استفاده از مدل‌های زبانی بزرگ برای متن‌کاوی، بازیابی اطلاعات، و تحلیل متن

آزاده شاکری

دانشکده مهندسی برق و کامپیوتر - دانشگاه تهران

چکیده

در سال‌های اخیر، پیشرفت‌های چشمگیر در پردازش زبان طبیعی (NLP) و مدل‌های زبانی بزرگ (LLMs)، مانند GPT-4 و دیگر مدل‌های مبتنی بر ترنسفورمر، منجر به تحولات گسترده‌ای در کاربردهای متن‌کاوی و تحلیل داده‌های متنی شده است. این مدل‌ها که بر پایه معماری‌های پیشرفته شبکه‌های عصبی عمیق ساخته شده‌اند، توانایی ما را در درک معانی، تولید متن و استخراج دانش از داده‌های متنی به طرز قابل توجهی افزایش داده‌اند. این پژوهش بر استفاده از این مدل‌های پیشرفته برای کاربردهای مختلف کاوش داده‌های متنی تمرکز دارد.

داده‌های متنی، به‌عنوان یکی از غنی‌ترین منابع اطلاعاتی، روزانه در قالب‌های مختلفی از جمله مقالات خبری، پست‌های رسانه‌های اجتماعی، ایمیل‌ها، و محتوای انجمن‌های پرسش و پاسخ تولید می‌شود. با افزایش حجم این داده‌ها، نیاز به ابزارهای هوشمند برای تحلیل و استخراج دانش از آن‌ها بیشتر از هر زمان دیگری احساس می‌شود. این پژوهش به بررسی کاربردهای مختلفی مانند بازیابی اطلاعات پیشرفته، سیستم‌های مکالمه هوشمند، شناسایی محتوای توهین‌آمیز، خلاصه‌سازی متن و سیستم‌های توصیه‌گر مکالمه‌ای می‌پردازد. این کاربردها با استفاده از مدل‌های زبانی پیشرفته و شبکه‌های عصبی عمیق، به‌ویژه معماری‌هایی مانند ترنسفورمر و روش‌های مبتنی بر یادگیری انتقالی پیاده‌سازی و ارزیابی می‌شوند.

کلیدواژه‌ها: مدیریت داده‌های متنی، متن‌کاوی، یادگیری انتقالی، بازیابی اطلاعات، یادگیری عمیق، مدل‌های زبانی بزرگ

Leveraging Large Language Models for Text Mining, Information Retrieval, and Text Analysis

Abstract

In recent years, significant advancements in natural language processing (NLP) and large language models (LLMs), such as GPT-4 and other Transformer-based models, have led to transformative changes in text mining and textual data analysis applications. These models, built on advanced deep neural network architectures, have greatly enhanced our ability to understand semantics, generate text, and extract knowledge from textual data. This research focuses on utilizing these advanced models for various applications in textual data exploration.

Textual data, as one of the richest sources of information, is generated daily in various formats, including news articles, social media posts, emails, and content from question-and-answer forums. With the increasing volume of such data, the need for intelligent tools to analyze and extract knowledge from it has become more pressing than ever. This study explores diverse applications, including advanced information retrieval, intelligent conversational systems, detection of offensive content, text summarization, and conversational recommendation systems. These applications are implemented and evaluated using advanced language models and deep neural networks, particularly architectures like Transformers and methods based on transfer learning.

Keywords: Text data management, Text mining, Transfer learning, Information retrieval, Deep learning, Large language models.

مقدمه و بیان مسئله

در عصر حاضر، داده‌های متنی به‌عنوان یکی از اصلی‌ترین و غنی‌ترین منابع اطلاعاتی در جهان دیجیتال شناخته می‌شوند. با گسترش فناوری‌های دیجیتال و افزایش استفاده از شبکه‌های اجتماعی، حجم داده‌های متنی تولیدشده به‌طور نمایی رشد کرده است. این داده‌ها نه‌تنها شامل محتوای متنی سنتی مانند مقالات علمی، کتاب‌ها، و ایمیل‌ها هستند، بلکه پست‌های شبکه‌های اجتماعی، نظرات کاربران، و محتوای تولیدشده در پلتفرم‌های اشتراک‌گذاری نیز بخش عمده‌ای از آن‌ها را تشکیل می‌دهند. این حجم عظیم داده‌ها، فرصت‌های بی‌شماری را برای استخراج دانش و تحلیل رفتارهای انسانی فراهم می‌کند، اما در عین حال چالش‌های جدیدی را نیز در زمینه‌ی پردازش و تحلیل این داده‌ها ایجاد کرده است.

در سال‌های اخیر، پیشرفت‌های چشمگیر در حوزه‌ی پردازش زبان طبیعی و ظهور مدل‌های زبانی بزرگ[1] مانند GPT-4، BERT، T5، و دیگر مدل‌های مبتنی بر معماری Transformer، تحولات بنیادینی را در این زمینه ایجاد کرده‌اند. این مدل‌ها با استفاده از معماری‌های پیشرفته‌ی شبکه‌های عصبی عمیق، به‌ویژه معماری Transformer، توانسته‌اند درک معنایی، تولید متن، و استخراج دانش از داده‌های متنی را به‌طور قابل‌توجهی بهبود بخشند. این پیشرفت‌ها نه‌تنها در کاربردهای سنتی مانند ترجمه‌ی ماشینی و تحلیل احساسات، بلکه در حوزه‌های نوظهوری مانند تولید متن خلاقانه، سیستم‌های مکالمه‌ای هوشمند، و تشخیص محتوای مضر نیز تأثیرگذار بوده‌اند.

یکی از مهم‌ترین دستاوردهای این مدل‌ها، توانایی آن‌ها در یادگیری انتقالی[2] است. این قابلیت به مدل‌ها اجازه می‌دهد تا دانش کسب‌شده از یک حوزه‌ی خاص را به حوزه‌های دیگر انتقال دهند و با حداقل داده‌ی آموزشی، عملکردی قابل‌قبول ارائه دهند. این ویژگی به‌ویژه در کاربردهایی مانند تشخیص اخبار جعلی، تحلیل احساسات، و خلاصه‌سازی متون بسیار مفید بوده است. علاوه بر این، مدل‌های زبانی بزرگ توانسته‌اند با استفاده از داده‌های عظیم و روش‌های پیشرفته‌ی آموزش، درک عمیق‌تری از زبان طبیعی به دست آورند و در نتیجه، عملکرد بهتری در وظایف پیچیده‌ی پردازش زبان طبیعی از خود نشان دهند.

در این پژوهش، تمرکز اصلی بر استفاده از مدل‌های زبانی بزرگ و شبکه‌های عصبی عمیق در کاربردهای متنوع کاوش داده‌های متنی است. این کاربردها شامل بازیابی اطلاعات پیشرفته، سیستم‌های مکالمه‌ای هوشمند، تشخیص محتوای توهین‌آمیز، خبره‌یابی، و سیستم‌های توصیه‌گر مبتنی بر متن هستند. با استفاده از معماری‌های پیشرفته‌ی Transformer و روش‌های مبتنی بر یادگیری انتقالی، این پژوهش به دنبال ارائه‌ی راه‌حل‌های نوین برای چالش‌های موجود در حوزه‌ی متن‌کاوی است.

علاوه بر این، در این پژوهش به بررسی چالش‌های خاصی که در پردازش داده‌های متنی به‌ویژه در زبان‌های کم‌منبع مانند فارسی وجود دارد، پرداخته می‌شود. با توجه به این که بسیاری از مدل‌های زبانی بزرگ بر روی زبان‌هایی مانند انگلیسی آموزش دیده‌اند، تطبیق این مدل‌ها با زبان‌های دیگر نیازمند روش‌های خاصی مانند Fine-tuning و استفاده از داده‌های چندزبانه است. این پژوهش به بررسی روش‌های بهینه‌سازی این فرآیندها و بهبود عملکرد مدل‌ها در زبان‌های کم‌منبع نیز می‌پردازد.

---

[1] Large Language Model
[2] Transfer Learning

در ادامه‌ی این پژوهش، ابتدا مروری بر پیشرفت‌های اخیر در حوزه‌ی مدل‌های زبانی بزرگ و شبکه‌های عصبی عمیق ارائه می‌شود. سپس، کاربردهای متنوع این فناوری‌ها در حوزه‌ی کاوش داده‌های متنی بررسی شده و چالش‌ها و فرصت‌های پیش‌رو در این زمینه تحلیل می‌شوند. در نهایت، نتایج حاصل از پیاده‌سازی و ارزیابی مدل‌های پیشنهادی ارائه خواهد شد. این پژوهش به دنبال آن است که با استفاده از آخرین دستاوردهای حوزه‌ی پردازش زبان طبیعی و یادگیری عمیق، گامی مؤثر در جهت بهبود روش‌های کاوش داده‌های متنی و ارائه‌ی راه‌حل‌های نوین برای چالش‌های موجود بردارد.

کاربردهایی از کاوش و بازیابی داده‌های متنی و پژوهش‌های پیشین

➢ بازیابی پرسش بین‌زبانی در انجمن‌های پرسش و پاسخ

امروزه اطلاعات موجود در انجمن‌های پرسش و پاسخ در کنار اطلاعات موجود در وب به کمک جستجوگران آمده و بسیاری از افراد نیازهای اطلاعاتی خود را در این انجمن‌ها مرتفع می‌کنند. مهمترین هدف یک انجمن پرسش و پاسخ، پاسخ دادن به پرسش‌های پرسش شده‌ی اخیر در کوتاهترین زمان ممکن است. در بسیاری از مواقع پرسش‌های کاربران تکراری و پاسخ‌های آنها در انجمن موجود است و کافیست که پرسش مشابه پرسش کاربر بازیابی شده و پاسخ آن بلافاصله مورد استفاده قرار گیرد. در شرایطی که پرسش ورودی به یک زبان و پرسش‌هایی که بازیابی باید از میان آنها صورت گیرد به زبان دیگری باشند مسئله تبدیل به مسئله بازیابی پرسش دوزبانه خواهد شد. با توجه به اینکه در بسیاری از زبان‌ها، انجمن‌های پرسش و پاسخ تک‌زبانه قدرتمندی وجود ندارد که کاربران بتوانند پاسخ پرسش‌های خود را از آنها دریافت کنند، مسئله‌ی بازیابی پرسش در محیط دوزبانه و امکان استفاده از منابع اطلاعاتی موجود در زبان‌های دیگر، اهمیت بیشتری پیدا می‌کند. از آنجا که افراد نیازهای اطلاعاتی یکسان را به صورتهای مختلفی مطرح می‌کنند و یا از کلمات مشابه برای طرح پرسش‌های متفاوت استفاده می‌نمایند، مسئله‌ی بازیابی پرسش با مشکل شکاف واژگانی روبرو است. استفاده از فراداده در کنار متن پرسش‌ها می‌تواند اطلاعات کامل‌تری از متن پرسش‌ها را برای رفع این مشکل در اختیار سیستم بازیابی قرار دهد[۶-۸].

در این رساله، روش‌های ITRLM ، و RankFusion-X با استفاده از فراداده برای بهبود عملکرد بازیابی پرسش در انجمن‌های پرسش و پاسخ ارائه شده‌اند. روش ITRLM با استفاده از دو روش مبتنی بر ترنسفورمر برای بکارگیری فراداده، مدل‌زبانی مبتنی بر ترجمه را در فرآیند بازیابی پرسش‌ها در یک زبان بهبود می‌دهد. در این روش از مدل BERT برای بهره‌گیری از اطلاعات مربوط به دسته پرسش، و از مدل زبانی بزرگ با تولید تقویت شده با بازیابی برای بکارگیری پاسخ پرسش‌ها استفاده می‌شود. برای ارزیابی روش ITRLM از مجموعه داده‌ای SemEval-2017 استفاده شده است و معیار MAP برابر با ۵۱.۴۷ را در محیط تک‌زبانه حاصل می‌کند.

روش ارائه شده دوم با نام RankFusion-X به بازیابی پرسش در محیط دوزبانه می‌پردازد. در هسته روش RankFusion-X، یک بازرتبه‌بند مبتنی بر مدل زبانی بزرگ قرار دارد که از مدل‌های GPT-3.5 و GPT-4o-mini استفاده می‌کند و با کمک فراداده موجود در انجمن به بازرتبه‌بندی پرسش‌های کاندیدا می‌پردازد. برای افزایش دقت و بهره‌گیری از فراداده، سه مؤلفه تکمیلی به مدل‌زبانی بزرگ اضافه شده‌اند: یادگیری در زمینه بین‌زبانی (X-ICL)، prompting با دانش تولید

شده، و ترکیب رتبه‌بندی مبتنی بر فیلد، که باعث بهبود دقت رتبه‌بندی در محیط‌های دوزبانه می‌شوند. به منظور ارزیابی روش پیشنهادی، ابتدا نسخه دوزبانه‌ی فارسی-انگلیسی مجموعه داده‌ای استاندارد SemEval 2017-task 3 ایجاد شده است. نتایج بدست آمده از آزمایش‌های انجام شده نشان می‌دهد RankFusion-X با مدل GPT-4o-mini با کمک مولفه‌های پیشنهادی، با معیار MAP برابر با ۵۵.۶۲، بهترین نتیجه را در بازیابی پرسش بین‌زبانی در انجمن‌های پرسش و پاسخ حاصل می‌کند[۹-۱۱].

➢ استفاده از نظرات کاربران برای استنباط دانش ذهنی در سیستم های توصیه‌گر مکالمه‌ای

دسته‌ای از سوالاتی که کاربران قبل از خرید یک محصول یا رزرو یک خدمات با آن‌ها مواجه می‌شوند، سوالاتی هستند که در پایگاه‌های داده شامل داده‌های واقعی مثل اندازه، قیمت، امتیاز و مشخصات محصول مورد نظر وجود ندارند و مربوط به ویژگی‌های ذهنی هستند. به عنوان مثال، در مورد کیفیت سرویس‌دهی یک هتل یا رستوران، شرایط لغو رزرو، میزان امنیت یک خودرو یا مناسب خانواده بودن اتمسفر یک رستوران، مواردی هستند که پاسخ آن‌ها رو باید در نظرات، تجربیات و پرسش‌های پرتکرار کاربران در مورد محصول یا خدمات مورد نظر جستجو کرد. کاری که افراد برای دستیابی به پاسخ مورد نظرشان می‌کنند هم این هست که قبل از خرید اون محصول، به سراغ بررسی‌ها و نظرات دیگر کاربران رفته و از تجربیات آن‌ها برای این منظور استفاده می‌کنند.

این دانش‌ها که مبنای آن تجربیات، نظرات و احساسات افراد هست رو به دانش ذهنی تعبیر کرده و قصد داریم در این پژوهش به استنباط این دانش‌ها از سخنان کاربران در سیستم‌های توصیه‌گر مکالمه‌ای پرداخته تا این سیستم‌ها بتوانند پاسخ مناسبی برای این دسته از سوالات کاربران داشته باشند.

روش کار و معماری سیستم توصیه‌گر مکالمه‌ای مبتنی بر دانش ذهنی[۳] به صورتی که در ادامه معرفی می‌شود، خواهد بود.

مرحله‌ی اول:

ابتدا با توجه به محتوای مکالمه و آخرین سخن کاربر، ابتدا تشخیص داده میشه که آیا برای پاسخ به این گفته‌ی کاربر نیاز به استنباط دانش ذهنی وجود داشته و یا می‌توانیم از سیستم‌های توصیه‌گر مکالمه‌ای مرسوم برای تولید پاسخ مناسب استفاده کنیم. چالشی که در این مرحله هنوز هم در مقالات وجود داره، قابلیت تعمیم این تشخیص برای وضعیت‌ها و دامنه‌های دیده‌نشده هست که قصد داریم از قابلیت تعمیم و دانش نهفته در مدل‌های زبانی بزرگ برای این مسئله‌ی طبقه‌بندی دوکلاسه و طبقه‌بندی گفته‌های کاربران استفاده کنیم.

مرحله‌ی دوم:

پس از تشخیص نیاز به استنباط دانش ذهنی، دانش‌های مرتبط با محتوای مکالمه و گفته‌ی کاربر، از منابع دانش ذهنی و واقعی[۴] استخراج شده و نوعی رتبه‌بندی برای استخراج این برش‌های دانش صورت می‌گیرد. در این مرحله نیز چالش تعمیم‌پذیری

[3] Subjective
[4] Factual

سیستم مطرح خواهد بود. روشی که در مقالات برای این مسئله ارائه شده بود داده‌افزایی با استفاده از مدل‌های زبانی بزرگ بوده و مشکلی که وجود داره، لزوم استناد برش‌های دانش بر واقعیت هست که این روش از داده‌افزایی تا الان نتوانسته این تضمین را فراهم کند. مخصوصاً این که در این مسئله نظرات کاربران مطرح هست، نیاز به داده‌های واقعی خواهد بود. به همین دلیل، قصد داریم از داده‌ها و نظرات کاربران درباره‌ی موجودیت‌ها (در این مسئله، هتل‌ها و رستوران‌ها) که به صورت عمومی در سطح وب قرار دارند برای غنی‌سازی منبع دانش ذهنی استفاده کنیم.

مرحله‌ی سوم:

در نهایت، گزیده‌ای از برش‌های دانش استخراج‌شده یا چکیده‌ای از اونها را به عنوان پاسخ سیستم تولید می‌کنیم. برای ارزیابی پاسخ تولید شده، میتونیم از روش‌های خودکار مانند BLEU و ROUGE و یا روش‌هایی که از مدل‌های Pre-trained برای ارزیابی پاسخ‌ها استفاده می‌کنند و وابسته به بافتار هم هستند مانند BERTScore استفاده کنیم. با توجه به ماهیت تعاملی سیستم‌های توصیه‌گر مکالمه‌ای، قصد داریم از روش‌های شبیه‌سازی کاربر هم برای ارزیابی پاسخ‌های تولیدشده توسط سیستم استفاده کنیم. با دانش حال حاضر ما نمونه‌ای از کاربرد این روش‌ها برای مسئله‌ی استنباط دانش Subjective در این سیستم‌ها مشاهده نشده و برای ارزیابی پاسخ‌ها از معیارهایی که گفته شد و یا ارزیابی انسانی استفاده می‌شود.

موردی که اخیراً در حال مطالعه درباره‌اش هستم، بحث RAG[5] هست که در مرحله‌ی تولید پاسخ سیستم از برش‌های دانش استخراج‌شده میتونه مؤثر باشه. مفاهیم اولیه رو در این مورد یاد گرفتم و دارم سعی می‌کنم یک نمونه از اون رو پیاده‌سازی کنم که هم آشنایی بیشتری با روند کارش پیدا کنم و هم به مرور پیاده‌سازی یک CRS رو با استفاده ازش انجام بدم.

همچنین در مرحله‌ی بعدی دنبال منابعی خواهیم بود که نظرات کاربران را درباره موجودیت‌ها پیدا کرد به صورتی که به منابع دانش Subjective اضافه کنیم. موجودیت‌هایی که در این مجموعه‌دادگان وجود دارد، ۱۱۰ هتل و ۳۳ رستوران هست که از دادگان MultiWOZ جمع‌آوری شده‌اند.

➢ استفاده از مدل‌های زبانی بزرگ برای هدایت پرسش در انجمن‌های پرسش و پاسخ

در سال‌های اخیر سکوهای برخطی مانند Stack Exchange، Stack Overflow توجه زیادی بر روی اینترنت به دست آورده‌اند. این سکوها یا انجمن‌های آنلاین پرسش‌وپاسخ برخی بر محوریت یک موضوع خاص مانند کدنویسی و برخی دربرگیرنده‌ی مجموعه‌ای گسترده از موضوعات هستند. انجمن‌های پرس و پاسخ این توانایی را به کاربر می‌دهند که در بستر انجمن پرسش‌های خاص خود را پیمایش و یا مطرح کند و به دنبال جواب خود باشد.

هدف این انجمن‌ها فراهم آوردن هر چه سریع‌تر و دقیق‌تر پاسخ برای کاربران با استفاده از هدایت پرسش به کاربران خبره و در نهایت تبادل دانش و اطلاعات میان کاربران خود است. یکی از مشکلات این انجمن‌ها نرخ پایین کاربران فعال علاقه‌مند به ارسال

[5] Retrieval Augmented Generation

پاسخ برای پرسش‌ها است، این انجمن‌ها می‌توانند با مشکل عدم پویایی و سرعت در پاسخ مواجه باشند که پس از مدتی موجب دل‌سرد شدن کاربر نسبت به دریافت پاسخ و کاهش فعالیت انجمن می‌شود.

هدف از مسیریابی پرسش[۶]، هدایت‌کردن پرسش جدید به خبره یا کاربران کاندیدا با دانش مورد نیاز پرسش است که بیشترین احتمال ارسال پاسخ را دارد [۱۲]. به طور خاص سامانه‌های پرسش و پاسخ می‌توانند با بررسی تاریخچه‌ی پاسخ‌ها و امتیازات پاسخ‌های هر کاربر و ایجاد یک معیار شباهت بین پرسش جدید با تاریخچه‌ی گذشته‌ی خبرگان، کاندیداها یا خبره‌ی مورد نظر خود را انتخاب کنند و پیشنهاد دهند.

همچنین در سال‌های اخیر با افزایش داده‌ها و توسعه‌ی شبکه‌های عصبی عمیق توجه زیادی به استفاده از شبکه‌های عصبی عمیق در بازیابی خبره و رتبه‌بندی و پیمایش کاندیدای خبره با استفاده از شبکه‌های عصبی مانند شبکه‌های عصبی کدگذار[۷]، شبکه‌های عصبی پیچشی[۸] و مدل‌های انتشاری گرافی[۹] شده است [۱۳-۱۵].

اهداف:

باتوجه‌به پیشرفت روزافزون مدل‌های زبانی بزرگ[۱۰]، هدف از این پژوهش پیداکردن چارچوبی هدفمند برای استفاده از مدل‌های زبانی بزرگ و کاربرد آن‌ها برای خبره‌یابی یا هدایت پرسش به کاندیدای خبره است. باتوجه‌به بیشینه‌ی مطرح شده مجموعه‌ای از سؤالات برای کاربرد مدل‌های زبانی بزرگ مطرح می‌شود:

- آیا امکان استخراج ویژگی با استفاده از مدل‌های زبانی بزرگ برای هر کاربر وجود دارد؟
- آیا برای پیداکردن علایق کاربران یا خبرگان، مدل‌های زبانی بزرگ می‌توانند کارساز باشند؟
- آیا مدل‌های زبانی امکان بسط‌دادن علایق و دانش کاربرانی با تاریخچه‌ی محدود و کوتاه را دارند؟
- آیا برای بازیابی و رتبه‌بندی خبرگان کاندیدای هر سؤال می‌توان از مدل‌های زبانی بزرگ استفاده کرد؟

سامانه‌های مکالمه‌ای جویای اطلاعات

طراحی و ساخت سامانه‌هایی که قابلیت گفتگو با انسان را دارند از چالش‌های اصلی علم هوش مصنوعی است و چنین سامانه‌هایی تکامل بعدی رابط‌های کامپیوتر-انسانی هستند. دسترسی به مجموعه داده‌های مکالمه‌ای بزرگ و پیشرفت‌های اخیر در یادگیری عمیق پیاده‌سازی سامانه‌های مکالمه‌ای مبتنی بر داده را امکان‌پذیر کرده‌اند. سامانه مکالمه‌ای یا سامانه گفتگو، به سامانه‌های هوشمندی گفته می‌شود که با استفاده از واسطهای کاربری مختلف مانند صوت یا متن و در قالب مکالماتی به زبان طبیعی با انسان تعامل داشته باشد، به گونه‌ای که سامانه در هر زمان با توجه به تاریخچه تعاملات کاربر با سامانه تا آن لحظه و درخواست فعلی کاربر، مناسب‌ترین پاسخ را به کاربر ارائه دهد. اگرچه با وجود پیشرفت‌های روز افزون در حوزه یادگیری عمیق و بهره‌گیری از آن در سامانه‌های پردازش زبان طبیعی طراحی این سامانه‌ها نیز رشد

[6] Question Routing
[7] Encoder Neural Networks
[8] Convolutional Neural Networks
[9] Graph Diffusion Models
[10] Large language models

چشمگیری داشته است، با این حال همچنان توانایی این سامانه‌ها در ایجاد یک مکالمه پیوسته و روان با انسان با محدودیت‌های زیادی مواجه است. بنابراین برقراری یک سامانه مکالمه‌ای خودکار بین انسان و رایانه یکی از مسائل مهم و قابل توجه در علوم کامپیوتر به خصوص هوش مصنوعی است که روز به روز بر اهمیت آن افزوده می‌شود.

سامانه‌های مکالمه‌ای را با توجه به هدفی که برای آن طراحی شده‌اند، می‌توان به سه دسته تقسیم کرد: (۱) سامانه‌های پرسش و پاسخ مکالمه‌ای، (۲) سامانه‌های مبتنی بر وظیفه و (۳) سامانه‌های غیر وظیفه‌گرا یا همان چت‌بات‌ها. در سامانه‌های پرسش و پاسخ کاربران می‌توانند درخواست‌های خود را به صورت سوالاتی به زبان طبیعی از یک پایگاه دانش بزرگ یا مجموعه‌ای از اسناد مشخص مطرح کنند و سامانه در هر مرحله با توجه به مجموعه سوال‌ها و جواب‌های قبلی و همچنین منبع دانش مورد نظر به سوال کاربر پاسخ دهد. حالت ساده‌تر این نوع از سامانه‌ها، همان ماشین‌های درک مطلب هستند که با توجه به متنی که در اختیار سامانه قرار گرفته است قادرند به سوالات مرتبط با آن متن پاسخ دهند. در این نوع از سامانه‌ها هر سوال کاربر به طور مجزا و مستقل از پرسش و پاسخ‌های قبلی، توسط سامانه پردازش شده و پاسخ مرتبط از روی متن استخراج می‌شود. دسته دوم، سامانه‌های مبتنی بر وظیفه هستند که در آن سامانه به کاربر کمک می‌کند تا بتواند وظیفه مشخصی را انجام دهد. برای مثال پیدا کردن یک محصول خاص و خرید اینترنتی آن، رزرو هتل یا رستوران، خرید بلیط هواپیما و سفارش غذا نمونه‌هایی از وظایفی است که این نوع از سامانه‌ها کاربران را در انجام آن راهنمایی می‌کنند. سامانه‌های غیر وظیفه گرا یا همان چت‌بات‌ها، با هدف پشبرد مکالمه با کاربر با ارائه پاسخ‌هایی منطقی، جملات خبری، جملات احساسی، طرح سوالات و یا درخواست اطلاعات بیشتر از کاربر طراحی می‌شوند. در این سامانه‌ها معمولا مکالمات بر روی دامنه باز بوده و موضوعات مطرح شده در آن‌ها محدوده مشخصی ندارد، مانند توییتر و Weibo. در این میان نوع خاصی از مکالمات هستند که در آن کاربران به دنبال اطلاعات خاصی هستند و تمام مکالمات انجام شده در آن در راستای رفع نیاز اطلاعاتی کاربر است. در این نوع از سامانه‌ها که مبتنی بر اطلاعات هستند دستیار مکالمه‌ای در صورتی که سوال کاربر نامفهوم باشد با طرح سوالاتی سعی می‌کند اطلاعات بیشتری از کاربر جمع‌آوری نماید. سپس با دریافت پاسخ‌های کاربر و تحلیل آن‌ها در نهایت اطلاعات مورد نیاز کاربر را در اختیارش قرار می‌دهد. به این نوع از سامانه‌ها، سامانه‌های مکالمه‌ای جستجوی اطلاعات گفته می‌شود. در این پژوهش تمرکز بر روی طراحی این نوع از سامانه‌های مکالمه‌ای است.

یکی از ویژگی‌های بارز مکالمه حافظه‌دار بودن آن است، به این معنی که می‌توان به گفته‌های قبلی ارجاع کرد [۱۴]. یکی از بزرگ‌ترین کاستی‌های سامانه‌های موجود در صنعت عدم توانایی آن‌ها در نگهداری بافتار مکالمه است [۱۵]. یک سامانه‌ی کارآمد باید بتواند از گفته‌های قبلی کاربر در رفع نیاز اطلاعاتی کاربر استفاده کند و کاربر باید این قابلیت را داشته باشد که بتواند به گفته‌های قبلی خود یا سامانه ارجاع دهد. برای بهره‌گیری از اطلاعات موجود در این بافتار باید روشی برای مدل‌سازی آن در نظر گرفته شود.

مدل‌سازی بافتار در زمینه‌های مختلفی از بازیابی اطلاعات و پردازش زبان طبیعی کاربرد دارند. یک نمونه در چت‌بات‌های جویش اطلاعات مانند چت‌بات پشتیبانی مشتری است. برای آموزش این چت‌بات‌ها از آرشیو بزرگ مکالمه‌های قبلی بین دو انسان استفاده می‌شود و چت‌بات می‌تواند برای پاسخ کاربر از این آرشیو یک پاسخ مناسب انتخاب کند یا یک جمله‌ی

جدید تولید کند [۱۶]. یک نمونه‌ی دیگر از کاربردهای مدل‌سازی بافتار در سامانه‌های سوال و جواب چند نوبتی هستند. در این سامانه‌ها به کاربر قابلیت سوال پرسیدن راجع به یک متن یا یک گراف دانش داده می‌شود و سامانه باید به کاربر پاسخی به زبان طبیعی برگرداند [۱۷].

سامانه‌های مکالمه‌ای را با توجه به روشی که برای ارائه پاسخ استفاده می‌کنند می‌توان به طور کلی به دو دسته مبتنی بر تولید پاسخ و مبتنی بر بازیابی پاسخ تقسیم نمود. در مدل مبتنی بر تولید، سامانه با تحلیل مکالمات قبلی کاربر با رایانه سعی بر تولید یک پاسخ مناسب با درخواست فعلی کاربر را دارد. این در حالی است که در سامانه‌های مبتنی بر بازیابی، بهترین پاسخ از میان مجموعه‌ای از پاسخ‌های کاندید موجود انتخاب می‌شود. اخیرا مطالعات زیادی در این حوزه صورت گرفته است [۱۸-۲۱] که در آن‌ها تاثیر روش‌های یادگیری عمیق در بهبود کارایی این سامانه‌ها بررسی شده است.

➢ تشخیص پیام‌های توهین‌آمیز در فضای بین زبانی با مدل‌سازی گرافی

تولید محتوای توهین‌آمیز[11] که در سال‌های اخیر در ارتباطات برخط افزایش یافته است، جرم محسوب می‌شود. عوامل مختلفی در گسترش جملات تنفرآمیز[12] و توهین‌آمیز وجود دارد. از یک طرف در اینترنت و به طور خاص شبکه‌های مجازی، به علت ماهیت مخفی که این محیط‌ها ایجاد می‌کنند، مردم رفتار خشونت‌آمیز بیشتری از خود نشان می‌دهند. از طرف دیگر، مردم تمایل بیشتری به بیان عقایدشان به صورت برخط دارند، در نتیجه انتشار محتوای توهین‌آمیز تسهیل می‌شود. از آنجا که این نوع تعاملات متعصبانه در بستر شبکه‌های اجتماعی می‌تواند برای جامعه مضر باشد، سکوهای شبکه اجتماعی می‌توانند از ابزارهای تشخیص و پیش‌گیری در این مورد بهره ببرند.

محتوای تنفرآمیز پدیده پیچیده‌ای است که به طور ذاتی مربوط به روابط درون‌گروهی می‌شود و با ظرافت‌های زبان ارتباط دارد. این پدیده از دید شبکه‌های اجتماعی مختلف، تعاریف تا حدی نزدیک و البته متفاوتی دارد. در یک تعریف جامع، محتوایی تنفرآمیز قلمداد می‌شود که بیان تهاجمی یا هدف تحقیر داشته باشد به طوری که خشونت و نفرت را ضد گروه‌ها بر اساس ویژگی‌های خاص نظیر ظاهر فیزیکی، مذهب، ملیت، گرایش جنسی، هویت جنسی و سایر موارد که ممکن است با زبان متفاوتی حتی به شکل ظریفی با طنز بیان شود، گسترش می‌دهد. تصمیم‌گیری در مورد اینکه آیا بخشی از متن حاوی محتوای تنفرآمیز است، حتی برای انسان‌ها نیز ساده نیست. در نتیجه نیازمند ارائه راه‌کارهایی برای بهبود دقت روش‌های تشخیص این گونه محتواها هستیم [۲۲].

در کارهای پژوهشی پیشین، مفاهیم تا حدی مرتبط با این موضوع تحت عنوان آزار و اذیت سایبری[13]، زبان سوء استفاده[14]، تبعیض[15]، دشنام[16]، و افراطی‌گرایی[17] آورده شده است [۲۳, ۲۴]. هر کدام از این مفاهیم از جهتی به پیام‌های توهین‌آمیز و تنفرآمیز شبیه و از جهت خاصی نیز متفاوتند. بررسی هر کدام از مفاهیم مرتبط و چالش‌های مربوط به دسته‌بندی آن‌ها،

[11] Offensive
[12] Hate speech
[13] Cyberbullying
[14] Abusive language
[15] Discrimination
[16] Profanity
[17] Extermism

دقت مدل‌ها را افزایش می‌دهد. همچنین تشخیص موارد فوق در فضایی که کاربران به چندین زبان پیام ارسال می‌کنند نیز از چالش‌های دیگر موجود در این حوزه است.

در اکثر روش‌های ارائه شده برای تشخیص خودکار سخنان نفرت‌انگیز و جملات توهین‌آمیز به بررسی مساله در یک زبان و با استفاده از ویژگی‌های زبانی پرداخته شده است. عمده روش‌های ارائه شده از یک روش یادگیری ماشین برای استخراج و دسته‌بندی محتوای توهین‌آمیز استفاده کرده‌اند. در حالی که تحقیقات در این زمینه به سرعت در حال افزایش است، یکی از چالش‌های مهم آن است که بیشتر روش‌ها برای چند زبان غنی از منابع پیشنهاد شده‌اند. از طرفی در چند پژوهش سعی شده است تا با استفاده از روش‌های یادگیری انتقالی، دانش مدل‌های آموزش شده در زبان‌های غنی از منابع را به زبان‌های کم‌منابع تعمیم دهند. حال آنکه روش‌های یادگیری انتقالی لزوما در همه موارد باعث بهبود دقت طبقه‌بندی بر روی داده زبان کم‌منبع نبوده‌اند. در این پیشنهاد پژوهشی، روش‌های مبتنی بر مدل‌سازی داده با گراف چندلایه و ناهمگون و سپس یادگیری بازنمایی از روی ساختار گراف ارائه شده است. در یک رویکرد داده‌های هر زبان به طور مستقل توسط گراف مدل‌سازی شده و در یک چارچوب مبتنی بر بازسازی یال‌های درون‌لایه‌ای و میان‌لایه‌ای، سعی در بهبود یادگیری بازنمایی‌ها شده است. در هر لایه گراف، یال‌هایی از جنس ارتباط بین کلمه و پیام، کلمه و کلمه و هشتگ و پیام در نظر گرفته شده است. ارتباطات بین‌لایه‌ای از طریق اتصال میان کلمات توهین‌آمیز معادل، در زبان‌های مختلف ایجاد می‌شوند. جهت غنی‌تر کردن ارتباطات بین لایه‌ای، استفاده از یک پایگاه دانش، به عنوان پیشنهاد مطرح شده است. در رویکردی دیگر، پیشنهاد یادگیری یال‌های میان‌لایه‌ای ارائه شده است تا به نوعی ساختار گراف روی هر مجموعه داده را کامل‌تر کند. به منظور ارزیابی روش‌های پیشنهادی، از چندین مجموعه داده در زبان‌های غنی از منابع و کم‌منابع استفاده شده است. نتایج اولیه آزمایش‌ها نشان می‌دهد که روش‌های پیشنهادی، دقت طبقه‌بندی پیام‌های توهین‌آمیز در زبان کم‌منبع را افزایش داده‌اند.

➢ مدل‌سازی کاربران برای پیش‌بینی واکنش در شبکه‌های اجتماعی با کمک مدل‌های زبانی بزرگ

شبکه‌های اجتماعی برخط، بستری برای برقراری ارتباط، تعامل و تولید محتوای کاربران هستند. به دنبال رشد سریع فناوری و دسترسی آسان به اینترنت، حضور و فعالیت کاربران در شبکه‌های اجتماعی رشد چشم‌گیری داشته است. کاربران با اهداف مختلفی چون تعامل و برقراری ارتباط با سایر کاربران، تولید محتوا، بیان احساسات و نظرات، پیگیری اخبار و کاوش اطلاعات جدید در موضوعات مختلف از شبکه‌های اجتماعی بهره می‌برند. به دست آوردن مدلی مناسب برای فهم و پیش‌بینی رفتار کاربران، اهمیت بسیاری در تحلیل‌ها و وظایف مختلف پژوهشی شبکه‌های اجتماعی دارد و حجم بالای فعالیت و داده‌های تولید شده، چالش‌ها و فرصت‌های فراوانی را در این زمینه به دنبال داشته است [۲۵].

به طور کلی پژوهش‌های حوزه رفتاری شبکه‌های اجتماعی را می‌توان به دو دسته تقسیم کرد؛ رفتار انفرادی کاربر و رفتار دسته جمعی کاربران به عنوان یک شبکه که در پی تعامل و ارتباطات کاربران پدید می‌آید. به دست آوردن مدلی کارا از کاربر که بتواند با توجه به مواردی مثل شرایط شبکه، سابقه کاربر، ویژگی‌ها و ترجیحات شخصی او، رفتار و واکنش‌هایش را پیش‌بینی

کند از اهمیت بالایی در مسائل متعدد برخوردار است. برخی از این مسائل عبارتند از پیش‌بینی الگوی پخش اخبار در شبکه، شناسایی و جلوگیری از پخش اخبار جعلی، بهبود سامانه‌های پیشنهاد دهنده[18]، تبلیغات بهینه و بازاریابی [۲۶].

مدل‌های زبانی بزرگ[19] در سال‌های اخیر پیشرفت شگفت‌انگیزی داشته‌اند و در مسائل مختلفی مانند سامانه‌های پیشنهاد دهنده، ویژگی‌های قدرتمند و بی‌سابقه‌ای در استدلال، استنتاج و یادگیری مفاهیم پیچیده و سطح بالا از خود نشان داده‌اند. لازم به ذکر است که بسیاری از این کاربردها، نیازی به آموزش مدل با داده‌های ویژه آن وظیفه ندارند. توانایی‌های این مدل‌ها در استدلال علّی، منطقی و قیاسی فرصت‌های جدیدی را برای یادگیری مفاهیم پیچیده و نهان در ساختارهای گرافی و هم‌چنین سازوکارهای شناختی انسان ایجاد کرده است [۵۱-۵۲].

اقدامات انجام شده:

مجموعه داده به طور کامل در دو زبان فارسی و انگلیسی جمع آوری شده است. موضوع مجموعه‌های داده انتخابات ۱۴۰۳ ایران و در نسخه انگلیسی، انتخابات ۲۰۲۴ آمریکا است. روی مجموعه داده فارسی یک مقاله با موضوع به دست آوردن پروفایل مناسب از کاربر آماده شده و در کنفرانس ICWSM 2025 تحت بررسی است. هم‌چنین مجموعه داده انگلیسی انتخابات آمریکا نیز در کنفرانس ICWSM 2025 ثبت شده و تحت بررسی است. به طور موازی و در راستای فعالیت در حوزه علوم اجتماعی محاسباتی و اهمیت مبحث فرهنگ در شبکه‌های اجتماعی، یک مقاله با این موضوع آماده شد و در NAACL 2025 قبول شده است. در قدم‌های بعدی از پروفایل‌های تولید شده استفاده شده و وظیفه پیش بینی واکنش کاربر در شبکه‌های اجتماعی به کمک مدل‌های زبانی بزرگ انجام خواهد شد.

➢ خبره‌یابی

مسئله‌ی خبره‌یابی تا به‌حال در زمینه‌های مختلفی همچون پیشنهاد دکتر به بیماران [۲۷]، یافتن استاد راهنمای مناسب برای دانشجویان [۲۸] و مسیریابی سوال [۲۹] مورد استفاده قرار گرفته است. اما یکی از مهم‌ترین کاربردهای آن انتخاب افراد مناسب از بین مجموعه‌ای از داوطلبان برای یک موقعیت شغلی [۳۰] داده شده می‌باشد. در اینجا به‌طور کلی هدف این است که با در اختیار داشتن لیستی از حوزه‌های مورد نیاز یک موقعیت شغلی، یک رتبه‌بندی از داوطلبان بر اساس میزان دانششان در حوزه‌های ذکر شده تولید شود. تا به حال تحقیقات زیادی در این زمینه انجام شده است [۳۱-۳۲]، اما بسیاری از آنها قادر به برآورده کردن نیازهای اصلی شرکت‌ها نیستند و بدون توجه به نیازهای اصلی آنها اقدام به بازیابی افراد خبره می‌نمایند.

امروزه بسیاری از شرکت‌ها به دنبال استخدام افراد خبره‌ای هستند که بتوانند بر اساس متدلوژی‌های توسعه‌ی نرم‌افزار همچون متدلوژی چابک در قالب یک تیم بر روی یک پروژه داده شده فعالیت نمایند. در یک تیم ایده‌آل که مبتنی بر متدلوژی چابک است، تعدادی افراد خبره‌ی تی-شکل وجود دارد، به‌طوریکه هر یک در یک حوزه‌ی مورد نیاز تیم خبره و در سایر حوزه‌های مورد نیاز دانش عمومی دارد [۳۳]. اخیرا تعدادی تحقیق در این زمینه انجام شده است. در این

[18] Recommender Systems
[19] Large Language Models

تحقیقات، هدف اصلی، یافتن یک رتبه‌بندی از افراد خبره‌ی تی-شکل است که در یک حوزه‌ی خاص خبره باشند [۳۶-۳۴]. در اینجا به منظور ساده‌سازی، توجهی به دانش عمومی کاربران نشده و مسئله‌ی بازیابی تیم‌های چابک نادیده گرفته شده است. رستمی و نشاطی [۳۷] برای اولین بار به مسئله‌ی تشکیل یک تیم چابک با بازیابی افراد خبره‌ی تی-شکل پرداختند. آنها با استفاده از مدل‌های احتمالاتی مولد که مبتنی بر تعداد اسناد داوطلبان در حوزه‌های مورد نیاز تیم بودند اقدام به بازیابی تیم‌های چابک نمودند. با این حال، آنها هیچ توجه‌ای به محتوی اسناد داوطلبان نداشتند و دانش داوطلبان را بدون توجه به محتوی اسنادشان تخمین زدند.

رویکرد ما برای حل این مسئله به این صورت است که ابتدا با استفاده از یک روش پیشنهادی، بهترین داوطلبان ممکن را برای هرکدام از جایگاه‌های تیم انتخاب می‌کنیم. سپس با استفاده از یک مدل برنامه‌ریزی خطی عدد صحیح، بهترین تیم چابک ممکن را از بین داوطلبان انتخاب شده تشکیل می‌دهیم. برای آنکه بتوان بهترین داوطلبان ممکن را برای هر جایگاه از تیم پیش‌بینی کرد، مدل پیشنهادی ما بایستی بتواند به طور همزمان از دانش حاصل از شکل خبرگی و سطح دانش تخصصی و عمومی هر داوطلب در حوزه های مهارتی مورد نیاز جایگاه، برای تخمین مرتبط بودن او به آن جایگاه استفاده کند. ما تا به حال با دو رویکرد به طراحی چنین مدلی پرداخته‌ایم:

رویکرد اول-مدل رمزگذار چندگانه: در اینجا از یک مدل رمزگذار استفاده می‌شود که با یک رویکرد یادگیری با نظارت به تخمین احتمال مرتبط بودن هر داوطلب می‌پردازد. این مدل در واقع از ترکیب بازنمایی‌های برداری حاصل از تعدادی رمزگذار، برای تخمین احتمال ذکرشده استفاده می‌کند. در اینجا، از یک رمزگذار شکل خبرگی برای تولید بازنمایی مربوط به شکل خبرگی داوطلب و از تعدادی رمزگذار سطح دانش برای تولید بازنمایی‌هایی که نشان دهنده‌ی سطح دانش تخصصی/عمومی داوطلب در حوزه‌های مهارتی مورد نیاز باشد، استفاده می‌شود. برای طراحی رمزگذار شکل خبرگی، ما از یک شبکه عصبی بازگشتی که از فرآیند توزیع تاپیک‌های اسناد داوطلب در گذر زمان بهره می‌برد، استفاده کرده‌ایم تا بتوانیم پروفایل داوطلب را به‌صورت یک بازنمایی برداری تولید کنیم. همچنین برای طراحی رمزگذارهای سطح دانش، بازنمایی برداری پروفایل داوطلب را با بازنمایی مربوط به هر حوزه‌ی مهارتی مورد نیاز تیم مقایسه کرده‌ایم تا دانش داوطلب در هر حوزه را به‌صورت یک بازنمایی برداری تخمین بزنیم.

رویکرد دوم- مدل احتمالاتی: در اینجا با یک رویکرد یادگیری بدون نظارت به تخمین احتمال مرتبط بودن هر داوطلب می‌پردازیم. این مدل در واقع با الهام‌گیری از مشخصه‌های اصلی یک تیم چابک، احتمال مرتبط بودن داوطلب به یک جایگاه از تیم را بسته به برقراری همزمان سه شرط پوشش، ارتباط و بهینگی برای حضور در آن جایگاه می داند. در اینجا، برای تخمین احتمال برقراری شروط پوشش و ارتباط، از یک رمزگذار سطح دانش و برای تخمین احتمال برقراری شرط بهینگی، از یک رمزگذار شکلی خبرگی استفاده می‌شود. ما در این رویکرد، برای طراحی رمزگذار سطح دانش، از یک رویکرد مبتنی بر توجه استفاده کرده‌ایم که با بررسی میزان توجه یک حوزه‌ی مهارتی مورد نیاز تیم به پروفایل داوطلب، یک بازنمایی شخصی‌سازی شده از حوزه‌ی مهارتی را ایجاد می‌کند. سپس با مقایسه‌ی این بازنمایی با پروفایل داوطلب، ما به تخمین سطح دانش داوطلب در آن حوزه می‌پردازیم. همچنین برای طراحی رمزگذار شکل خبرگی، ما

ویژگی‌های شکل خبرگی داوطلب را با تحلیل مجموعه بازنمایی‌های برداری مربوط به دانش داوطلب در حوزه‌های مهارتی یاد می‌گیریم، سپس با داشتن این ویژگی‌ها، توزیع شکل خبرگی داوطلب را تخمین می‌زنیم.

➢ تشخیص تشخیص سخنان تنفرآمیز با در نظر گرفتن نمایه و تعاملات بین کاربران

این پژوهش با هدف تشخیص سخنان تنفرآمیز در محیط‌های شبکه‌های اجتماعی و با در نظر گرفتن تعاملات بین کاربران (مانند دنبال کردن و دنبال شدن) و همچنین ساخت پروفایل کاربران بر اساس ویژگی‌های استخراج‌شده از متون منتشرشده قبلی انجام شده است. در ادامه، فعالیت‌های انجام‌شده در این پژوهش به‌صورت سوم شخص ارائه می‌شود.

در طول سال جاری، پژوهشگر اقدام به جمع‌آوری و برچسب‌زنی یک مجموعه‌داده فارسی‌زبان در حوزه تشخیص سخنان تنفرآمیز کرده است. با توجه به اینکه این مجموعه‌داده هنوز در مرحله آماده‌سازی نهایی برای انجام آزمایش‌ها قرار دارد، پژوهشگر به‌موازات آن، تحقیقاتی را بر روی یک مجموعه‌داده انگلیسی‌زبان مبتنی بر شبکه‌های اجتماعی انجام داده است. این تحقیقات شامل نمونه‌برداری از مجموعه‌داده، بررسی روش‌های شناسایی کامیونیتی‌ها در شبکه‌های اجتماعی و انجام تحلیل احساسات[۲۰] بر روی متون منتشرشده توسط کاربران بوده است.

همچنین، مطالعاتی را در زمینه مهندسی پیش‌الگو[۲۱] برای بهبود فرآیند برچسب‌زنی مجموعه‌داده انجام شده است که نتایج آن می‌تواند در مراحل بعدی پژوهش مورد استفاده قرار گیرد. علاوه بر این، پژوهشگر به بررسی معماری‌های پیشرفته‌ای مانند RAG[22] پرداخته که قرار است پس از تکمیل فرآیند برچسب‌زنی مجموعه‌داده فارسی‌زبان (که پیش‌بینی می‌شود به زودی به اتمام برسد)، بر روی مدل پیاده‌سازی شود.

در مجموع، این پژوهش با ترکیب روش‌های تحلیل شبکه‌های اجتماعی، استخراج ویژگی‌های متنی و استفاده از معماری‌های پیشرفته یادگیری ماشین، در تلاش است تا راه‌حلی کارآمد برای تشخیص سخنان تنفرآمیز در محیط‌های پیچیده شبکه‌های اجتماعی ارائه دهد.

[20] Sentiment Analysis
[21] Prompt Engineering
[22] Retrieval-Augmented Generation

جست‌وجوی مکالمه‌ای با تمرکز بر شفاف‌سازی پرس‌وجوی کاربر و شبیه‌سازی کاربر

۱- سیستم‌های جست‌وجوی مکالمه‌ای و پرسش‌های شفاف‌ساز

کاربران در هنگام ارسال پرس‌وجوی خود به موتورهای جست‌وجوی اطلاعات، به طور واضح نیاز اطلاعاتی خود را بیان نمی‌کنند و یا نمی‌دانند چگونه آن را به درستی بیان کنند. این امر موجب ایجاد پرس‌وجوهای مبهم یا چندوجهی می‌شود و هدف و نیاز اطلاعاتی کاربر مبهم باقی می‌ماند. موتورهای جست‌وجو تلاش می‌کنند تا با متنوع سازی نتایج جست‌وجو با این مسئله مقابله کنند و تمامی جنبه‌های ممکن یک پرس‌وجوی مبهم را در لیستی از نتایج ارائه دهند. به این ترتیب تا حدی مسئله‌ی پرس‌وجوهای مبهم برطرف می‌شود.

در سیستم‌های جست‌وجویی که دارای صفحه نمایش کوچکی هستند و یا در دستیاران صوتی نظیر Siri، Alexa و Google Assistan ، استفاده از صفحاتی که نتایج متنوع جست‌وجو را لیست می‌کنند، با مشکل مواجه می‌شود؛ زیرا در این سیستم‌ها کاربران نمی‌توانند به آسانی لیست نتایج مرتبط را دنبال کنند و نتیجه‌ی مورد نظر خود را پیدا کنند. یکی از راهکارهایی که برای رسیدگی به این مسئله وجود دارد استفاده از پرسش‌های شفاف‌ساز است. در این رویکرد، با بهره‌گیری از پرسش‌های شفاف‌ساز، اطلاعات بیشتری از کاربر به‌دست می‌آید و از این اطلاعات در جهت رفع ابهام پرس‌وجوی او استفاده می‌شود.

مسئله‌ی پرس‌وجوهای مبهم و استفاده از پرسش‌های شفاف‌ساز برای رسیدگی به آن، موجب شده است که سیستم‌های جست‌وجو به سمت مکالمه‌ای شدن سوق پیدا کنند. در سیستم‌های جست‌وجوی مکالمه‌ای، سیستم در یک تعامل ترکیبی با استفاده از پرسش‌های شفاف‌ساز در چندنوبت تلاش می‌کند تا هدف کاربر را مشخص کند و در نتیجه ابهام موجود در پرس‌وجوی او را برطرف نماید. علاوه بر مزایای کارکردی که جست‌وجوی مکالمه‌ای دارد، مطالعات نشان داده است که کاربران از پاسخ به پرسش‌های شفاف‌ساز در طی یک مکالمه لذت می‌برند و این امر موجب افزایش رضایت و اطمینان آن‌ها می‌شود [۴۵].

پرسش‌های شفاف‌ساز منجر به مشخص شدن هدف کاربر شده و در نهایت منجر به بازیابی مرتبط‌ترین نتایج منطبق با نیاز اطلاعاتی کاربر، بجای نمایش لیستی از نتایج متنوع می‌شوند. اما با سه چالش اساسی مواجه هستندکه ما در طی پژوهش خود به بررسی این چالش‌ها می‌پردازیم:

۱- یکی از بزرگترین چالش‌های تحقیق در این زمینه، کمبود مجموعه داده‌های با کیفیت است.

۲- ارزیابی پرسش‌های شفاف‌ساز در سیستم‌های جست‌وجوی مکالمه‌ای نیز با چالش روبرو است. چراکه مجموعه داده‌ها در این زمینه کم است و ارزیابی انسانی نیز پرهزینه است.

۳- کیفیت پرسش‌های شفاف‌ساز بر فرآیند بازیابی نتایج مرتبط و همچنین بر رفتار جست‌وجوی کاربران، رضایت و توانایی آن‌ها در یافتن نتایج مرتبط تأثیرگذار است. بررسی‌ها نشان می‌دهد که سیستم‌ها باید تنها زمانی که پرسش‌های شفاف‌ساز از کیفیت خوبی برخوردار هستند، از آن‌ها استفاده کنند؛ در این صورت پرسش‌های شفاف‌ساز منجر به بهبود عملکرد کاربران و افزایش رضایت آن‌ها می‌شوند. باید توجه کرد که پرسیدن پرسش‌های شفاف‌ساز با کیفیت پایین منجر به تأثیرات منفی می‌شود. به طور خلاصه، عدم استفاده از پرسش‌های شفاف‌ساز بهتر از استفاده از پرسش‌های با کیفیت پایین است، زیرا این امر می‌تواند تأثیر منفی بر رضایت و عملکرد کاربران داشته باشد [۴۶].

۲- مجموعه داده‌های مورد بررسی در سیستم‌های جست و جوی مکالمه‌ای

همانطوری که بیان کردیم یکی از چالش‌های مهم در این پژوهش مجموعه داده‌های مناسب می‌باشد. در ادامه دو مجموعه داده‌ی مهم در این زمینه‌ی تحقیقاتی که ما مورد بررسی قرار دادیم را بیان می‌کنیم:

مجموعه داده‌ی ClariQ

این مجموعه داده بر پایه‌ی مجموعه داده‌ی TREC از سال ۲۰۰۹ تا ۲۰۱۴ است و در مقاله [۴۷] مورد بررسی قرار گرفته است. در این مجموعه داده این مسئله که "چه زمانی" نیاز است از پرسش‌های شفاف‌ساز استفاده شود نیز مورد بررسی قرار گرفته است. در این مجموعه داده از پرسش‌های شفاف‌ساز برای روشن شدن نیاز اطلاعاتی کاربر استفاده شده است اما با بررسی‌هایی که ما در طی پژوهش انجام دادیم متوجه شدیم که پرسش‌های اولیه در این مجموعه داده مناسب مکالمه‌ای شدن روش بازیابی اطلاعات نیستند و بنابراین ما در حال حاضر این مجموعه داده را کنار گذاشته‌ایم.

مجموعه داده‌ی TREC iKAT 2023

مجموعه داده‌ی TREC iKAT 2023، در مقاله‌های [۴۹] مورد بررسی قرار گرفته است. این مجموعه داده، مجموعه داده‌ی نسبتا کوچکی است اما بسیار ارزشمند است و گامی جدید در زمینه‌ی جست‌وجوی مکالمه‌ای می‌باشد. ارائه دهندگان این مجموعه داده معتقد هستند که در سیستم‌های جست و جوی مکالمه‌ای، کاربرانی که ویژگی‌های متفاوتی دارند، باید پاسخ های متفاوتی بر اساس ویژگی‌های خاصی خود دریافت کنند. به عنوان مثال فرض کنید پرس‌وجوی اولیه‌ی کاربر مربوط به جایگزین‌های شیر گاو است، سه کاربر با سه ویژگی متفاوت می‌توانند این پرس‌وجو را ارسال کنند:

۱- آلیس، یک گیاهخوار سرسخت، به دنبال جایگزینی سالم و غیرحیوانی است.

۲- باب، یک طرفدار محیط زیست، به دنبال جایگزینی با کلسیم بالا و کمترین آسیب به محیط زیست است.

۳- چارلی، که به تازگی به دیابت مبتلا شده، به دنبال جایگزینی کم‌قند است.

ویژگی‌های خاص هر کدام از این سه کاربر بر روند مکالمه و نتایج جست‌وجو تأثیر می‌گذارد و پاسخ‌های سیستم جست‌وجوی مکالمه‌ای به هر یک از این افراد، با توجه به زمینه شخصی و سابقه مکالمه آن‌ها باید متفاوت باشد. برای بررسی این موضوع در مجموعه داده‌ی TREC iKAT 2023 در کنار مکالمه‌ها، ویژگی‌های خاص هر کاربر به صورت پایگاه دانش متنی شخصی (PTKB) نیز قرار داده شده است. این پایگاه داده شامل ویژگی‌های خاص هر فرد است که با توجه به آن‌ها روند مکالمه و نتایج بازیابی برای هر فرد متفاوت است. این مجموعه داده شامل ۳۶ دور مکالمه بر روی ۲۰ موضوع مختلف است که ۱۱ مکالمه مربوط به داده‌های آموزشی و ۲۵ مکالمه مربوط به داده‌های ارزیابی می‌باشد.

نکته‌ی مهم دیگری که در این مجموعه داده وجود دارد این است که پرس‌وجوهای موجود در این مجموعه داده از نوع جست وجوی تصمیم‌گیری هستند؛ به این معنا که کاربر قصد دارد تا در طی مکالمه با جست‌وجویی که انجام می‌دهد یک تصمیم مناسب اخذ نماید. ما معتقدیم که به دلیل این ویژگی، استفاده از پرسش‌های شفاف‌ساز در این مجموعه داده بیشتر معنا پیدا می‌کند. بنابراین ما در حال حاضر در حال کار بر روی این مجموعه داده هستیم.

۳- ارزیابی سیستم‌های جست‌وجوی مکالمه‌ای با شبیه‌سازی کاربر

ارزیابی سیستم‌های جست‌وجوی مکالمه‌ای نیز به دلیل چالش‌های منحصر به فردی که این زمینه‌ی تحقیقاتی دارد، به یک حوزه تحقیقاتی تبدیل شده است. یکی از روش‌های ارزیابی این سیستم‌ها شبیه‌سازی رفتار کاربر است. در این روش شبیه‌ساز کاربر می‌تواند در طی چند دوره با سیستم جست‌وجو تعامل داشته باشد و به پرسش‌های شفاف‌ساز پاسخ دهد. سیستم نیز با توجه به پاسخ‌های کاربر شبیه‌سازی شده نتایج بازیابی را بازمی‌گرداند و یا از پرسش شفاف‌ساز استفاده می‌کند. بنابراین در این روش در عین حال که می‌توان یک سیستم را به صورت پویا در برابر یک شبیه‌ساز (ثابت) ارزیابی کرد مقیاس‌پذیری و کم‌هزینه بودن بودن نیز وجود دارد [۴۸]. ما در طی این پژوهش تلاش می‌کنیم که با شبیه‌سازی ویژگی‌های کاربر در طی مکالمه (نظیر صبر کاربر، استراتژی توقف کاربر و اثرات پیش‌زمینه‌ای (Priming effects) کاربر [۵۰]) گامی مهم در جهت ارزیابی سیستم‌های جست و جوی مکالمه برداریم.

تشخیص سخنان توهین‌آمیز و تنفرآمیز در شبکه‌های اجتماعی با در نظر گرفتن زمینه‌های موثر بر کاربر

استفاده از سکوهای رسانه‌های اجتماعی برخط تحول بزرگی در نحوه‌ی ارتباط، اشتراک‌گذاری اطلاعات و تعامل میان افراد ایجاد کرده است. این فضاها به بخشی جدایی‌ناپذیر از زندگی روزمره تبدیل شده‌اند و بستری برای بیان دیدگاه‌ها، برقراری ارتباط و شکل‌گیری جوامع مختلف فراهم آورده‌اند. طی دهه‌ی اخیر، استفاده از شبکه‌های اجتماعی با رشد چشمگیری همراه بوده و به یکی از سریع‌ترین ابزارهای ارتباطی بدل شده است، چراکه پیام‌ها تقریباً در لحظه ارسال و دریافت می‌شوند. بااین‌حال، در کنار تمام مزایای این بسترها، نگرانی‌های فزاینده‌ای درباره‌ی گسترش و تأثیر سخنان توهین‌آمیز و تنفرآمیز در این فضاها وجود دارد.

افزایش تعاملات در رسانه‌های اجتماعی، هم‌زمان به افزایش انتشار محتوای توهین‌آمیز و تنفرآمیز منجر شده که چالش‌های جدی برای افراد و جامعه به همراه داشته است. در نتیجه، شناسایی و مقابله‌ی مؤثر با این نوع محتوا اهمیتی دوچندان یافته است. بااین‌حال، تشخیص خودکار سخنان توهین‌آمیز و تنفرآمیز با چالش‌های متعددی روبه‌رو است. پویایی زبان، تغییرات مداوم در شیوه‌ی بیان و تفاوت‌های ظریف زبانی، فرایند شناسایی را پیچیده‌تر می‌کند. برای عبور از این موانع، نیاز به رویکردهای نوآورانه‌ای است که بتوانند ساختار پیچیده و متغیر این نوع محتوا را در بسترهای مختلف تحلیل کنند[۱].

بیشتر پژوهش‌های انجام‌شده در این حوزه، تنها بر تحلیل محتوای متنی متمرکز بوده‌اند و کمتر به تأثیر ساختار تعاملات میان کاربران و عوامل زمینه‌ای پرداخته‌اند. علاوه بر این، عمده‌ی تحقیقات بر روی زبان انگلیسی صورت گرفته و مطالعات کمی به زبان فارسی اختصاص داده شده است. در این پژوهش، رویکردی تازه معرفی می‌شود که بر اهمیت نقش تعاملات کاربران و عوامل زمینه‌ای در تشخیص سخنان توهین‌آمیز و تنفرآمیز تأکید دارد. تحلیل الگوهای ارتباطی کاربران و تأثیر شبکه‌های اجتماعی می‌تواند درک عمیق‌تری از نحوه‌ی انتشار این نوع محتوا ارائه دهد. همچنین، در نظر گرفتن عوامل زمینه‌ای، دقت مدل‌های تشخیصی را بهبود می‌بخشد و امکان تحلیل جامع‌تری از این پدیده را فراهم می‌کند[۲-۳].

در این پژوهش، مجموعه داده‌ای به زبان فارسی در دست تهیه است که علاوه بر محتوای متنی، ساختار تعاملات میان کاربران را نیز لحاظ می‌کند. مدل پیشنهادی بر اساس این مجموعه داده آموزش داده خواهد شد. در مرحله‌ی ارزیابی اولیه، زیرمجموعه‌ای از داده‌های شبکه‌ی اجتماعی پاساژ جمع‌آوری و برچسب‌گذاری شده است. این مجموعه داده شامل متن و تعاملات کاربران بوده و مدل‌های پایه بر روی آن اجرا و بررسی شده‌اند[۴-۵].

برای استخراج مجموعه داده به این صورت عمل شده است به منظور حفظ جامعیت مجموعه داده براساس شبکه ارتباطات بین کاربران، انجمن‌های کاربران شناسایی گردید، از هر انجمن براساس تراکم آن کاربران انتخاب گردید. سپس متون منتشر شده توسط کاربران استخراج شد. با استفاده از مدل‌های زبانی بزرگ بنا داریم برچسب اولیه‌ای به آن نسبت دهیم، سپس برچسب‌ها توسط ناظران انسانی بازبینی و اصلاح خواهد شد تا مجموعه داده‌ای استاندارد ایجاد گردد.

نتایج اولیه‌ی این پژوهش نشان می‌دهد که لحاظ کردن تعاملات میان کاربران، می‌تواند به بهبود عملکرد مدل‌های تشخیص سخنان توهین‌آمیز و تنفرآمیز کمک کند. برای بررسی این فرضیه، آزمایش‌هایی روی مجموعه داده‌ی ایجادشده انجام شده و تأثیر تعاملات کاربران بر عملکرد مدل مورد تأیید قرار گرفته است. همچنین، بررسی‌ها نشان می‌دهد که در نظر گرفتن سوابق فردی کاربران می‌تواند دقت تشخیص را افزایش دهد. نتایج آزمایش‌های انجام‌شده نیز تأثیر این عامل را بر بهبود مدل تأیید کرده‌اند.

در ادامه‌ی پژوهش، جزئیات بیشتری از هر یک از این مؤلفه‌ها مورد بررسی قرار خواهد گرفت و مدل ترکیبی پیشنهادی تکمیل خواهد شد.

➢ شخصی‌سازی و تنوع‌بخشی پیشنهادات در سیستم‌های توصیه‌گر مکالمه‌ای با تمرکز بر ترجیحات کاربر

سیستم‌های توصیه‌گر مکالمه‌ای به عنوان یکی از حوزه‌های مهم در تعاملات انسان و ماشین، نقش اساسی در ارائه پیشنهادات هوشمند و متناسب با نیازهای کاربران ایفا می‌کنند. چالش‌های اساسی این سیستم‌ها شامل تشخیص ترجیحات کوتاه‌مدت و بلندمدت کاربران، ارائه پیشنهادات متنوع و شخصی‌سازی‌شده، و تفسیر شفاف پیشنهادات است. هدف این پژوهش توسعه یک مدل توصیه‌گر مکالمه‌ای است که بتواند نیازهای کاربران را در لحظه شناسایی کند، از داده‌های تعاملی و دانش‌محور برای بهبود دقت و تنوع پیشنهادات بهره ببرد، و توضیحاتی شفاف برای پیشنهادات ارائه دهد.

در راستای این هدف، این پژوهش به طراحی و پیاده‌سازی یک مدل شخصی‌سازی‌شده برای توصیه‌های مکالمه‌ای می‌پردازد که ترکیبی از روش‌های یادگیری عمیق، گراف‌های دانش، و مدل‌های زبانی بزرگ را به کار می‌گیرد. مدل پیشنهادی از داده‌های مکالمه‌ای (مانند ReDial) و منابع دانش خارجی مانند IMDB و DBpedia برای غنی‌سازی اطلاعات استفاده خواهد کرد. معیارهای ارزیابی شامل دقت پیشنهادات، تنوع، قابلیت تفسیر، و همخوانی پیشنهادات با ترجیحات کاربر خواهند بود [۳۸-۴۴].

در حال حاضر، تمرکز اصلی این پژوهش یافتن یک مدل بیس‌لاین مناسب برای مقایسه عملکرد روش‌های پیشنهادی است. برای این منظور، مدل‌های کلاسیک توصیه‌گر مکالمه‌ای بررسی شده و یکی از آن‌ها برای پیاده‌سازی و ارزیابی اولیه انتخاب خواهد شد. این مرحله به ما کمک می‌کند تا عملکرد مدل پیشنهادی را در مقایسه با روش‌های موجود تحلیل کنیم و نقاط ضعف و قوت آن را شناسایی نماییم.

جمع بندی

در این پژوهش، تمرکز بر روی استفاده از شبکه‌های عصبی در چند کاربرد مختلف کاوش داده‌های متنی و بازیابی اطلاعات است. به طور خاص در این پژوهش، سامانه‌های مکالمه‌ای جویای اطلاعات، تشخیص محتوای توهین آمیز در فضای چندزبانه، تشخیص سخنان توهین‌آمیز و تنفرآمیز در شبکه‌های اجتماعی با در نظر گرفتن زمینه‌های موثر بر کاربر، مدل‌سازی کاربران برای پیش‌بینی واکنش در شبکه‌های اجتماعی با کمک مدل‌های زبانی بزرگ، رویکرد چند اظهاره جهت بررسی لزوم پیشنهاد سوال شفاف‌سازی و خبره‌یابی مورد بررسی قرار گرفت.

مراجع

1. F. Alkomah and X. Ma, “A Literature Review of Textual Hate Speech Detection Methods and Datasets,” Information, vol. 13, no. 6, p. 273, May 2022, doi: 10.3390/info13060273.

2 F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, “Resources and benchmark corpora for hate speech detection: a systematic review,” Lang. Resour. Eval., vol. 55, no. 2, pp. 477–523, Jun. 2021, doi: 10.1007/s10579-020-09502-8.

3. S. Hinduja and J. W. Patchin, “Bullying, Cyberbullying, and Suicide,” Arch. Suicide Res., vol. 14, no. 3, pp. 206–221, Jul. 2010, doi: 10.1080/13811118.2010.494133.

4. U. Belavusau, “Fighting Hate Speech through EU Law,” Amst. Law Forum, vol. 4, no. 1, p. 20, Dec. 2012, doi: 10.37974/ALF.208.

5. S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, “Hate speech detection: Challenges and solutions,” PLOS ONE, vol. 14, no. 8, p. e0221152, Aug. 2019, doi: 10.1371/journal.pone.0221152.6. Sun W, Yan L, Ma X, Wang S, Ren P, Chen Z et al. (2023) Is chatgpt good at search? Investigating large language models as re-ranking agents. The 2023 Conference on Empirical Methods in Natural Language Processing. pp

6. Sun W, Yan L, Ma X, Wang S, Ren P, Chen Z et al. (2023) Is chatgpt good at search? Investigating large language models as re-ranking agents. The 2023 Conference on Empirical Methods in Natural Language Processing. pp

7. Wang L, Yang N, Wei F (2023) Query2doc: Query expansion with large language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp 9414-23
8. Zahedi MS, Rahgozar M, Zoroofi RA (2020) Hca: Hierarchical compare aggregate model for question retrieval in community question answering. Information Processing & Management 57(6):102318
9. Zhu Y, Yuan H, Wang S, Liu J, Liu W, Deng C et al. (2023) Large language models for information retrieval: A survey. arXiv preprint arXiv:230807107
10. Tanwar E, Dutta S, Borthakur M, Chakraborty T (2023) Multilingual llms are better cross-lingual in-context learners with alignment. The 61st Annual Meeting Of The Association For Computational Linguistics. pp
11. Qin L, Chen Q, Zhou Y, Chen Z, Li Y, Liao L et al. (2024) Multilingual large language model: A survey of resources, taxonomy and frontiers. arXiv preprint arXiv:240404925

12. T. C. Zhou, M. R. Lyu, and I. King, “A classification-based approach to Question Routing in Community Question Answering,” WWW’12 - Proceedings of the 21st Annual Conference on World Wide Web Companion, pp. 783–790, 2012, doi: 10.1145/2187980.2188201.

13. H. Liu, Z. Lv, Q. Yang, D. Xu, and Q. Peng, “ExpertBert: Pretraining Expert Finding,” International Conference on Information and Knowledge Management, Proceedings, pp. 4244–4248, Oct. 2022, doi: 10.1145/3511808.3557597.

14. J. Wang, J. Sun, H. Lin, H. Dong, and S. Zhang, “Predicting best answerers for new questions: An approach leveraging convolution neural networks in community question answering,” Communications in Computer and Information Science, vol. 669, pp. 29–41, 2016, doi: 10.1007/978-981-10-2993-6_3.

15. V. Krishna and N. Antulov-Fantulin, “Temporal-Weighted Bipartite Graph Model for Sparse Expert Recommendation in Community Question Answering,” UMAP 2023 - Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, pp. 156–163, Jun. 2023, doi: 10.1145/3565472.3592957.

16. Radlinski, F. and N. Craswell. *A theoretical framework for conversational search*. in *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 2017. ACM.

17. Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A.C., Zitouni, I. and Anastasakos, T., *Understanding user satisfaction with intelligent assistants*. in *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. 2016. ACM.

18. Lowe, R.T., Pow, N., Serban, I.V., Charlin, L., Liu, C.W. and Pineau, J., *Training end-to-end dialogue systems with the ubuntu dialogue corpus.* Dialogue & Discourse, 2017. 8(1): p. 31-65.

19. Gao, J., M. Galley, and L. Li. *Neural approaches to conversational AI*. in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 2018. ACM.

20. Chen, H., Liu, X., Yin, D. and Tang, J., *A survey on dialogue systems: Recent advances and new frontiers.* ACM SIGKDD Explorations Newsletter, 2017. 19(2): p. 25-35.

21. Yan, R., Y. Song, and H. Wu. *Learning to respond with deep neural networks for retrieval-based human-computer conversation system*. in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2016. ACM.

22. Fortuna, P. and S. Nunes, *A survey on automatic detection of hate speech in text.* ACM Computing Surveys (CSUR), 2018. 51(4): p. 85.

23. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. and Chang, Y., *Abusive language detection in online user content*. in *Proceedings of the 25th international conference on world wide web*. 2016. International World Wide Web Conferences Steering Committee.

24. Davidson, T., Warmsley, D., Macy, M. and Weber, I., *Automated Hate Speech Detection and the Problem of Offensive Language*. in *ICWSM*. 2017.

25. N. A. Patel and N. Nanavati, “A State of the Art Review on User Behavioral Issues in Online Social Networks,” *Recent Advances in Computer Science and Communications*, vol. 16, no. 2, May 2022, doi: 10.2174/2666255815666220513162448

26. R. M. Safari, A. M. Rahmani, and S. H. Alizadeh, “User behavior mining on social media: a systematic literature review,” *Multimed Tools Appl*, vol. 78, no. 23, pp. 33747–33804, Dec. 2019, doi: 10.1007/s11042-019-08046-6.

27. Hu, J., Zhang, X., Yang, Y., Liu, Y., & Chen, X. (2020). New doctors ranking system based on vikor method. *International Transactions in Operational Research*, *27*(2), 1236–1261.

28. Alarfaj, F., Kruschwitz, U., Hunter, D., & Fox, C. (2012). Finding the right supervisor: expert-finding in a university domain. In *Proceedings of the NAACL HLT 2012 Student Research Workshop* (pp. 1–6).

29. Zhang, X., Cheng, W., Zong, B., Chen, Y., Xu, J., Li, D., & Chen, H. (2020). Temporal context-aware representation learning for question routing. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 753–761).

30. Liang, S. (2019). Unsupervised semantic generative adversarial networks for expert retrieval. In *The World Wide Web Conference* (pp. 1039–1050).

31. Montandon, J. E., Silva, L. L., & Valente, M. T. (2019). Identifying experts in software libraries and frameworks among github users. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)* (pp. 276–287). IEEE.

32. Liang, S. (2019). Unsupervised semantic generative adversarial networks for expert retrieval. In *The World Wide Web Conference* (pp. 1039–1050).

33. Ambler, S. (2012). *Agile database techniques: Effective strategies for the agile software developer*. John Wiley & Sons.

34. Gharebagh, S. S., Rostami, P., & Neshati, M. (2018). T-shaped mining: A novel approach to talent finding for agile software teams. In *European conference on information retrieval* (pp. 411–423). Springer.

35. Dehghan, M., Biabani, M., & Abin, A. A. (2019). Temporal expert profiling: With an application to t-shaped expert finding. *Information Processing & Management*, *56*(3), 1067–1079.

36. Dehghan, M., Rahmani, H. A., Abin, A. A., & Vu, V.-V. (2020). Mining shape of expertise: A novel approach based on convolutional neural network. *Information Processing & Management*, *57*(4), 102239.

37. Rostami, P., & Neshati, M. (2019). T-shaped grouping: Expert finding models to agile software teams retrieval. *Expert Systems with Applications*, *118*, 231–245.

38. Y. Wang *et al.*, "Enhancing Recommender Systems with Large Language Model Reasoning Graphs," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.10835

39. L. Wu, Z. Qiu, Z. Zheng, H. Zhu, and E. Chen, "Exploring Large Language Model for Graph Data Understanding in Online Job Recommendations," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.05722

40. Zhou, K., Li, J., Li, C., & Zhang, Y. (2020). Towards Topic-Guided Conversational Recommender System. Renmin University of China.

41. Gao, C., Li, C., Mei, Q., & Ren, X. (2021). Advances and Challenges in Conversational Recommender Systems: A Survey. AI Open, 2, 100-115.

42. Feng, Y., Liu, S., Xue, Z., Cai, Q., Hu, L., Jiang, P., Gai, K., & Sun, F. (2024). A Large Language Model Enhanced Conversational Recommender System. arXiv preprint arXiv:2308.06212.

43. Wang, C., Wang, X., He, X., et al. (2018). Toward Privacy-Preserving Personalized Recommendation Services. Engineering, 4(1), 21-29.

44. Salemi, A., Li, Z., Ren, X., & Yang, Q. (2024). LaMP: When Large Language Models Meet Personalization.

45. Zamani, H., Dumais, S., Craswell, N., Bennett, P. and Lueck, G., 2020, April. Generating clarifying questions for information retrieval. In Proceedings of the web conference 2020 (pp. 418-428).

46. J. Zou, M. Aliannejadi, E. Kanoulas, M. S. Pera and Y. Liu, "Users Meet Clarifying Questions: Toward a Better Understanding of User Interactions for Search Clarification," ACM Trans. Inf. Syst., vol. 41, no. 1, p. 16, 2023.

47. M. Aliannejadi, J. Kiseleva, . A. Chuklin, J. Dalton and M. Burtsev, "ConvAI3: Generating Clarifying Questions for Open-Domain Dialogue Systems (ClariQ)," in arXiv, 2009.11352, 2020.

48. Sekulić, I., Aliannejadi, M. and Crestani, F., 2021. User engagement prediction for clarification in search. In Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR

2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43 (pp. 619-633). Springer International Publishing.

49. M. Aliannejadi, Z. Abbasiantaeb, S. Chatterjee, J. Dalton and L. Azzopardi, "TREC iKAT 2023: A Test Collection for Evaluating Conversational and Interactive Knowledge Assistants," arXiv preprint arXiv:2405.02637, 2024.

50. X. Fu, A. Lipani,"Priming and Actions: An Analysis in Conversational Search Systems," in SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023.