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 ۱۴۰۴

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

فهرست جداول

فهرست تصاویر

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

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

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 موضوع

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.

فصل ۴

نتایج جدید

در این فصل نتایج جدید به‌دست‌آمده در پایان‌نامه توضیح داده می‌شود. در صورت نیاز می‌توان نتایج جدید را در قالب چند فصل ارائه نمود. همچنین در صورت وجود پیاده‌سازی، بهتر است نتایج پیاده‌سازی را در فصل مستقلی پس از این فصل قرار داد.

فصل ۵

نتیجه‌گیری

در این فصل، ضمن جمع‌بندی نتایج جدید ارائه‌شده در پایان‌نامه یا رساله، مسائل باز باقی‌مانده و همچنین پیشنهادهایی برای ادامه‌ی کار ارائه می‌شوند.

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.

واژه‌نامه

الف

ابتکاری ..... heuristic

ابعاد بالا ..... high dimensions

اریب ..... bias

آستانه ..... threshold

اصل لانه‌ی کبوتری ..... pigeonhole principle

ان‌پی‌ـسخت ..... NP-Hard

انتقال ..... transition

ب

برخط ..... online

برنامه‌ریزی خطی ..... linear programming

بهینه ..... optimum

بیشینه ..... maximum

پ

پرت ..... outlier

پرسمان ..... query

پوشش ..... cover

پیچیدگی ..... complexity

ت

تنظیم براساس دستورالعمل .... Insturction Tuning

ج

جداپذیر ..... separable

جعبه سیاه ..... black box

جویبار داده ..... data stream

ح

حدی ..... extreme

حریصانه ..... greedy

خ

خوشه ..... cluster

خطی ..... linear

د

دستور ..... Prompt

درخت تفکر ..... Tree-of-Thoughts

درخواست سامانه ..... System Prompt

ر

رأس ..... vertex

رسمی ..... formal

ز

زنجیره تفکر ..... Chain-of-Thoughts

س

سرشکن ..... amortized

سلسه‌مراتبی ..... hierarchichal

ش

شبه کد ..... pseudocode

شیء ..... object

ص

صدق‌پذیری ..... satisfiability

غ

غلبه ..... dominate

ف

فاصله ..... distance

فضا ..... space

ق

قطعی ..... deterministic

ک

کارا ..... efficient

کاندیدا ..... candidate

کمینه ..... minimum

م

مدل‌های زبانی بزرگ .... Large Language Models

مهندسی درخواست ..... Prompt Engineering

ن

نابه‌جایی ..... inversion

ناوردا ..... invariant

نقطه‌ی مرکزی ..... center point

نیم‌فضا ..... half space

هـ

هزینه‌ی آشوب ..... price of anarchy (POA)

ی

یادگیری چند نمونه ..... Few-Shot Learning

یادگیری درون‌متنی ..... In-Context Learning

پیوست آ

مطالب تکمیلی

پیوست‌های خود را در صورت وجود می‌توانید در این قسمت قرار دهید.

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