دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی
مهندسی کامپیوتر

تشخیص اهمیت اخبار فارسی با استفاده از مدل‌های زبانی بزرگ

نگارش

شایان صالحی

استاد راهنما

دکتر مهدی جعفری

بهمن ۱۴۰۳

سپاس

از استاد بزرگوارم، دکتر جعفری به خاطر زحمات و راهنمایی‌هایی که در طول این پروژه داشته‌اند متشکرم و همچنین از دانشجوی دکترا ایشان، آقای معین سلیمی به خاطر زمان و راهنمایی‌هایی که برای پیش‌بردن پروژه انجام داده‌اند قدردانم.

چکیده

این پروژه به بررسی قدرت تشخیص اهمیت یک خبر فارسی توسط مدل‌های زبانی بزرگ پرداخته و قدرت یادگیری از محتوا، قدرت استدلال و قدرت تفکر آن را ارزیابی کرده است. در ابتدا، از دادگان علائم‌گذاری‌شده توسط افراد در حوزه‌های مختلف از جمله ورزشی، سیاسی، اجتماعی، پزشکی و فرهنگی استفاده و محیطی برای ارزیابی مدل‌های زبانی بزرگ توسعه داده شده است. در این محیط مدل‌های مختلف موجود بررسی و ارزیابی شده و در نهایت با تمام حالات مختلف و شرایط مخلتف، قدرت تحلیل آنها در زبان فارسی و انگلیسی بررسی شده است. این پروژه نشان‌داده که دستورهای[1] شامل زنجیره تفکر[2] و درخت تفکر[3] باعث بهبود کارایی مدل‌ها و همچنین روش نتظیم نمادها[4] باعث حساسیت بسیار زیاد به پرسش داده شده و محتوای آن می‌شود.

کلیدواژه‌ها: مدل‌های زبانی بزرگ، پردازش زبان‌های طبیعی، یادگیری ماشین، تشخیص اهمیت اخبار

[1] Prompt
[2] Chain-of-Thoughts
[3] Tree-of-Thoughts
[4] Symbol Tuning

فهرست مطالب

۱ مقدمه **۱**

۱-۱ تعریف مسئله چگونگی بررسی مهم بودن یک خبر ۱

۱-۲ اهمیت موضوع تشخیص اهمیت اخبار ۲

۱-۳ ادبیات به کار رفته در این پژوهش ۲

۱-۴ اهداف پژوهش ۲

۲ مفاهیم اولیه **۴**

۲-۱ مدل‌های زبانی بزرگ ۴

۲-۲ یادگیری درونی ۴

۲-۳ تنظیم بر اساس دستورالعمل ۵

۲-۴ درخواست‌های سامانه و کاربر ۵

۲-۵ مهندسی درخواست ۶

۳ کارهای پیشین **۷**

۳-۱ تشخیص اهمیت اخبار ۷

۳-۱-۱ رویکردهای کلاسیک ۷

۳-۱-۲ رویکردهای تشخیص اهمیت با استفاده از یادگیری ماشین ۸

۳-۱-۳ رویکرد یادگیری عمیق ۸

۳-۱-۴ استفاده از مدل‌های زبانی بزرگ ۸

۳-۲ تنظیم بر اساس دستورالعمل و تنظیم نمادین ۹
۳-۳ یادگیری چندنمونه‌ای در تشخیص اهمیت اخبار ۹
۳-۴ تشخیص اهمیت اخبار فارسی ۹
۳-۵ سامانه‌های درخواست وابسته به پرسش ۱۰

۴ روش پیشنهادی ۱۱
۴-۱ انواع دستورالعمل‌های توسعه‌داده شده ۱۱
۴-۱-۱ دستورالعمل‌های وانیلا ۱۱
۴-۱-۲ دستورالعمل‌های چندنمونه‌ای ۱۲
۴-۱-۳ دستورالعمل‌های زنجیره تفکر ۱۳
۴-۱-۴ دستعورالعمل‌ها درخت تفکر ۱۴
۴-۲ رویکرد دستورالعمل سیستمی و کاربر ۱۴
۴-۳ تحلیل قدرت استدلال در حالت چند زبانی ۱۵
۴-۴ رویکرد تنظیم نمادها ۱۵
۴-۵ تنظیم براساس دستورالعمل و حالت‌های مختلف آن ۱۶
۴-۶ انواع متن بررسی شده از اخبار ۱۷
۴-۷ نحوه خروجی گرفتن از مدل‌های زبانی بزرگ ۱۷
۴-۷-۱ مدل‌های زبانی بزرگ استفاده شده ۱۷
۴-۷-۲ نحوه بهینه خروجی گرفتن از مدل‌های زبانی بزرگ ۱۸
۴-۸ چرخه کلی خروجی گرفتن‌ها ۱۸
۴-۹ ساختار نتایج ۱۹
۴-۱۰ نحوه تحلیل نتایج ۲۰

۵ نتایج جدید ۲۲
۵-۱ نتایج تنظیم نمادین ۲۲

۵-۱-۱ تنظیم نمادین با افزودن تعریف اخبار غیرمهم ۲۴
۵-۱-۲ نتایج دستور نمادین در مدل جما۲ . . . . . ۲۵
۵-۲ نتایج رویکرد جداسازی دستور سیستمی و کاربر در زبان‌های فارسی و انگلیسی ۲۶
۵-۳ نتایج دسته‌بندی‌ها مختلف خبری ۲۹
۵-۴ نتایج دستورالعمل‌های درخت تفکر و زنجیره‌های تفکر ۳۰
۵-۵ نتایج دستورالعمل‌های درخت و زنجیره‌های تفکر در حالت یادگیری چندنمونه‌ای . . . ۳۱
۵-۶ سیستم انتخاب دستورالعمل وابسته به ورودی ۳۱

۶ نتیجه‌گیری ۳۳

مراجع ۳۵

واژه‌نامه ۳۸

آ مطالب تکمیلی ۳۹

آ-۱ دستورالعمل‌های به کار گرفته شده ۳۹
آ-۱-۱ دستورهای وانیلا یا خام ۳۹
آ-۱-۲ دستورهای نمادین ۴۱
آ-۱-۳ دستورها با رویکرد یادگیری چند نمونه‌ای ۴۲
آ-۱-۴ دستورهای مخصوص دسته‌های متخلف خبری ۴۲
آ-۱-۵ دستعورالعمل‌های زنجیره‌های تفکر ۴۵
آ-۱-۶ دستورالعمل‌های درخت تفکر ۴۵
آ-۲ نتایج اضافی‌تر ۴۶
آ-۲-۱ نتایج بخش دستور نمادین ۴۶
آ-۲-۲ نتایج دستور نمادین مدل جما۲ . . . . . ۴۸
آ-۲-۳ نتایج دستور سیستمی انگلیسی ۴۸

آـ۲-۴ نتایج دستور سیستمی انگلیسی در حالت کل متن خبری ۴۸
آـ۲-۵ نتایج اضافی‌تر دستورهای درخت تفکر و زنجیره تفکر ۵۰

فهرست تصاویر

۴-۱ حالت‌های مختلف یک دستورالعمل و برچسب‌های آن ۱۶

۴-۲ ساختار نتایج ذخیره شده و مشخص‌کردن برچسب‌های $l - i$ پیش‌بینی شده از مدل . . ۲۰

۴-۳ نمونه‌ای از نوع نتیجه و دقت‌سنجی اعلامی ۲۰

۴-۴ نمونه‌ای از بررسی ماتریس درهم‌ریختگی ۲۱

۵-۱ نتایج به دست آمده از حالت دستور نمادین در $K = ۰$ ۲۲

۵-۲ نتایج به دست آمده از حالت دستور نمادین در $K = ۵$ ۲۳

۵-۳ نتایج به دست آمده از حالت دستور نمادین در $K = ۵۰$ ۲۳

۵-۴ نمودار تغییرات تعداد برچسب‌های پیش‌بینی شده از حالت صفر نمونه تا حالت ۵۰ نمونه ۲۳

۵-۵ نمودار تغییرات تعداد برچسب‌های پیش‌بینی شده از با افزودن تعاریف اخبار غیرمهم . ۲۴

۵-۶ جدول دقت‌ها در حالات نمونه‌های متخلف در کل دادگان تست ۲۵

۵-۷ جدول دقت‌ها در حالات نمونه‌های متخلف برای مدل جما۲ . . . . . ۲۵

۵-۸ نمودار تغییرات تعداد برچسب‌های پیش‌بینی شده در مدل جما۲ . . . . . ۲۶

۵-۹ دقت‌های به دست آمده در حالت دستور سیستمی به زبان فارسی ۲۷

۵-۱۰ دقت‌های به دست آمده در حالت دستور سیستمی به زبان انگلیسی ۲۷

۵-۱۱ ماتریس درهم‌ریختگی در حالت دستور سیستمی به زبان انگلیسی ۲۸

۵-۱۲ نمودار معیارهای دقت‌سنجی در دسته‌بندی‌های مختلف خبری ۲۹

۵-۱۳ جدول معیارهای دقت‌سنجی در دسته‌بندی‌های مختلف خبری ۲۹

۵-۱۴ جدول دقت‌های دستورهای زنجیره تفکر و درخت تفکر ۳۰

۵-۱۵ جدول دقت‌های دستورهای زنجیره تفکر و درخت تفکر در دسته‌بندی‌های مختلف . . ۳۰
۵-۱۶ جدول دقت‌های دستورهای زنجیره و درخت تفکر در حالت $K =$ ۲۰ . . . . . ۳۱
۵-۱۷ جدول دقت‌های دستورهای زنجیره و درخت تفکر در حالت $K =$ ۲۰ در دسته‌بندی‌ها ۳۱
۵-۱۸ جدول دقت تشخیص اهمیت خبر در حالت استفاده از طبقه‌بند برای انتخاب دستور . . ۳۲
۵-۱۹ ماتریس درهم‌ریختگی پنج دستور منتخب ۳۲

آ_۱ مقایسه معیارهای دقت‌سنجی در Kهای مختلف ۴۶
آ_۲ مقایسه تعداد برچسب‌های تولید شده در کل دادگان تست ۴۷
آ_۳ مقایسه معیارهای دقت‌سنجی در Kهای مختلف بر روی کل دادگان تست ۴۷
آ_۴ مقایسه معیارهای دقت‌سنجی در Kهای مختلف در مدل جما۲ . . . . . ۴۸
آ_۵ دقت‌های به دست آمده در حالت دستور سیستمی به زبان انگلیسی با برچسب‌های عددی ۴۹
آ_۶ دقت‌های مدل در حالت کل متن خبری ۴۹
آ_۷ ماتریس درهم‌ریختگی در حالت کل متن خبری به عنوان ورودی ۵۰
آ_۸ جدول دقت‌های دستورهای زنجیره تفکر و درخت تفکر در دادگان آموزش ۵۰
آ_۹ جدول دقت‌های دستورهای زنجیره و درخت تفکر در دادگان آموزش براساس دسته‌بندی ۵۱
آ_۱۰ جدول دقت‌های دستورهای زنجیره و درخت تفکر در دادگان ارزیابی ۵۱

فصل ۱

مقدمه

در دنیای رو به پیش رفت روزمره، حجم عظیمی از اخبار شبانه‌روز به سمت کاربران روانده می‌شود. در این حین می‌دانیم که بسیاری از این اخبار مبنای درستی نداشته و بسیاری نیز برای کاربران بسیار اهمیت کمی دارد. با معرفی یک بستر که بتوان به وسیله آن اخبار مهم به خصوص با توجه به فرهنگ ایرانیان تشخیص داده خود یک چالش بزرگ اما بسیار کاربردی است. در اینجا با استفاده و بهره‌گیری از مدل‌های زبانی بزرگ و دانش که توسط آنها جمع‌آوری شده است به انجام این امر پرداختیم. در ادامه همچنین چالش‌های این مدل‌ها و منطبق نبودن آن طبق فرهنگ و عادات ایرانیان بررسی می‌کنیم و با ارائه روش یادگیری چند نمونه[1]، این مشکل را برای طرف می‌کنیم.

۱-۱ تعریف مسئله چگونگی بررسی مهم بودن یک خبر

مسئله به این شکل تعریف می‌شود که یک خبر در هر دسته‌ای که قرار داشته باشد یا دارای اهمیت بالا یا برچسب ۱ و یا دارای اهمیت پایین و برچسب ۰ است. با دادگان جمع‌آوری شده و برچسب‌گذاری‌های انسانی روی آنها، به ۵۵۰۹ داده آموزش و ۱۱۸۰ داده تست و ازیابی رسیده، که با با استفاده از آنها مدل‌ها توصیه نمونه براساس شباهت تعریف شده است و هدف آن است که مدل بتواند اهمیت خبر (۰ یا ۱) را تشخیص دهد و به کاربر اعلام کند.

[1] Few-Shot Learning

۱-۲ اهمیت موضوع تشخیص اهمیت اخبار

از اهمیت این کار و محیط توسعه داده‌شده می‌توان به موارد زیر اشاره کرد:

- تسهیل پیگیری اخبار برای کاربران، از آنجایی که این محیط توان تشخیص اخبار مهم در دسته‌ها مختلف را داشته، می‌توان برای کاربران صرفا اخبار مهم را دسته‌بندی کرده و آنها با خواندن این اخبار در وقت خود نسبت به خواندن مطالب بی‌اهمیت صرفه‌جویی خواهند کرد.
- بررسی قدرت استدلال و تفکر مدل‌های زبانی بزرگ، از آنجایی تشخیص اهمیت یک خبر کار نسبتا پیچیده‌ای برای این مدل‌ها احتساب می‌شود، این بستر فراهم شده است که قدرت استدلال و تحلیل مدل‌های مختلف در شرایط گوناگون ارزیابی و اعلام شود.
- در این کار، روش‌هایی برای بهبود و بهینه کردن دقت این مدل‌ها پیشنهاد و بررسی شده که در جنبه‌های دیگری غیر از تشخیص اخبار مهم می‌توان کمک کننده باشد و به کار گرفته شود. از جمله اینها مسئله طبقه‌بندی و یادگیری محتوای دستور یا درخواست داده شده به مدل‌های زبانی بزرگ است.

۱-۳ ادبیات به کار رفته در این پژوهش

از آنجایی که بخش‌هایی از این پروژه الهام گرفته و ادامه کار تنظیم نمادها[۱] بوده از ادبیات این کار نیز در اینجا استفاده شده است. نتظیم نمادها عبارت است از روشی که به جای برچسب‌های اصلی که در اینجا همان ۰ یا ۱ هستند، یک رشته از نمادها همانند !، #، & و کاراکترهای دیگر جایگزین شود و مدل نتواند به دانش پیشینه خود اتکا کند.

۱-۴ اهداف پژوهش

اهداف این پژوهش صورت گرفته به دو قسمت کلی تقسیم می‌شود:

- ابتدا با ساختار و تعریف اخبار مهم پرداخته شده است، در این پژوهش نتیجه‌ها و بررسی‌های انجام شده حاکی این موضوع است که اخبار در دسته‌های گوناگون و برای اشخاص با فرهنگ‌های مختلف اهمیت متفاوتی دارد. بنابراین انجام یک مسئله طبقه‌بندی روی آنها کار آسانی نبوده و با بهبودهای

انجام شده در این پژوهش، مسیری برای پژوهش‌های بعدی در جهت رسیدن که دقت بالا با درنظر گرفتن تمام این شرایط فراهم کند.

- مدل‌های زبان بزرگ که کانون اصلی توجه این پژوهش بوده است در این مسئله خاص به طور کامل بررسی شده و تمامی نقاط ضعف و قوت این مدل‌ها در تشخیص اهمیت اخبار بررسی شده است. همچنین تفاوت قدرت استدلال این مدل‌ها در زبان فارسی با انگلیسی مورد مقایسه قرار گرفته که خود می‌تواند مورد استناد برای پژوهش‌های آینده در این زمینه قرار گیرد.

فصل ۲

مفاهیم اولیه

در اینجا به مفاهیم اصلی به کار برده شده در این پروژه، و بررسی کاربرد و پیشینه آن می‌پردازیم.

۲-۱ مدل‌های زبانی بزرگ

مدل‌های زبانی بزرگ[1] به سامانه‌های هوش مصنوعی گفته می‌شوند که بر اساس پردازش زبان طبیعی طراحی شده‌اند و قادر به تولید و درک متن‌های انسانی در مقیاس وسیع هستند. این مدل‌ها با استفاده از حجم بسیار زیادی از داده‌های متنی آموزش می‌بینند و می‌توانند وظایف متنوع زبانی، از جمله ترجمه، خلاصه‌سازی و پاسخ به پرسش‌ها را انجام دهند.

مفهوم مدل‌های زبانی از دهه ۱۹۸۰ با ظهور الگوریتم‌های احتمالاتی ساده آغاز شد. با معرفی شبکه‌های عصبی در دهه ۱۹۹۰ و توسعه یادگیری عمیق در دهه ۲۰۱۰، مدل‌هایی مانند ترانسفورمرها و سیستم‌هایی نظیر جی‌پی‌تی (نسل اول تا سوم) و مدل‌های مشابه توانستند به کارایی فوق‌العاده‌ای دست یابند[۲]. افزایش قدرت محاسباتی و دسترسی به داده‌های بیشتر، این پیشرفت‌ها را تسهیل کرد.

۲-۲ یادگیری درونی

یادگیری درون‌متنی[2] به توانایی یک مدل زبانی اشاره دارد که بتواند بر اساس نمونه‌هایی که در همان متن ورودی ارائه می‌شود، وظایف جدیدی را یاد بگیرد. در این روش، نیاز به آموزش دوباره مدل وجود ندارد،

[1] Large Language Models
[2] In-Context Learning

بلکه مدل از اطلاعات داده شده در همان لحظه استفاده می‌کند[۳].

این مفهوم در اوایل دهه ۲۰۲۰ با توسعه مدل‌هایی مانند جی‌پی‌تی ۳ به وضوح مطرح شد. این مدل‌ها نشان دادند که بدون نیاز به آموزش دوباره، می‌توانند تنها با ارائه نمونه‌هایی در ورودی، وظایف مختلفی را انجام دهند. این پیشرفت‌ها نقطه عطفی در ساده‌سازی استفاده از مدل‌های زبانی محسوب می‌شوند.

۲-۳ تنظیم بر اساس دستورالعمل

تنظیم بر اساس دستورالعمل[۳] فرآیندی است که در آن یک مدل هوش مصنوعی با استفاده از داده‌هایی آموزش می‌بیند که حاوی دستورالعمل‌های خاصی برای انجام وظایف مختلف هستند[۴]. هدف این روش بهبود عملکرد مدل در درک و اجرای دستورالعمل‌هاست.

ایده این روش از مفاهیم یادگیری انتقالی نشأت گرفته است. در سال‌های اخیر، با توجه به توانایی مدل‌های بزرگ زبانی در تعمیم وظایف، محققان تلاش کردند تا این مدل‌ها را با داده‌های حاوی دستورالعمل بهبود دهند. پروژه‌هایی مانند اجرای دستور عمل در مدل‌های جی‌پی‌تی[۵] نشان‌دهنده موفقیت این رویکرد هستند.

۲-۴ درخواست‌های سامانه و کاربر

درخواست‌های سامانه[۴] و کاربر به متونی اطلاق می‌شود که برای هدایت مدل زبانی به سمت تولید پاسخ مناسب استفاده می‌شوند. درخواست سامانه معمولاً وظیفه مشخص‌کردن قواعد کلی را دارد، در حالی که درخواست کاربر هدف یا سؤال خاصی را بیان می‌کند[۶].

این مفهوم با گسترش استفاده از مدل‌های زبانی در تعاملات انسانی به وجود آمد. اولین تلاش‌ها برای تعریف و تمایز این دو نوع درخواست در توسعه رابط‌های کاربری تعاملی و چت‌بات‌ها مشاهده شد. این ایده در مدل‌های زبانی بزرگ تکامل یافت.

[3] Insturction Tuning
[4] System Prompt

۲-۵ مهندسی درخواست

مهندسی درخواست[5] به هنر و دانش طراحی درخواست‌ها برای هدایت مدل‌های زبانی جهت تولید پاسخ‌های دقیق و مفید اشاره دارد. این فرآیند شامل ایجاد ورودی‌هایی است که بتوانند بهترین نتیجه ممکن را از مدل دریافت کنند.

این مفهوم با ظهور مدل‌های زبانی پیچیده و نیاز به بهره‌برداری بهتر از توانایی‌های آن‌ها مطرح شد. در سال‌های اخیر، مقالات و ابزارهای بسیاری برای استانداردسازی و بهبود این فرآیند ارائه شده است[۷]. مهندسی درخواست در زمینه‌های مختلف، از پژوهش گرفته تا صنعت، نقش کلیدی ایفا می‌کند.

[5] Prompt Engineering

فصل ۳

کارهای پیشین

همواره در طول زمان بررسی اهمیت اخبار چه در زبان فارسی و چه در زبان انگلیسی یک دغدغه و یک کار مبهم بوده است. از آنجایی که اهمیت یک خبر وابسته به عوامل مختلف همانند فرهنگ، موقعیت جغرافیایی، سلایق شخصی و دیدگاه‌های کاربران بوده در نگاه اول به نظر این کار، ناممکن می‌رسد. اما پژوهش‌های اخیر نشان‌داده است که با استفاده از دادگان‌های برچسب‌گذاری شده و استفاده از یادگیری چند نمونه‌ای می‌توان به نتایج قابل‌قبولی برای این قسمت رسید.

در اینجا به روش‌های مختلف که در گذشته برای بررسی اهمیت اخبار توسعه داده‌شده است پرداخته شده است و سپس مسیرهای مختلف بررسی و آنالیز این طبقه‌بندی را در مدل‌های زبانی بزرگ بیان شده است.

۳-۱ تشخیص اهمیت اخبار

این بخش به بررسی روش‌های مختلفی می‌پردازد که در طول زمان برای تشخیص اهمیت اخبار استفاده شده‌اند. این روش‌ها شامل رویکردهای کلاسیک، یادگیری ماشین و هوش مصنوعی، یادگیری عمیق و در نهایت مدل‌های زبانی بزرگ هستند.

۳-۱-۱ رویکردهای کلاسیک

در روش‌های کلاسیک، تشخیص اهمیت اخبار بیشتر بر اساس معیارهای دستی انجام می‌شد. از معیارهایی مانند طول خبر، تعداد دفعات ذکر شدن یک موضوع در منابع مختلف، یا تحلیل‌های آماری ساده برای این

کار استفاده می‌شد[۸]. این روش‌ها به دلیل محدودیت در قابلیت درک معنایی متون، کارایی پایینی در مسائل پیچیده داشتند.

۳-۱-۲ رویکردهای تشخیص اهمیت با استفاده از یادگیری ماشین

با ظهور الگوریتم‌های یادگیری ماشین[۱]، از مدل‌هایی مانند ماشین بردار پشتیبان[۲]، دسته‌بند بیزین ساده و جنگل‌های تصادفی[۳][۹] برای تحلیل اخبار و تشخیص اهمیت آن‌ها استفاده شد[۱۰]. این روش‌ها از ویژگی‌های استخراج‌شده مانند تعداد کلمات کلیدی، میزان تعامل خبرها و تعداد نقل‌قول‌ها بهره می‌بردند.

۳-۱-۳ رویکرد یادگیری عمیق

با توسعه یادگیری عمیق[۴]، استفاده از شبکه‌های عصبی مانند شبکه‌های بازگشتی و شبکه‌های توجه‌محور برای درک معنایی متون و تحلیل اخبار رواج یافت. مدل‌هایی نظیر LSTM و GRU توانستند با درک وابستگی‌های طولانی‌مدت در متن[۱۱]، عملکرد چشمگیری ارائه دهند.

۳-۱-۴ استفاده از مدل‌های زبانی بزرگ

مدل‌های زبانی بزرگ با پیش‌آموزش بر داده‌های گسترده، توانایی تحلیل متون خبری را با دقت بالا فراهم کرده‌اند[۲]. این مدل‌ها با توجه به پیکربندی و حجم داده‌های آموزشی، قادرند وظایف مختلف را به‌صورت چندمنظوره انجام دهند.

در کارهای پیشین انجام شده بررسی شده که رفتار مدل‌های زبانی بزرگ در تشخیص اخبار جعلی چطور بوده است. به خصوص در پژوهش‌های قبلی[۱۲] با تعریف بازیگر خوب و بد، سعی بر ارزیابی این نوع اخبار داشته و نشان می‌دهد که این مدل‌ها توانایی مناسب جهت تشخیص اخبار جعلی در شرایط از پیش‌تعریف شده مناسب خواهند داشت.

همچنین کارهای فراتری نسبت به صرفا اتکا کردن به پردازش متن انجام شده است، به طوری که با بهره‌گیری همزمان از مدل‌های تصویری مانند CLIP اهمیت اخبار براساس محتوای تصویری، ویدیوی و صوتی به همراه متن آنها نیز بررسی شود[۱۳].

[1] Machine Learning
[2] SVM
[3] Random Forest Tree
[4] Deep Learning

۳-۲ تنظیم بر اساس دستورالعمل و تنظیم نمادین

تنظیم بر اساس دستورالعمل به آموزش مدل‌های زبانی بزرگ با داده‌هایی اشاره دارد که شامل دستورالعمل‌های دقیق برای انجام وظایف هستند[۱۴]. این روش باعث می‌شود مدل‌ها بتوانند وظایف مشخصی مانند دسته‌بندی اهمیت اخبار را با دقت بیشتری انجام دهند. از سوی دیگر، تنظیم نمادین[۵] شامل استفاده از اطلاعات ساختاریافته مانند نمودارهای دانش یا نمایش‌های معنایی برای تقویت عملکرد مدل‌ها است.

۳-۳ یادگیری چندنمونه‌ای در تشخیص اهمیت اخبار

در این روش، مدل‌ها با تعداد بسیار کمی از نمونه‌های آموزشی، وظایف خود را یاد می‌گیرند. این ویژگی در تحلیل اخبار و تشخیص اهمیت آن‌ها، به‌ویژه در مواقعی که داده‌های آموزشی محدود است، کاربرد دارد. مدل‌هایی مانند Aya توانسته‌اند نشان دهند که تنها با چند نمونه ورودی می‌توانند وظایف پیچیده را انجام دهند[۱۵].

۳-۴ تشخیص اهمیت اخبار فارسی

در گذشته، تلاش‌هایی برای توسعه مدل‌های تشخیص اهمیت اخبار فارسی صورت گرفته است. این تلاش‌ها بیشتر بر اساس روش‌های یادگیری ماشین کلاسیک بوده و از ویژگی‌های زبانی خاص فارسی مانند ریشه‌یابی و تحلیل صرفی بهره گرفته‌اند[۱۶]. با این حال، استفاده از مدل‌های زبانی بزرگ برای زبان فارسی هنوز در مراحل ابتدایی قرار دارد.

همچنین این پژوهش، ادامه مسیر کار خبرچین[۱۷] بوده که با استفاده از مدل‌های مبنی بر معماری ترانسفورمر سعی داشته که به بررسی اهمیت اخبار بپردازد. در این پژوهش سعی شده با استفاده از مدل‌های زبانی بزرگ رویکرد کلی‌تری نسبت به بررسی اهمیت اخبار ارائه شود که بتواند بستر جامع‌تری برای طبقه‌بندی این حوزه فراهم کند.

[5] Symbol Tuning

۳-۵ سامانه‌های درخواست وابسته به پرسش

این سامانه‌ها با طراحی درخواست‌هایی که وابسته به موضوع پرسش هستند، قادرند نتایج بهینه‌ای در تشخیص اهمیت اخبار ارائه دهند. برای این منظور، از مهندسی درخواست استفاده می‌شود تا مدل‌های زبانی بزرگ بتوانند بر اساس متن ورودی و هدف پرسش، خروجی مطلوبی تولید کنند[۱۸].

پژوهش‌های اخیری در این زمینه انجام شده است که نشان می‌دهد استفاده از دستورهای مختلف بسته به ورودی کاربر می‌تواند نتایج ثمربخش‌تر به ارقام بیاورد. اگرچه در یکسری پژوهش‌های به این رویکرد یادگیری تقویتی[۶] الحاق می‌شود[۱۹] اما بیشتر یک سیستم در پس‌زمینه بوده که بتواند بهترین دستور را با توجه به ورودی و محتوا تشخیص دهد.

[6] Reinforcement Learning

فصل ۴

روش پیشنهادی

در این قسمت به روش‌های توسعه داده‌شده و نحوه به دست آمدن نتایج و خروجی‌ها می‌پردازیم. فرآیند به این صورت طی می‌شود که ابتدا دستورالعمل مناسب براساس شرط‌های مشخص شده انتخاب می‌شود و سپس در صورت نیاز نمونه‌های مشابه به خبر مورد نظر در دستور آمده و سپس مدل‌های زبانی بزرگ ۸ یا ۹ میلیارد پارامتر روی کارت گرافیکی بالا آمده و به صورت دسته‌های ۱۰ تایی نتایج از خروجی این مدل‌ها ذخیره شده و تحلیل‌ها روی آن انجام می‌شود.

۴-۱ انواع دستورالعمل‌های توسعه‌داده شده

در اینجا به چهار حالت پرامپت‌ها یا همان دستعورالعمل‌ها را که به‌کار گرفته شده است پرداخته می‌شود و کاربرد و اهداف هرکدام بررسی می‌شود. یک دستورالعمل p_i از یک مجموعه $P = \{p_۱, p_۲, \ldots, p_n\}$ انتخاب می‌شود و براساس آن نتایج خروجی که به صورت برچسب‌های $l_j =< ۰, ۱ >$ مشخص می‌شود خروچی گرفته می‌شود.

۴-۱-۱ دستورالعمل‌های وانیلا

دستورالعمل‌های وانیلا[1] که همان دستورالعمل‌های خام بوده صرفا اطلاعات مورد نیاز را برای مدل‌های زبانی بزرگ تهیه و توضیح می‌دهد. در ابتدا وظیفه طبقه‌بندی انجام شده توضیح داده‌می‌شود و سپس مشخص می‌شود که حتما خروجی به حالت $l_j =< ۰, ۱ >$ می‌بایست باشد و چندین بار روی این موضوع

[1] Vanila

تاکید می‌شود تا مطمئن شویم خروجی این مدل‌ها صرفا یک برچسب باشد. سپس توضیح اخبار مهم براساس دسته‌های مختلف تهیه شده و درنهایت خبری که می‌خواهیم طبقه‌بندی شود در پایان این دستور می‌آید. برای نمونه می‌توان یک دستور را به شکل زیر مشاهده کنیم:

برای این وظیفه‌ی طبقه‌بندی، از شاخه‌های فکری زیر استفاده کنید تا تصمیم بگیرید که آیا خبر «مهم» (۱) است یا «غیر مهم» (۰):

۱. ابتدا بررسی کنید که آیا موضوع خبر می‌تواند بخش بزرگی از کاربران فارسی‌زبان را تحت تأثیر قرار دهد و به گستردگی احتمالی آن توجه کنید. ۲. سپس محتوای موضوع را از نظر اهمیت اقتصادی، سیاسی و اجتماعی تحلیل کنید.

برای اخبار اقتصادی: عواملی مانند تورم، وضعیت مسکن و روندهای بورس که برای کاربران عادی اهمیت دارند را مدنظر قرار دهید.

برای اخبار سیاسی: ارزیابی کنید که آیا محتوا به سیاست‌های کلان ایران، تغییرات مهم در دولت، یا تعاملات جهانی مربوط می‌شود.

برای اهمیت اجتماعی: بررسی کنید که آیا خبر شامل رویدادهای ورزشی محبوب یا موضوعاتی با جذابیت گسترده است.

۳. بررسی کنید که آیا جذابیت خبر عمومی است یا فقط برای مخاطبان خاصی جذابیت دارد.

اگر خبر از نظر گسترده‌ای مهم است، آن را با «۱» برچسب بزنید. در غیر این صورت، اگر بیشتر برای مخاطبان خاص جذاب است، «۰» را انتخاب کنید. تحلیل این سناریوها را به پایان رسانده و طبقه‌بندی نهایی را به‌صورت زیر ارائه دهید:

طبقه بندی نهایی: «۰ یا ۱»

که این نمونه آورده شده شامل مفهوم درخت‌های تفکر نیز می‌شود. نمونه‌های بیشتر این نوع دستورالعمل‌ها را می‌توانید در قسمت مطالب تکمیلی مشاهده کنید.

۴-۱-۲ دستورالعمل‌های چندنمونه‌ای

در این نوع دستورالعمل‌ها علاوه برای بر محتوای خود پرامپت، چندین نمونه براساس خبر خواسته شده نیز آورده می‌شود. فرض کنید خبری که اکنون می‌خواهیم طبقه بندی کنیم به صورت t_i از مجموعه دادگان تست یعنی $T = \{t_۱, t_۲, \ldots, t_n\}$ باشد. آنگاه به ازای هر هر تایتل یا عنوان خبر در دادگان آموزش خود متن و همچنین نسخه برداری آن را به وسیله مدل TF-IDF خواهیم داشت که به این صورت تعریف می‌شود:

$$S_{<Title>} = \{s_۱, s_۲, \ldots, s_n\} \xrightarrow{TF-IDF} V_{<Title,TF-IDF>} = \{v_۱, v_۲, \ldots, v_n\} \qquad (۱-۴)$$

که در آن s_i مشخص کننده یک نمونه و v_j نشان‌دهنده بردار شده است. پس از این اقدام با روش $cosine\ similarity$ از تمامی بردارها، براساس k انتخابی یا همان تعداد نمونه‌ها، شبیه‌ترین عناوین خبرها را همراه برچسب اهمیت آنها قرار می‌دهیم.

$$E_{<samples>} = \{< s_i, l_i > | s_i \in \text{argmax}_k \left(cosine(V_{<Title,TF-IDF>}, t_i)\right)\} \qquad (۲-۴)$$

این نمونه‌های کمک می‌کند که مدل مشاهده کند که هرکدام از اخبار شبیه به خبر داده شده به چه صورت از نظر اهمیت ارزیابی شده است و سپس تصمیم نهایی خود را براساس آن تغییر دهد. نمونه‌های در دستور

العمل به این صورت قرار می‌گیرد:

نمونه‌ها: به نمونه‌های زیر نگاه کنید و بر اساس آنها تشخیص دهید که کدام خبرها مهم هستند و با عدد '۰' نمایش داده می‌شوند و کدام خبرها غیرمهم هستند و با عدد '۱' نمایش داده می‌شوند.

نمونه: هشدار؛ بارش باران و برف در این استان‌ها طبقه بندی نهایی: ۱

نمونه: بارش برف و باران در جاده‌های ۱۴ استان طبقه بندی نهایی: ۱

نمونه: پیش‌بینی بارش برف و باران در این استان‌ها| گرم‌ترین و سردترین شهرها کدامند؟ طبقه بندی نهایی: ۱

نمونه: بارش برف و باران از فردا در این استان‌ها شروع می‌شود طبقه بندی نهایی: ۱

نمونه: پیش‌بینی بارش ۵ روزه در ۸ استان طبقه بندی نهایی: ۱

نمونه: پیش‌بینی رگبار و رعدوبرق طی ۵ روز آتی در برخی نقاط کشورا وزش باد شدید در شرق طبقه بندی نهایی: ۱

نمونه: بارش باران در نقاط مختلف کشور طبقه بندی نهایی: ۱

نمونه: هشدار! دیگر بارش برف و باران هم تاثیر چندانی در آلودگی هوا ندارد طبقه بندی نهایی: ۰

نمونه: ثبت بیش از ۲۳ میلیون سفر نوروزی فقط در روز گذشته | بارش برف و باران در جاده‌های ۴ استان طبقه بندی نهایی: ۰

نمونه: بارش برف و باران در جاده‌های ۵ استان | رانندگان قبل از سفر حتما با این سامانه تماس بگیرند طبقه بندی نهایی: ۰

نمونه: هوای تهران در آخر هفته | چه خبر از بارش برف و باران؟ طبقه بندی نهایی: ۰

نمونه: ورود سامانه بارشی ؛ بارش باران در ۸ استان طی ۲۴ ساعت آینده | تهران و ۳ استان دیگر منتظر وزش باد شدید باشند طبقه بندی نهایی: ۱

نمونه: بارش باران و تگرگ در آذربایجان غربی طبقه بندی نهایی: ۰

نمونه: تشدید بارش‌ها از فردا در کشور ؛ بارش برف در تهران | کاهش ۴ تا ۸ درجه‌ای دما در نوار شمالی طبقه بندی نهایی: ۱

نمونه: بارش برف و باران در این ۱۱ استان | آماده‌باش امدادگران هلال احمر | ورود سامانه بارشی جدید به کشور از امروز طبقه بندی نهایی: ۱

نمونه: تداوم بارش در برخی استان‌ها | وزش باد شدید و خیزش گرد و خاک در شرق کشور طبقه بندی نهایی: ۱

نمونه: هشدار هواشناسی | افزایش آلودگی هوا در ۲ کلانشهر | وزش باد شدید و ارتفاع امواج در ۳ استان طبقه بندی نهایی: ۰

نمونه: پیش بینی بارش ها در کشور | امروز پنجشنبه فقط ۲ استان بارندگی ندارد | گرمترین و سردترین شهرهای کشور طبقه بندی نهایی: ۱

نمونه: عکس ا اقدام عجیب عراق در جعل نام خلیج فارس طبقه بندی نهایی: ۱

نمونه: باخت برانکو به عراق در فینال جام خلیج‌فارس طبقه بندی نهایی: ۰

۴-۱-۳ دستورالعمل‌های زنجیره تفکر

در این نوع دستورها، از مدل زبانی بزرگ خواسته می‌شود که صرفا به یک زاویه دید اتکا نکند و موقعیت‌های مختلف را بررسی کند. بعضا در این حالت به مدل اجازه می‌دهیم که خروجی خود را فراتر از برچسب $l_j =< ۰, ۱ >$ ببرد و سعی کند تفکرات خود را بسازد و سپس طبقه‌بندی نهایی را در یک قالب مشخص بیان کند که سپس بتوان آن را استخراج کرد. برای نمونه یک نوع از این پرامپت‌های استفاده در زبان انگلیسی را می‌توانید به صورت زیر مشاهده کنید:

The goal is to classify news items into 'important' (1) or 'not important' (0). To classify accurately, follow these steps:

1. Identify if the news topic could be relevant to a large Persian-speaking audience.

2. Assess if it pertains to significant economic events (currency or inflation changes, housing updates, etc.), critical political events (government actions, international relations), or socially impactful themes that could affect many people.

3. Finally, determine if the story has widespread appeal or is only relevant to a niche audience.

If the news is of broad interest and covers the themes above, label it as '1'. Otherwise, label it as '0'. Upon completing the scenario analysis, output the final classification using the format below:

Final Classification: [1 or 0]

۴-۱-۴ دستورالعمل‌ها درخت تفکر

در این حالت از دستورها، علاوه بر زنجیره تفکر از مدل خواسته می‌شود که در یک یا دو تا از زنجیره‌ها، حالت‌های دیگر را به صورت درختی بررسی کند. در اینجا، از آنجایی که هدف ما طبقه‌بندی اخبار است خواسته شده که در انواع دسته‌بندی خبرها همانند اقتصادی، فرهنگی، سلامت و بهداشت و ورزشی مدل در خصوص اهمیت خبر فکر کرده و سپس اظهار نظر کند.

۴-۲ رویکرد دستورالعمل سیستمی و کاربر

یکی از رویکردهایی که در این پژوهش مورد بررسی قرار گرفته بررسی تاثیر تعریف قاعده‌مند دستور سیستمی و کاربری و جداسازی آن و تاثیر آن برای دقت خروجی مدل‌های زبانی بزرگ است. دستورالعمل سیستمی، دستوری است که معمولا ثابت است و وظایف اصلی را برای مدل شرح می‌دهد که طبق آن به دستور کاربر پاسخ دهد.

در این پژوهش قواعد اصلی، تعریف مهم و غیرمهم بودن خبر، نمونه‌ها و شرح اصلی به صورت دستورالعمل سیستمی تعریف شده و خود خبری که می‌خواهیم مورد طبقه‌بندی قرار دهیم به صورت دستورالعمل کاربر[2] تعریف شده است. این رویکرد با رویکرد دستورالعمل یکپارچه مقایسه شده و در قسمت نتایج مشاهده شده که تعریف جداگانه آنها به بهبود کارایی مدل‌ها کمک شایانی می‌کند.

[2] User Prompt

۴-۳ تحلیل قدرت استدلال در حالت چند زبانی

در این پژوهش این سوال مورد هدف قرار گرفته است که تقاوتی بین دستورالعمل به زبان انگلیسی و فارسی وجود دارد؟ آیا مدل‌های زبان بزرگ فرهنگ و موقعیت جغرافیای کشور ایران را بیشتر در زبان فارسی ذخیره و استدلال می‌توانند بکنند و یا اینکه در حالتی که به صورت زبان انگلیسی به آنها دستور دهیم قدرت استدلال بیشتری خواهند داشت؟ با بررسی دستورهای سیستمی در فصل نتایج به هر دو زبان، نتایج به دست آمده حاکی است که در یکسری شرایط خاص و وابسته به مدل استفاده شده، تفاوت عملکرد متفاوت است به طوری که در دستورهای وانیلا در زبان انگلیسی قدرت استدلال بیشتر و در حالت‌های درخت تفکر و زنجیره تفکر به نظر می‌آید در زبان فارسی قدرت تفکر بیشتری داریم.

۴-۴ رویکرد تنظیم نمادها

این رویکرد که الهام گرفته از روش معرفی شده در مقاله[۱] هست بر این عنوان تمرکز می‌کند که به جای قرار دادن تعریف برچسب‌های اصلی برای یک مدل زبانی بزرگ یا همان $l_j =< ۰, ۱ >$ به جای برچسب‌های اصلی، برچسب‌های نمادین و بی‌ارتباط استفاده شود و قدرت مدل در این سناریو بررسی شود.

در نگاه اول شاید سوال پیش بیایید که چرا اصلا این رویکرد موثر واقع خواهد شد؟ و چرا قرار ندادن واژه‌های «مهم» و «غیرمهم» و گمراه کردن مدل با سمبل‌های غیر مرتبط در نهایت می‌تواند به دقت مدل کمک کند؟ جواب این سوال‌ها را باید اینطور داد که تنظیم نمادها این هدف را دنبال می‌کند که مدل زبانی بزرگ نتواند به دانش پیشینه خود که روی آن آموزش دیده تکیه کند و تمامی خروجی‌ها را براساس نمونه‌های قرار داده شده در دستور یادبگیرد و پیش‌بینی کند.

در سناریو تشخیص اهمیت اخبار، بسیاری از اوقات مشاهده می‌شود که این مدل‌های رویکردی غیر از فرهنگ مربوط به کشور ایران در پیش‌بینی اهمیت اخبار به کار می‌برد زیرا که بیشتر داده‌های آموزش خود به زبان انگلیسی بوده و شاید یک چیز بی‌اهمیت برای یک فرد ایرانی، در فرهنگ غرب بسیار مهم واقع شود.

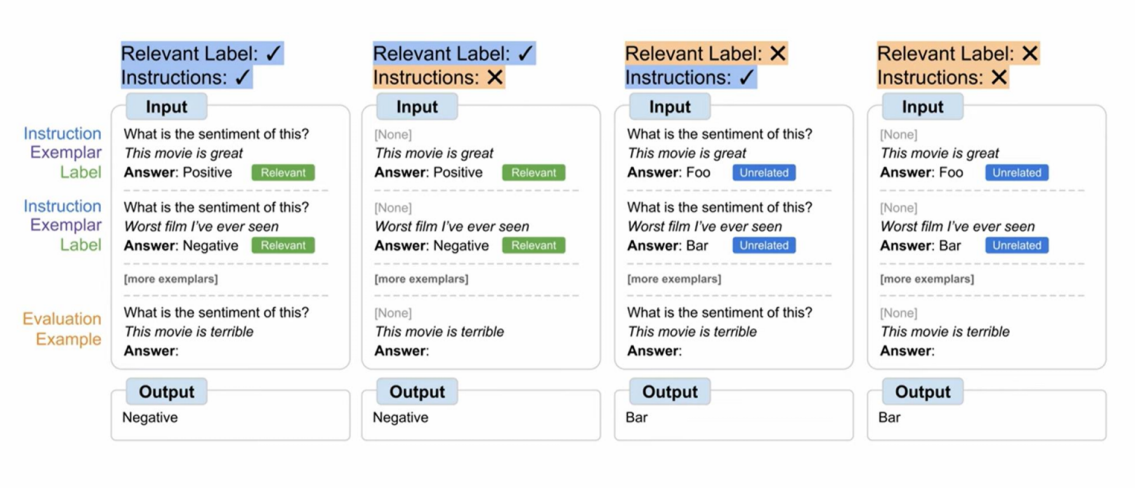
در این شرایط نیاز داریم که مدل بسیار بیشتر به مثال‌ها و نمونه‌های قرار داده شده توجه کند. در این پژوهش به جای برچسب‌های «مهم» و «غیرمهم» از برچسب‌های «۵۸» و «۴۷» استفاده شده است تا مدل را از دانش پیشینه خود محروم کند. یعنی در پرامپت قرار گرفته برای مدل هیچ واژه «مهم» یا «غیرمهم» وجود نداشته و مدل براساس مثال‌ها و تعاریف‌ها باید مفهوم دو برچسب «۵۸» و «۴۷» را یادبگیرد و سپس برچسب نهایی را براساس این دو پیش‌بینی کند.

$$l_j =< ۰, ۱ > \xrightarrow{SymbolTuning} l_{<Symbolic>} =< ۴۷, ۵۸ > \qquad (۴-۳)$$

یکی از کاربردهای دیگر این روش نیز سنجیدن حساسیت مدل‌ها به دستور داده شده است، از آنجایی که مدل نمی‌تواند به دانش پیشینه خود تکیه کند بسیار بیشتر به محتوای دستور حساس‌تر شده و با سنجیدن آنها به وسیله K های مختلف نمونه‌ها می‌توان رویکردهای جالبی از عملکرد این مدل‌ها و حساسیتی که به پراپمت نشان می‌دهد بررسی کرد.

۴-۵ تنظیم براساس دستورالعمل و حالت‌های مختلف آن

در رویکردی که در پرامپت یک شرح دستور و نمونه می‌آوریم چهار حالت ممکن طبق ۴-۱ داریم.



شکل ۴-۱: حالت‌های مختلف یک دستورالعمل و برچسب‌های آن

که همانطور که مشخص است چهار حالت داریم که در یکسری از حالت شرح خود وظیفه کامل مشخص است و در دیگر حالت‌ها صرفا به آوردن نمونه اتکا شده است. و همچنین حالت‌های دیگر به وسیله آوردن برجسب‌های مرتبط و غیرمرتبط شکل می‌گیرد.

در این کار، دو حالت شرح وظیفه به همراه برچسب مرتبط و شرح وظیفه به همراه برچسب غیرمرتبط بررسی شده است. آز آنجایی که بدون شرح وظیفه درک اینکه برچسب‌ها چه مفهمومی برای مدل زبانی بزرگ در تشخیص اهمیت خبر دارد بسیار پیچیده و مبهم می‌شود، به بررسی اینگونه حالات در این پژوهش نپرداخته شده است.

۴-۶ انواع متن بررسی شده از اخبار

همانطور که می‌دانیم یک متن خبری منتشر شده در سایت‌های دارای یک عنوان و یک متن است که هرکدام از آنها را می‌توان به صورت جداگانه بررسی کرد. علاوه براینها دادگان جمع‌شده از سایت‌های خبری شامل یک خلاصه بوده که از مدل‌های زبانی بزرگ گرفته شده است. بنابراین نمونه خبری t_i که می‌خواهیم نوع آن را بررسی کنیم می‌تواند در سه حالت زیر قرار بگیرد.

$$Type(t_i) \in \{T_{<Title>}, T_{<Summery>}, T_{<Text>}\} \tag{۴-۴}$$

در خصوص نمونه‌های e_i که برای یادگیری مدل مورد استفاده قرار می‌گیرد صرفا از عنوان خبرها یا همان $T_{<}Title>$ استفاده شده است زیرا که برای مدل‌های زبانی بزرگ دچار محدودیت برای ورودی هستیم و در حالت‌های $K = ۲۰$ و $k = ۵۰$ دچار این محدودیت شده اگر بخواهیم از نوع‌های $T_{<Summery>}$ و یا $T_{<Text>}$ استفاده کنیم، بنابراین مجموعه E همواره از توزیع عنوان خبرها خواهد بود.

۴-۷ نحوه خروجی گرفتن از مدل‌های زبانی بزرگ

یکی از چالش‌های اصلی خروجی گرفتن از این مدل‌ها مقدار رم گرافیکی مورد استفاده قرار گرفته است. به طوری که عملا خروجی گرفتن از مدل‌های بزرگتر از ۱۲ میلیاردی را غیرممکن می‌سازد.

در این پژوهش با استفاده از روش و تکنینک 4bit Quantized و لود کردن پارامترها در مقدار Float4 به جای Float32 می‌توانیم در مقدار رم گرفیکی ۱۶ گیگ مدل را بارگذاری کرده و از آن خروجی گرفت.

۴-۷-۱ مدل‌های زبانی بزرگ استفاده شده

در این کار، از دو مدل‌های زبانی بزرگ موجود به صورت عمومی یعنی مدل آیا ۲۳[3] با ۸ میلیارد پارامتر و مدل جما ۲ با ۹ میلیارد[4] پارامتر استفاده شده است.

همچنین کارت گرافیکی استفاده شده در اینجا، کارت گرافیکی P100 با مقدار رم گرافیکی ۱۶ گیگ بوده و تمامی خروجی‌ها براساس این گرفته شده است.

[3] Aya23
[4] Gemma2-9b-instruct

۴-۷-۲ نحوه بهینه خروجی گرفتن از مدل‌های زبانی بزرگ

در این پژوهش از معماری Fast Attension[۲۰] به جای معماری توجه ساده استفاده شده زیرا که این معماری دارای سرعت بسیار بالاتر به وسیله بهبود ضرب‌های ماتریسی از پیش تعریف شده است.

همچنین خروجی و ورودی‌های I/O به کمترین حالت خود رسیده تا بتوان در سریع‌ترین حالت ممکن از کل دادگان تست خروچی گرفت.

۴-۸ چرخه کلی خروجی گرفتن‌ها

در نهایت شبه‌کد زیر، روند کلی ساخت دستورالعمل‌ها با توجه به نمونه‌های و ورودی خبر به عنوان دستور کاربر نشان می‌دهد. در این چرخه از دو حلقه بسته به K انتخابی بهره گرفته می‌شود و برای به حداقل رساندن I/O نتایج به صورت دسته‌های ۲۰تایی ذخیره می‌شود.

الگوریتم ۱ ساخت دستور و خروجی گرفتن برچسب پیش‌بینی شده توسط مدل زبانی بزرگ

ورودی: خبر t_i از دادگان تست به عنوان دستور کاربر

خروجی: برچسب $l_j = \{< ۰, ۱ > \ or \ < ۴۷, ۵۸ >\}$ به عنوان پیش‌بینی مدل از اهمیت خبر

۱: انتخاب مدل M و پرامپت p_i از مجموعه $P = \{p_۱, \ldots, p_n\}$

۲: قراردادن مقدار K به عنوان تعداد نمونه

۳: انتخاب $Type(t_i)$ از مجموعه $\{T_{<Title>}, T_{<Summery>}, T_{<Text>}\}$

۴: تا وقتی مجموعه T اتمام نشود:

۵: اگر $K \neq ۰$:

۶: ساخت فضای برداری V از تمامی فضای نمونه

۷: $V_{<Title,TF-IDF>} = \{v_i \,|\, v_i = R_{TF-IDF}(e_i \in E)\} \leftarrow$

۸: ساخت فضای برداری $v^{(t_i)}$ از ورودی t_i

۹: تا وقتی K نمونه انتخاب نشده است:

۱۰: انتخاب e_i از مجموعه E

۱۱: $e_i = \{< s_i, l_i > | s_i \in \text{argmax}\left(cosine(V_{<Title,TF-IDF>}, v^{(t_i)})\right)\} \leftarrow$

۱۲: ساخت افزودن مجموعه نمونه $E^{(K)}$ به پرامپت p_i

۱۳: گرفتن خروجی برچسب از مدل $l_i = M(p_i, t_i) \leftarrow$

۱۴: ذخیره نتایج برچسب‌ها l_i به همراه پرامپت‌های p_i به صورت دسته‌های ۲۰تایی

۱۵: ذخیره دادگان و نتایج کلی

که درنهایت خروجی برای تحلیل داده‌ای و به دست آمدن دقت‌ها فراهم می‌شود.

۴-۹ ساختار نتایج

نتایج به دست آمده در یک فایل CSV ذخیره شده که تمامی حالات مختلف K را در خود داشته و هچنین عنوان خبر، برچسب اصلی t_i، برچسب‌های پیش‌بینی شده $l_i^{(K)}$ و دسته‌بندی خبر یا همان c_i در خود خواهد داشت.

برای نمونه ساختار نتایج به دست آمده را می‌توان در شکل ۴-۲ مشاهده کرد.

text	text_type	real_tag	category	predicted_k_0	predicted_k_20
1178 unique values	1 unique value	0 1	سیاسی 27% اجتماعی 22% Other (601) 51%	0 1	0 1
برگزاری دادگاه پرونده کثیرالشاکی شرکت کاغذی «آذین خودرو ساریا»	only_title	0	اجتماعی	0.0	0.0
خبر جدید وزیر بهداشت درباره بازگشایی مدارس در مهر ماه	only_title	1	اجتماعی	1.0	0.0
شکایت باشگاه استقلال از عیسی آل کثیر	only_title	1	ورزشی	0.0	0.0
دولت فرانسه مسئول عواقب اهانت بی‌شرمانه علیه مقدسات مسلمانان جهان است	only_title	0	سیاسی	0.0	1.0
تامین روشنایی کنارگذر نواب	only_title	0	اجتماعی	0.0	0.0
شما نظر دهید/ ریشه و پیامدهای خشونت‌های اخیر در مدارس چیست ؟	only_title	0	اجتماعی	1.0	1.0
استان کشور متاثر از شرایط ۲۰ جوی / رهاسازی ۲۶۸ خودرو از برف و کولاک	only_title	0	اجتماعی	0.0	1.0
همه علیه کنعانی زادگان، حتی اپوزیسیون ها!	only_title	0	ورزشی	1.0	1.0
رضا فیاضی به دلیل ابتلا به کرونا بستری شد	only_title	0	فرهنگی و هنری	0.0	0.0

شکل ۴-۲: ساختار نتایج ذخیره شده و مشخص‌کردن برچسب‌های $l-i$ پیش‌بینی شده از مدل

۴-۱۰ نحوه تحلیل نتایج

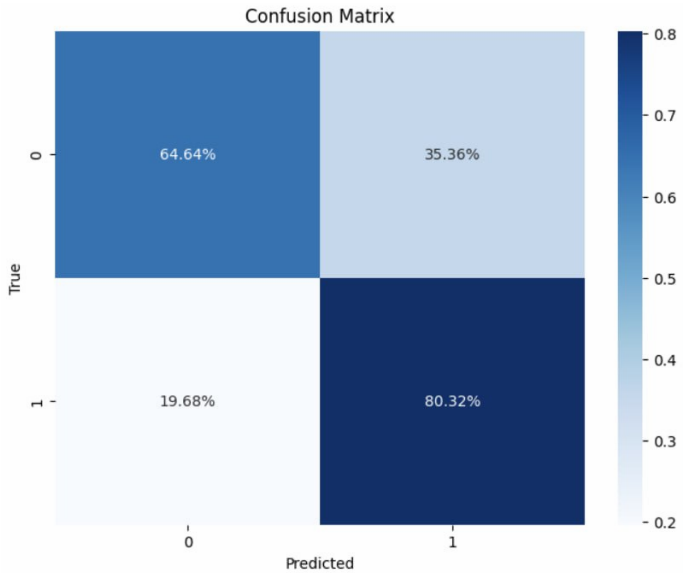
از آنجایی که در دادگان تست، ناترازی داده‌ای داریم به صورتی که حدود یک پنجم داده‌ها مهم و مابقی غیرمهم بوده است (که نشان‌دهنده توزیعی از خبرها برای یک کاربر دارد) معیاری که برای دقت‌سنجی ما بسیار اهمیت دارد Macro F1-Score خواهد بود.

همچنین تمامی نتایج در حالت K-fold نیز مورد بررسی قرار گرفته است که مطمئن شود واریانس درستی از نتایج داده‌ها وجود داشته باشد.

```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.92      0.65      0.76      4324
           1       0.38      0.80      0.52      1184

    accuracy                           0.68      5508
   macro avg       0.65      0.72      0.64      5508
weighted avg       0.81      0.68      0.71      5508

  Number of '1' labels: 2480
  Number of '0' labels: 3028
```

شکل ۴-۳: نمونه‌ای از نوع نتیجه و دقت‌سنجی اعلامی

و در آخر توزیع ماتریس درهم‌ریختگی[5] آن نیز بررسی می‌شود که بتوان متوجه شد در چه حوزه یک دستور خوب عمل کرده و در چه حوزه‌ای ضعف داشته است.



شکل ۴-۴: نمونه‌ای از بررسی ماتریس درهم‌ریختگی

[5]Confusion Matrix

فصل ۵

نتایج جدید

در اینجا به نتایج به دست آمده و تحلیل کامل آنها می‌پردازیم. ابتدا به دستورهای نمادین توسعه داده شده می‌پردازیم، سپس به نقش دستور سیستمی و تاثیر آن بر روی دقت مدل پرداخته، قدرت استدلال مدل‌ها را در زبان فارسی و انگلیس بررسی کرده و در نهایت به بررسی نتایج حاصل از پرامپت درخت تفکر و زنجیره تفکر می‌پردازیم.

۵-۱ نتایج تنظیم نمادین

همانطور که در قسمت‌های قبل اشاره شد، در اینجا برچسب‌های اصلی یعنی $l_i =< ۰,۱ >$ را با برچسب‌های نمادین جایگذاری کردی و تاثیر آنها را در طول انتخاب K های مختلف بررسی کنیم. (تمامی نتایج زیر در مدل ۲۳Aya گرفته شده است.)

با توجه به ۵-۱ می‌توان فهمید که دقت بسیار پایین بوده است (تمامی اعداد گزارش شده براساس دقت برای پیش‌بینی درست برچسب «مهم» اخبار است) و دلیل آن به این خاطر بوده که مدل تمامی اخبار را «غیرمهم» پیش‌بینی کرده چرا که هیچ تعریف و نمونه‌ای از خبر «غیرمهم» نداشته است.

K = 0	Accuracy	Precision	Recall	F1-Score	# of '58'	# of '47'
Title	17%	14%	93%	24%	96	5

شکل ۵-۱: نتایج به دست آمده از حالت دستور نمادین در $K = ۰$

در حالتی که پنج نمونه به عنوان مثال برای مدل تهیه شده است می‌بینیم که به حالت متعادل‌تری رسیدیم

و مدل کم‌کم در حال یادگیری آن است که چه چیز «غیرمهم» است.

K = 5	Accuracy	Precision	Recall	F1-Score	# of ’58’	# of ’47’
Title	47%	16%	64%	25%	58	43

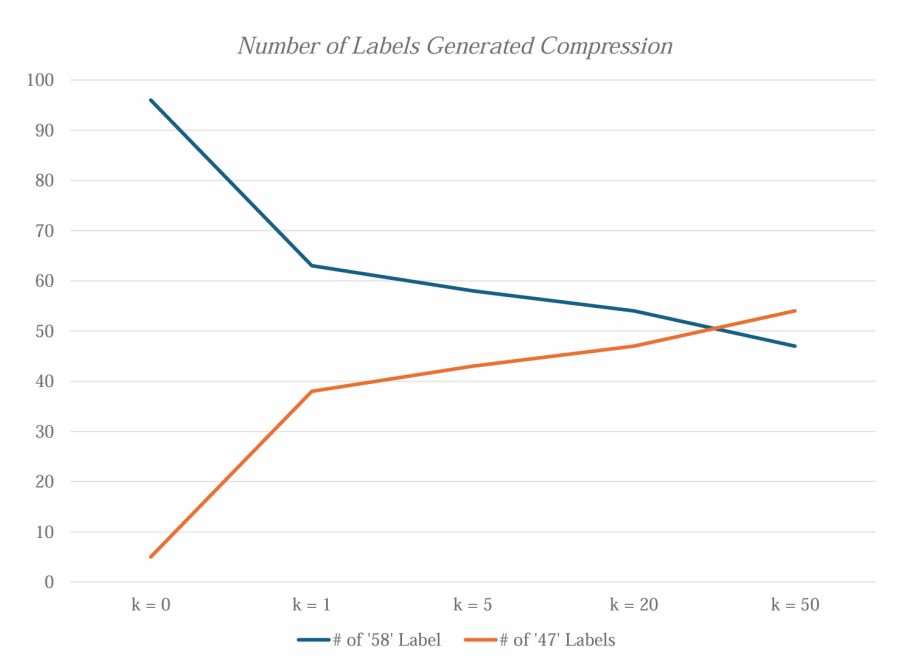
شکل ۵-۲: نتایج به دست آمده از حالت دستور نمادین در $K = ۵$

حال با توجه به شکل ۵-۳ می‌توان دید در حالتی که $K = ۵۰$ باشد مدل تعریف و نمونه‌های مشابه e_i غیرمهم بیشتری دیده و به این سمت می‌رود که اخبار بیشتری را «غیرمهم» تشخیص دهد.

K = 50	Accuracy	Precision	Recall	F1-Score	# of ’58’	# of ’47’
Title	55%	17%	57%	26%	47	54

شکل ۵-۳: نتایج به دست آمده از حالت دستور نمادین در $K = ۵۰$

در نهایت در شکل ۵-۴ می‌توان دید که در طول افزودن نمونه‌ها، مدل بیشتر و بیشتر با مفاهیم اخبار «غیرمهم» آشنا شده و اعتماد بیشتر برای «غیرمهم» پیش‌بینی کردن دارد.

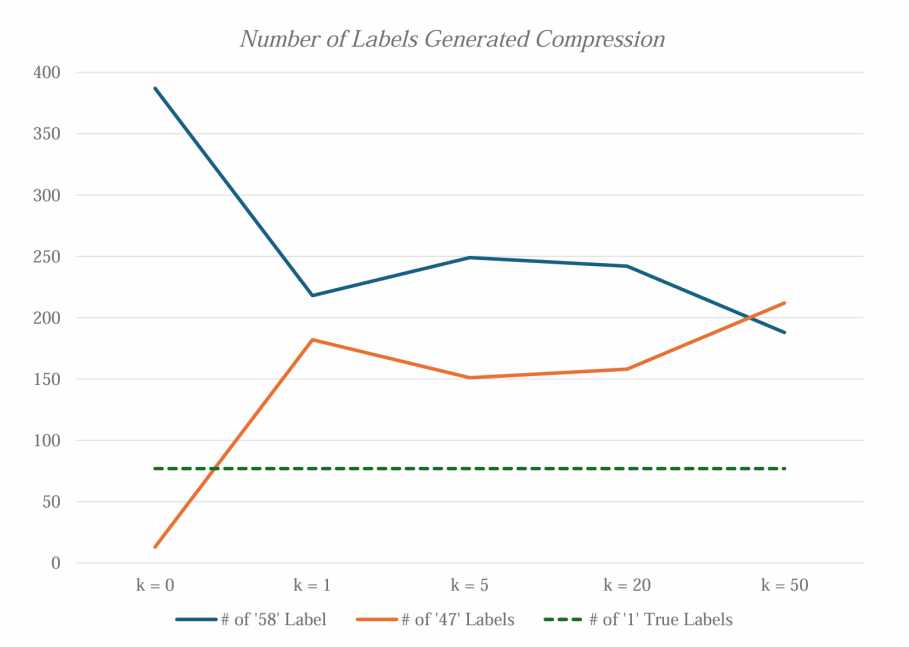


شکل ۵-۴: نمودار تغییرات تعداد برچسب‌های پیش‌بینی شده از حالت صفر نمونه تا حالت ۵۰ نمونه

همانطور که مشخص است خط آبی بیانگر تعداد اخبار «مهم» پیش‌بینی شده و خط قرمز تعداد اخبار «غیرمهم» را بیان می‌کند.

۵-۱-۱ تنظیم نمادین با افزودن تعریف اخبار غیرمهم

همانطور که در قسمت قبل بیان شد، از آنجایی که در دستورالعمل ورودی هیچ تعریفی از خبر «غیرمهم» نداریم مدل به این سمت می‌رود که تمامی اخبار را مهم ببیند. یک بهبودی که داده شده است این است که محتوای اخبار «غیرمهم» یا همان برچسب «۴۷» نیز به دستور اضافه شده است و همانگونه که مشاهده می‌شود منجر به رفتار در نمودار ۵-۵ می‌شود.



شکل ۵-۵: نمودار تغییرات تعداد برچسب‌های پیش‌بینی شده از با افزودن تعاریف اخبار غیرمهم

در این نمودار خط سبز رنگ بیانگر تعداد واقعی اخبار «مهم» در دادگان تست ما هست. همانطور که می‌توان دید، با افزودن تعاریف اخبار غیر مهم، مدل اعتماد بیشتری برای پیش‌بینی غیرمهم بودن خبر دارد زیرا تا حدودی می‌دانید یک خبر «غیرمهم» چیست و با افزودن نمونه‌ها در هر قدم، تصویر کلی بهتری از آن پیدا کرده و همواره به نوار سبز رنگ نزدیک‌تر می‌شود.

تنها رفتار عجیب مشاهده شده در حالت $K = ۱$ بوده که با شیب خیلی بیشتری در قیاس با بقیه حالت‌های در خصوص کمتر پیش‌بینی کمتر اخبار مهم رخ داده است. دلیل آن با توجه به بررسی‌های صورت گرفته به این خاطر است که در حالت تک مثاله، مدل بسیار حساس به برچسب مثال داده شده می‌شود و صرفا بر آن تکیه می‌کند و در عمل یادگیری ندارد. یعنی اگر برچسب نمونه «۴۷» باشد خروجی را «۴۷» اعلام می‌کند و اگر «۵۸» باشد همان اعلام می‌کند.

در صورتی که با افزوده شدن مثال‌ها، این حساسیت کمتر شده و مدل رویکرد کلی‌تری نسبت به تشخیص اهمیت اخبار دارد.

و درنهایت نتایج کل پایگاه تست در این حالت را می‌توان در ۵-۶ مشاهده نمود.

Title Only	Accuracy	Precision	Recall	F1-Score	# of '58'	# of '47'
k = 0	19%	17%	97%	28%	1121	40
k = 1	53%	21%	67%	32%	609	552
k = 5	49%	21%	77%	33%	701	460
k = 20	46%	20%	73%	31%	711	450
k = 50	58%	22%	62%	33%	537	624
Tr Labels					196	983

شکل ۵-۶: جدول دقت‌ها در حالات نمونه‌های متخلف در کل دادگان تست

۵-۱-۲ نتایج دستور نمادین در مدل جما۲

حال پس از بررسی مدل آیا۲۳، به بررسی مدل جما۲ با ۷ میلیارد پارامتر می‌پردازیم. در ۵-۷ می‌توانید تمامی دقت‌های به دست آمده در K های مختلف را مشاهده کنید.

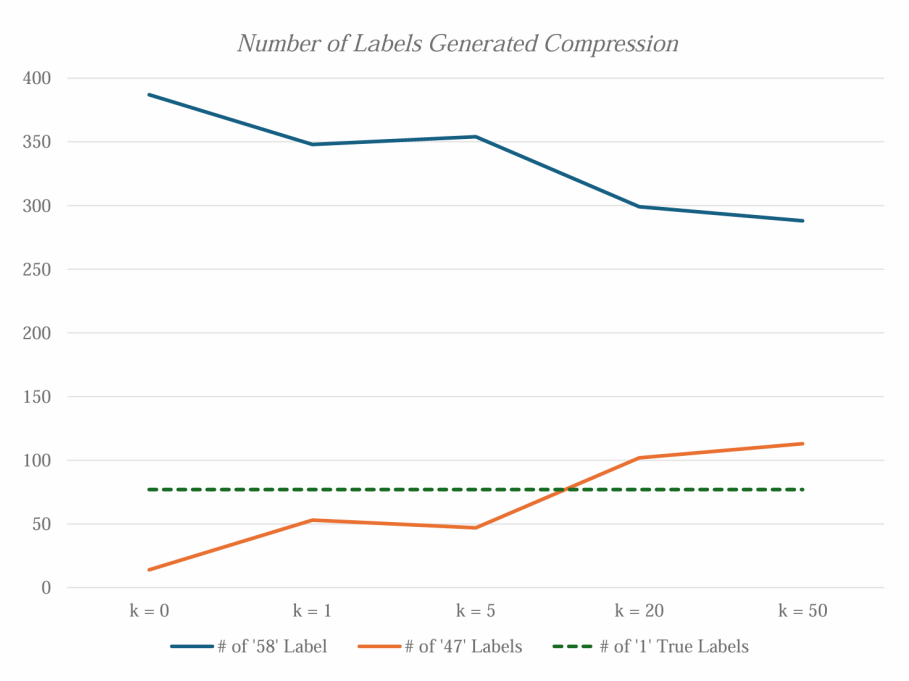
Title Only	Accuracy	Precision	Recall	F1-Score	# of '58'	# of '47'
k = 0	23%	20%	100%	33%	387	14
k = 1	29%	20%	91%	33%	348	53
k = 5	29%	21%	95%	34%	354	47
k = 20	39%	22%	86%	35%	299	102
k = 50	40%	22%	81%	34%	288	133
Tr Labels					77	323

شکل ۵-۷: جدول دقت‌ها در حالات نمونه‌های متخلف برای مدل جما۲

و همچنین نمودار تغییرات را می‌توان در شکل ۵-۸ مشاهده کرد.

همانطور که از شکل و جدول مشخص است، به نظر می‌آید مدل جما۲، خیلی آرام‌تر و شیب کمتری نسبت به مثال‌های قرار داده شده، نتایج خود را تغییر می‌دهد. این بیانگر آن است که این مدل قدرت یادگیری درون‌متنی و نسبت به محتوا دستور کمی دارد و عملا به مثال‌های تا حد زیادی بی‌توجهی می‌کند.

یکی از مزایای استفاده از تنظیم‌های نمادین این است که حساسیت مدل‌ها به دستور داده شده بررسی شود زیرا که این مدل‌های زبانی بزرگ در حساس‌ترین حالت خود قرار داشته و همچنین در حالت یادگیری چندنمونه‌ای دید که آیا واقعا براساس نمونه‌ها یادگیری صورت می‌گیرد یا نه. که به خصوص در این نتایج می‌توان دید که مدل آیا۲۳ یادگیری نسبتا خوبی از نمونه‌ها داشته در صورتی که مدل جما۲ این رفتار را



شکل ۵-۸: نمودار تغییرات تعداد برچسب‌های پیش‌بینی شده در مدل جما۲

نشان نمی‌دهد.

همچنین این نکته را نیز باید خاطر نشان کرد که در حالت‌های $N(E) = ۵۰$ به خاطر تعداد بسیار زیاد نمونه‌ها، در حالت‌های جزئی رفتارهای متفاوت و عجیبی مشاهده است که دلیل آن می‌تواند کم‌توجهی یا کم‌رنگ‌تر شدن شرح تسک در دستور داده شده به مدل‌ها باشد.

۵-۲ نتایج رویکرد جداسازی دستور سیستمی و کاربر در زبان‌های فارسی و انگلیسی

مورد دیگری که در این پژوهش مورد بررسی قرار گرفته است تاثیر جداسازی دستور سیستمی و کاربر است. همانطور که در شکل ۵-۹ مشخص است در حالتی که دستورالعمل سیستمی فارسی جداگانه‌ای تعریف کرده‌ایم به چنین دقت‌هایی رسیده‌ایم.

```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.93      0.46      0.62       324
           1       0.27      0.86      0.42        77

    accuracy                           0.54       401
   macro avg       0.60      0.66      0.52       401
weighted avg       0.81      0.54      0.58       401

  Number of '1' labels: 241
  Number of '0' labels: 160
```

شکل ۵-۹: دقت‌های به دست آمده در حالت دستور سیستمی به زبان فارسی

اما در خصوص دستور سیستمی به زبان انگلیسی ماجرا کاملا متفاوت است، به طوری که شاهد بهبود ۱۶ درصدی در حوزه F1-Score هستیم. همانطور که در شکل ۵-۱۰ می‌توان این مسئله را با جزییات بیشتری دید.

```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.88      0.89      0.88       324
           1       0.49      0.47      0.48        77

    accuracy                           0.81       401
   macro avg       0.68      0.68      0.68       401
weighted avg       0.80      0.81      0.80       401

  Number of '1' labels: 73
  Number of '0' labels: 328
```

شکل ۵-۱۰: دقت‌های به دست آمده در حالت دستور سیستمی به زبان انگلیسی

این نتایج را می‌توان با زاویه دیدهای مختلفی تحلیل کرد. اما چندین علتی که چنین رویکرد و رفتاری مشاهده شده است را می‌توان به این صورت بیان کرد.

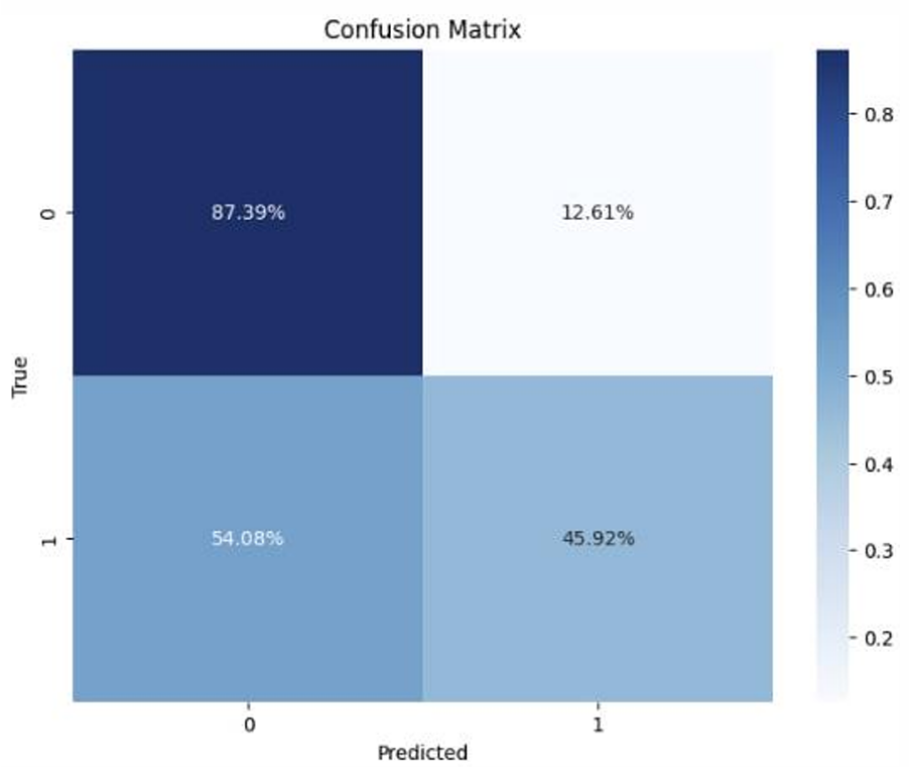
- از آنجایی که بسیاری از این مدل‌های زبانی بزرگ در حالت دستوری[1] آموزش دیده‌اند و در این حین دستور سیتمی و کاربری جدایی داشته‌اند. در زمانی که ما شرح وظیفه و توضیحات مربوط به آن و همچنین نمونه‌ها را جداسازی کنیم و به صورت دستور سیستمی ارائه دهیم، مدل درک بهتری نسبت

[1] Instruct

به آنها خواهد داشت.

- در حالتی که تست‌ها انجام شده، شرح وظیفه و توضیحات به زبان انگلیسی بوده در صورتی که خود نمونه‌ها و خبر اصلی به زبان فارسی بوده است. یکی از دلایلی که شاهد نتایج بهتر در حالت دستور سیستمی انگلیسی هستیم آن است که این نوع مدل‌ها شرح وظیفه را در زبان انگلیسی بهتر متوجه می‌شوند زیرا در این حالت نیز آموزش دیده‌اند اما چون اخبار به زبان فارسی وارد شده‌اند، پیشینه دانش خود در زبان فارسی را نیز لحاظ می‌کنند.
- همچنین به نظر می‌رسد قدرت استدلال و تفکر این مدل‌ها در زبان انگلیسی بیشتر بوده و همچنین دقت و توجه بیشتری نسبت به دستورالعمل ورودی در زبان انگلیسی نسبت به زبان فارسی دارند.

همچنین با بررسی ماتریس درهم‌ریختگی در ۵-۱۱ می‌توان دید این مدل‌ها زمانی به دقت کلی قابل قبول خواهند رسید که تعداد زیادی از اخبار را «غیرمهم» پیش‌بینی کرده و در حوزه اخبار «مهم» نیز دقت قابل‌قبولی داشته باشند. اما با این حال تشخیص خبر «مهم» جدی‌ترین چالش برای این مدل‌ها هست.

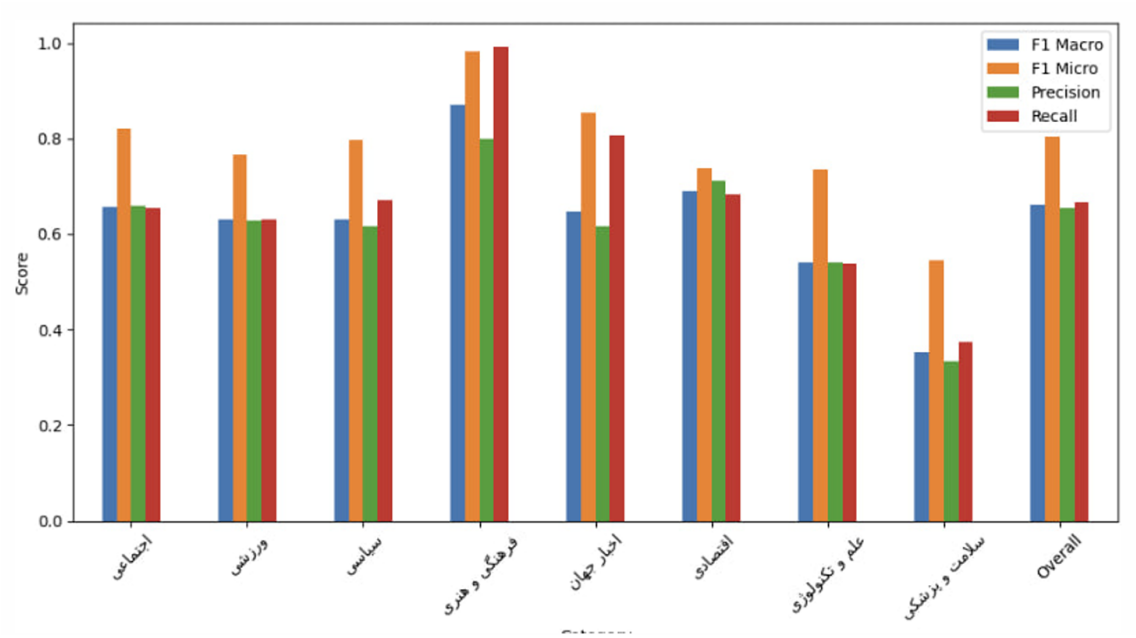


شکل ۵-۱۱: ماتریس درهم‌ریختگی در حالت دستور سیستمی به زبان انگلیسی

۵-۳ نتایج دسته‌بندی‌ها مختلف خبری

یکی دیگر از موردهایی که در تشخیص اهمیت اخبار بسیار حائز اهمیت است، دسته‌بندی آنها و گزارش دقت در هرکدام از دسته‌ها است.

همانطور که در نمودار ۵-۱۲ مشاهده می‌کنید بهترین دقت به دست آمده در دسته‌بندی‌های مختلف نمایان است.



شکل ۵-۱۲: نمودار معیارهای دقت‌سنجی در دسته‌بندی‌های مختلف خبری

با توجه به نمودار می‌توان فهمید که این مدل‌ها در حوزه فرهنگی و هنری دقت بالایی در تشخیص نوع خبر داشته‌اند در صورتی که در حوزه سلامت از خود ضعف نشان‌داده‌اند.

	F1 Macro	F1 Micro	Precision	Recall
اجتماعی	0.6571	0.8213	0.6587	0.6555
ورزشی	0.6302	0.7651	0.6286	0.6319
سیاسی	0.6315	0.7968	0.6164	0.67
فرهنگی و هنری	0.8706	0.9829	0.8	0.9912
اخبار جهان	0.6465	0.8551	0.6162	0.8058
اقتصادی	0.6912	0.7377	0.7109	0.6831
علم و تکنولوژی	0.5402	0.7361	0.5417	0.5391
سلامت و پزشکی	0.3529	0.5455	0.3333	0.375
Overall	0.6605	0.8049	0.6554	0.6665

شکل ۵-۱۳: جدول معیارهای دقت‌سنجی در دسته‌بندی‌های مختلف خبری

جزییات بیشتر دقت‌ها را می‌توان در جدول ۵-۱۳ مشاهده نمود.

۵-۴ نتایج دستورالعمل‌های درخت تفکر و زنجیره‌های تفکر

همانطور که در فصل کارهای پیشین بررسی کرده‌ایم، دستورهای درخت تفکر و زنجیره‌های تفکر دستورهایی هستند که مدل‌های زبانی بزرگ را به تفکر قدم به قدم و بررسی حالت‌های مختلف دعوت می‌کنند.

Title Only	Accuracy	Precision	Recall	F1-Score	F1-Score('1')	F1-Score('0')
Prompt 18	53%	57%	63%	49%	35%	63%
Prompt 19	72%	62%	68%	63%	44%	81%
Prompt 21	46%	57%	61%	44%	34%	54%
Prompt 24	77%	55%	54%	54%	22%	86%
Prompt 25	74%	54%	54%	54%	24%	84%

شکل ۵-۱۴: جدول دقت‌های دستورهای زنجیره تفکر و درخت تفکر

در جدول ۵-۱۴ می‌توان نتایج دستورهای آزمایش شده را دید. تمامی این نتایج در حالت صفر نمونه انجام شده است و همچنین دستور ۱۸ یک دستور وانیلا خام، دستور ۱۹ زجیره تفکر در زبان انگلیسی، دستور ۲۱ درخت تفکر در زبان انگلیسی، دستور ۲۴ زنجیره تفکر در زبان فارسی و در نهایت دستور ۲۵ نمایانگر درخت تفکر در زبان فارسی است. این دستورها، پنج منتخب از میان ۲۵ دستور نوشته شده قبلی براساس دقت‌سنجی‌های انجام شده در تست محدود است.

همانطور که مشخص است این نوع دستورها به صورت کلی عملکرد بهتری داشته‌اند. عملکرد دستور درخت تفکر و زنجیره تفکر در زبان فارسی تا حد زیادی مشابه بوده در صورتی که برای زبان انگلیسی، به نظر میرسد دستور زنجیره تفکر بسیار بهتر از درخت تفکر عمل کرده است.

Title Only	Social	Sports	Political	Cultural	Global	Economic	Scientific	Health
Prompt 18	45%	55%	40%	41%	35%	59%	58%	63%
Prompt 19	56%	61%	49%	49%	41%	62%	58%	80%
Prompt 21	45%	49%	35%	35%	34%	48%	49%	63%
Prompt 24	48%	50%	58%	47%	46%	51%	56%	69%
Prompt 25	54%	51%	54%	60%	43%	52%	46%	69%

شکل ۵-۱۵: جدول دقت‌های دستورهای زنجیره تفکر و درخت تفکر در دسته‌بندی‌های مختلف

در جدول ۵-۱۵ نیز می‌توان دقت هرکدام از این دستورها را در دسته‌بندی‌های مختلف خبری مشاهده نمود. در اینجا دستور ۲۴ معادل فارسی دستور ۱۹ و دستور ۲۵ معادل فارسی دستور ۲۱ بوده است.

۵-۵ نتایج دستورالعمل‌های درخت و زنجیره‌های تفکر در حالت یادگیری چندنمونه‌ای

حال همان دستورهای قبلی که در حالت صفر نمونه بوده‌اند را در حالت چندنمونه‌ای و قرار دادن نمونه‌های مشابه e_i از مجموعه $E^{(K)}$ براساس K انتخابی قرار می‌دهیم. در جدول ۵-۱۶ می‌توان نتایج را برای حالت ۲۰ نمونه مشاهده نمود.

Title Only	Accuracy	Precision	Recall	F1-Score	F1-Score('1')	F1-Score('0')
Prompt 18	66%	62%	71%	60%	43%	76%
Prompt 19	76%	64%	72%	65%	47%	83%
Prompt 21	70%	63%	72%	62%	45%	79%
Prompt 24	82%	68%	70%	69%	49%	89%
Prompt 25	82%	68%	69%	69%	48%	89%

شکل ۵-۱۶: جدول دقت‌های دستورهای زنجیره و درخت تفکر در حالت $K = ۲۰$

که در دسته‌بندی‌های مختلف خبری، تمامی دقت‌های به دست آمده نیز در جدول ۵-۱۷ نمایان است.

Title Only	Social	Sports	Political	Cultural	Global	Economic	Scientific	Health
Prompt 18	66%	57%	48%	61%	48%	64%	58%	71%
Prompt 19	65%	62%	56%	76%	55%	69%	62%	89%
Prompt 21	63%	56%	53%	67%	52%	67%	60%	80%
Prompt 24	66%	68%	66%	68%	68%	70%	57%	42%
Prompt 25	69%	65%	66%	66%	63%	72%	50%	68%

شکل ۵-۱۷: جدول دقت‌های دستورهای زنجیره و درخت تفکر در حالت $K = ۲۰$ در دسته‌بندی‌ها

۵-۶ سیستم انتخاب دستورالعمل وابسته به ورودی

نتایج و کار دیگری که در این پژوهش انجام شده است، معرفی یک سیستم طبقه‌بندی براساس دستور است، به طوری که وابسته به عنوان خبر ورودی می‌تواند بهترین دستور را برای دقت بالاتر انتخاب کند.

در این خصوص باید گفت هرکدام از دستورهای بررسی شده در حالات مختلف و دسته‌بندی‌های مختلف رفتار متفاوتی نشان می‌دهند. این مورد به خصوص در جدول ۵-۱۷ در دسته‌بندی سلامت و بهداشت قابل مشاهده است. به طوری که هرکدام از دستورها رفتار متفاوتی را از خود نشان داده‌اند.

حال با استفاده از یک طبقه‌بندی کننده و با استفاده از نتایج داده‌های آموزش آورده شده در قسمت مطالب تکمیلی، می‌توان یک طبقه‌بندی‌کننده آموزش داد که در هرحالت بتواند بهترین دستور از پنج دستور

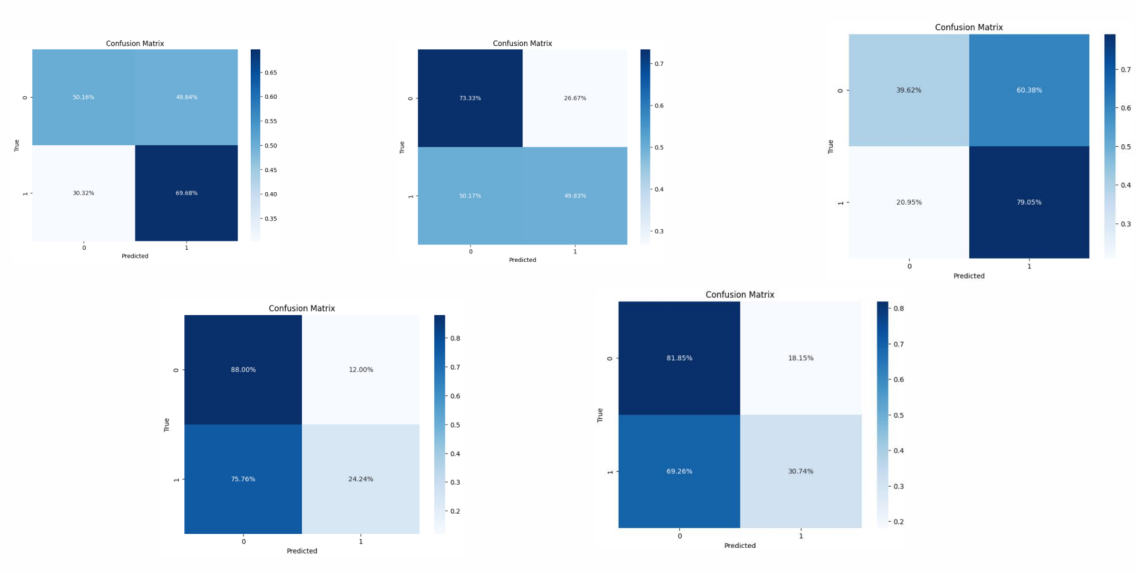
بررسی شده، یعنی دستورهای ۱۸، ۱۹، ۲۱، ۲۴ و ۲۵ را انتخاب کند.

با به کارگیری این طبقه‌بند نتایج جدول ۵-۱۸ حاصل می‌شود.

Title Only	Accuracy	Precision	Recall	F1-Score	F1-Score('1')	F1-Score('0')
Q-Dependent	86%	75%	72%	73%	55%	92%

شکل ۵-۱۸: جدول دقت تشخیص اهمیت خبر در حالت استفاده از طبقه‌بند برای انتخاب دستور

که همانگونه که مشخص است دقتی فراتر از تمام حالت‌های قبلی به دست آورده است و همچنین این دقت در حالت صفر نمونه بوده است.



شکل ۵-۱۹: ماتریس درهم‌ریختگی پنج دستور منتخب

یکی از دلایل موفقیت این سیستم توسعه داده شده، استفاده از رفتارهای متخلف دستورها در حالات گوناگون است. همانطور که در شکل ۵-۱۹ مشخص است، ماتریس درهم‌ریختگی هرکدام از این دستورها رفتار متفاوتی نشان‌داده و طبقه‌بند می‌تواند براساس ورودی تشخیص دهد که چه دستوری برای این شرایط بهتر خواهد شد.

این طبقه‌بند با انتخاب درست دستور توانسته ۱۰ درصد نسبت به بهترین دستور در حالت صفرنمونه و ۴ درصد در حالت چندنمونه مقدار F1-Score را بهبود ببخشد.

اما نکته حائز اهمیت در این سیستم آن است که دستورهای منتخب رفتارهای متفاوت و ماتریس‌های درهم‌ریختگی گوناگونی نسبت به پیش‌بینی برچسب‌های «مهم» و «غیرمهم» داشته باشد، زیراکه اگر تمامی دستورها رفتار یکسانی نشان دهند طبقه‌بندی نمی‌تواند از گستردگی فضا استفاده کند و در نهایت بهبودی نیز حاصل نمی‌شود.

فصل ۶

نتیجه‌گیری

در این پژوهش به بررسی ابعاد مختلف استفاده از مدل‌های زبانی بزرگ در تشخیص اهمیت اخبار پرداخت شد. از بررسی جابه‌جایی برچسب‌ها با مقدار نمادین آنها تا معرفی سیستم طبقه‌بند انتخاب دستورالعمل براساس ورودی، همه و همه بیانگر این است که زاویه‌دیدهای مختلف نسبت به این مسئله باعث آشکار شدن ابعاد کشف‌نشده از مدل‌های زبانی بزرگ و رسیدن به دقت‌های بالاتر در درستی تشخیص خبر مهم از غیرمهم می‌شود.

همچنین این نکته را باید مدنظر گرفت که وظیفه تشخیص اهمیت یک خبر، بسیار کار پیچیده و مبهمی است زیرا که برای فرهنگ‌های مختلف، سلایق مختلف، شخصیت‌های مختلف و حتی دسته‌بندی‌های مختلف اهمیت یک خبر دچار دگرگونی است. در این مسئله حتی بحث زمان هم مطرح است یعنی یک خبر در یک حوزه زمانی می‌تواند مهم باشد در صورتی که برحه دیگه اهمیت خود را از دست بدهد.

در طول این پژوهش سعی شد با استفاده از دادگان برچسب‌گذاری جمع‌آوری‌شده از حدود ۱۱ هزار خبر تمامی نتایج گرفته‌شده و بررسی شود و مشخصا در یک زاویه دید دیگر و در دادگان دیگر ممکن است تعریف اهمیت خبر متفاوت باشد.

مسیری آینده این پژوهش می‌تواند به جست‌وجوی عمیق‌تر در خصوص سیستم‌های انتخاب دستعورالعمل وابسته به ورودی کاربر بیانجامد خصوصا در حالت چندنمونه‌ای، به طوری که بالاترین دقت در این پژوهش از طریق این سیستم‌ها در حالت صفرنمونه گرفته شد و به نظر می‌رسد که این نوع سیستم‌ها می‌تواند در حالت‌های دیگر و به خصوص در حالت یادگیری چندنمونه‌ای به دقت‌های بالاتر دست یابد.

در نهایت این پروژه سعی کرده بسیاری از ابعاد استفاده از یک مدل زبانی بزرگ به عنوان یک تشخیص دهنده را بررسی کند و نتایج آن را اعلام کند به طوری که مسیر را برای پژوهش‌های آینده در حوزه استفاده

از مدل‌های زبانی بزرگ در تشخیص «مهم» یا «غیرمهم» بودن یک خبر را هموارتر سازد.

Bibliography

[1] J. Wei, L. Hou, A. Lampinen, X. Chen, D. Huang, Y. Tay, X. Chen, Y. Lu, D. Zhou, T. Ma, and Q. V. Le. Symbol tuning improves in-context learning in language models, 2023.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[4] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners, 2022.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022.

[6] S. Gao, A. Sethi, S. Agarwal, T. Chung, and D. Hakkani-Tur. Dialog state tracking: A neural reading comprehension approach, 2019.

[7] L. Reynolds and K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm, 2021.

[8] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[9] M. Felicilda, A. Geriane, V. Agustin, M. C. Blanco, J. Morano, L. Mahusay, and J. Guialil. Enhancement of random forest algorithm applied in fake news detection. *World Journal of Advanced Research and Reviews*, 22:1075–1079, 05 2024.

[10] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, page 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[12] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, and P. Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113, Mar. 2024.

[13] J. Wang, Z. Zhu, C. Liu, R. Li, and X. Wu. Llm-enhanced multimodal detection of fake news. *PLOS ONE*, 19, 10 2024.

[14] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. Bers, S. Biderman, L. Gao, T. Wolf, and A. M. Rush. Multitask prompted training enables zero-shot task generalization, 2022.

[15] P. Zheng, H. Chen, S. Hu, B. Zhu, J. Hu, C.-S. Lin, X. Wu, S. Lyu, G. Huang, and X. Wang. Few-shot learning for misinformation detection based on contrastive models. *Electronics*, 13(4), 2024.

[16] M. Heydari, M. Khazeni, and M. A. Soltanshahi. Deep learning-based sentiment analysis in persian language. In *2021 7th International Conference on Web Research (ICWR)*, page 287–291. IEEE, May 2021.

[17] H. H. Hemati, A. Lagzian, M. S. Sartakhti, H. Beigy, and E. Asgari. Khabarchin: Automatic detection of important news in the persian language, 2023.

[18] L. Reynolds and K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm, 2021.

[19] H. Sun, A. Hüyük, and M. van der Schaar. Query-dependent prompt evaluation and optimization with offline inverse rl, 2024.

[20] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022.

واژه‌نامه

ت

تنظیم براساس دستورالعمل ... Insturction Tuning

تنظیم نمادین ... Symbol Tuning

ج

جنگل‌های تصادفی ... Random Forest Tree

د

دستور ... Prompt

درخت تفکر ... Tree-of-Thoughts

درخواست سامانه ... System Prompt

دستوری ... Instruct

ز

زنجیره تفکر ... Chain-of-Thoughts

ص

صدق‌پذیری ... Satisfiability

صفر نمونه ... Zero Shot

ق

قطعی ... Deterministic

م

مدل‌های زبانی بزرگ ... Large Language Models

مهندسی درخواست ... Prompt Engineering

ماشین بردار پشتیبان ... SVM

ماتریس درهم‌ریختگی ... Confusion Matrix

و

دستورالعمل خام ... Vanila Prompt

دستعورالعمل کاربر ... User Prompt

ی

یادگیری چند نمونه ... Few-Shot Learning

یادگیری درون‌متنی ... In-Context Learning

یادگیری ماشین ... Machine Learning

یادگیری عمیق ... Deep Learning

یادگیری تقویتی ... Reinforcement Learning

پیوست آ

مطالب تکمیلی

در اینجا تمام محتوای تکمیلی پژوهش از جمله دستورالعمل‌های استفاده شده در مراحل مخلتف برای مدل‌های زبانی بزرگ و نتایج دسته‌بندی شده قرار گرفته شده است.

آـ ۱ دستورالعمل‌های به کار گرفته شده

در این بخش انواع دستورهای نوشته شده در این پژوهش آمده و شرح داده می‌شود.

آـ ۱-۱ دستورهای وانیلا یا خام

ابتدایی‌ترین دستورهای نوشته شده برای مدل‌های زبانی بزرگ به طوری که شامل تعریف اخبار مهم و شرح وظیفه است.

هدف، داشتن یک دسته‌بند دودویی است که با گرفتن هر متن ورودی، کلاس آن را در خروجی مشخص می‌کند. کلاس‌ها شامل دو دسته‌ی غیرمهم و مهم هستند. یعنی خبر نوع غیرمهم و خبر نوع مهم.

شرح تسک:

متن یا خبری را مهم می‌گوییم اگر که برای بیش‌تر کاربران فارسی‌زبان اهمیت بالایی داشته باشد. یا به عبارت دیگر، جمعیت زیاد و بزرگی از ایرانیان مایل باشند که آن متن یا خبر را بخوانند و یا برای یکدیگر بفرستند.

اگر خبری مربوط به یک قشر کوچک یا جامعه‌ی خاصی از کاربران باشد و یا ارزش خواندن کمی داشته باشد و یا خاص نباشد، آن خبر از نوع غیرمهم است.

در صورتی که متن ورودی از نوع مهم باشد، کلاس مهم خواهد بود و در صورتی که غیرمهم باشد، کلاس غیرمهم خواهد بود.

برخی از مفاهیم از نوع مهم عبارت‌اند از: یارانه و سهام و مواردی که قرار است پول به مردم برسد مهم هستند ثبت نام مسکن و خانه و اخبار مربوط به وام‌ها و... ثبت نام خودرو افزایش و کاهش های شدید و زیاد قیمت ارز یا طلا و سکه و یا تورم

سیاسی: اخبار جنگ، برجام، توافق های ایران، تحریم های ایران، خبرهای جنگ‌های بزرگ منطقه‌ای، عزل و نصب مقامات بلندپایه ایرانی، این‌ها همگی مهم هستند

ورزشی: اخبار مربوط به تیم‌های معروف و پرطرفدار ایرانی و همین‌طور اروپایی مهم است

تمام اخبار بالا از نوع مهم بوده و اخبار دسته‌های دیگر که کمتر خواننده دارند را از نوع غیرمهم در نظر می‌گیریم.

با توجه به متن زیر تنها در یک عدد پاسخ بده که باتوجه به مفاهیمی که در بالا مطرح شد و قدرت استنتاجی که خودت داری، آیا متن مهم حساب می‌شود یا غیرمهم. (مهم یا غیرمهم):

در خروجی فقط مجاز هستی مهم یا عدد غیرمهم بنویسی. بدون هیچ توضیح اضافه‌ای.

و نسخه انگلیسی آن به این صورت نوشته و مورد بررسی قرار گرفته است:

The goal is to have a binary classifier that, by receiving any input text, determines its class in the output. The classes include two categories: 'not important' and 'important', meaning news type 'not important' and news type 'important'.

Task description:

We label a text or news as 'important' if it is of high importance to most Persian-speaking users. In other words, if a large population of Iranians are likely to read, share, or be interested in it, it is classified as 'important'.

If the news pertains to a small group or a specific community of users, has little reading value, or is not significant, it is classified as 'not important'.

If the input text is of type 'important', the output class will be 'important'; if it is 'not important', the output class will be 'not important'.

Some concepts that fall under type 'important' are: Subsidies, stocks, and matters that involve receiving money are important. Housing and home registrations, news related to loans, etc. Car registrations Significant fluctuations in currency, gold, coins, or inflation rates

Politics: News about war, the JCPOA, Iran's agreements, Sanctions on Iran, News of major regional wars, Dismissal and appointment of high-ranking Iranian officials, These are all important.

Sports: News about famous and popular Iranian teams as well as European teams is important.

All the above news are classified as type 'important', and other news categories that have fewer readers are considered as type 'not important'.

A text or news is classified as 'not important' if it pertains to a specific small section of the society. News that does not engage a broad spectrum of the community is type 'not important'. For example: Sports: News about non-famous clubs and small events are of type 'not important'. Politics: News about non-prominent figures that do not affect the Iranian society is of type 'not important'. Social: News that does not engage a large section of society is type 'not important'.

Based on the following text, respond with only a single label that, considering the concepts discussed above and your own inferential ability, indicates whether the text should be classified as 'important' or 'not important'. ('important' or 'not important'):

You are only allowed to write the label 'important' or 'not important' in the output, without any additional explanation.

که شامل تعاریف اخبار مهم برای دسته‌های مختلف خبری به صورت اضافه‌تر است.

آ ـ ۱ - ۲ دستورهای نمادین

در اینجا در دستور نوشته شده، هیچ اسمی از «مهم» و «غیرمهم» بودن و تعاریف آنها برده نشده و صرفا از برچسب‌های نمادین «۵۸» و «۴۷» استفاده شده است.

هدف، داشتن یک دسته‌بند دودویی است که با گرفتن هر متن ورودی، کلاس آن را در خروجی مشخص می‌کند. کلاس‌ها شامل دو دسته‌ی ۴۷ و ۵۸ هستند. یعنی خبر نوع ۴۷ و خبر نوع ۵۸.

شرح تسک:

متن یا خبری را ۵۸ می‌گوییم اگر که برای بیش‌تر کاربران فارسی‌زبان اهمیت بالایی داشته باشد. یا به عبارت دیگر، جمعیت زیاد و بزرگی از ایرانیان مایل باشند که آن متن یا خبر را بخوانند و یا برای یکدیگر بفرستند.

اگر خبری مربوط به یک قشر کوچک یا جامعه‌ی خاصی از کاربران باشد و یا ارزش خواندن کمی داشته باشد و یا خاص نباشد، آن خبر از نوع ۴۷ است.

در صورتی که متن ورودی از نوع ۵۸ باشد، کلاس ۵۸ خواهد بود و در صورتی که ۴۷ باشد، کلاس ۴۷ خواهد بود.

برخی از مفاهیم از نوع ۵۸ عبارت‌اند از: یارانه و سهام و مواردی که قرار است پول به مردم برسد مهم هستند ثبت نام مسکن و خانه و اخبار مربوط به وام‌ها و... ثبت نام خودرو افزایش و کاهش های شدید و زیاد قیمت ارز یا طلا و سکه و یا تورم

سیاسی: اخبار جنگ، برجام، توافق های ایران، تحریم های ایران، خبرهای جنگ‌های بزرگ منطقه‌ای، عزل و نصب مقامات بلندپایه ایرانی، این‌ها همگی مهم هستند

ورزشی: اخبار مربوط به تیم‌های معروف و پرطرفدار ایرانی و همین‌طور اروپایی مهم است

تمام اخبار بالا از نوع ۵۸ بوده و اخبار دسته‌های دیگر که کمتر خواننده دارند را از نوع ۴۷ در نظر می‌گیریم.

با توجه به متن زیر تنها در یک عدد پاسخ بده که باتوجه به مفاهیمی که در بالا مطرح شد و قدرت استنتاجی که خودت داری، آیا متن ۵۸ حساب می‌شود یا ۴۷ . (۵۸ یا ۴۷):

در خروجی فقط مجاز هستی عدد ۵۸ یا عدد ۴۷ بنویسی. بدون هیچ توضیح اضافه‌ای.

در نتایج به دست آمده، متوجه شدیم که در این دستور صرفا تعریف اخبار «مهم» یا همان «۵۸» آمده است و برای همین مدل در رویکر بدون نمونه صرفا تمامی اخبار را مهم پیش‌بینی می‌کند. برای حل این مشکل دستور زیر با اضافه شدن تعاریف اخبار غیرمهم یا همان «۴۷» نوشته و به کار گرفته شد.

هدف، داشتن یک دسته‌بند دودویی است که با گرفتن هر متن ورودی، کلاس آن را در خروجی مشخص می‌کند. کلاس‌ها شامل دو دسته‌ی ۴۷ و ۵۸ هستند. یعنی خبر نوع ۴۷ و خبر نوع ۵۸.

شرح تسک:

متن یا خبری را ۵۸ می‌گوییم اگر که برای بیش‌تر کاربران فارسی‌زبان اهمیت بالایی داشته باشد. یا به عبارت دیگر، جمعیت زیاد و بزرگی از ایرانیان مایل باشند که آن متن یا خبر را بخوانند و یا برای یکدیگر بفرستند.

اگر خبری مربوط به یک قشر کوچک یا جامعه‌ی خاصی از کاربران باشد و یا ارزش خواندن کمی داشته باشد و یا خاص نباشد، آن خبر از نوع ۴۷ است.

در صورتی که متن ورودی از نوع ۵۸ باشد، کلاس ۵۸ خواهد بود و در صورتی که ۴۷ باشد، کلاس ۴۷ خواهد بود.

برخی از مفاهیم از نوع ۵۸ عبارت‌اند از: یارانه و سهام و مواردی که قرار است پول به مردم برسد مهم هستند ثبت نام مسکن و خانه و اخبار مربوط به وام‌ها و... ثبت نام خودرو افزایش و کاهش های شدید و زیاد قیمت ارز یا طلا و سکه و یا تورم

سیاسی: اخبار جنگ، برجام، توافق های ایران، تحریم های ایران، خبرهای جنگ‌های بزرگ منطقه‌ای، عزل و نصب مقامات بلندپایه ایرانی، این‌ها همگی مهم هستند

ورزشی: اخبار مربوط به تیم‌های معروف و پرطرفدار ایرانی و همین‌طور اروپایی مهم است

تمام اخبار بالا از نوع ۵۸ بوده و اخبار دسته‌های دیگر که کمتر خواننده دارند را از نوع ۴۷ در نظر می‌گیریم.

متن یا خبری را ۴۷ می‌گویند که مربوط به بخش خاص و کوچکی از جامعه باشد. اخباری که گستره‌ی وسیعی از جامعه را درگیر نکند، اخبار از نوع ۴۷ هستند. برای نمونه: ورزشی: اخبار مربوط به باشگاه‌های غیر معروف و رخدادهای کوچک از نوع ۴۷ هستند. سیاسی: اخبار مربوط به شخصیت‌های غیرمشهور که تاثیری روی جامعه‌ی ایران ندارد از نوع ۴۷ هستند. اجتماعی: اخباری که گستره‌ی وسیعی از جامعه را درگیر نمی‌کند از نوع ۴۷ هستند.

با توجه به متن زیر تنها در یک عدد پاسخ بده که باتوجه به مفاهیمی که در بالا مطرح شد و قدرت استنتاجی که خودت داری، آیا متن ۵۸ حساب می‌شود یا ۴۷ . (۵۸ یا ۴۷):

در خروجی فقط مجاز هستی عدد ۵۸ یا عدد ۴۷ بنویسی. بدون هیچ توضیح اضافه‌ای.

آ ـ ۱ - ۳ دستورها با رویکرد یادگیری چند نمونه‌ای

در این نوع پرامپت‌ها، چندین نمونه و مثال برای یادگیری و شباهت‌سنجی در اختیار مدل زبانی بزرگ قرار می‌گیرد که یک نمونه از این نوع دستورها در اینجا قرار داده شده است.

هدف، داشتن یک دسته‌بند دودویی است که با گرفتن هر متن ورودی، کلاس آن را در خروجی مشخص می‌کند. کلاس‌ها شامل دو دسته‌ی ۱ یا ۰ هستند. ۱ یعنی خبر مهم است و ۰ یعنی خبر مهم نیست.

شرح تسک:

متن یا خبری را مهم یا تاثیرگذار می‌گوییم اگر که برای بیش‌تر کاربران فارسی‌زبان اهمیت بالایی داشته باشد. یا به عبارت دیگر، جمعیت زیاد و بزرگی از ایرانیان مایل باشند که آن متن یا خبر را بخوانند و یا برای یکدیگر بفرستند. اگر خبری مربوط به یک قشر کوچک یا جامعه‌ی خاصی از کاربران باشد، آن خبر مهم نیست. در صورتی که متن ورودی مهم باشد، کلاس ۱ خواهد بود و در صورتی که مهم نباشد، کلاس ۰ خواهد بود

برخی از مفاهیم مهم و از کلاس ۱ عبارت‌اند از: یارانه و سهام و مواردی که قرار است پول به مردم برسد مهم هستند ثبت نام مسکن و خانه و اخبار مربوط به وام‌ها و... ثبت نام خودرو افزایش و کاهش های شدید و زیاد قیمت ارز یا طلا و سکه و یا تورم

سیاسی: اخبار جنگ، برجام، توافق های ایران، تحریم های ایران، خبرهای جنگ‌های بزرگ منطقه‌ای، عزل و نصب مقامات بلندپایه ایرانی، این‌ها همگی مهم هستند

ورزشی: اخبار مربوط به تیم‌های معروف و پرطرفدار ایرانی و همین‌طور اروپایی مهم است

نمونه‌ها: چند نمونه پایین را ببین و باتوجه به آن‌ها به سوال پایین پاسخ بده

SAMPLES

از روی نمونه‌های بالایی یاد بگیر و خروجی را مشخص کن (فقط ۰ یا ۱). حال با توجه به «نمونه‌های بالا»، برای متن زیر تنها در یک واژه پاسخ بده که باتوجه به مفاهیمی که در بالا مطرح شد و قدرت استنتاجی که خودت داری، آیا متن مهم (تاثیرگذاری) حساب می‌شود یا خیر. (۱ یا ۰):

در خروجی فقط مجاز هستی عدد ۱ یا عدد ۰ بنویسی. بدون هیچ توضیح اضافه‌ای.

آ ـ ۱ - ۴ دستورهای مخصوص دسته‌های متخلف خبری

یک نگاه به تحلیل موضوع اهمیت اخبار این است که هردسته برای خود می‌بایست به صورت مستقل ارزیابی شود. با این نگاه پرامپت‌های مختلفی برای دسته‌های بندی نوشته شده است که در ادامه برای نمونه، چهار

دسته از این دستورها قرار داده شده است.

دستور مخصوص دسته‌بندی ورزشی از اخبار:

The goal is to have a binary classifier that, by receiving any input text, determines its class in the output. The classes include two categories: 'not important' and 'important', meaning sports news type 0 and sports news type 1.

Task description:

We label a sports news text as 1 if it is of high importance to most Persian-speaking users. In other words, if a large population of Iranians are likely to read, share, or be interested in it, it is classified as 1.

If the sports news pertains to a small group or a specific community of users, has little reading value, or is not significant, it is classified as 0.

If the input text is of type 1, the output class will be 1; if it is 0, the output class will be 0.

Some concepts that fall under sports news type 1 are: Matches, transfers, or achievements involving famous and popular Iranian football teams, such as Persepolis, Esteghlal, and Sepahan. News related to Iranian athletes who are internationally recognized or have significant achievements in global competitions, such as the Olympics or World Championships. Major events in European football, particularly those involving teams like Barcelona, Real Madrid, Manchester United, etc., which have a large following in Iran. News regarding Iranian athletes in sports that hold national pride, such as wrestling, weightlifting, or volleyball. Updates on Iran's national teams in any sport, particularly during significant tournaments like the World Cup, Asian Games, or Olympic qualifiers.

Samples: Look at the following examples and, based on them, answer the question below.

SAMPLES HERE

Learn from the above examples and determine the output (only 1 or 0).

Now, based on the "above examples," respond with only a single word that, considering the concepts discussed above and your own inferential ability, indicates whether the following text should be classified as 1 or 0. (1 or 0):

You are only allowed to write the label 1 or 0 in the output, without any additional explanation.

برای دسته‌بندی اجتماعی نیز داریم:

Task description:

We label a social news text as '1' if it is of high importance to most Persian-speaking users. In other words, if a large population of Iranians are likely to read, share, or be interested in it, it is classified as '1'.

If the social news pertains to a small group or a specific community of users, has little reading value, or is not significant, it is classified as '0'.

If the input text is of type '1', the output class will be '1'; if it is '0', the output class will be '0'.

Some concepts that fall under social news type '1' are: News about significant social movements or protests within Iran. Changes in laws or regulations that impact daily life, such as those related to education, employment, or public services. News involving prominent Iranian social figures, celebrities, or influential personalities who are widely recognized. Social issues like poverty, inequality, or social justice matters that resonate widely within the Iranian society.

Samples: Look at the following examples and, based on them, answer the question below.

SAMPLES HERE

Learn from the above examples and determine the output (only '1' or '0').

Now, based on the "above examples," respond with only a single word that, considering the concepts discussed above and your own inferential ability, indicates whether the following text should be classified as '1' or '0'. ('1' or '0'):

You are only allowed to write the label '1' or '0' in the output, without any additional explanation.

همچنین برای دسته‌بندی سیاسی داریم:

The goal is to have a binary classifier that, by receiving any input text, determines its class in the output. The classes include two categories: 'not important' and 'important', meaning political news type '0' and political news type '1'.

Task description:

We label a political news text as '1' if it is of high importance to most Persian-speaking users. In other words, if a large population of Iranians are likely to read, share, or be interested in it, it is classified as '1'.

If the political news pertains to a small group or a specific community of users, has little reading value, or is not significant, it is classified as '0'.

If the input text is of type '1', the output class will be '1'; if it is '0', the output class will be '0'.

Some concepts that fall under political news type '1' are: News about significant international agreements or treaties involving Iran, such as the JCPOA (Joint Comprehensive Plan of Action). Updates on sanctions imposed on or lifted from Iran by other countries or international organizations. Coverage of major regional or global conflicts, particularly those involving Iran or affecting its geopolitical standing. Elections, both domestic and international, that have a substantial impact on Iran's political landscape. Legislative changes or government decisions that affect the broader population, such as those related to civil liberties, national security, or economic policies. Coverage of protests or significant political movements within Iran that resonate with a large segment of the population.

Samples: Look at the following examples and, based on them, answer the question below.

Learn from the above examples and determine the output (only '1' or '0').

Now, based on the "above examples," respond with only a single word that, considering the concepts discussed above and your own inferential ability, indicates whether the following text should be classified as '1' or '0'. ('1' or '0'):

و در نهایت برای دسته‌بندی علمی خواهیم داشت:

The goal is to have a binary classifier that, by receiving any input text, determines its class in the output. The classes include two categories: 'not important' and 'important', meaning science and technology news type '0' and science and technology news type '1'.

Task description:

We label a science and technology news text as '1' if it is of high importance to most Persian-speaking users. In other words, if a large population of Iranians are likely to read, share, or be interested in it, it is classified as '1'.

If the science and technology news pertains to a small group or a specific community of users, has little reading value, or is not significant, it is classified as '0'.

If the input text is of type '1', the output class will be '1'; if it is '0', the output class will be '0'.

Some concepts that fall under science and technology news type '1' are: Major advancements or discoveries in science that have global significance or specific implications for Iran, such as breakthroughs in medicine, physics, or environmental science.

News about technological innovations, particularly those developed in Iran or by Iranian scientists, that have the potential to impact industries or society at large. Significant developments in information technology, cybersecurity, and artificial intelligence that are relevant to Iranian interests or that could influence the global technology landscape. Reports on major scientific conferences or events where Iranian scientists or technologists are recognized or play a significant role. Developments in the tech industry, particularly regarding companies or startups that are driving innovation in Iran or globally influential tech giants that have a major impact on the Iranian market.

Samples: Look at the following examples and, based on them, answer the question below.

You are only allowed to write the label '1' or '0' in the output, without any additional explanation.

آ-۱-۵ دستورالعمل‌های زنجیره‌های تفکر

این نوع دستورها مدل را به بررسی گام‌به‌گام موضوع و تفکر در خصوص آن دعوت می‌کند. در اینجا نمونه‌ای نوشته شده و به کار برده شده این دستور را مشاهده می‌کنید.

هدف این است که اخبار را به دو دسته «مهم» (۱) و «غیر مهم» (۰) طبقه‌بندی کنیم. برای طبقه‌بندی دقیق، مراحل زیر را دنبال کنید:

۱. بررسی کنید که آیا موضوع خبر می‌تواند برای جمعیت زیادی از فارسی‌زبانان مرتبط باشد یا خیر.

۲. ارزیابی کنید که آیا خبر به رویدادهای مهم اقتصادی (تغییرات ارز یا تورم، به‌روزرسانی‌های مسکن و غیره)، رویدادهای سیاسی حیاتی (اقدامات دولت، روابط بین‌الملل) یا موضوعات اجتماعی تاثیرگذار که ممکن است افراد زیادی را تحت تاثیر قرار دهد، مربوط می‌شود.

۳. در نهایت، مشخص کنید که آیا خبر جذابیت عمومی دارد یا تنها برای مخاطبان خاصی قابل توجه است.

اگر خبر به موضوعات ذکر شده مربوط باشد و علاقه‌مندی گسترده‌ای را به خود جلب کند، آن را با «۱» برچسب بزنید. در غیر این صورت، «۰» را انتخاب کنید. پس از بررسی این سناریوها، طبقه‌بندی نهایی را در قالب زیر ارائه کنید:

طبقه بندی نهایی: «۰ یا ۱»

آ-۱-۶ دستورالعمل‌های درخت تفکر

این نوع دستورها کاربرد این را داشته که در مرحله‌ای به صورت درختی و جست‌وجوی سطحی یک موضوعی را بررسی و سپس نتیجه را اعلام کند.

For this classification task, follow these branches of thought to decide if the news item is 'important' (1) or 'not important' (0):

1. First, examine if the topic could affect a large portion of Persian-speaking users, considering its potential reach.

2. Next, break down the topic's content into economic, political, and social relevance. - For economic news: Consider factors like inflation, housing, and stock trends relevant to everyday users. - For political news: Evaluate if the content relates to Iran's major policies, high-profile government changes, or global interactions. - For social relevance: Check if it covers popular sports or events with a broad appeal.

3. Assess if the story's appeal is universal or niche.

Samples: Look at the following examples and, based on them, understand which news is considered important or '1' and which ones are considered not important or '0'.

SAMPLES

Learn from the above examples and determine the output.

Assign '1' if the story is broadly significant or '0' if it appeals mainly to a niche audience. Once you've reviewed these scenarios, provide the final classification formatted as follows:

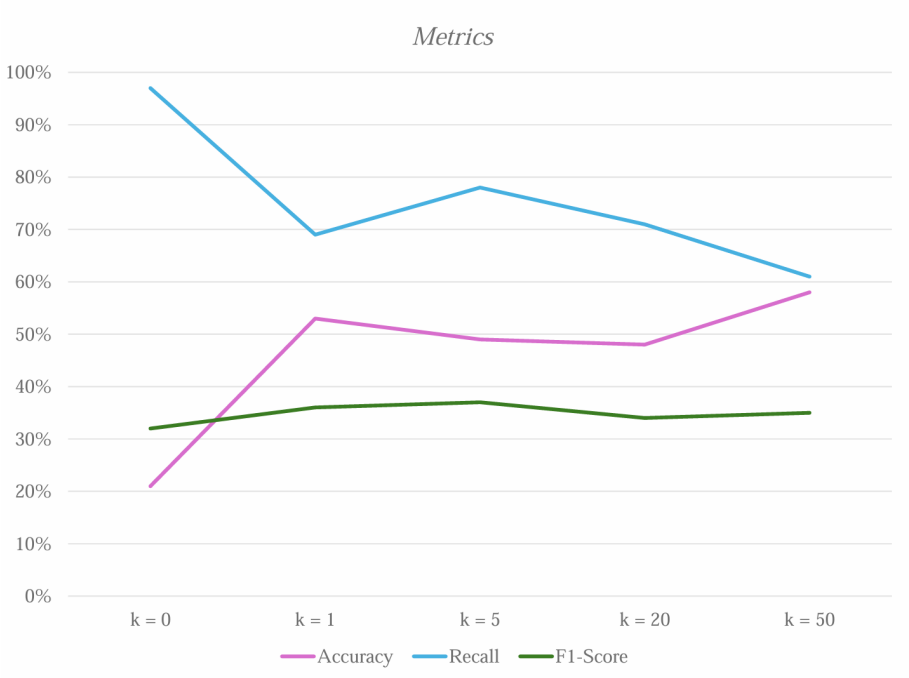
Final Classification: [1 or 0]

آ_۲ نتایج اضافی‌تر

در این بخش به نتایج اضافه‌تر در دسته‌های گوناگون خبری از جمله سیاسی، اقتصادی، وزشی و فرهنگی اشاره شده است.

آ_۲-۱ نتایج بخش دستور نمادین

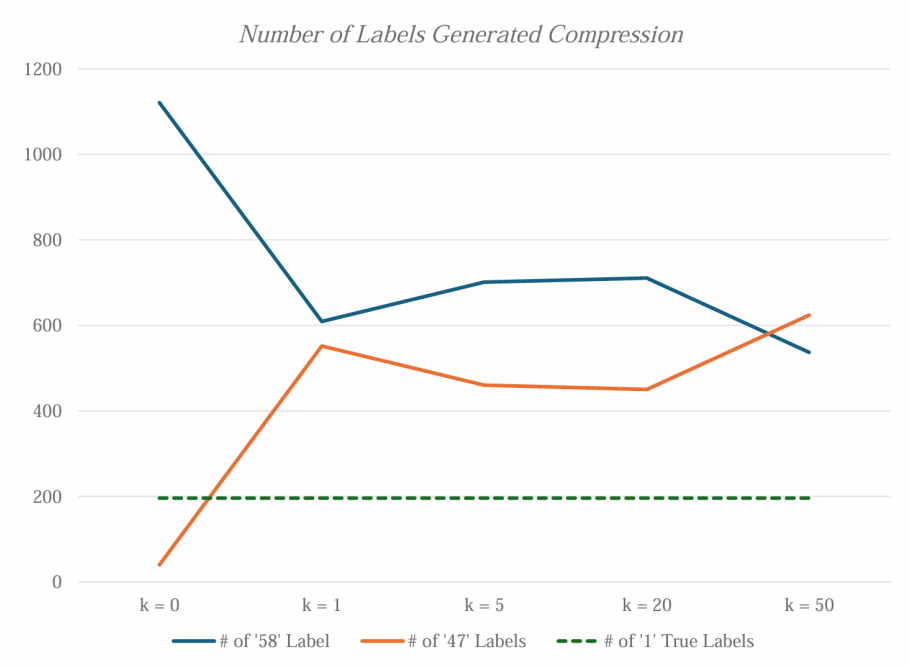
در اینجا به علاوه نمودار آورده شده در فصل نتایج جدید، می‌توان رفتارهای Recall ، Accuracy و F1-Score را در شکل آ_۱ مشاهده نمود.



شکل آ_۱: مقایسه معیارهای دقت‌سنجی در Kهای مختلف

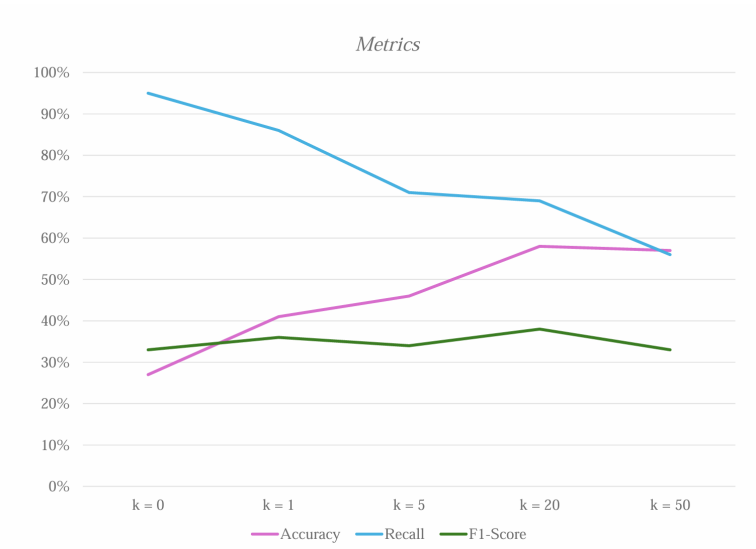
همچنین در شکل آ_۲ می‌توان دید که همان رفتار کلی در کل دادگان تست نسبت به ۱۰۰ داده‌ای

ابتدایی آن تکرار شده است.



شکل آ_۲: مقایسه تعداد برچسب‌های تولید شده در کل دادگان تست

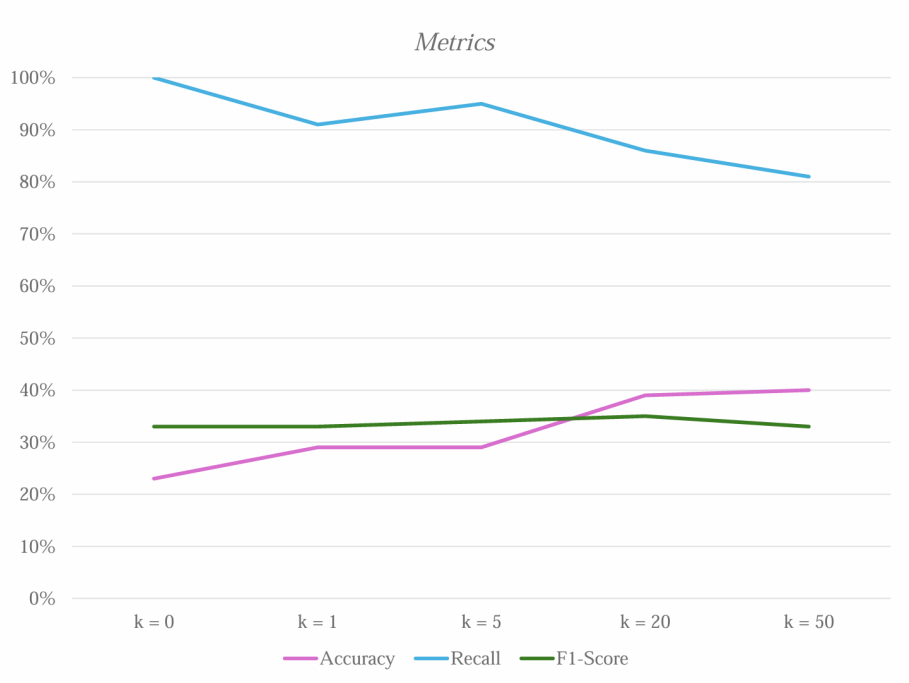
و در شکل آ_۳ می‌توان دید که معیارهای دقت‌سنجی رفتارهای اندکی متفاوت در کل دادگان تست نشان داده است.



شکل آ_۳: مقایسه معیارهای دقت‌سنجی در Kهای مختلف بر روی کل دادگان تست

آ-۲-۲ نتایج دستور نمادین مدل جما۲

علاوه بر نمودار و جدول نمایان شده در فصل نتایج جدید، یک قیاس از رفتار معیارها از مدل جما۲ در طول تعداد نمونه‌های مختلف را می‌توان در نمودار آ-۴ مشاهده نمود.



شکل آ-۴: مقایسه معیارهای دقت‌سنجی در Kهای مختلف در مدل جما۲

آ-۲-۳ نتایج دستور سیستمی انگلیسی

افزون بر نتایج ارائه شده در فصل نتایج جدید، یک حالت دیگری که تست شده است جابه‌جایی برچسب‌های «مهم» و «غیرمهم» با اعداد «۰» و «۱» هست. زمانی که این تغییرات انجام می‌شود دقت‌ها مدل به صورت آ-۵ در خواهد آمد.

همانگونه که مشخص است تغییرات خیلی اندکی نسبت به حالت برچسب‌های مفهومی داریم و این به آن معناست که در فضای دانش مدل‌های زبانی بزرگ، برچسب‌های «۱» و «۰» همان مفاهیم «مهم» و «غیرمهم» را نشان‌داده و در سطح توکن خود همان مفاهیم را نشان می‌دهند.

آ-۲-۴ نتایج دستور سیستمی انگلیسی در حالت کل متن خبری

یکی دیگر از پژوهش‌های که انجام شد در خصوص بررسی کل متن خبری بود. در حالت عادی همواره صرفا عنوان خبر داده می‌شد که برچسب اهمیت آن پیش‌بینی شود. اما در حالتی که کل متن خبر داده شود

```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.87      0.89      0.88       324
           1       0.48      0.43      0.45        77

    accuracy                           0.80       401
   macro avg       0.67      0.66      0.67       401
weighted avg       0.79      0.80      0.80       401

  Number of '1' labels: 69
  Number of '0' labels: 332
```

شکل آ_۵: دقت‌های به دست آمده در حالت دستور سیستمی به زبان انگلیسی با برچسب‌های عددی

به دقت‌های شکل آ_۶ می‌رسیم.

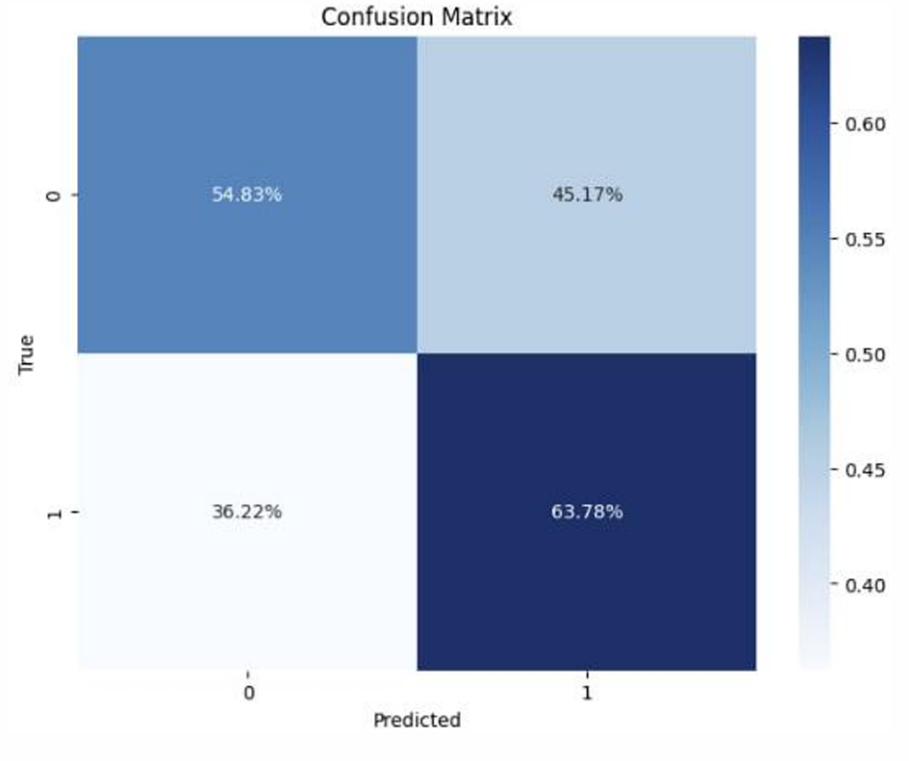
```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.88      0.55      0.68       983
           1       0.22      0.64      0.33       196

    accuracy                           0.56      1179
   macro avg       0.55      0.59      0.50      1179
weighted avg       0.77      0.56      0.62      1179

  Number of '1' labels: 569
  Number of '0' labels: 610
```

شکل آ_۶: دقت‌های مدل در حالت کل متن خبری

دلیل این اتفاق به این خاطر است که مدل در نمونه‌های خود صرفا عنوان‌های خبر را می‌بیند و از آنها یاد می‌گیرد و زمانی که به عنوان دستور کاربر یک متن خبر را بدهیم باعث می‌شود که وارد فضای ناشناخته‌تری بشود و نتواند به درستی پیش‌بینی کند.

دلیل دیگر آن نیز طولانی شدن محتوای ورودی مدل و کم جلوه شدن نمونه‌ها و شرح وظیفه است که در نهایت باعث می‌شود مدل دقت کمتری را حاصل کند.

همچنین ماتریس درهم‌ریختگی این حالت در شکل آ_۷ قابل مشاهده است.



شکل آ_۷: ماتریس درهم‌ریختگی در حالت کل متن خبری به عنوان ورودی

آ_۲-۵ نتایج اضافی‌تر دستورهای درخت تفکر و زنجیره تفکر

در فصل نتایج جدید به نتایج دادگان تست در این حالت پرداختیم. اما از آنجایی که برای توسعه سیستم انتخاب دستور براساس ورودی نیاز به تمامی دقت‌های این نوع دستورها در دادگان آموزش داشتیم. نتایج این بخش در جدول آ_۸ قابل مشاهده است.

Title Only	Accuracy	Precision	Recall	F1-Score	F1-Score('1')	F1-Score('0')
Prompt 18	54%	57%	60%	51%	40%	63%
Prompt 19	68%	59%	62%	59%	40%	78%
Prompt 21	48%	57%	59%	47%	40%	55%
Prompt 24	74%	58%	56%	57%	29%	84%
Prompt 25	71%	56%	56%	56%	31%	82%

شکل آ_۸: جدول دقت‌های دستورهای زنجیره تفکر و درخت تفکر در دادگان آموزش

نکته حائز اهمیت آن است که تمامی این دقت‌ها در حالت صفر نمونه[1] انجام شده است و دقت‌های مختلف آن را در دسته بندی‌های مختلف را می‌توان در جدول آ_۹ مشاهده نمود.

و در نهایت تمامی دقت‌ها برای دادگان ارزیابی در جهت ساخت یک طبقه‌بندی برای انتخاب

[1]Zero Shot

Title Only	Sports	Political	Economic	Global	Social	Cultural	Scientific	Health
Prompt 18	56%	43%	59%	43%	49%	41%	49%	52%
Prompt 19	59%	54%	62%	48%	57%	49%	54%	58%
Prompt 21	52%	41%	51%	40%	44%	37%	47%	49%
Prompt 24	55%	56%	54%	56%	55%	48%	52%	49%
Prompt 25	55%	54%	54%	56%	58%	49%	58%	49%

شکل آـ ۹: جدول دقت‌های دستورهای زنجیره و درخت تفکر در دادگان آموزش براساس دسته‌بندی

دستورالعمل مناسب براساس ورودی کاربر در جدول آـ ۱۰ می‌توان دید.

Title Only	Accuracy	Precision	Recall	F1-Score	F1-Score(‘1’)	F1-Score(‘0’)
Prompt 18	51%	56%	59%	48%	36%	61%
Prompt 19	66%	57%	60%	56%	36%	77%
Prompt 21	45%	56%	58%	44%	36%	52%
Prompt 24	76%	58%	57%	57%	30%	85%
Prompt 25	72%	57%	57%	57%	31%	83%

شکل آـ ۱۰: جدول دقت‌های دستورهای زنجیره و درخت تفکر در دادگان ارزیابی

Abstract

This work examines the capability of large language models (LLMs) to messure the importance of Persian news, evaluating their learning ability from content, reasoning skills, and overall cognitive capacities. Initially, annotated datasets were collected from various domains, including sports, politics, social issues, medicine, and culture, to develop an evaluation framework for LLMs. Within this framework, various existing models were analyzed and assessed under different scenarios and conditions to evaluate their analytical performance in both Persian and English. The findings indicate that prompts incorporating Chain-of-Thoughts and Tree-of-Thoughts significantly improve the models' performance. Additionally, the Symbol Tuning method enhances sensitivity to the input queries and their content.

Keywords: Large Language Models, Natural Language Processing, Machine Learning, News Importance Detection

Sharif University of Technology

Department of Computer Engineering

B.Sc. Thesis

News Importance Detection With the Use of Large Langauge Models

By:

Shayan Salehi

Supervisor:

Dr. Mahdi Jafari

January 2025