

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 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.

* Corresponding Author: Mohammad.Rabiei@eyc.ac.ir

How to Cite: Designing a system for matching the name and field of activity of companies based on artificial intelligence, *Journal of Business Intelligence Management Studies*, 14(54), 1-44. DOI: 10.22054/ims.2025.85270.2608

Accepted:

Received:

eISSN:

ISSN: 2821-0964

1. 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.

2. 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; Zarehshahi 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.

3. 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.

4. 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.

5. 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.

اماره انتشار



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

استادیار، گروه مهندسی برق و کامپیوتر، دانشکده فنی و مهندسی، دانشگاه ایوان کی، ایوان کی، سمنان،

محمد ربیعی * ID

ایران

چکیده

تایید نام در فرآیند ثبت تاسیس شرکت باعث می شود از ثبت نام شرکتها که با زمینه فعالیت آنها همخوانی ندارد جلوگیری بعمل آید. در این مقاله بمنظور 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. داده های این پژوهش از سازمان ثبت اسناد و املاک کشور جمع آوری گردیده است. در روش پیاده سازی ابتدا از فیلترهای اولیه نامگذاری شرکت استفاده شده است. سپس با استفاده از ترکیب روش آریا برت به عنوان یک تکنیک تعبیه کلمات به تبدیل نام پیشنهادی شرکت به بردار پرداخته می شود. در مرحله ای موازی زمینه فعالیت شرکت را با استفاده از فستکس به بردار عددی و تلفیق بردار بدست آمده با الگوریتمهای یادگیری عمیق حافظه کوتاه و بلند مدت دو طرفه بر اساس یک لایه توجه اضافه می گردد. جهت ارزیابی نتایج از معیار شباهت کسینوسی و معیار روج (۱ و ۲) استفاده شده است. پس از تایید پذیرش نام شرکت و زمینه فعالیت، از روش خوشه بندی دیبی اسکن برای خوشه بندی نام شرکت در دسته های فعالیت استفاده می شود.

نتایج تحقیق نشان می دهد که مقادیر دقت در بخش بردار سازی زمینه فعالیتهای شرکت برای معیار روج ال مقدار ۰٫۷۹۸۲ و مقادیر دقت و فراخوانی نهایی مدل به ترتیب ۰٫۸۵۱۲، ۰٫۸۳۱۷ محاسبه گردید. همچنین ضریب همبستگی بین شباهت کسینوسی محاسبه شده بین نام پیشنهادی و زمینه فعالیت شرکت با مقدار ۹۳ درصد بر اساس معیارهای تعیین نام نشان دهنده کارکرد درست مدل می باشد.

کلیدواژه ها: ارتباط معنایی، ثبت شرکت، شباهت کسینوسی، متن کاوی، یادگیری عمیق.

مقدمه

پردازش متون، به عنوان حوزه ای پیشرو در داده کاوی، به استخراج الگوها و دانش پنهان از منابع متنی می پردازد. از جمله وظایف اصلی پردازش متون می توان به طبقه بندی اسناد، خوشه بندی، استخراج مفاهیم کلیدی، تحلیل معنایی، تلخیص و شناسایی ارتباطات مفهومی میان عناصر متنی اشاره کرد. در این راستا، تعیین میزان تشابه متنی، به ویژه در سطوح معناشناختی، لغوی و ساختاری، نقشی اساسی در بهینه سازی کاربردهای فناوری ایفا می کند و به شدت مورد توجه قرار گرفته است. تحلیل شباهت های متنی، به ویژه در سطوح معناشناختی و واژگانی، امکان بررسی عمیق تر متون را فراهم می کند. اندازه گیری شباهت بین متون مختلف، می تواند به شناسایی مفاهیم هم معنی کمک نماید. این قابلیت در سامانه های پردازش زبان طبیعی مانند طبقه بندی متون، تلخیص خودکار و سیستم های پرسش-پاسخ اهمیت بسزایی دارد. تحقیقات نوین در حوزه پردازش زبان طبیعی، روش های متنوعی برای سنجش تشابه معناشناختی میان کلمات و جملات ارائه کرده اند که شامل روش های مبتنی بر پایگاه های واژگانی، مدل های بردار کلمه مبتنی بر شبکه های عصبی بازگشتی و شبکه های عصبی عمیق است. امروزه، الگوریتم های یادگیری ماشین و یادگیری ژرف به عنوان ابزارهای حیاتی در تجزیه و تحلیل متون، به طور گسترده ای در کاربردهای مختلف تحلیل محتوای متنی به کار گرفته می شوند.

انتخاب نام مناسب برای شرکت، مسئله ای حساس و مرحله ای تعیین کننده 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 می کند. کارشناسان اداره، پس از بررسیهای اولیه لغوی با بررسی نام های ثبت شده قبلی، مشابهت نام پیشنهادی را ارزیابی کرده و در صورت نداشتن مشابه، آن را تأیید می کنند.

آمارها نشان می دهد که از میان ۳۴۸،۹۲۷ درخواست نام گذاری شرکت در ده ماه نخست سال ۱۴۰۰، حدود ۲۲۷،۹۲۶ نام رد شده و تنها ۲۰ درصد از نام ها تأیید شده اند (Baigi et al. 2023). مدت زمان میانگین برای تعیین نام تقریباً ۴ روز است که بخش عمده آن به دلیل رد نام ها و پیشنهاد نام های جدید از سوی متقاضیان به طول می انجامد.

تحلیل شباهت متنی درباره سامانه ثبت شرکتها می‌تواند چالش‌های متعددی را در حوزه‌های کسب و کار مرتفع کند. از جمله کاربردهای کلیدی این حوزه، ارزیابی نام‌های پیشنهادی برای ثبت شرکت‌ها و مؤسسات غیرتجاری در سازمان ثبت اسناد و املاک است. در این مقاله، هدف اصلی روشی برای شناسایی میزان شباهت معناشناختی میان نام‌های پیشنهادی و زمینه فعالیت‌های اعلان شده برای شرکت‌ها و مؤسسات غیرتجاری معرفی می‌گردد. پس از بررسی‌های انجام شده 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 تایید صحیح تر نام‌های پیشنهادی متقاضیان ثبت شرکت و خوشه بندی 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 این که 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 ثبت نام‌های 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 انجام شده پرداخته است. در بخش ۳ مدل ارائه شده و روش پیشنهادی بیان شده است. بخش ۴ نتایج و ارزیابی روش پیشنهادی و بخش ۵ در برگیرنده نتیجه‌گیری و پیشنهاد‌های مطالعاتی می‌باشد.

پیشینه پژوهش

فعالیت شرکت بر انتخاب نام شرکت تاثیر ویژه ای 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 نه تنها

به یادسپاری بهتر کمک می کند بلکه نقش مؤثری در بازاریابی و جذب مخاطب ایفا می نماید. نام ثبت شده یک شرکت، به عنوان هویت حقوقی آن، در مرجع ثبت شرکت ها به تأیید رسیده و یا در فرآیند تأسیس می باشد. در مقابل، نام پیشنهادی، نامی است که متقاضی در زمان تأسیس یا درخواست تغییر نام به مرجع ثبت شرکت ها ارائه می نماید. همچنین، تعریف زمینه فعالیت شرکت به منظور تعیین مرزها و حدود فعالیت های شرکت، نشان دهنده 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 نام های اشخاص حقوقی ثبت شده، اجتناب از کاربرد اصطلاحات دولتی و عدم استفاده از عناوین غیرقانونی و واژگان بیگانه و نامهای حکومتی، از الزامات مهم به شمار می آید. درباره برخی از موارد لازم الاجرا بغیر از موارد فوق در جدول ۱- توضیحاتی داده شده است. رابط زیر که در فرهنگستان با نام انواع جناس شناخته می شود باید رعایت کامل گردند.

جدول ۱. نکات اولیه مورد بررسی و ارزیابی جهت تعیین نام تجاری

اسم شرکت	اسمی که هم در معنی و هم تلفظ با اسامی قبلی برابری می کند	فناوران صنعت روز-فناوران صنعت روز
اسم یکسان نوشتاری	معانی متفاوت برای دو اسم کاملاً یکسان	مَلک، مُلک، مَلک
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 هر کدام مزایا و محدودیت‌های مختص خود را دارا هستند. مزیت اصلی تحلیل شباهت لغوی سرعت بالا است، اما در متون بازنویسی‌شده و تغییر یافته، 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 منهن 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. حسینی مقدم و همکاران (Hosseini et al. 2021) رویکردی برای سنجش شباهت معنایی در متون کوتاه فارسی که از برنامه های تلویزیونی بصورت جفتی شامل ۳۵۲۶۶ جفت جمله ساده ارائه داده اند. روش پیشنهادی شامل سه مرحله است. اولین گام جمع آوری داده ها و ساختن یک پیکره موازی است. در مرحله بعدی یعنی مرحله پیش پردازش، داده ها نرمال می شوند. نتایج این تحقیق نشان می دهد که روش پیشنهادی با معیار اندازگیری اف، به دقت ۷۸,۳۷ درصد برای داده ها ۲ واژه پشت سر هم در یک جمله و ۶۵,۹۸ درصد برای داده های چهار واژه پشت سر هم دست یافته است.

صادقی پور و همکاران (Sadidpour et al. 2022) با بهره گیری از شبکه های حافظه کوتاه بلند مدت دوطرفه^۱، روشی را برای ارزیابی میزان شباهت معنایی جملات فارسی ارائه داده اند. در این روش، نگاشت برداری جملات به فضای برداری با استفاده از شبکه های عصبی انجام شده و نتایج آن بر اساس امتیازات انسانی سنجیده شده است؛ دقت این سیستم پیشنهادی حدود ۸۹,۲ درصد ارزیابی شده است.

تحقیقات غفوری و همکاران (Ghafouri, Abbasi, and Naderi 2023) بر آریابرت^۲ انجام پذیرفت. آنها آریابرت را به عنوان یک مدل زبان از پیش آموزش دیده برای فارسی معرفی کردند. همچنین به مطالعه در زمینه کمبود داده های متنی متنوع و مدل های از پیش آموزش دیده کارآمد در زبان فارسی پرداختند. آریابرت بر روی مجموعه داده های متنوعی از متون فارسی، از جمله متون محاوره ای، رسمی و ترکیبی مانند توییت ها، اخبار، شعرها، متون پزشکی و موارد دیگر آموزش داده شد که مجموعاً بیش از ۳۲ گیگابایت است. برخلاف دیگر مدل هایی که از برت استفاده می کنند، آریابرت از معماری روبرتا^۲ استفاده می کند. در مقایسه با مدل های موجود زبان فارسی در بین وظایف مختلف پردازش زبان طبیعی برتری قابل توجه آریابرت را با میانگین بهبود ۳٪ در تحلیل احساسات ۰,۶۵٪، در طبقه بندی و ۲,۳٪ در تشخیص موضع نشان داده شده است.

1. AriaBert
2. 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 مسابقه ای را برای پیش بینی درصد متن انتشار آگهی فروش خودرو در فروش واقعی آن ترتیب دادند. با استفاده از کتابخانه هزم^۳ و مدل های زبانی پیشرفته ام برت^۴ و پارس برت^۵ به تجزیه و تحلیل داده‌ها، تنظیم دقیق مدل و تنظیمات آموزشی پرداختند. نقاط قوت و ضعف مدل را برجسته داده کاوی و تکنیک های یادگیری ماشین را مورد بحث قرار داده اند.

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 قابل توجهی توانایی پوشش واژگانی و درک معنایی را دارند. توسعه و پاک‌سازی متنوع و انسجام واژگان می‌تواند به‌طور مستقیم کیفیت مدل‌های بازاریابی 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 شبکه های اجتماعی کار کردند. آنها در مطالعات خود از مدل زبان پارسبرت به عنوان مدل پایه استفاده کرده و برای آنالیز سریعتر، جملات نشانه گذاری شده را به بردارهای معنی دار تبدیل نمودند. از طریق لایه های ادغام، بردارهای بهینه استخراج و در نتیجه بردارهای ۷۶۸ بعدی برای هر جمله ایجاد کرده اند. نتایج نشان می‌دهد که مدل پیشنهادی به ضریب همبستگی پیرسون تقریبی ۰٫۶۸ دست یافت که از مدل‌های قبلی فستکس^۶ و یک مدل شبکه عصبی پیچشی^۷ پیشی گرفته است و نشان‌دهنده عملکرد برتر آن نسبت به مدل‌های قبلی است.

1. Bidirectional encoder representations from transformers (BERT)
2. RoBERTa
3. Hazm
4. mBERT
5. ParsBERT
6. FastText
7. Convolutional Neural Network (CNN)

در تحقیقات شباهتهای معنایی محاوره ای سجادی (Sadjadi et al. 2024) در زمینه جملات غیر رسمی و عامیانه در شبکه های اجتماعی برای بررسی شباهت جملات از مدل ترانسفورمر^۱ استفاده کرده است.

این گروه پایگاه داده جدیدی به نام فارس سیم^۲ با تعداد ۱۰۴ میلیون جمله کوتاه محاوره ای را برای تحقیقات خود تهیه نموده اند. مدل ارایه شده بر اساس ترانسفورمر با تعداد لایه های توجه بالا بر اساس معیارهای ضریب پیرسون ۰,۷۷ و اسپیرمن ۰,۶۴ بهتر از برت و پارس برت چند زبانه عمل می کند. همچنین نقطه برتری دیگر نتایج را در مدل زبان بزرگ از پیش آموزش دیده شده برای استفاده در سایر کارهای پردازش زبان طبیعی در متن محاوره ای و به عنوان نشانه ای برای کلمات غیررسمی کمتر شناخته شده معرفی نموده اند.

همچنین عبدوس و همکاران برای نخستین بار یک ارتباط دوسویه و دوزبانه با نام PESTS شامل ۵۳۷۵ جفت جمله فارسی-انگلیسی با برچسب گذاری انسانی تهیه کردند. این پایگاه داده امکان آموزش و ارزیابی مستقیم مدل های شباهت معنایی میان زبانی را بدون اتکا به ترجمه ماشینی فراهم می کند و از انتشار خطاهای ترجمه جلوگیری می شود. 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 با هدف 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 گسترده ای از وبلاگ های فارسی با سبک های مختلف آموزش داده شده است. در ارزیابی های انجام شده روی ۱۲ دیتاست در وظایف مختلف مانند تحلیل احساس، تشخیص موجودیت نامدار، استنتاج طبیعی زبان، پرسش پاسخ و تشخیص پارافریز، نتایج نشان داد که FaBERT در اغلب وظایف، عملکرد بهتری نسبت به مدل های هم رده مانند (mBERT و ParsBERT) دارد.

1. Transformer
2. Farsim

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

در شکاف تحقیقات انجام شده، مطالعات کمی در مورد تشخیص خودکار تناقض در تطابق معنایی نام شرکت و زمینه فعالیت های شرکت در فرآیندهای ثبت وجود دارد. با این حال، روش های تشخیص تضاد در برنامه های مختلف پردازش زبان طبیعی در دنیا وجود دارد و هر روز در حال گسترش می باشد. پایگاه های داده بزرگ و استاندارد برای فارسی در حوزه شباهت معنایی محدود است و این امر دقت مدل ها را تحت تأثیر قرار می دهد. در پژوهش های پیشین نیز کمبود منابع متنی غنی فارسی به عنوان یک چالش اصلی مطرح شده است.

بخش عمده‌ی داده‌های بازیابی شده 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 متن کاوی بر اساس فیلتر سازی اولیه و تولید بردار ویژگی متن گلاو¹ و استفاده از الگوریتم حافظه کوتاه و بلند مدت دو طرفه و یک لایه توجه² 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

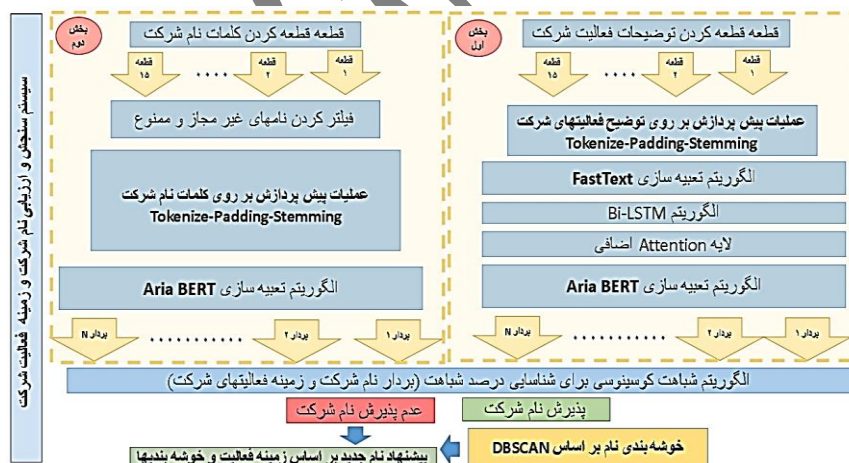
روش پیاده سازی

این پژوهش کاربردی و در دسته تحقیقات مطالعه موردی دسته بندی می گردد. برای پیاده سازی این مدل، از داده های ثبت شده شامل نام های تایید شده و رد شده در بانک اطلاعاتی ثبت شرکت ها استفاده شده است.

این مدل با ترکیب رویکردهای یادگیری ماشین و تکنیک های پردازش زبان طبیعی برای بردار سازی کلمات توسعه یافته و با بهره گیری از روش های خوشه بندی به شناسایی نزدیک ترین نام ها و محاسبه شباهت معنایی می پردازد.

ساختار مدل پیشنهادی شامل دو بخش تحلیل نام پیشنهادی و ارزیابی زمینه فعالیت شرکت است. هدف از این مدل، تسریع و بهینه سازی فرآیند ثبت شرکت ها و موسسات غیر تجاری است. در شکل ۱- جزئیات روش پیاده سازی مدل بررسی شباهت معنایی نام شرکت و زمینه فعالیت شرکت در ثبت موسسات غیر تجاری نشان داده شده است. همچنین تعیین میزان شباهت معنایی نام پیشنهادی با زمینه فعالیت شرکت در نهایت توسط الگوریتم شباهت کوسینوسی محاسبه شده است.

شکل ۱. روش پیاده سازی مدل بررسی شباهت معنایی نام شرکت ها و موسسات غیر تجاری



در مرحله نخست، به منظور تبدیل فعالیت های شرکت به بردارهای عددی و استخراج ارتباط معنایی میان واژگان، توضیحات متنی فعالیت ها پس از انجام عملیات پیش پردازش شامل Tokenizing، Padding، و Stemming آماده سازی شدند. سپس این داده ها با استفاده از مدل FastText که برای زبان فارسی آموزش داده شده است، به بردارهای چندبعدی تبدیل گردید. به دلیل ماهیت جمله محور و وجود تنوع واژگانی در فعالیت های شرکت، بردارهای FastText برای استخراج معنای عمیق تر

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 به صورت یک بردار واحد در ماتریس ویژگی ذخیره شد. در بخش دوم، ابتدا کلمات تشکیل دهنده 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 به دست 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 تعیین شده 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 DBSCAN¹ برای دسته‌بندی نام‌های شرکت بر اساس حوزه فعالیت استفاده شد.

این کار موجب شد نام‌های مشابه در خوشه‌های معنایی مشترک قرار گیرند و امکان پیشنهاد نام‌های جایگزین مرتبط فراهم شود. برای ارزیابی کیفیت خوشه‌بندی، از شاخص Davies-Bouldin (Ros, Riad & Guillaume, 2023)² استفاده شد که میزان جدایی خوشه‌ها و نزدیکی درون خوشه‌ای را با دقت مناسبی نشان می‌دهد. FastText بردار هر واژه را بر اساس n-gram‌های 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. با این حال، FastText به تنهایی ترتیب واژه‌ها و ساختار جمله را مدل نمی‌کند. برای این هدف، بردارهای FastText به لایه 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 تاسیسات») وزن بیشتری بدهد.

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

2. Davies-Bouldin

- بدین ترتیب، نمایش نهایی جمله به صورت یک بردار وزن‌دهی شده از کل توالی به دست می‌آید. به طور 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 تاییدشده، ردشده و اطلاعات حوزه فعالیت شرکت‌ها از پایگاه داده ثبت شرکت‌ها و موسسات غیرتجاری استخراج شده است. 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 به صورت صرفاً تصادفی انجام نشد. در عوض، کل رکوردها بر اساس تاریخ ثبت مرتب و سپس به صورت 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 از جمله تاییدشده، 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 به منظور آماده‌سازی برای مراحل پردازش نگهداری شده‌اند.

پیش پردازش

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. در نهایت، پس از اعمال عملیات پیش‌پردازشی، تعداد ۱،۷۴۶،۹۷۷ نام به‌عنوان مجموعه داده نهایی انتخاب شدند.

جدول ۲. مرحله پیش پردازش یکسان سازی

اصلاح نوشتاری بر اساس حروف فارسی و عربی	حرف الف چندین شکل آ- ا- ا- ا- ا-
توجه به عدم باقیماندن ایست واژه ها در متن	
توجه به فاصله ها و نیم فاصله ها و حذف آنها	
توجه به حداقل ۴ سیلاب و حداکثر کمتر از ده سیلاب	نام های درخواستی معمولاً ۴ سیلابی می باشند
توجه به ساختار کامات و ریشه آنها	حستجو جهت پیدا کردن ریشه هر کلمه

تعبیه سازی کلمات

تعبیه کلمات به‌عنوان یک روش کلیدی در نگاشت متون به بردارهای عددی به شمار می‌آید. از آنجایی که سیستم‌ها و رایانه‌ها قادر به تحلیل و پردازش مستقیم متون و اسناد نیستند، تبدیل این متون به بردارهای عددی قابل فهم برای رایانه‌ها از اهمیت بالایی برخوردار است. در این راستا، تکنیک تعبیه کلمات به کار 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. به‌طور کلی، نزدیکی این بردارها به یکدیگر بیانگر شباهت بیشتر میان کلمات است.

خوشه بندی بردار نام ها

پس از تبدیل نام‌های ثبت شده به بردارهای عددی، به دلیل عدم وجود برجسب برای این بردارها، ضرورت دارد از الگوریتم‌های خوشه‌بندی به منظور دسته بندی این بردارها بهره

بردارى شود. در این تحقیق، از روش خوشه‌بندى دىبى اسکن برای خوشه‌بندى بردارهای مرتبط با یکدیگر استفاده شده است. این الگوریتم با شناسایی ساختارهای دانه‌ای در داده‌ها، امکان گروه‌بندى 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^۱، روج دو^۲ و روج ال^۳ استفاده شده است (Barbella and Tortora 2022). هرچه میزان شباهت افزایش یابد، می‌توان نتیجه‌گیری کرد که فاصله‌ی بین دو شی 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 خلاصه سازی آنها با استفاده از روج ان تعداد n -گرم تطبیق بین متن تولید شده توسط مدل و مرجع تولید شده توسط انسان را اندازه‌گیری می‌کنیم. در زمینه ارزیابی نتایج خلاصه سازی 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.

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

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

اما به دلیل 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 داده ای متشکل از ۱۷۶۱۱۰۰ نام از سامانه ثبت شرکت ها و موسسات غیرتجاری جمع آوری کردیم. مجموعه داده به دو دسته تقسیم شده اند. ۱۲۷۰۵۰۰ نام تایید شده و ۴۹۰۶۰۰ نام تایید نشده/غیر مجاز یارد شده می باشد. جدول ۳ توزیع داده ها را برای آموزش مدل، اعتبارسنجی و آزمایش نهایی نشان می دهد ۸۰٪ برای آموزش، ۱۰٪ برای اعتبار سنجی، و ۱۰٪ باقی مانده برای آزمایش استفاده می شود.

جدول ۳. پایگاه داده مورد استفاده 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 ساخته شده بردار نام پیشنهادی و بردار زمینه فعالیت شرکت به عناوین ورودی به مدل داده می شود. در جدول ۴- پارامترهای بدست آمده تحت روش آریا برت برای نام شرکت و فستکس برای زمینه فعالیتهای شرکت جهت آموزش بر روی مجموعه داده آورده شده است. لازم به ذکر است که برای آریا برت مقدار اندازه دسته ها را ۵۰ و الگوریتم بهینه ساز آدام^۲ و دراپ اوت^۳ مقدار ۰,۲ در نظر گرفته شده است.

1. Batch Size
2. Adam
3. Dropout

جدول ۴. پارامترها و مقادیر روشهای ساخت بردار کلمات

مقادیر تنظیمی		پارامترها
Aria BERT	Fast text	
۰,۰۰۰۵	۰,۰۶	نرخ یادگیری
ابعاد پیش فرض ۷۶۸	۱۰۰	ابعاد بردار
۱۰ یا مقدار واقعی پروژه شما؛ مقدار ۲۰۰ برای BERT صحیح نیست و باید منطقی باشد	۶۰	تعداد epoch
در معماری BERT وجود ندارد	۵	Window Size
—	۳	MinCount
—	Skip-Gram	مدل
۱۲۸ یا مقدار واقعی	—	Max Sequence Length
۳۲	—	Batch Size

همان گونه که در جدول ۴ - نشان داده شده است، پارامترهای مدل FastText شامل مقادیر پایه ای مانند window size، mincount و ساختار skip-gram هستند که مختص مدل های تعبیه ساز واژگان مبتنی بر زیر واحدهای واژگانی می باشند. در مقابل، مدل AriaBERT یک مدل ترنسفورمری از پیش آموزش دیده 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 از جمله learning rate، max sequence length و batch size گزارش شده اند. این تفکیک کمک می کند تفاوت ماهوی میان مدل های embedding ایستا (FastText) و embedding زمینه مند (AriaBERT) به طور شفاف نشان داده شود. جهت آزمایش و تست روش پیشنهادی بصورت تصادفی چند نمونه نام پیشنهادی متقاضی و زمینه فعالیت شرکت به مدل ساخته شده ارسال گردید و میزان ارتباط معنایی این دو پارامتر با استفاده از معیار شباهت کسینوس، روج ها مورد سنجش و اندازه گیری قرار گرفته شده است. در جدول ۵ - نام و زمینه فعالیت چند شرکت جهت آزمایش آورده شده است. میزان درصد ارتباط معنایی نشان می دهد که نام پیشنهادی با زمینه فعالیت شرکت چند درصد از لحاظ معنایی با هم ارتباط دارند.

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

نمونه	نام پیشنهادی	خلاصه سازی شده زمینه فعالیت
۱	گردشگری تفریحی ستاره مارال	گردشگری و تفریحی و فرهنگی و زیارتی
۲	یاران توسعه نرم افزار	خدمات تاسیسات و لوله کشی آب و گاز
۳	یاران توسعه نرم افزار	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 شده اند.

لایه توجه معمولاً با یک تابع سافت مکس^۱ برای عادی سازی وزن ها با اطمینان از مجموع آنها برابر با یک و همچنین تابع تبدیل مانند تابع تانژانت هایپربولیک^۲ برای تنظیم وزن ورودی ها جفت می شود. افزودن لایه توجه به مدل، تمرکز را بر روی مؤلفه های مهم متن خروجی افزایش می دهد. جدول ۶- پارامترهای لایه توجه را نشان می دهد.

جدول ۶. 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 می دهد اما بر پارامترهای داخلی واحد حافظه کوتاه و بلند دوطرفه تأثیر نمی گذارد. در جدول ۷- نتایج ارزیابی زمینه فعالیتها و نام شرکت بر اساس روج ۱، روج ۲ و روج ال 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 شرکت ها بعد از خلاصه سازی 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 نشان دهنده این می باشد که این دو پارامتر ارتباط معنایی نزدیکی با همدیگر دارند.

در واقع در این حالت مدل پیشنهادی یاد گرفت که بین کلمات سیاحتی، توریستی، گردشگری و تفریحی ارتباط معنایی وجود دارد که نشان دهنده کارکرد درست روش پیشنهادی می باشد.

در نمونه ۲ " یاران توسعه نرم افزار " به عنوان نام پیشنهادی و " خدمات تاسیسات " به عنوان زمینه فعالیت شرکت به مدل ارسال شد و میزان ارتباط این دو پارامتر ۲۸ درصد توسط مدل محاسبه گردید.

این مقدار نشان دهنده این می باشد که این دو پارامتر فاقد ارتباط معنایی با همدیگر دارند چرا که نام داده پرداز نوین ارتباط معنایی با حوزه فناوری اطلاعات دارد و با زمینه فعالیت انتخابی خدمات تاسیسات، ارتباطی ندارد.

در ادامه نام " یاران توسعه نرم افزار " یک بار دیگر به مدل به همراه زمینه فعالیت " فناوری اطلاعات " ارسال گردید که درصد ارتباط معنایی بین این دو پارامتر ۸۰ درصد تعیین گردید که نشان دهنده مرتبط بودن این دو پارامتر با هم می باشد.

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

نمونه	نام پیشنهادی	خلاصه سازی 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 نرم افزار	خدمات تاسیسات و لوله کشی آب و گاز	۲۸ درصد فاقد اعتبار معنایی	۰	۰,۴۲۲
۳	یاران توسعه نرم افزار	فناوری اطلاعات و پایگاه داده و ذخیره داده	۸۰ درصد اعتبار معنایی	۱	۰,۹۴۲

فرآیند خوشه بندی نام و بررسی خوشه ها

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

جدول ۹. ارزیابی الگوریتم های خوشه بندی با استفاده از معیار ارزیابی دیویس -بولدین

نام شرکت	زمینه فعالیت	نام شرکتهای مشابه در خوشه های ارزیابی شده
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 نرم افزار	تاسیسات	زرین هوشمند داده پرداز داده پرداز رایان ابتکار پارس هوشمند داده پرداز

برای ارزیابی مؤثر نتایج تحقیق، تعریف معیارها برای مقایسه نتایج در مطالعات مختلف، از جمله تحقیقات قبلی، بسیار مهم است. استاندارد کردن گزارش خروجی از این پلتفرم ضروری است و سایر محققان را قادر می سازد تا یافته ها را مورد استفاده و ارزیابی قرار دهند. ایجاد تعادل در تعداد معیارهای ارزیابی بسیار مهم است. این در حالی که معیارهای بیشتر می تواند دقت نتیجه را افزایش دهد و پیچیدگی بیش از حد ممکن است تصمیم گیری برای عملکرد بهینه را به چالش بکشد. در مدل توسعه یافته، ادغام شبکه توجه و حافظه کوتاه و بلند مدت دوطرفه نه تنها محدودیت های آریا برت در تعداد نشانه های ورودی را برطرف نموده ایم، بلکه نتایج مطلوب تری را در معیار اعتبار سنجی بدست آورده ایم. به منظور اعتبارسنجی عملکرد مدل، یک ارزیابی انسانی مطابق استانداردهای SemEval انجام شد. از میان کل داده ها، ۲۰۰ نمونه به طور تصادفی انتخاب گردید و

توسط دو کارشناس رسمی اداره ثبت شرکت ها مورد بررسی 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 نشان دهنده توافق قوی میان کارشناسان است. خروجی مدل با میانگین برچسب های انسانی مقایسه شد و ضریب همبستگی پیرسون ۰,۹۳ به دست آمد. این نتایج نشان می دهد که مدل توانسته است منطق تصمیم گیری انسانی را با دقت بالایی بازنمایی کند. همچنین نتایج اعتبار سنجی در جدول ۱۰- ارائه شده است که تجزیه و تحلیل مقایسه ای نمرات را با سایر مطالعات تحقیقاتی نشان می دهد. بر اساس بررسی های انجام شده و مقایسه با سایر مطالعات تحقیقاتی، نتایج حاکی از بهبود قابل توجهی در عملکرد الگوریتم توسعه یافته است. در مطالعه ای که حسینی مقدم و همکارانش در سال ۲۰۲۱ انجام دادند، آنها الگوریتم شبکه عصبی را گسترش دادند و بهترین امتیاز ۷۵ را در معیار ارزیابی کننده اندازه اف به دست آوردند. قابل ذکر است، تحقیق فعلی بهبود عملکرد قابل توجهی ۱۷ درصدی را نسبت به این معیار نشان می دهد. در تحقیقی که حاجی غلامرضا و همکارانش در مورد تطبیق فضای بردار انجام دادند.

جدول ۱۰. مقایسه بهترین نتایج الگوریتم با الگوریتم های توسعه یافته 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 بلند مدت دوطرفه-آریا پرت	دقت پیش بینی ۹۳ درصد

بهترین امتیاز ۸۸ درصد را در معیار دقت به دست آوردند. قابل ذکر است، تحقیق فعلی بهبود عملکرد قابل توجهی ۴ درصدی را نسبت به این معیار نشان می دهد. در مطالعه حاضر، بهبود عملکرد قابل توجهی را مشاهده می نمایم، همانطور که نتایج به دست آمده نشان می دهد علاوه بر این بهبود به خوشه بندی نامهای شرکتها و بسته بندی آنها و همچنین پیشنهاد ارایه نام شرکتها مشابه هم در این تحقیق دست یافته ایم.

بحث و نتیجه گیری

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 بردار سازی و استفاده از لایه های توجه خاص بر روی مدلهای یادگیری عمیق استفاده شده است. بر همین اساس ابتدا بخش زمینه فعالیتهای شرکت را بر اساس متن کاوی، الگوریتمهای یادگیری عمیق و یک لایه توجه اضافی و خلاصه سازی متن به بردارهایی تبدیل کرده ایم. در بخش دوم نام شرکت را به بردار تبدیل نمودیم.

تلفیق روشهای پرت فارسی (آریا پرت) و بردار سازیهای عددی و یک لایه توجه اضافی به جهت خلاصه سازیهای فعالیتهای شرکت یک ماتریس به طول کلمات نام شرکت و یک ماتریس زمینه فعالیتهای شرکت ایجاد نمودیم. سپس با اعمال ارزیابی نتایج با تکنیک روج (۱ و ۲ و ال) پرداختیم.

جهت آزمایش روش پیشنهادی چند مقدار برای نام پیشنهادی و زمینه فعالیت بصورت تصادفی به مدل ساخته شده ارسال و نتیجه در ستون درصد ارتباط معنایی آورده شده است. درصد ارتباط معنایی با مقدار شباهت کسینوسی محاسبه گردید و نشان دهنده این است که نام پیشنهادی با زمینه فعالیت شرکت چند درصد از لحاظ معنایی با هم ارتباط دارند. نتایج بدست آمده نشان دهنده این است که زمانی که نام پیشنهادی با زمینه فعالیت ارتباط معنایی نزدیکی با یکدیگر داشته باشند مدل به درستی درصد بالایی را بر می گرداند و زمانی که فاقد ارتباط معنایی باشند درصد پائینی را بر می گرداند. همچنین ضریب همبستگی بین میزان ارتباط معنایی بدست آمده توسط مدل و نظر واقعی کارشناس ثبت شرکت محاسبه گردید و نتایج نشان دهنده ۹۳ درصد ضریب همبستگی می باشد. در مقایسه ای کلی بر اساس جدول ۷، معیار ROUGE-L با دقت بالاتر نسبت به ROUGE-1 و ROUGE-2 نشان می دهد که مدل در حفظ ساختار معنایی فعالیت های شرکت عملکرد قوی تری دارد. جدول ۸ نیز نشان می دهد که شباهت کسینوسی میان نام و حوزه فعالیت در نمونه های مرتبط بسیار بالا (۹۲٪ و ۸۰٪) و در نمونه نامرتبط پایین (۲۸٪) بوده و همبستگی ۹۳٪ با نظر کارشناس صحت مدل را تأیید می کند. در جدول ۹ نیز نتایج خوشه بندی DBSCAN بیانگر آن است که مدل توانسته نام ها را براساس نزدیکی معنایی در خوشه های درست قرار دهد و نام های مشابه را در حوزه فعالیت مشترک گروه بندی کند. محدودیتهای پژوهش بر اساس پایگاه داده استفاده شده به زبان فارسی برای سازمان ثبت شرکتها می باشد. تنها به میزان مشابهت زمینه فعالیت شرکت و نام شرکت پرداخته شده است.

پیشنهاد می شود در کارهای آتی از الگوریتمهای یادگیری برخطو ترانسفورمرها استفاده شود. همچنین به بررسی معنایی نام شرکت و معنایی زمینه فعالیت های 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 می تواند دقت کلی را افزایش دهد. در پژوهش های جدید نشان داده شده که یادگیری مبتنی بر چند فعالیت همزمان در پردازش زبان طبیعی برای زبان های کم منبع عملکرد بهتری دارد.



Mohammad Rabiei

<http://orcid.org/0000-0001-8728-8825>

۲۷ | مطالعات مدیریت کسب و کار هوشمند | سال چهاردهم | شماره ۴۵ | زمستان

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.

طراحی سیستم سنجش تطابق نام و حوزه فعالیت شرکت ها بر اساس نام و زمینه...؛ ربیعی | ۲۸

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>

DOI:



Journal of Business Intelligence Management Studies is licensed under
a Creative Commons Attribution-NonCommercial 4.0 International License.

آماره انتشار