Business Intelligence Management Studies
Vol. 15, No.55, Spring 2026; P 307-336
ims.atu.ac.ir
DOI: 10.22054/ims.2026.83957.2573

ATU PRESS

Original Research
Received: 25 January 2025 Revised: 19 November 2025 Accepted: 03 January 2026
eISSN: 2821-0816
ISSN: 2821-0964

Designing a System for Matching the Name and Field of Activity of Companies Based on Artificial Intelligence

Mohammad Rabiei *

Associate Professor, Department of Electrical and Computer Engineering, Faculty of Engineering, Eyvanekey University, Eyvanekey, Semnan, Iran.

Abstract

Selecting the right company name isn't just a formality; it actually needs to fit what the company does. But the way things usually work, registration systems depend on manual checks or simple rule-based filters that only look at how similar the names are on the surface. This wastes time, leads to a lot of rejections, and makes things harder for everyone involved. That's where semantic analysis steps in. Instead of just matching words, it compares meaning, which is already a standard in natural language processing for things like search engines, text summarization, and even sentiment analysis. In this study, there's a new deep learning framework for checking how well a proposed company name matches its intended field. The method uses the AriaBERT model to turn company names into contextual vectors, and FastText to do the same with descriptions of company activities. Then, a deep learning setup with Bi-LSTM layers and an attention mechanism processes these vectors to pick up on the key semantic connections. The model measures how closely things match using cosine similarity and ROUGE metrics. To top it off, DBSCAN clustering groups together company names that relate to similar activities. The results speak for themselves: the model hit ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.7623, 0.7413, and 0.7982. The accuracy

* Corresponding Author: Mohammad.Rabiei@eyc.ac.ir
How to Cite: Rabiei, M. (2026). Designing a System for Matching the Name and Field of Activity of Companies Based on Artificial Intelligence, *Journal of Business Intelligence Management Studies*, 15(55), 307-336. DOI: 10.22054/ims.2026.83957.2573

landed at 0.8512, with a recall of 0.8317. The name-clustering function of the method also lets it recommend similar-sounding names for companies based on their area of activity.

Introduction

Digital tech has exploded in the last few years, and with it, we're swimming in unstructured text, think emails, reports, web pages, you name it. Businesses now lean hard on text processing and semantic analysis to make sense of all this data. Now, let's talk company names. Picking the right name is a huge step for any business or nonprofit. The name has to fit the field you're working in truly. Right now, most registration systems check proposed names either by hand or with rule-based tools. They mostly look for word overlap or make sure the name follows the rules. Because of this, tons of names get rejected. It's slow and creates extra work for everyone. And honestly, even though matching a company name to its actual business is important, there's no smart, automated way to check if the name really fits. That's where this study comes in.

We're rolling out an AI-driven method that actually evaluates how well a proposed company name matches its field of activity semantically, not just on the surface. Using text vectorization and deep learning, our approach aims to make the whole process faster and smarter, cut down on bad registrations, and help registration systems make better calls.

1. Theoretical Foundations and Hypothesis Development

Semantic similarity is a big topic in natural language processing. Basically, it's about figuring out how much two pieces of text, words, phrases, or whole sentences actually mean the same thing.

Neural networks, recurrent, convolutional, and now especially transformers, capture way more nuance by paying attention to context and connections across longer stretches of text. For Persian, these advances are even more crucial because the language is morphologically complex and there aren't big annotated datasets lying around .Other models, like ParsBERT and FaBERT, underscore how much difference good, clean pretraining data makes, especially for short or casual texts (Masumi et al., 2024; Zareshahi et al., 2024).

Given all this, this study starts from the idea that combining contextual embeddings with attention-based sequence models can do a

better job

of matching company names with what these companies say they do. So, the main hypothesis is this:

H1: A hybrid approach that mixes contextual embeddings with attention mechanisms significantly boosts the accuracy of detecting whether company names actually fit their stated fields of activity.

2. Materials & Methods

This study takes an applied research approach, focusing on a real-world case: making company name registration faster and more accurate.

They worked with both approved and rejected company names, along with the fields of activity those companies declared, which gave them plenty of material for analyzing meaning. The model itself has two main parts. First, it looks at the meaning of the company name.

Second, it checks how that name lines up with the declared field of activity. The whole point is to see if these two pieces of text actually fit together, which helps people make better decisions when registering new companies or institutions. As you can see in Figure 1, the model compares the company name and the activity field vectors using cosine similarity. To get there, they start by running the activity descriptions through a bunch of text cleaning steps, things like tokenization, padding, stemming, and normalizing Persian characters.

Still, FastText misses some context, like word order or how sentences are built, so it feeds its vectors into a BiLSTM layer to pick up on that. Then, they use an attention mechanism to highlight the most important words. After that, they run everything through an AriaBERT embedding layer, which gives them a single, context-rich vector for each activity. If the score's high enough above a set threshold, the name is considered a good fit for the declared activity, so it's accepted. If not, it's rejected. Finally, they use the DBSCAN algorithm to cluster company names based on how similar they are in meaning.

3. Results & Discussion

For the activity descriptions, I used FastText to create embeddings. To check how the model performed, I relied on cosine similarity and ROUGE metrics. I calculated ROUGE-1, ROUGE-2, and ROUGE-L scores for both company names and activity descriptions after

vectorization. The results showed the model didn't just keep track of word overlap, it actually picked up on deeper semantic meaning. ROUGE-L, in particular, scored high on precision, recall, and F1 for both company names and activity descriptions. Also clustered the company name vectors using DBSCAN, with a radius of 0.5 and at least five samples per cluster. To make sure the model's reliable, I ran a human evaluation following SemEval standards on 200 random samples. Experts agreed with each Most of the time, Cohen's Kappa came in at 0.87. Even better, model-generated similarity scores matched expert judgments with a correlation of 0.93.

4. Conclusion

A company's name isn't just a tag you slap on a business. Sometimes it tells you right away what the company does, other times it's just a catchy word with no clear meaning. In this study, we built an AI system to see how well a proposed company name matches up with what the business actually does.

When we compared these scores to the opinions of real company registration experts, the system matched them 93% of the time. That's a solid result—it shows the AI is doing what it's supposed to. Digging a bit deeper, ROUGE-L ended up being the best measure compared to ROUGE-1 and ROUGE-2.

Now, there are some limits. The system only works with Persian-language data from the national company registration system, and it just focuses on matching names to activity fields. Looking ahead, we want to try out online learning and more advanced transformer models. AI is moving fast, after all.

***Keywords*:** Company registration, Cosine similarity, Deep learning, Semantic relation, Text mining.

مقاله پژوهشی تاریخ ارسال: ۱۴۰۳/۱۱/۰۶ تاریخ بازنگری: ۱۴۰۴/۰۸/۲۸ تاریخ پذیرش: ۱۴۰۴/۱۰/۱۳ ISSN: 2821-0964 eISSN: 2821-0816

مطالعات مدیریت کسب و کار هوشمند
سال پانزدهم، شماره ۵۵، بهار ۱۴۰۵، ص ۳۰۷ تا ۳۳۶
ims.atu.ac.ir
DOI: 10.22054/ims.2026.83957.2573

طراحی سیستم سنجش تطابق نام و حوزه فعالیت شرکت‌ها بر اساس هوش مصنوعی

محمد ربیعی * استادیار، گروه مهندسی برق و کامپیوتر، دانشکده فنی و مهندسی، دانشگاه ایوان کی، ایوان کی، سمنان، ایران.

چکیده

تأیید نام در فرآیند ثبت تأسیس شرکت باعث می‌شود از ثبت‌نام شرکت‌ها که با زمینه فعالیت آن‌ها همخوانی ندارد جلوگیری به عمل آید. در این مقاله به‌منظور بررسی درصد تطبیق نام پیشنهادی متقاضیان ثبت شرکت با زمینه فعالیت شرکت روشی نوین بر اساس الگوریتم‌های یادگیری عمیق ترکیبی BiLSTM و یک لایه توجه تکمیلی برای افزایش دقت تشخیص ارائه شده است. داده‌های این پژوهش از سازمان ثبت‌اسناد و املاک کشور جمع‌آوری گردیده است. در روش پیاده‌سازی ابتدا از فیلترهای اولیه نام‌گذاری شرکت استفاده شده است. سپس با استفاده از ترکیب‌روش آریا برت به‌عنوان یک تکنیک تعبیه کلمات به تبدیل نام پیشنهادی شرکت به بردار پرداخته می‌شود. در مرحله‌ای موازی زمینه فعالیت شرکت را با استفاده از فستتکس به بردار عددی و تلفیق بردار به‌دست‌آمده با الگوریتم‌های یادگیری عمیق حافظه کوتاه و بلندمدت دوطرفه بر اساس یک لایه توجه اضافه می‌گردد. جهت ارزیابی نتایج از معیار شباهت‌کسینوسی و معیار روج (۱ و ۲ و آل) استفاده شده است. پس از تأیید پذیرش نام شرکت و زمینه فعالیت، از روش خوشه‌بندی دیبی اسکن برای خوشه‌بندی نام شرکت در دسته‌های فعالیت استفاده می‌شود. نتایج تحقیق نشان می‌دهد که مقادیر دقت در بخش بردار سازی زمینه فعالیت‌های شرکت برای معیار روج آل مقدار ۷۹۸۲.۰ و مقادیر دقت و فراخوانی نهایی مدل به ترتیب ۰٫۸۵۱۲، ۸۳۱۷.۰ محاسبه گردید. همچنین ضریب همبستگی بین شباهت کسینوسی محاسبه‌شده بین نام پیشنهادی و زمینه فعالیت شرکت با مقدار ۹۳ درصد بر اساس معیارهای تعیین نام نشان‌دهنده کارکرد درست مدل می‌باشد.

کلیدواژه‌ها: ارتباط معنایی، ثبت شرکت، شباهت کسینوسی، متن‌کاوی، یادگیری عمیق.

* نویسنده مسئول : Mohammad.Rabiei@eyc.ac.ir

مقدمه

پردازش متون، به‌عنوان حوزه‌ای پیشرو در داده‌کاوی، به استخراج الگوها و دانش پنهان از منابع متنی می‌پردازد. ازجمله وظایف اصلی پردازش متون می‌توان به طبقه‌بندی اسناد، خوشه‌بندی، استخراج مفاهیم کلیدی، تحلیل معنایی، تلخیص و شناسایی ارتباطات مفهومی میان عناصر متنی اشاره کرد. در این راستا، تعیین میزان تشابه متنی، به‌ویژه در سطوح معناشناختی، لغوی و ساختاری، نقشی اساسی در بهینه‌سازی کاربردهای فناوری ایفا می‌کند و به‌شدت موردتوجه قرار گرفته است. تحلیل شباهت‌های متنی، به‌ویژه در سطوح معناشناختی و واژگانی، امکان بررسی عمیق‌تر متون را فراهم می‌کند. اندازه‌گیری شباهت بین متون مختلف، می‌تواند به شناسایی مفاهیم هم‌معنی کمک نماید. این قابلیت در سامانه‌های پردازش زبان طبیعی مانند طبقه‌بندی متون، تلخیص خودکار و سیستم‌های پرسش-پاسخ اهمیت بسزایی دارد. تحقیقات نوین در حوزه پردازش زبان طبیعی، روش‌های متنوعی برای سنجش تشابه معناشناختی میان کلمات و جملات ارائه کرده‌اند که شامل روش‌های مبتنی بر پایگاه‌های واژگانی، مدل‌های بردار کلمه مبتنی بر شبکه‌های عصبی بازگشتی و شبکه‌های عصبی عمیق است. امروزه، الگوریتم‌های یادگیری ماشین و یادگیری ژرف به‌عنوان ابزارهای حیاتی در تجزیه‌وتحلیل متون، به‌طور گسترده‌ای در کاربردهای مختلف تحلیل محتوای متنی به کار گرفته می‌شوند.

انتخاب نام مناسب برای شرکت، مسئله‌ای حساس و مرحله‌ای تعیین‌کننده در فرآیند تأسیس یک شرکت یا مؤسسه غیرتجاری است. در این فرآیند، متقاضی، نام پیشنهادی خود را همراه با حوزه فعالیت شرکت به اداره تعیین نام ارسال می‌کند. کارشناسان اداره، پس از بررسی‌های اولیه لغوی با بررسی نام‌های ثبت‌شده قبلی، مشابهت نام پیشنهادی را ارزیابی کرده و در صورت نداشتن مشابه، آن را تأیید می‌کنند.

آمارها نشان می‌دهد که از میان ۳۴۸، ۹۲۷ درخواست نام‌گذاری شرکت در ده ماه نخست سال ۱۴۰۰، حدود ۲۲۷، ۹۲۶ نام رد شده و تنها ۲۰ درصد از نام‌ها تأیید شده‌اند (Baigi et al. 2023). مدت‌زمان میانگین برای تعیین نام تقریباً ۴ روز است که بخش عمده

آن به دلیل رد نام‌ها و پیشنهاد نام‌های جدید از سوی متقاضیان به طول می‌انجامد.

تحلیل شباهت متنی درباره سامانه ثبت شرکت‌ها می‌تواند چالش‌های متعددی را در حوزه‌های کسب‌وکار مرتفع کند. ازجمله کاربردهای کلیدی این حوزه، ارزیابی نام‌های پیشنهادی برای ثبت شرکت‌ها و مؤسسات غیرتجاری در سازمان ثبت‌اسناد و املاک است. در این مقاله، هدف اصلی روشی برای شناسایی میزان شباهت معناشناختی میان نام‌های پیشنهادی و زمینه فعالیت‌های اعلان‌شده برای شرکت‌ها و مؤسسات غیرتجاری معرفی می‌گردد. پس از بررسی‌های انجام‌شده بر اساس پایگاه داده سامانه ثبت شرکت‌ها و الگوریتم‌های یادگیری عمیق مدلی تدوین گردید تا به بهبود روند بررسی نام مناسب کمک نماید. همچنین از اهداف فرعی این پژوهش می‌توان به تأیید صحیح‌تر نام‌های پیشنهادی متقاضیان ثبت شرکت و خوشه‌بندی نام‌ها و فعالیت‌های مشترک شرکت‌ها در خوشه‌های مرتبط اشاره نمود. با توسعه فناوری‌های نوین و افزایش چشمگیر حجم داده‌های دیجیتال، داده‌های متنی بدون ساختار به‌طور گسترده‌ای اهمیت یافته‌اند. تدوین روش‌ها و الگوریتم‌های کارآمد برای استخراج دانش از این داده‌ها بسیار مهم شده است. با توجه به این‌که تاکنون هیچ فیلتری برای ارزیابی تطابق نام شرکت با حوزه فعالیت وجود نداشته است، برخی شرکت‌ها با نام‌هایی که تطابق مفهومی مناسبی با حوزه کاری خود ندارند، به ثبت رسیده‌اند. ضرورت این پژوهش را می‌توان جلوگیری از به ثبت رسیدن نام‌های مناسب بر مبنای الگوریتم‌های هوش مصنوعی در سامانه ثبت شرکت‌ها بر شمرد. این پژوهش، با بهره‌گیری از الگوریتم‌های بردارسازی عددی و یادگیری ژرف، امکان ارزیابی دقیق‌تر تطابق نام‌های پیشنهادی با حوزه فعالیت شرکت‌ها را فراهم می‌کند. این روش می‌تواند از ثبت‌نام‌های نامرتبط جلوگیری کرده و از این طریق به بهبود فرآیند انتخاب نام در سامانه ثبت شرکت‌ها کمک کند. این مقاله در بخش ۲ به ادبیات تحقیق و پیشینه پژوهش و مطالعات انجام‌شده پرداخته است. در بخش ۳ مدل ارائه‌شده و روش پیشنهادی بیان شده است. بخش ۴ نتایج و ارزیابی روش پیشنهادی و بخش ۵ دربرگیرنده نتیجه‌گیری و پیشنهادهای مطالعاتی می‌باشد.

پیشینه پژوهش

فعالیت شرکت بر انتخاب نام شرکت تأثیر ویژه‌ای دارد. نام شرکت‌ها به‌عنوان یکی از عناصر کلیدی در بازنمایی هویت و ساختار سازمانی، در ایجاد تصویر ذهنی اولیه مخاطبان نقش بسزایی دارد. انتخاب نام معنادار و مرتبط با حوزه فعالیت شرکت نه‌تنها

به یادسپاری بهتر کمک می‌کند بلکه نقش مؤثری در بازاریابی و جذب مخاطب ایفا می‌نماید. نام ثبت‌شده یک شرکت، به‌عنوان هویت حقوقی آن، در مرجع ثبت شرکت‌ها به تأیید رسیده و یا در فرآیند تأسیس می‌باشد. در مقابل، نام پیشنهادی، نامی است که متقاضی در زمان تأسیس یا درخواست تغییر نام به مرجع ثبت شرکت‌ها ارائه می‌نماید. همچنین، تعریف زمینه فعالیت شرکت به‌منظور تعیین مرزها و حدود فعالیت‌های شرکت، نشان‌دهنده هویت و اهداف بنیادی آن است. از منظر حقوقی و رعایت اصول تجاری، در انتخاب نام شرکت، استفاده از واژگان ایرانی و معنادار، پرهیز از مشابهت با نام‌های اشخاص حقوقی ثبت‌شده، اجتناب از کاربرد اصطلاحات دولتی و عدم استفاده از عناوین غیرقانونی و واژگان بیگانه و نام‌های حکومتی، از الزامات مهم به شمار می‌آید. درباره برخی از موارد لازم‌الاجرا به‌غیراز موارد فوق در جدول ۱- توضیحاتی داده شده است. رایط زیر که در فرهنگستان با نام انواع جناس شناخته می‌شود باید رعایت کامل گردند.

جدول ۱. نکات اولیه موردبررسی و ارزیابی جهت تعیین نام تجاری

اسم شرکت	اسمی که هم در معنی و هم تلفظ با اسامی قبلی برابری می‌کند	فناوران صنعت روز-فناوران صنعت روز
اسم یکسان نوشتاری	معانی متفاوت برای دو اسم کاملاً یکسان	مَلِک، مُلک، مَلَک
اسم یکسان لفظی	دو اسم در تلفظ مثل هم در نوشتار متفاوت	تهران و طهران
اسم‌های برتر	حالت عالی قرار دادن برای برخی از اسم‌ها	تفصیلی، عالی مثل برترین

تحلیل تشابه معنایی، فرآیندی پیچیده و چندلایه در حوزه پردازش زبان طبیعی است که به شناسایی هم‌پوشانی معنایی میان ساختارهای جملات مختلف می‌پردازد. این تحلیل با محاسبه نزدیکی معنایی و سنجش درجه تطابق محتوایی واژگان در جملات، امکان تعیین

هم‌معنایی و مفهوم‌گرایی مشترک را میان متون مختلف فراهم می‌آورد.

پژوهش‌های متعدد نشان داده‌اند که روش‌های متنوعی برای ارزیابی میزان شباهت میان متون وجود دارند که در کاربردهایی چون کشف و جلوگیری از سرقت ادبی و تحلیل متون، کارآمدی خود را به اثبات رسانده‌اند. روش‌های رایج در این حوزه عبارت‌اند از: تحلیل شباهت لغوی بر مبنای مقایسه واژگان، محاسبه شباهت تکرار واژگان مشترک و ارزیابی مفهومی از طریق تحلیل معنای کلمات که هرکدام مزایا و محدودیت‌های مختص خود را دارا هستند. مزیت اصلی تحلیل شباهت لغوی سرعت‌بالا است، اما در متون بازنویسی‌شده و تغییریافته، روش‌های مفهومی دقت بیشتری را به همراه دارند (Khan and Anjum 2023).

در حوزه معیارهای سنجش تشابه (Dogan, Goru Dogan, and Bozkurt 2023)، معیارهای فاصله‌ای نقش اساسی در ارزیابی میزان هم‌پوشانی میان موجودیت‌های معنایی بازی می‌کنند در اصل، معیارهای فاصله و شباهت رابطه‌ای معکوس دارند؛ یعنی با افزایش شباهت، میزان فاصله مفهومی کاهش می‌یابد. از معیارهای معمول در این حوزه می‌توان به شباهت کسینوسی اشاره کرد که با محاسبه کسینوس زاویه میان دو بردار کلمه‌ای یا بردار جمله‌ای تعیین می‌کند که آیا این بردارها از هم‌جهتی مفهومی برخوردارند یا خیر. این معیار به‌ویژه در تحلیل شباهت متون و مستندات بسیار کاربردی است. علاوه بر آن، معیارهایی چون فاصله اقلیدسی، فاصله منهتن و شاخص شباهت ژاکارد نیز در حوزه تحلیل معنایی متون به‌طور گسترده به کار گرفته شده‌اند. در سال‌های اخیر، تحقیقات در حوزه پردازش زبان طبیعی به توسعه روش‌های پیشرفته‌ای برای محاسبه شباهت معنایی میان واژگان و جملات منجر شده است. این روش‌ها شامل تکنیک‌های مبتنی بر هم‌وقوعی واژگان، روش‌های مبتنی بر پایگاه داده‌های واژگانی و روش‌های مبتنی بر یادگیری عمیق مانند شبکه‌های عصبی بازگشتی و شبکه‌های عصبی پیچشی است. در ادامه، به بررسی برخی از این تحقیقات با تمرکز بر زبان فارسی پرداخته خواهد شد. حسینی‌مقدم و همکاران (Hosseini et al. 2021) رویکردی برای سنجش شباهت معنایی در متون کوتاه

فارسی که از برنامه‌های تلویزیونی به‌صورت جفتی شامل ۳۵۲۶۶ جفت جمله ساده ارائه داده‌اند. روش پیشنهادی شامل سه مرحله است. اولین گام جمع‌آوری داده‌ها و ساختن یک پیکره موازی است. در مرحله بعدی یعنی مرحله پیش‌پردازش، داده‌ها نرمال می‌شوند. نتایج این تحقیق نشان می‌دهد که روش پیشنهادی با معیار اندازه‌گیری اف، به‌دقت ۳۷.۷۸ درصد برای داده‌ها[1] واژه پشت سر هم در یک جمله و ۶۵,۹۸ درصد برای داده‌های چهار واژه پشت سر هم دست یافته است.

صادقی پور و همکاران (Sadidpour et al. 2022) با بهره‌گیری از شبکه‌های حافظه کوتاه بلندمدت دوطرفه، روشی را برای ارزیابی میزان شباهت معنایی جملات فارسی ارائه داده‌اند. در این روش، نگاشت برداری جملات به فضای برداری با استفاده از شبکه‌های عصبی انجام‌شده و نتایج آن بر اساس امتیازات انسانی سنجیده شده است؛ دقت این سیستم پیشنهادی حدود ۲.۸۹ درصد ارزیابی شده است.

تحقیقات غفوری و همکاران (Ghafouri, Abbasi, and Naderi 2023) بر آریابرت[2] انجام پذیرفت. آن‌ها آریابرت را به‌عنوان یک مدل زبان از پیش آموزش‌دیده برای فارسی معرفی کردند. همچنین به مطالعه در زمینه کمبود داده‌های متنی متنوع و مدل‌های از پیش آموزش‌دیده کارآمد در زبان فارسی پرداختند. آریابرت بر روی مجموعه داده‌های متنوعی از متون فارسی، ازجمله متون محاوره‌ای، رسمی و ترکیبی مانند توییت‌ها، اخبار، شعرها، متون پزشکی و موارد دیگر آموزش داده شد که مجموعاً بیش از ۳۲ گیگابایت است. برخلاف دیگر مدل‌هایی که از برت استفاده می‌کنند، آریابرت از معماری روبرتا[3] استفاده می‌کند. در مقایسه با مدل‌های موجود زبان فارسی در بین وظایف مختلف پردازش زبان طبیعی برتری قابل‌توجه آریابرت را با میانگین بهبود ٪۳ در تحلیل احساسات ٪۰,۶۵ در طبقه‌بندی و ٪۳ در تشخیص موضع نشان داده شده است.

1 Bidirectional encoder representations from transformers (BERT)
2 AriaBert
3 Roberta

در سال ۲۰۲۳ مهربان و احدیان (Mehrban and Ahadian 2023) در ارزیابی تبلیغات متنی فارسی، با تمرکز بر تأثیر اینترنت بر تجارت مدرن و ارزش داده‌های تراکنش برای بهبود بازاریابی، همکاری کردند. آن‌ها سایت دیوار را بررسی کردند و مسابقه‌ای را برای پیش‌بینی درصد متن انتشار آگهی فروش خودرو در فروش واقعی آن ترتیب دادند. با استفاده از کتابخانه هزم[1] و مدل‌های زبانی پیشرفته‌ام برت[2] و پارس برت[3] به تجزیه‌وتحلیل داده‌ها، تنظیم دقیق مدل و تنظیمات آموزشی پرداختند. نقاط قوت و ضعف مدل را برجسته داده‌کاوی و تکنیک‌های یادگیری ماشین را موردبحث قرار داده‌اند.

منیری و همکاران با بررسی جامع چالش‌ها و فرصت‌های موجود در زبان فارسی، پرداختند. نتایج مطالعه بیانگر آن است که مدل‌های ترنسفورمرهای آموزش‌دیده بر داده‌های گسترده فارسی به‌طور قابل‌توجهی توانایی پوشش واژگانی و درک معنایی را دارند. توسعه و پاک‌سازی متنوع و انسجام واژگان می‌تواند به‌طور مستقیم کیفیت مدل‌های بازیابی اطلاعات فارسی را ارتقا دهد و در جست‌وجو، تحلیل محتوا و پردازش زبان طبیعی فارسی را هموار سازد (Moniri et al., 2024).

زارع شاهی و همکاران (Zareshahi, Javadzade, and Bastami 2024) بر روی آنالیز جملات کوتاه فارسی در شبکه‌های اجتماعی کار کردند. آن‌ها در مطالعات خود از مدل زبان پارسبرت به‌عنوان مدل پایه استفاده کرده و برای آنالیز سریع‌تر، جملات نشانه‌گذاری شده را به بردارهای معنی‌دار تبدیل نمودند. از طریق لایه‌های ادغام، بردارهای بهینه استخراج و درنتیجه بردارهای ۷۶۸ بعدی برای هر جمله ایجاد کرده‌اند. نتایج نشان می‌دهد که مدل پیشنهادی به ضریب همبستگی پیرسون تقریبی ۰,۶۸ دست یافت که از مدل‌های قبلی فستکس[4] و یک مدل شبکه عصبی پیچشی[5] پیشی گرفته است و نشان‌دهنده عملکرد برتر آن نسبت به مدل‌های قبلی است.

---

1 Hazm
2 mBERT
3 ParsBERT
4 FastText
5 Convolutional Neural Network (CNN)

در تحقیقات شباهت‌های معنایی محاوره‌ای سجادی(Sadjadi et al. 2024) در زمینه جملات غیررسمی و عامیانه در شبکه‌های اجتماعی برای بررسی شباهت جملات از مدل ترانسفورمر[1] استفاده کرده است.

این گروه پایگاه داده جدیدی به نام فارس سیم[2] با تعداد ۱۰۴ میلیون جمله کوتاه محاوره‌ای را برای تحقیقات خود تهیه نموده‌اند. مدل ارائه‌شده بر اساس ترنسفورمر با تعداد لایه‌های توجه بالا بر اساس معیارهای ضریب پیرسون ۷۷.۰ و اسپیرمن ۰,۶۴ بهتر از برت و پارس برت چند زبانه عمل می‌کند. همچنین نقطه برتری دیگر نتایج را در مدل زبان بزرگ از پیش آموزش دیده‌شده برای استفاده در سایر کارهای پردازش زبان طبیعی در متن محاوره‌ای و به‌عنوان نشانه‌ای برای کلمات غیررسمی کمتر شناخته‌شده معرفی نموده‌اند.

همچنین عبدوس و همکاران برای نخستین‌بار یک ارتباط دوسویه و دوزبانه با نام PESTS شامل ۵۳۷۵ جفت‌جمله فارسی–انگلیسی با برچسب‌گذاری انسانی تهیه کردند. این پایگاه‌داده امکان آموزش و ارزیابی مستقیم مدل‌های شباهت معنایی میان‌زبانی را بدون اتکا به ترجمه ماشینی فراهم می‌کند و از انتشار خطاهای ترجمه جلوگیری می‌شود. نتایج نشان داد که با استفاده از این پایگاه، عملکرد مدل XLM-RoBERTa در معیار پیرسون از ۸۵,۸۷٪ به ۹۵,۶۲٪ افزایش یافته است (Abdous et al., 2024).

مدل FaBERT باهدف پوشش شکاف موجود در پردازش زبان طبیعی فارسی و افزایش توانایی در درک متون رسمی و غیررسمی، توسط Masumi et al. (۲۰۲۴) معرفی شده است. این مدل یک نسخه BERT-base است که از ابتدا بر روی کورپوس HmBlogs شامل حجم گسترده‌ای از وبلاگ‌های فارسی با سبک‌های مختلف آموزش داده شده است. در ارزیابی‌های انجام‌شده روی ۱۲ دیتاست در وظایف مختلف مانند تحلیل احساس، تشخیص موجودیت نامدار، استنتاج طبیعی زبان، پرسش‌پاسخ و تشخیص پارافرایز، نتایج نشان داد که FaBERT در اغلب وظایف، عملکرد بهتری نسبت به مدل‌های

1 Transformer
2 Farsim

هم‌رده مانند (ParsBERT و mBERT) دارد.

باوجود حجم کوچک مدل، پایداری بالایی در نتایج ارائه می‌کند، نشان می‌دهد که استفاده از داده‌های متنوع، محاوره‌ای و پاکسازی‌شده اثر چشمگیری بر افزایش دقت مدل‌های زبانی فارسی دارد. به‌طورکلی، این پژوهش ثابت می‌کند که آموزش مدل‌های زبانی فارسی بر روی مجموعه‌داده‌های غنی و چندسبکی می‌تواند به‌طور مستقیم کیفیت آن‌ها را در وظایف متنوع NLU افزایش دهد.

در شکاف تحقیقات انجام‌شده، مطالعات کمی در مورد تشخیص خودکار تناقض در تطابق معنایی نام شرکت و زمینه فعالیت‌های شرکت در فرآیندهای ثبت وجود دارد. بااین‌حال، روش‌های تشخیص تضاد در برنامه‌های مختلف پردازش زبان طبیعی در دنیا وجود دارد و هر روز در حال گسترش می‌باشد. پایگاه‌های داده بزرگ و استاندارد برای فارسی در حوزه شباهت معنایی محدود است و این امر دقت مدل‌ها را تحت تأثیر قرار می‌دهد. در پژوهش‌های پیشین نیز کمبود منابع متنی غنی فارسی به‌عنوان یک چالش اصلی مطرح شده است.

بخش عمده‌ی داده‌های بازیابی‌شده مربوط به چند حوزه پرتکرار (فناوری، بازرگانی، خدمات) بوده و برخی حوزه‌ها داده کافی نداشته‌اند؛ این امر ممکن است منجر به سوگیری در بردارسازی واژگان شود. از سوی دیگر نام‌ها در ایران تحت تأثیر فرهنگ، قومیت، شهر، جنسیت و ساختارهای واژگانی خاص هستند و این سوگیری‌ها می‌توانند نتایج مدل‌های NLP را تحت تأثیر قرار دهند.

روش‌های پیشنهادی در این پژوهش شامل متن‌کاوی بر اساس فیلترسازی اولیه و تولید بردار ویژگی متن گلاو[1] و استفاده از الگوریتم حافظه کوتاه و بلندمدت دوطرفه و یک‌لایه توجه[2] اضافه با استفاده از پارس برت است. در بخش دوم الگوریتم با استفاده از معیار شباهت کوسینوسی به خوشه بندی نام شرکتها براساس زمینه فعالیت پرداخته خواهد شد.

1 GloVe
2 Attention layer

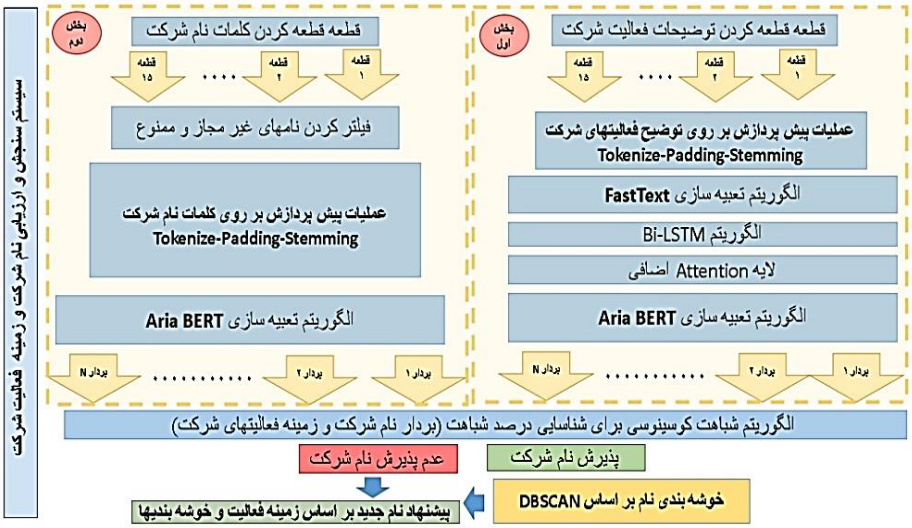
روش پیاده‌سازی

این پژوهش کاربردی و در دسته تحقیقات مطالعه موردی دسته‌بندی می‌گردد. برای پیاده‌سازی این مدل، از داده‌های ثبت‌شده شامل نام‌های تأییدشده و رد شده در بانک اطلاعاتی ثبت شرکت‌ها استفاده شده است.

این مدل با ترکیب رویکردهای یادگیری ماشین و تکنیک‌های پردازش زبان طبیعی برای بردارسازی کلمات توسعه‌یافته و با بهره‌گیری از روش‌های خوشه‌بندی به شناسایی نزدیک‌ترین نام‌ها و محاسبه شباهت معنایی می‌پردازد.

ساختار مدل پیشنهادی شامل دو بخش تحلیل نام پیشنهادی و ارزیابی زمینه فعالیت شرکت است. هدف از این مدل، تسریع و بهینه‌سازی فرآیند ثبت شرکت‌ها و مؤسسات غیرتجاری است. در شکل ۱- جزئیات روش پیاده‌سازی مدل بررسی شباهت معنایی نام شرکت و زمینه فعالیت شرکت در ثبت مؤسسات غیرتجاری نشان داده شده است. همچنین تعیین میزان شباهت معنایی نام پیشنهادی با زمینه فعالیت شرکت درنهایت توسط الگوریتم شباهت کوسینوسی محاسبه شده است.

شکل ۱. روش پیاده‌سازی مدل بررسی شباهت معنایی نام شرکت‌ها و مؤسسات غیرتجاری



در مرحله نخست، به‌منظور تبدیل فعالیت‌های شرکت به بردارهای عددی و استخراج ارتباط معنایی میان واژگان، توضیحات متنی فعالیت‌ها پس از انجام عملیات پیش‌پردازش شامل Tokenizing، Padding و Stemming آماده‌سازی شدند. سپس این داده‌ها با استفاده از مدل FastText که برای زبان فارسی آموزش داده شده است، به بردارهای چندبعدی تبدیل گردید. به دلیل ماهیت جمله‌محور و وجود تنوع واژگانی در فعالیت‌های شرکت، بردارهای FastText برای استخراج معنای عمیق‌تر وارد یک لایه BiLSTM شدند تا وابستگی‌های معنایی بلندمدت حفظ شود. در ادامه، یک لایه Attention اضافه شد تا مدل بتواند بر واژگان مهم‌تر در متن فعالیت تمرکز کند. خروجی این ماژول سپس در یک لایه تعبیه‌ساز مبتنی بر AriaBERT تقویت شد تا ترکیبی از ویژگی‌های زبانی مبتنی بر مدل‌های آماری و مدل‌های ترنسفورمری فراهم آید. نتیجه نهایی هر فعالیت به‌صورت یک بردار واحد در ماتریس ویژگی ذخیره شد. در بخش دوم، ابتدا کلمات تشکیل‌دهنده نام پیشنهادی شرکت استخراج و پس از پیش‌پردازش (Tokenizing، Padding و Stemming)، نام‌هایی که در فهرست نام‌های تکراری، غیرمجاز یا ممنوعه قرار داشتند فیلتر شدند. سپس بردارسازی نام شرکت با استفاده از AriaBERT انجام گرفت تا با حفظ ساختار معنایی کلمات، یک نمایش برداری دقیق از نام پیشنهادی به‌دست آید. این خروجی یک ماتریس برداری برای هر نام است. در مرحله بعد، جهت سنجش شباهت معنایی بین «بردار نام شرکت» و «بردار حوزه فعالیت»، از الگوریتم شباهت کسینوسی استفاده شد. هرگاه امتیاز شباهت از آستانه تعیین‌شده بیشتر باشد، نام پیشنهادی با زمینه فعالیت شرکت «مرتبط» تلقی شده و قابل پذیرش است. در غیر این صورت، نام پیشنهادی فاقد ارتباط معنایی بوده و رد می‌شود. در بخش نهایی مدل، از الگوریتم خوشه‌بندی DBSCAN[1] برای دسته‌بندی نام‌های شرکت بر اساس حوزه فعالیت استفاده شد.

این کار موجب شد نام‌های مشابه در خوشه‌های معنایی مشترک قرار گیرند و امکان

1 Density-based spatial clustering of applications with noise (DBSCAN)

پیشنهاد نام‌های جایگزین مرتبط فراهم شود. برای ارزیابی کیفیت خوشه‌بندی، از شاخص Davies–Bouldin (Ros, Riad & Guillaume, 2023)[1] استفاده شد که میزان جدایی خوشه‌ها و نزدیکی درون‌خوشه‌ای را با دقت مناسبی نشان می‌دهد.

FastText بردارِ هر واژه را بر اساس n-gram های کاراکتری می‌سازد؛ این ویژگی آن را برای زبان فارسی – که دارای صرف پیچیده و شکل‌های نوشتاری متنوع است – مناسب می‌کند و مشکل OOV را کاهش می‌دهد. بااین‌حال، FastText به‌تنهایی ترتیب واژه‌ها و ساختار جمله را مدل نمی‌کند. برای این هدف، بردارهای FastText به لایه‌ی BiLSTM داده شده‌اند تا وابستگی‌های دوطرفه‌ی بین واژگان و الگوهای نحوی–معنایی در سطح جمله استخراج شوند. سپس یک لایه Attention افزوده شده است تا مدل بتواند بر واژگانی که در تشخیص حوزه‌ی فعالیت نقش کلیدی دارند (مثلاً «گردشگری»، «داده‌پرداز»، «پیمانکاری تأسیسات») وزن بیشتری بدهد.

بدین ترتیب، نمایش نهایی جمله به‌صورت یک بردار وزن‌دهی‌شده از کل توالی به‌دست می‌آید. به‌طور خلاصه، ترکیب AriaBERT با FastText و BiLSTM+Attention صرفاً یک چیدمان تصادفی از مدل‌ها نیست، بلکه پاسخی است به سه نیاز هم‌زمان:

۱-پوشش بهتر واژگان و ساختار صرفی زبان فارسی در حوزه‌ی فعالیت‌ها (FastText)،

۲- مدل‌سازی توالی و وابستگی‌های معنایی در سطح جمله (BiLSTM+Attention)،

۳- استخراج نمایش زمینه‌مند و غنی برای نام‌های کوتاه و ترکیب آن‌ها در یک فضای مشترک معنایی (AriaBERT + cosine similarity).

جمع‌آوری مجموعه داده

در این پژوهش، تعداد ۱, ۷۶۱, ۰۵۵ نام شامل نام‌های تأییدشده، ردشده و اطلاعات حوزه فعالیت شرکت‌ها از پایگاه داده ثبت شرکت‌ها و مؤسسات غیرتجاری استخراج شده است.

1 Davies-Bouldin

با توجه به ماهیت زمانی داده‌های ثبت شرکت‌ها و امکان تکرار نام‌ها و حوزه‌های مشابه در بازه‌های زمانی نزدیک، برای جلوگیری از نشت اطلاعات، تقسیم داده‌ها به سه مجموعه آموزش، اعتبارسنجی و آزمون به‌صورت صرفاً تصادفی انجام نشد. در عوض، کل رکوردها بر اساس تاریخ ثبت مرتب و سپس به‌صورت زمانی تفکیک شدند؛ به این صورت که سهم آموزش از بخش ابتدایی بازه زمانی و مجموعه‌های اعتبارسنجی و آزمون از بخش‌های میانی و انتهایی بازه انتخاب گردید. این روش باعث می‌شود مدل بر داده‌های گذشته آموزش ببیند و بر روی داده‌های جدیدتر آزمون شود و بدین ترتیب، شرایط استقرار واقعی مدل در سامانه ثبت شرکت‌ها نام‌های موجود با برچسب‌های مختلف ازجمله تأییدشده، ردشده و نام‌های غیرمجاز (اعلام‌شده توسط نهادهایی نظیر مراجع قضایی و وزارت فرهنگ و ارشاد اسلامی و همچنین فرهنگستان زبان و فایلی از مجلس شورای اسلامی) در فایل‌های مجزا به‌منظور آماده‌سازی برای مراحل پردازش نگهداری شده‌اند.

پیش‌پردازش

پیش‌پردازش پردازش زبان طبیعی محسوب می‌شود. در این مرحله، عملیات پیش‌پردازشی شامل استانداردسازی یکی از مراحل اساسی و اثرگذار کدهای یونیکد فارسی و عربی، اصلاح و یکسان‌سازی انواع فاصله‌های موجود، حذف فواصل اضافی در ابتدا و انتهای نام‌ها و حذف نام‌های پرت بر روی داده‌ها اعمال گردید. جزئیات برخی از این عملیات در جدول ۲- نشان داده شده است. درنهایت، پس از اعمال عملیات پیش‌پردازشی، تعداد ۱,۷۴۶,۹۷۷ نام به‌عنوان مجموعه داده نهایی انتخاب شدند.

جدول ۲. مرحله پیش‌پردازش یکسان‌سازی

اصلاح نوشتاری بر اساس حروف فارسی و عربی	حرف الف چندین شکل أ -آ -إ -ا -أ
توجه به عدم باقیماندن ایست واژه‌ها در متن	
توجه به فاصله‌ها و نیم فاصله‌ها و حذف آن‌ها	
توجه به حداقل ۴ سیلاب و حداکثر کمتر از ده سیلاب	نام‌های درخواستی معمولاً ۴ سیلابی می‌باشند
توجه به ساختار کامات و ریشه آن‌ها	جستجو جهت پیدا کردن ریشه هر کلمه

تعبیه سازی کلمات

تعبیه کلمات به‌عنوان یک روش کلیدی در نگاشت متون به بردارهای عددی به شمار می‌آید. ازآنجایی‌که سیستم‌ها و رایانه‌ها قادر به تحلیل و پردازش مستقیم متون و اسناد نیستند، تبدیل این متون به بردارهای عددی قابل‌فهم برای رایانه‌ها از اهمیت بالایی برخوردار است. در این راستا، تکنیک تعبیه کلمات به‌کار می‌رود که به تولید بردارهای چندبعدی می‌پردازد و هدف آن ثبت و ضبط معنای واژه‌ها و محتوای آن‌ها از طریق مقادیر عددی است. هر مجموعه‌ای از اعداد می‌تواند به‌عنوان یک «بردار کلمه» معتبر تلقی شود؛ اما تنها مجموعه‌هایی از این بردارها برای کاربردهای خاص ما مفید خواهند بود که روابط میان آن‌ها و محتوای متنوع کلمات را به‌طور طبیعی منتقل سازند. به‌طورکلی، نزدیکی این بردارها به یکدیگر بیانگر شباهت بیشتر میان کلمات است.

خوشه‌بندی بردار نام‌ها

پس از تبدیل نام‌های ثبت‌شده به بردارهای عددی، به‌دلیل عدم وجود برچسب برای این بردارها، ضرورت دارد از الگوریتم‌های خوشه‌بندی به‌منظور دسته‌بندی این بردارها بهره

برداری شود. در این تحقیق، از روش خوشه‌بندی دیبی اسکن برای خوشه‌بندی بردارهای مرتبط با یکدیگر استفاده شده است. این الگوریتم با شناسایی ساختارهای دانه‌ای در داده‌ها، امکان گروه‌بندی نام‌ها و زمینه فعالیت‌های مشابه را فراهم می‌آورد.

ارزیابی مدل

جهت ارزیابی نتایج روش پیشنهادی از معیار شباهت کسینوسی و دقت و فراخوانی برای نمرات روج یک[1]، روج دو[2] و روج آل[3] استفاده شده است Barbella and Tortora (2022). هرچه میزان شباهت افزایش یابد، می‌توان نتیجه‌گیری کرد که فاصله‌ی بین دوشی

1 ROUGE-1
2 ROUGE-2
3 ROUGE-L

کاهش یافته است. معیار شباهت کسینوسی بر اساس محاسبه کسینوس زاویه میان دو بردار، تعیین می‌کند که آیا دو بردار تقریباً در یک راستا قرار دارند. این معیار اغلب برای سنجش شباهت اسناد در تحلیل متن استفاده می‌شود.

بر اساس رابطه ۱) در صورت انطباق دو بردار زاویه بین دو بردار صفر است و نتیجه آن شباهت کامل است.

از سوی مخالف اگر زاویه را ۱۸۰ درجه ارزیابی نماییم دو بردار در کمترین میزان شباهت می‌باشند. از طرف دیگر در زمینه بررسی فعالیت‌های شرکت و عملیات خلاصه‌سازی آن‌ها با استفاده از روج ان تعداد-n گرم تطبیق بین متن تولیدشده توسط مدل و مرجع تولیدشده توسط انسان را اندازه‌گیری می‌کنیم. در زمینه ارزیابی نتایج خلاصه‌سازی فعالیت‌های شرکت از روج استفاده می‌شود از مزایای آن می‌توانیم به همبستگی مثبت با ارزیابی انسان و محاسبه ارزان و مستقل از زبان اشاره نمود. برای ارزیابی بخش خلاصه‌سازی و توصیف فعالیت شرکت‌ها، از معیار ROUGE استفاده شد؛ چراکه ROUGE مبتنی بر هم‌پوشانی n-gram بوده و برای سنجش شباهت در سطح واژگان است.

$$\cos(x,y)= x.y/ \|x\|.\|y\| \qquad (۱$$

اما به‌دلیل ماهیت معنایی مسئله—یعنی سنجش ارتباط بین نام شرکت و حوزه فعالیت—ROUGE معیار قابل اتکایی برای تحلیل سطح مفهوم و رابطه معنایی محسوب نمی‌شود. بنابراین، در بخش ارزیابی اصلی مدل، از معیارهای حوزه Semantic Similarity شامل Cosine Similarity برای سنجش نزدیکی بردارها، Pearson و Spearman برای همبستگی معنایی و همچنین شاخص‌های سنجش عملکرد تصمیم‌یار استفاده شده است.

یافته‌ها

مجموعه داده‌ای متشکل از ۱۷۶۱۱۰۰ نام از سامانه ثبت شرکت‌ها و مؤسسات غیرتجاری جمع‌آوری کردیم. مجموعه داده به دو دسته تقسیم شده‌اند. ۱۲۷۰۵۰۰ نام تأییدشده و ۴۹۰۶۰۰ نام تأیید نشده/غیرمجاز یا رد شده می‌باشد. جدول ۳ توزیع داده‌ها را برای

آموزش مدل، اعتبارسنجی و آزمایش نهایی نشان می‌دهد ۸۰٪ برای آموزش، ۱۰٪ برای اعتبار سنجی و ۱۰٪ باقی‌مانده برای آزمایش استفاده می‌شود.

جدول ۳. پایگاه داده مورداستفاده در تحقیق

پایگاه داده	آموزش (۸۰ درصد)	اعتبار سنجی (۱۰ درصد)	آزمایش (۱۰ درصد)
نام و زمینه فعالیت شرکت ۱۷۶۱۱۰۰	۱۴۱۰۰۰۰	۱۷۵۵۵۰	۱۷۵۵۵۰

بررسی شباهت نام شرکت و زمینه فعالیت شرکت

با توجه به اینکه نام و زمینه فعالیت ارسالی از چند کلمه تشکیل شده است، از طریق بردار کلمات ساخته‌شده بردار نام پیشنهادی و بردار زمینه فعالیت شرکت به عناوین ورودی به مدل داده می‌شود. در جدول ۴-پارامترهای به‌دست‌آمده تحت روش آریا برت برای نام شرکت و فستتکس برای زمینه فعالیت‌های شرکت جهت آموزش بر روی مجموعه داده آورده شده است. لازم به ذکر است که برای آریابرت مقداراندازه دسته‌ها[1] را ۵۰ و الگوریتم بهینه‌ساز آدام[2] و دراپ اوت[3] مقدار ۰,۲ در نظر گرفته شده است.

جدول ٤. پارامترها و مقادیر روش‌های ساخت بردار کلمات

پارامترها	مقادیر تنظیمی	
	Fast text	Aria BERT
نرخ یادگیری	۰,۰۶	۰,۰۰۰۵
ابعاد بردار	۱۰۰	ابعاد پیش‌فرض ۷۶۸
تعدادepoch	۶۰	۱۰ یا مقدار واقعی پروژه شما؛ مقدار ۲۰۰ برای BERT صحیح نیست و باید منطقی باشد
Window Size	۵	در معماری BERT وجود ندارد
MinCount	۳	—
مدل	Skip-	—

1 Batch Size
2 Adam
3 Dropout

پارامترها	مقادیر تنظیمی	
	Fast text	Aria BERT
	Gram	
Max Sequence Length	—	۱۲۸ یا مقدار واقعی
Batch Size	—	۳۲

همان‌گونه که در جدول ۴ - نشان داده شده است، پارامترهای مدل FastText شامل مقادیر پایه‌ای مانند window size، mincount و ساختار skip-gram هستند که مختص مدل‌های تعبیه‌ساز واژگان مبتنی بر زیرواحدهای واژگانی می‌باشند. در مقابل، مدل AriaBERT یک مدل ترنسفورمری از پیش‌آموزش‌دیده بر پایه معماری BERT است.

بنابراین پارامترهای کلاسیک FastText در آن وجود ندارد. به همین دلیل، در ستون AriaBERT تنها پارامترهای مربوط به مرحله fine-tuning ازجمله learning rate، max sequence length و batch size گزارش شده‌اند. این تفکیک کمک می‌کند تفاوت ماهوی میان مدل‌های embedding ایستا (FastText) و embedding زمینه‌مند (AriaBERT) به‌طور شفاف نشان داده شود. جهت آزمایش و تست روش پیشنهادی به‌صورت تصادفی چند نمونه نام پیشنهادی متقاضی و زمینه فعالیت شرکت به مدل ساخته‌شده ارسال گردید و میزان ارتباط معنایی این دو پارامتر با استفاده از معیار شباهت‌کسینوس، روج‌ها موردسنجش و اندازه‌گیری قرارگرفته شده است. در جدول ۵ - نام و زمینه فعالیت چند شرکت جهت آزمایش آورده شده است. میزان درصد ارتباط معنایی نشان می‌دهد که نام پیشنهادی با زمینه فعالیت شرکت چند درصد ازلحاظ معنایی باهم ارتباط دارند.

جدول ۵. نام و زمینه فعالیت شرکت درخواستی ارسالی به مدل

نمونه	نام پیشنهادی	خلاصه‌سازی شده زمینه فعالیت
۱	گردشگری تفریحی ستاره مارال	گردشگری و تفریحی و فرهنگی و زیارتی
۲	یاران توسعه نرم‌افزار	خدمات تأسیسات و لوله‌کشی آب و گاز
۳	یاران توسعه نرم‌افزار	فناوری اطلاعات و پایگاه داده و ذخیره داده

بسته به ساختار کلی مدل، الگوریتم‌های حافظه کوتاه و بلند دوطرفه و الگوریتم توجه بر اساس پارامترهای جدول ۶- تنظیم شده‌اند.

لایه توجه معمولاً با یک تابع سافت مکس[1] برای عادی‌سازی وزن‌ها با اطمینان از مجموع آن‌ها برابر با یک و همچنین تابع تبدیل مانند تابعتانژانت هایپربولیک[2] برای تنظیم وزن ورودی‌ها جفت می‌شود .افزودن لایه توجه به مدل، تمرکز را بر روی مؤلفه‌های مهم متن خروجی افزایش می‌دهد. جدول ۶-فراپارامترهای لایه توجه را نشان می‌دهد.

جدول ٦. پارامترهای الگوریتم حافظه کوتاه و بلند دوطرفه و لایه توجه

الگوریتم حافظه کوتاه و بلند دوطرفه			لایه توجه	
Epoch Number	۲۰۰		Number of Heads	۴
Batch Size	۵۰		Weight Dimensions	۶۴
Learning Rate	۰,۰۰۰۱		Dropout	۰,۲
Activation Function	ReLU		Masking Strategy	padding

لایه توجه معمولاً نمایش توالی خروجی از حافظه کوتاه و بلند دوطرفه را افزایش می‌دهد اما بر پارامترهای داخلی واحد حافظه کوتاه و بلند دوطرفه تأثیر نمی‌گذارد.

در جدول ۷- نتایج ارزیابی زمینه فعالیت‌ها و نام شرکت بر اساس روج، روج و روج آل محاسبه شده است.

1 softmax
2 tanh

جدول ۷. مقایسه نتایج ارزیابی زمینه فعالیت شرکت و نام شرکت در الگوریتم پیشنهادی

ارزیابی	Rouge-1			Rouge-2			Rouge-L		
	دقت	فراخوانی	معیار اف	دقت	فراخوانی	معیار اف	دقت	فراخوانی	معیار اف
بخش اول (فعالیت‌های شرکت)	۰,۷۷	۰,۸۷	۰,۶۷	۰,۷۳	۰,۸۱	۰,۶۱	۰,۸۲	۰,۷۴	۰,۷۷
بخش دوم (نام شرکت)	۰,۷۵	۰,۹۱	۰,۷۳	۰,۷۷	۰,۸۵	۰,۶۲	۰,۷۴	۰,۸۲	۰,۷۹

این ارزیابی بر روی داده‌های زمینه فعالیت شرکت‌ها بعد از خلاصه‌سازی و بردار سازی نهایی انجام شده است.

بر اساس جدول ۸ –بعد از محاسبه شباهت کسینوسی میزان ارتباط نام با زمینه فعالیت شرکت ۹۲ درصد توسط مدل محاسبه شد که این مقدار نشان‌دهنده این می‌باشد که این دو پارامتر ارتباط معنایی نزدیکی با همدیگر دارند.

درواقع در این حالت مدل پیشنهادی یاد گرفت که بین کلمات سیاحتی، توریستی، گردشگری و تفریحی ارتباط معنایی وجود دارد که نشان‌دهنده کارکرد درست روش پیشنهادی می‌باشد.

در نمونه ۲ «یاران توسعه نرم‌افزار» به‌عنوان نام پیشنهادی و «خدمات تأسیسات» به‌عنوان زمینه فعالیت شرکت به مدل ارسال شد و میزان ارتباط این دو پارامتر ۲۸ درصد توسط مدل محاسبه گردید.

این مقدار نشان‌دهنده این می‌باشد که این دو پارامتر فاقد ارتباط معنایی با همدیگر دارند چراکه نام داده‌پرداز نوین ارتباط معنایی با حوزه فناوری اطلاعات دارد و با زمینه فعالیت انتخابی خدمات تأسیسات، ارتباطی ندارد.

در ادامه نام «یاران توسعه نرم‌افزار» یک‌بار دیگر به مدل به همراه زمینه فعالیت «فناوری اطلاعات» ارسال گردید که درصد ارتباط معنایی بین این دو پارامتر ۸۰ درصد

تعیین گردید که نشان‌دهنده مرتبط بودن این دو پارامتر باهم می‌باشد.

جدول ۸. ارزیابی درصد شباهت نام و زمینه فعالیت شرکت بر اساس شباهت کسینوسی

نمونه	نام پیشنهادی	خلاصه‌سازی شده زمینه فعالیت	درصد شباهت کسینوسی	نظر کارشناس	روش پیرسون
۱	گردشگری تفریحی ستاره مارال	گردشگری و تفریحی و فرهنگی و زیارتی	۹۲ درصد اعتبار معنایی	۱	۰,۹۷۱
۲	یاران توسعه نرم‌افزار	خدمات تأسیسات و لوله‌کشی آب و گاز	۲۸ درصد فاقد اعتبار معنایی	۰	۰,۴۲۲
۳	یاران توسعه نرم‌افزار	فناوری اطلاعات و پایگاه داده و ذخیره داده	۸۰ درصد اعتبار معنایی	۱	۰,۹۴۲

فرآیند خوشه‌بندی نام و بررسی خوشه‌ها

در بخش قبلی، همان‌طور که بررسی گردید نام‌ها و زمینه‌های فعالیت ثبت‌شده در بانک اطلاعاتی به بردارهای عددی تبدیل گردیدند و برای هر یک از آن‌ها یک بردار عددی به طول ۱۰ و بدون برچسب ایجاد شد. با توجه به تنوع الگوریتم‌های موجود برای خوشه‌بندی داده‌ها، این تحقیق از روش خوشه‌بندی دیبی اسکن به‌عنوان یک رویکرد مبتنی بر چگالی است. به‌منظور بهینه‌سازی فرآیند خوشه‌بندی، پارامترهای کلیدی این الگوریتم، شامل شعاع و حداقل تعداد نقاط برای تشکیل هر خوشه، به‌دقت تنظیم گردید. در این راستا، مقدار شعاع برابر با ۰,۵ و حداقل تعداد نقاط لازم برای ایجاد هر خوشه به‌عنوان ۵ در نظر گرفته شد. این تنظیمات به‌منظور دستیابی به بهترین نتایج ممکن در خوشه‌بندی داده‌های موردمطالعه به کار گرفته و به ارزیابی مؤثری از ساختار داده‌ها منجر گردید. در این مرحله مطابق با جدول ۹- این خوشه‌بندی آورده شده است.

جدول ۹. ارزیابی الگوریتم‌های خوشه‌بندی با استفاده از معیار ارزیابی دیویس –بولدین

نام شرکت	زمینه فعالیت	نام شرکت‌های مشابه در خوشه‌های ارزیابی‌شده
گردشگری تفریحی ستاره مارال	گردشگری و تفریحی	گردشگری و تفریحی آترین سیاحت پویا مجتمع تفریحی گردشگری و خدماتی سپاس ایرانیان خدمات گردشگری رفاهی اسکان
یاران توسعه نرم‌افزار	تأسیسات	زرین هوشمند داده‌پرداز داده‌پرداز رایان ابتکار پارس هوشمند داده‌پرداز

برای ارزیابی مؤثر نتایج تحقیق، تعریف معیارها برای مقایسه نتایج در مطالعات مختلف، ازجمله تحقیقات قبلی، بسیار مهم است .استاندارد کردن گزارش خروجی از این پلتفرم ضروری است و سایر محققان را قادر می‌سازد تا یافته‌ها را مورداستفاده و ارزیابی قرار دهند. ایجاد تعادل در تعداد معیارهای ارزیابی بسیار مهم است. این درحالی‌که معیارهای بیشتر می‌تواند دقت نتیجه را افزایش دهد و پیچیدگی بیش‌ازحد ممکن است تصمیم‌گیری برای عملکرد بهینه را به چالش بکشد. در مدل توسعه‌یافته، ادغام شبکه توجه و حافظه کوتاه و بلندمدت دوطرفه نه‌تنها محدودیت‌های آریا برت در تعداد نشانه‌های ورودی را برطرف نموده‌ایم، بلکه نتایج مطلوب‌تری را در معیاراعتبار سنجی به دست آورده‌ایم. به‌منظور اعتبارسنجی عملکرد مدل، یک ارزیابی انسانی مطابق استانداردهای SemEval انجام شد. از میان کل داده‌ها، ۲۰۰ نمونه به‌طور تصادفی انتخاب گردید و

توسط دو کارشناس رسمی اداره ثبت شرکت‌ها موردبررسی قرار گرفت. کارشناسان برای هر نام پیشنهادی دو نوع برچسب تخصیص دادند:

۱) برچسب دودویی ارتباط (مرتبط/نامرتبط)،

۲) امتیاز شدت ارتباط از ۰ تا ۱۰۰.

جهت سنجش پایایی بین ارزیابان، ضریب توافق Cohen's Kappa محاسبه شد که مقدار ۰٫۸۷ به دست آمد و نشان‌دهنده توافق قوی میان کارشناسان است. خروجی مدل با میانگین برچسب‌های انسانی مقایسه شد و ضریب همبستگی پیرسون ۰٫۹۳ به دست آمد. این نتایج

نشان می‌دهد که مدل توانسته است منطق تصمیم‌گیری انسانی را با دقت بالایی بازنمایی کند. همچنین نتایج اعتبار سنجی در جدول ۱۰- ارائه‌شده است که تجزیه‌وتحلیل مقایسه‌ای نمرات را با سایر مطالعات تحقیقاتی نشان می‌دهد. بر اساس بررسی‌های انجام شده و مقایسه با سایر مطالعات تحقیقاتی، نتایج حاکی از بهبود قابل‌توجهی در عملکرد الگوریتم توسعه‌یافته است. در مطالعه‌ای که حسینی مقدم و همکارانش در سال ۲۰۲۱ انجام دادند، آن‌ها الگوریتم شبکه عصبی را گسترش دادند و بهترین امتیاز ۷۵ را در معیار ارزیابی‌کننده اندازه اف به دست آوردند .قابل‌ذکر است، تحقیق فعلی بهبود عملکرد قابل‌توجهی ۱۷ درصدی را نسبت به این معیار نشان می‌دهد .در تحقیقی که حاجی غلامرضا و همکارانش در مورد تطبیق فضای بردار انجام دادند.

جدول ۱۰. مقایسه بهترین نتایج الگوریتم با الگوریتم‌های توسعه‌یافته مشابه

تحقیقات	موضوع	مدل	نتایج
حسینی مقدم در سال (۲۰۲۱ Hosseini et al. 2021)	شباهت معنایی متون کوتاه فارسی	شبکه عصبی	مقدار اف را برای ۴ تگ ۷۵ درصد محاسبه کردند
مینا حاجی غلامرضا در سال (۱۴۰۱ Sadidpour et al. 2022)	نقش فضای برداری در شناخت شباهت معنایی جملات فارسی	الگوریتم یادگیری عمیق	دقت ۸۸ درصد محاسبه شده است
روش پیشنهادی برای ثبت شرکت‌ها	شباهت نام و موضوع فعالیت شرکت‌ها	یادگیری عمیق حافظه کوتاه بلندمدت دوطرفه-آریا برت	دقت پیش‌بینی ۹۳ درصد

بهترین امتیاز ۸۸ درصد را در معیار دقت به دست آوردند. قابل‌ذکر است، تحقیق فعلی بهبود عملکرد قابل‌توجهی ۴ درصدی را نسبت به این معیار نشان می‌دهد .در مطالعه حاضر، بهبود عملکرد قابل‌توجهی را مشاهده می‌نماییم، همان‌طور که نتایج به‌دست‌آمده نشان می‌دهد علاوه بر این بهبود به خوشه‌بندی نام‌های شرکت‌ها و دسته‌بندی آن‌ها و همچنین پیشنهاد ارائه نام شرکت‌های مشابه هم در این تحقیق دست‌یافته‌ایم.

بحث و نتیجه‌گیری

نام شرکت معرف شخصیت و هویت آن است. نام شرکت گاه به زمینه فعالیت شرکت بستگی دارد،گاه صرفاً اسم خاص است. در این مقاله با طراحی سیستم تشخیص و ارزیابی میزان ارتباط معنایی نام شرکت با زمینه فعالیت شرکت بر اساس بردارسازی و استفاده از لایه‌های توجه خاص بر روی مدل‌های یادگیری عمیق استفاده شده است. بر همین اساس ابتدا بخش زمینه فعالیت‌های شرکت را بر اساس متن‌کاوی، الگوریتم‌های یادگیری عمیق و یک لایه توجه اضافی و خلاصه‌سازی متن به بردارهایی تبدیل کرده‌ایم. در بخش دوم نام شرکت را به بردار تبدیل نمودیم.

تلفیق روش‌های برت فارسی (آریا برت) و بردارسازی‌های عددی و یک لایه توجه اضافی به جهت خلاصه‌سازی‌های فعالیت‌های شرکت یک ماتریس به طول کلمات نام شرکت و یک ماترس زمینه فعالیت‌های شرکت ایجاد نمودیم. سپس با اعمال ارزیابی نتایج با تکنیک روج (۱ و ۲ و آل) پرداختیم.

جهت آزمایش روش پیشنهادی چند مقدار برای نام پیشنهادی و زمینه فعالیت به‌صورت تصادفی به مدل ساخته‌شده ارسال و نتیجه در ستون درصد ارتباط معنایی‌آورده شده است. درصد ارتباط معنایی با مقدار شباهت کسینوسی محاسبه گردید و نشان‌دهنده این است که نام پیشنهادی با زمینه فعالیت شرکت چند درصد ازلحاظ معنایی باهم ارتباط دارند. نتایج به‌دست‌آمده نشان‌دهنده این است که زمانی که نام پیشنهادی با زمینه فعالیت ارتباط معنایی نزدیکی با یکدیگر داشته باشند مدل به‌درستی درصد بالایی را بر می‌گرداند و زمانی که فاقد ارتباط معنایی باشند درصد پایینی را بر می‌گرداند. همچنین ضریب همبستگی بین میزان ارتباط معنایی به‌دست‌آمده توسط مدل و نظر واقعی کارشناس ثبت شرکت محاسبه گردید و نتایج نشان‌دهنده ۹۳ درصد ضریب همبستگی می‌باشد. در مقایسه‌ای کلی بر اساس جدول ۷، معیار ROUGE-L با دقت بالاتر نسبت به ROUGE-1 و ROUGE-2 نشان می‌دهد که مدل در حفظ ساختار معنایی فعالیت‌های شرکت عملکرد قوی‌تری دارد. جدول ۸ نیز نشان می‌دهد که شباهت کسینوسی میان نام و حوزه فعالیت در

نمونه‌های مرتبط بسیار بالا (۹۲٪ و ۸۰٪) و در نمونه نامرتبط پایین (۲۸٪) بوده و همبستگی ۹۳٪ با نظر کارشناس صحت مدل را تأیید می‌کند. در جدول ۹ نیز نتایج خوشه‌بندی DBSCAN بیانگر آن است که مدل توانسته نام‌ها را بر اساس نزدیکی معنایی در خوشه‌های درست قرار دهد و نام‌های مشابه را در حوزه فعالیت مشترک گروه‌بندی کند. محدودیت‌های پژوهش بر اساس پایگاه داده استفاده شده به زبان فارسی برای سازمان ثبت شرکت‌ها می‌باشد. تنها به میزان مشابهت زمینه فعالیت شرکت و نام شرکت پرداخته شده است.

پیشنهاد می‌شود در کارهای آتی از الگوریتم‌های یادگیری برخط و ترانسفورمرها استفاده شود. همچنین به بررسی معنایی نام شرکت و معنایی زمینه فعالیت‌های شرکت با سایر الگوریتم‌های ترکیبی یادگیری عمیق بر اساس تغییرات سریع مدل‌های زبانی بزرگ نیز پرداخته شود. ترکیب هم‌زمان وظایفی مانند «بردارسازی نام»، «تشخیص حوزه فعالیت» و «تشابه معنایی» در یک معماری واحد می‌تواند دقت کلی را افزایش دهد. در پژوهش‌های جدید نشان داده شده که یادگیری مبتنی بر چند فعالیت هم‌زمان در پردازش زبان طبیعی برای زبان‌های کم منبع عملکرد بهتری دارد.

تعارض منافع

تعارض منافع نداریم.

ORCID
Mohammad Rabiei **http://orcid.org/0000-0001-8728-8825**

References

1. Abdous, M., Piroozfar, P., & Minaei Bidgoli, B. (2024). PESTS: Persian–English cross-lingual corpus for semantic textual similarity. *Language Resources & Evaluation*. https://doi.org/10.1007/s10579-024-09759-3
2. Baigi, S. F. M., Sarbaz, M., Sobhani-Rad, D., & Kimiafar, K. (2023). A comparative study of rehabilitation information systems in 8 countries: A literature review. *Iranian Rehabilitation Journal*, 21(1), 1–16. https://doi.org/10.32598/irj.21.1.1766.1
3. Barbella, M., & Tortora, G. (2022). Rouge metric evaluation for text summarization techniques. *SSRN*. https://doi.org/10.2139/ssrn.4120317
4. Dogan, M. E., Dogan, T. G., & Bozkurt, A. (2023). The use of artificial intelligence (AI) in online learning and distance education processes: A systematic review of empirical studies. *Applied Sciences*, 13(5). https://doi.org/10.3390/app13053056
5. Ghafouri, A., Abbasi, M. A., & Naderi, H. (2023). AriaBERT: A pre-trained Persian BERT model for natural language understanding. *arXiv*. https://arxiv.org/abs/2304.04632
6. Hosseini, Z. S. M. E., Izadi, M., Tavakoli, M., et al. (2021). Designing a deep neural network model for finding semantic similarity between short Persian texts using a parallel corpus.
7. Khan, S., & Anjum, M. A. I. (2023). Words in mental lexicon: A comparative analysis of word association (WA) responses of Pakistani L1 and Afghan L2 speakers of Urdu. *Journal of Communication and Cultural Trends*, 5(1), 86–105. https://doi.org/10.32350/jcct.51.05
8. Masumi, M., Majd, S. S., Shamsfard, M., & Beigy, H. (2024). FaBERT: Pre-training BERT on Persian blogs. *arXiv*. https://doi.org/10.48550/arXiv.2402.06617
9. Mehrban, A., & Ahadian, P. (2023). Evaluating BERT and ParsBERT for analyzing Persian advertisement data. *arXiv*. https://arxiv.org/abs/2305.02426
10. Moniri, S., Schlosser, T., & Kowerko, D. (2024). Investigating the challenges and opportunities in Persian language information retrieval through standardized data collections and deep learning. *Computers*. https://doi.org/10.3390/computers13020045
11. Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies–Bouldin index for clustering evaluation. *Neurocomputing*, 528, 125–139. https://doi.org/10.1016/j.neucom.2023.01.043

12. Sadidpour, S. S., Haji Gholamreza, M., Mohammadzadeh, M. R., Mohammadi, M. R., & Keivanrad, M. A. (2022). Investigation of the semantic similarity of Persian sentences using vector space adaptation and deep learning.
13. Sadjadi, S. M., Rajabi, Z., Rabiei, L., & Moin, M.-S. (2024). FarSSiBERT: A novel transformer-based model for semantic similarity measurement of Persian social networks informal texts. *arXiv*. https://arxiv.org/abs/2407.19173
14. Zareshahi, A., Javadzade, M. A., & Bastami, E. (2024). Measuring semantic similarity of Persian sentences using ParsBERT model. In 2024 10th International Conference on Artificial Intelligence and Robotics (*QICAR*) (pp. 316–321). https://doi.org/10.1109/QICAR61538.2024.10496609

استناد به این مقاله: ربیعی، محمد. (۱۴۰۵). طراحی سیستم سنجش تطابق نام و حوزه فعالیت شرکت‌ها بر اساس هوش مصنوعی، *مطالعات مدیریت کسب وکار هوشمند*، ۱۵(۵۵)، ۳۰۷-۳۳۶. DOI: 10.22054/ims.2026.83957.2573