Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

ISSN: 2783-2538

Research Article

Topic Extraction from Persian Texts Using BERTopic Framework, Language Embedding Models, and Text Clustering

Zahra Fallahi [a], Mohammad Rahmanimanesh [b,*]

[a] *Department of Electrical and Computer Engineering, Faculty of Electrical and Computer Engineering, Semnan University, Semnan, Iran*
[b] *Department of Electrical and Computer Engineering, Faculty of Electrical and Computer Engineering, Semnan University, Semnan, Iran*

PAPER INFO

Paper history:

Received:
Revised:
Accepted:

Keywords:

Topic Extraction;
BERTopic;
NMF;
LDA;
LSA;
Persian Text.

ABSTRACT

With the growth of information, extracting knowledge from textual collections has become essential. Topic modeling is an unsupervised machine learning technique that uncovers the hidden themes in documents. In this paper, inspired by BERTopic, we present an unsupervised method for topic modeling on Persian texts. The proposed approach employs the LaBSE language embedding model to convert texts into embedding vectors, then reduces their dimensions using UMAP, and finally groups similar texts into clusters using the K-Means algorithm. Next, by forming a cluster-token matrix and applying a topic representation technique, various topics are extracted from each cluster. We compared LaBSE model with other language embedding models including XLM-R, ParsBERT, Paraphrase-multilingual-MiniLM-L12-v2, Shiraz, and HooshvareLab (RoBERTa). We also compared the K-Means and HDBSCAN clustering algorithms. For evaluation, the AsreIran dataset was used, and both the coherence evaluation metric (NPMI) and human evaluation confirmed the proposed method's performance. In HDBSCAN, Hooshvare (RoBERTa) yielded the best coherence, while ParsBERT excelled in human evaluation. In K-Means, Paraphrase-multilingual-MiniLM-L12-v2 performed best in terms of coherence and LaBSE in human evaluation. The superiority of K-Means over HDBSCAN was also verified. Furthermore, using the AsreIran and Tasnim datasets separately, the proposed method was compared with non-negative matrix factorization, latent Dirichlet allocation, and latent semantic analysis, with results demonstrating its outstanding performance.

*** Corresponding author.**
E-mail address: rahmanimanesh@semnan.ac.ir

How to cite this article:

مقاله پژوهشی

استخراج موضوع از متون فارسی با استفاده از چهارچوب BERTopic، مدل‌های تعبیه زبانی و خوشه بندی متن

زهرا فلاحی[۱]، محمد رحمانی منش[۲،*]

اطلاعات مقاله

دریافت مقاله:
بازنگری مقاله:
پذیرش مقاله:

واژگان کلیدی:
استخراج موضوع،
BERTopic،
NMF،
LDA،
LSA،
متن فارسی.

چکیده

با رشد اطلاعات، استخراج دانش از مجموعه‌های متنی اهمیت یافته است. استخراج موضوع، تکنیکی بدون نظارت در یادگیری ماشین است که مضامین پنهان اسناد را کشف می‌کند. در این مقاله، با الهام از BERTopic، روشی بدون نظارت برای استخراج موضوع از متون فارسی ارائه شده است. روش پیشنهادی از مدل تعبیه‌زبانی LaBSE برای تبدیل متون به بردارهای تعبیه استفاده می‌کند. سپس، با استفاده از UMAP، ابعاد بردارهای تعبیه را کاهش می‌دهد و پس از آن، با استفاده از الگوریتم خوشه‌بندی K-Means، متون مشابه را در خوشه‌های یکسان قرار می‌دهد. سپس با تشکیل ماتریس خوشه-توکن و تکنیک بازنمایی موضوعات، موضوعات مختلف را به ازای هر خوشه از متن استخراج می‌کند. ما مدل LaBSE را با مدل‌های تعبیه‌زبانی XLM-R، ParsBERT، Paraphrase-multilingual-MiniLM-L12-v2، Shiraz و HooshvareLab (RoBERTa) مقایسه کردیم. همچنین مقایسه‌ای بین الگوریتم‌های K-Means و HDBSCAN انجام دادیم. برای ارزیابی روش پیشنهادی، از مجموعه داده عصر ایران استفاده شد. معیار ارزیابی انسجام (NPMI) و معیار ارزیابی انسانی عملکرد روش پیشنهادی را تأیید کردند. در الگوریتم HDBSCAN، مدل Hooshvare (RoBERTa) بر اساس معیار انسجام، و مدل ParsBERT بر اساس ارزیابی انسانی بهترین نتایج را ارائه داد. در K-Means، مدل Paraphrase-multilingual- MiniLM-L12-v2 مطابق معیار انسجام و LaBSE مطابق ارزیابی انسانی، نتایج بهتری داشت. برتری K-Means نسبت به HDBSCAN نیز تأیید شد. همچنین با استفاده از دو مجموعه داده عصر ایران و تسنیم به صورت جداگانه، روش پیشنهادی با مدل‌های فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان و تحلیل معنایی پنهان مقایسه شد. نتایج مقایسه، عملکرد برجسته روش پیشنهادی را نشان می‌دهد.

DOI: https://doi.org/10.22075/jme.2024.31282.2495

۱- مقدمه[۱]

استخراج موضوع یک تکنیک یادگیری ماشینی بدون نظارت است که به شناسایی موضوعات موجود در مجموعه

* پست الکترونیک نویسنده مسئول: rahmanimanesh@semnan.ac.ir
۱. دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران
۲. دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران

استناد به این مقاله: نحوه استناد فارسی در اینجا درج گردد.

گسترده‌ای از اسناد می‌پردازد. این روش به سازماندهی، درک و خلاصه کردن مجموعه‌های بزرگی از اطلاعات متنی کمک می‌کند. همچنین کاربردهای مهمی در بسیاری از زمینه ها مانند پردازش زبان طبیعی و بازیابی اطلاعات دارد. تحقیقات در حوزه مدل‌سازی موضوعی از اواخر دهه ۱۹۹۰ با معرفی روش‌هایی مانند تحلیل معنایی پنهان[۲] [۱] آغاز شد که به کشف روابط معنایی بین کلمات و اسناد می‌پرداخت. در سال ۲۰۰۳، تخصیص دیریکله پنهان[۳] [۲] ارائه شد و به یکی از پرکاربردترین روش‌ها تبدیل شد. فاکتورسازی ماتریس غیرمنفی[۴] [۳] نیز برای کاهش ابعاد و شناسایی موضوعات پنهان معرفی شد. در دهۀ پس از ۲۰۱۰، با ظهور شبکه‌های عصبی، امکان تحلیل الگوهای متنی پیچیده فراهم شد. در سال‌های اخیر، پیشرفت در مدل‌های زبانی ازپیش آموزش‌دیده در مدل‌سازی موضوع نقش مهمی ایفا نموده است. برای مثال، BERTopic [۴] یک تکنیک مدل‌سازی موضوعی است که می‌تواند از مدل‌های تعبیه زبانی مبتنی‌بر ترنسفورمرها مانند BERT و روش فراوانی وزنی کلمه کلیدی[۵] مبتنی‌بر کلاس، برای ایجاد خوشه‌های متراکم استفاده کند. این مدل، همچنین از UMAP[۶] [۵] برای کاهش ابعاد تعبیه‌ها قبل از خوشه‌بندی استفاده می‌کند. BERTopic به‌طور پیش‌فرض برای خوشه‌بندی از HDBSCAN[۷] استفاده می‌کند [۶]. آزمایش‌ها نشان می‌دهد که BERTopic عملکرد رقابتی و پایدار را در انواع وظایف نشان می‌دهد [۷]. با جداسازی فرآیند خوشه‌بندی اسناد و تولید بازنمایی موضوع، انعطاف‌پذیری قابل‌توجهی در این مدل ارائه می‌شود که امکان استفاده آسان را فراهم می‌کند.

در زبان فارسی، با توجه به تعداد اندک تحقیقات صورت گرفته، پژوهش‌های بیشتر برای درک جامع فرصت‌ها و چالش‌های مرتبط با تجزیه و تحلیل خودکار متن و ارائه راه حل، ضروری است [۸]. این نوع شناسایی دانش در تعیین اولویت‌های پژوهشی و تدوین برنامه‌های راهبردی سیاست‌گذاران این حوزه، اهمیت بالایی دارد و مسیر همواری را برای برنامه‌ریزان و سیاست‌گذاران علمی-پژوهشی برای تدوین برنامه‌های راهبردی فراهم می‌کند [۹]. در این پژوهش، قصد داریم پس از انجام پیش‌پردازش بر روی مجموعه داده خبری، عملیات استخراج موضوع را بر روی این مجموعه داده با الهام از چهارچوب BERTopic انجام داده و راه حلی جهت دست‌یابی به بهترین موضوعات و بهبود نتایج حاصل از اعمال این مدل در متون فارسی ارائه نماییم. سپس مقایسه‌ای بین عملکرد روش پیشنهادی و الگوریتم‌های تخصیص دیریکله پنهان، تحلیل معنایی پنهان و فاکتورسازی ماتریس غیرمنفی انجام می‌دهیم که به ایجاد درک بهتری از چالش‌های موجود و در نهایت در رسیدن به روشی مناسب و دقیق در این زمینه کمک نماید. ارزیابی‌ها تأیید کردند که روش پیشنهادی این مقاله در استخراج موضوع از متون فارسی عملکرد قابل توجهی دارد، زیرا موضوعات استخراج شده از انسجام و کیفیت بالایی برخوردار هستند. نتایج همچنین لزوم استفاده از پیش‌پردازش متون را نشان داده و نیز بر اهمیت استفاده از مدل‌های پیشرفته زبانی در تحلیل متون تأکید دارد. در مقایسۀ روش پیشنهادی با فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان و تحلیل معنایی پنهان، معیارهای ارزیابی انسجام و تنوع نشان دادند که روش پیشنهادی به طور کلی نتایج بهتری را در مقایسه با سه مدل دیگر نشان می‌دهد.

ادامه این مقاله به شرح زیر سازماندهی شده است: بخش ۲ یک نمای کلی از مطالعات قبلی در این زمینه ارائه می دهد. بخش ۳ مبانی نظری و بخش ۴ روش پیشنهادی پژوهش را ارائه می‌کند. بخش ۵ به آزمایش‌ها، تنظیمات پیاده‌سازی، ارائه نتایج و ارزیابی می پردازد. بخش ۶ به بحث و تحلیل پیرامون نتایج پرداخته و چالش‌ها و محدودیت‌های پیش آمده در طول پژوهش را عنوان می‌کند. در نهایت، بخش ۷ نتیجه‌گیری نهایی را ارائه می‌نماید.

۲- مرور کارهای پیشین

در این بخش، مروری بر مطالعات پیشین و پژوهش‌های صورت گرفته در زمینه استخراج موضوع از متون فارسی انجام می‌دهیم.

آنوپا و همکاران [۱۰] روشی مقیاس‌پذیر برای استخراج مفاهیم و یادگیری سلسله مراتب مفهومی از متن‌های بدون ساختار ارائه دادند، که با استفاده از مدل‌های موضوعی

[2] Latent Semantic Analysis (LSA)
[3] Latent Dirichlet Allocation (LDA)
[4] Non-negative Matrix Factorization (NMF)
[5] frequency-inverse document frequency (TF-IDF)

[6] Uniform Manifold Approximation and Projection
[7] Hierarchical Density-Based Spatial Clustering of Applications with Noise

هدایت می‌شود. این روش مفاهیم را از موضوعات کشف شده آماری استخراج کرده و سلسله‌مراتبی از مفاهیم می‌سازد. نتایج آزمایش‌ها نشان داد که این روش، عملکرد بهتری نسبت به برخی روش‌های موجود دارد. ابراهیمی و همکاران [۱۱]، با استفاده از تکنیک تحلیل هم‌واژه و تخصیص دیریکله پنهان، به استخراج موضوع از چکیده‌ها و مقالات پرداختند. نتایج نشان داد که تخصیص دیریکله پنهان به‌خوبی می‌تواند موضوعات مقالات را شناسایی کند. امیرخوانی، آذری جعفری و همکاران [۱۲]، مجموعه داده FarsTail را برای استنتاج زبان طبیعی فارسی ارائه کردند. این مجموعه، شامل ۱۰۳۶۷ نمونه بوده و به روشی مشابه مجموعه داده SciTail طراحی شده است. روش‌های مختلف از جمله ParsBERT برای نمایش جملات ورودی به کار رفتند و نتایج نشان داد که بهترین دقت آزمون با مدل مبتنی‌بر BERT به دست آمد که نشان‌دهنده فضای زیادی برای بهبود مدل‌ها در کاربردهای زبان فارسی بود. فراهانی و همکاران [۱۳]، از مدل ParsBERT برای پردازش زبان طبیعی فارسی استفاده کردند و مدل را در سه کار تجزیه و تحلیل احساسات، طبقه‌بندی متن و شناسایی موجودیت نامدار ارزیابی کردند. ParsBERT توانست نتایج بهتری نسبت به مدل‌های چندزبانه ارائه دهد و در تمامی کارها به بهترین عملکرد دست یافت. ابوزاید و الخلیفه [۱۴]، BERTopic را با مدل‌های مختلف زبان عربی از پیش آموزش دیده برای تعبیه متن آزمایش کردند و نتایج آن را با تخصیص دیریکله پنهان و فاکتورسازی ماتریس غیرمنفی مقایسه کردند. آن‌ها به این نتیجه رسیدند که BERTopic عملکرد بهتری در مقایسه با سایر روش‌ها دارد و نتایج بهتری در استخراج موضوعات در زبان عربی ارائه می‌دهد. شائو [۱۵]، از رویکرد گراف معنایی کپسول برای تشخیص موضوع از اخبار آنلاین استفاده کرد. این مدل با استفاده از گراف‌ها به نمایش روابط معنایی بین کلمات در اخبار پرداخت و نیز توانست بهبودهایی در دقت، recall و امتیاز F1 در مقایسه با مدل‌های سنتی ارائه دهد. گارسیا و برتون [۱۶]، در مطالعه‌ای بر روی توییت‌های پرتغالی مرتبط با کووید-۱۹، تجزیه و تحلیل احساسات و تشخیص موضوع را انجام دادند. آن‌ها مدل‌های تعبیه‌شده مانند SBert، mUSE و Fast-Text را برای استخراج ویژگی ترکیب کردند. همچنین از الگوریتم GSDMM که برای متون کوتاه مناسب است، استفاده کردند تا هر سند را به یک موضوع واحد مرتبط کنند. مارتن گروتندورست [۳]، مدل BERTopic را برای استخراج موضوع با استفاده از مدل‌های زبانی پیشرفته توسعه داد. BERTopic تعبیه سند را با مدل‌های زبان مبتنی‌بر ترانسفورماتور از پیش آموزش‌دیده ایجاد می‌کند و پس از خوشه‌بندی اسناد، با روش c-TF-IDF به ایجاد موضوعات مختلف می‌پردازد. BERTopic در بین مدل‌های سنتی و جدید، رقابتی عمل کرده و موضوعات منسجمی را تولید می‌کند. اِگِر و یو [۷]، پست‌های توییتر را برای ارزیابی عملکرد چهار مدل مختلف موضوعی، از جمله تخصیص دیریکله پنهان، فاکتورسازی ماتریس غیرمنفی، Top2Vec و BERTopic، بررسی کردند. نتایج نشان دادند که BERTopic و فاکتورسازی ماتریس غیرمنفی عملکرد بهتری در تجزیه و تحلیل داده‌های توییتر داشتند. دهقانی و ابراهیمی [۱۷]، از ترکیب ParsBERT و تخصیص دیریکله نهفته برای مدل‌سازی موضوع‌ها و چکیده‌های مقالات استفاده کردند. کلمات کلیدی به بردارهای ۷۶۸ بعدی تبدیل شدند و با استفاده از تخصیص دیریکله پنهان دسته‌بندی شدند. این روش توانست ارتباط و همبستگی بین موضوعات را به خوبی نشان دهد و به شناسایی جهت‌های تحقیقاتی آینده کمک کند. اسنوی و همکاران [۱۸]، مدلی را با ترکیب مدل موضوعی مبتنی و MPNet ارائه دادند که به نتایج قابل توجهی در انسجام موضوعات دست یافت. این مدل از تخصیص دیریکله پنهان و سایر مدل‌های موضوعی سنتی عملکرد بهتری نشان داد و انسجام بالای ۰/۷۰۹۱ را به ثبت رساند. سامسیر و همکاران [۱۹]، با استفاده از تکنیک‌های خوشه‌بندی بدون نظارت و مدل BERTopic با مدل تعبیه MiniLM-L6-v2، ۱۳۰۲۷ چکیده از مقالات را پردازش نمودند. نتایج نشان داد که BERTopic توانایی تولید موضوعات منسجم و نمایش ارتباط بین مطالعات را دارد. کرمانی [۸]، به بررسی نحوه استفاده کاربران ایرانی توییتر از مدل‌های موضوعی برای تحلیل فریم‌های سیاسی در دوران کووید-۱۹پرداخت و عنوان نمود که تخصیص دیریکله پنهان در شناسایی فریم‌هایی که برجستگی کمتری دارند، محدودیت‌هایی دارد. لذا تفسیر انسانی در تعیین درجه همسویی خروجی‌های تخصیص دیریکله پنهان با فریم‌های شبکه ضروری است. رنجبر و همکاران [۲۰]، یک چهارچوب برای تشخیص موضوع در پست‌های تلگرامی فارسی ارائه دادند. این چهارچوب از Node2Vec برای تعبیه گراف و از

استناد به این مقاله: نحوه استناد فارسی در اینجا درج گردد.

HDBSCAN برای خوشه‌بندی پست‌ها استفاده کرد. در نهایت نتایج بهتری در دقت، بازیابی و معیار F1 نسبت به روش‌های دیگر به دست آمد. وانگ، چن و همکاران [۶]، مدل BERTopic را برای استخراج موضوعات در حوزه کتابداری و اطلاع‌رسانی (LIS) به کار بردند و نشان دادند که این مدل می‌تواند خوشه‌های متراکمی ایجاد کند و کلمات مناسبی را به عنوان نماینده برای هر موضوع انتخاب کند. وو و همکاران [۲۱]، یک بررسی جامع از مدل‌های موضوع عصبی (NTM) ارائه دادند. آن‌ها مدل‌های NTM را برای متون کوتاه و اسناد بین‌زبانی معرفی کردند و چالش‌ها و کاربردهای آن را برای آینده برجسته کردند. NTMها در این مقاله به‌عنوان ابزارهای مؤثری برای تحلیل متون پیشنهاد شده‌اند.

در نتیجه، کاوش مدل‌ها و روش‌های مختلف برای مدل‌سازی موضوعی، پیشرفت‌های بزرگ و رویکردهای متنوع در این زمینه را برجسته می‌کند. از روش سنتی کیسه کلمات گرفته تا مدل‌های پیچیده یادگیری عمیق، هر تکنیک مزایا و چالش‌های منحصر به فردی را ارائه می‌دهد. استفاده از مدل‌هایی مانند تخصیص دیریکله پنهان، تحلیل معنایی پنهان و BERTopic در مرور زمان، ماهیت در حال تکامل پردازش زبان طبیعی، اهمیت نوآوری و انطباق مداوم در پردازش زبان طبیعی و ظرفیت موجود برای رسیدگی به پیچیدگی‌های زبان فارسی را نشان می‌دهد. تحقیقات آتی باید بر اصلاح بیشتر این مدل‌ها و گسترش کاربرد آن‌ها در حوزه‌های مختلف در زبان فارسی متمرکز شوند.

۳- مبانی نظری

در این قسمت به بیان مفاهیم اولیه و تعاریف الگوریتم‌ها و روش‌های استفاده‌شده در پژوهش می‌پردازیم.

۳-۱- پیش‌پردازش داده‌ها

پیش‌پردازش متن یکی از مراحل اساسی در پردازش زبان طبیعی است که به بهبود عملکرد مدل‌های یادگیری ماشین کمک می‌کند [۲۲]. پیش‌پردازش متن، شامل چندین گام می‌باشد که عبارت‌اند از:

- توکن‌سازی[8]: در این مرحله، متن به واژه‌ها یا به‌طور دقیق‌تر به واحدهای کوچک‌تری به نام توکن شکسته می‌شود [۱۱].
- لماتیزاسیون[9]: در این مرحله، کلمات به ریشه آن‌ها تبدیل می شوند [۱۱].
- حذف علائم نگارشی: در این مرحله، علائم نگارشی حذف می‌شوند [۱۱].
- حذف کلمات توقف[10]: به‌طورکلی منظور از کلمات توقف کلماتی هستند که تعداد حضور آن‌ها در متن بسیار زیاد است و ارزش چندانی به معنای محتوا نمی‌افزایند، بنابراین در مرحله پیش‌پردازش حذف می‌شوند [۱۱].
- نرمال‌سازی[11]: نرمال‌ساز، همه متون را قبل از اعمال هر رویکردی بر روی آن‌ها به صورت استاندارد نشان می‌دهد. بنابراین، تمام علائم نگارشی، حروف، فاصله بین کلمات، اختصارات و غیره، بدون تغییر در معنای متن به شکل استاندارد خود تغییر می‌کنند [۲۳].

۳-۲- BERTopic

BERTopic، یک تکنیک مدل‌سازی موضوع است که می‌تواند از تعبیه‌های مبتنی‌بر ترنسفورمرها استفاده کند. همچنین، خوشه‌های متراکم ایجاد می‌کند که امکان تفسیر آسان موضوعات را فراهم می‌کند. در BERTopic، نخست تبدیل اسناد ورودی به نمایش‌های عددی انجام می‌شود. BERTopic، در بین مراحل مختلف استخراج موضوع تا حدی استقلال را در نظر می‌گیرد [۴]، لذا از هر مدل تعبیه‌ای می‌توان استفاده نمود. به‌منظور کاهش ابعاد بردارهای حاصل از مرحله تعبیه زبانی، در BERTopic به‌طور پیش‌فرض از UMAP استفاده می‌شود [۴]. سپس بر روی بردارهای حاصل از مرحله قبل، خوشه‌بندی انجام می‌شود. برای خوشه‌بندی، به‌طور پیش‌فرض از HDBSCAN استفاده می‌شود. اما می‌توان از مدل‌های دیگری نیز استفاده نمود. HDBSCAN توسعه‌ای از الگوریتم DBSCAN است که با معرفی رویکرد سلسله مراتبی، بهبود قابل‌توجهی در تحلیل داده‌ها ارائه می‌دهد. استفاده از روش خوشه‌بندی سلسله‌مراتبی در

[8] Tokenization
[9] lemmatization
[10] Stop words
[11] Normalization

HDBSCAN، علاوه بر انعطاف‌پذیری، توانایی تحلیل بهتر داده‌های پیچیده را فراهم می‌کند. همچنین، این الگوریتم از نظر محاسباتی کارآمد است و می‌تواند بر روی داده‌های با ابعاد بالا به‌خوبی عمل کند [۲۴]. اگرچه با HDBSCAN نمی‌توانیم به صورت اختیاری تعداد خوشه‌های مدنظر خود را مشخص کنیم، ولی BERTopic با استفاده از قابلیت کاهش ابعاد خود می‌تواند تعداد موضوعات ایجاد شده را با ادغام بردارهای مشابه‌ترین موضوعات، به تعداد مورد نظر ما کاهش دهد [۲۵].

در مرحله بعد، مدل‌سازی موضوعات بر اساس خوشه‌بندی اسناد انجام می‌شود. در BERTopic، برای به دست آوردن نمایش دقیق موضوعات از ماتریس مجموعه کلمات، TF-IDF برای کار در سطح خوشه به جای سطح سند، تنظیم می‌شود. این نمایش TF-IDF تنظیم شده، "c-TF-IDF" نامیده می‌شود و آنچه را که اسناد یک خوشه را از اسناد در خوشه دیگر متفاوت می‌کند، در نظر می‌گیرد و در نهایت موضوعات استخراج می‌شوند [۴].

۳-۳- مدل‌های تعبیه زبانی

۳-۳-۱- Hooshvare (RoBERTa)

مدل roberta-fa-zwnj-base از HooshvareLab قادر است با استفاده از کاراکتر zero-width non-joiner متن‌های فارسی را به‌طور موثرتری تجزیه و تحلیل کند. این مدل با استفاده از مجموعه داده‌های متنوع متنی و یک مجموعه واژگان جدید آموزش دیده است، که باعث بهبود عملکرد آن در پردازش زبان فارسی شده است [۲۶]. این مدل بر پایه RoBERTa [۲۷] عمل می‌کند.

۳-۳-۲- Shiraz

مدل زبانی Shiraz، بر پایهٔ معماری MobileBERT آموزش داده شده و شامل بیش از ۲۵ میلیون پارامتر است. این مدل، سرعت اجرایی بیش از ۵۰۰ درصدی، بدون از دست دادن قابل توجه دقت و کارایی نسبت به سایر مدل‌های زبانی فارسی دارد. این مدل، بر روی سه وظیفه شامل شناسایی موجودیت نامدار، تحلیل احساسات و تشخیص احساسات ارزیابی شده است. طبق [۲۸]، این مدل ۴/۳ برابر کوچک‌تر و ۵/۵ برابر سریع‌تر از BERT-base است. برای توسعه این مدل زبانی، از مجموعه داده Divan با بیش از ۱۶۴ میلیون سند و بیش از ۱۰ میلیارد توکن، استفاده شده است. این مجموعه داده، برآمده از بسترهای مختلف، مانند سایت‌های خبری، گروه‌ها و کانال‌های تلگرامی، پست‌های سایت‌های پرطرفدار ورزشی، حقوقی، تاریخی، تکنولوژی و... است [۲۹].

۳-۳-۳- LaBSE

یک مدل از خانواده Sentence-Transformers است. LaBSE، BERT چندزبانه برای ایجاد تعبیه جملات، برای ۱۰۹ زبان است. پیشرفته‌ترین کار برای بسیاری از وظایف پردازش زبان طبیعی تک زبانه و چندزبانه، پیش آموزش مدل زبانی نقاب‌دار و به دنبال آن تنظیم دقیق وظایف خاص است. این مدل، پیش‌آموزش برای مدل زبانی نقاب‌دار و مدل زبان ترجمه[۱۲] را با کار رتبه‌بندی ترجمه با استفاده از رمزگذارهای دوطرفه ترکیب می‌کند. تعبیه‌های چند زبانه به دست آمده، میانگین دقت ۸۳.۷ درصد را در آزمون بازیابی متن دو زبانه به دست آورده است [۳۰].

۳-۳-۴- Paraphrase-multilingual-MiniLM-L12-v2

یک مدل از خانواده Sentence-Transformers است که جملات و پاراگراف‌ها را به یک فضای برداری متراکم با ابعاد ۳۸۴ تبدیل می‌کند. این مدل، برای وظایفی مانند خوشه‌بندی یا جستجوی معنایی استفاده می‌شود. این مدل، از معماری MiniLM استفاده می‌کند که یک نسخه کوچک‌تر و بهینه‌تر از مدل‌های بزرگ‌تر مانند BERT است. شامل ۱۲ لایه ترانسفورمر است و از روش mean pooling برای ایجاد تعبیه‌های جملات استفاده می‌کند. همچنین، می‌تواند جملات را در بیش از ۵۰ زبان مختلف پردازش کند [۳۱].

۳-۳-۵- XLM-R

یک مدل از خانواده Sentence-Transformers است که جملات و پاراگراف‌ها را به یک فضای برداری متراکم ۷۶۸ بعدی نگاشت می‌کند و می‌تواند برای کارهایی مانند خوشه‌بندی یا جستجوی معنایی استفاده شود. این مدل مبتنی‌بر مدل RoBERTa فیس‌بوک است که در سال ۲۰۱۹ منتشر شد [۳۲]. این یک مدل زبان ماسک‌شده مبتنی‌بر Transformer است که با استفاده از بیش از دو ترابایت داده CommonCrawl فیلتر شده، روی صد زبان آموزش داده شده است. این مدل، به طور قابل توجهی بهتر از BERT چندزبانه (mBERT) در انواع معیارهای چندزبانی، از جمله دقت است. XLM-R به‌ویژه در

[12] Translation Language Model (TLM)

زبان‌های کم‌منبع خوب عمل می‌کند. XLM-R با مدل‌های تک‌زبانه قوی در معیارهای GLUE و XNLI بسیار رقابتی است [۳۳].

ParsBERT -۶-۳-۳

الگوریتم ParsBERT یک مدل تک‌زبانه برای برداری کردن متون فارسی است که مفاهیم و معنای جملات فارسی را در نظر می‌گیرد [۱۷]. این مدل بر اساس معماری BERT-base طراحی شده است [۳۴] و این توانایی را دارد که تفاوت‌های ظریف و پیچیدگی‌های زبان فارسی را بشناسد [۱۳]. ParsBERT در مقایسه با روش‌های تعبیه مختلف مانند word2vec، fastText، ELMo و LASER نیز به دقت بالاتری دست یافته است [۱۲]. این مدل روی مجموعه‌های بزرگ فارسی با سبک‌های نوشتاری مختلف از موضوعات متعدد (مانند علمی، رمان، اخبار) با بیش از ۳.۹ میلیون سند، ۷۳ میلیون جمله و ۱.۳ میلیارد کلمه از قبل آموزش داده شده است [۱۳]. فراهانی از مجموعه داده FarsTail برای آموزش یک مدل SentenceTransformers (با استفاده از ParsBERT)، به‌عنوان پایه‌ای برای برنامه‌های کاربردی دیگر مانند جستجوی معنایی، خوشه‌بندی، استخراج اطلاعات، خلاصه سازی، مدل‌سازی موضوع و برخی موارد دیگر، استفاده نموده است [۳۵].

۳-۴- الگوریتم K-Means

الگوریتم K-Means یکی از ساده‌ترین و پرکاربردترین روش‌ها در خوشه‌بندی داده‌ها به خصوص در حوزه یادگیری بدون نظارت است. روند عملکرد K-Means با تعیین کردن تعداد خوشه‌ها (k) آغاز می‌شود و سپس بر اساس معیار فاصله، نقاط داده به نزدیک‌ترین مرکز خوشه‌ها اختصاص می‌یابند. در مرحله بعد، مراکز خوشه‌ها با محاسبه میانگین مختصات نقاط هر خوشه به‌روز می‌شوند و این روند تا زمانی که تغییرات محسوسی در مراکز خوشه‌ها ایجاد نشود، ادامه می‌یابد. این الگوریتم به دلیل سادگی و سرعت بالا در اجرا، برای مجموعه داده‌های حجیم مناسب است [۳۶].

۳-۵- تخصیص دیریکله پنهان

این روش فرض می‌کند که هر سند یک ترکیب از موضوعات مختلف است و کلمات موجود در هر سند، احتمال تعلق به این موضوعات را دارند [۳۷]. تخصیص دیریکله پنهان با استفاده از الگوریتم‌هایی مانند نمونه‌گیری گیبز، به تدریج ساختار موضوعی را بهبود می‌دهد، اگرچه کشف تعداد بهینه موضوعات همچنان یک چالش است. فرآیند یادگیری به‌طور مداوم تخصیص موضوعات را به‌روزرسانی می‌کند تا سندها به ترکیبی بهینه از موضوعات تخصیص یابند [۲].

۳-۶- تجزیه و تحلیل معنایی پنهان

این روش بر شناسایی مفاهیم نهفته در داده‌های متنی متمرکز است [۱] و از فاکتورگیری ماتریس برای کاهش ابعاد و شناسایی ساختارهای پنهان استفاده می‌کند. برای عملکرد بهتر، تحلیل معنایی پنهان به مجموعه‌ای بزرگ از اسناد نیاز دارد که اصطلاحات مرتبط در آن‌ها در چهارچوب‌های مختلف به کار رفته باشد [۳۷].

۳-۷- فاکتورسازی ماتریس غیرمنفی

فاکتورسازی ماتریس غیرمنفی یک الگوریتم غیر احتمالی است [۳۸] که برای تجزیه ماتریس داده‌ها به دو ماتریس با رتبه پایین‌تر استفاده می‌شود. این روش، بر اساس مفهوم فاکتورسازی عمل کرده و معمولاً با داده‌های حاصل از اعمال TF-IDF بر روی متون، برای استخراج موضوعات استفاده می‌شود [۳].

۴- روش پیشنهادی

این مقاله با الهام از [۴] روشی بدون نظارت برای استخراج موضوع از متون فارسی پیشنهاد می‌کند. این روش شامل مراحل پیش‌پردازش متن، تعبیه‌سازی متن، کاهش ابعاد بردارهای تعبیه، خوشه‌بندی، و در نهایت بازنمایی موضوعات می‌باشد که در ادامه بررسی می‌شود. چهارچوب کلی پیشنهادی در شکل (۱) ارائه شده است.

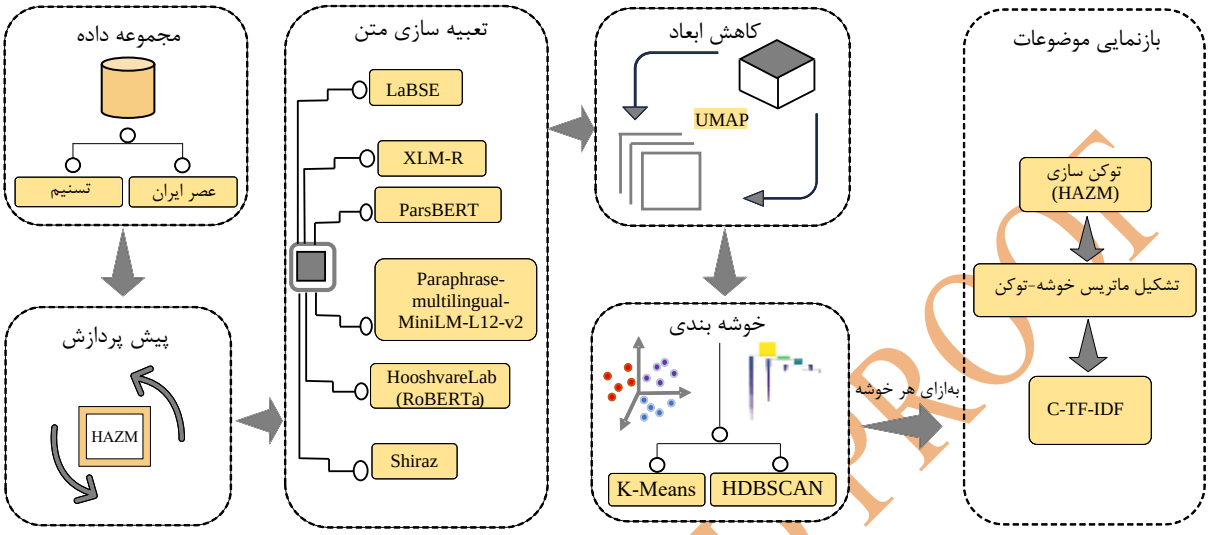
۴-۱- پیش‌پردازش داده‌ها

در ابتدا پیش‌پردازش متون صورت می‌گیرد. مراحل پیش‌پردازش شامل حذف سطرهای خالی، حذف سطرهای تکراری، حذف نشانی‌های اینترنتی[۱۳]، حذف علائم نگارشی، حذف کلمات توقف، لماتیزاسیون، حذف اعداد و کاراکترهای انگلیسی و نرمالسازی متن می‌باشد. در این مقاله، ما از ابزار HAZM برای پیش‌پردازش متن استفاده می کنیم که برای پردازش متون فارسی، به‌ویژه برای متون رسمی، بسیار مورد استفاده قرار می‌گیرد [۲۲].

۴-۲- تعبیه سازی متن

[13] URL

پس از پیش‌پردازش، از مدل تعبیه جمله LaBSE برای تعبیه‌سازی متون، استفاده می‌نماییم. سپس مقایسه ای بین مدل تعبیه پیشنهادی و مدل‌های تعبیه جمله مبتنی‌بر ترنسفورمرها مانند XLM-R، ParsBERT، Paraphrase-multilingual-MiniLM-L12-v2 و همچنین مدل های تعبیه کلمه مبتنی‌بر ترنسفورمرها مانند(RoBERTa) HooshvareLab و Shiraz انجام می‌دهیم.



شکل ۱- چهارچوب کلی روش پیشنهادی

۴-۳- کاهش ابعاد

از آنجایی که بردارهای تعبیه اغلب ابعاد بالایی دارند، خوشه‌بندی دشوار می‌شود. یک راه حل این است که ابعاد بردارهای تعبیه را به یک فضا با ابعاد قابل اجرا کاهش دهیم [۴]. به این منظور، در این مقاله از الگوریتم UMAP استفاده می‌نماییم. این الگوریتم، با سرعت بالایی اجرا می‌شود و می‌تواند هم جزئیات محلی و هم ساختار کلی داده‌ها را در ابعاد پایین‌تر حفظ کند. لذا، باعث می‌شود ویژگی‌های اصلی داده‌ها بهتر نمایان شوند. علاوه‌براین، چون UMAP هیچ محدودیت محاسباتی در مورد ابعاد تعبیه ندارد، می‌توان از آن در مدل‌های زبانی با بردارهایی با ابعاد متفاوت استفاده کرد [۵].

۴-۴- خوشه بندی

پس از کاهش ابعاد بردارهای تعبیه، آن‌ها را در گروه‌هایی از بردارهای تعبیه مشابه دسته‌بندی می‌کنیم. هرچه تکنیک خوشه‌بندی کارآمدتر باشد، نمایش موضوع دقیق‌تر است. به منظور شناسایی الگوریتم خوشه‌بندی مناسب، از الگوریتم‌های HDBSCAN و K-Means، به‌طور جداگانه استفاده می‌کنیم.

الگوریتم HDBSCAN، به‌طور خودکار خوشه‌هایی با چگالی‌های مختلف را شناسایی کرده و نقاط پرت را حذف می‌کند. این الگوریتم برای شناسایی ساختارهای پیچیده داده‌ها طراحی شده است. الگوریتم K-Means بر اساس تقسیم مجموعه داده به چند خوشه متمایز و بدون همپوشانی عمل می‌کند، لذا هر نقطه را مجبور می‌کند در یک خوشه قرار گیرد. بنابراین، بر خلاف HDBSCAN، هیچ نقطه پرتی ایجاد نخواهد شد. عدم ایجاد نقاط پرت معایبی دارد، زیرا وقتی هر نقطه مجبور به قرار گرفتن در یک خوشه می‌شود، به این معنی است که خوشه به احتمال زیاد دارای نویز است، که می‌تواند به نمایش موضوع آسیب برساند [۴]. الگوریتم K-Means به دلیل سادگی و سرعت بالا در اجرا، برای مجموعه داده‌های حجیم مناسب است [۳۶].

۴-۵- بازنمایی موضوعات

پس از مرحله خوشه‌بندی، تحلیل متن باید به صورت جداگانه در سطح هر خوشه انجام شود. لذا با استفاده از CountVectorizer، پس از انجام توکن‌سازی مختص زبان فارسی، ماتریس خوشه-توکن ساخته می‌شود.

برای به‌دست آوردن نمایش دقیق موضوعات از ماتریس خوشه-توکن، c-TF-IDF برای کار در سطح خوشه تنظیم می‌شود تا امتیاز اهمیت هر کلمه در هر خوشه را به‌دست آوریم و یک موضوع به هر خوشه اختصاص داده شود [۴].

محاسبه c-TF-IDF برای کلمه x در خوشه c، مطابق زیر است:

$$W_{x,c} = || tf_{x,c} || \times log\left(1+\frac{A}{f_x}\right) \quad (۱)$$

که $tf_{x,c}$ فراوانی کلمه x در خوشه c، f_x فراوانی کلمه x در تمام خوشه ها، و A میانگین تعداد کلمات در خوشه مربوطه است. در انتها به ازای هر خوشه، کلمات با بیشترین وزن در ماتریس خوشه-توکن به‌عنوان موضوعات مهم آن خوشه انتخاب می‌شوند.

۵- آزمایش‌ها، نتایج و ارزیابی

۵-۱- مجموعه داده

در این مقاله، از مجموعه داده‌های تسنیم با ۶۳۴۹۳ سطر [۳۹] و عصر ایران با ۱۳۲۵۸ سطر [۴۰] استفاده می‌کنیم. محتوای هر دو مجموعه داده، اخبار می‌باشد. در این دو مجموعه داده، موضوع مربوط به هر خبر با برچسب مربوطه از قبل مشخص شده است. اگرچه ما در هیچ کدام از مدل‌های مورد استفاده برای استخراج موضوع، از این برچسب‌ها استفاده نمی‌کنیم، ولی وجود آن‌ها برای تعیین تعداد موضوعات مورد انتظار از مدل‌ها و نیز ارزیابی عملکرد مدل‌ها بر اساس قضاوت انسانی، برای ما مفید است.

مطابق برچسب های موضوعی درمی‌یابیم مجموعه داده عصر ایران، دارای هشت موضوع کلی اجتماعی، اقتصادی، بین المللی، سیاسی، علوم و فناوری، هنری و فرهنگی، ورزشی و پزشکی می‌باشد. مجموعه داده تسنیم، دارای هشت دسته‌ی اقتصادی، اجتماعی، بین‌الملل، فرهنگی هنری، ورزش، رسانه، استان‌ها و دسته سیاسی می‌باشد.

۵-۲- تنظیمات پیاده سازی

پیاده سازی روش پیشنهادی در محیط آنلاین Google Colab با پردازنده‌های گرافیکی نوع T4 و Python3.8 انجام شده است. تنظیمات پارامترهای مربوط به مدل‌ها و الگوریتم‌های استفاده شده در پژوهش به صورت خلاصه در جدول ۱ ارائه شده است

۵-۳- معیارهای ارزیابی

ارزیابی مدل‌های موضوعی به دلیل ماهیت بدون نظارت آن‌ها و نبود معیارهای استاندارد، چالش‌برانگیز است. در این مقاله از سه معیار ارزیابی استفاده شده است:

- **معیار ارزیابی انسجام:** معیار انسجام موضوعی با استفاده از اطلاعات متقابل نقطه‌ای نرمال‌شده[14] [۴۱] محاسبه شد که محدوده آن بین ۱- تا ۱ بود. مقادیر بالاتر نشان‌دهنده ارتباط بیشتر کلمات است [۳]. این معیار با کتابخانه Gensim پیاده‌سازی شده است [۱۴].
- **معیار تنوع**[15]: این معیار، درصد کلمات غیرتکراری در میان کلمات برتر موضوعات را نشان می‌دهد [۴۲]. این معیار دارای محدوده [۰, ۱] است و توسط کتابخانه OCTIS[16] محاسبه می‌شود. مقادیر بالاتر، نشان‌دهنده تنوع بیشتر موضوعات است [۴].
- **معیار انسانی:** با توجه به عدم وجود معیارهای دقیق و مناسب برای ارزیابی روش‌های استخراج موضوع، در این مقاله ما معیاری بر اساس قضاوت انسانی برای ارزیابی این روش‌ها پیشنهاد می کنیم. در معیار قضاوت انسانی، بررسی می‌کنیم که از ۸ موضوع استخراج‌شده، چند موضوع به یک حوزه واحد تعلق داشته و با موضوعات مورد انتظار ما از مجموعه داده تطابق دارد. امتیاز این معیار ارزیابی، از عدد ۸ محاسبه می‌شود.

۵-۴- آزمایش روش پیشنهادی

این آزمایش را به دو روش مختلف انجام می دهیم. در هر بار اجرای روش پیشنهادی با استفاده از مدل‌های ثابت و روی مجموعه داده ثابت، اندکی تفاوت در نتایج موضوعات قابل مشاهده است. لذا برای ارزیابی هر مدل تعبیه، پنج بار مدل پیشنهادی را به طور کامل اجرا می‌کنیم و در نهایت میانگین مقادیر حاصل از معیارهای ارزیابی را برای این پنج اجرا محاسبه و بررسی می‌نماییم.

۵-۴-۱- روش اول: عملکرد روش پیشنهادی با الگوریتم HDBSCAN

در این بخش، آزمایشی به وسیله‌ی مجموعه داده عصر ایران انجام می‌دهیم. در این روش از مدل‌های تعبیه‌سازی جمله مبتنی‌بر ترنسفورمرها مانند ParsBERT، XLM-R، LaBSE و Paraphrase-multilingual-MiniLM-L12-v2، و همچنین مدل‌های تعبیه کلمه مبتنی‌بر ترنسفورمرها مانند Hooshvare (RoBERTa) و Shiraz، به طور جداگانه استفاده می‌نماییم. در این آزمایش، از

[14] Normalized pointwise mutual information (NPMI)
[15] Diversity
[16] Optimizing and Comparing Topic models Is Simple

الگوریتم HDBSCAN برای خوشه‌بندی استفاده می‌کنیم. HDBSCAN قادر است تعداد خوشه‌ها را به صورت خودکار مشخص نموده و نقاط پرت را نیز شناسایی کند. از آنجایی که موضوعات مجموعه داده مطابق برچسب‌های اولیه در ۸ دسته قرار دارند، ما نیز تعداد موضوعات را ۸ عدد تنظیم می‌کنیم. البته ما از این برچسب‌ها در اجرای مدل‌ها استفاده نمی‌کنیم، بلکه فقط برای تعیین تعداد موضوعات در ابتدا و نیز برای ارزیابی نتایج مطابق قضاوت انسانی، در استنباط ذهنی از آن‌ها کمک می‌گیریم. در نهایت، مقادیر حاصل را به وسیله‌ی معیار انسجام و معیار انسانی ارزیابی می‌نماییم.

۵-۴-۲- روش دوم: عملکرد روش پیشنهادی با الگوریتم K-Means

آزمایشی مشابه آزمایش پیش انجام می‌دهیم، با این تفاوت که برای خوشه‌بندی، از الگوریتم K-Means استفاده می‌کنیم. این الگوریتم هیچ نقطه پرتی ایجاد نمی‌کند. به وسیله آن قادر هستیم تعداد خوشه‌ها را به اختیار ۸ عدد انتخاب کنیم. در نهایت، موضوعات استخراج شده را مطابق معیار انسجام و نیز معیار انسانی ارزیابی می‌نماییم.

نتایج اجرای این دو آزمایش از روش پیشنهادی، در جدول ۲ نشان داده شده است.

۵-۵- مقایسه روش پیشنهادی با مدل‌های فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان و تحلیل معنایی پنهان

۵-۵-۱- آزمایش اول

در این آزمایش، عملکرد روش پیشنهادی و سه مدل فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان و تحلیل معنایی پنهان، برای استخراج ۸ موضوع ارزیابی شد. روش پیشنهادی، با مدل‌های تعبیه LaBSE و ParsBERT (به‌طور جداگانه)، به همراه الگوریتم خوشه‌بندی K-Means ارزیابی شد. در این آزمایش، از مجموعه داده عصر ایران استفاده نمودیم. هر آزمایش سه بار تکرار شد و میانگین نتایج به‌عنوان معیار نهایی گزارش شد. ارزیابی نتایج با استفاده از معیار انسجام موضوعی و معیار انسانی انجام گرفت. نتایج حاصل از این آزمایش در جدول ۳ ارائه شده است. مقادیر حاصل از جدول ۳ را با استفاده از شکل (۵) نمایان ساختیم. مطابق شکل (۵) در حالت کلی، بهترین نتایج از روش پیشنهادی (با مدل تعبیه LaBSE) و مدل فاکتورسازی ماتریس غیرمنفی به دست می‌آید. ضعیف‌ترین عملکرد نیز مربوط به مدل تخصیص دیریکله پنهان می‌باشد.

۵-۵-۲- آزمایش دوم

این آزمایش شامل بررسی عملکرد روش پیشنهادی و مدل‌های فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان و تحلیل معنایی پنهان در استخراج موضوعاتی با تعداد مختلف (۸، ۱۵، ۲۵، ۵۰، ۷۵، ۱۰۰ و ۱۲۵) بود. در روش پیشنهادی، از مدل‌های تعبیه LaBSE (بهترین مدل تعبیه طبق جدول ۲) و Shiraz (سریع‌ترین مدل)، به‌همراه الگوریتم خوشه‌بندی K-Means استفاده شد. در این آزمایش از مجموعه داده تسنیم استفاده نمودیم. هر آزمایش سه بار تکرار شد و نتایج نهایی میانگین‌گیری شدند. ارزیابی نتایج این آزمایش بر اساس معیار انسجام و معیار تنوع صورت گرفت. نتایج حاصل در جدول ۴ و ۵ ارائه شده است. همچنین مقادیر حاصل از این دو جدول را با استفاده از شکل‌های (۶) و (۷) نمایان ساختیم.

۵-۶- نتایج آزمایش‌ها

نتایج در ابتدا نشان داد که انجام پیش‌پردازش بر روی متون، تأثیر مثبتی بر موضوعات حاصل دارد. برای مثال، حفظ علائم نگارشی در داده‌های ورودی، تأثیر منفی بر عملکرد روش پیشنهادی داشته و منجر به کاهش امتیاز نتایج، مطابق معیار انسجام برای مدل‌های تعبیه مختلف شد.

نتایج عملکرد مرحله اول و دوم آزمایش روش پیشنهادی با مدل‌های تعبیه مختلف و الگوریتم‌های خوشه‌بندی HDBSCAN و K-Means، به صورت خلاصه در جدول ۲ نشان داده شده است.

جدول ۱- تنظیمات پارامترهای مهم مدل‌ها و الگوریتم‌های استفاده شده در پژوهش

مدل / الگوریتم	کتابخانه / ابزار	پارامتر	مقدار	توضیح / دلیل انتخاب
HDBSCAN	hdbscan	min_cluster_size	۸	حداقل اندازه خوشه برای تطابق با موضوعات مجموعه داده.
		metric	فاصله اقلیدسی	- محاسبات سریع و سازگار با پیاده‌سازی پیش فرض HDBSCAN - عملکرد مناسب برای بردارهای کاهش بُعد یافته پس از تعبیه‌سازی - نتایج تجربی مناسب.
		min_samples	۵	حداقل تعداد نقاط برای شناسایی یک نقطه به‌عنوان هسته متراکم و تعیین حساسیت به نویز است.
K-Means	scikit-learn	n_clusters	متغیر	تعداد موضوعات
		metric	فاصله اقلیدسی	- محاسبات سریع و سازگار با پیاده‌سازی پیش فرض K-Means - عملکرد مناسب برای بردارهای کاهش بُعد یافته پس از تعبیه‌سازی - نتایج تجربی مناسب.
UMAP	UMAP	n_neighbors	۱۵	حفظ تعادل بین ساختار محلی و کلی تعبیه‌ها.
		n_components	۵	کاهش ابعاد با حفظ اطلاعات برای خوشه‌بندی مؤثر.
		metric	فاصله کسینوسی	اندازه‌گیری شباهت معنایی بین ویژگی‌ها در متن.
		min_dist	۰/۰	حفظ ساختار محلی داده‌ها در فضای تعبیه شده.
CountVectorizer	scikit-learn	ngram_range	۱	در نظر گرفتن کلمات منفرد به‌عنوان موضوع.
		min_df	۲	حذف کلماتی با تکرار کم برای کاهش نویز.
	HAZM	tokenizer	–	سازگار با زبان فارسی و متن رسمی.
LDA	gensim	num_topics	متغیر	تعداد موضوعات
LSA	scikit-learn	max_df	۰/۹۵	حذف کلمات بسیار رایج برای کاهش نویز.
		min_df	۵	حذف کلمات با تکرار کم برای کاهش پیچیدگی.
		n_iter	۱۰۰	اطمینان از همگرایی مطلوب در تجزیه مقادیر تکین.
NMF	scikit-learn	max_df	۰/۹۵	حذف کلمات بسیار رایج برای کاهش نویز.
		min_df	۲	حذف کلمات با تکرار کم برای حفظ کیفیت موضوعات.
		Random State	۱	بازتولیدپذیری نتایج آزمایش‌ها.

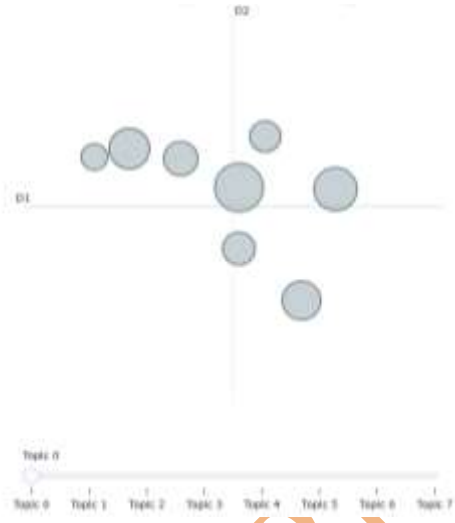
جدول ۲- عملکرد مدل‌های تعبیه بر روی مجموعه داده عصر ایران برای دست‌یابی به ۸ موضوع مطابق معیار ارزیابی انسجام و معیار انسانی، مقادیر جدول برای هر مدل تعبیه میانگین پنج اجرا می‌باشد.

الگوریتم	معیار ارزیابی	LaBSE	Shiraz	ParsBERT	XLM-R	paraphrase-multilingual-MiniLM-L12-v2	Hooshvare (RoBERTa)
HDBSCAN	انسجام	۰/۰۹	۰/۱۰	۰/۱۱	۰/۰۶۵	۰/۱۱	**۰/۱۲۴**
	معیار انسانی	۵/۲	۴/۸	**۵/۴**	۳/۹	۴/۷	۴/۲
K-Means	انسجام	۰/۱۱	۰/۱۱۸	۰/۰۹۲	۰/۱۰۴	**۰/۱۳۶**	۰/۱۰۲
	معیار انسانی	**۸**	۷	۷/۵	۷	۷/۲	۷/۵

نتایج جدول ۲، نشان‌دهنده‌ی اختلاف اندکی در مقادیر معیار انسجام برای مدل‌های تعبیه مختلف است. با استفاده از الگوریتم HDBSCAN، مطابق معیار انسجام مدل تعبیه Hooshvare(RoBERTa) و مطابق معیار انسانی، مدل ParsBERT به بهترین نتایج دست می‌یابد. همچنین با الگوریتم K-Means، مطابق معیار انسجام مدل paraphrase-multilingual-MiniLM-L12-v2 و مطابق معیار انسانی مدل LaBSE به بهترین نتایج دست می‌یابد.

۵-۷- تجسم موضوعات

در شکل (۲) تا (۴)، سه نمایش از عملکرد مدل پیشنهادی با مدل تعبیه‌سازی LaBSE را به همراه الگوریتم خوشه‌بندی K-Means برای استخراج ۸ موضوع کلی از مجموعه داده عصر ایران ارائه می‌دهیم.



شکل ۲- نمودار دو بعدی نقشه فاصله موضوعات در روش پیشنهادی با مدل تعبیه‌سازی LaBSE و الگوریتم خوشه‌بندی K-Means

	Topic	Count	Name	Representation	Representative_Docs
0	0	822	کاربر_موبایل_اپل_دوربین_0	[...کاربر, موبایل, اپل, دوربین, نمایشگر, استفاده]	[...نازگی دم گوشیگوش جنبدی سری اکسپریا لیست سای]
1	1	1379	دانشگاه_علمی_پزشکی_آموزش_1	[...دانشگاه, علمی, پزشکی, آموزش, کشور, بهداشت, حل]	[...گزارش خبرنگار گروه علمی دانشگاهی خبرگزاری فار]
2	2	1129	تیم_بازی_بازیکن_فوتبال_2	[...تیم, بازی, بازیکن, فوتبال, ملی, باشگاه, لیگ]	[...گزارش خبرنگار توپ نور گروه ورزش باشگاه خبرنگا]
3	3	2783	خبرنگار_تهران_کشور_اسلامی_3	[...خبرنگار, تهران, کشور, اسلامی, مردم, اجتماعی]	[...گزارش خبرنگار اجتماعی باشگاه خبرنگار اخمد مسج]
4	4	1759	فیلم_سینما_جشنواره_نمایش_4	[...فیلم, سینما, جشنواره, نمایش, خبرنگار, برنامه]	[... گزارش باشگاه خبرنگار نقل روابط عمومی سی‌امین]
5	5	1939	بیماری_استفاده_درمان_مصرف_5	[...بیماری, استفاده, درمان, مصرف, بدن, قرار, خون]	[...گزارش خبرنگار علمی باشگاه خبرنگار دانشمندان م]
6	6	2197	آمریکا_کشور_ایران_سوریه_6	[...آمریکا, کشور, ایران, سوریه, گزارش, گروه, تروز]	[...گزارش خبرنگار سیاست خارجی گروه سیاسی باشگاه خ]
7	7	1250	هزار_قیمت_بازار_کشور_7	[...هزار, قیمت, بازار, کشور, اقتصادی, تومان, افزا]	[... اظهار نکته مصوبات امروز دولت توجه قانون دولت]

شکل ۳- نمایش توزیع اسناد و موضوعات حاصل از روش پیشنهادی، با مدل تعبیه‌سازی LaBSE و الگوریتم خوشه‌بندی K-Means

شکل ۴- امتیازات c-TF-IDF برای کلمات برتر موضوعات حاصل از روش پیشنهادی، با مدل تعبیه‌سازی LaBSE و الگوریتم خوشه‌بندی K-Means

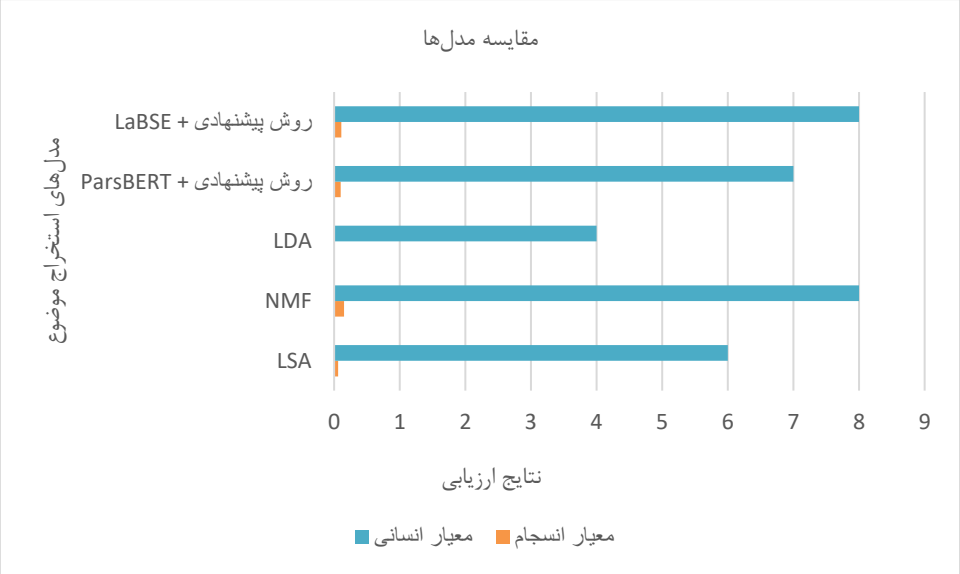
۵-۸- مقایسه عملکرد الگوریتم‌های HDBSCAN و K-Means در روش پیشنهادی

مطابق جدول ۲، الگوریتم‌های خوشه‌بندی HDBSCAN و K-Means رفتار متفاوتی را از خود نشان می‌دهند. در زیر به بیان این تفاوت‌ها می‌پردازیم:

- **تعداد موضوعات**: K-Means به هر سند حتما موضوعی را اختصاص می‌دهد. به وسیله‌ی K-Means به هر ۸ موضوع مورد انتظار از مجموعه داده، می‌توانیم دست یابیم، ولی HDBSCAN معمولاً به تعداد قابل توجهی از اسناد، موضوعی اختصاص نمی‌دهد. زیرا تعدادی نقطه پرت ایجاد می‌شود و در نتیجه ۷ موضوع حاصل می‌شود.
- **پایداری نتایج**: نتایج نشان می‌دهد اگر روش پیشنهادی را با الگوریتم HDBSCAN روی تعبیه‌های مشخصی چندین بار اجرا کنیم، در هر بار اجرا نتایج کمی متفاوت است. اما در نتایج حاصل از عملکرد الگوریتم K-Means تفاوت کمتری مشاهده می‌شود و اغلب نتایج حاصل از چندین بار اجرای روش پیشنهادی، مقدار نسبتاً ثابتی است.
- **نتایج معیارهای ارزیابی**: مطابق معیار انسانی، در K-Means نسبت به HDBSCAN به نتایج مطلوب‌تری دست می‌یابیم. همچنین میانگین تمام مقادیر به دست آمده از معیار انسجام، در جدول ۲ برای HDBSCAN و K-Means به ترتیب ۰/۱۰ و ۰/۱۱ می‌باشد، که اگرچه اختلاف بسیار اندک است، اما نشان می‌دهد K-Means طبق معیار انسجام نیز بهتر عمل نموده است

جدول ۳- ارزیابی استخراج ۸ موضوع از مجموعه داده عصر ایران، مقادیر جدول میانگین سه مرتبه اجرا هستند.

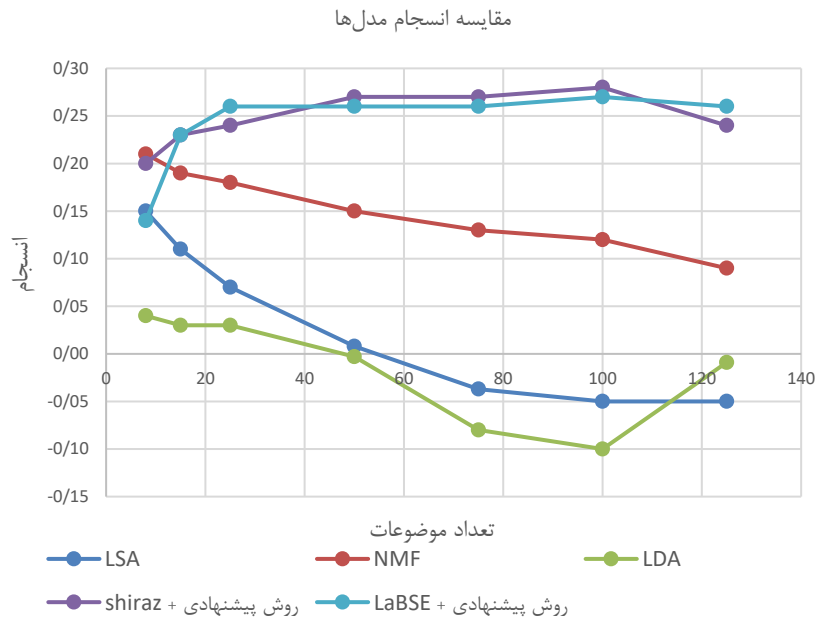
معیار ارزیابی	LSA	NMF	LDA	+ روش پیشنهادی ParsBERT	+ روش پیشنهادی LaBSE
انسجام	۰/۰۶	**۰/۱۵**	۰/۰۱	۰/۱	۰/۱۱
معیار انسانی	۶	**۸**	۴	۷	**۸**



شکل ۵- مقایسه نتایج حاصل از ارزیابی مدل‌های فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان، تحلیل معنایی پنهان و روش پیشنهادی با مدل‌های تعبیه ParsBERT و LaBSE در ۸ موضوع

جدول ۴- ارزیابی استخراج موضوعات از مجموعه داده تسنیم به‌وسیله معیار انسجام. مقادیر جدول، میانگین سه مرتبه اجرا هستند.

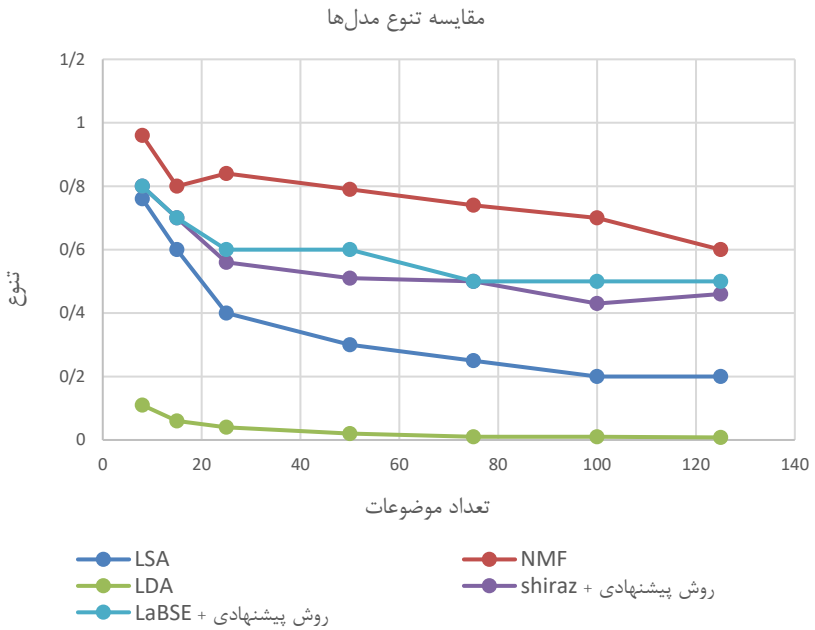
تعداد موضوع	LSA	NMF	LDA	روش پیشنهادی + Shiraz	روش پیشنهادی + LaBSE
۸	۰/۱۵	**۰/۲۱**	۰/۰۴	۰/۲	۰/۱۴
۱۵	۰/۱۱	۰/۱۹	۰/۰۳	**۰/۲۳**	**۰/۲۳**
۲۵	۰/۰۷	۰/۱۸	۰/۰۳	۰/۲۴	**۰/۲۶**
۵۰	۰/۰۰۸	۰/۱۵	-۰/۰۰۳	**۰/۲۷**	۰/۲۶
۷۵	-۰/۰۳۷	۰/۱۳	-۰/۰۸	**۰/۲۷**	۰/۲۶
۱۰۰	-۰/۰۵	۰/۱۲	-۰/۱	**۰/۲۸**	۰/۲۷
۱۲۵	-۰/۰۵	۰/۰۹	-۰/۰۰۹	۰/۲۴	**۰/۲۶**



شکل ۶- مقایسه موضوعات نتایج حاصل از مدل‌های فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان، تحلیل معنایی پنهان و روش پیشنهادی با مدل‌های تعبیه Shiraz و LaBSE به‌وسیله ارزیابی با معیار انسجام

جدول ۵- ارزیابی استخراج موضوعات از مجموعه داده تسنیم به وسیله معیار ارزیابی تنوع. مقادیر جدول، میانگین سه مرتبه اجرا هستند.

تعداد موضوع	LSA	NMF	LDA	روش پیشنهادی + Shiraz	روش پیشنهادی + LaBSE
۸	۰/۷۶	**۰/۹۶**	۰/۱۱	۰/۸	۰/۸
۱۵	۰/۶	**۰/۸**	۰/۰۶	۰/۷	۰/۷
۲۵	۰/۴	**۰/۸۴**	۰/۰۴	۰/۵۶	۰/۶
۵۰	۰/۳	**۰/۷۹**	۰/۰۲	۰/۵۱	۰/۶
۷۵	۰/۲۵	**۰/۷۴**	۰/۰۱	۰/۵	۰/۵
۱۰۰	۰/۲	**۰/۷**	۰/۰۱	۰/۴۳	۰/۵
۱۲۵	۰/۲	**۰/۶**	۰/۰۰۸	۰/۴۳	۰/۵



شکل ۷- مقایسه موضوعات حاصل از مدل‌های فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان، تحلیل معنایی پنهان و روش پیشنهادی با مدل‌های تعبیه Shiraz و LaBSE به‌وسیله ارزیابی با معیار تنوع

همانطور که شکل (۶) نشان می‌دهد، به طور کلی مطابق معیار انسجام، روش پیشنهادی (با هر دو مدل تعبیه Shiraz و LaBSE) عملکرد بهتری نسبت به سایر مدل‌های استخراج موضوع دارد و بیشترین مقدار انسجام را در تعداد ۱۰۰ موضوع به دست می‌آورد. مدل فاکتورسازی ماتریس غیرمنفی، در ابتدا برای تعداد موضوعات کم (۸ موضوع)، عملکرد بهتری نشان می‌دهد و به بیشترین مقدار انسجام در بین مدل‌ها دست می‌یابد، اما با افزایش تعداد موضوعات، کارایی آن کاهش می‌یابد. همچنین موضوعات حاصل از مدل تحلیل معنایی پنهان، در ابتدا به انسجام نسبتاً خوبی دست می‌یابند، اما میزان انسجام با افزایش تعداد موضوعات به شدت کاهش می‌یابد. مطابق این معیار، ضعیف‌ترین عملکرد نیز به مدل تخصیص دیریکله پنهان اختصاص می‌یابد.

مطابق معیار تنوع، مدل فاکتورسازی ماتریس غیرمنفی و سپس روش پیشنهادی با مدل تعبیه LaBSE، متنوع‌ترین موضوعات را استخراج می‌کنند. با این حال، با افزایش تعداد موضوعات، میزان تنوع در همه مدل‌ها کاهش می‌یابد. بیشترین تکرار در موضوعات حاصل از مدل تخصیص دیریکله پنهان مشاهده می‌شود.

۶- بحث

ارزیابی ها با معیار انسجام و معیار انسانی، تأیید کردند که روش پیشنهادی این مقاله در استخراج موضوع از متون فارسی عملکرد قابل توجهی دارد. همچنین با استفاده از الگوریتم HDBSCAN، به ترتیب مطابق معیار انسجام، مدل تعبیه Hooshvare(RoBERTa) و مطابق معیار انسانی، مدل ParsBERT به بهترین نتایج دست می‌یابد. همچنین، با الگوریتم K-Means، مطابق معیار انسجام، مدل paraphrase-multilingual-MiniLM-L12-v2 و مطابق معیار انسانی مدل LaBSE به بهترین نتایج دست می‌یابد. نتایج همچنین، برتری K-Means را نسبت به HDBSCAN در هر دو معیار ارزیابی نشان داد. زیرا مطابق معیار انسانی، با استفاده از K-Means می‌توان با توجه به موضوعات حاصل، به هدف خود یعنی موضوعات مورد انتظار از مجموعه داده رسید. مطابق معیار انسجام نیز، میانگین تمام مقادیر به دست آمده در جدول ۱ برای HDBSCAN عدد ۰/۱۰ و برای K-Means عدد ۰/۱۱ می‌باشد که نشان‌دهنده برتری عملکرد K-Means است.

برای مقایسه نتایج روش پیشنهادی با مدل‌های فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان و تحلیل معنایی پنهان، از مجموعه داده عصر ایران با معیار ارزیابی انسجام و معیار ارزیابی انسانی استفاده نمودیم. بهترین نتایج در حالت کلی، از مدل پیشنهادی (با مدل تعبیه LaBSE) و مدل

فاکتورسازی ماتریس غیرمنفی به دست می‌آید. ضعیف‌ترین عملکرد نیز مربوط به مدل تخصیص دیریکله پنهان می‌باشد.

در دومین مقایسه روش پیشنهادی با مدل‌های فاکتورسازی ماتریس غیرمنفی، تخصیص دیریکله پنهان و تحلیل معنایی پنهان، از مجموعه داده تسنیم و معیارهای ارزیابی انسجام و تنوع استفاده نمودیم. روش پیشنهادی، به طور کلی نتایج بهتری را در مقایسه با سه مدل دیگر نشان داد. زیرا به طور کلی مطابق معیار انسجام، روش پیشنهادی (با هر دو مدل تعبیه Shiraz و LaBSE) عملکرد بهتری نسبت به سایر مدل‌های استخراج موضوع دارد. مدل‌های فاکتورسازی ماتریس غیرمنفی و تحلیل معنایی پنهان برای تعداد موضوعات کم، عملکرد خوبی را نشان می‌دهند. اما با افزایش تعداد موضوعات، کارایی آن‌ها کاهش می‌یابد. مطابق معیار انسجام، ضعیف‌ترین عملکرد به مدل تخصیص دیریکله پنهان اختصاص می‌یابد. به‌علاوه، مطابق معیار تنوع، مدل فاکتورسازی ماتریس غیرمنفی و سپس روش پیشنهادی با مدل تعبیه LaBSE متنوع‌ترین موضوعات را استخراج می‌کنند. با افزایش تعداد موضوعات، میزان تنوع در همه مدل‌ها کاهش می‌یابد. بیشترین تکرار در موضوعات حاصل از مدل تخصیص دیریکله پنهان مشاهده می‌شود. محققان معتقدند که اطلاعات آماری ناکافی برای استخراج ویژگی عامل اساسی در پشت موضوعات تکراری است [۴۳]. به‌طور کلی، در مدل تخصیص دیریکله پنهان کشف تعداد بهینه موضوعات کار دشواری است [۷]. همچنین موضوعات متعددی که توسط مدل تخصیص دیریکله پنهان در پژوهش ما شناسایی شده‌اند، مشابه [۷]، اطلاعات عمومی یا نامربوط را به دست می‌دهند. محققان معمولاً موافق هستند که مدل فاکتورسازی ماتریس غیرمنفی، با متون کوتاه به خوبی کار می‌کند [۴۴]. [۱۴] و [۷] نیز برتری روش پیشنهادی و مدل فاکتورسازی ماتریس غیرمنفی بر مدل تخصیص دیریکله پنهان را تایید می‌نمایند.

اگرچه هر سند ممکن است بیش از یک موضوع داشته باشد، روش پیشنهادی بر خلاف سه الگوریتم دیگر، به هر سند فقط یک موضوع اختصاص می‌دهد.

۶-۱- محدودیت‌های پژوهش

پژوهش حاضر با چند محدودیت همراه بوده است، از جمله:

- عدم وجود ابزارهای مؤثر برای پیش‌پردازش زبان فارسی، به‌ویژه در زمینه ریشه‌یابی.
- محدودیت منابع محاسباتی که باعث افزایش زمان اجرای مدل‌ها می‌شود.
- عدم وجود معیارهای دقیق برای ارزیابی مدل‌سازی موضوع.
- کمبود مجموعه داده‌های مناسب برای پردازش زبان فارسی.
- تفاوت لهجه‌ها و سبک‌های نوشتاری که دقت مدل‌ها را کاهش می‌دهد، هرچند به دلیل رسمی بودن متون استفاده‌شده در مجموعه داده‌های خبری، این موضوع تأثیر کمتری در این پژوهش داشته است.

۷- نتیجه‌گیری

این مقاله با هدف ارائه روشی کارآمد برای استخراج موضوعات از متون فارسی انجام شد. با الهام از چهارچوب BERTopic و استفاده از روش‌های تعبیه‌زبانی پیشرفته مبتنی بر ترنسفورمرها، مشخص شد که این روش در مقایسه با مدل‌های سنتی، مانند تخصیص دیریکله پنهان، تحلیل معنایی پنهان و فاکتورسازی ماتریس غیرمنفی، عملکرد بهتری در معیارهای انسجام و تنوع موضوعات دارد. نتایج نشان دادند که روش پیشنهادی با الگوریتم خوشه‌بندی K-Means و مدل تعبیه LaBSE، بهترین انسجام و تنوع را ارائه می‌دهد، در حالی که مدل فاکتورسازی ماتریس غیرمنفی تنها در تعداد کم موضوعات عملکرد بهتری داشت.

این پژوهش برتری روش پیشنهادی را به‌عنوان یک رویکرد نوین برای مدل‌سازی موضوعات متون فارسی تأیید می‌کند و بر اهمیت استفاده از مدل‌های پیشرفته زبانی در تحلیل متون تأکید دارد. یافته‌ها همچنین نیاز به بهبود روش‌های پیش‌پردازش زبان فارسی و گسترش مجموعه داده‌های متنوع‌تر برای پردازش زبان طبیعی را برجسته می‌کنند. این نتایج می‌توانند مسیر جدیدی برای تحقیقات آینده در زمینه متن‌کاوی و تحلیل خودکار متون فارسی فراهم کنند.

تعارض منافع

نویسندگان اعلام می کنند که در مورد انتشار این مقاله تعارض منافع وجود ندارد.

تاییدیه اخلاقی

نویسندگان متعهد می‌شوند که مطالب این مقاله را در هیچ مجله دیگری به چاپ نرسانده اند.

مشارکت های نویسندگان

زهرا فلاحی: روش‌شناسی، تحقیق، اعتبار سنجی، نرم افزار، نگارش پیش‌نویس اصلی، بررسی و ویرایش، منابع.

محمد رحمانی منش: روش‌شناسی، اعتبار سنجی، نگارش پیش‌نویس اصلی، راهنمایی.

منابع مالی

در انجام پژوهش هیچ منبع مالی مورد استفاده قرار نگرفته است.

مراجع

[1] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” *Machine Learning*, vol. 42, pp. 177–196, 2001.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Jan. 2003.

[3] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.

[4] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, Mar. 2022. [Online]. Available: http://arxiv.org/abs/2203.05794

[5] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform Manifold Approximation and Projection,” *J Open Source Softw*, vol. 3, no. 29, p. 861, Sep. 2018, doi: 10.21105/joss.00861.

[6] Z. Wang, J. Chen, J. Chen, and H. Chen, "Identifying interdisciplinary topics and their evolution based on BERTopic," *Scientometrics*, pp. 1–26, 2023, doi: 10.1007/s11192-023-04776-5.

[7] R. Egger and J. Yu, “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts,” *Frontiers in Sociology*, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.

[8] H. Kermani, “Framing the Pandemic on Persian Twitter: Gauging Networked Frames by Topic Modeling,” *American Behavioral Scientist*, 2023, doi: 10.1177/00027642231207078.

[9] S. Ghasemi and A. H. Jadidinejad, “Persian text classification via character-level convolutional neural networks,” in 2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN), Apr. 2018, pp. 1–6.

[10] V. S. Anoop, S. Asharaf, and P. Deepak, “Unsupervised Concept Hierarchy Learning: A Topic Modeling Guided Approach,” in *Procedia Computer Science*, Elsevier B.V., 2016, pp. 386–394. doi: 10.1016/j.procs.2016.06.086.

[11] F. Ebrahimi, M. Dehghani, and F. Makkizadeh, “Analysis of Persian Bioinformatics Research with Topic Modeling,” *Biomed Res Int*, vol. 2023, 2023, doi: 10.1155/2023/3728131.

[12] H. Amirkhani, M. AzariJafari, S. Faridan-Jahromi, Z. Kouhkan, Z. Pourjafari, and A. Amirak, “Farstail: A Persian natural language inference dataset,” *Soft Computing*, pp. 1–13, 2023, doi: 10.1007/s00500-023-08486-2.

[13] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, “ParsBERT: Transformer-based model for Persian language understanding,” Neural Processing Letters, vol. 53, pp. 3831–3847, 2021, doi: 10.1007/s11063-021-10528-4.

[14] A. Abuzayed and H. Al-Khalifa, “BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique,” *Procedia Comput Sci*, vol. 189, pp. 191–194, Jan. 2021, doi: 10.1016/J.PROCS.2021.05.096.

[15] K. Xiao, Z. Qian, and B. Qin, “A graphical decomposition and similarity measurement approach for topic detection from online news,” *Inf Sci (N Y)*, vol. 570, pp. 262–277, Sep. 2021, doi: 10.1016/j.ins.2021.04.029.

[16] K. Garcia and L. Berton, “Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA,” *Appl Soft Comput*, vol. 101, Mar. 2021, doi: 10.1016/j.asoc.2020.107057.

[17] M. Dehghani and F. Ebrahimi, “ParsBERT topic modeling of Persian scientific articles about COVID-19,” *Inform Med Unlocked*, vol. 36, Jan. 2023, doi: 10.1016/j.imu.2022.101144.

[18] M. H. Asnawi, A. A. Pravitasari, T. Herawan, and T. Hendrawati, “The Combination of Contextualized Topic Model and MPNet for User Feedback Topic Modeling,” *IEEE Access*, vol. 11, pp. 130272–130286, 2023, doi: 10.1109/ACCESS.2023.3332644.

[19] S. Samsir, R. S. Saragih, S. Subagio, R. Aditiya, and R. Watrianthos, "BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory," J. Media Inform. Budidarma, vol. 7, no. 3, pp. 1514–1520, Jul. 2023, doi: 10.30865/mib.v7i3.6426.

[20] M. Ranjbar-Khadivi, S. Akbarpour, M.-R. Feizi-Derakhshi, and B. Anari, “Persian topic detection based on Human Word association and graph embedding,” arXiv preprint arXiv:2302.09775, 2023.

[21] X. Wu, T. Nguyen, and A. T. Luu, “A survey on neural topic models: methods, applications, and challenges,” *Artif Intell Rev*, vol. 57, no. 2, Feb. 2024, doi: 10.1007/s10462-023-10661-7.

[22] M. Khazeni, M. Heydari, and A. Albadvi, “Persian Slang Text Conversion to Formal and Deep Learning of Persian Short Texts on Social Media for Sentiment Classification,” arXiv preprint arXiv:2403.06023, 2024.

[23] P. Ahmadi, M. Tabandeh, and I. Gholampour, “Persian text classification based on topic models,” in 2016 24th Iranian Conference on Electrical Engineering (ICEE), May 2016, pp. 86–91.

[24] L. McInnes, "How HDBSCAN Works," [Online]. Available: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html. [Accessed: Aug. 8, 2024].

[25] M. Grootendorst, "Topic Reduction," BERTopic Documentation. [Online]. Available: https://maartengr.github.io/BERTopic/getting_started/topicreduction/topicreduction.html. [Accessed: Aug. 8, 2024].

[26] HooshvareLab, "roberta-fa-zwnj-base: A Pretrained RoBERTa Model for Persian Text," Hugging Face, 2024. [Online]. Available: https://huggingface.co/HooshvareLab/roberta-fa-zwnj-base. [Accessed: Aug. 8, 2024].

[27] Y. Liu, et al., “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

[28] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, “MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices,” Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.02984

[29] R. SalehiChegeni, "Lifeweb AI team, Lifeweb language models," GitHub, 2024. [Online]. Available: https://github.com/lifeweb-ir/LM. [Accessed: June. 8, 2024].

[30] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, “Language-agnostic BERT Sentence Embedding,” Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.01852

[31] S. Takahashi, "paraphrase-multilingual-MiniLM-L12-v2," GitHub, 2024. [Online]. Available: https://github.com/shinichiro-takahashi-sbr/paraphrase-multilingual-MiniLM-L12-v2/blob/main/README.md. [Accessed: Aug. 8, 2024].

[32] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," arXiv preprint arXiv:1908.10084, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084. [Accessed: September 7, 2024].

[33] A. Conneau *et al.*, “Unsupervised Cross-lingual Representation Learning at Scale,” Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.02116

[34] A. Ghafouri, M. A. Abbasi, and H. Naderi, “AriaBERT: A Pre-trained Persian BERT Model for Natural Language Understanding,” 2023, doi: 10.21203/rs.3.rs-3558473/v1.

[35] M. Farahani, “m3hrdadfi/sentence-transformers.” Accessed: Aug. 08, 2024. [Online]. Available: https://github.com/m3hrdadfi/sentence-transformers

[36] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967.

[37] A. Gandhi, “Topic Modeling with LSA, PLSA, LDA & lda2Vec,” Jun. 2021, Accessed: Aug. 08, 2024. [Online]. Available: https://nanonets.com/blog/topic-modeling-with-lsa-plsa-lda-lda2vec/

[38] O. Oshriyeh, “Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications,” *Information Technology & Tourism*, vol. 25, no. 1, pp. 133–136, Mar. 2023, doi: 10.1007/s40558-023-00243-2.

[39] A. Pourmand, "TasnimNews Dataset (Farsi - Persian)," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/amirpourmand/tasnimdataset. [Accessed: July 15, 2024].

[40] HooshvareLab, "Persian News," 2020. [Online]. Available: https://hooshvare.github.io/docs/datasets/tc#persian-news. [Accessed: July 30, 2024].

[41] G. Bouma, “Normalized (pointwise) mutual information in collocation extraction,” in *Proceedings of GSCL 30*, 2009, pp. 31–40.

[42] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, “Topic Modeling in Embedding Spaces,” Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.04907

[43] G. Cai, G. Guoray, F. Sun, and Y. Sha, “Interactive Visualization for Topic Model Curation,” in IUI Workshops, 2018.

[44] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, “Experimental explorations on short text topic mining between LDA and NMF based Schemes,” *Knowl Based Syst*, vol. 163, pp. 1–13, Jan. 2019, doi: 10.1016/j.knosys.2018.08.011.