

## طراحی سیستم توصیه‌گر مقالات فارسی و انگلیسی به کمک مدل زبانی BERT با تمرکز بر چکیده مقاله، عنوان و کلمات کلیدی

احسان ابراهیمی\*<sup>۱</sup>، محمدعلی جوادزاده<sup>۲</sup>، علی امیری<sup>۳</sup>

<sup>۱</sup> دانشجوی کارشناسی ارشد هوش مصنوعی دانشگاه جامع امام حسین(ع)، تهران، ایران، [ebrahimiehsan@ihu.ac.ir](mailto:ebrahimiehsan@ihu.ac.ir)

<sup>۲</sup> استادیار دانشگاه جامع امام حسین(ع)، تهران، ایران، [javadzade@ihu.ac.ir](mailto:javadzade@ihu.ac.ir)

<sup>۳</sup> پژوهشگر دانشگاه جامع امام حسین(ع)، تهران، ایران [aamirij@ihu.ac.ir](mailto:aamirij@ihu.ac.ir)

### چکیده

رشد سریع علوم مختلف در سال‌های اخیر، و رشد و فراگیر شدن شبکه جهانی اینترنت دو عاملی است که باعث شده تا روزانه در فضای اینترنت چندین هزار مقاله در رشته‌های مختلف انتشار یابد، و در دسترس همگان قرار گیرد. این مسئله چالش سربار اطلاعات را برای محققین به هنگام یافتن مقاله متناسب هدفشان به وجود آورده است. متخصصین علم هوش مصنوعی با توسعه سیستم‌های توصیه‌گر تا حد خوبی مشکل سربار اطلاعات را رفع کرده‌اند. اکثر پژوهش‌های ارائه شده در این حوزه برای زبان انگلیسی توسعه یافته، و تا کنون در زبان فارسی پژوهشی به منظور گسترش سامانه توصیه‌گر با هدف توصیه مقالات فارسی ارائه نشده است. هدف از این پژوهش ارائه تحقیقات پایه و توسعه سامانه توصیه‌گر مقالات فارسی است، تا نقطه شروعی برای سایر محققین و توسعه دهندگان باشد. در این پژوهش، پیشنهاد مقالات بر اساس شباهت معنایی عنوان، چکیده و کلمات کلیدی مقاله (با وزن دهی مناسب) با سوابق مطالعاتی کاربر است. پس از پیاده‌سازی و قرار گرفتن مدل به صورت آزمایشگاهی بر روی سرور داخلی دانشگاه بزرگمهر قاین، مدل توسط جمعی از دانشجویان کارشناسی نرم‌افزار کامپیوتر این دانشگاه مورد ارزیابی قرار گرفت، که در مجموع دقت سامانه ۷۹٪ ارزیابی شد.

**واژه‌های کلیدی:** سیستم‌های توصیه‌گر، مدل زبانی، BERT، پردازش زبان طبیعی.

## ۱. مقدمه

دو عامل رشد سریع علوم مختلف در سال‌های اخیر و فراگیر شدن استفاده از کامپیوتر و استفاده از شبکه جهانی اینترنت باعث شده که روزانه در فضای اینترنت چندین هزار مقاله در رشته‌های مختلف انتشار یابد، و در دسترس همگان قرار گیرد [1-4]. این مسئله چالشی را برای متخصصین علوم مختلف ایجاد می‌کند، زیرا باید در انبوهی از مقالات انتشار یافته به دنبال علایق و حوزه علمی خود باشند تا همیشه بتوانند یک متخصص باقی بمانند. این حجم زیاد اطلاعات و استفاده روزانه از محتوای اینترنت، برای تعداد زیادی از کاربران سردرگمی به هنگام پیدا کردن محتوای مورد نیاز را پدید آورده است. این چالش را حوزه‌ای به نام سیستم‌های توصیه‌گر از علم هوش مصنوعی رفع می‌کند [5-7]. گسترش استفاده از این سیستم‌ها در حوزه‌هایی مانند تجارت الکترونیک، آموزش الکترونیک، کتابخانه الکترونیک، کسب و کارهای الکترونیک و برنامه‌های کاربردی مانند سایت‌های موسیقی، فیلم، مقالات، گردشگری و از این دسته، شواهدی از عملکرد خوب و محبوبیت این نوع سیستم‌ها است [8].

سیستم‌های توصیه‌گر با تحلیل رفتار کاربر اقدام به پیشنهاد مناسب‌ترین محتوا به وی می‌کنند. این سیستم‌ها بر مبنای اطلاعات مورد استفاده برای توصیه مقاله به کاربران، به گروه‌های مختلفی طبقه‌بندی می‌شوند. سه طبقه‌بندی کلی برای این سیستم‌ها عبارت‌اند از [9-12]:

- فیلترسازی محتوا محور<sup>۱</sup>
- فیلترسازی مشارکتی<sup>۲</sup>
- فیلترسازی هیبریدی<sup>۳</sup>

برای پیاده‌سازی یک سیستم توصیه‌گر، چالش‌های بسیاری وجود دارد که بسته به نوع سامانه باید به آنها توجه شود. برخی از برجسته‌ترین این چالش‌ها عبارت‌اند از:

- مشکل شروع سرد<sup>۴</sup> [13]
- مشکل پراکندگی داده<sup>۵</sup> [13]
- مشکل مقیاس پذیری<sup>۶</sup> [10]
- مشکل تعصب روی دانش پیشین<sup>۷</sup> [13]
- مشکل تنوع<sup>۸</sup> [14]
- حریم خصوصی<sup>۹</sup> [15]

امروزه سیستم‌های توصیه‌گر بر پایه متن در حال توسعه هستند. سیستم‌های توصیه‌گر برای تجزیه و تحلیل متن و توصیه بر پایه متون از تکنیک‌های مختلف پردازش زبان طبیعی استفاده می‌کنند [16]. یوهوی ژانگ و همکاران سامانه توصیه‌گری بر مبنای مدل‌های زبانی ارائه کردند. تحقیقات آنها نشان می‌دهد که می‌توان مدل‌های زبانی از پیش آموزش دیده مانند GPT<sup>۱۰</sup> و BERT را در سیستم‌های توصیه‌گر به کار گرفت. زیرا مدل‌های زبانی از پیش آموزش دیده، نمایش کلی متن

<sup>1</sup> Content Based

<sup>2</sup> Collaborative Filtering

<sup>3</sup> Hybrid Filtering

<sup>4</sup> Cold Start

<sup>5</sup> SParsity

<sup>6</sup> Scalability

<sup>7</sup> Over Specialization

<sup>8</sup> Diversity

<sup>9</sup> Privacy

<sup>10</sup> Generative Pre-trained Transformer

را یاد می‌گیرند و می‌توانند دانش موجود در متون را رمزگذاری کنند. از این رو از آنها می‌توان برای تشخیص تشابه میان متون استفاده کرد [17]. BERT یک مدل زبانی با مفهوم ساده و قدرت بالا برای فهم زبان و انجام دقیق‌تر وظایف NLP است [18].

در حوزه فارسی پایگاه داده‌ای<sup>۱</sup> شامل مقالات و برجسب آنها برای توسعه سیستم‌های توصیه‌گر وجود ندارد همچنین تا به این لحظه تحقیقات پایه‌ای برای این مسئله در زبان فارسی وجود ندارد تا نقطه شروعی برای سایر محققین باشد. به همین دلیل تحقیقات این حوزه محدود به محتویات لاتین است. در این پژوهش برای اولین بار سیستم توصیه‌گری را معرفی کرده‌ایم که به کاربران فارسی زبان، مقالات مناسبی در جهت علایق آنها پیشنهاد می‌دهد. هدف از ارائه این سیستم کاهش زمان و افزایش دقت به هنگام جستجوی مقالات فارسی توسط دانش‌پژوهان ایرانی است. در این پژوهش پیشنهاد مقالات بر اساس شباهت معنایی عنوان، چکیده و کلمات کلیدی مقاله (با وزن دهی مناسب) است. در این راستا با چالش‌های اصلی زیر روبه‌رو هستیم:

- جمع‌آوری پایگاه داده‌ای از مقالات شامل عنوان، چکیده و کلمات کلیدی
- یافتن شباهت معنایی در متون طولانی (چکیده)
- یافتن مقادیر مناسب برای وزن دهی به عنوان، چکیده و کلمات کلیدی

## ۲. روش تحقیق:

برای رسیدن به اهداف پیش‌بینی شده، روش تحقیق در این پژوهش به سه گام زیر تقسیم می‌شود:

### گام اول:

با توجه به این که هدف از این پژوهش ارائه سامانه توصیه‌گر مقالات فارسی است، پس اولین پیش‌نیاز برای پیاده‌سازی و آزمایش الگوریتم‌های این حوزه، مجموعه داده مناسب است. مجموعه داده مناسب برای این پژوهش به معنی وجود دادگانی شامل عنوان، کلمات کلیدی و چکیده از هزاران مقاله فارسی است. برای حل این چالش ابتدا به بررسی پایگاه داده‌های فارسی جمع‌آوری شده در اینترنت، پرداخته شد. در موارد بررسی شده فقط یک نمونه مشابه در سایت دیتا هارت<sup>۲</sup> یافت شد که دادگانی شامل چکیده، عنوان و کلمات کلیدی ۵۰۰ مقاله بود. به دلیل کم بودن رکوردها، این یک مجموعه داده مناسب برای این پژوهش نبود. با توجه به در دسترس نبودن دادگان مناسب و نیاز الزامی این پژوهش به دادگان مناسب فارسی، در گام اول به جمع‌آوری دادگان پرداخته شده است. برای جمع‌آوری دادگان، سایت سیویلیکا<sup>۳</sup> انتخاب شده است، زیرا بزرگ‌ترین و بروزترین بانک مقالات فارسی است، که شامل ده‌ها هزار مقاله فارسی و انگلیسی است. به منظور استخراج مقالات از چهارچوب<sup>۴</sup> scrapy در زبان برنامه نویسی پایتون استفاده شده است [19].

برای این هدف یک خزنده<sup>۵</sup> مناسب سایت سیویلیکا به کمک scrapy نوشته شده است. این برنامه بعد از به دست آوردن لینک ده هزار مقاله شروع به جمع‌آوری عنوان، چکیده و کلمات کلیدی مقالات می‌کند و در نهایت در یک فایل CSV

<sup>1</sup> Dataset

<sup>2</sup> <http://dataheart.ir>

<sup>3</sup> <https://civilica.com/>

<sup>4</sup> Framework

<sup>5</sup> Crawler

آنها را ذخیره می‌کند. روند جمع‌آوری حدود ۳۰ دقیقه با یک کامپیوتر خانگی<sup>۱</sup> و اینترنت ۱۶ مگابایت بر ثانیه به طول انجامید، و حجم فایل حاصل ۲۱ مگابایت شد. چند رکورد از فایل به دست آمده به شرح زیر است (شکل ۱).

url	keywords	abstract	title
/https://civilica.com/doc/1494432	احساس تعهد، انعطاف پذیری شناختی فرسودگی ...	هدف این پژوهش تعیین رابطه انعطاف پذیری شناختی ...	ارتباط انعطاف پذیری شناختی و احساس تعهد با فرس...
/https://civilica.com/doc/1494452	برنامه ریزی، مدیریت شهری، گردشگری شه...	صنعت گردشگری یکی از پویاترین و بی نظیرترین ص...	نقش مدیریت شهری در گردشگری شهری
/https://civilica.com/doc/1494453	آسیبهای اجتماعی، فضاهای شهری، نقش کل...	در دنیای امروز که پیشرفت تکنولوژی و رشد است ...	جایگاه مدیریتی اجتماعی زنان در فضاهای شهری
/https://civilica.com/doc/1494433	مهارت های فردی، توسعه فردی، عملکرد...	توسعه فردی، تلاش و سرمایه گذاری برای رشد مه...	بررسی و مطالعه توسعه فردی، تلاش و سرمایه گذاری...
/https://civilica.com/doc/1494442	بازی درمانی، الگردگی، اضطراب، شن...	پژوهش حاضر با هدف بررسی اثربخشی بازی درمانی...	اثربخشی بازی درمانی مبتنی بر رویکرد شناختی-رفت...

شکل ۱: چند رکورد از پایگاه داده جمع‌آوری شده

مرحله بعدی تمیز کردن داده‌های جمع‌آوری شده است. برای این منظور کاراکترهای خاصی به شرح زیر از فایل حذف شده‌اند.

- یونی کد چپ به راست (u200e)
- خطوط جدید اضافه (r\n)
- نیم فاصله‌ها (u200b)
- خط تیره‌ای که در انتهای خط به منظور شکست کلمه ظاهر می‌شود. (xad)
- اصلاح spaceهایی که با کد html قرار داده شده‌اند.

یک نوع داده نويز که در فایل جمع‌آوری شده وجود داشت، مقالات با عنوان فارسی اما چکیده انگلیسی یا بالعکس بود. که با نوشتن یک حلقه ساده این داده‌ها نیز از فایل نهایی حذف شدند. در نهایت پس از اصلاحات صورت گرفته و حذف داده‌های نويز، دادگانی با ۹۵۸۵ مقاله به دست آمد. که شامل ۸۳۶۹ مقاله فارسی و ۱۲۱۶ مقاله انگلیسی است.

## گام دوم:

هدف این گام از پژوهش، یافتن میزان شباهت بین عناوین، چکیده‌ها و کلمات کلیدی مقالات با یکدیگر است. برای این منظور گام‌های زیر طی شده‌اند:

- پیش پردازش متن<sup>۲</sup>
- یافتن تعبیه برداری<sup>۳</sup> از عنوان، کلمات کلیدی و چکیده
- یافتن شباهت بین بردارها

پیش پردازش متن یک کار مهم و مرحله حیاتی در پردازش متن است [20]. در گام ابتدایی با ابزار langid<sup>۴</sup> فارسی یا انگلیسی بودن مقاله بررسی شده [21] و سپس برای پیش پردازش زبان فارسی از ابزار هضم استفاده شده است [22]. که برای زبان فارسی این مراحل به شرح زیر است:

- حذف ایست واژه

<sup>1</sup> Core i7 (4710HQ) / 8GB Ram

<sup>2</sup> Text Preprocessing

<sup>3</sup> Embedding Vector

<sup>۴</sup> یک کتابخانه برنامه‌نویسی در زبان پایتون که متون را بر اساس زبان متن کلاس‌بندی می‌کند.

- لمتایز<sup>۱</sup>
- اصلاح ک و ی عربی
- حذف کلمات انگلیسی
- حذف اعداد
- اصلاح پرانتز
- اصلاح فاصله‌های اضافی

و برای پیش پردازش مقالات انگلیسی از ابزار nltk استفاده شده است [23]. که مراحل آن به شرح زیر است:

- حذف ایست واژه
- لمتایز
- حذف کلمات فارسی
- حذف اعداد
- اصلاح پرانتز
- اصلاح فاصله‌های اضافی

بعد از پیش پردازش دادگان، برای تعبیه سازی جملات به بردارهای عددی، از مدل زبانی BERT استفاده شده است. برای جانمایی جملات فارسی از ورژن ۳ مدل از پیش آموزش دیده ParsBERT<sup>۲</sup> استفاده شده است [24]. برای این منظور با استفاده از اکوسیستم هاگینگ فیس<sup>۳</sup> مدل مربوطه را بارگیری و جملات فارسی را به بردارهای ۷۶۸ بعدی تعبیه کرده ایم [25]. و به صورت مشابه برای تعبیه سازی جملات انگلیسی از مدل BERT-Base<sup>۴</sup> از پیش آموزش داده شده در اکوسیستم هاگینگ فیس استفاده شده است. تعبیه سازی در سه بخش به شرح زیر صورت گرفته است.

### بخش اول: تعبیه سازی عناوین

عناوین بسته به فارسی یا انگلیسی بودن با BERT مربوط به خود تعبیه سازی شد و در نهایت یک ماتریس  $9585 \times 768$  بعدی به دست آمد.

### بخش دوم: تعبیه سازی کلمات کلیدی

از آنجایی که مقالات فارسی غالباً تعدادی از کلمات کلیدی شان انگلیسی هستند برای هر مقاله کلمات کلیدی به دو دسته فارسی و انگلیسی تقسیم می شود و با مدل زبانی متناسب تعبیه سازی به دست می آید، که در نهایت یک تنسور<sup>۵</sup> با ابعاد  $9585 \times 2 \times 768$  بعدی حاصل می شود.

### بخش سوم: تعبیه سازی چکیده ها

از آنجایی که جملات چکیده با یک رابطه معنادار به دنبال یکدیگر ظاهر می شوند، برای افزایش دقت، مدل زبانی BERT تنظیم دقیق<sup>۱</sup> شده است. تنظیم دقیق مدل BERT به یکی از چهار روش زیر ممکن است [26]:

<sup>1</sup> Lemmatization

<sup>2</sup> <https://huggingface.co/bert-base-uncased>

<sup>3</sup> <https://huggingface.co/>

<sup>4</sup> <https://huggingface.co/bert-base-uncased>

<sup>5</sup> Tensor

- یک جفت جمله و یک برچسب عددی که نشان دهنده میزان شباهت دو جمله با یکدیگر است.
- یک جفت جمله مثبت (شبیه به هم) بدون برچسب. (مانند جملات پشت سر هم در یک متن یا سوال و پاسخ یا جفت سوالات شبیه یا متن و خلاصه آن)
- یک جمله با یک برچسب عدد صحیح که نشان دهنده کلاسی است که این جمله به آن تعلق دارد.
- یک سه گانه (جمله لنگر<sup>۲</sup> - جمله مثبت - جمله منفی)

با توجه به دادگان جمع آوری شده، ما برای تنظیم دقیق مدل BERT از روش دوم استفاده کرده ایم. به این صورت که ParsBERT را با جملات چکیده مقالات فارسی آموزش داده ایم. هر دو جمله متوالی در یک چکیده را به عنوان جفت جمله مثبت در نظر گرفته شده است. و به صورت مشابه برای BERT-Base همین مراحل را برای چکیده مقالات انگلیسی انجام داده ایم. مراحل تنظیم دقیق مدل با گوگل کولب<sup>۳</sup> رایگان همراه با پردازنده گرافیکی، برای ParsBERT حدود ۱۰۰ دقیقه و برای BERT-Base حدود ۲۵ دقیقه به طول انجامید. عمل تنظیم دقیق با ۵ دوره<sup>۴</sup> برای ParsBERT و ۸ دوره برای BERT-Base انجام شد. تعداد دوره با توجه به منابع محدود برای آموزش مدل در نظر گرفته شده است. (متخصصین هاگینگفیس تعداد ۱۰ دوره را پیشنهاد می دهند [26]) همچنین برای آموزش مدل از تابع هزینه MultipleNegativesRankingLoss استفاده شد.

برای تعبیه سازی چکیده ابتدا حاشیه<sup>۵</sup> مناسب برای تعداد جملات چکیده را به دست آورده ایم. به عبارتی بیشترین تعداد جمله در چکیده ها به دست آمده است. که در دادگان ما این عدد برابر ۳۱ است. سپس عمل تعبیه سازی را با در نظر گرفتن این حاشیه و مدل زبانی تنظیم دقیق شده، انجام داده ایم، که در نهایت یک تنسور  $9585 \times 31 \times 768$  بُعدی حاصل شد.

جدول ۱ نشان دهنده مدت زمان اجرای الگوریتم های ذکر شده به کمک گوگل کولب نسخه رایگان به همراه پردازش گرافیکی است.

جدول ۱: مدت زمان اجرای الگوریتم های تعبیه سازی متون توسط گوگل کولب

عملیات	زمان (بر حسب دقیقه)
تعبیه سازی عناوین	2
تعبیه سازی چکیده ها	4
تعبیه سازی کلمات کلیدی	3

برای به دست آوردن میزان شباهت از معیار شباهت کسینوسی<sup>۶</sup> استفاده شده است. زیرا این معیار در معیارهای ارزیابی مختلف عملکرد خوبی دارد [27,28]. برای عناوین پس از محاسبه میزان شباهت یک ماتریس قطری شباهت  $9585 \times 9585$  بُعدی به دست آمد. (برای مثال درایه ی (۵، ۶) میزان شباهت عنوان مقاله ۵ با مقاله ۶ را نشان می دهد).

<sup>1</sup> Fine Tune

<sup>2</sup> Anchor Sentence

<sup>3</sup> Google Colab

<sup>4</sup> Epoch

<sup>5</sup> Padding

<sup>6</sup> Cosine Similarity

برای کلمات کلیدی نیز به صورت مشابه عمل شده است، که در نهایت یک تنسور  $۹۵۸۵ \times ۲ \times ۷۶۸$  بُعدی حاصل شد. با میانگین گیری وزن دار بین بعد مربوط به انگلیسی و بعد مربوط به فارسی یک ماتریس شباهت کلی  $۹۵۸۵ \times ۹۵۸۵$  بُعدی به دست آمد. (وزن ها همان نسبت کلمات کلیدی فارسی و انگلیسی به کل کلمات کلیدی یک مقاله است)

در شباهت یابی چکیده، یک مقاله را به عنوان مقاله لنگر و یک جمله از چکیده آن را به عنوان جمله لنگر در نظر گرفته ایم سپس شباهت این جمله را با تمام جملات چکیده مقاله دوم به دست آورده ایم. در نهایت بزرگ ترین شباهت را به عنوان میزان شباهت جمله لنگر با چکیده مقاله دوم در نظر گرفته ایم و این کار را برای مقاله سوم، چهارم و تا آخر انجام داده ایم. پس از اتمام، جمله دوم از مقاله لنگر را به عنوان جمله لنگر در نظر گرفته می شود و این روند ادامه پیدا کرد تا تک تک مقالات به عنوان لنگر انتخاب شدند و با سایر مقالات بررسی شدند. در نهایت یک تنسور  $۹۵۸۵ \times ۳۱ \times ۷۶۸$  بُعدی حاصل شد. که برای مثال درایه (۵, ۰, ۰) آن نشان دهنده میزان شباهت جمله اول از چکیده مقاله اول با تمام جملات چکیده مقاله ششم است. سپس از میزان شباهت جملات یک چکیده میانگین گرفته شد تا در نهایت یک ماتریس شباهت  $۹۵۸۵ \times ۹۵۸۵$  بُعدی حاصل شد.

پس از اتمام مراحل فوق سه ماتریس  $۹۵۸۵ \times ۹۵۸۵$  بُعدی به دست آمد که میزان شباهت عنوان ها، کلمات کلیدی مقالات مختلف و چکیده ها را نشان می دهد. حال چالش پیش رو پیدا کردن وزن مناسب برای میانگین وزن دار گرفتن بین این ماتریس ها است. وزن های مناسب برای این میانگین گیری با توجه به نکات زیر تعیین شد:

- پژوهشگران برای انتخاب عنوان مقاله دقیق ترین کلمات را استفاده می کنند و از استفاده از کلمات نامربوط پرهیز می کنند. و به طور کلی یک عنوان نشان دهنده زمینه و روش تحقیق است.
- کلمات کلیدی حوزه های استفاده شده در یک مقاله را نشان می دهند، اما مانند عنوان، روش تحقیق یک مقاله را بیان نمی کنند.
- چکیده مقالات مختلف پر از عبارات مشابه هستند که در پیش بینی حوزه مقاله تاثیری ندارند. برای مثال در اکثر مقالات از عباراتی چون "در این پژوهش قصد داریم ....."، "هدف از انجام این پژوهش ....."، "چالش موجود در تحقیقات قبلی ..... " استفاده می شود. تجربه در شباهت یابی مقالات نشان می دهد که دقت شباهت یابی در چکیده مقالات به اندازه دقت شباهت یابی در کلمات کلیدی و عنوان نیست. دلیل این امر وجود عبارت های مرسوم در مقالات است.


با توجه به نکات فوق و نظرسنجی از اساتید این حوزه، این نتیجه حاصل شد که از وزن ۰.۲ برای چکیده، ۰.۳ برای کلمه کلیدی و ۰.۵ برای عنوان استفاده شود. در نهایت با میانگین گیری وزن دار یک ماتریس شباهت کلی  $۹۵۸۵ \times ۹۵۸۵$  بُعدی به دست آمد.

از ماتریس به دست آمده فوق هم برای شباهت یابی مبتنی بر مورد<sup>۱</sup> و هم برای شباهت یابی بر اساس سوابق نمایه<sup>۲</sup> هر کاربر استفاده شده است. به این صورت که در گام اول یک سایت با باطن<sup>۳</sup> پایتون طراحی می شود. در صفحه جزئیات هر مقاله ۶ تا از مرتبط ترین مقالات از ماتریس شباهت به دست آمده استخراج می شود و با درصد شباهتشان به کاربر نمایش داده می شود. این توصیه مستقل از سوابق مطالعاتی کاربر است و به حل مشکل شروع سرد مربوط به مورد جدید کمک می کند. (شکل ۲)

<sup>1</sup> Item

<sup>2</sup> Profile

<sup>3</sup> Backend

<p>ارتباط انعطاف پذیری شناختی و احساس تعهد با فرسودگی شغلی در زمان شیوع کووید-۱۹</p> <p>آدرس این مقاله در سیولیکا</p>	 
<p>کلمات کلیدی: احساس تعهد ، انعطاف پذیری شناختی فرسودگی شغلی ، کووید- ۱۹.</p>	<p>چکیده:</p>
<p>هدف این پژوهش تعیین رابطه انعطاف پذیری شناختی و احساس تعهد با فرسودگی شغلی در زمان شیوع کووید-۱۹ بود. پژوهش از نظر هدف، کاربردی و از لحاظ روششناسی توصیفی-همبستگی بوده است. جامعه آماری شامل کلیه کارکنان داروخانه های شهر رشت بودند که از میان آنان ۱۷۰ نفر به روش نمونه گیری در دسترس انتخاب شدند. در این مطالعه جهت گردآوری داده های پژوهش از پرسشنامه های انعطاف پذیری شناختی دنیس و وندروال (۲۰۱۰)، تعهد سازمانی آلن و مایر (۱۹۹۷) و فرسودگی شغلی ماسلش (۱۹۸۱) استفاده گردید. جهت تجزیه و تحلیل داده ها از آزمون رگرسیون چندگانه و نرم افزار SPSS ۲۰ استفاده شد. نتایج بدست آمده نشان داد بین انعطاف پذیری شناختی و فرسودگی شغلی کارکنان ارتباط منفی و بین احساس تعهد و فرسودگی شغلی کارکنان ارتباط مثبت وجود دارد. همچنین بین انعطاف پذیری شناختی با فرسودگی شغلی در زمان شیوع کووید-۱۹ رابطه وجود دارد. همچنین حدود ۳۴ درصد از تغییرات فرسودگی شغلی کارکنان از طریق انعطاف پذیری شناختی و بدبینی کارکنان قابل پیشبینی بود. (<math>p &lt; 0.05</math>) در زمان شیوع کووید-۱۹ جهت ارائه عملکرد شایسته کارکنان داروخانه ها و کنترل سطح فرسودگی شغلی آنان، باید به شاخصهای انعطاف پذیری شناختی و احساس تعهد توجه مناسب گردد.</p>	
<p>موارد مشابه با این مقاله:</p>	
<p>نقش حساسیت اضطرابی و تشدید جسمانی در پیشبینی اضطراب کووید-۱۹ دانشجویان</p> <p>اضطراب کووید-۱۹ ، اضطراب سلامت ، تشدید جسمانی حسی ، حساسیت اضطرابی</p> <p>22.59.13 2022-09-12</p>	<p>بررسی وضعیت اضطراب ناشی از کووید-۱۹ و ارتباط آن با برخی متغیرهای جمعیت شناختی</p> <p>اضطراب ناشی از کرونا ویروس ، کووید- ۱۹ ، متغیرهای جمعیت شناختی</p> <p>22.59.14 2022-09-12</p>
<p>رابطه اضطراب مرگ با اختلال سازگاری: نقش واسطه ای خودبیمارلگاری در دوران شیوع کرونا</p> <p>اضطراب مرگ ، خودبیمار لگاری ، اختلال سازگاری</p> <p>22.59.10 2022-09-12</p>	<p>مقایسه بیانگری هیجانی- اجتناب شناختی و مهارت های اجتماعی در دانش آموزان با و بدون مشکلات یادگیری خاص</p> <p>بیانگری هیجانی ، اجتناب شناختی ، مهارت های اجتماعی</p> <p>22.59.11 2022-09-12</p>
<p>رابطه بین هوش سازمانی و تنظیم شناختی هیجانی با فرسودگی شغلی در کارکنان بهزیستشهر مرودشت</p> <p>هوش سازمانی ، تنظیم شناختی هیجانی ، فرسودگی شغلی.</p> <p>22.59.06 2022-09-12</p>	<p>مقایسه راهبردهای تنظیم شناختی هیجان و ذهن آگاهی بیماران کووید-۱۹ و افراد غیربیمار در شرایط قرنطینه خانگی</p> <p>راهبردهای تنظیم شناختی هیجان ، ذهن آگاهی ، قرنطینه خانگی ، کووید- ۱۹</p> <p>22.59.07 2022-09-12</p>

شکل ۲: صفحه جزئیات یک مقاله

هر کاربر توانایی امتیاز دهی از ۱ تا ۵ به مقالات مطالعاتی خود را دارد. یعنی جمع آوری اطلاعات از کاربر به صورت صریح است [29]. سپس برای هر کاربر یک نمایه ساخته شده و در نمایه وی دو دسته مقاله به او توصیه می شود. دسته اول ۴ مقاله پیشنهادی که بیشترین شباهت را به محبوب ترین و جدیدترین مقالات مطالعه شده توسط هر کاربر دارند. دسته دوم، ۲ مقاله پیشنهادی که پر بازدید، محبوب و جدید هستند. پیشنهادهای دسته دوم برای حل مشکل حباب فیلتری<sup>۱</sup> (مشکل تعصب روی دانش پیشین) و مشکل شروع سرد مربوط به کاربر جدید است. (شکل ۳)

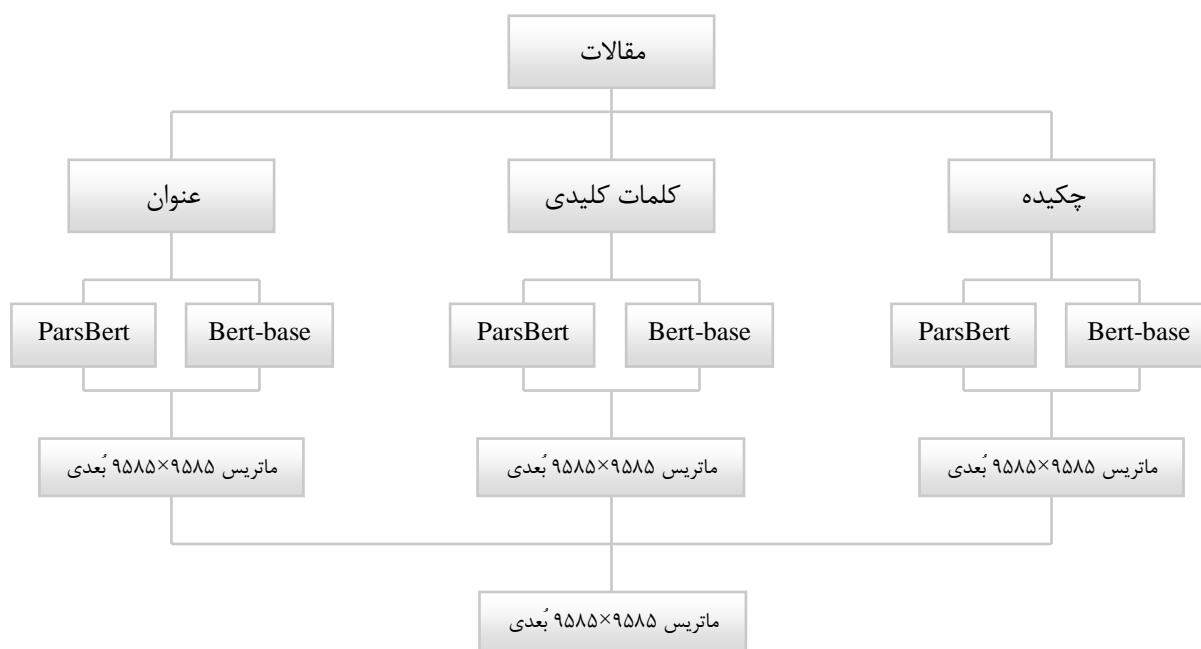
<sup>1</sup> Filter Bubbles



با توجه به سوابق مطالعاتی شما توصیه می‌کنیم موارد زیر را نیز بررسی کنید.			
<p>اثر بخشی درمان شناختی و رفتاری بر نشخوار فکری و کاهش افسردگی در نوجوانان دختر شهر شیراز</p> <p>درمان شناختی و رفتاری ، نشخوار فکری ، افسردگی</p>		<p>مقایسه راهبردهای تنظیم شناختی هیجان و ذهن آگاهی بیماران کووید-۱۹ و افراد غیرمبتلا در شرایط قرنطینه خانگی</p> <p>راهبردهای تنظیم شناختی هیجان ، ذهن‌آگاهی ، قرنطینه خانگی ، کووید-۱۹</p>	
<p>مطالعه و بررسی خود تنظیمی هیجانی، انعطاف پذیری شناختی و تفکر فعال روشنگرانه و تأثیرگذاری آن بر میزان بهره‌وری در سازمان</p> <p>خود تنظیمی هیجانی ، انعطاف پذیری شناختی ، تفکر فعال روشنگرانه ، بهره‌وری سازمان</p>		<p>بررسی و مطالعه همسوسازی فعالیت‌های مدیریت منابع انسانی از طریق شناخت ویژگی‌های سازمان‌های تعاونی و تشویق حرفه‌گرایی</p> <p>همسوسازی ، فعالیت‌ها ، مدیریت منابع انسانی ، شناخت ویژگی‌ها ، سازمان‌های تعاونی ، تشویق حرفه‌گرایی</p>	
توصیه می‌کنیم موارد زیر هم بررسی شود.			
<p>بررسی و مطالعه توسعه فردی، تلاش و سرمایه‌گذاری برای رشد مهارت‌های فردی در شرایط مختلف سنی، تحصیلی و موقعیتی</p> <p>مهارت‌های فردی ، توسعه فردی ، عملکرد سازمانی ، منابع انسانی</p>		<p>اثر بخشی بازی درمانی مبتنی بر رویکرد شناختی-رفتاری بر افسردگی و اضطراب دانش آموزان دختر پایه سوم ابتدایی شهر ساری</p> <p>بازی درمانی ، افسردگی ، اضطراب ، شناختی ، رفتاری</p>	

شکل ۳: توصیه تعدادی مقاله در صفحه نمایه کاربر

به طور کلی در این پژوهش روند زیر برای به دست آوردن ماتریس شباهات طی شده است. (شکل ۴)



شکل ۴: مراحل طی شده برای به دست آوردن ماتریس شباهت مقالات

## گام سوم

در گام سوم به ارزیابی مدل ارائه شده پرداخته شده است. با توجه به اینکه یادگیری مدل پیش‌بینی شده از نوع بدون نظارت<sup>۱</sup> است، و پایگاه داده‌ای بر چسب‌دار برای ارزیابی مدل ارائه شده وجود ندارد؛ ارزیابی مدل مذکور بر مبنای ارزیابی انسانی است. به این صورت که بعد از پیاده‌سازی، مدل به صورت آزمایشگاهی روی بستر داخلی دانشگاه بزرگمهر قاین قرار گرفت و توسط جامعه‌ای از دانشجویان کارشناسی نرم‌افزار کامپیوتر مدل مورد ارزیابی قرار گرفت.

ارزیابی از دانشجویان با پرس کردن فرم نظرسنجی کتبی و از نوع کمی صورت گرفته است. ابتدا شیوه کارکرد با سامانه، هدف از پیاده‌سازی سامانه، مواردی که باید مورد ارزیابی قرار گیرند و فرضیه‌های مسئله برای آزمایش کنندگان توضیح داده شد. و سپس از آنها تقاضا شد پس از تعامل با سامانه فرم‌های ارزیابی را پر کنند. حاصل ارزیابی توسط دانشجویان به شرح زیر است.

جدول ۲: ارزیابی مدل ارائه شده توسط ۷۰ دانشجویان

میانگین تعداد مقاله بررسی شده توسط دانشجویان	۳ مقاله
میانگین زمان سپری شده تعامل دانشجویان با سامانه	۱۹ دقیقه
میانگین رضایت دانشجویان از آسان بودن تعامل با سامانه	۹۲٪
میانگین رضایت دانشجویان از مقالات توصیه شده در نمایه کاربر	۷۹٪
میانگین رضایت دانشجویان از مقالات پیشنهادی مرتبط با هر مقاله	۸۵٪

## ۳. نتیجه:

سیستم‌های توصیه‌گر امروزه یک زمینه تحقیقاتی مهم هستند [13]. زیرا رشد روز افزون داده‌ها و کاربران اینترنت، باعث شده اکثر محققان دیگر در کتابخانه‌های فیزیکی به دنبال منابع آموزشی نباشند. ظهور سیستم‌های توصیه‌گر به حل مشکل سربار اطلاعات کمک زیادی کرده‌اند. در این مقاله برای اولین بار یک تحقیقات پایه در رابطه با سیستم توصیه‌گر مقالات فارسی با بیان جزئیات کامل ارائه شده است. در این پژوهش با استفاده از مدل زبانی BERT و ابزارهای موجود برای زبان فارسی، سیستم توصیه‌گری پیاده‌سازی شد که مورد توجه ارزیاب‌ها قرار گرفت. پس از پیاده‌سازی و اجرای مدل پیشنهادی به صورت آزمایشگاهی و ارزیابی مدل به صورت ارزیابی نمونه بزرگ به این نتیجه رسیدیم که مدل پیشنهادی ۷۹٪ رضایت کاربران را در جهت هدف توسعه این سیستم به دست آورده است.

<sup>1</sup> Unsupervised

## ۴. پیشنهادات

در ادامه این پژوهش پیشنهاد می‌شود که راه حلی به منظور به‌روز رسانی ماتریس شباهت با ورود مقالات جدید، به این سامانه افزوده شود. همچنین توصیه می‌شود که مدل پایه BERT با دامنه وسیعی از هزاران مقاله آموزش داده شود. با این کار مدل BERT تعبیه سازی دقیق‌تری از جملات مقالات علمی فارسی خواهد داشت.

## منابع:

- [1] H. Benbya, N. Nan, H. Tanriverdi and Y. Yoo, (Dec 2020), "Complexity and Information Systems Research in the Emerging Digital World," MIS Quarterly, 44(1), pp. 1-17.
- [2] P. Zikopoulos, C. Eaton, and IBM, (2011), "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", 1st. McGraw-Hill Osborne Media.
- [3] F. Liu, P. Fontelo, (2018), "A review of recent publication trends from top publishing countries". Syst Rev 7, 147.
- [4] F. Isinkayea, Y. Folajimi and B. Ojokoh, (2015), "Recommendation systems: Principles, methods and evaluation", *Egyptian Informatics Journal*, pp. 261–273.
- [5] J. K. Tarus, Z. Niu, and D. Kalui, (2018), "A hybrid recommender system for e-learning based on context awareness and sequential pattern mining," *Soft Comput.*, vol. 22, no. 8, pp. 2449–2461.
- [6] A. Abdi, N. Idris, R. M. Alguliyev, and R. M. Aliguliyev, (2017), "Query-based multi-documents summarization using linguistic knowledge and content word expansion," *Soft Comput.*, vol. 21, no. 7, pp. 1785–1801.
- [7] R. Mu, (2018), "A Survey of Recommender Systems Based on Deep Learning", in *IEEE Access*, vol. 6, pp. 69009-69022.
- [8] J. Lu, D. Wu, M. Mao, W. Wang and G. Zhang, (2015), "Recommender system application developments: A survey," *Decision Support Systems*, pp. 12-32.
- [9] Ç. Erion, M. Morisio, (2017), "Hybrid recommender systems: A systematic literature review," *Intelligent Data Analysis* 21.6, pp. 1487-1524.
- [10] L. S. Sree, T. Lakshmi, (2014), "Recommendation systems: Issues and challenges," *International Journal of Computer Science and Information Technologies* 5.4, pp. 5771-5772.
- [11] Aljunid, Mohammad, D. Manjaiah, (2017), "A survey on recommendation systems for social media using big data analytics," *International Journal of Latest Trends in Engineering and Technology*, pp. 48-58.
- [12] S. Lalita, A. Gera, (2013), "A survey of recommendation system: Research challenges," *International Journal of Engineering Trends and Technology (IJETT)* 4.5, pp. 1989-1992.
- [13] Mohamed, M. Hussien, M. Helmy Khafagy and M. H. Ibrahim, (2019), "Recommender systems challenges and solutions survey," 2019 International Conference on Innovative Trends in Computer Engineering (ITCE), pp. 149-155.
- [14] Maan, Tanya, S. Gupta and A. Mishra, (2018), "A Survey On Recommendation System," international conference on recent innovations in management, engineering, science and technology (RIMEST 2018), pp. 543-549.
- [15] H. Mahmoud, A. Hegazy and M. Khafagy, (2018), "An approach for big data security based on Hadoop distributed file system," 2018 International Conference on Innovative Trends in Computer Engineering (ITCE), pp. 109-114.

- [16] W. Guo, H. Gao, J. Shi, B. Long, L. Zhang, B.-C. Chen and D. Agarwal, (July 2019), "Deep Natural Language Processing for Search and Recommender Systems," Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, p. 3199–3200.
- [17] Y. Zhang, H. Ding, Z. Shui, Y. Ma, J. Zou, A. Deoras and H. Wang, (2021), "Language Models as Recommender Systems: Evaluations and Limitations," I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop.
- [18] J. Devlin, M.-W. Chang, K. Lee, και K. Toutanova, (2018), "BERT Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv.
- [19] X. Ma and M. Yan, (2021), "Design and Implementation of Craweper Based on Scrapy", Journal of Physics: Conference Series, vol. 2033, no. 1, p. 012204, Sep.
- [20] V. Gurusamy and S. Kannan, (2014), "Preprocessing Techniques for Text Mining", 10.
- [21] M. Lui and T. Baldwin, (2012), "langid.py: An Off-the-shelf Language Identification Tool", in Proceedings of the ACL 2012 System Demonstrations, pp. 25–30.
- [22] "Python library for digesting Persian text", github, 2018. [Online]. Available: <https://github.com/sobhe/hazm>. [Accessed: 2020].
- [23] E. Loper and S. Bird, (2002), "NLTK: The Natural Language Toolkit", arXiv [cs.CL].
- [24] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, (2021), "ParsBERT: Transformer-based model for Persian language understanding", Neural Process. Lett., vol. 53, no. 6, pp. 3831–3847.
- [25] T. Wolf et al., (2019), "HuggingFace's transformers: State-of-the-art natural language processing", arXiv [cs.CL].
- [26] O. Espejel, "Train and Fine-Tune Sentence Transformers Models", huggingface. [Online]. Available: <https://huggingface.co/blog/how-to-train-sentence-transformers>.
- [27] M. Wijewickrema, V. Petras, and N. Dias, (2019), "Selecting a text similarity measure for a content-based recommender system: A comparison in two corpora", Electron. libr., vol. ahead-of-print, no. ahead-of-print.
- [28] F. Fkih, (2021), "Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison", J. King Saud Univ. - Comput. Inf. Sci.
- [29] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, (2007), "User Profiles for Personalized Information Access", in The Adaptive Web, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 54–89.

## Designing a recommender system for Persian and English articles using the BERT language model, focusing on article abstract, title and keywords

Ehsan Ebrahimi<sup>1</sup>, Mohammad Ali JavadZadeh<sup>2</sup>, Ali Amiri<sup>3</sup>

<sup>1</sup> Master's student in Artificial Intelligence, Imam Hossein University, ebrahimiehsan@ihu.ac.ir

<sup>2</sup> Assistant Professor of Imam Hossein University, javadzade@ihu.ac.ir

<sup>3</sup> Researcher at Imam Hossein University, aamirij@ihu.ac.ir

**Abstract**— Two factors of the rapid growth of various sciences in recent years and the growth and spread of the global Internet network have caused several thousand articles in various fields to be published on the Internet every day, and become available to everyone. This problem creates a challenge for the specialists of different sciences, because they have to look for their interests and scientific field in a large number of published articles so that they can always remain an expert. In this research, for the first time, we intend to introduce a recommender system that suggests suitable articles to Persian language users according to their interests. The purpose of this research is to provide basic research for this category of issues in Persian language, to be a starting point for other researchers. In this study, we want to propose articles based on the semantic similarity of the title, abstract and keywords of the article (with appropriate weighting) with the user's academic records. After the implementation and placement of the model as a pilot on the internet platform. A group of undergraduate students evaluated the model's computer software, and the overall accuracy of the system was evaluated at 79%.

**Keywords:** Recommender Systems, Language Model, BERT, Natural Language Processing