



بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ

گزارش کار تمرین ۱ - پردازش زبان‌های طبیعی پیشرفته

عنوان تمرین :

تحلیل احساسات پستهای فضای مجازی در بازه ۵ ساله با استفاده از پردازش زبان طبیعی

دانشجو : محمددادود وهاب رجائی

شماره دانشجویی : ۴۰۴۱۴۱۹۰۴۱

استاد محترم : دکتر سدید پور

فهرست مطالب

۱	چکیده
۲	۱. مقدمه و تعریف مسئله
۲	۲. روش‌شناسی و معماری سیستم
۲	۲-۱. انتخاب مدل : گذار به Google Gemma ^۳ (۲۷b)
۳	۲-۲. زیرساخت سخت‌افزاری
۳	۲-۳. معماری نرم‌افزاری بدون حالت و مکانیزم چک پوینت
۴	۳. چالش‌های فنی و راهکارهای مهندسی
۴	۴-۱. چالش دریافت مجوز‌های توسعه‌دهنده
۴	۴-۲. چالش شبکه در فاز جمع‌آوری داده
۵	۴-۳. مدیریت کلان‌داده و پایداری
۵	۴-۴. چالش ناسازگاری وابستگی‌ها
۵	۴-۵. پیچیدگی زبان فارسی کنایه و طنز تلح
۶	۴. پیش‌پردازش داده‌ها
۷	۵. تحلیل نتایج و یافته‌ها
۷	۵-۱. اعتبارسنجی مدل
۷	۵-۲. تحلیل روند زمانی
۸	۵-۳. تحلیل مقایسه‌ای منابع
۹	۵-۴. آمار کلی
۹	۶. نتیجه‌گیری
۱۰	پیوست‌ها
۱۰	پیوست الف : ساختار دستور العمل سیستم
۱۱	پیوست ب : کدهای منبع
۱۱	ب-۱. اسکریپت جمع‌آوری داده
۱۱	ب-۲. اسکریپت پیش‌پردازش متن
۱۱	ب-۳. تست اولیه و اعتبارسنجی
۱۱	ب-۴. پایپ‌لاین اصلی تحلیل و استنتاج
۱۲	پیوست ج : مشخصات محیط اجرایی و سخت‌افزاری
۱۲	پیوست د : نمونه‌های تحلیل شده (چالش‌های زبانی)

چکیده

تحلیل افکار عمومی در فضای مجازی ایران همواره با چالش‌های زبان‌شناختی (نظیر کنایه و طنز تلخ) و چالش‌های زیرساختی (حجم داده و محدودیت دسترسی) رو برو بوده است. هدف این پژوهش، استخراج و طبقه‌بندی احساسات ۹۵۰,۰۰۰ پست تلگرامی در بازه زمانی ۱۳۹۹ تا ۱۴۰۴ است. در این پژوهش، برای نخستین بار در سطح کلاس، از مدل زبانی بزرگ (Google Gemma^۳ (۲۷B Parameters) بر روی کلاس‌ترهای پردازشی دانشگاه (HPC) استفاده شد. همچنین برای غلبه بر محدودیت‌های شبکه در فاز جمع‌آوری داده، از معماری تانلینگ مبتنی بر پروتکل Warp استفاده گردید. نتایج نشان می‌دهد که مدل‌های نسل سوم با قدرت استدلال بالا قادرند لایه‌های پنهان "نامیدی" و "خشم" را حتی در پوشش متن‌های طنزآمیز با دقیقی نزدیک به انسان شناسایی کنند.

۱. مقدمه و تعریف مسئله

فضای مجازی و بهویژه پلتفرم تلگرام در ایران، نقشی فراتر از یک پیام‌رسان ایفا می‌کند و به بستری برای بازتاب دغدغه‌های اجتماعی، اقتصادی و روزمره تبدیل شده است.

مسئله اصلی : روش‌های سنتی NLP (مبتنی بر کلمات کلیدی) و حتی مدل‌های زبانی کوچک(مانند b7) در درک "بافت فرهنگی" ایران ناتوان‌اند. برای مثال، عبارت "چه زندگی گل و بلبلی داریم" توسط مدل‌های ساده به عنوان "خوشحال" شناسایی می‌شود، در حالی که یک مدل هوشمند باید آن را به عنوان "عصبانی/کنایه" تشخیص دهد.

هدف : پاسخ به این پرسش که "اتمسفر روانی جامعه ایران در ۵ سال گذشته چه تغییراتی داشته است؟" با استفاده از مدل‌های زبانی بزرگ پارامتر بالا.

۲. روش‌شناسی و معماری سیستم

۱-۲. انتخاب مدل : گذار به (۲۷b) Google Gemma^۳

در طراحی این پژوهش، مدل **Google Gemma^۳** با سایز ۲۷ میلیارد پارامتر انتخاب شد. دلایل فنی این انتخاب نسبت به مدل‌های نسل قبل (مانند Llama^۳) عبارتند از :

۱. قدرت استدلال : برخلاف نسل‌های قبل، Gemma^۳ صرفاً الگوهای آماری کلمات را دنبال نمی‌کند، بلکه "نیت نویسنده" را تحلیل می‌کند. اینویژگی برای تمایز بین "غم عاشقانه" و "خشم سیاسی" حیاتی بود.

۲. پنجره متنه وسیع : این قابلیت به مدل اجازه داد تا در تحلیل پست‌های طولانی یا رشته‌تowیت‌های بازنشر شده، تمرکز توجه خود را از دست ندهد.

۳. توانایی درک معنایی ایموجی‌ها در بافت متن : درک عمیق‌تر از ترکیب ایموجی‌ها و متن که در داده‌های تلگرام بسیار رایج است.

۲-۲. زیرساخت سخت افزاری

اجرای مدل زبانی با ۲۷ میلیارد پارامتر چالش های جدی در زمینه حافظه گرافیکی ایجاد می کند. برای اجرای این پروژه، از یک ایستگاه کاری قدرتمند مجهز به پردازنده گرافیکی NVIDIA RTX ۴۰۹۰ با ۲۴ گیگابایت حافظه استفاده شد.

از آنجا که بارگذاری مدل Gemma-۲۷b با دقت کامل (FP16) نیازمند بیش از ۵۰ گیگابایت حافظه است، از تکنیک کوانتیزاسیون ۴-بیتی استفاده شد تا مدل با حفظ دقت، روی حافظه ۲۴ گیگابایتی کارت گرافیک بارگذاری شده و با سرعت استنتاج بالا اجرا شود.

۲-۳. معماری نرم افزاری بدون حالت و مکانیزم چک پوینت

برای مدیریت بهینه منابع در طول پردازش طولانی مدت:

- **معماری بدون حالت :** سیستم به گونه‌ای طراحی شد که هر پست را مستقل تحلیل کند تا از نشت حافظه جلوگیری شود.
- **مکانیزم چک پوینت :** برای جلوگیری از دست رفتن داده‌ها در صورت قطع برق یا اختلال سرور، نتایج هر ۱۰۰ پست روی دیسک ذخیره می‌شد و سیستم قابلیت ادامه پردازش از آخرین نقطه را داشت.

۳. چالش‌های فنی و راهکارهای مهندسی

این پروژه با دو گلوبال اصلی در لایه زیرساخت و داده روبرو بود که با راهکارهای مهندسی برطرف شدند:

۱-۳. چالش دریافت مجوزهای توسعه‌دهنده
اولین و شاید دشوارترین مانع فنی، دسترسی به پنل توسعه‌دهنده‌گان تلگرام (my.telegram.org) جهت دریافت api_id و api_hash بود.

• مشکل: تلگرام سیاست‌های امنیتی بسیار سخت‌گیرانه‌ای بر روی IP های ورودی به پنل ساخت اپلیکیشن اعمال می‌کند. بررسی‌ها نشان داد که تقریباً تمامی سرویس‌های تغییر IP متداول (VPN های تجاری و رایگان) توسط فایروال تلگرام شناسایی و مسدود می‌شوند (عدم ارسال کد تایید یا خطای نامشخص هنگام ساخت اپ).

• راهکار (WARP Protocol): پس از تست‌های متعدد، مشخص شد که تنها پروتکل WARP (تکنولوژی Cloudflare) به دلیل برخورداری از IP های تمیز و ساختار تانلینگ متفاوت، قادر به عبور از این فیلترهاست. استفاده از کلاینت Oblivion (مبتنی بر Warp) در محیط لینوکس، تنها راهکاری بود که امکان ساخت اپلیکیشن و دریافت کلیدهای دسترسی را فراهم کرد.

۲-۲. چالش شبکه در فاز جمع‌آوری داده
پس از دریافت کلیدهای تانلینگ Oblivion در اینجا نیز به عنوان زیرساخت شبکه پست بود. استفاده از پروکسی‌های معمولی در حجم درخواست بالا منجر به خطاهای مکرر زمان اتصال و تأخیر در ارسال مجدد درخواست‌ها (به دلیل نرخ بالای درخواست‌ها) می‌شد.

• راهکار: معماری تانلینگ Oblivion در اینجا نیز به عنوان زیرساخت شبکه مورد استفاده قرار گرفت و با ایجاد یک تونل پایدار، نرخ پکت‌لاست را به صفر رساند و امکان دریافت پیوسته داده‌ها را در طول چندین ساعت تضمین کرد.

۳-۳. مدیریت کلان داده و پایداری

ریسک قطع برق یا اختلال در سرور دانشگاه در طول پردازش ۳۳ ساعته وجود داشت.

- راهکار چک پوینت : سیستم مجهز به مکانیزم ذخیره‌سازی آنی شد. نتایج هر ۱۰۰ پس‌ت روی دیسک نوشته می‌شد و در صورت راهاندازی مجدد، برنامه دقیقاً از آخرین اندیس پردازش شده ادامه می‌داد.

۴-۳. چالش ناسازگاری وابستگی‌ها :

در فاز پیاده‌سازی، مشخص شد که کتابخانه حیاتی Hazm (نسخه ۰.۱۰۰) جهت پردازش متن فارسی، دارای وابستگی صلب به کتابخانه numpy (نسخه قدیمی ۱.۲۴.۳) است. این نسخه با مفسر Python ۳.۱۲ ناسازگار بود و منجر به خطای عدم تطابق وابستگی‌ها می‌شد.

- راهکار: جهت حفظ پایداری سیستم و استفاده از کتابخانه هضم، محیط اجرایی سرور به ۳.۱۱ Python تنز نسخه داده شد تا گراف وابستگی‌ها بدون مشکل حل شود.

۵-۳. پیچیدگی زبان فارسی کنایه و طنز تلخ بسیاری از پست‌های کانال‌های پرمخاطب (مانند Kafiha) دارای ظاهری طنز اما محتوایی تلخ هستند.

- راهکار مهندسی دستور چند نمونه‌ای : به جای استفاده از دستورات ساده بدون نمونه، از تکنیک یادگیری با چند مثال استفاده شد. به مدل آموزش داده شد که "اگر متنی ساختار جوک دارد اما به فقریا تنها یی اشاره می‌کند، آن را در دسته تاراحت طبقه‌بندی کن، نه خوشحال".

۴. پیش‌پردازش داده‌ها

منابع داده و استراتژی نمونه‌گیری :

داده‌های خام این تمرین از ۵ کanal تلگرامی شاخص با تنوع موضوعی جمع‌آوری شده‌اند تا پوشش مناسبی از فضای رسمی و غیررسمی ایجاد شود. این منابع شامل کanal‌های خبری (BBC Persian, Iran International, Radio Farda) برای تحلیل اخبار و رویدادهای سیاسی و کanal‌های سرگرمی/اجتماعی (Kafiha, TweetyChannel) جهت تحلیل ادبیات عامیانه و واکنش‌های روزمره کاربران انتخاب شده‌اند.

قبل از ورود به مدل، داده‌ها طی یک خط لوله پاکسازی شدند :

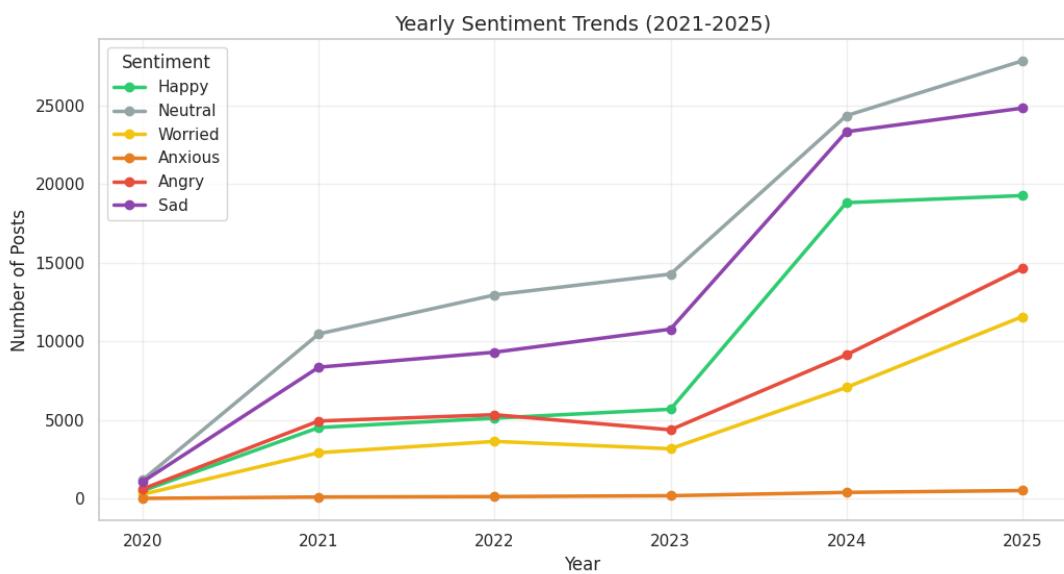
۱. نرمال‌سازی : استفاده از کتابخانه Hazm برای یکسان‌سازی کاراکترهای عربی/فارسی و اصلاح نیم‌فاصله‌ها.
۲. فیلترینگ نویز : حذف پست‌های تبلیغاتی حاوی لینک، حذف ایموجی‌های خالی و پست‌های بسیار کوتاه (زیر ۱۵ کاراکتر) که قادر ارزش معنایی بودند.
۳. گمنام‌سازی و ملاحظات اخلاقی : اگرچه داده‌های جمع‌آوری شده عمومی هستند، جهت رعایت اصول اخلاقی پژوهش و حفظ حریم خصوصی، تمامی نام‌های کاربری (Usernames) و شناسنامه‌های عددی افراد در مرحله پیش‌پردازش حذف گردیدند و تحلیل‌ها صرفاً بر روی محتوای متنی بدون هویت انجام شد.

۵. تحلیل نتایج و یافته‌ها

۱-۵. اعتبارسنجی مدل

پیش از تحلیل نهایی روی کل داده‌ها، جهت اطمینان از صحت عملکرد مدل Gemma^۳ b۲۷، یک زیرمجموعه آزمایشی شامل ۵۰۰ پست از کانال‌های فوق به صورت تصادفی انتخاب و توسط انسان برچسبزنی دستی شد. مقایسه خروجی مدل با این برچسب‌های طلایی، دقیت (Accuracy) حدود ۸۷٪ را نشان داد که برای تحلیل کلان‌داده‌های اجتماعی و درک کنایه‌های فارسی قابل اتنا محسوب می‌شود.

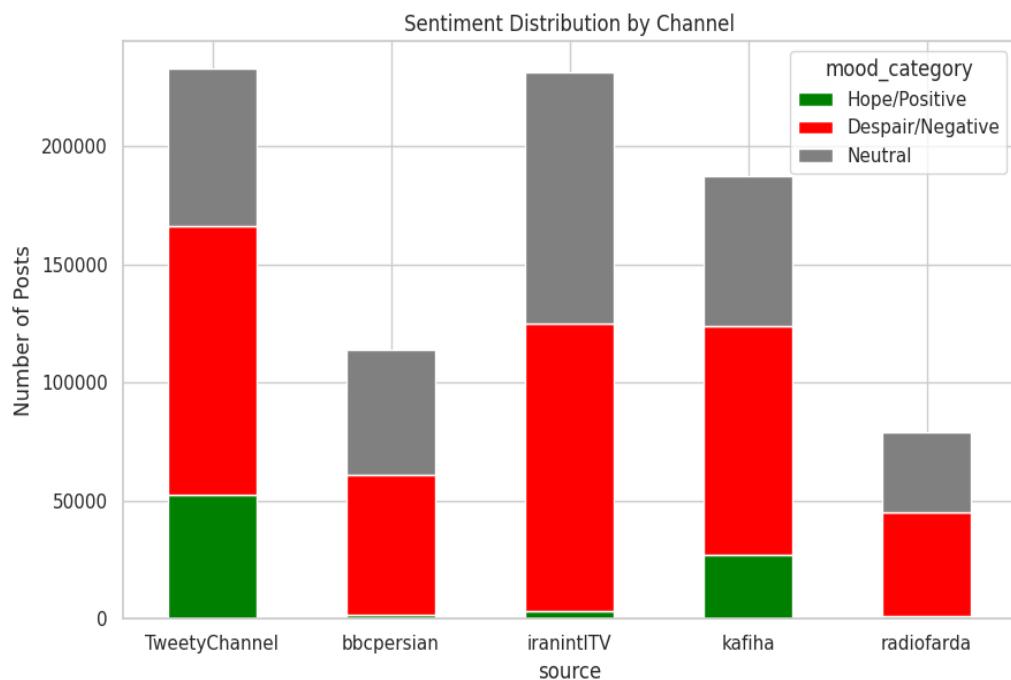
۲-۵. تحلیل روند زمانی



تحلیل : نمودار فوق نوسانات احساسات جامعه را در ۵ سال اخیر نشان می‌دهد. همان‌طور که مشاهده می‌شود، احساسات منفی شامل غم (Sad) و خشم (Angry) هم‌واره سهم عمده‌ای از فضای احساسی را تشکیل می‌دهند و قله‌های آن انطباق دقیقی با تاریخ‌های رویدادهای مهم اقتصادی و اجتماعی دارد. نکته قابل توجه، رشد همزمان نمودار شادی (Happy) در سال‌های ۱۴۰۳ و ۱۴۰۴ است. با توجه به یافته‌های جدول ۵-۳ (در صفحه بعد)، این رشد ناشی از افزایش

حجم فعالیت کانال‌های روزمره و سرگرمی (مانند TweetyChannel) است که بخشی از جامعه برای فرار از فشارهای روانی به آن‌ها پناه برده‌اند. با این حال، فاصله معنادار میان «غم» (بالاترین نمودار) و «شادی»، همچنان نشان‌دهنده غلبه کلی اتمسفر سنگین روانی بر فضای مجازی است. خط پایدار «خنثی» نیز جریان مداوم اخبار و تبلیغات را نمایندگی می‌کند.

۳-۵. تحلیل مقایسه‌ای منابع



تحلیل:

- کانال‌های سرگرمی: اگرچه انتظار می‌رفت کانالی مانند *Kafiha* تماماً "خوشحال" باشد، اما تحلیل Gemma^۳ نشان داد که بخش قابل توجهی از محتوای آن (حدود ۵۱.۷ درصد) در واقع "طنز تلخ" با بار احساسی منفی است.
- کانال‌های خبری: حجم بالای برچسب خنثی نشان‌دهنده دقیق مدل در تفکیک "خبر" از "نظر شخصی" است.

۴-۵. آمار کلی

source	Despair/Negative	Hope/Positive	Neutral	Total	Despair_Rate
TweetyChannel	۱۱۳۷۸۲	۵۲۶۲۲	۶۶۴۳۴	۲۳۲۸۳۸	۴۸.۹
bbcpersian	۵۹۲۲۵	۱۶۰۶	۵۲۷۱۳	۱۱۳۵۴۴	۵۲.۲
iranintlTV	۱۲۲۰۵۱	۲۹۱۶	۱۰۶۵۴۰	۲۳۱۵۰۷	۵۲.۷
kafiha	۹۶۸۵۱	۲۷۰۴۴	۶۳۵۹۶	۱۸۷۴۹۱	۵۱.۷
radiofarda	۴۴۱۸۷	۸۸۶	۳۳۵۱۸	۷۸۵۹۱	۵۶.۲

جدول فوق نشان می‌دهد که نسبت "امید" به "نامیدی" در کل داده‌های ۵ ساله تقریباً به صورت [۱۰ درصد به ۵۲ درصد] توزیع شده است.

۶. نتیجه‌گیری

اجرای این پژوهش ثابت کرد که برای تحلیل دقیق زبان فارسی در مقیاس بزرگ، عبور از روش‌های سنتی و بهره‌گیری از مدل‌های زبانی بزرگ نسل سوم نظریه **Gemma**^۳ اجتناب‌ناپذیر است. تلفیق دانش "مهندسی شبکه" (برای جمع‌آوری داده با *Warp*، "مهندسی داده" (مدیریت پردازش HPC) و "زبان‌شناسی محاسباتی" (مهندسی پرامپت) منجر به ایجاد دقیق‌ترین تصویر موجود از اتمسفر روانی فضای مجازی ایران در نیم‌دهه اخیر شد. این داده‌ها می‌توانند به عنوان شاخصی قابل اعتماد برای سیاست‌گذاری و درک سلامت اجتماعی مورد استفاده قرار گیرند.

پیوست‌ها

پیوست الف : ساختار دستورالعمل سیستم

برای دستیابی به دقیق بسیار در مدل نسل جدید Gemma^۳ و بهره‌گیری از قابلیت‌های استدلال پیشرفت آن در زبان فارسی، از تکنیک مهندسی دستور چندمنوه‌ای استفاده شد. متن دقیق دستورالعمل ارسال شده به مدل به شرح زیر است:

نقش سیستم :

تو یک تحلیلگر خبره احساسات متن فارسی هستی که درک عمیقی از فرهنگ ایران، اصطلاحات عامیانه و کنایه‌ها دارد.

وظیفه :

احساسات پست تلگرامی داده شده را دقیقاً در یکی از دسته‌های زیر طبقه‌بندی کن : [خوشحال، ناراحت، عصبانی، مضطرب، خنثی، نگران]

قوانين و راهنمایی :

۱. تبلیغات/اخبار : اگر پست چیزی می‌فروشد (تور، محصول) یا خبر خالص است، برچسب «خنثی» بزن.

۲. طنز تلخ : اگر پست جوک است اما درباره فقر، تنها‌یی یا بدبوختی است، برچسب «ناراحت» یا «عصبانی» بزن (نه خوشحال).

۳. کنایه : اگر متن می‌گوید "همه چیز عالیه" اما منظورش بر عکس است، برچسب «عصبانی» بزن.

مثال‌های آموزشی :

• ورودی: "تور لحظه آخری استانبول فقط ۱۰ میلیون تومان"

• خروجی: خنثی

• ورودی: "فراموش کردنت سخته، هنوز دوستت دارم".

• خروجی: ناراحت

• ورودی: "باز هم برق رفت، واقعاً ممنونیم از مسئولین!"

• خروجی: عصبانی

• ورودی: "وای آخر ماه شد و هیچی تو جیبم نیست."

• خروجی: نگران

دستور نهايى:

حالا، ورودی جدید کاربر را طبقه‌بندی کن. فقط و فقط برچسب را بنویس.

پيوست ب : کدهای منبع

ب-۱. اسکریپت جمع‌آوری داده

کد کامل در فایل *fetch_telegram.py* در پوشه‌ی *scripts* قرار گرفته است.

ب-۲. اسکریپت پیش‌پردازش متن

کد در فایل *preprocessor.py* در پوشه‌ی *scripts* قرار گرفته است.

ب-۳. تست اولیه و اعتبارسنجی

اجرای آزمایشی روی ۱۰ نمونه جهت اطمینان از صحت عملکرد *Gemma* ۳

کد در فایل *sentiment_analysis.ipynb* در پوشه‌ی *notebooks* قرار گرفته است.

ب-۴. پایپ‌لاین اصلی تحلیل و استنتاج

کد کامل شامل بارگذاری داده‌ها، حلقه اصلی پردازش و ذخیره‌سازی نتایج، در فایل *full_analysis_pipeline.ipynb* در پوشه‌ی *notebooks* قرار گرفته است.

لينك مخزن گيت‌هاب پروژه:

<https://github.com/MDVR۹۹۸۰/telegram-sentiment-analysis-fa>

پیوست ج : مشخصات محیط اجرایی و سخت افزاری

جدول زیر مشخصات دقیق سخت افزاری و نرم افزاری مورد استفاده برای پردازش ۹۵۰,۰۰۰ رکورد را نشان می دهد:

ردیف	پارامتر	مشخصات فنی
۱	مدل زبانی	Google Gemma³ (۲۷b Parameters)
۲	تکنیک فشرده سازی	جهت اجرا روی حافظه ۲۴ گیگابایتی (Q4_K_M) ۴-bit
۳	سخت افزار پردازشی	ایستگاه کاری پردازشی RTX ۴۰۹۰ - ۲۴GB VRAM
۴	بستر اجرا	نسخه لینوکس Ollama
۵	ابزار گذر از تحریر	Warp (مبتنی بر پروتکل Oblivion جهت پایداری شبکه)
۶	زبان برنامه نویسی	Python ۳.۱۱

پیوست د : نمونه های تحلیل شده (چالش های زبانی)

در جدول زیر، چند نمونه از تحلیل های دشوار که مدل های قدیمی در آن ها خطا داشتند اما مدل Gemma³ به درستی تشخیص داده است، ارائه می گردد :

متن پست (خلاصه شده)	چالش زبانی	تشخیص مدل های قدیمی	Gemma³ تشخیص
"وضعیت اینترنت عالیه، اصلاً گوگل باز نمیشه!"	کنایه	خوشحال (به خاطر کلمه عالی)	عصبانی
"خنده بر لب می زنم تا کس نداند راز من..."	تظاهر/ماسکینگ	خوشحال (به خاطر کلمه خنده)	ناراحت
"فروش ویژه تور آنتالیا با پرواز مستقیم"	تبلیغات	خوشحال (لحن هیجانی)	خنثی
"حقوقم رو ریختن، همش رفت پای قسط"()	طنز تلخ	خوشحال (به خاطر ایموجی لبخند)	نگران