Business Intelligence Management Studies
Vol. 15, No.55, Spring 2026; P 375-396
ims.atu.ac.ir
DOI: 10.22054/ims.2026.88566.2683

ATU PRESS

Original Research
Accepted: 31 May 2026
Revised: 31 May 2026
Received: 04 October 2025
eISSN: 2821-0816
ISSN: 2821-0964

Proposing an Artificial Intelligence Model for Extracting Persian Question–Answer in the Tourism Domain

Shirin Amin — Department of Information Technology Management, KI.C, Islamic Azad University, Kish, Iran

Mohammad Ali Afshar Kazemi * — Department of Industrial Management, CT.C, Islamic Azad University, Tehran, Iran

Akbar Alam Tabriz — Department of Industrial Management and Information Technology, Faculty of Management and Accounting, Shahid Beheshti University, Tehran, Iran

Seyed Mohammad Ali Khatami Firouzabadi — Department of Operations Management and Information Technology, Faculty of Management and Accounting, Allameh Tabataba'i University , Tehran, Iran

Abstract

Given the importance of the tourism industry in the country's economy and culture, and the shortage of domain-specific datasets for training Persian language models, this research aims to fill the existing gap. The objective of this study is to improve the performance of Persian tourism-related question answering systems and to provide accurate responses to queries concerning Iranian attractions and destinations. To achieve this, a collection of reliable tourism texts was gathered and processed, and using advanced natural language processing approaches, including Prompt Engineering, approximately 20,000 question–answer records were

This article is taken from a doctoral dissertation in Department of Information Technology Management, KI.C., Islamic Azad University, Kish, Iran.

* Corresponding Author: moh.afsharkazemi@iauctb.ac.ir

How to Cite: Amini. Sh., Afshar Kazemi. M., Alam Tabriz. A., Khatami Firouzabadi. S. M. (2026). Proposing an Artificial Intelligence Model for Extracting Persian Question–Answer in the Tourism Domain, *Journal of Business Intelligence Management Studies*, 15(55), 375-396. DOI: 10.22054/ims.2026.88566.2683

generated. In addition, the application of the innovative Answer Window strategy enhanced the quality of the produced data. Subsequently, the ToKA-BERT model was fine-tuned with this dataset, and evaluation results revealed its significant superiority in specialized tourism question answering compared to general-purpose Persian QA models. Both the dataset and the final model have been published on the Hugging Face platform and can serve as valuable resources for developing intelligent systems in the tourism industry and enhancing user experience.

Introduction

This study was designed to propose an artificial intelligence model for Persian question answering in the tourism domain. The authors highlight the significance of the tourism industry in Iran's economy and culture, as well as the shortage of domain-specific datasets for the Persian language, underscoring the necessity of developing efficient datasets and models in this field. In recent years, Natural Language Processing (NLP) has witnessed remarkable growth, with question answering (QA) systems being one of its most important applications. Nevertheless, the lack of high-quality datasets in Persian has constrained the advancement of such systems. Tourists frequently require precise information regarding attractions, routes, and travel conditions, and the absence of a specialized model in this context diminishes their user experience

Literature Review

Globally, datasets such as SQuAD, Natural Questions, and HotpotQA are recognized as key resources for QA research, while in Persian only limited efforts, such as PersianQuAD, ParSQuAD, and PeQA, have been undertaken. In the area of language models, approaches based on BERT and Persian-specific models such as ParsBERT and TookaBERT have been introduced. Despite these advances, no domain-specific dataset or model has yet been developed for Persian tourism-related QA

Methodology

More than one thousand tourism-related articles were initially collected from the Alibaba website, subsequently cleaned, and divided into 2,000 paragraphs. Leveraging large language models and prompt engineering techniques, over 20,000 structured question–answer records were

generated following the SQuAD format. To improve the precision of answer localization, the innovative "Answer Window" strategy was employed. For model training, TookaBERT-Large, with 340 million parameters, was selected as the base model and fine-tuned with the generated dataset

Results

The results revealed that the developed model significantly outperformed existing general-purpose Persian models. In evaluations using the Exact Match (EM) and F1 metrics, the proposed model achieved scores of 67% and 82%, respectively, whereas the best-performing competitor (xlmr-large-qa-fa) achieved 41% and 72%. These outcomes demonstrate the effectiveness of combining a domain-specific dataset with fine-tuning.

Comparison of TookaBERT-Large with other Persian QA models:

Model Name	EM (%)	F1 (%)
xlmr-large-qa-fa	41.22	71.82
mdeberta-v3-base-squad2	34.63	66.89
distilbert-fa-zwnj-base	27.51	52.75
parsbert-persian-QA	16.98	43.68
TookaBERT-Large (this study)	67.00	82.00

The proposed model thus performed substantially better than its competitors

Discussion

The study underscores the importance of employing domain-specific data and advanced architectures, arguing that techniques such as prompt engineering and the Answer Window strategy played a pivotal role in enhancing data quality and model accuracy. Furthermore, the proposed model was able to provide more accurate responses to tourism-related queries, including historical information, routes, and visiting conditions

Conclusion

The findings indicate that developing a Persian dataset in the tourism domain and fine-tuning TookaBERT resulted in a powerful model for Persian question answering systems. Beyond practical applications in the tourism industry—such as chatbots and information systems—this model can also serve as a framework for similar studies in other

specialized domains. Although the results are promising, expanding the diversity of data sources and increasing dataset scale may lead to further improvements in model performance in the future.

Keywords: ToKA-BERT, Intelligent Question-Answering, Iranian Tourism, Natural Language Processing (NLP), SQuAD-based Persian Dataset, Prompt Engineering, Answer Window

مقاله پژوهشی تاریخ ارسال: ۱۴۰۴/۰۷/۱۲ تاریخ بازنگری: ۱۴۰۵/۰۳/۱۰ تاریخ پذیرش: ۱۴۰۵/۰۳/۱۰
ISSN: 2821-0964 eISSN: 2821-0816

مطالعات مدیریت کسب و کار هوشمند
سال پانزدهم، شماره ۵۵، بهار ۱۴۰۵، ص ۳۷۵ تا ۳۹۶
ims.atu.ac.ir
DOI: 10.22054/ims.2026.88566.2683

ارائه یک مدل هوش مصنوعی برای استخراج پرسش و پاسخ‌های فارسی در حوزه گردشگری

شیرین امینی — گروه مدیریت فناوری اطلاعات، واحد بین‌المللی کیش، دانشگاه آزاد اسلامی، کیش، ایران

محمدعلی افشار کاظمی * — گروه مدیریت صنعتی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران

اکبر عالم تبریز — گروه مدیریت صنعتی و فناوری اطلاعات، دانشکده مدیریت و حسابداری، دانشگاه شهید بهشتی، تهران، ایران

سید محمدعلی خاتمی فیروزآبادی — گروه مدیریت عملیات و فناوری اطلاعات، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبائی، تهران، ایران

چکیده

با توجه به اهمیت صنعت گردشگری در اقتصاد و فرهنگ کشور و کمبود منابع داده‌ای تخصصی برای آموزش مدل‌های زبانی فارسی، این تحقیق درصدد پر کردن این خلأ برآمد. هدف از این پژوهش، ارتقای عملکرد سیستم‌های پاسخگویی به سؤالات گردشگری فارسی و ارائه پاسخ‌های دقیق به پرس‌وجوهای مرتبط با جاذبه‌ها و مقاصد گردشگری ایران بوده است. بدین منظور، مجموعه‌ای از متون معتبر گردشگری گردآوری و پردازش شد و با بهره‌گیری از رویکردهای پیشرفته پردازش زبان طبیعی، ازجمله مهندسی پرامپت، حدود ۲۰ هزار رکورد پرسش و پاسخ تولید گردید. علاوه‌براین، استفاده از راهبرد نوآورانه پنجره پاسخ موجب ارتقای کیفیت داده‌های تولیدشده شد. در ادامه، مدل توکابرت با این مجموعه‌داده فاین‌تیون گردید و نتایج ارزیابی‌ها نشان داد که عملکرد آن در حوزه پرسش و پاسخ تخصصی گردشگری نسبت به مدل‌های عمومی زبان فارسی برتری قابل‌توجهی دارد. مجموعه‌داده و مدل نهایی در پلتفرم هاگینگ فیس

مقاله حاضر برگرفته از رساله دکتری گروه مدیریت فناوری اطلاعات، واحد بین‌المللی کیش، دانشگاه آزاد اسلامی، کیش، ایران است.

* نویسنده مسئول : moh.afsharkazemi@iauctb.ac.ir

منتشرشده و می‌توانند به‌عنوان منبعی ارزشمند برای توسعه سیستم‌های هوشمند در صنعت گردشگری و بهبود تجربه کاربران مورداستفاده قرار گیرند.

کلیدواژه‌ها: پرسش و پاسخ هوشمند، گردشگری ایران، پردازش زبان طبیعی، مجموعه‌داده اسکویاد فارسی، مهندسی پرامپت، پنجره پاسخ.

مقدمه

پردازش زبان طبیعی[1] یا NLP به‌عنوان یکی از شاخه‌های کلیدی هوش مصنوعی، در سال‌های اخیر پیشرفت‌های چشمگیری داشته است. یکی از مهم‌ترین کاربردهای پردازش زبان طبیعی، پاسخگویی به سؤالات[2] یا QA توسط مدل‌هایی است که برای این منظور تولید شدند و هدف آن ارائه پاسخ‌های دقیق و خودکار به پرس‌وجوهای مطرح‌شده توسط کاربران می‌باشد. مدل‌های QA در حوزه‌های مختلفی ازجمله آموزش، خدمات مشتری و گردشگری کاربرد دارند. بااین‌حال، تولید چنین مدل‌هایی در زبان فارسی با چالش‌های متعددی مواجه است که مهم‌ترین آن‌ها کمبود مجموعه داده[3] تخصصی و باکیفیت است. گردشگران ممکن است سؤالاتی مانند «بهترین زمان بازدید از کاخ گلستان چیست؟» یا «چگونه می‌توان به چشمه‌های باداب سورت دسترسی پیدا کرد؟» داشته باشند. مدل‌های QA می‌توانند با ارائه پاسخ‌های دقیق به این سؤالات، تجربه کاربری را بهبود بخشند و به افزایش رضایت گردشگران کمک کنند؛ اما در حال حاضر، مجموعه داده و مدل‌های QA تخصصی در حوزه گردشگری در زبان فارسی وجود ندارد و مدل‌های QA عمومی فارسی عملکرد محدودی در حوزه گردشگری دارند. هدف اصلی این پژوهش، تولید یک مجموعه داده تخصصی در حوزه گردشگری به زبان فارسی و به‌تبع آن تولید مدل QA مناسب سؤالات گردشگری به زبان فارسی است. در این راستا به‌طور خلاصه اقداماتی به این شرح انجام گردید: ابتدا مجموعه‌ای از مقالات گردشگری از وب‌سایت علی‌بابا جمع‌آوری شد. سپس، این داده‌ها با استفاده از تکنیک‌های پیش‌پردازش، جهت افزایش کیفیت پاک‌سازی شدند و ۲۰۰۰ پاراگراف تولید گردید. در ادامه، با بهره‌گیری از مدل‌های زبانی بزرگ، ۲۰،۰۰۰ رکورد پرسش و پاسخ از متن این پاراگراف‌ها استخراج شد. برای افزایش دقت، از رویکرد خلاقانه پنجره پاسخ استفاده شد که امکان شناسایی دقیق‌تر موقعیت پاسخ‌ها در متن را فراهم کرد. پس از تولید مجموعه داده، مدل QA مبتنی بر آن آموزش

1 Natural Language Processing
2 Question Answering
3 Dataset

داده شـد و کارایی آن با اسـتفاده از معیارهای اسـتاندارد مانند معیار F1[1] و انطباق دقیق[2] با سایر مدل‌های QA موجود فارسی مقایسه شد. درنهایت مجموعه داده و مدل QA تولیدشده در مخزن HuggingFace[3] جهت استفاده عموم بارگذاری شد.

پیشینه پژوهش

تحقیقات در زمینه QA و تولید مجموعه داده‌های مرتبط در سال‌های اخیر رشد قابل‌توجهی داشـته اسـت. مجموعه داده SQuAD[4] (Rajpurkar et al., 2016) که در زبان انگلیسـی توسـعه‌یافته، یکی از برجسـته‌ترین نمونه‌هاسـت و به‌عنوان معیاری اسـتاندارد برای ارزیابی مدل‌های QA شـناخته می‌شـود. بااین‌حال، در زبان فارسـی، تلاش‌های مشـابه بسـیار محدود بوده و بیشـتر تحقیقـات بـه ترجمـه مجموعـه داده‌هـای موجود یـا تولید مجموعه داده‌های کوچک با دامنه محدود پرداخته‌اند. در حوزه گردشـگری، در زبان فارسـی تقریباً هیچ منبع تخصـصـی وجود ندارد. توسـعه مجموعه‌داده‌های بزرگ‌مقیاس و باکیفیت برای پیشـرفت تحقیقـات QA حیـاتی بوده اسـت. چنـدین مجموعه‌داده معیار این حوزه را شـکل داده و معیـارهـای ارزیـابی اسـتـانـداردی برای مقـایسـه رویکردهـای مختلف فراهم کرده‌انـد. مجموعه‌داده SQuAD یکی از تأثیرگذارترین معیارها در QA اسـتخراجی اسـت. نسـخه SQuAD 1.1 شـامل بیش از ۱۰۰،۰۰۰ جفت پرسـش و پاسـخ مبتنی بر مقالات ویکی‌پدیا اسـت که هر پاسـخ یک بخش متنی پیوسـته از پاراگراف مربوطه اسـت. این مجموعه‌داده برای آزمون قابلیت‌های درک مطلب طراحی‌شـده و پیشـرفت‌های قابل‌توجهی در توسـعه مدل‌های QA به دنبال داشـته اسـت. نسـخه SQuAD 2.0 با افزودن پرسـش‌های بی‌پاسـخ، مجموعه‌داده اصلی را گسترش داد و وظیفه را واقع‌گرایانه‌تر و چالش‌برانگیزتر کرد (Tran et al., 2023). این افزونه از مدل‌ها می‌خواهد نه‌تنها مکان پاسـخ‌ها را در متن تعیین کنند، بلکه تشـخیص دهند چه زمانی پرسـش‌ها بر اسـاس زمینه ارائه‌شـده قابل پاسـخ‌گویی نیستند

1 F1 Score
2 Exact Match (EM)
3 https://huggingface.co/shirinamini67
4 Stanford Question Answering Dataset

(Tran et al., 2023). این مجموعه‌داده شـامل بیش از ۳۰۰،۰۰۰ پرسـش طبیعی همراه با مقالات ویکی‌پدیا اسـت و محیط ارزیابی واقع‌گرایانه‌تری فراهم می‌کند (Kwiatkowski et al., 2019). پرسـش‌ها در مقایسـه با مجموعه‌داده‌های مردمی، تنوع و پیچیدگی زبانی بیشتری نشان می‌دهند (Alberti et al., 2019). پرسش‌های طبیعی به معیاری استاندارد برای ارزیابی سیستم‌های QA دامنه باز و توانایی آن‌ها در مدیریت نیازهای اطلاعاتی واقعی تبدیل شده است(Kwiatkowski et al., 2019; Rosenthal et al., 2024) . TriviaQA با تمرکز بر پرسـش‌های trivia که نیازمند اسـتدلال در چندین جمله هسـتند، پیچیدگی بیشـتری را معرفی می‌کند (Joshi et al., 2017). این مجموعه‌داده شـامل بیش از ۶۵۰،۰۰۰ سـه‌تایی پرسـش، پاسـخ و شـواهد اسـت که پرسش‌ها توسط علاقه‌مندان به trivia نوشته‌شده و اسناد شـواهد به‌طور مسـتقل جمع‌آوری شـده‌اند.[1] TriviaQA با تنوع واژگانی و نحوی بالا بین پرسـش‌ها و جملات پاسـخ و شـواهد مشـخص می‌شـود و نیاز به قابلیت‌های اسـتدلالی پیچیده‌تری دارد. HotpotQA به‌طور خاص قابلیت‌های استدلال چندمرحله‌ای را هدف قرار می‌دهد و مدل‌ها را ملزم به یافتن و اسـتدلال بر روی چندین سند پشتیبان می‌کند (Yang et al., 2018) . این مجموعه‌داده شـامل ۱۱۳،۰۰۰ جفت پرسـش و پاسـخ مبتنی بر ویکی‌پدیا اسـت. مجموعه‌داده MS MARCO[2] یکی دیگر از مشـارکت‌های مهم در تحقیقات QA اسـت(Bajaj et al., 2018). برخـلاف SQuAD، پرسـش‌هـای MS MARCO از پرس‌وجوهای واقعی کاربران اسـتخراج‌شـده و پاسـخ‌ها توسط انسان تولید شده‌اند، نه اینکه مسـتقیماً از متن اسـتخراج شـوند. این طراحی، مجموعه‌داده را چالش‌برانگیزتر می‌کند، زیرا پاسخ‌ها ممکن است به‌صورت کلمه به کلمه در متن ارائه‌شده ظاهر نشوند.

معیارهای ارزیابی اسـتاندارد برای QA استخراجی شامل تطابق دقیق EM[3] و امتیاز F1 هستند (Fajcik et al., 2021 ; Soh & Zhao, 2024). تطابق دقیق درصد پیش‌بینی‌هایی را که دقیقاً با پاسـخ حقیقت زمینه‌ای مطابقت دارند اندازه‌گیری می‌کند، درحالی‌که امتیاز F1

---

۱ TriviaQA یک دیتاست استاندارد و شناخته‌شده در حوزه پرسش و پاسخ است که شامل پرسش‌هایی با پاسخ‌های کوتاه است که از منابع مختلف جمع‌آوری شده‌اند

2 Microsoft MAchine Reading COmprehension

3 Exact Match

معیاری ملایم‌تر ارائه می‌دهد که با محاسبه میانگین هارمونیک دقت و فراخوان در سطح توکن عمل می‌کند. این معیارها به‌طور گسترده در معیارهای QA پذیرفته شده‌اند و امکان مقایسه مداوم عملکرد مدل‌ها را فراهم می‌کنند. توسعه مدل‌های زبانی خاص فارسی برای پیشبرد عملکرد QA حیاتی بوده است. چندین مدل قابل‌توجه توسعه یافته‌اند. ParsBERT اولین مدل BERT تک‌زبانه برای فارسی است که بر روی یک پیکره[1] بزرگ از متن فارسی آموزش دیده است(Farahani et al., 2021). ParsBERT در مقایسه با مدل‌های چندزبانه، عملکرد برتری در وظایف مختلف NLP فارسی، ازجمله پاسخ‌دهی به پرسش‌ها، نشان داده است. SINA-BERT به‌طور خاص بر حوزه پزشکی متمرکز است و کمبود مدل‌های زبانی فارسی در کاربردهای مراقبت‌های بهداشتی را برطرف می‌کند(Taghizadeh et al., 2021). این مدل اهمیت آموزش خاص حوزه برای وظایف QA تخصصی را نشان می‌دهد. FaBERT یک مدل BERT فارسی را معرفی می‌کند که بر روی پیکره HmBlogs پیش‌آموزش دیده است و شامل متون فارسی غیررسمی و رسمی می‌شود(Masumi et al., 2024) .FaBERT بهبود عملکرد در وظایف مختلف پایین‌دستی[2]، ازجمله پاسخ‌دهی به پرسش‌ها، نشان داده است. TookaBERT پیشرفت‌های اخیر در مدل‌سازی زبان فارسی را نشان می‌دهد و با دو نسخه به عملکرد پیشرفته در چندین وظیفه NLU فارسی دست یافته است(SadraeiJavaheri et al., 2024). مدل بزرگ‌تر TookaBERT بهبودهای قابل‌توجهی نسبت به مدل‌های BERT فارسی موجود نشان می‌دهد. چندین سیستم QA تخصصی برای فارسی توسعه یافته‌اند. PerAnSel یک سیستم نوین مبتنی بر شبکه عصبی عمیق را معرفی می‌کند که به‌طور خاص برای پاسخ‌دهی به پرسش‌های فارسی طراحی شده است(Mozafari et al., 2022). این سیستم با موازی‌سازی روش‌های ترتیبی و مبتنی بر ترانسفورمر، نظم آزاد کلمات فارسی را مدیریت می‌کند و عملکرد قوی در چندین مجموعه‌داده فارسی به دست می‌آورد. FarsNewsQA یک سیستم QA مبتنی بر یادگیری عمیق برای مقالات خبری فارسی توسعه می‌دهد(Kazemi

1 corpus
2 Downstream Task

et al., 2023). این کار کاربرد تکنیک‌های QA در کاربردهای حوزه خبری فارسی را نشان می‌دهد. درحالی‌که پیشرفت‌های قابل‌توجهی در QA فارسی عمومی حاصل شده است، حوزه‌های تخصصی مانند گردشگری تا حد زیادی ناشناخته باقی مانده‌اند. صنعت گردشگری حوزه کاربردی حیاتی است که در آن سیستم‌های QA می‌توانند با کمک به گردشگران در دسترسی به اطلاعات درباره جاذبه‌ها، اقامتگاه‌ها و مکان‌های فرهنگی، ارزش قابل‌توجهی ارائه دهند. عدم وجود مجموعه‌داده‌های QA خاص گردشگری در فارسی، توسعه سیستم‌های تخصصی برای این حوزه را محدود کرده است. مدل‌های QA فارسی عمومی موجود هنگام اعمال بر پرس‌وجوهای مرتبط با گردشگری، اثربخشی محدودی نشان می‌دهند و نیاز به داده‌های آموزشی و مدل‌های خاص حوزه را برجسته می‌کنند. توسعه سیستم‌های QA متقابل زبانی و چندزبانه با توجه به تلاش محققان برای گسترش قابلیت‌های QA به چندین زبان، موردتوجه قرار گرفته است (Wang et al., 2019). رویکردهای مختلفی بررسی شده‌اند، ازجمله روش‌های مبتنی بر ترجمه، آموزش مدل‌های چندزبانه و یادگیری انتقال متقابل زبانی. بااین‌حال، مطالعات به‌طور مداوم نشان می‌دهند که مدل‌های تک‌زبانه آموزش‌دیده بر داده‌های خاص زبان، از جایگزین‌های چندزبانه عملکرد بهتری دارند. این یافته اهمیت توسعه مجموعه‌داده‌ها و مدل‌های خاص زبان برای عملکرد بهینه QA را تقویت می‌کند.

روش

در این بخش، روش‌ها و فرآیندهایی که برای انجام این پژوهش به کار گرفته شده‌اند، با جزئیات کامل تشریح می‌شوند. بخش حاضر به سه زیرمجموعه تقسیم می‌شود: تولید مجموعه داده، تولید مدل و ارزیابی کارایی. تولید مجموعه داده یکی از مهم‌ترین مراحل این پژوهش بود که باهدف ایجاد منبعی باکیفیت و جامع برای آموزش و ارزیابی مدل‌های QA در حوزه گردشگری انجام شد. این فرآیند شامل جمع‌آوری داده‌ها از منابع معتبر، پردازش و پاک‌سازی متون و تولید سؤالات و پاسخ‌ها بود. جمع‌آوری داده‌ها گام اولیه و اساسی در تولید مجموعه داده بود ابتدا مجموعه‌ای از مقالات گردشگری از وب‌سایت

علی‌بابا جمع‌آوری شـد. این سـایت به دلیل ارائه محتوای غنی و متنوع درباره جاذبه‌های گردشـگری ایران، نکات سـفر و اطلاعات فرهنگی و تاریخی، به‌عنوان منبع اصـلی انتخاب شـد. بیش از ۱۰۰۰ مقاله با اسـتفاده از تکنیک‌های وب‌کاوی اسـتخراج شـدند. این مقالات موضـوعات مختلفی ازجمله تاریخچه مکان‌ها، شـرایط بازدید و نکات کاربردی را در بر می‌گرفتند که برای ایجاد یک مجموعه داده جامع و متنوع بسـیار مناسـب بودند. پس از جمع‌آوری مقـالات، نیاز به پردازش و پاک‌سـازی آن‌ها به‌منظور آماده‌سـازی برای تولید مجموعه داده در قالب SQuAD احسـاس شد. مقالات وب اغلب حاوی عناصر غیرضروری مانند تبلیغات و لینک‌ها بودند که باید حذف می‌شـدند. این مرحله با تقسـیم‌بندی متون به تکه‌های کوچک‌تر آغاز شـد. هر تکه به حداکثر ۲۰۰۰ کاراکتر محدود شـد و همپوشـانی ۵۰ کاراکتری برای حفظ پیوسـتگی معنایی در نظر گرفته شـد. این کار منجر به تولید ۲۰۰۰ تکه متنی یا پاراگراف شد که برای مراحل بعدی آماده بودند. پاک‌سازی متون با استفاده از مـدل زبـانی gpt-4o mini و سـرویس Gretel[1] انجـام شـد. هـدف این بود کـه متون به پاراگراف‌هایی منسـجم و روان تبدیل شوند که برای تولید سؤالات و پاسخ‌ها مناسب باشند. پرامپتی[2] طراحی شـد کـه مـدل را بـه بازنویسـی متون با لحنی مدرن و محاوره‌ای هدایت می‌کرد و از افزودن اطلاعات غیرضروری جلوگیری می‌کرد. این فرآیند نه‌تنها کیفیت متون را بهبود داد، بلکه آن‌ها را به شکلی استاندارد و قابل‌استفاده برای مجموعه داده تبدیل کرد. مرحله نهایی تولید مجموعه داده، ایجاد رکوردهای SQuAD شـامل سـؤالات، پاسـخ‌ها و بـخش‌هـای متنی مرتبط بود. برای این منظور، از مـدل زبـانی DeepSeek-R1-Distill-Llama-70B اسـتفاده شـد که به دلیل توانایی بالای خود در پردازش زبان فارسـی انتخاب شـده بود. این مدل با پرامپتی دقیق هدایت شـد تا سـؤالاتی مرتبط با نیازهای گردشـگران، مانند سـاعات بازدید یا هزینه‌ها، تولید کند. درنهایت، دیتاسـتی با ۲۰،۰۰۰ رکورد ایجاد شد که هر رکورد شـامل سـؤال، پاسـخ و بخش متنی مرتبط بود. سـاختار نهایی مجموعه داده تولیدشده شامل ستون‌های زیر بود:

1 https://gretel.ai
2 prompt

- Title: عنوان متن زمینه (در این مورد، نام جاذبه گردشگری یا موضوع مرتبط).
- Context: متن زمینه که اطلاعات اصلی را ارائه می‌دهد.
- Question: سؤالی که از متن زمینه استخراج شده است.
- Answers: شامل متن پاسخ و موقعیت آن (start_index و end_index) در متن زمینه.

که کاملاً منطبق بر ساختار مجموعه داده SQuAD است.

تولید مدل یکی از مهم‌ترین و بنیادی‌ترین مراحل این پژوهش بود که هدف آن توسعه یک سیستم پاسخگویی به سؤالات پیشرفته در حوزه گردشگری به زبان فارسی بود. این فرآیند شامل انتخاب یک مدل زبانی پایه و تنظیم دقیق[1] آن بود. انتخاب مدل پایه برای این پژوهش از اهمیت بالایی برخوردار بود، زیرا این مدل باید توانایی درک متون پیچیده فارسی و ارائه پاسخ‌های دقیق به سؤالات را داشته باشد. پس از بررسی گزینه‌های مختلف، مدل TookaBERT-Large به‌عنوان مدل پایه انتخاب شد. این مدل توسط تیم PartAI توسعه‌یافته و در پلتفرم Hugging Face در دسترس است[2]. TookaBERT-Large یک مدل زبانی مبتنی بر معماری BERT است که به‌طور خاص برای زبان فارسی طراحی و بهینه‌سازی شده است. این مدل پایه، ۳۴۰ میلیون پارامتر دارد که آن را به یکی از بزرگ‌ترین و قدرتمندترین مدل‌های زبانی موجود برای زبان فارسی تبدیل می‌کند. این مدل بر روی مجموعه داده‌های گسترده و متنوع فارسی، ازجمله مقالات ویکی‌پدیای فارسی، متون خبری، محتوای وب و در کل بیش از ۵۰۰ گیگابایت متن فارسی پیش‌آموزش[3] شده است. این مدل در مقایسه با سایر مدل‌های موجود، مانند ParsBERT یا XLM-RoBERTa، در وظایف مختلف پردازش زبان طبیعی فارسی عملکرد بهتری نشان داده است. برای تنظیم دقیق مدل TookaBERT-Large پایه از کدهای مورداستفاده از مخزن رسمی Hugging Face[4] استفاده شد و برای سازگاری با TookaBERT-Large و

1 Fine tuning
2 https://huggingface.co/PartAI/TookaBERT-Large
3 Pre training
4 https://github.com/huggingface/transformers/tree/main/examples/pytorch/question-answering#fine-tuning-bert-on-squad10

مجموعه داده فارسی تغییراتی در آن‌ها اعمال شد.

یافته‌ها

ارزیابی کارایی مدل توسعه‌یافته یکی از مهم‌ترین بخش‌های این پژوهش بود که هدف آن سنجش دقت و کیفیت پاسخ‌های مدل TookaBERT-Large و مقایسه آن با مدل‌های QA فارسی موجود بود. برای یک ارزیابی علمی معتبر، مدل‌هایی که حمایت علمی منتشرشده در مجلات همتا (peer-reviewed) یا اعتبار پژوهشی نداشتند، از مقایسه کنار گذاشته شدند. ارزیابی نهایی شامل مدل‌های SOTA مبتنی بر ترنسفورمر معتبر و همچنین مدل‌های پایه پرسش و پاسخ فارسی پرکاربرد است که امکان مقایسه‌ای منصفانه و معنی‌دار با مدل پیشنهادی TookaBERT-Large را فراهم می‌کند. برای اطمینان از یک مقایسه علمی معتبر، تنها مدل‌های مبتنی بر ترنسفورمر معتبر و با سوابق مستند پژوهشی در ارزیابی گنجانده شدند. مدل‌هایی که در دسته SOTA قرار می‌گیرند، نمایانگر پیشرفته‌ترین معماری‌های پرسش و پاسخ چندزبانه و تک‌زبانه هستند، درحالی‌که مدل‌های پایه (baseline) به سیستم‌های پرسش و پاسخ فارسی پرکاربرد اشاره دارند که به‌عنوان نقاط مرجع عمل می‌کنند. پیاده‌سازی‌های غیرعلمی یا آموزشی برای حفظ اعتبار ارزیابی کنار گذاشته شدند. این طراحی امکان ارزیابی منصفانه مدل پیشنهادی TookaBERT-Large در برابر هم مدل‌های SOTA قدرتمند و هم سیستم‌های رایج پرسش و پاسخ فارسی را فراهم می‌کند. این ارزیابی نه‌تنها عملکرد مدل ما را نشان داد، بلکه به ما امکان داد تا برتری‌های آن را نسبت به رقبا به‌صورت کمی و کیفی تحلیل کنیم. برای سنجش عملکرد مدل‌ها، از دو معیار استاندارد در حوزه QA استفاده شد:

- معیار Exact Match (EM): این معیار نشان می‌دهد که آیا پاسخ پیش‌بینی‌شده توسط مدل دقیقاً با پاسخ صحیح در مجموعه داده مطابقت دارد یا خیر. این یک معیار سخت‌گیرانه است که تنها در صورت تطابق کامل، امتیاز مثبت می‌دهد. برای مثال، اگر پاسخ صحیح «تهران» باشد و مدل «تهران» را پیش‌بینی کند، امتیاز ۱ می‌گیرد، اما اگر «شهر تهران» را پیش‌بینی کند، امتیاز ۰ می‌گیرد.

- معیار F1 Score: این معیار میانگین هارمونیک دقت[1] و فراخوان[2] را در سطح توکن محاسبه می‌کند. این انعطاف‌پذیرتر از EM است و همپوشانی جزئی بین پاسخ پیش‌بینی‌شده و پاسخ صحیح را نیز در نظر می‌گیرد. برای مثال، اگر پاسخ صحیح «موزه ملی ایران» باشد و مدل «موزه ملی» را پیش‌بینی کند، F1 Score بر اساس تعداد توکن‌های مشترک («موزه» و «ملی») محاسبه می‌شود.

این معیارها به‌طور گسترده در ارزیابی سیستم‌های QA، به‌ویژه در مجموعه داده‌های SQuAD استفاده می‌شوند و امکان مقایسه دقیق و استاندارد بین مدل‌ها را فراهم می‌کنند. ارزیابی بر روی مجموعه آزمون که شامل ۲،۰۰۰ رکورد از مجموعه داده بود، انجام شد. این مجموعه جدا از داده‌های آموزشی و اعتبارسنجی بود تا عملکرد مدل‌ها در شرایط واقعی و بر روی داده‌های نادیده سنجیده شود. هر مدل بر روی این مجموعه اجرا شد و پاسخ‌های پیش‌بینی‌شده با پاسخ‌های صحیح مقایسه شدند. برای اجرای ارزیابی، از اسکریپت‌های استاندارد Hugging Face استفاده شد که معیارهای EM و F1 Score را به‌صورت خودکار محاسبه می‌کردند. این اسکریپت‌ها به‌گونه‌ای طراحی شده‌اند که نتایج قابل‌اعتماد و قابل تکرار باشند.

نتایج ارزیابی برای تمامی مدل‌ها در جدول زیر ارائه شده است. این جدول شامل امتیازات EM و F1 Score برای هر مدل است و مدل ما در ردیف آخر قرار دارد.

1 precision
2 recall

جدول ۱. نتایج ارزیابی مدل‌ها

ردیف	نام مدل	نوع مدل	EM (%)	F1 (%)
1	xlmr-large-qa-fa	SOTA (Multilingual)	41.22	71.82
2	mdeberta-v3-base-squad2	SOTA (Transformer-based)	34.63	66.89
3	xlm-roberta-base-finetuned	SOTA (Transformer-based)	37.26	65.25
4	mdeberta-v3-base-squad2-onnx	SOTA (Optimized Variant)	34.63	66.89
5	distilbert-fa-zwnj-base	Baseline (Lightweight)	27.51	52.75
6	Persian-QA-Bert-V1	Baseline (Persian QA)	30.59	58.63
7	parsbert-persian-QA	Baseline (ParsBERT-based)	16.98	43.68
8	tara-roberta-base-fa-qa	Baseline (Domain-adapted)	8.86	30.12
9	TookaBERT-Large (Proposed)	Proposed Model	67.00	82.00

نتایج نشـان می‌دهند که مدل TookaBERT-Large با امتیاز 67.00% EM و F1 Score 82.00%، به‌طور قابل‌توجهی از سـایر مدل‌ها پیشـی گرفته است. بهترین مدل بعدی، xlmr-large-qa-fa، با 41.22% EM و 71.82% F1 Score ، اختلاف قابل‌توجهی با مدل ما دارد. این برتری نشـان‌دهنده موفقیت فرآیند تنظیم دقیق و اسـتفاده از مجموعه داده تخصـصـی اسـت. برتری مدل TookaBERT-Large به عوامل متعددی بازمی‌گردد که برخی از آن‌ها عبارت‌اند از: مدل ما بر روی دیتاسـت SQuAD فارسی در حوزه گردشگری تنظیم شد که شـامل سؤالات و پاسخ‌های مرتبط با جاذبه‌های گردشگری ایران بود. این در حالی است که سایر مدل‌های مورد ارزیابی بر روی داده‌های عمومی‌تر یا ترجمه‌شده آموزش دیده‌اند. این تطبیق تخصـصـی به مدل امکان داد تا الگوهای زبانی و مفاهیم خاص حوزه گردشـگری را بهتر درک کنـد. معمـاری قدرتمند TookaBERT-Large با ۲۴ لایه و ۳۴۰ میلیون پارامتر، ظرفیت بالاتری برای یادگیری الگوهای پیچیده دارد. در مقابل، مدل‌هایی مانند distilbert-fa-zwnj-base با ۶ لایه و ۶۶ میلیون پارامتر، برای سـرعت بیشـتر طراحی شـده‌اند، اما در دقـت عملکرد ضـعیف‌تری دارنـد و درنهایت توکن‌سـازی بهینه به مدل کمک کرد تا با

نیازهای زبانی و معنایی کاربران فارسی‌زبان سازگار شود. در مقایسه با xlmr-large-qa-fa که بهترین مدل رقیب TookaBERT-Large بود، مدل ما در هر دو معیار EM و F1 Score برتری داشت. این به دلیل اندازه مدل ما نیست زیرا مدل ما ۳۴۰ میلیون پارامتر در مقابل ۵۵۰ میلیون پارامتر دارد، بلکه به تنظیم دقیق تخصصی و استفاده از داده‌های باکیفیت‌تر مربوط می‌شود. اگرچه یک مدل چندزبانه قدرتمند است، اما به دلیل تمرکز بر زبان‌های متعدد، دقت کمتری در جزئیات زبانی فارسی دارد. نتایج این ارزیابی نشان‌دهنده پتانسیل بالای مدل‌های زبانی تنظیم‌شده برای زبان فارسی در حوزه‌های تخصصی است. امتیاز EM 67%و F1 Score 82% نشان می‌دهند که TookaBERT-Large می‌تواند به‌عنوان یک سیستم QA قابل‌اعتماد در کاربردهای واقعی، مانند چت‌بات‌های گردشگری یا سیستم‌های اطلاع‌رسانی، استفاده شود. همچنین، این نتایج معیاری برای تحقیقات آینده در زمینه QA فارسی فراهم می‌کند و بر اهمیت تولید داده‌های تخصصی و تنظیم دقیق تأکید دارد. اگرچه مدل ما عملکرد برتری داشت، اما هنوز جای بهبود وجود دارد و می‌توان با افزایش تنوع منابع داده‌ها و حجم رکوردها دقت را بیشتر کرد.

بحث و نتیجه‌گیری

این پژوهش موفق به تولید یک مجموعه داده در قالب SQuAD با ۲۰،۰۰۰ رکورد تخصصی در حوزه گردشگری شد که با روش‌های پیشرفته ایجاد شده است. مدل TookaBERT-Large نیز با موفقیت تنظیم شد و به عملکردی برتر دست یافت. این دستاوردها نه‌تنها به تقویت تحقیقات NLP کمک می‌کند، بلکه کاربردهای عملی در صنعت گردشگری دارد. مجموعه داده تولیدشده با تمرکز بر کیفیت و تنوع، منبعی ارزشمند برای پژوهشگران است. تنظیم مدل نیز نشان داد که چگونه می‌توان با استفاده از داده‌های تخصصی، عملکرد سیستم‌های هوشمند را بهبود داد.

اهمیت این پژوهش در ارائه منابعی است که می‌توانند در تحقیقات و صنعت به کار گرفته شوند. دیتاست به‌عنوان یک معیار استاندارد و مدل به‌عنوان ابزاری برای سیستم‌های هوشمند مانند چت‌بات‌ها قابل‌استفاده است. این ابزارها می‌توانند تجربه گردشگران را بهبود

دهند و به توسعه صنعت توریسم کمک کنند. کاربردهای این پژوهش فراتر از گردشگری است و می‌تواند در حوزه‌های دیگر نیز الهام‌بخش باشد. تمرکز بر زبان فارسی و رفع کمبود منابع، این پروژه را به یک نمونه موفق تبدیل کرده است که می‌تواند الگویی برای پروژه‌های مشابه باشد. تحقیقات آینده می‌توانند بر بهبود کیفیت داده‌ها و افزایش مقیاس دیتاست تمرکز کنند. همچنین، همکاری با صنعت برای پیاده‌سازی این ابزارها می‌تواند تأثیر واقعی آن‌ها را نشان دهد.

از منظر کسب‌وکار هوشمند، نتایج این پژوهش فراتر از یک دستاورد فنی صرف قابل تفسیر است. سیستم‌های پاسخ‌گویی هوشمند مبتنی بر این مدل می‌توانند به‌عنوان زیرساخت اطلاعاتی در اکوسیستم گردشگری دیجیتال مورداستفاده قرار گیرند و نقش مؤثری در کاهش هزینه‌های پشتیبانی مشتری، افزایش سرعت دسترسی به اطلاعات و بهبود کیفیت تعامل با گردشگران ایفا کنند. به‌ویژه در پلتفرم‌های گردشگری آنلاین، چنین مدل‌هایی می‌توانند بخشی از زنجیره ارزش داده‌محور باشند که تصمیم‌گیری مدیران و ارائه خدمات شخصی‌سازی‌شده را تسهیل می‌کنند.

در همین راستا، پیاده‌سازی عملی این مدل در قالب یک چت‌بات گردشگری نشان داد که استفاده از مدل‌های QA تخصصی می‌تواند منجر به افزایش رضایت کاربران و بهبود تجربه تعامل انسان–سیستم شود. این موضوع حاکی از آن است که بهره‌گیری از هوش مصنوعی زبانی در گردشگری، صرفاً یک نوآوری فناورانه نیست، بلکه ابزاری برای خلق ارزش اقتصادی و رقابتی در کسب‌وکارهای مبتنی بر پلتفرم محسوب می‌شود.

از منظر مدیریتی، نتایج این پژوهش می‌تواند برای مدیران صنعت گردشگری، توسعه‌دهندگان پلتفرم‌های هوشمند و سیاست‌گذاران حوزه تحول دیجیتال مفید باشد؛ زیرا نشان می‌دهد که سرمایه‌گذاری در داده‌های بومی و مدل‌های زبانی خاص زبان می‌تواند مزیت رقابتی پایداری ایجاد کند. همچنین، مجموعه‌داده و مدل ارائه‌شده می‌توانند به‌عنوان زیرساختی برای توسعه خدمات هوشمند دیگر مانند سیستم‌های توصیه‌گر، تحلیل نیاز گردشگران و پشتیبانی تصمیم‌گیری مورداستفاده قرار گیرند.

درنهایت، اگرچه مدل پیشنهادی عملکرد قابل‌توجهی از خود نشان داده است، اما توسعه آتی آن از طریق تنوع‌بخشی به منابع داده، مشارکت مستقیم بازیگران صنعت گردشگری و ارزیابی‌های میدانی گسترده‌تر می‌تواند به افزایش اثربخشی تجاری و مدیریتی این سیستم‌ها منجر شود. این مسیر، زمینه‌ساز پژوهش‌های آینده در تقاطع هوش مصنوعی، گردشگری و کسب‌وکار هوشمند خواهد بود.

سپاسگزاری

با سپاس فراوان از اساتید بزرگوارم که بسیار یاری گر من بوده‌اند.

تعارض منافع

تعارض منافع ندارم.

ORCID

Shirin Amin **https://orcid.org/0009-0000-4424-3969**
Mohammad Ali Afshar Kazemi **https://orcid.org/0000-0003-4327-8320**
Akbar Alam Tabriz **https://orcid.org/0000-0002-4562-8032**
Seyed Mohammad Ali Khatami Firouzabadi **https://orcid.org/0000-0003-0903-6184**

References

1. Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., & Wang, T. (2018). MS MARCO: A human generated machine reading comprehension dataset (No. arXiv:1611.09268). *arXiv*. https://doi.org/10.48550/arXiv.1611.09268
2. Fajcik, M., Jon, J., & Smrz, P. (2021). Rethinking the objectives of extractive question answering (No. arXiv:2008.12804, Version 4). *arXiv*. https://doi.org/10.48550/arXiv.2008.12804
3. Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). ParsBERT: Transformer-based model for Persian language understanding. *Neural Processing Letters*, *53*(6), 3831–3847. https://doi.org/10.1007/s11063-021-10528-4
4. Joshi, M., Choi, E., Weld, D., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1601–1611).AssociationforComputationalLinguistics. https://doi.org/10.18653/v1/P17-1147
5. Kazemi, A., Zojaji, Z., Malverdi, M., Mozafari, J., Ebrahimi, F., Abadani, N., Varasteh, M. R., & Nematbakhsh, M. A. (2023). FarsNewsQA: A deep learning-based question answering system for the Persian news articles. *Information Retrieval Journal*, *26*(1), 3. https://doi.org/10.1007/s10791-023-09417-2
6. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova,
7. Masumi, M., Majd, S. S., Shamsfard, M., & Beigy, H. (2024). FaBERT: Pre-training BERT on Persian blogs (No. arXiv:2402.06617).*arXiv*. https://doi.org/10.48550/arXiv.2402.06617
8. Mozafari, J., Kazemi, A., Moradi, P., & Nematbakhsh, M. A. (2022). PerAnSel: A novel deep neural network-based system for Persian question answering. *Computational Intelligence and Neuroscience*,*2022*(1),3661286. https://doi.org/10.1155/2022/3661286
9. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of

text(No.arXiv:1606.05250).*arXiv*. https://doi.org/10.48550/arXiv.1606.05250

10. Rosenthal, S., Sil, A., Florian, R., & Roukos, S. (2024). CLAPNQ: Cohesive long-form answers from passages in natural questions for RAG systems (No. arXiv:2404.02103). *arXiv*. https://doi.org/10.48550/arXiv.2404.02103
11. SadraeiJavaheri,M.,Moghaddaszadeh,A.,Molazadeh,M.,Naeiji, F.,Aghababaloo,F.,Rafiee,H.,Amirmahani,Z.,Abedini,T.,Sheikhi, F. Z., & Salehoof, A. (2024). TookaBERT: A step forward forPersianNLU(No.arXiv:2407.16382).*arXiv*. https://doi.org/10.48550/arXiv.2407.16382
12. Soh, Y. J., & Zhao, J. (2024). A step towards mixture of grader: Statistical analysis of existing automatic evaluation metrics (No. arXiv:2410.10030,Version1).*arXiv*. https://doi.org/10.48550/arXiv.2410.10030
13. Taghizadeh, N., Doostmohammadi, E., Seifossadat, E., Rabiee, H. R., & Tahaei, M. S. (2021). SINA-BERT: A pre-trained language model for analysis of medical texts in Persian (No. arXiv:2104.07613).*arXiv*. https://doi.org/10.48550/arXiv.2104.07613
14. Tran, S. Q., Do, G.-H., Do, P. N.-T., Kretchmar, M., & Du, X. (2023). AGent: A novel pipeline for automatically creating unanswerablequestions(No.arXiv:2309.05103).*arXiv*. https://doi.org/10.48550/arXiv.2309.05103
15. Wang, Z., Ng, P., Ma, X., Nallapati, R., & Xiang, B. (2019). Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP)*(pp.5878–5882). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1599
16. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in NaturalLanguageProcessing*(pp.2369–2380).Associationfor Computational Linguistics. https://doi.org/10.18653/v1/D18-1259

اســتناد به این مقاله: امینی، شــیرین.، افشــار کاظمی، محمدعلی.، عالم تبریز، اکبر.، خاتمی فیروزآبادی، ســید. محمدعلی. (۲۰۲۶). ارائه یک مدل هوش مصــنوعی برای اســتخراج پرســش–پاســخ فارســی در حوزه گردشــگری، *مطالعات مدیریت کسب وکار هوشمند*، ۱۵(۵۵)، ۳۷۵-۳۹۶. DOI: 10.22054/ims.2026.88566.2683