Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال دوم، شماره دوم، پاییز و زمستان ۱۴۰۳، صفحات ۲۰۱ – ۱۸۵

مدل‌سازی موضوعی متنی بر روی ریز نوشته‌های رسانه‌های اجتماعی فارسی

زینب متقی‌نیا[۱]، محمدرضا فیضی درخشی[۲*]

[۱]دانشجوی دکتری، آزمایشگاه سیستم‌های پردازش هوشمند رایانه‌ای، گروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

[۲]استاد، آزمایشگاه سیستم‌های پردازش هوشمند رایانه‌ای، گروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۳/۱۲/۱۰

تاریخ پذیرش:
۱۴۰۴/۲/۲۱

کلیدواژه‌ها:
مدل‌سازی موضوع، ریزنوشته، رسانه اجتماعی، تشخیص موضوع، تعبیه متن

نویسنده‌مسئول:
mfeizi@tabrizu.ac.ir

چکیده

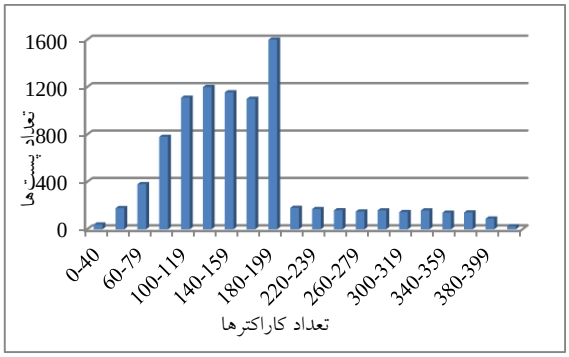
ظهور رسانه‌های اجتماعی فرصت‌های فزاینده‌ای برای اشتراک افکار کاربران فراهم می‌کنند. روزانه میلیاردها ریزنوشته در رسانه‌های اجتماعی تولید می‌شود که تحلیل آن‌ها در حوزه متن‌کاوی و تحلیل محتوا امری ضروری است. استخراج موضوعات دقیق از ریزنوشته‌ها در مقیاس بزرگ کاری مهم و چالش‌برانگیز است. مطالعات اندکی در زمینه تشخیص موضوع در ریزنوشته‌های فارسی انجام شده است و الگوریتم‌های موجود قابل‌توجه نیستند. ازاین‌رو بر آن شدیم در حوزه تشخیص موضوع در زبان فارسی به مطالعه بپردازیم. مدل‌سازی موضوعی از روش‌های تشخیص موضوع است که گروه‌هایی از کلمات را به‌عنوان موضوع از اسناد استخراج می‌کنند. اخیراً مدل‌های موضوع عصبی بهبودهایی برای افزایش انسجام مدل‌سازی موضوعی نشان داده‌است. همچنین، تعبیه‌های متنی مدل‌های عصبی را ارتقا داده‌اند. بدین‌سبب، در این تحقیق دو مدل موضوعی متنی ترکیبی و مدل موضوعی متنی ZeroShot برای استخراج موضوع در ریزنوشته‌های شبکه‌های اجتماعی فارسی ارائه شده‌است. این دو مدل بازنمایی متنی از پیش آموزش‌دیده‌شده BERT فارسی را در مدل‌های موضوع عصبی گنجانده‌اند. نتایج آزمایش‌ها نشان می‌دهد که این دو روش نسبت به روش‌های مورد مقایسه با بالاترین مقدار F1-Measure، تنوع موضوع و امتیاز انسجام بالا بهترین عملکرد را از خود نشان می‌دهند. همچنین مدل موضوعی متنی ZeroShot نسبت به مدل موضوعی متنی ترکیبی نتایج بهتری از لحاظ معیارهای ارزیابی داشته‌است.

doi : 10.22034/ABMIR.2025.22849.1106

E-ISSN: 2821-2037

۱- مقدمه

در دهه‌های اخیر تغییرات شگرفی در دنیای اینترنت رخ‌داده است. یکی از این تغییرات عظیم، ظهور سکو[1]های رسانه‌های اجتماعی است. میلیون‌ها کاربر برخط[2] فعال ممکن است روزانه میلیاردها پست رسانه‌های اجتماعی تولید کنند [۱] . داده‌های غنی و گسترده، رسانه‌های اجتماعی[3] را به منابع اطلاعاتی مهم مبدل ساخته که دارای پست‌هایی با ماهیت کوتاه تحت عنوان ریزنوشته[4] هستند. رسانه اجتماعی تلگرام[5] را می‌توان نمونه‌ای از سامانه ریزنوشته‌نگاری[6] به شمار آورد که در آن کاربران محتوای خود را عمدتاً در قالب پست‌های کوتاه (حداکثر ۳۰۰ کلمه) به اشتراک می‌گذارند. این محتوا می‌تواند شامل تصاویر، **GIF**، لینک‌ها، اینفوگرافیک‌ها، فیلم‌ها و کلیپ‌های صوتی باشد [۲]. شکل ۱ طول کاراکتر پست‌های تلگرام را نشان می‌دهد. بدین ترتیب تلگرام نیز مانند سایر رسانه‌های اجتماعی منبع ارزشمندی برای تحلیل داده‌ها [۳]–[۸] به‌خصوص درزمینه تشخیص موضوع خواهدبود. تشخیص موضوعات از چنین ریز نوشته‌هایی برای بسیاری از وظایف پردازش زبان طبیعی[7] از اهمیت بالایی برخوردار است.



شکل (۱): طول پست‌های تلگرام برحسب کاراکتر [۹]

مدل‌های موضوعی[8] کاربردهای وسیعی در حوزه‌های مختلف متن‌کاوی دارند [۱۰]–[۱۵]. بااین‌وجود، با افزایش تعداد داده‌های تولیدشده توسط کاربر در ریزنوشته‌ها، روش‌های مدل‌سازی موضوعی به ابزاری کلیدی برای استخراج دانش و موضوعات نهفته از ریزنوشته‌های رسانه‌های اجتماعی مبدل گشته‌اند. مدل‌سازی موضوعی داده‌های رسانه‌های اجتماعی در بسیاری از زمینه‌ها ازجمله روزنامه‌نگاری [۱۶]، بهداشت عمومی [۱۷]، برنامه‌ریزی شهری [۱۸]، علوم سیاسی [۱۹]، سیستم‌های اطلاعاتی [۲۰] وسایر موارد ارائه شده است. رویکرد تشخیص موضوع از ریز نوشته‌ها شامل سه کار فرعی کلیدی است: (الف) استخراج داده، (ب) پیش پردازش و (ج) روش مدل‌سازی موضوع. در مرحله اول داده‌های بدون ساختار و نیمه ساختاریافته از منابع داده جمع‌آوری می‌شوند. پس‌ازآن، مرحله پیش پردازش به پاک‌سازی مجموعه داده‌ها می‌پردازد. درنهایت، مدل‌سازی موضوع برای استخراج مجموعه‌ای از موضوعات استفاده می‌شود. اگرچه مدل‌سازی موضوع به‌طور گسترده در بسیاری از تحلیل‌های جریان متن برای شناسایی موضوعات در ریزنوشته‌های به زبان مختلف استفاده شده است [۲۱]–[۲۵]، پژوهش‌های کم‌تری به بررسی این روش در زبان فارسی پرداخته‌اند. مدل‌های موضوعی، گروه‌هایی از کلمات را از اسناد استخراج می‌کنند که امیدواریم تفسیر آن‌ها به‌عنوان یک موضوع به درک بهتر داده‌ها کمک کند. بااین‌حال، گروه‌های کلمه به دست آمده اغلب منسجم[9] نیستند و تفسیر آن‌ها را سخت‌تر می‌کند. اخیراً، مدل‌های موضوع عصبی بهبودهایی را در انسجام کلی نشان داده‌اند. به‌طور هم‌زمان، تعبیه‌های متنی، وضعیت مدل‌های عصبی جدید را به‌طورکلی ارتقا داده‌اند. در این پژوهش، با توجه به ویژگی‌های خاص زبان فارسی و ماهیت متون کوتاه در سکوهایی مانند تلگرام، بازنمایی‌های متنی پیشرفته مبتنی بر زبان فارسی را با مدل‌های موضوع عصبی و روش‌های پیش‌پردازش اختصاصی فارسی ترکیب می‌کنیم. این ترکیب هدفمند، منجر به طراحی چارچوبی یکپارچه برای مدل‌سازی موضوعی متنی می‌شود که قادر است موضوعات پنهان در ریزنوشته‌های فارسی را به صورت بدون ناظر شناسایی کند. این چارچوب با بهره‌گیری از

[1] Platform
[2] Online
[3] Social Media
[4] Microblog
[5] Telegram
[6] Microblogging Service
[7] Natural Language Processing (NLP)
[8] Topic Models
[9] Coherent

نقاط قوت هر دو رویکرد و پیش‌پردازش اختصاصی، چالشهای منحصربه‌فرد فارسی را حل کرده و با تولید تعبیه‌های معنایی عمیق از شبکه عصبی برای مدل‌سازی توزیع موضوعی در متون کوتاه استفاده می‌کند.

ساختار مقاله به شرح زیر است: بخش ۲ چالش‌های زبان فارسی در ریزنوشته‌ها را توصیف می‌کند. بخش ۳ به بررسی پژوهش‌های مرتبط پیشین در این زمینه می‌پردازد. جزئیات مدل‌های پیشنهادی در بخش ٤ ارائه می‌شود. بخش ۵ یک نمای کلی از مجموعه داده، معیارهای ارزیابی و نتایج تجربی را ارائه می‌دهد. درنهایت، مقاله در بخش ٦ نتیجه‌گیری می‌شود.

۲- چالش‌های زبان فارسی در ریزنوشته‌ها

در سال ۱۳۷۱ شمسی، ایران دومین کشور خاورمیانه بود که اتصال اینترنت را برقرار کرد [۲٦]. امروزه، اینترنت به منبع اصلی اطلاعات در ایران تبدیل شده و چالش‌ها و فرصت‌های متعددی را به همراه آورده است. اثربخشی سیستم‌های پردازش زبان طبیعی به دلیل تفاوت‌های ساختاری زبانی، بین زبان‌ها می‌تواند متفاوت باشد. بسیاری از پژوهش‌های مطرح در زبان‌های دیگر (مانند انگلیسی) می‌توانند به عنوان پایه نظری برای پژوهش‌های فارسی مورد استفاده قرار گیرند. بااین‌حال، تفاوت‌ها و چالش‌های زبان فارسی سبب می‌شود که این روش‌ها نیازمند سازگاری و تطبیق باشند. چالش‌های زبان فارسی در پردازش ریزنوشته‌ها شامل موارد زیر است:

- تفاوت‌های ساختاری ازجمله نویسه‌های منحصربه‌فرد، عدم فاصله‌گذاری استاندارد
- پیچیدگی‌های صرفی-نحوی از جمله صیغه‌های فعل متعدد،ساختار جمله‌بندی متفاوت
- کمبود داده‌های آموزشی
- ماهیت محاوره‌ای و کوتاه نویسی ریزنوشته‌ها
- وجود اختصارات، ایموجی‌ها، لینک‌ها، کلمات غیرفارسی، اشتباهات املایی
- عدم سازگاری با ابزارهای جهانی پیش‌پردازش همچون NLTK

بنابراین، هدف این پژوهش نادیده گرفتن دستاوردهای جهانی نیست، بلکه تاکید بر ضرورت ارائه روش‌های اختصاصی همسو با چالش‌های زبان فارسی است. تمرکز بر بهبود پیش‌پردازش، طراحی مدل‌های هوشمند و توسعه منابع داده‌ای برای ریزنوشته‌های فارسی، می‌تواند گامی کلیدی در غلبه بر این چالش‌ها باشد.

۳- کارهای پیشین

روش‌های تشخیص موضوع به‌طورکلی به دو دسته روش‌های سندمحور[1] و روش‌های ویژگی‌محور[2] تقسیم می‌شوند [۲۷]. روش‌های سندمحور، شباهت‌های بین اسناد را مقایسه می‌کنند و سپس اسناد مشابه را گروه‌بندی می‌کنند [۲۸]–[۳٤]. روش‌های ویژگی‌محور، گروه‌های کلمه را که به‌طور هم‌زمان در مجموعه اسناد ظاهر می‌شوند، می‌یابند [۳۵]–[٤۲]. مدل موضوع احتمالاتی[3] زیرمجموعه‌ای از روش‌های ویژگی‌محور هستند و بر اساس این ایده است که اسناد ترکیبی از موضوعات هستند که درآن هر موضوع با پیش‌بینی توزیع احتمال بر روی کلمات شناسایی می‌شود [٤۳]–[٤٦]. این مدل در سال‌های اخیر به‌طور گسترده برای استخراج موضوعات از اسناد متنی استفاده می‌شود.

۳-۱ روش‌های سندمحور

در روش‌های سندمحور، هر موضوع شامل مجموعه‌ای از اسناد است. شباهت بین اسناد با استفاده از معیارهای شباهت مقایسه می‌شود و اسنادی که دارای میزان شباهت بالاتر از حد آستانه تعیین‌شده هستند با هم گروه‌بندی می‌شوند. هر یک از گروه‌های تولیدشده نشان‌دهنده یک موضوع است. در پژوهش‌های [۳۰]–[۳۲]، روش‌هایی برای تشخیص موضوعات در ریزنوشته‌های توییتر ارائه شده است که در آن‌ها پیام‌های مشابه بر اساس فراوانی کلمه- معکوس فراوانی سند[4] گروه‌بندی می‌شوند. به عنوان نمونه، نویسندگان در پژوهش [۳۳] مدل تشخیص اولین داستان[5] مبتنی بر

[1] Document-Pivot
[2] Feature-Pivot
[3] Probability Topic Model
[4] Term Frequency-Inverse Document Frequency (TF-IDF)
[5] First Story Detection (FSD)

LSH[1]را پیشنهاد کردند. این روش از TF-IDF و هم‌رخدادی[2] کلمات سند برای یافتن شباهت و خوشه‌های اسناد استفاده می‌کند. هرچند روش‌های سندمحور قادرند معنای اسناد را به صورت جامع‌تری بازنمایی کنند، اما ممکن است از پیچیدگی محاسباتی و نویز بالایی رنج ببرند. به‌عنوان‌مثال، ازآنجایی‌که ریز نوشته‌ها جملات کوتاهی هستند، بردارهای تولیدشده با استفاده از TF-IDF ممکن است بردارهای پراکنده باشند؛ بنابراین، تضمین عملکرد روش‌های سندمحور دشوار است.

۳-۲ روش‌های ویژگی‌محور

روش‌های ویژگی‌محور، تحلیل را از خوشه‌بندی مستقیم اسناد به خوشه‌بندی اصطلاحات[3] یا کلمات کلیدی منتقل می‌کنند. در این رویکرد، هر موضوع به‌صورت گروهی از کلمات بیان می‌شود. در پژوهش [۳۵] امتیاز df-idft برای هر چندبخشی[4] در بازه زمانی i بر اساس فراوانی سند تعریف می‌شود. خوشه‌هایی با بالاترین امتیاز df-idft به عنوان موضوعات در آن بازه زمانی شناسایی می‌شوند. Twevent [۳۸] یک سیستم تشخیص رویداد مبتنی بر بخش[5] برای توییت‌ها است. این چارچوب از سه جزء اصلی تشکیل شده است: بخش‌بندی توییت، تشخیص بخش رویداد و خوشه‌بندی بخش رویداد. در بخش‌بندی توییت، هر توییت با استفاده از الگوریتم پیشنهادی در [۴۷] به بخش‌هایی که همپوشانی ندارند تقسیم می‌شود. سپس بخش‌های انفجاری[6] با مدل‌سازی فراوانی یک بخش به‌عنوان یک توزیع گاوسی براساس پنجره زمانی ثابت از پیش تعریف شده (به عنوان مثال، یک روز یا یک ساعت) شناسایی می‌شوند. سپس بخش‌های رویداد مربوط به رویداد یکسان باهم گروه‌بندی می‌شوند تا رویداد را توسط مؤلفه خوشه‌بندی رویداد تشکیل دهند. پس از خوشه‌بندی، ویکی‌پدیا برای شناسایی رویدادهای واقعی و استخراج خبرسازترین بخش‌ها برای توصیف رویدادهای شناسایی‌شده مورداستفاده قرار می‌گیرد. در تحقیق [۴۸] الگوریتم تشخیص موضوع مبتنی بر AGF[7] در رسانه اجتماعی توییتر ارائه شده است. در این روش پس از یافتن الگوهای پرتکرار با استفاده از tf-idf، مقادیر AGF بین هر جفت الگو با استفاده از پارامترهای Kr (رتبه‌بندی کلمه کلیدی[8]) و CIMAWA (مفهوم تقلید از توانایی ذهنی تداعی واژه‌ها[9]) محاسبه می‌شود. سپس با استفاده از مقادیر AGF خوشه‌هایی از الگوهای پرتکرار ایجاد می‌شود که به‌عنوان موضوع نام‌گذاری می‌شوند. روش مبتنی بر HWA[10] [۹] مدلی مبتنی بر درک انسان از تداعی واژه‌ها است که ترکیبی از مدل تعبیه کلمه و خوشه‌بندی است. این الگوریتم از سه روش، هم‌رخدادی کلمات کلیدی، رابطه معنایی بین کلمات و درک انسان از تداعی کلمات استفاده می‌کند و بر ساختارهای زبانی تأکید دارد. HWA از ۵ گام تشکیل شده است: در ابتدا جریان داده پیش پردازش شده و سپس کلمات کلیدی رتبه‌بندی شده و میزان هم‌رخدادی و AGF[11] محاسبه شده و الگو استخراج می‌شود. در گام بعدی تعبیه کلمات محاسبه و بردار الگو استخراج می‌شود. در گام چهارم فاصله الگو محاسبه و خوشه‌بندی الگو صورت می‌گیرد. درنهایت، خوشه‌ها رتبه‌بندی شده و موضوعات استخراج می‌شوند. روش پیشنهادی در پژوهش [۴۹]، روش تشخیص موضوعی را پیشنهاد می‌دهد که ترکیبی از HWA، روش‌های تعبیه گراف و روش‌های خوشه‌بندی است. در ابتدا هم‌رخدادی کلمات در پست‌های تلگرام استخراج شده و سپس محاسبات با استفاده از HWA انجام می‌شود که منجر به تولید نمودار AGF می‌شود. این نمودار به الگوریتم تعبیه گراف DeepWalk [۵۰] یا Node2Vec [۵۱] وارد می‌شود. درنهایت، پس از کاهش ابعاد با استفاده از الگوریتم UMAP [۵۲]، بردارهای به‌دست‌آمده برای خوشه‌بندی موضوعات مشابه با استفاده از الگوریتم‌های HDBSCAN [۵۳] یا Kmeans [۵۴] در کنار هم قرار می‌گیرند. نویسندگان در [۵۵] مفهوم سودمندی بالا[12] را معرفی کرده‌اند که در روش‌های تشخیص

[1] Locality Sensitive Hashing
[2] Co-Occurrence
[3] Terms
[4] N-Gram
[5] Segment
[6] Bursty
[7] Topic Detection Using AGF (TDA)
[8] Keyword Rating
[9] Concept For The Imitation Of The Mental Ability Of Word Association
[10] Human Word Association
[11] Association Gravity Force
[12] Pattern Mining With High Utility

موضوع خوشه‌بندی الگوهای با سودمندی بالا[1] [۳۷] و استخراج الگوهای با سودمندی بالا[2] [۵۶] مورداستفاده قرارگرفته است. روش HUPC در گام نخست گروهی از الگوهای نماینده را از جریان داده‌های ریزنوشته استخراج می‌کند و سپس این الگوها را در خوشه‌هایی از موضوعات گروه‌بندی می‌کند. برای کاهش تعداد الگوهای مورداستفاده در فرآیند خوشه‌بندی، تنها الگوهای نماینده استخراج می‌شوند که (۱) کل جریان داده را بتوانند تا حد امکان پوشش دهند و (۲) افزونگی‌ها در بین الگوها را تا حد امکان کم‌تر کنند. همچنین، روش HUPM برای شناسایی موضوعات در توییتر، بر پایه استخراج الگوهای با سودمندی بالا پیشنهاد شده است که هم‌زمان فراوانی و سودمندی[3] را در نظر می‌گیرد. در این روش، سودمندی کلمات در توییت‌ها بر اساس نرخ رشد فراوانی تعریف شده و گروه‌هایی از کلمات با فراوانی و سودمندی بالا توسط HUPM استخراج می‌شوند.

مدل‌های موضوعی، مانند تخصیص پنهان دیریکله[4] [۴۳]، نسخه‌های بهبودیافته آن [۵۷]، [۵۸] و روش تجزیه نامنفی ماتریس[5] [۵۹]، [۶۰]، در چندین سال گذشته به‌طور گسترده برای استخراج موضوعات مورداستفاده قرار گرفته‌اند. در این مدل‌ها، هر موضوع با یک توزیع چندجمله‌ای از کلمات نمایش داده می‌شود که با روش‌هایی مانند نمونه‌گیری گیبس[6] یا استنتاج متغیر[7] و براساس اطلاعات هم‌رخدادی کلمه در متن استخراج می‌گردد. بااین‌حال، طول محدود متون کوتاه، در مقایسه با متون معمولی، باعث کاهش هم‌رخدادی کلمات و درنتیجه مشکل پراکندگی[8] و نامسنجم بودن موضوعات تولیدشده می‌شود.

در سال‌های اخیر، مدل‌های موضوعی عصبی به‌طور فزاینده‌ای در استفاده از شبکه‌های عصبی برای بهبود روش‌های مدل‌سازی موضوعی موجود موفقیت نشان داده‌اند [۶۱]–[۶۷]. گنجاندن تعبیه‌های کلمه در مدل‌های سنتی مانند LDA، قابلیت استفاده از این بازنمایی‌های قدرتمند را نشان داد [۶۸]–[۷۱]. تعبیه‌های کلمه [۷۲] با هدف آموزش بازنمایی معنایی از کلمه در اسناد با مقیاس بزرگ (به‌عنوان‌مثال، ویکی‌پدیا) به‌طور گسترده موردمطالعه قرارگرفته‌اند. تعبیه‌های کلمه برای مقابله با مشکل پراکندگی داده‌ها در مدل موضوعی نیز اعمال شده است [۶۸]، [۷۳]–[۷۵]. در این مدل‌ها از شباهت معنایی تعبیه‌های کلمه برای بهبود کیفیت موضوعات استخراج شده استفاده می‌شود. در پژوهش [۶۵] روش BERTopic معرفی شده است که متن را با استفاده از تعبیه‌های متنی به بردارهای عددی تبدیل می‌کند، سپس ابعاد بردارها با استفاده از الگوریتم UMAP[9] [۵۲] کاهش می‌یابد. بردارهای کاهش‌یافته با استفاده از روش HDBSCAN[10] [۷۶] خوشه‌بندی می‌نماید. در نهایت با بهره‌گیری از روش c-TF-IDF [۷۷] موضوعات استخراج می‌شوند. در این مطالعه، با تلفیق توانمندی‌های مدل‌سازی موضوعی و تعبیه‌های کلمه و روش‌های پیش پردازش اختصاصی فارسی، رویکردی برای تشخیص موضوع در ریزنوشته‌های رسانه‌های اجتماعی فارسی ارائه شده است.

۴- روش پیشنهادی

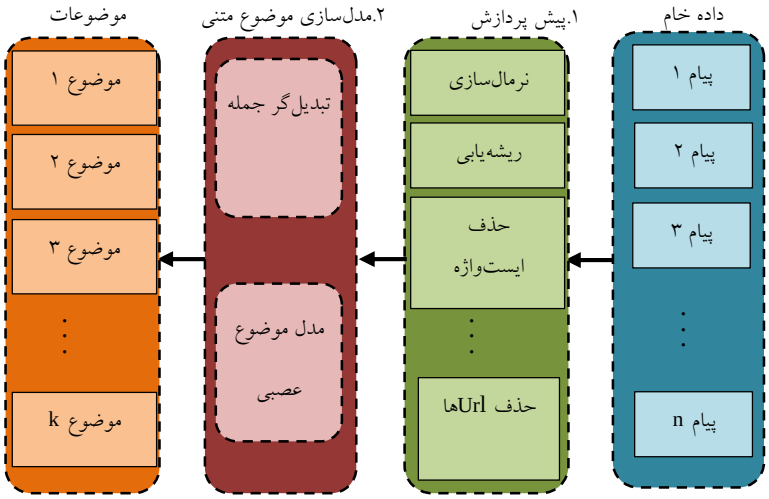
در این بخش، رویکردهای مدل‌سازی موضوع متنی[11] پیشنهادی خود را برای کشف ویژگی‌های معنی‌دار پنهان از اسناد متنی کوتاه مورد بحث قرار می‌دهیم. روند کلی روش پیشنهادی در شکل(۲ ترسیم شده است. در ابتدا، پیش پردازش بر روی داده‌های خام استخراج شده انجام می‌شود. سپس دو مدل موضوعی متنی شامل مدل موضوع متنی ترکیبی[12] و مدل موضوع متنی ZeroShot[13] اعمال می‌شود. مدل موضوع متنی از دو جزء تبدیل‌گر[14] جمله و مدل موضوع عصبی تشکیل شده است. در نهایت، موضوعات به‌عنوان خروجی استخراج می‌شود. در ادامه به تشریح هریک از این بخش‌ها می‌پردازیم.

[1] High Utility Pattern Clustering (HUPC)
[2] High Utility Pattern Mining (HUPM)
[3] Utility
[4] Latent Dirichlet Allocation (LDA)
[5] Non-Negative Matrix Factorization (NMF)
[6] Gibbs Sampling
[7] Variational Inference
[8] Sparsity

[9] Uniform Manifold Approximation and Projection
[10] Hierarchical Density-Based Spatial Clustering of Applications with Noise
[11] Contextualized Topic Models (CTM)
[12] CombinedTM
[13] ZeroShotTM
[14] Transformer

٤-١ پیش پردازش

پیش پردازش متن، فرآیند آماده‌سازی داده‌های متنی است تا ماشین‌ها بتوانند از همان داده‌ها برای انجام وظایفی همچون تجزیه‌وتحلیل، پیش‌بینی و سایر موارد استفاده کنند. مراحل مختلفی در پیش پردازش متن وجود دارد. در روش پیشنهادی، مراحل زیر به عنوان بخشی از فرایند پیش پردازش اعمال شده‌اند:



شکل(۲): روند کلی روش پیشنهادی

- **نرمال‌سازی**: متن را به فرم استاندارد تبدیل می‌کند، مانند اصلاح فاصله‌گذاری‌ها، یکسان‌سازی پیشوندها و پسوندها، حذف نشانه‌های اعراب از متن و غیره.
- **ریشه‌یابی**: ریشه کلمات را پیدا می‌کند و هدف آن ساده‌سازی و استانداردسازی کلمات است.
- حذف مواردی شامل:

موارد حذف‌شده	مثال
کاراکترهای ویژه	؛{}
ایموجی‌ها	☺☹
Urlها	https://google.com
حروف غیرفارسی	a-z, A-Z
اعداد	0-9

- **حذف ایست‌واژه‌ها**: رایج‌ترین کلمات موجود در زبان که اطلاعات زیادی به متن اضافه نمی‌کنند ایست‌واژه نامیده می‌شوند که معمولاً در مرحله پیش پردازش حذف می‌شوند. «و»، «که»، «اگر»، «ولی» نمونه‌هایی از ایست‌واژه‌ها در زبان فارسی هستند.

٤-٢ مدل موضوعی متنی[1]

مدل‌های موضوعی متنی خانواده‌ای از مدل‌های موضوعی هستند که از نمایش‌های زبانی از پیش آموزش دیده‌شده (مانند BERT) برای بهبود مدل‌سازی موضوع استفاده می‌کنند. در این پژوهش، از دو مدل موضوع متنی ترکیبی و مدل موضوع متنی ZeroShot برای تشخیص موضوع استفاده کردیم.

٤-٢-١ مدل موضوع متنی ترکیبی

مدل موضوع متنی ترکیبی [۷۸] یک مدل موضوع عصبی است که هر سند در مجموعه را هم به‌عنوان یک بردار کیفی از کلمات[2] و هم به‌عنوان یک بردار متراکم که توسط یک تبدیل‌گر از قبل آموزش دیده تولید می‌شود، نشان می‌دهد. این مدل از دو جزء اصلی ساخته شده است: (۱) تبدیل‌گر جمله SBERT [۷۹] و (۲) مدل موضوع عصبی[3] ProdLDA [۸۰].

[1] Contextualized Topic Models (CTM)
[2] Bag-Of-Words (BOW)
[3] Neural Topic Model

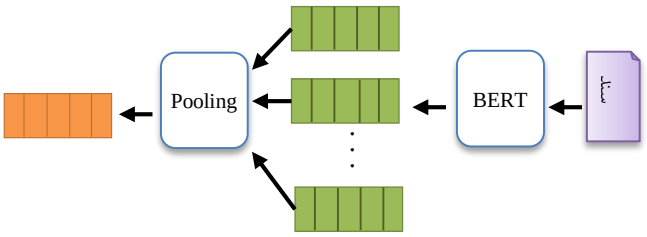
٤-٢-١-١ تبدیل‌گر جمله

تبدیل‌گر جمله به‌عنوان ابزاری قدرتمند برای رمزگذاری جملات در بردارهایی با ابعاد بالا که به‌عنوان تعبیه‌ها شناخته می‌شوند، ظهور کرده‌اند. در این مرحله، مدل با استفاده از تعبیه‌های متنی سند بدست آمده از تبدیل‌گر جمله SBERT گسترش داده می‌شود (شکل(۳). SBERT توسعه اخیر BERT بوده و امکان تولید سریع تعبیه‌های جمله را فراهم می‌کند. این چارچوب به ما اجازه می‌دهد جملات را با استفاده از مدل‌های زبانی از پیش آموزش‌دیده به نمایش‌های برداری متراکم تبدیل کنیم که در کاربردهای مبتنی بر تعبیه جمله عملکرد خوبی دارند [۸۱], [۸۲]. نمایش‌های سند از طریق یک لایه پنهان با ابعادی مشابه اندازه واژگان، همراه با نمایش BoW نمایش داده می‌شوند. برای این کار ما از تبدیل‌گر جمله ParsBERT [۸۳] که یک مدل تک‌زبانه فارسی بر اساس معماری BERT گوگل [٨٤] است، بهره برده‌ایم. مدل بر اساس معماری مدل BERT با تعداد کل ۱۱۰ میلیون پارامتر است و بر روی BERT BASE با ۱۲ لایه پنهان، ۱۲ سر توجه[1] و اندازه پنهان[2] ۷٦۸ ساخته شده است. این مدل بر 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. این تبدیل‌گر، جملات و پاراگراف‌ها را به یک فضای برداری متراکم ۷٦۸ بعدی نگاشت می‌کند و آن را برای بازیابی اطلاعات مرتبط با متن برای ایجاد پاسخ‌های دقیق[3] و منسجم[4] در برنامه‌های مختلف بسیار مؤثر می‌کند. طبق تحقیقات فراهانی و همکاران، مدل ParsBert از BERT چندزبانه[5] [۸۵] و مدل‌های قبلی در برخی از کارهای پایین‌دستی NLP فارسی، مانند طبقه‌بندی متن و تحلیل احساسات، پیشی می‌گیرد [۸۳]. ParsBERT نسبت به BERT چندزبانه بر روی مجموعه گسترده‌تر و متنوع‌تری از مجموعه داده‌های فارسی از پیش آموزش‌دیده شده است و وزن آن را سبک‌تر کرده است.

[1] Attention Head
[2] Hidden Sizes
[3] Accurate
[4] Coherent
[5] Multilingual

٤-٢-١-٢ مدل موضوع عصبی

ProdLDA[6] رویکرد مدل‌سازی موضوع عصبی مبتنی بر رمزگذار خودکار متغیر[7] [۸٦] است و به ما این امکان را می‌دهد که از قابلیت‌های یادگیری عمیق در مدل‌سازی متن استفاده کنیم [٦٤]. این چارچوب یک شبکه استنتاج عصبی یا رمزنگار[8] را برای نگاشت مستقیم بازنمایی BoW [۸۷] به یک بازنمایی پنهان پیوسته[9] آموزش می‌دهد. سپس، یک شبکه رمزگشا[10] BoW را با تولید کلمات آن از بازنمایی نهفته سند بازسازی می‌کند. این چارچوب به‌صراحت دیریکله را قبل از استفاده از توزیع‌های گاوسی تقریب می‌کند [۸۸]. علاوه بر این، ProdLDA توزیع چندجمله‌ای را بر روی کلمات جداگانه جایگزین می‌کند [۸۹].
شکل ٤ به‌طور خلاصه معماری مدل موضوع متنی ترکیبی را نمایش می‌دهد که دو نمایش آماری (BoW) و معنایی (متنی) را برای یادگیری بازنمایی‌های موضوعی اسناد ادغام می‌کند.



شکل(۳): معماری تبدیل‌گر جمله (بر پایه مدل BERT)

٤-٢-٢ مدل موضوع متنی ZeroShot

مدل موضوعی متنی ZeroShot [۹۰] یک معماری مدل‌سازی عصبی متنی مشابه مدل موضوعی متنی ترکیبی است که در آن بازنمایی‌های سند ورودی مبتنی بر BoW با تعبیه‌های متنی جایگزین می‌شود (شکل ۵). مدل‌های موضوعی عصبی، بازنمایی BoW سند را که حاوی اطلاعات ارزشمندی است، به‌عنوان ورودی می‌گیرند. بااین‌حال، ساختار این اطلاعات پس از عبور از

[6] Latent Dirichlet Allocation With Products Of Experts
[7] Variational Autoencoder (VAE)
[8] Encoder
[9] Continuous Latent
[10] Decoder

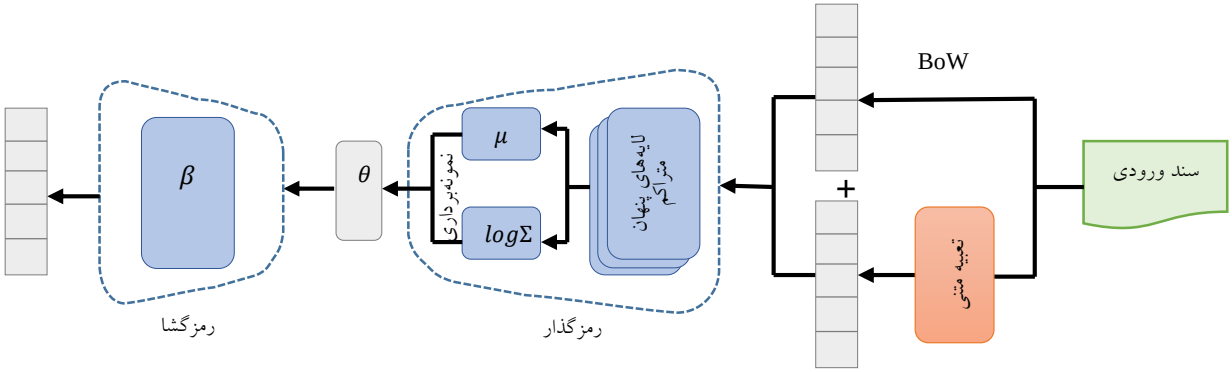
اولین لایه پنهان در معماری عصبی، از بین می‌رود. بنابراین، در این مدل اطلاعات متنی جایگزین بازنمایی BoW می‌گردد. این مدل با استفاده از بازنمایی اسناد ورودی که ترتیب کلمات و اطلاعات متنی را در نظر می‌گیرد، آموزش داده شده و بر یکی از محدودیت‌های اصلی مدل‌های مبتنی بر BoW غلبه می‌کند.

۵- مجموعه داده‌ها و نتایج پیاده‌سازی

در این بخش از تحقیق به تشریح جامع مجموعه داده مورد استفاده، معیارهای ارزیابی و جزئیات پیاده‌سازی می‌پردازیم.

۵-۱ مجموعه داده تلگرام

به‌منظور ارزیابی عملکرد الگوریتم پیشنهادی، از مجموعه داده تلگرام با نام Sep-General-Tel-01[1] [۹۱] استفاده شده است.



شکل (٤): معماری مدل موضوع متنی ترکیبی: ادغام نمایش آماری و بازنمایی معنایی

این مجموعه داده توسط آزمایشگاه سیستم‌های پردازش هوشمند رایانه‌ای (سپهر)[2] دانشگاه تبریز ارائه گردیده و به‌منظور رعایت اصول حریم خصوصی، تنها داده‌های مربوط به کانال‌ها و گروهای عمومیدر بازه زمانی ۱۲ دی‌ماه ۱۳۹۵ تا ۱۲ بهمن‌ماه ۱۳۹۵ جمع‌آوری شده است. این مجموعه داده شامل ۱۰۲۰۹ رکورد از پیام‌های ارسال‌شده است که با توجه به ماهیت تحقیق، تنها متن پیام‌ها مورد استفاده قرار گرفته است. این بانک اطلاعاتی در شصت پنجره ۱۲ ساعته تقسیم‌بندی شده است که شامل دو موضوع فوق‌العاده داغ[3] «رحلت آیت‌الله هاشمی رفسنجانی» و «آتش‌سوزی ساختمان پلاسکو» است. ۹ مورد از این پنجره‌ها که بیش‌ترین ارزش خبری و بهترین تطابق را با دو موضوع فوق‌العاده داغ دارند به‌عنوان داده مرجع[4] انتخاب و برچسب‌گذاری شده است. این پنجره‌ها شامل پنجره‌های [۱٤، ۱٥،۱٦، ۱۷، ۱۸، ۳۷، ۳۸، ۳۹، ٤۰] هستند.

خلاصه‌ای از جزییات این مجموعه داده در جدول ۱ ارائه شده است.

جدول (۱): جزییات مجموعه داده تلگرام

تعداد پست‌ها	۱۰۲۰۹
تعداد موضوعات فوق‌العاده داغ	۲
تعداد زیر موضوعات	۸۱
تعداد پنجره‌ها	٦۰
تعداد پنجره‌های داده مرجع	۹
میانگین تعداد پست‌های هر پنجره	۱۷۰
اندازه هر پنجره (ساعت)	۱۲
تعداد پست‌های برچسب خورده	۲۳٦٥

[1] داده‌های مرتبط از طریق نشانی زیر در دسترس هستند. https://doi.org/10.17632/372rnwf9pc
[1]Computerized Intelligence Systems Laboratory (ComInSyS)

[3] Super-hot
[4] Ground Truth (GT)

۵-۲ معیارهای ارزیابی

در این بخش از تحقیق به تشریح معیارهای ارزیابی موردنیاز برای سنجش عملکرد روش پیشنهادی در مقایسه با سایر روش‌های موجود خواهیم پرداخت.

۵-۲-۱ فراخوانی موضوع[1]

فراخوانی موضوع برابر است با تعداد موضوعاتی که به‌درستی از موضوعات داده مرجع استخراج شده‌اند و با استفاده از رابطه ۱ محاسبه می‌شوند.

$$Topic\ Recall = \frac{|successfully\ detected\ GT\ topics|}{|GT\ topics|} \quad (۱)$$

شود. مجموع موضوعات کشف‌شده و موضوعات کشف‌نشده برابر ۹ و تعداد موضوعات مرجع داده که کشف شده است برابر ۵ خواهد بود، همچنین مجموع موضوعات منطبق و غیرمنطبق با موضوعات مرجع داده برابر ۱۰ و تعداد موضوعات منطبق با موضوعات مرجع داده ۷ به دست می‌آید، بنابراین فراخوانی موضوع برابر ۵/۹ و دقت موضوع برابر ۷/۱۰ خواهد بود.

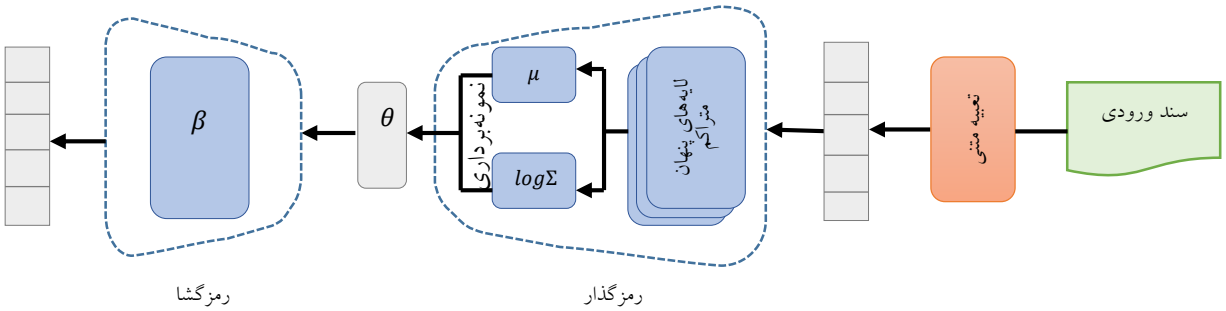
شکل (۶): توصیف بصری مفاهیم فراخوانی موضوع و دقت موضوع

۵-۲-۲ دقت موضوع[2]

دقت موضوع نسبت تعداد موضوعات استخراج شده که با موضوعات داده مرجع مطابقت دارند به کل موضوعات استخراج شده است و با استفاده از رابطه ۲ به دست می‌آید.

$$Topic\ Precision = \frac{|topics\ matches\ to\ GT|}{|extracted\ topics|} \quad (۲)$$

شکل (۶) مفهوم فراخوانی موضوع و دقت موضوع را به‌صورت بصری توصیف می‌کند. به‌عنوان‌مثال، فرض کنید که نه مرجع داده ارائه شده است. اگر هفت موضوع از بین ده موضوع استخراج شده توسط مدل با پنج موضوع داده مرجع مطابقت داده



شکل (۵): معماری مدل موضوعی متنی ZeroShot

۵-۲-۳ F1-Measure موضوع

F1-Measure موضوع درواقع ترکیبی از معیارهای دقت و فراخوانی موضوع است و با استفاده از رابطه ۳ به دست می‌آید.

$$Topic\ F1-measure = 2 * \frac{Topic\ Precision * Topic\ Recall}{Topic\ Precision + Topic\ Recall} \quad (۳)$$

کیفیت موضوعات استخراج شده ازنظر تنوع موضوع[3] و انسجام موضوع[4] ارزیابی می‌شوند. تنوع موضوع [۹۲] میزان منحصربه‌فرد بودن کلمات در بین تمامی موضوعات را با استفاده از معیار معکوس همپوشانی مبتنی بر رتبه[5] اندازه‌گیری می‌کند. معیار اطلاعات متقابل نقطه‌ای نرمال‌شده[6] [۹۳]، [۹۴] انسجام موضوع را

[1] Topic Recall
[2] Topic Precision
[3] Topic Diversity
[4] Topic Coherence
[5] Inverted Rank-biased Overlap (IRBO)
[6] Normalized Pointwise Mutual Information (NPMI)

به‌صورت داخلی با استفاده از یک پنجره کشویی برای شمارش الگوهای هم‌رخدادی کلمه می‌سنجد.

۵-۲-٤ معکوس همپوشانی مبتنی بر رتبه

معکوس همپوشانی مبتنی بر رتبه[۹۵]–[۹۷] میزان تنوع موضوعات تولیدشده توسط مدل را ارزیابی می‌کند. IRBO ،ده کلمه برتر هر دو موضوع را مقایسه می‌کند و به موضوعات اجازه می‌دهد کلمات کاملاً متفاوتی داشته باشند. این معیار از رتبه‌بندی وزنی استفاده می‌کند. در حالت ایده‌ال، ما انتظار موضوعاتی را داریم که مفاهیم یا ایده‌های جداگانه‌ای را نشان می‌دهند. موضوعاتی با کلمات مشترک در رتبه‌بندی‌های مختلف کم‌تر از موضوعاتی که کلمات مشابهی را در بالاترین رتبه‌ها به اشتراک می‌گذارند جریمه می‌شوند. IRBO عددی در بازه [۰،۱] را برمی‌گرداند. هر چه این معیار بالاتر باشد بهتر است [۹۸]. IRBO نزدیک به صفر یک موضوع زائد[1] را نشان می‌دهد و موارد نزدیک به یک موضوعات متنوع‌تری (کاملاًمتفاوت) را نشان می‌دهد. مقدار IRBO یک نشان می‌دهد که همه کلمات موضوع متفاوت هستند.

۵-۲-۵ اطلاعات متقابل نقطه‌ای نرمال‌شده

NPMI انسجام موضوع را با اندازه‌گیری احتمال هم‌رخدادی کلمات موضوع محاسبه می‌کند. اگر $\boldsymbol{p(w_i, w_j)}$ نشان‌دهنده احتمال هم‌رخدادی دو کلمه $\boldsymbol{w_i}$ و $\boldsymbol{w_j}$ در یک پنجره متنی لغزشی بولی باشد و $\boldsymbol{p(w_i)}$ احتمال حاشیه‌ای کلمه $\boldsymbol{w_i}$ باشد، امتیاز NPMI برابر رابطه ٤ است [۹۹].

$$NPMI(w_i, w_j) = \left(\frac{\log \frac{p(w_i, w_j) + \epsilon}{p(w_i).p(w_j)}}{-\log\left(p(w_i, w_j) + \epsilon\right)} \right) \qquad (٤)$$

در رابطه ٤، $\boldsymbol{\epsilon}$ یک عدد ثابت مثبت کوچک برای جلوگیری از صفر است. $\boldsymbol{NPMI(w_i, w_j)}$ در [۱+, ۱−] قرار دارد که ۱− نشان می‌دهد که کلمات هرگز هم‌زمان رخ نمی‌دهند و ۱+ نشان می‌دهد که آن‌ها همیشه هم‌رخداد هستند.

۵-۳ نتایج پیاده‌سازی

روش‌های مورداستفاده برای مقایسه عملکرد تشخیص موضوع شامل روش‌های Twevent [۳۸]، HUPC [۳۷]، HUPM [٥٦]، AGF [٤٨]، HWA based [۹]، AGF+ Node2Vec+ Hdbscan [٤٩]، AGF + deep walk + Hdbscan [٤٩]، AGF + Node2Vec+ kmeans [٤٩]، AGF + deep walk + kmeans [٤٩] و BERTopic [٦٥] می‌باشند. به یاد داشته باشید که یک مدل موضوعی یک مدل احتمالی[2] است و هر بار اگر با مقادیر یکسانی از ابرپارامترها اجرا شود (مثلاً تعداد موضوعات یکسان) نتایج متفاوتی تولید می‌کند. به همین دلیل، ما مدل‌های پیشنهادی را با تعداد موضوع یکسان برای ۵ بار اجرا کرده‌ایم. جدول ۲ عملکرد روش پیشنهادی و روش‌های مورد مقایسه را برحسب معیارهای فراخوانی موضوع، دقت موضوع و F1-measure موضوع نمایش می‌دهد. نتایج تجربی برای مجموعه داده تلگرام نشان می‌دهد که ZeroShotTM با امتیاز F1=0.7918 بهترین عملکرد را دارد و تعادل مناسبی بین فراخوانی‌موضوع (۰,۷۷۷۸) و دقت موضوع (۰,۸۰٦٤) ایجاد کرده است و نشان‌دهنده توانایی در کشف موضوعات مرتبط و دقیق است. CTMدر رتبه دوم قرار می‌گیرد (F1 =0.7156) و نشان می‌دهد روش‌های ZeroShotTM و CTM به دلیل استفاده از دانش زمینه‌ای و الگوریتم‌های مبتنی بر تبدیل‌گر، عملکرد بهتری نسبت به روش‌های سنتی و ترکیبی دارند. روش BERTopic دارای مقدار فراخوانی و دقت موضوع قابل توجهی و متعادل است. اگرچه **این روش در رتبه سوم قرار می‌گیرد، همچنان با روش‌های** پیشنهادی رقابت می‌کند. روش‌های سنتی (HUPC, HUPM, Twevent) فراخوانی موضوع بسیار پایین (مثلاً HUPM با ۰,۱۷۲٤) دارند که نشان می‌دهد موضوعات کمی را تشخیص می‌دهند. همچنین، دقت موضوع بالا (مثلاً HUPM با ۰,۹٤) اما F1پایین به دلیل ناتوانی در کشف موضوعات متنوع است. HUPMو HUPC با وجود دقت بالا، به دلیل فراخوانی بسیار ضعیف، برای کاربردهای واقعی مناسب نیستند. ترکیب AGF با الگوریتم‌های گراف:(Node2Vec, DeepWalk) بهبود نسبی

[1] Redundant

[2] Probabilistic Model

نسبت به AGF خالص ایجاد می‌کند (F1=0.636) و بهترین ترکیب AGF+Node2Vec+Hdbscan :با.F1=0.6308 است. اما همچنان عملکرد این روش پایین‌تر از CTM و ZeroShotTM. می‌باشد. AGF بالاترین فراخوانی (۰٫۸۲۷۵) اما دقت پایین (۰٫۵۱۶٤) دارد و این بدین معناست که تشخیص موضوعات زیاد اما با نویز همراه است.

جدول(۲):مقایسه عملکرد روش‌های تشخیص موضوع براساس معیارهای ارزیابی تشخیص موضوع

روش	فراخوانی موضوع	دقت موضوع	F1-measure موضوع
Twevent [۳۸]	۰٫٤۱۳۷۹۳	۰٫۵۲۷۲۷۳	۰٫٤٦۳٦۹۱
HUPC [۳۷]	۰٫۲۹۳۱۰۳	۰٫۹۰۸۰٤٦	۰٫٤٤۳۱٦۱
HUPM [۵٦]	۰٫۱۷۲٤۱٤	۰٫۹٤۱۱۷٦	۰٫۲۹۱٤۳۹
AGF [٤۸]	۰٫۸۲۷۵۸٦	۰٫۵۱٦٤۸٦	۰٫٦۳٦۰۳
HWA based [۹]	۰٫۵٦۱۹٦۵	۰٫۸۳۸۸۸۹	۰٫٦۷۳۰۵۵
AGF+Node2 Vec+Hdbscan [٤۹]	۰٫۵۱۰۲۹۳	۰٫۸۲۵۹۲٦	۰٫٦۳۰۸۳۱
AGF+deep walk+Hdbsca n [٤۹]	۰٫۵۰۷۰۷۹	۰٫۷۷٤۰۱۸	۰٫٦۱۲۷۳۸
AGF+Node2 Vec+kmeans [٤۹]	۰٫٤۸۰۱۳٤	۰٫۸۲۵۵۵۵	۰٫٦۰۷۱۵٤
AGF+deep walk+kmeans [٤۹]	۰٫٤٦۵۵۱۷	۰٫۸۱۸۱۸۲	۰٫۵۹۳٤۰۷
BERTopic [٦۵]	۰٫٦۳۱۷۳	۰٫۷٦٦٤	۰٫٦۹۲٦
CTM	۰٫٦٦٦۷	۰٫۷۷۲۷	۰٫۷۱۵٦
ZeroShotTM	۰٫۷۷۷۸	۰٫۸۰٦٤	۰٫۷۹۱۸

جدول ۳ میزان معکوس همپوشانی مبتنی بر رتبه که نشان‌دهنده تنوع موضوعات است را برای روش‌های مدل‌سازی موضوعی LDA، CTM و ZeroshotTm نشان می‌دهد. ما میزان IRBO را برای تعداد موضوعات ۱۰ تا ۲۰ محاسبه کردیم. بالاترین مقدار IRBO زمانی است که تعداد موضوعات برابر ۱۲ است. همان‌گونه که نتایج نشان می‌دهد ZeroShotTM با میانگین تقریباً ۰٫۹۹۲۱ بهترین عملکرد را دارد و در بیشتر موارد (به ویژه ۱۲ موضوع) به مقدار تقریباً ۰٫۹۹۸۵ می‌رسد. CTM نیز با میانگین تقریباً ۰٫۹۸۹۹ عملکردی نزدیک به ZeroShotTM دارد و در ۱۲ موضوع به ۰٫۹۹٤۲ می‌رسد.

جدول(۳): معکوس همپوشانی مبتنی‌بر رتبه روش‌های پیشنهادی

تعداد موضوع	CTM	ZeroShotTM
۱۰	۰٫۹۸۸۵۹۲٦۳٦٦۵۷۱٤۳	۰٫۹۹۱۳۱۷۹۸۸۳۸۵۷۱٤۳
۱۱	۰٫۹۹۳۱٦٦۵٤۳۷۲٤٦۷۵	۰٫۹۹۸۵۹۱۱۹۸۲۲۱۸۱۸۲
۱۲	۰٫۹۹٤۲۷۳۱۵۵۹٦۲۰۱۲۹	۰۹۹۸۵۹۰٤۷۳۵۲۹۵٤۵٤
۱۳	۰٫۹۸۵۷۵۹٤٦۷۹۹۵۸۷۹۱	۰٫۹۹۸٤۷٦۹۰٦۵۰۸۵۱٦۵
۱٤	۰٫۹۸۸۹۹۱۰۸۵٦۵۵٤۹٤۵	۰٫۹۹۳۳۳۱٦٤۹۸۰۷۷۷۰۸
۱۵	۰٫۹۹۱۰۱۳٤۸۸۹۰٦٤۳٦٤	۰٫۹۹۱۳۹۰۱۳٤۵۱۵۵۷۱٤۳
۱٦	۰٫۹۸۸۵۱۱۰۲۹۸۵۲۰۲۳۸	۰٫۹۸۹۰۸٤۸۷٦۱۳٤۱۰۷۱
۱۷	۰٫۹۸۸۱۵۷٤۷۹۹۱۵۷۵٦۳	۰٫۹۸۲۷۵۸۰۲۰۵٦۵۷۰۳۸
۱۸	۰٫۹۹۱۵٤۹۰۸۰٦٤۵۷۹۸۳	۰٫۹۹۰۳۹۵۷٤۹۵۰٤٦۲۱۸
۱۹	۰٫۹۹۱۷٤۱۳٦۰۸۷۰۳۰۰۷	۰٫۹۸۹۸٤۸۵۲۱۱۱۰۱۵۰٤
۲۰	۰٫۹۹۰۱٦٦۷۳۲۱۱۲۸۹٤۷	۰٫۹۸۹٦۷۰۲۸۲٦۰۸۹۰۹۸
میانگین	۰٫۹۸۹۹۱٦	۰٫۹۹۲۱۳۲

جدول ٤ مقادیر اطلاعات متقابل نقطه‌ای نرمال‌شده که نشان‌دهنده انسجام موضوعات است را برای روش‌های مدل‌سازی موضوعی LDA، CTM و ZeroshotTm نمایش می‌دهد. همان‌گونه که مشاهده می‌کنیم ZeroShotTM بالاترین مقدار NPMI (انسجام موضوعی) را به دست می‌آورد.

جدول(٤): اطلاعات متقابل نقطه‌ای نرمال‌شده روش‌های پیشنهادی

تعداد موضوع	CTM	ZeroShotTM
۱۰	-۰٫۰۰۷۸٤۲٤۷۰۰۷٦۸٤۵۸۹۱	-۰٫۰٤۹۹٦۸۸٤۵۳۹۰۹۲۱۷۱٦
۱۱	-۰٫۰٤٦۷۸۷۱۵۵۵٦۷٤۹۸۷۲٦	-۰٫۰۳۰۱۲۹۸۷۲٤۸۱۲٤۱۹۸۲
۱۲	-۰٫۰۳٦۷۵۸۵۲۸۷۳۰۸۰۹۸	-۰٫۰۵۸۳٦۰٦۵۵۳۳۵۸۲۲۵۸
۱۳	-۰٫۰٦۲٦۹۱۲۲٦۱٦۱۵۷۱۷۷	۰٫۰۱۹۹۸۰۷۱۷۲۰۳۲۹۰۱۰٦
۱٤	-۰٫۰۱۷۲۳۹٤۲۷۵۲۷۲۷۵۷۲	-۰٫۰۲۵۷۲۸٦٦٦۰٤۳٦۳٤۸۰۵
۱۵	-۰٫۰۳٤۵٦۹۹٤٦۹۷۸۸٦۳۷۱	-۰٫۰٦۵۳۰۷٦۹۲٤۲٦۲٤۵۸۲
۱٦	-۰٫۰۷۹۳۲٦۷۵۳٤٦۷۳٦۲۱۹	-۰٫۰٤۱٦۵۱٦۳۸۳٤۹۸۹۷۵۲
۱۷	-۰٫۰۱۳۱۳۸۱۰۸٦٦۷۰۳۸۸۳۸	-۰٫۰۱٦۰۸۹۷۸٦۳۹۲۹۰۷٦۳
۱۸	-۰٫۰۱۳۱۷٦٤۳۹۱۷۳٦۲۵۱۹	-۰٫۰۰۰۵۱۹٦۷٦۱۰۷٦٦۰٤۰۵۱
۱۹	-۰٫۰۳۹۵٤٦۵۷۳۷۸۲٦٤۵۸۷	-۰٫۰۳۹۳۲۱۹۰۹٤۹٤٦۵۹۵۲
۲۰	-۰٫۰۲۷۷۲٦۹۳۱۳۱۳۸۵۱٤٤۵	۰٫۰٦۷۷٤۹۹۱٦۷۸۸٤۰٦٦۸
میانگین	-۰٫۰۳٤٤٤	-۰٫۰۲۱۷٦

جدول ۵ میزان انسجام موضوع و تنوع موضوع را برای روش‌های مدل‌سازی موضوعی مورد مقایسه قرار داده است. همانگونه که در جدول ۵ مشاهده می‌کنید میزان NPMI در روش LDA پایین‌ترین مقدار را دارد. این روش مبتنی بر Bag-of-Words است و روابط معنایی را درک نمی‌کند.از سوی دیگر، روش NMF عملکرد کمی بهتر از LDA دارد.. CTM و ZeroShot مقدار NPMI بهتری از LDA را دارا می‌باشد. روش BERTopic از تعبیه‌های کلمات مبتنی بر BERT استفاده می‌کند که ارتباط معنایی بین کلمات را بهتر درک می‌کند و بهترین مقدار NPMI را دارد. در روش LDA موضوعات تمایل به اشتراک‌گذاری کلمات پرتکرار دارند و بنابراین این روش معیار IRBO پایینی دارد. روش BERTopic از الگوریتم‌های خوشه‌بندی مدرن (مثل UMAP+HDBSCAN) استفاده می‌کند که تنوع موضوعات را افزایش می‌دهد. زوش CTM و ZeroShot با ترکیب تعبیه‌های معنایی پیشرفته مثل SBERT با توزیع موضوعی مبتنی بر واریانس ، افزونگی را کاهش می‌دهد بهترین تنوع در داده‌های غیرساختاریافته دارند.

جدول(۵): مقایسه عملکرد روش‌های مدل‌سازی موضوع براساس میانگین معیارهای NPMI و IRBO

روش	میانگین NPMI	میانگین IRBO
LDA [۴۳]	۰,۸۶۳	-۰,۰۹۳
NMF [۶۰]	۰,۸۷۲	-۰,۱۱
BERTopic [۶۵]	۰,۹۹٤	-۰,۰۵۱
CTM	۰,۹۸۹	-۰,۰۳٤
ZeroShotTM	۰,۹۹۲	-۰,۰۲۲

نتایج بدست آمده نشان می دهد که دو روش CTM و ZeroShot در پردازش ریزنوشته های فارسی عملکرد بهتری به نسبت سایر روش‌ها دارند. مدل ZeroShotTM هنوز نتایجی بسیار مشابه به نتایج CombinedTM ارائه می‌دهد و اساساً می‌توانند برای کارهای مشابه استفاده شوند. این موضوع به‌ویژه در زبان فارسی که داده‌های آموزشی برچسبخورده کم‌تری در دسترس است، اهمیت بالایی دارد.

٦- نتیجه‌گیری

حجم زیادی از داده‌های موجود در سکوهای رسانه‌های اجتماعی، فرصت‌های جدیدی را برای استخراج اطلاعات در مورد دنیای واقعی فراهم می‌کند. با توجه به محبوبیت و کاربرد گسترده این سکوها، می‌توان از آن‌ها برای استنباط جنبه‌های مهم در مورد کاربران آن خدمات و رویدادهای محیطی تاثیرگذار بر آن‌ها بهره برد. در این میان، تشخیص موضوع در رسانه‌های اجتماعی یکی از موارد چالش‌برانگیز است. در این پژوهش، با تلفیق تعبیه‌های متنی پیشرفته، مدل های موضوعی و پیش پردازش اختصاصی متون کوتاه فارسی، دو روش مدل‌سازی موضوع متنی برای تشخیص موضوع در رسانه اجتماعی تلگرام فارسی توسعه دادیم. ارزیابی عملکرد این روش‌ها بر روی مجموعه داده تلگرام فارسی نشان می‌دهد که این دو روش مقدار F1-Measure بیش‌تر، انسجام موضوع نسبتاً بالا(همانطورکه توسط NPMI اندازه‌گیری می‌شود) و تنوع موضوع قابل مقایسه (همانطورکه توسط IRBO بدست آمده است) نسبت به روش‌های مورد مقایسه به دست آورده‌اند و به طور قابل توجهی کیفیت موضوعات کشف‌شده را بهبود می‌بخشند. پیشرفت‌های آینده در مدل‌های زبان باعث تولید مدل‌های موضوعی با عملکرد بهتر خواهد شد.

References

[1] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?,” in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 591–600.

[2] H.-T. Liao, K. Fu, and S. A. Hale, “How much is said in a microblog? A multilingual inquiry based on Weibo and Twitter,” in *Proceedings of the ACM Web Science Conference*, 2015, pp. 1–9.

[3] T. Lin, W. Tian, Q. Mei, and H. Cheng, “The dual-sparse topic model: mining focused topics and focused terms in short text,” in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 539–550.

[4] J. Qiang, P. Chen, W. Ding, T. Wang, F. Xie, and X. Wu, “Topic discovery from heterogeneous texts,” in *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2016, pp. 196–203.

[5] T. Shi, K. Kang, J. Choo, and C. K. Reddy, “Short-text topic modeling via non-negative matrix factorization enriched with local word-

context correlations," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1105–1114.
[6] T. Ramamoorthy, V. Kulothungan, and B. Mappillairaju, "Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India," *Front. Artif. Intell.*, vol. 7, p. 1329185, 2024.
[7] F. Zhang, W. Gao, Y. Fang, and B. Zhang, "Enhancing short text topic modeling with FastText embeddings," in *2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 2020, pp. 255–259.
[8] M. Asgari-Chenaghlu, M.-R. Feizi-Derakhshi, L. Farzinvash, M.-A. Balafar, and C. Motamed, "Topic detection and tracking techniques on Twitter: a systematic review," *Complexity*, vol. 2021, no. 1, p. 8833084, 2021.
[9] M. R. Khadivi, S. Akbarpour, M.-R. Feizi-Derakhshi, and B. Anari, "A Human Word Association based model for topic detection in social networks," *arXiv Prepr. arXiv2301.13066*, 2023.
[10] P. Kherwa and P. Bansal, "Topic modeling: a comprehensive review," *EAI Endorsed Trans. scalable Inf. Syst.*, vol. 7, no. 24, 2019.
[11] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, p. 101582, 2020.
[12] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Inf. Syst.*, vol. 112, p. 102131, 2023.
[13] R. Churchill and L. Singh, "The evolution of topic modeling," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–35, 2022.
[14] J. Boyd-Graber, Y. Hu, and D. Mimno, "Applications of topic models," *Found. Trends® Inf. Retr.*, vol. 11, no. 2–3, pp. 143–296, 2017.
[15] 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, p. 18, 2024.
[16] C. Jacobi, W. Van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," in *Rethinking Research Methods in an Age of Digital Journalism*, Routledge, 2018, pp. 89–106.
[17] A. T. Han, L. Laurian, and J. Dewald, "Plans versus political priorities: Lessons from municipal election candidates' social media communications," *J. Am. Plan. Assoc.*, vol. 87, no. 2, pp. 211–227, 2021.
[18] N. N. Haghighi, X. C. Liu, R. Wei, W. Li, and H. Shao, "Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service," *Public Transp.*, vol. 10, pp. 363–377, 2018.
[19] C. A. Bail *et al.*, "Exposure to opposing views on social media can increase political polarization," *Proc. Natl. Acad. Sci.*, vol. 115, no. 37, pp. 9216–9221, 2018.
[20] H. Pousti, C. Urquhart, and H. Linger, "Researching the virtual: A framework for reflexivity in qualitative social media research," *Inf. Syst. J.*, vol. 31, no. 3, pp. 356–383, 2021.
[21] E. Schubert, M. Weiler, and H.-P. Kriegel, "Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 871–880.
[22] C. K. Vaca, A. Mantrach, A. Jaimes, and M. Saerens, "A time-based collective factorization for topic discovery and monitoring in news," in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 527–538.
[23] C.-H. Lee, T.-F. Chien, and H.-C. Yang, "An automatic topic ranking approach for event detection on microblogging messages," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*, 2011, pp. 1358–1363.
[24] Y. Du, Y. Yi, X. Li, X. Chen, Y. Fan, and F. Su, "Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet Allocation," *Eng. Appl. Artif. Intell.*, vol. 87, p. 103279, 2020.
[25] X. Liu, Y. Gao, Z. Cao, and G. Sun, "LDA-based Topic Mining of Microblog Comments," in *Journal of Physics: Conference Series*, 2021, vol. 1757, no. 1, p. 12118.
[26] M. Sadeghi and J. Vegas, "How well does Google work with Persian documents?," *J. Inf. Sci.*, vol. 43, no. 3, pp. 316–327, 2017.
[27] Z. Mottaghinia, M.-R. Feizi-Derakhshi, L. Farzinvash, and P. Salehpour, "A Review of Approaches for Topic Detection in Twitter," *J. Exp. Theor. Artif. Intell.*, 2021.
[28] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," in *Proceedings of the international AAAI conference on web and social media*, 2011, vol. 5, no. 1, pp. 438–441.
[29] X. Zhou and L. Chen, "Event detection over twitter social media streams," *VLDB J.*, vol. 23, no. 3, pp. 381–400, 2014.
[30] B. O'Connor, M. Krieger, and D. Ahn, "TweetMotif: Exploratory Search and Topic

Summarization for Twitter.," in *ICWSM*, 2010, pp. 384–385.
[31] S. Phuvipadawat and T. Murata, "Breaking news detection and tracking in Twitter," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, 2010, vol. 3, pp. 120–123.
[32] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: news in tweets," in *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, 2009, pp. 42–51.
[33] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 181–189.
[34] M.-R. Feizi-Derakhshi, Z. Mottaghinia, and M. Asgari-Chenaghlu, "Persian Text Classification Based on Deep Neural Networks," *Soft Comput. J.*, vol. 11, no. 1, 2022.
[35] L. M. Aiello *et al.*, "Sensing trending topics in Twitter," *IEEE Trans. Multimed.*, vol. 15, no. 6, pp. 1268–1282, 2013.
[36] S. Gaglio, G. Lo Re, and M. Morana, "Real-time detection of twitter social events from the user's perspective," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 1207–1212.
[37] J. Huang, M. Peng, and H. Wang, "Topic detection from large scale of microblog stream with high utility pattern clustering," in *Proceedings of the 8th Workshop on Ph. D. Workshop in Information and Knowledge Management*, 2015, pp. 3–10.
[38] C. Li, A. Sun, and A. Datta, "Twevent: segment-based event detection from tweets," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp.
[39] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*,
[40] G. Petkos, S. Papadopoulos, L. Aiello, R. Skraba, and Y. Kompatsiaris, "A soft frequent pattern mining approach for textual topic detection," in *Proceedings of the 4th international conference on web intelligence, mining and semantics (WIMS14)*, 2014, pp. 1–10.
[41] J. Weng and B.-S. Lee, "Event detection in twitter," in *Proceedings of the international aaai conference on web and social media*, 2011, vol. 5, no. 1, pp. 401–408.
[42] W. Zhang, T. Yoshida, X. Tang, and Q. Wang, "Text clustering using frequent itemsets," *Knowledge-Based Syst.*, vol. 23, no. 5, pp. 379–388, 2010.
[43] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.
[44] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999.
[45] H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, "Enriching text representation with frequent pattern mining for probabilistic topic modeling," *Proc. Am. Soc. Inf. Sci. Technol.*, vol. 49, no. 1, pp. 1–10, 2012.
[46] D. Quercia, H. Askham, and J. Crowcroft, "Tweetlda: supervised topic classification and link prediction in twitter," in *Proceedings of the 4th Annual ACM Web Science Conference*, 2012, pp. 247–250.
[47] C. Li *et al.*, "Twiner: named entity recognition in targeted twitter stream," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 721–730.
[48] A. Benny and M. Philip, "Keyword Based Tweet Extraction and Detection of Related Topics," *Procedia Comput. Sci.*, vol. 46, pp. 364–371, 2015.
[49] M. Ranjbar-Khadivi, S. Akbarpour, M.-R. Feizi-Derakhshi, and B. Anari, "Persian topic detection based on Human Word association and graph embedding," *arXiv Prepr. arXiv2302.09775*, 2023.
[50] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701–710.
[51] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.
[52] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv Prepr. arXiv1802.03426*, 2018.

[53] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, 2013, pp. 160–172.
[54] Y.-W. Seo and K. Sycara, *Text clustering for topic detection*. Carnegie Mellon University, the Robotics Institute, 2004.
[55] M. Liu and J. Qu, "Mining high utility itemsets without candidate generation," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 55–64.
[56] H.-J. Choi and C. H. Park, "Emerging topic detection in twitter stream based on high utility pattern mining," *Expert Syst. Appl.*, vol. 115, pp. 27–36, 2019.
[57] Z. Chen and B. Liu, "Topic modeling using topics from many domains, lifelong learning and big data," in *International conference on machine learning*, 2014, pp. 703–711.
[58] Z. Chen and B. Liu, "Mining topics in documents: standing on the shoulders of big data," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1116–1125.
[59] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Adv. Neural Inf. Process. Syst.*, vol. 13, 2000.
[60] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421–2456, 2011.
[61] Z. Cao, S. Li, Y. Liu, W. Li, and H. Ji, "A novel neural topic model and its supervised extension," in *Proceedings of the AAAI Conference on artificial intelligence*, 2015, vol. 29, no. 1.
[62] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano, and A. Candelieri, "OCTIS: Comparing and optimizing topic models is simple!," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2021, pp. 263–270.
[63] H. Larochelle and S. Lauly, "A neural autoregressive topic model," *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
[64] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, and W. Buntine, "Topic modelling meets deep neural networks: A survey," *arXiv Prepr. arXiv2103.00498*, 2021.
[65] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv Prepr. arXiv2203.05794*, 2022.
[66] D. Angelov, "Top2vec: Distributed representations of topics," *arXiv Prepr. arXiv2008.09470*, 2020.
[67] H. Rahimi, H. Naacke, C. Constantin, and B. Amann, "ANTM: Aligned Neural Topic Models for Exploring Evolving Topics," in *Transactions on Large-Scale Data-and Knowledge-Centered Systems LVI: Special Issue on Data Management-Principles, Technologies, and Applications*, Springer, 2024, pp. 76–97.
[68] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *arXiv Prepr. arXiv1810.06306*, 2018.
[69] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical Word Embeddings.," in *AAAI*, 2015, pp. 2418–2424.
[70] J. Qiang, P. Chen, T. Wang, and X. Wu, "Topic modeling over short texts by incorporating word embeddings," in *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II 21*, 2017, pp. 363–374.
[71] M. Shi, J. Liu, D. Zhou, M. Tang, and B. Cao, "WE-LDA: a word embeddings augmented LDA model for web services clustering," in *2017 ieee international conference on web services (icws)*, 2017, pp. 9–16.
[72] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv Prepr. arXiv1301.3781*, 2013.
[73] R. Das, M. Zaheer, and C. Dyer, "Gaussian LDA for topic models with word embeddings," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 795–804.
[74] W. Gao, M. Peng, H. Wang, Y. Zhang, Q. Xie, and G. Tian, "Incorporating word embeddings into topic modeling of short text," *Knowl. Inf. Syst.*, vol. 61, pp. 1123–1145, 2019.
[75] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic Modeling for Short Texts with Auxiliary Word Embeddings," *Sigir*, no. September, pp. 165–174, 2016.
[76] L. McInnes and J. Healy, "Accelerated hierarchical density based clustering," in *2017 IEEE international conference on data mining workshops (ICDMW)*, 2017, pp. 33–42.
[77] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text

categorization," in *ICML*, 1997, vol. 97, pp. 143–151.
[78] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence," *arXiv Prepr. arXiv2004.03974*, 2020.
[79] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv Prepr. arXiv1908.10084*, 2019.
[80] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," *arXiv Prepr. arXiv1703.01488*, 2017.
[81] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," *arXiv Prepr. arXiv2004.09813*, 2020.
[82] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks," *arXiv Prepr. arXiv2010.08240*, 2020.
[83] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "Parsbert: Transformer-based model for persian language understanding," *Neural Process. Lett.*, vol. 53, pp. 3831–3847, 2021.
[84] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
[85] T. Pires, "How multilingual is multilingual BERT," *arXiv Prepr. arXiv1906.01502*, 2019.
[86] D. P. Kingma, "Auto-encoding variational bayes," *arXiv Prepr. arXiv1312.6114*, 2013.
[87] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, pp. 43–52, 2010.
[88] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *International conference on machine learning*, 2016, pp. 1727–1736.
[89] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.
[90] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, "Cross-lingual contextualized topic models with zero-shot learning," *arXiv Prepr. arXiv2004.07737*, 2020.
[91] M. Ranjbar-Khadivi, M.-R. Feizi-Derakhshi, A. Forouzandeh, P. Gholami, A.-R. Feizi-Derakhshi, and E. Zafarani-Moattar, "Sep TD Tel01," 2022.
[92] F. Nan, R. Ding, R. Nallapati, and B. Xiang, "Topic modeling with wasserstein autoencoders," *arXiv Prepr. arXiv1907.12374*, 2019.
[93] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 2010, pp. 100–108.
[94] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*, 2013, pp. 13–22.
[95] G. Carbone and G. Sarti, "ETC-NLG: End-to-end topic-conditioned natural language generation," *IJCoL. Ital. J. Comput. Linguist.*, vol. 6, no. 6–2, pp. 61–77, 2020.
[96] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Trans. Inf. Syst.*, vol. 28, no. 4, pp. 1–38, 2010.
[97] S. Terragni, D. Nozza, E. Fersini, and M. Enza, "Which matters most? comparing the impact of concept and document relationships in topic models," in *Proceedings of the First Workshop on Insights from Negative Results in NLP*, 2020, pp. 32–40.
[98] K. Murakami, N. Itsubo, and K. Kuriyama, "Explaining the diverse values assigned to environmental benefits across countries," *Nat. Sustain.*, vol. 5, no. 9, pp. 753–761, 2022.
[99] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 530–539.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 2, No. 2, Autumn & Winter 2024, pp. 185 - 201

Contextual Topic Modeling of Persian Social Media Short Texts

Zeynab Mottaghinia[1], Mohammad-Reza Feizi-Derakhshi [2*]

[1]PhD Candidate, Computerized Intelligence Systems Laboratory, Department of Computer Engineering, University of Tabriz, Tabriz, Iran

[2]Professor, Computerized Intelligence Systems Laboratory, Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

Article Information

Original Research Paper

Received:
2025 February 28

Accepted:
2025 May 11

Keywords:
Topic modeling, Short text, Social media, Topic detection, Sentence embedding.

Corresponding Author*:
mfeizi@tabrizu.ac.ir

Abstract

The emergence of social media creates opportunities for users to share their thoughts. Billions of short texts are produced on social media daily, and their analysis in text mining and content analysis is essential. Detecting topics from short texts on a large scale is an important and challenging task. Few studies have been conducted on topic detection in Persian short texts, and the existing algorithms are not remarkable. Therefore, we decided to study the topic detection in Persian. Topic modeling is a topic detection technique that extracts groups of words as topics from documents. Recently, neural topic models have shown improvements in increasing the coherence of topic modeling. Also, text embeddings have enhanced neural models. For this reason, in this study, two combined topic models and the ZeroShot topic model are presented for topic detection in Persian social media short texts. These two models incorporate pre-trained BERT text representation into neural topic models. The experimental results show that these two methods outperform the comparison methods with the highest F1-measure, topic diversity, and coherence score. Also, the ZeroShot topic model has better results in terms of evaluation metrics than the combined topic model.

: 10.22034/ABMIR.2025.22849.1106

E-ISSN: 2821-2037