌توان عنصر بعدی یک زنجیره را تنها با استفاده از آخرین عنصر پیش‌بینی نمود. در سال ۱۹۱۳، مارکوف عمل پیش‌بینی برای حروف آینده را با استفاده از توالی‌های گذشته زنجیره در یک مجموعه داده ۲۰،۰۰۰ حرفی به کار برد. با این که مارکوف رایانه‌ای نداشت اما موفق به اثبات نظریه‌ای شد، که هنوز هم در هوش مصنوعی مورد استفاده قرار می‌گیرد.

کلود شانون[16] در [۷]، نتایج تحقیقات خود را تحت عنوان «تئوری ریاضی ارتباطات شانون» در سال ۱۹۴۸منتشر

12-Andrey Markov
13-Markov Decision Processes (MDPs)
14-Markov Chains
15-Markov Processes
16-Claude Shannon

نمـــود. او هنگام ایجاد رویکرد خود در مدل‌ســـازی توالی، چندین بار به نظریه مارکوف اشاره می‌کند. شانون زمینه‌ای را برای یک مدل ارتباطی بر اساس کدگذاری منبع، ارسال و کدگشایی یا کدگشایی معنایی در دریافت انجام داد.

آلن تورینـــگ[17] در [۸] نتایج تحقیقات خود تحت عنوان «دســـتگاه محاسباتی و هوش» را در ســـال ۱۹۵۰ منتشر کرد. آلن تورینگ این مقاله را بر اســـاس هوش دســـتگاه بـــر روی دســـتگاه تورینگ فوق‌العاده موفـــق که پیام‌های آلمانی را رمزگشـــایی می‌کرد، پایه‌گذاری کرد. واژه هوش مصنوعی برای اولین بار توسط جان مک‌کارتی[18] در سال ۱۹۵٦ مورد اســـتفاده قرار گرفت. با این حال، آلن تورینگ در دهه ۱۹٤۰ هوش مصنوعی را برای رمزگشایی پیام‌های رمزگذاری شده به زبان آلمانی پیاده‌سازی کرد.

از نخستین سیســـتم‌های پرسش و پاسخ سیستمی به نام بیس‌بال[19] بود که در سال ۱۹٦۱ برای پاسخ به سؤالات کاربران در رابطه با لیگ سالانه در آمریکا طراحی شد[۹]. برنامه رایانه‌ای مذکور قادر بود سؤالات کاربران در زمینه تاریـــخ و زمان برگزاری مســـابقات و همچنیـــن امتیازات کســـب شده پاسخ دهد. کاربران ســـؤالات را با استفاده از کارت منگنه مطرح کرده و سیســـتم کلمات و اصطلاحات موجود را در یک لغت‌نامه جستجو و با تحلیل ساختاری و دستورزبانی پرسش پاسخ مناسب را تهیه می‌نمود.

جـــان هاپفیلـــد[20] در ســـال۱۹۸۲، شـــبکه‌های عصبی بازگشـــتی[21] معروف به شـــبکه‌های هاپفیلد یا شـــبکه‌های عصبی «انجمنی» را معرفی کـــرد. هاپفیلد برای این کار از لیتل[22] که در ســـال ۱۹۷٤ «وجود حالات مداوم در مغز» را نوشته است الهام گرفته است. شبکه‌های عصبی بازگشتی تکامـــل یافتند و LSTMها پدیدار شـــد. شـــبکه‌های عصبی بازگشتی حالات مداوم یک دنباله را به‌طور مؤثر به خاطر دارد:

17-Alan Turing
18-John McCarthy
19-BASEBALL
20-John Hopfield
21-Recurrent Neural Networks (RNNs)
22-Little

یان‌لی‌کان[23] در دهه ۱۹۸۰ شـــبکه عصبی هم‌آمیختی[24] را طراحی کرد. او این شـــبکه عصبـــی را برای توالی متن اعمال کرد به‌طوری‌که سپس این شبکه‌های عصبی به‌طور گسترده‌ای برای انتقال و مدل‌سازی توالی استفاده شدند. آن‌ها همچنین بر اساس "وضعیت‌های ماندگار" اطلاعات را بر اساس لایه به لایه جمع می‌کنند.

در ســـال ۱۹۹۰ یان‌لی‌کان لی‌نـــت- ۵[25] را ایجاد کرد و منجر به بسیاری از مدل‌های شبکه‌های عصبی هم‌آمیختی شد که امروزه می‌شناسیم. پس‌ازآن توسعه‌دهندگان برای تجزیه‌وتحلیـــل توالی‌های طولانی نیاز بـــه افزایش قدرت رایانه داشتند و با استفاده از رایانه‌های قدرتمندتر، راه‌هایی برای بهینه‌سازی شیب[26] پیدا می‌کردند. در این زمان به نظر می‌رســـید هیچ‌چیز دیگری نمی‌تواند برای پیشرفت بیشتر انجام شود.

پس از ســـی سال در دســـامبر ســـال ۲۰۱۷، مبدل‌ها نوآوری حیـــرت‌آوری را به همـــراه آوردنـــد و امتیازات چشمگیری را در مجموعه داده‌های استاندارد ایجاد کردند. وسوانی[27] و همکاران در سال ۲۰۱۷ نتایج کارهای خود را با عنوان "توجه، تمام چیزی که نیاز دارید" در [۱۰] منتشر کردند. آن‌ها معماری شـــبکه عصبـــی جدیدی به نام مبدل بر اساس سازوکار توجه ارایه نمودند و عملکرد مدل‌هایی مانند جی‌پی‌تی و برت با تکیه‌بر این معماری کاملاً بهتر از شـــبکه‌های قبلی بود. این رویکرد به‌طور گسترده از سایر رویکردها پیشی گرفت و تقریباً تمام مدل‌های اخیر مبتنی بر معماری مبدل هستند. این یافته‌ها در تیم تحقیقاتی گوگل و پروژه گوگل برین[28] انجام شد و سریع‌تر از معماری‌های قبلی، آموزش‌دیده و با کســـب نتایج ارزیابی بالاتر به یک مؤلفه اصلی پردازش زبان طبیعی تبدیل شدند.

دولین[29] و همکاران در [۱۱] بیان کردند پیشـــرفت‌های تجربی اخیر ناشـــی از انتقال یادگیری بـــا مدل‌های زبانی

23-Yann Le Cun
24-Convolutional Neural Network (CNN)
25-LENET-5
26-Gradients
27-Vaswani
28-Google Brain
29-Devlin

نشـــان داده است که پیش آموزشـــی غنی و بدون نظارت بخش جدایی‌ناپذیر بســـیاری از سیســـتم‌های درک زبان اســـت. همچنین ایـــن نتایج وظایـــف پایین‌دســـت را قادر می‌سازند تا از معماری‌های تک‌جهتی عمیق بهره‌مند شوند و ســـهم اصلی ما تعمیم بیشتر این یافته‌ها به معماری‌های دوسویه عمیق اســـت. مدل پیش‌آموزش‌دیده این امکان را فراهم می‌کند تا مجموعه وسیعی از وظایف پردازش زبان طبیعی انجام شـــود. آن‌ها در ادامه یک مدل چندزبانه برت معرفی کردند که از زبان فارسی نیز پشتیبانی می‌کند و بر پایه مبدل‌ها اســـت. این مدل برخلاف مدل‌های ارایه شـــده زبانی اخیر برای آموزش دوســـویه عمیـــق، از متن بدون برچســـب با شرطی‌سازی مشترک در بخش چپ و راست در همه لایه‌ها طراحی شـــده است. مدل برت آموزش‌دیده می‌تواند تنها با یک لایه خروجی اضافی تنظیم شده تا مدل برای طیف وســـیعی از وظایف به کار گرفته شود. وظایف موردنظر مانند پاسخ به سؤال و استنتاج زبان است.

دایی[۳۰] و همکاران در [۱۲] نشـــان دادند که مدل زبانی از قبل آموزش داده شـــده برای بهبود بسیاری از کارهای پردازش زبان طبیعی مؤثر اســـت. آن‌ها در این پایان‌نامه دو روش بـــرای اســـتفاده از داده‌های بدون برچســـب به منظور بهبود شبکه تکرارشونده ارایه نمودند. اولین روش پیش‌بینی آنچه در دنباله بعدی می‌آید که یک مدل زبانی در پردازش زبان طبیعی است و روش دوم استفاده از کدگذار است که توالی ورودی را به یک توالی ورودی برداری قابل پیش‌بینی نگاشت می‌کند.

درویشی[۳۱] و همکاران در [۱۳] یک مجموعه داده پاسخ به سؤال فارســـی به نام پی‌کیو ای دی[۳۲] از داده‌های درک مطلب در مقالات ویکی‌پدیا فارســـی ارایه داده‌اند. پرسش شـــامل ۸۰،۰۰۰ ســـؤال همراه با پاســـخ‌های است که ۲۵ درصد ســـؤالات غیرقابل پاسخ هســـتند. سؤالات غیرقابل پاسخ موجب می‌شوند مدل زمانی که پاسخ مناسبی وجود نداشته باشد، از پاسخ امتناع کند. آن‌ها هدف از انتشار این مجموعـــه داده‌ها را انجام تحقیقـــات در زمینه درک مطلب فارسی و توسعه سیستم‌های پاسخ به سؤال فارسی بیان نموده‌اند. آزمایش‌های انجام شـــده امتیـــاز ۷۴/۸٪ تطابق دقیق[۳۳] و امتیاز ۸۷/۶٪ در معیار اف۱[۳۴]را نشان می‌دهد که می‌تواند به‌عنوان نتایج پایه برای تحقیقات بیشتر در مورد پرسش و پاسخ فارسی مورد استفاده قرار گیرد.

دوان[۳۵] و همکاران در [۱۴] به بررسی مدل‌های زبان از پیش آموزش‌دیده پرداخته‌اند. بررســـی در تحقیقات شامل معرفی تاریخ توسعه و دستاوردهای مدل‌های زبان از پیش آموزش‌دیـــده، ارایه چندین مدل زبان از پیش آموزش‌دیده مانند برت و مدل‌های چندحالته همراه با نمودار دانش برای تولید زبان طبیعی معرفی شـــده اســـت. در نهایت بررسی آینـــده زبان‌های از پیش آموزش‌دیده انجام شـــده اســـت. انتظار می‌رود این تحقیق یـــک راهنمای عملی برای درک، استفاده و توسعه زبان‌های از پیش آموزش‌دیده فراهم کند.

سینگ[۳۶] و همکاران در [۱۵] خلاصه‌ای جامع و مفصل از مدل‌های اصلی زبانی ارایه می‌دهند که منجر به ایده‌های نوآورانـــه در پردازش زبان طبیعی شـــده اســـت. از زمان راه‌اندازی سازوکار توجه و معماری مبدل پیشرفت پردازش زبان طبیعی نمایی شـــده است. آن‌ها یک نقشه ذهنی سطح بالا از طبقه‌بندی مدل ارایه نموده‌اند. این طبقه‌بندی در درجه اول مبتنی بر معماری‌های مشتق شده از مبدل‌ها است که برای کارهـــای تخصصی مانند درک و تولید زبان، کاهش اندازه مـــدل، بازیابی اطلاعات، مدل‌ســـازی توالی طولانی و ســـایر تکنیک‌های کاهش مدل عمومی ایجاد شده است. مدل‌های زبان اخیر در درجه اول با دســـتیابی به عملکرد بالاتر به‌ســـوی منابع محاســـباتی عظیم هدایت می‌شوند. بنابراین، مقیاس‌بندی مدل مســـیر طبیعی در صنعت است. تلاش‌هـــای قابل‌توجهی برای مهندســـی مدل‌های با اندازه معقول و محاســـبه توجه کارآمد برای سرعت بخشیدن به همگرایی مدل منجر به تأخیر کمتر در مدل‌ها انجام شـــده

30- Dai
31- Darvishi
32- PQuAD

33-Exact Match (EM)
34-F1- score
35-Duan
36-Singh

است. در این تحقیقات یک روش مؤثر برای مدل‌های بزرگ به‌منظور دستیابی به راندمان محاسباتی معرفی می‌شود.

سیار[۳۷] و همکاران در [۱۶] برنامه‌ای به‌منظور شناسایی خطرات امنیت ملی در برنامه‌های وب ارایه نمودند. آن‌ها بیان کردند با توجه به این که برنامه‌های وب شامل طیف وسیعی از اطلاعات شخصی، مالی، دفاعی و سیاسی هستند ممکن است حادثه‌ای مانند ویکی‌لیکس[۳۸] را رقم بزنند. از این رو ممکن است ثبت چنین اطلاعاتی به یک مشکل حیاتی تبدیل شده و به خطر امنیت ملی تبدیل شود. در تحقیقات آن‌ها یک راه‌حل برای این مشکل با ارایه مدلی جدید انجام شده که قادر به تشخیص درخواست‌های اچ‌تی‌تی‌پی[۳۹] عادی و درخواست‌های اچ‌تی‌تی‌پی غیرعادی است. این مدل از تکنیک‌های کدگذار پردازش زبان طبیعی دوسویه مبدل برت و روش‌های یادگیری عمیق استفاده می‌کند. نتایج تحقیقات در طبقه‌بندی درخواست‌های عادی و غیرعادی نشان داد روش پیشنهادی به نرخ موفقیت بیش از ۹۹/۹۸٪ و امتیاز اف۱ بیش از ۹۸/۷۰٪ دست می‌یابد. همچنین زمان تشخیص به‌طور قابل‌توجهی کم‌تر از سایر روش‌های ارایه شده است.

فراهانی[۴۰] و همکاران در [۱۷] یک برت یک‌زبانه برای زبان فارسی به نام پارس‌برت[۴۱] ارایه کردند که عملکرد بهتری در مقایسه با سایر معماری‌ها و مدل‌های چندزبانه دارد. همچنین با توجه به محدود بودن مقادیر داده‌ها برای وظایف پردازش زبان طبیعی در زبان فارسی، یک مجموعه داده بزرگ برای کارهای مختلف و پیش آموزش مدل تدوین شده است. پارس برت در تمام مجموعه داده‌ها نمرات بالاتری کسب کرده و عملکرد آن فراتر از برت چندزبانه و سایر کارهای قبلی در تحلیل احساسات، طبقه‌بندی متن و شناسایی موجودیت‌های نامدار است. با توجه به این که اکثر مدل‌ها معمولاً بر انگلیسی متمرکز هستند، زبان‌های دیگر به مدل‌های چندزبانه با منابع محدود وابسته هستند. از این‌رو پارس‌برت در زبان فارسی می‌تواند گزینه مناسبی برای کارهای پردازش زبان طبیعی باشد.

طاهر[۴۲] و همکاران در [۱۸] یک مدل به نام بهشتی ان‌ای‌آر[۴۳] بر پایه برت به منظور شناسایی موجودیت‌های نامدار[۴۴] ارایه نموده‌اند. این مدل در ارزیابی وظیفه ۷ از نسیوآرال[۴۵] ۲۰۱۹ که در ارتباط با شناسایی موجودیت‌های نامدار در زبان فارسی است، رتبه دوم را کسب کرده است. نتایج شامل ۸۳/۵ و ۸۸/٤ در امتیاز اف۱ از مجموعه داده کنلل[۴۶] در سطح عبارت و کلمه است.

بلبل‌زبان یک مدل تولید شعر فارسی است که در [۱۹] قابل دسترس برای عموم قرار دارد. این مدل با توجه با ۲۵٦ نویسهٔ قبلی قادر به تولید شعر با درک قافیه و ریتم است اما شعر تولید شده در عین حال ممکن است دارای انسجام معنایی ضعیف باشد.

آلام[۴۷] و همکاران در [۲۰] یک کتابخانه منبع باز پایتون به‌منظور استخراج اطلاعات تهدید سایبری از متن بدون ساختار از منابع ناهمگن اینترنت ارایه کردند که برای شناسایی موجودیت‌های نامدار در امنیت سایبری کاربرد دارد. این ابزار که سای‌ان‌ای‌آر[۴۸] نام دارد، مبتنی بر مبدل‌ها است و برای استخراج موجودیت‌های مرتبط با امنیت سایبری شاخص‌های مختلف فنی و مدل‌های موجود عمومی را ترکیب می‌کند. سای‌ان‌ای‌آر رویدادها را از دامنه وسیعی از پیکره اطلاعات تهدید استخراج می‌کند. کاربران می‌توانند برای رفع نیازهایشان پیش‌بینی‌ها را از چندین رویکرد مختلف ترکیب کنند. این کتابخانه به‌طور عمومی منتشر شده است.

گوا[۴۹] و همکاران در [۲۱] روشی را برای غلبه بر چالش استخراج اطلاعات حساس از داده‌های حجیم بدون ساختار ارایه کرده‌اند. این روش از استخراج داده‌های حساس

37-Seyyar
38-WikiLeaks
39-HTTP
40-Farahani
41-ParsBERT

42-Taher
43-Beheshti-NER
44-Named Entity Recognition
45-NSURL
46-CONLL
47-Alam
48-CyNER
49-Guo

مبتنی بر الگو (آی‌پی[50]، پســت الکترونیکی، کلید و ...) و از مدل برت برای اســتخراج اطلاعات حساس استفاده نموده است. در واقع آخرین نسخه برت برای انجام تعبیه کلمات اســتفاده شده و اطلاعات حســاس را با استفاده از شبکه عصبی BiLSTM و سازوکار توجه استخراج می‌کند. نتایج ارزیابی اف۱ در این کار نشــان داد که روش پیشــنهادی با دقت ۹۹/۱۵ درصد بســیار موفق‌تر از روش‌های فردی تحلیل محتوا عمل کرده اســت. به علاوه حدود یک‌میلیون متن مورد تحلیل قرارگرفته اســت و ثابت شــده که روش پیشــنهادی می‌تواند اطلاعات حساس را از داده‌های بدون ساختار به‌طور مؤثر استخراج کند.

ژوانگ[51] و همــکاران در [۲۲] مطرح کردند برای ورود مقدار زیــادی داده‌ها به کامپیوتر به این امید که بتواند یاد بگیــرد. ابتدا باید داده‌ها را آماده کنیم تا برای رایانه راحت باشد الگوها و استدلال‌ها را کشف کند. همچنین نیاز است با اضافه کردن فراداده‌های مربوط به مجموعه داده اصلی به این هدف دست یابیم. در این مقاله استخراج اطلاعات متنی ساخت‌یافته از متن نامرتب یا غیر ساخت‌یافته موردتوجه قرارگرفته و به استخراج اطلاعات بر اساس پردازش زبان طبیعی، سیســتم بازیابی اطلاعات وب، اســتخراج روابط متنی و عملکرد پردازش خطلوله‌ای[52] پرداخته شده است.

۳. سیستم پیشنهادی

سیستم پیشــنهادی یک مدل هوش مصنوعی است که از یک مبدل آموزش‌دیــده در زمینه تحلیل داده‌های بدون ســاختار در حوزه شــبکه‌های کامپیوتری تشــکیل شده اســت. آموزش مبدل بر روی ســؤالات احصاء شــده در رابطه با امنیت شبکه‌های کامپیوتری انجام شده و سیستم می‌تواند از طریق پایگاه دانش خود و یا با دریافت گزارش به‌صورت مستقیم، پاسخ سؤالات را از متن بدون ساختار اســتخراج نماید. سؤالات آموزشــی در قالب یک مجموعه داده شامل۳/۳۵۶ رکورد ایجاد شده است.

با توجه به این که نتایج تحقیقات نشان داده‌اند مبدل‌هایی که بر روی داده‌های مهندســی نرم‌افــزار آموزش دیده‌اند دارای عملکرد بهتری در حوزه تخصصی هســتند، ۱/۹۹۸ رکــورد مربوط بــه داده‌های فناوری اطلاعــات و ۶/۸۰۳ رکورد مربوط به داده‌های عمومی از سایر مجموعه‌داده‌ها برای افزایش درک عمومی مدل بــه مجموعه داده افزوده شده است. در مجموع رکوردهای مجموعه داده به ۱۲/۱۵۷ رسیده است.

چنــد مــدل مختلف مبدل بــر روی ایــن مجموعه‌داده آموزش داده شــده و عملکــرد آنها نیــز در بخش بعدی مقایسه شده است.

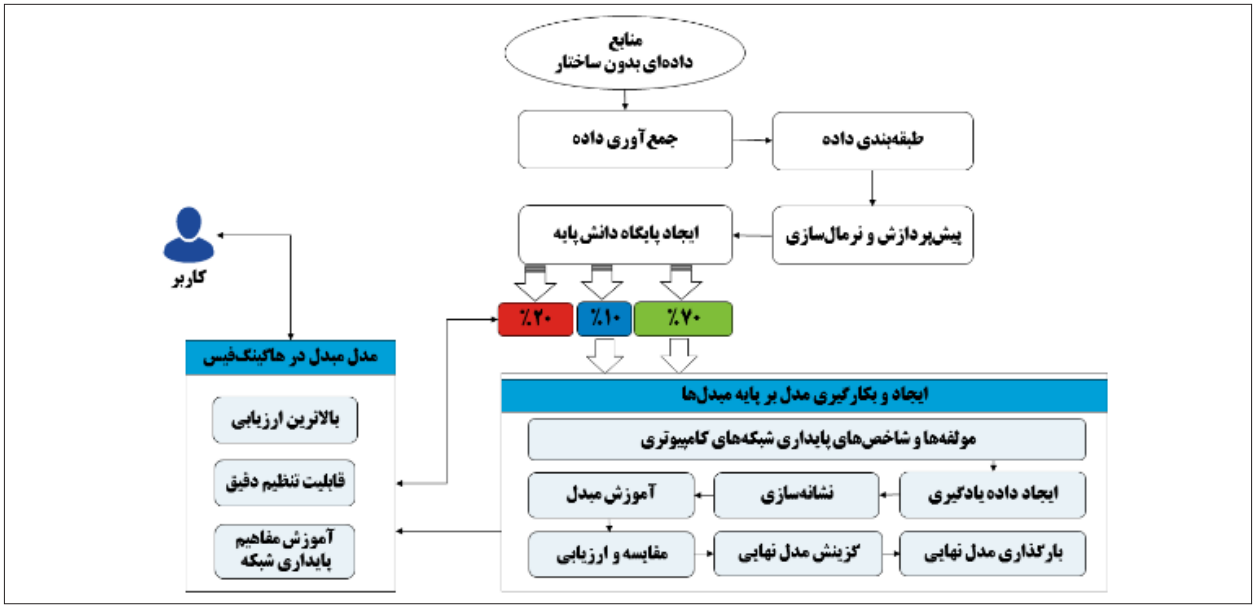
۳-۱. سیستم‌های پاسخ به سوال

یک روش مناســب با یک مدل متوســط اغلــب نتایج کارآمدتری نســبت به یک روش ناقص با مدل عالی ایجاد می‌کنــد. به‌عنوان‌مثــال، در یــک پروژه تولید موشــک و فضاپیما، پرســیدن یک سؤال از یک مدل هوش مصنوعی به معنای به دســت آوردن یک پاســخ دقیق اســت. فرض می‌شــود کاربر باید در مورد گــزارش صد صفحه‌ای در مورد وضعیت نازل و محفظه احتراق مجدد یک موشــک، سؤالی بپرسد. این ســؤال می‌تواند خاص باشد، مانند این که "آیا وضعیت خنک‌کننده قابل‌اعتماد اســت یا خیر؟" یک مدل هوش مصنوعی قابل‌اعتماد، در پس‌زمینه به یک پایگاه دانش داده‌ها و قوانین متصل می‌شــود تا پاسخ را بررسی کند. مدل هوش مصنوعی یک پاسخ به زبان طبیعی شفاف و قابل‌اعتمــاد و احتمالاً با صدای انســانی می‌تواند ایجاد نماید.

۳-۲. سخت‌افزار و محیط کدنویسی

برای دسترســی بــه قــدرت پردازشــی موردنیاز به منظور آموزش مدل از بن‌ســازهٔ گوگل کولب[53] اســتفاده شــده اســت. برای اســتفاده از گوگل کولب نیاز به نصب برنامه و یا پیکربندی‌های پیچیده نیســت و تنها کافی است یــک نوت‌بوک کولب در مرورگر راه‌اندازی و اجرا شــود.

50-IP
51-Zhuang
52-Pipeline
53-Google Colab



شکل (۱). روندنمای سیستم پیشنهادی

کولب امکان استفاده از سخت‌افزارهای شتاب‌دهنده مانند جی‌پی‌یو[54] و تی‌پی‌یو[55] را فراهم می‌کند. استفاده از شتاب‌دهنده برای همیشه رایگان نیست و بعد از آموزش چند مدل شتاب‌دهنده برای مدتی از کاربر گرفته می‌شود و برای پردازش بیشتر کولب‌پرو تهیه شده است. در ایران شرکت‌ها و مؤسساتی وجود دارند که با دور زدن تحریم اکانت کولب پرو را روی حساب ایمیل افراد شارژ می‌کنند. برای فضای ذخیره‌سازی از گوگل درایو استفاده شده و مجموعه داده‌ها و نتایج جمع‌آوری اطلاعات بر روی این محیط قرار داده شده است. گوگل درایو ۱۵ گیگابایت فضا روی بستر ابری در اختیار هر کاربر قرار می‌دهد که بر انجام تحقیقات اولیه و آموزش مدل کافی است.

۳-۳. روندنمای سیستم پیشنهادی

روندنمای سیستم پیشنهادی در شکل (۱) ترسیم شده است. همان‌طور که در روندنما مشخص است کاربر با ارایه درخواست به مدل هوش مصنوعی پاسخ سوالات را از پایگاه دانش داده‌های بدون ساختار استخراج می‌کند.

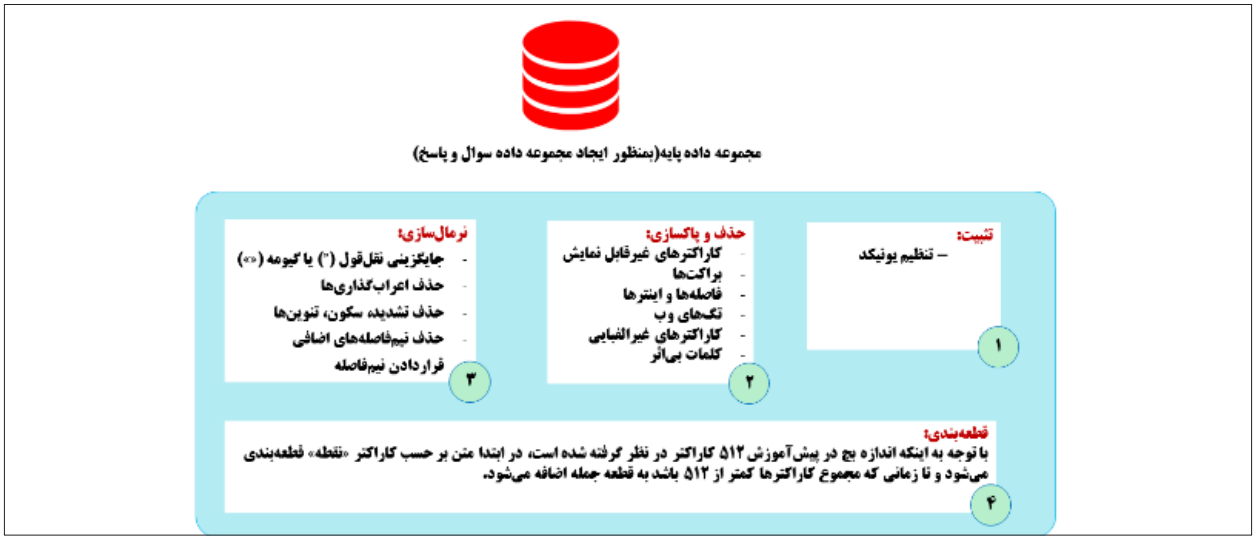
۳-۳-۱. منابع داده‌ای بدون ساختار

این منابع شامل وب، شبکه‌های اجتماعی، مجموعه‌داده‌های جمع‌آوری شده متنی و سایر موارد هستند. منابع برخط مانند وب توسط روش پیشنهادی جمع‌آوری و ذخیره می‌شوند تا در مراحل بعدی مورد استفاده قرار گیرند.

۳-۳-۲. جمع‌آوری داده‌ها

برخی از مجموعه داده‌ها آماده برای آموزش مدل وجود دارند. اما این مجموعه داده‌ها ایستا هستند و بدون ارایه روشی برای جمع‌آوری داده‌ها امکان به‌روزرسانی اخبار جدید در آن‌ها وجود ندارد. همچنین ممکن است این مجموعه داده‌ها تعداد محدودی از سبک‌های نوشتن و موضوعات را پوشش دهند. بنابراین، برای افزایش دامنه اطلاعات یادگیری روشی برای جمع‌آوری داده‌های جدید ارایه شده است. این روش بدون نیاز به کد نویسی و با استفاده از افزونه وب‌اسکرپر[56] در مرورگر گوگل کروم[57] انجام شده است. افزونه وب‌اسکرپر گوگل‌کروم امکان جمع‌آوری اطلاعات در بستر ابری[58] را داد. در جمع‌آوری اطلاعات در بستر ابری امکان استفاده از لیست پراکسی

54-GPU
55-TPU
56-Web scraper
57-Google chrome
58-Cloud



شکل (۲). پیش‌پردازش و نرمال‌سازی داده‌ها

وجود دارد. روش استفاده از بستر ابری به این شکل است که ابتدا نقشه وبگاه را در افزونه تنظیم نموده و سپس آن را در رابط وب بستر ابری وارد کرده و جمع‌آوری اطلاعات در بستر ابری آغاز می‌شود. برای استفاده بیش‌ازحد تعریف شده نیاز به پرداخت هزینه و تهیه حساب کاربری می‌باشد.

در مرحله جمع‌آوری اطلاعات حدود ۲٤,۰۰۰ متن خبر از وب‌گاه‌های اعتماد آنلاین، فردانیوز، خبرگزاری فارس و ... جمع‌آوری شده است. تمام این اخبار در حوزه سایبری نیستند و نیاز به طبقه‌بندی دارند. همچنین برخی از مجموعه داده‌های آماده مانند عصر ایران، خبرگزاری تسنیم، ایسنا و ... شناسایی و مورداستفاده قرار گرفت.

۳-۳-۳. طبقه‌بندی داده‌ها

در این مرحله یک طبقه‌بندی برای تفکیک اخبار مرتبط و اخبار غیر مرتبط با موضوع تحقیق کد نویسی شده است. منظور از اخبار غیرمرتبط اخباری هستند که در حوزه سایبری نیستند و آموزش آنها به مدل، خارج از حوزه این تحقیق است. به عنوان مثال از اخبار غیرمرتبط می‌توان به اخبار مذاکرات هسته‌ای اشاره نمود. این طبقه‌بندی که بر مبنای مبدل‌ها کد نویسی شده است با استفاده از مدل‌های مختلف بر روی ۹,٦٦٦ عنوان خبر در سه دوره[59] آموزش‌دیده است. طبقه‌بندی به‌صورت دو ردهٔ "مرتبط" و "غیر مرتبط" صورت می‌گیرد و اخباری که در ردهٔ "غیر مرتبط" قرار گیرند از فرآیندهای بعدی سیستم کنار گذاشته می‌شوند. نتایج ارزیابی این مدل(که روش و نتیجه آن به طور کامل در فصل پنجم تشریح شده است) نشان داده است کارایی این مدل بیش از ۹۵ ٪ برای اخبار غیرمرتبط و بیش از ۸۰ ٪ برای اخبار مرتبط است.

۳-۳-۴. پیش‌پردازش و نرمال‌سازی داده‌ها

در ابتدا به هر خبر یک برچسب تخصیص داده می‌شود. این برچسب بعدها برای یافتن سؤالات مربوطه کاربرد دارد. در مجموعه داده ایجاد شده در حدود ٤۲ نوع برچسب مختلف ایجاد شده است. پس از جمع‌آوری داده‌ها و قبل از ورود به مرحله پیش‌آموزش، یک سلسله‌مراتب عظیم از مراحل پردازشی از جمله تثبیت یونیکد، حذف و پاک‌سازی، نرمال‌سازی و قطعه‌بندی انجام می‌شود. مراحل انجام شده در شکل (۲) قابل مشاهده هستند.

۳-۳-۵. ایجاد پایگاه داده پایه

داده‌های جمع‌آوری شده پس از پیش‌پردازش و

59-Epoch

جدول (۲). شواهد امنیت در داده‌های بدون ساختار

مؤلفه	سؤال	مؤلفه	سؤال
امنیت زیرساخت و ارتباطات	آیا حادثه ناشی از اقدام سایبری بوده است؟	دفاع و امنیت شبکه	گروه رخنه‌گری به چه نامی است؟
	حادثه منجر به چه پیامدهایی شد؟		رخنه‌گر موفق به انجام چه‌کاری شدند؟
	میزان خسارات چگونه است؟		گروه هکری چه چیزی اعلام کرد؟
	اهمیت زیرساخت موردحمله چیست؟		فعالیت قبلی این گروه چه بوده است؟
	درباره زیرساخت چه توضیحاتی وجود دارد؟		منتسب به کجا شده است؟
	دشواری بازگردانی زیرساخت در چیست؟		پیچیدگی و گستردگی حادثه چگونه است؟
	چه بخشی حادثه‌دیده است؟		چه توضیحاتی در مورد گروه هکری وجود دارد؟
	حادثه چطور اتفاق افتاده است؟		از چه بدافزار یا روشی استفاده شد؟
امنیت اطلاعات	چه اطلاعاتی به دست رخنه‌گر افتاد؟		عملکرد بدافزار یا روش مورداستفاده چگونه است؟
	چه اطلاعاتی از بین رفته است؟		آیا حادثه از طریق عوامل نفوذی صحت دارد؟
	رخنه‌گرها با اطلاعات سرقت رفته چه‌کاری انجام دادند؟		چه تعداد افراد و مشتریان تحت تأثیر قرار گرفتند؟
	اطلاعات با چه قیمتی به فروش گذاشته شده است؟		اکنون وضعیت خدمات چگونه است؟
حاکمیت	هدف حمله چه بوده است؟		واکنش متولیان چگونه بوده است؟
	با چه شعاری اقدام به نفوذ شده است؟		واکنش بخش فنی نسبت به حمله چه بوده است؟
	آیا عامل حمله دستگیر شده است؟		قبلاً چه هشداری صادر شده بود؟
	واکنش مرکز افتا چگونه بوده است؟		چه اقدامات جبرانی صورت گرفته است؟
	واکنش سازمان پدافندغیر عامل چگونه بوده است؟		چه حادثه‌ای رخداده است؟
	واکنش مرکز ملی فضای مجازی چگونه بوده است؟		چه زمانی حادثه اتفاق افتاد؟
	واکنش مقامات چه بوده است؟		حادثه مقارن با چه زمانی بوده است؟
	واکنش‌ها در خارج کشور نسبت به حادثه چه بوده؟		در وقایع مرتبط شاهد چه چیزی بودیم؟
کاربران و رسانه	گزارش‌های فضای مجازی و رسانه چه بود؟		
	گزارش کارکنان و مشتریان چه بوده است؟		

نرمال‌سازی در پایگاه داده ذخیره می‌شوند تا در مراحل بعدی به منظور یافتن پاسخ سوالات مورد استفاده قرار گیرند. با توجه به حجم زیاد اطلاعات موجود برای یافتن پاسخ هر سوال قطعا نیاز است موضوعات مرتبط تشخیص داده شود. برای این منظور هر رکورد از پایگاه داده دارای یک برچسب است که می‌تواند به منظور پالایش اطلاعات مورد نیاز استفاده شود. در مرحله بعد داده‌ها به نسبت ۷۰، ۲۰ و ۱۰ تقسیم شده‌اند. از ۷۰ درصد داده‌ها برای آموزش مدل و ۱۰ درصد برای اعتبارسنجی و ۲۰ درصد باقیمانده برای آزمون مدل استفاده شده است.

۳-۳-۶. شواهد امنیت شبکه در منابع آزاد

برخی از شواهد امنیت شبکه که از منابع آزاد قابل احصاء است در این بخش استخراج شده است. از این شواهد برای آموزش به مدل هوش مصنوعی استفاده می‌شود تا مدل بتواند جواب را برای سایر موارد مشابه ارایه نماید. مؤلفه‌ها شامل پنج بخش به شرح زیر هستند:

- امنیت زیرساخت و ارتباطات
- امنیت اطلاعات
- دفاع و امنیت شبکه
- کاربران و رسانه‌ها
- حاکمیت(قانون‌گذار، سازمان‌ها و مراکز، مسئولین)

در جدول (۲) شواهدی در ارتباط با امنیت شبکه که از داده‌های بدون ساختار قابل احصاء است، فهرست شده است.

۳-۳-۷. ایجاد مجموعه داده(مجموعه داده هوشفا)

با توجه به این که هوش مصنوعی داده‌محور است

معمولاً محققان بخشی از وقت خود را صرف تهیه مجموعه داده مناسب برای مدل‌های مبدل می‌کنند. یک مدل آموزش‌دیده مانند شخصی که زبان آموخته رفتار خواهد کرد. داده باید با طی مراحل، اطلاعات جدیدی به مجموعه داده آموزش اضافه کنند و در نهایت مجموعه داده‌های آموزش می‌تواند به یک پایگاه دانش تبدیل شوند. رویکرد درست آموزش مبدل، پیاده‌سازی آن و ارایه نتایج قابل قبول در گرو دسترسی به مجموعه داده مناسب است. در این بخش یک مجموعه داده با استفاده از مؤلفه‌ها و شاخص‌های احصاء شده ایجاد شده است که از این مجموعه داده برای آموزش و ارزیابی مدل هوش مصنوعی استفاده می‌شود. با ایجاد این مجموعه داده مدل قادر است پاسخ سوالاتی که در حوزه امنیت شبکه‌های کامپیوتری احصاء شده است را از متن بدون ساختار، احصاء نموده و به کاربر نمایش دهد.

با توجه به این که انجمن‌های جمع‌سپاری[60] برای تولید مجموعه داده فارسی وجود ندارد و از طرفی ایجاد مجموعه داده به‌صورت دستی بسیار زمان‌بر است در این بخش روشی پیشنهاد می‌شود که بتوان سرعت ایجاد مجموعه داده را تا چند برابر افزایش داد. روش پیشنهادی به این صورت است که ابتدا یک مجموعه ۱۵۰ رکوردی از پرسش و پاسخ تولید می‌شود و سپس یک مدل با این مجموعه داده آموزش داده می‌شود. در ادامه مدل با استفاده از مجموعه داده پایه پرسش و پاسخ جدید تولید می‌کند. ملاک این که یک رکورد برای ذخیره‌سازی انتخاب شود، بالا بودن نمره است. همچنین رکوردهایی که نمره بسیار پایینی دارند به‌عنوان سؤالات بدون پاسخ در مجموعه داده ذخیره می‌شوند تا در مراحل بعدی آموزش، مدل تشخیص دهد چه سؤالاتی را نباید پاسخ بدهد. در ادامه خروجی تولید شده به‌صورت دستی تصحیح شده و مواردی که دارای خطا هستند، حذف می‌شوند. این فرآیند به‌طور نمایی ادامه می‌یابد و در گام دوم ۳۰۰ رکورد و در گام سوم ۶۰۰ رکورد تولید و تصحیح می‌شوند تا این که تعداد رکوردها به حد مورد نیاز برسد. در نهایت مجموعه داده مورد استفاده برای آموزش مدل پرسش و پاسخ به ترتیب زیر ایجاد شده است.

جدول (۳). رکوردهای موضوعی مجموعه داده هوشفا

امنیت شبکه	فناوری اطلاعات	عمومی	مجموع رکوردها
۳/۳۵۶	۱/۹۹۸	۶/۸۰۳	۱۲/۱۵۷

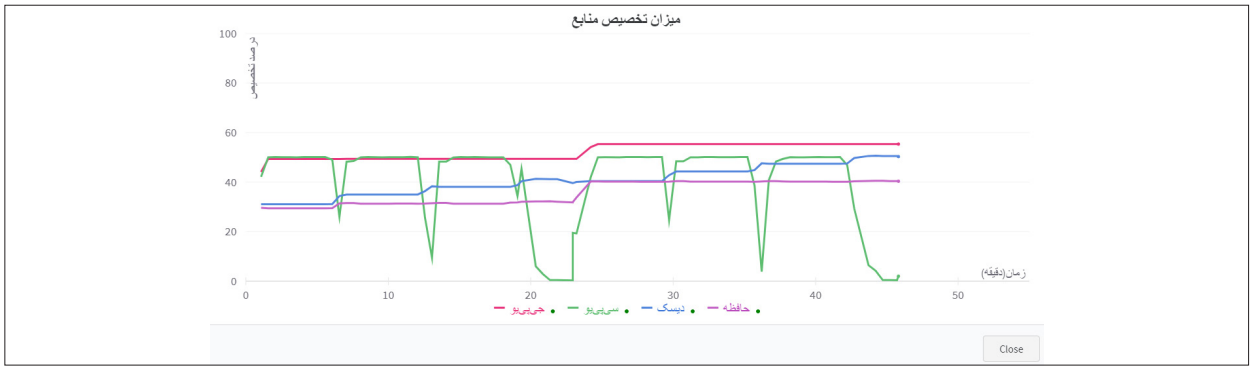
رکوردهای پایه شبکه مربوط به پرسش و پاسخ در زمینه امنیت شبکه‌های کامپیوتری است که با استفاده از روش پیشنهادی تولید شده است. رکوردهای فناوری اطلاعات مربوط به سایر مجموعه داده‌ها است که در حوزه فناوری و اطلاعات به‌صورت پرسش و پاسخ از قبل توسط سایر محققان تولید شده‌اند و در ترکیب مجموعه داده قرار گرفته است. رکوردهای عمومی مربوط به سایر مجموعه داده‌ها است که در حوزه عمومی به‌صورت پرسش و پاسخ از قبل توسط سایر محققان تولید شده‌اند و در ترکیب مجموعه داده قرار گرفته‌اند.

۳-۳-۸. آموزش مدل با مجموعه داده هوشفا

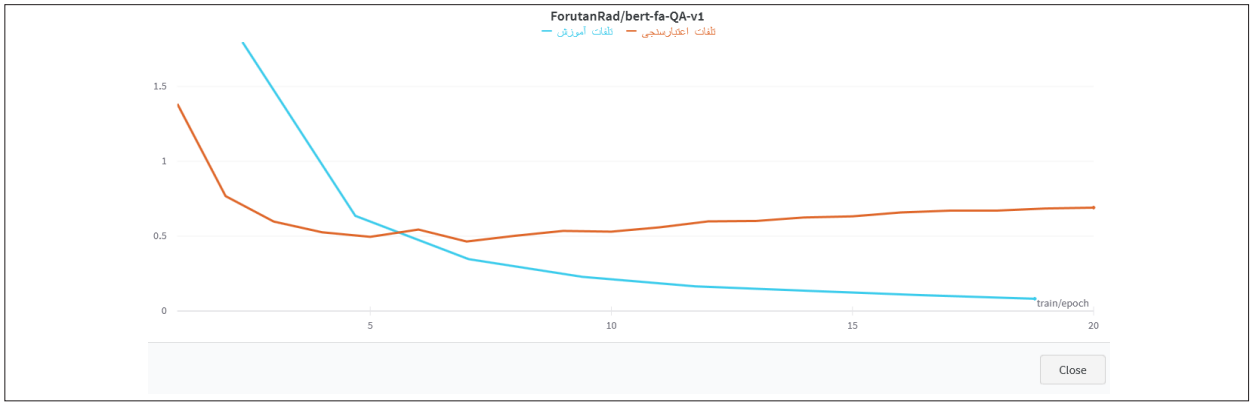
پس از ایجاد مجموعه داده مناسب یادگیری با زبان برنامه‌نویسی پایتون[61] اقدام به آموزش مدل مبدل شده است. آموزش مدل بر روی ۷۰ درصد از داده‌های موجود انجام شده است و از ۱۰ درصد داده‌ها برای اعتبارسنجی استفاده شده است. همچنین ۲۰ درصد داده‌ها نیز برای آزمون در نظر گرفته شده است. در این بخش مدل‌های هوش مصنوعی با استفاده از مجموعه داده‌ای که توسط مؤلفه‌ها و شاخص‌های امنیت شبکه‌های کامپیوتری در بخش‌های قبل ایجاد شده است آموزش داده می‌شوند. پس از آموزش مدل‌های مختلف(سه مدل در این تحقیق) ارزیابی عملکرد مدل‌ها با یکدیگر صورت می‌گیرد و بهترین مدل مشخص می‌گردد. در ابتدای آموزش و در دوره‌های

60-Crowdsourcing

61-Python



شکل (۳). منابع (مدل ForutanRad/bert-fa-QA-v1)



شکل (۴). دوره‌ها مدل(ForutanRad/bert-fa-QA-v1)

پایین، نرخ تلفات و اعتبارسنجی فاصله زیادی داشتند که با انجام اصلاحاتی از قبیل تنظیم نرخ یادگیری، استفاده از سایر توابع بهینه‌ساز، رفع نوفهٔ[۶۲] داده و سایر روش‌ها فاصله نرخ تلفات و اعتبارسنجی به حد قابل مناسب رسیده است.

- آموزش مدل ForutanRad/bert-fa-QA-v1

این مدل یک نمونه فارسی از مدل برت است. در آموزش این مدل از سخت‌افزار شتاب‌دهنده گرافیکی با مدل NVIDIA A100 (40 GB) دارای قدرت پردازشی 19.5 TeraFlops استفاده شده است. با قدرت پردازشی مذکور آموزش مدل در زمان ٤٥ دقیقه انجام شده است. میزان مشغولیت حافظه، دیسک، سی‌پی‌یو و جی‌پی‌یو در شکل (۳) ترسیم شده است.

همان‌طور که در شکل(۳) قابل مشاهده می‌باشد در این آموزش از جی‌پی‌یو استفاده شده است و میزان مشغولیت جی‌پی‌یو بیشتر از سایر منابع می‌باشد.

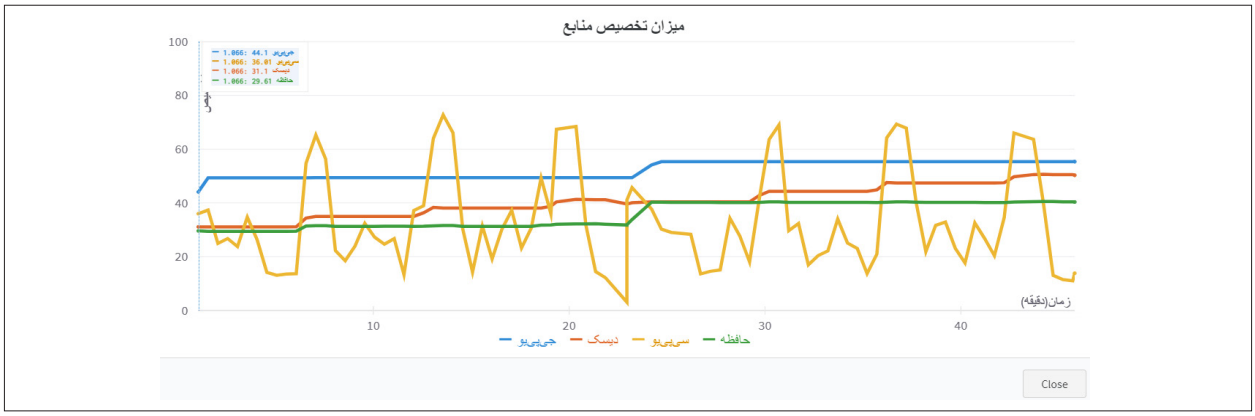
همان‌طور که در شکل (٤) نمایش داده شده است، تعداد دوره‌های آموزش بر روی ۲۰ دوره تنظیم شده است.

با توجه به تقاطع نمودار تلفات اعتبارسنجی و تلفات آموزش در دوره ششم نتیجه گیری می‌شود که مدل در آموزش شش دوره دارای بهترین نتیجه می‌باشد. لذا مدل نهایی بر روی آموزش بر مبنای شش دوره انجام می‌شود.

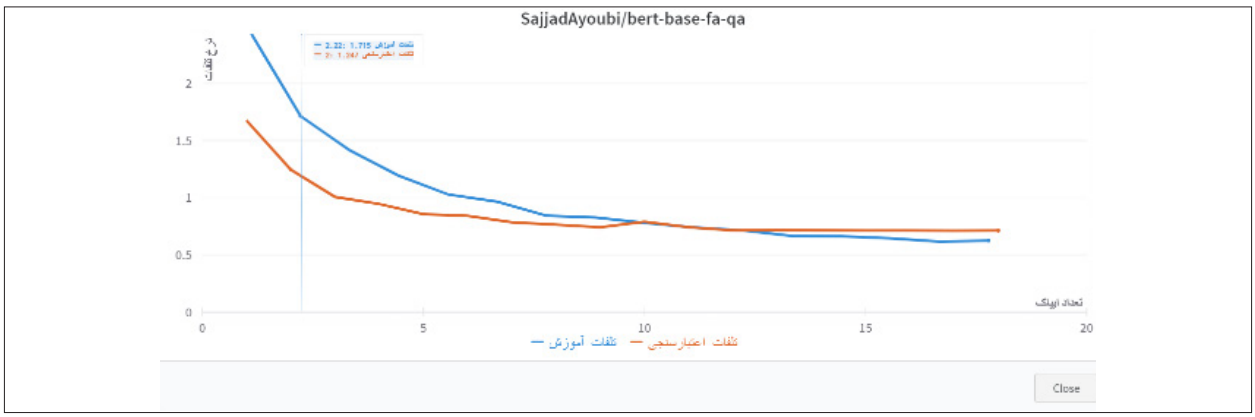
- آموزش مدل SajjadAyoubi_bert-base-fa-qa

یک مدل فارسی با 162M پارامتر است. در آموزش این مدل از سخت‌افزار شتاب‌دهنده گرافیکی با مدل NVIDIA A100 (40 GB) دارای قدرت پردازشی ۱۹,۵ TeraFlops استفاده شده است. با قدرت پردازشی مذکور آموزش مدل در زمان ٤٥ دقیقه انجام شده است. میزان مشغولیت حافظه، دیسک، سی‌پی‌یو و جی‌پی‌یو در شکل (٥) ترسیم شده است.

62-Noise



شکل (۵). منابع (مدل SajjadAyoubi_bert-base-fa-qa)



شکل (۶). دوره‌ها مدل(SajjadAyoubi_bert-base-fa-qa)

همان‌طور که در شکل(۵) قابل مشاهده می‌باشد در این آموزش از جی‌پی‌یو استفاده شده است و پیوستگی کارکرد و میزان مشغولیت جی‌پی‌یو بیشتر از سایر منابع می‌باشد.

همان‌طور که در شکل (٦) نمایش داده شده است، تعداد دوره‌های آموزش بر روی ۱۸ دوره تنظیم شده است.

با توجه به تقاطع نمودار تلفات اعتبارسنجی و تلفات آموزش در دوره دهم نتیجه‌گیری می‌شود که مدل در آموزش ده دوره دارای بهترین نتیجه می‌باشد. لذا مدل نهایی بر روی آموزش بر مبنای ده دوره انجام می‌شود.

- آموزش مدل SajjadAyoubi_xlm-roberta-large-fa-qa

یک مدل فارسی بزرگ با 558M پارامتر است. در آموزش این مدل از سخت‌افزار شتاب‌دهنده گرافیکی با مدل NVIDIA A100 (40 GB) دارای قدرت پردازشی 19.5 TeraFlops استفاده شده است. با قدرت پردازشی مذکور آموزش مدل در زمان ۹۰ دقیقه انجام شده است.

میزان مشغولیت جی‌پی‌یو نزدیک به ۱۰۰ درصد بوده است و در شکل (۷) ترسیم شده است.

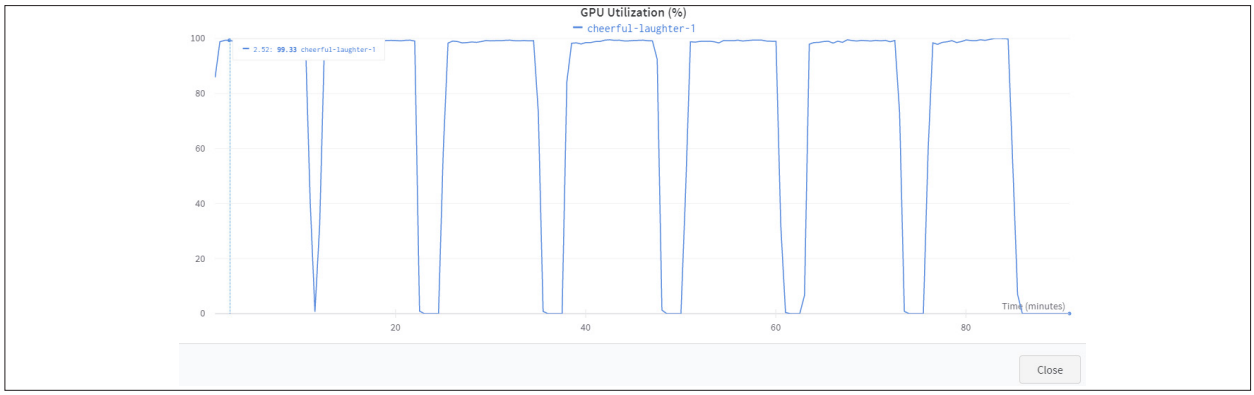
همان‌طور که در شکل (۸) نمایش داده شده است، تعداد دوره‌های آموزش بر روی ۲۰ دوره تنظیم شده است.

با توجه به تقاطع نمودار تلفات اعتبارسنجی و تلفات آموزش در دوره ۱۲ نتیجه گیری می‌شود که مدل در آموزش ۱۲ دوره دارای بهترین نتیجه می‌باشد. لذا مدل نهایی بر روی آموزش بر مبنای ۱۲ دوره انجام می‌شود.

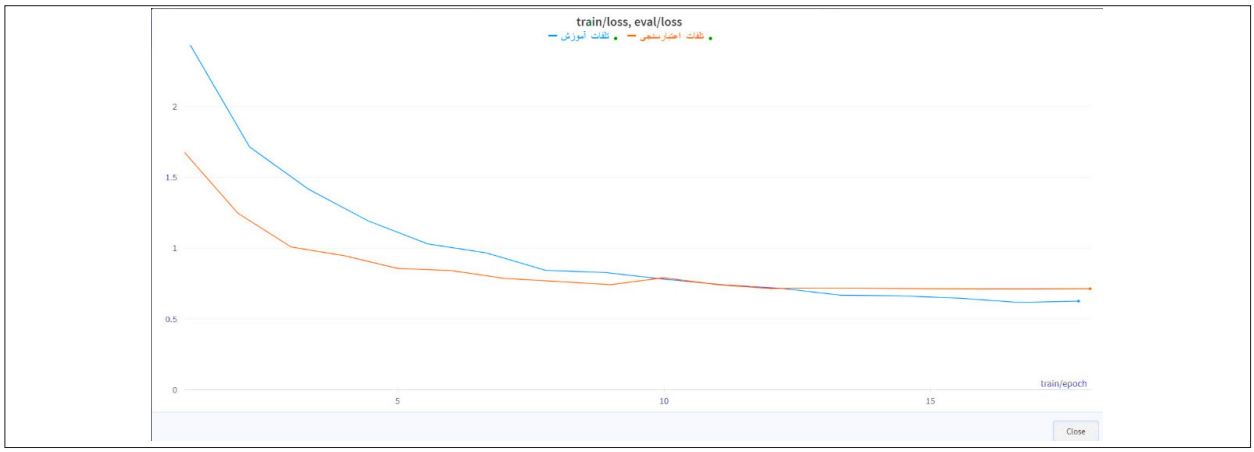
۳-۳-۹- کاربری سیستم پیشنهادی

بسته به این که کاربر، یک محقق، تحلیلگر و یا دانشمند داده باشد، در مواقعی نیاز دارد برای یافتن اطلاعات موردنظر اقیانوسی از اسناد را بررسی کند. پاسخ به سؤال شامل سؤالاتی است که پاسخ آن‌ها می‌تواند به‌عنوان یک متن در سند وجود داشته باشد.

مدل آموزش داده شده، به منظور پاسخ به سوال بر



شکل (۷). منابع (مدل SajjadAyoubi_xlm-roberta-large-fa-qa)



شکل (۸). دوره‌ها (مدلSajjadAyoubi_xlm-roberta-large-fa-qa)

روی هاگینگ‌فیس بارگذاری شده است و امکان استفاده از آن توسط خط لوله وجود دارد. در ادامه مثالی از کاربری مدل ارایه شده است:

توسط پرسمان زیر می‌توان پاســـخ را از متون بدون ساختار به‌دست آورد:

```
nlp_qa(context=sequence، question=
'اکنون وضعیت خدمات چگونه است؟'
)
```

خروجی به‌صورت زیر نمایش داده می‌شود:

```
{'answer':' ۳۱۰۰ جایگاه سوخت‌رسانی به سامانه
هوشمند وصل شـــده‌اند اما ۹۰۰ جایگاه همچنان امکان
فروش بنزین با نرخ آزاد دارند'،
'end':66،'score':0.918201259751591،'start':55}
```

۴. ارزیابی سیستم پیشنهادی

ارزیابی‌های استاندارد می‌توانند روشی اطمینان‌بخش برای ارزیابی و مقایســـه کارایی مدل‌های مبدل ارایه دهند. همان‌طور که در بخش ســـوم اشاره شده است ۲۰ درصد داده‌ها به عنوان داده‌های آزمون در نظر گرفته شده است.

۴-۱. معیارهای ارزیابی

مقایسه یک مدل مبدل با سایر مدل‌ها بدون معیارهای اندازه‌گیری استاندارد، غیرممکن اســـت. از این رو در این پژوهـــش از میان انبوه معیارهایی که برای ارزیابی مدل‌ها ارایه شـــده، از یک یا چند معیار دقت[۶۳]، صحت[۶۴]، پوشش[۶۵]، اف ۱[۶۶]، تطبیق دامنه دقیق[۶۷] استفاده شده است.

63-Accuracy
64-Precision
65-Recall
66-F1
67-Exact_match

جدول (۴). استعدادیابی مدل‌ها

داده	نام مـدل	تطبیق دقیق	اف۱
هوشفا	HooshvareLab_bert-fa-base-uncased	۱/۸۶	۲۰/۸۱
	SajjadAyoubi_bert-base-fa-qa	۴۷/۸۸	۲۱/۶۷
	SajjadAyoubi_xlm-roberta-large-fa-qa	۲۸/۱۵	۵۴/۵۷
	ForutanRad/bert-fa-QA-v1	۱۸/۷۱	۴۵/۹۸
	Koala_xlm-roberta-large-fa	۰/۰	۲۱/۲۶
	Koala_bert-large-uncased-fa	۱/۸۶	۲۶/۲۹
	xlm-roberta-base	۱۷/۳۹	۷/۲۳
	xlm-roberta-large	۰/۰۵	۲۲/۰۷
	bert-base-multilingual-cased	۰/۶۵	۲۴/۱۵

۴-۱-۱. معیار دقت

ایـــن معیار نمایانگر میـــزان پیش‌بینی صحیح خروجی توسط مدل است و می‌توان توسط آن حد آموزش صحیح مدل را مشـــاهده نمود. این معیار میـــزان نتایج صحیح را نســـبت به کل موارد مورد ســـنجش قرار می‌دهد. در این معیار اگر نتیجه برای زیرمجموعه مثبت باشد مقدار یک و اگر نادرست باشد مقدار صفر بازگردانده می‌شود.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

۴-۱-۲. معیار صحت

معیار صحت معیاری است که مشخص می‌کند الگوریتم چند درصد مثبت‌ها را درست پیش‌بینی کرده است. فرمول محاسبه این معیار به صورت زیر است:

$$Precision = \frac{T_p}{T_p + F_p}$$

۴-۱-۳. معیار پوشش

معیار پوشش برای محاســـبهٔ میزان پوشش روی کل داده‌ها اســـتفاده می‌شـــود. فرمول محاســـبه این معیار به صورت زیر است:

$$Recall = \frac{T_p}{T_p + F_N}$$

۴-۱-۴. معیار اف۱

معیار اف۱ یک رویکرد انعطاف‌پذیرتر را معرفی می‌کند کـــه می‌تواند در مواجهه با مجموعه داده‌های حاوی توزیع ردهٔ ناهموار مفید واقع شـــود. معیار اف۱ از میانگین وزنی دقت و فراخوانی به ترتیب زیر استفاده می‌کند:

$$F1 - score = 2 * (precision * recall) /(precision + recall)$$

$$P = \frac{T_p}{T_p + F_p}$$

$$R = \frac{T_p}{T_p + F_n}$$

از ایـــن‌رو می‌تـــوان نمـــره اف۱ را به‌عنـــوان میانگین هارمونیـــک[۶۸] (میانگین حســـابی) دقـــت(P) و فراخوان(R) مشاهده کرد:

$$F1Score = \frac{PxR}{P + R}$$

۴-۱-۵. معیار دامنه تطبیق دقیق

روش مورد اســـتفاده ایـــن معیار در ارزیابی بســـیار ســـاده است. در مقایسه بین رشته‌ها اگر دقیقاً همان رشته پیش‌بینی شـــده باشد نمره یک و در غیر این صورت نمره صفر لحاظ می‌گردد.

۴-۲. استعدادیابی مدل‌ها

با توجه بـــه تعدد مدل‌های موجـــود، دور از انتظار نیســـت که برخی از آنها بهتر از ســـایرین در آموزش باشـــند. از این رو نیاز اســـت توســـط آزمون بهترین مدل‌ها انتخاب شوند. برای انتخاب مناسب‌ترین مدل‌ها با استفاده از مجموعه داده آزمون هوشفا عملکرد مدل مورد آزمون قرار گرفته تا استعداد مدل مورد سنجش قرار گیرد.

از میان مدل‌های بررســـی شده مشخص است که سه مدل از استعداد خوبی به منظور آموزش در مجموعه داده هوشفا دارند.

۴-۳. نتایج ارزیابی مدل‌ها پس از آموزش با هوشفا

ارزیابی مجموعه داده با استفاده از کدها و کتابخانه‌های زبان برنامه‌نویسی پایتون انجام شده است. داده‌هایی که به

68-Harmonic

صورت مستقل از سایر مجموعه‌داده‌ها در مجموعه داده هوشفا تولید شده‌اند تعداد ۳,۳۵٦ رکورد است. در جدول زیر نتایج ارزیابی مدل‌های مختلف بر روی این بخش قابل مشاهده است. "نمره قبل از آموزش"، مربوط به زمانی است که مدل با استفاده از داده‌های هوشفا آموزش داده نشده و "نمره بعد از آموزش" مربوط به زمان است که مدل با استفاده از داده‌های هوشفا آموزش داده شده است. "میزان تغییرات" مربوط به نتایج بهبود یافته مدل بعد از آموزش می‌باشد.

جدول (۵). نتایج ارزیابی با آموزش مجموعه داده هوشفا

نام مدل	قبل از آموزش		بعد از آموزش		میزان بهبود	
	تطبیق دقیق	اف۱	تطبیق دقیق	اف۱	تطبیق دقیق	اف۱
SajjadAyoubi/ xlm-roberta-large-fa-qa	۱۸/۱۲	۵۰/۷۵	۶۱/۳۵	۶۹/۲۲	+۴۳/۲۳	+۱۸/۴۷
SajjadAyoubi/ bert-base-fa-qa	۵/۹۷	۴۳/۵۴	۲۶/۶۹	۵۱/۰۶	+۲۰/۷۲	+۷/۵۲
ForutanRad/ bert-fa-QA-v1	۴/۵۸	۴۲/۶۹	۲۶/۴۹	۵۱/۵۹	+۲۱/۹۱	+۸/۹

۴-۴. نتایج ارزیابی مدل طبقه‌بندی اخبار

چنان که بیان شد از مدل طبقه‌بندی اخبار برای تشخیص اخبار مرتبط و غیرمرتبط استفاده شده است. نتایج ارزیابی این مدل روی عناوین اخبار کارایی این مدل بیش از ۹۵٪ برای اخبار غیرمرتبط و بیش از ۸۰٪ برای اخبار مرتبط را نشان داده است. برچسب صفر مربوط به اخبار غیر مرتبط و برچسب یک مربوط به اخبار مرتبط است.

جدول (۶). نتایج ارزیابی مدل طبقه‌بندی اخبار

نام مدل	دقت	پارامتر	اف۱	پوشش	صحت
HooshvareLab_ bert-fa-zwnj-base	۰/۹۶	۰	۰/۹۸	۰/۹۸	۰/۹۷
		۱	۰/۸۳	۰/۸۰	۰/۸۷
HooshvareLab_ gpt2-fa	۰/۹۵	۰	۰/۹۷	۰/۹۹	۰/۹۶
		۱	۰/۸۰	۰/۷۱	۰/۹۱
xlm-roberta-base	۰/۹۲	۰	۰/۹۶	۱/۰	۰/۹۲
		۱	۰/۵۷	۰/۴۰	۰/۹۸

۵. بحث و نتیجه‌گیری

با توجه به چالش رشد سریع داده‌های بدون ساختار، در این پژوهش سیستمی بر پایه هوش مصنوعی به منظور تحلیل داده‌های بدون ساختار با تمرکز بر حوزه امنیت شبکه‌های کامپیوتری ارایه شده است. سیستم مذکور قادر به یادگیری و پاسخ به سوالات در حوزه‌های مختلف می‌باشد و در این پژوهش بر روی امنیت شبکه‌های کامپیوتری تمرکز شده است. سیستم مذکور از مقادیر زیادی داده‌های متنی برای آموزش استفاده نموده و با ایجاد درک از نقش کلمه در جمله (مشابه مغز انسان)، قادر به یادگیری، و پاسخ به سوالات است.

برای طراحی و پیاده‌سازی سیستم مذکور در ابتدا جمع‌آوری ماشینی داده‌های بدون ساختار از منابع وب انجام شده است و یک مدل طبقه‌بندی برای شناسایی داده‌های مرتبط با حوزه امنیت شبکه‌های کامپیوتری آموزش داده شده است. پس از انتخاب برخی موارد مرتبط به امنیت شبکه‌های کامپیوتری، تشکیل مجموعه داده سفارشی به نام هوشفا توسط پژوهشگر انجام شده است. این مجموعه داده سفارشی، شامل ۳/۳۵٦ رکورد جدید و ۸۸۰۱ رکورد از مجموعه داده‌های توسعه داده شده توسط سایر پژوهشگران است. در نهایت آموزش مدل‌های مختلف با استفاده از مجموعه داده سفارشی انجام شده است. نتایج آموزش در سیستم پیشنهادی توسط مجموعه داده سفارشی، بسته به مدل‌های انتخابی، میزان درک ارایه پاسخ را در معیار تطبیق دقیق به میزان ۲۰ تا ٤۳ درصد و در معیار اف۱ به میزان ۸ تا ۱۸ درصد بهبود بخشیده است.

۶. کارهای آتی

کارهایی را که در آینده می‌توان تحقیقات را ادامه و بهبود بخشید می‌توان به شرح زیر برشماری نمود:

۱. افزایش کمّی و کیفی شاخص‌ها

شاخص‌های احصاء شده در زمینه امنیت شبکه‌های کامپیوتری، مبنای ایجاد مجموعه داده یادگیری هستند و

افزایش کمّی و کیفی شاخص‌های منجر به ارایه نتایج بهتر می‌شود.

۲. ارتقای کمّی و کیفی مجموعه داده هوشفا

مجموعه داده یادگیری هوشفا به طور مستقیم در ایجاد درک مدل از مسایل مختلف تاثیر دارد و با افزایش کمّی و کیفی رکوردها درک مدل از مسایل مختلف افزایش می‌یابد.

۳. آموزش مدل‌های مختلف جدید

مدل‌هـــای مبدل مختلـــف دارای پارامترهـــای متفاوتی هستند و نتایج متفاوتی را ارایه می‌کنند. آموزش مدل‌های مختلف با مجموعه داده هوشفا و بررسی و مقایسه کارایی می‌تواند به بهبود محصول نهایی منجر شود.

٤. بهبود پارامترهای یادگیری

به‌کارگیری توابع بهینه‌ساز مختلف و دوره‌های متفاوت و ارزیابی نتایج حاصله می‌تواند منجر به بهبود درک مدل شود.

۵. آمـــوزش مدلی به منظور طبقه‌بنـــدی اخبار و منابع مرتبط با امنیت شبکه‌های کامپیوتری

اســـتفاده مدل هـــوش مصنوعی از مطالـــب مرتبط با موضوع مـــورد آموزش می‌تواند منجر به ارایه نتایج بهتر شود. این موضوع به‌خصوص در آموزش مدل‌های بر پایه متن مانند برت محسوس است.

۷. منابع

[1] M. A. Hassan, “Relational and NoSQL Databases: The Appropriate Database Model Choice,” in 2021 22nd International Arab Conference on Information Technology (ACIT), 2021: IEEE, pp. 1-6.

[2] A. Agarwal, C. Baechle, R. Behara, and X. Zhu, “A natural language processing framework for assessing hospital readmissions for patients with COPD,” IEEE journal of biomedical and health informatics, vol. 22, no. 2, pp. 588-596, 2017.

[3] N. N. Vo, S. Liu, X. Li, and G. Xu, “Leveraging unstructured call log data for customer churn prediction,” Knowledge-Based Systems, vol. 212, p. 106586, 2021.

[4] S. Crockett and C. Lee, “Does it Matter What They Said? A Text Mining Analysis of the State of the Union Addresses of USA Presidents,” in 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2012: IEEE، pp. 77-82.

[5] L. Tunstall,L. von Werra، and T. Wolf، Natural language processing with transformers. “ O’Reilly Media، Inc.”, 2022.

[6] ParsiNLU, “ParsiNLU.” [Online]. Available: https://github.com/persiannlp/parsinlu.

[7] C. E. Shannon, “A mathematical theory of communication,” The Bell system technical journal, vol. 27, no. 3, pp. 379-423, 1948.

[8] A. M. Turing and J. Haugeland, “Computing machinery and intelligence,” The Turing Test: Verbal Behavior as the Hallmark of Intelligence, pp. 29-56, 1950.

[9] O. Kolomiyets and M.-F. Moens, “A survey on question answering technology from an information retrieval perspective,” Information Sciences، vol. 181, no. 24, pp. 5412-5434, 2011.

[10] A. Vaswani et al., “Attention is all you need,” Advances in neural information processing systems, vol. 30، 2017.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805، 2018.

[12] A. M. Dai and Q. V. Le, “Semi-supervised sequence learning,” Advances in neural information processing systems، vol. 28، 2015.

[13] K. Darvishi, N. Shahbodagh, Z. Abbasiantaeb, and S. Momtazi, “PQuAD: A Persian Question Answering Dataset,” arXiv preprint arXiv:2202.06219، 2022.

[14] J. Duan, H. Zhao, Q. Zhou, M. Qiu, and M. Liu, “A study of pre-trained language models in natural language processing,” in 2020 IEEE International Conference on Smart Cloud (SmartCloud), 2020: IEEE, pp. 116-121.

[15] S. Singh and A. Mahmood, “The NLP cookbook: modern recipes for transformer based deep learning architectures, “ IEEE Access, vol. 9, pp. 68675-68702, 2021.

[16] Y. E. Seyyar, A. G. Yavuz, and H. M. Ünver, “An attack detection framework based on BERT and deep learning, “ IEEE Access, vol. 10, pp. 68633-68644, 2022.

[17] M. Farahani, M. Gharachorloo, M. Farahani، and M. Manthouri, “Parsbert: Transformer-based model for persian language understanding,” Neural Processing Letters، vol. 53, no. 6, pp. 3831-3847, 2021.

[18] E. Taher, S. A. Hoseini، and M. Shamsfard, “Beheshti-NER: Persian named entity recognition Using BERT,” arXiv preprint arXiv:2003.08875, 2020.

[19] A. Khashei, “Bolbol-Zaban.” [Online]. Available: http://www.bolbolzaban.com/.

[20] M. Tanvirul Alam, D. Bhusal, Y. Park, and N. Rastogi, “CyNER: A Python Library for Cybersecurity Named Entity Recognition, “ arXiv e-prints، p. arXiv: 2204.05754, 2022.

[21] Y. Guo, J. Liu، W. Tang, and C. Huang, “Exsense: Extract sensitive information from unstructured data, “ Computers & Security، vol. 102، p. 102156, 2021.

[22] W. Zhuang, “Architecture of Knowledge Extraction System based on NLP,” in 2021 International Conference on Aviation Safety and Information Technology,2021, pp. 294-297.