Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال دوم، شماره اول، بهار و تابستان ۱۴۰۳، صفحات ۵۰ - ۳۹

استخراج برچسب برای آگهی‌های وب‌سایت مبتنی بر درون‌سازی واژه

علی‌محمد زارع چاهوکی[1*]، محمدمهدی صدقیان [2]

[1]دانشیار دانشکده کامپیوتر، دانشگاه یزد، یزد، ایران

[2]دانشکده کامپیوتر، دانشگاه یزد، یزد، ایران

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۲/۰۸/۲۵

تاریخ پذیرش:
۱۴۰۳/۰۱/۲۱

کلیدواژه‌ها:
استخراج برچسب، بهینه‌سازی برای موتور جست‌وجو (سئو)، یادگیری عمیق، درون‌سازی واژه.

نویسنده‌مسئول:
chahooki@yazd.ac.ir

چکیده

در جامعه امروزی اخبار و آگهی‌ها، جایگاه به‌خصوصی در رشد و ترقی جامعه دارند. با مشخص‌کردن واژگان اصلی آگهی، می‌توان به مفهوم کلی آن پی برد. آماده‌سازی این واژگان به روش سنتی نیازمند صرف زمان و دانش تخصصی راجع به موضوع متن است. سایت ایده‌کاو، سامانه‌ای هست که به جمع‌آوری پیام‌ها و آگهی‌های تلگرام می‌پردازد. نیازمندی سامانه ایده‌کاو، استخراج کلمات کلیدی از آگهی‌های منتشر شده در تلگرام بوده است. کیفیت کلمات کلیدی استخراج شده، نقش بسزایی در بهبود سئو و آمار بازید آگهی‌ها دارد. با استفاده از الگوریتم‌های درون‌سازی، می‌توان صحبت‌های محاوره‌ای و ساختار معنایی متن را استخراج کرد، ازاین‌رو در تشخیص کلمات کلیدی در آگهی‌های تلگرام که اغلب به‌صورت عامیانه منتشر می‌شوند، مفید واقع می‌شود. در این پژوهش با استفاده از داده‌های سامانه ایده‌کاو مدلی از روش‌های درون‌سازی پیاده‌سازی شده است. نوآوری استفاده شده در این پژوهش از ترکیب‌کردن روش‌های درون‌سازی واژه، بسامد کلمات و جایگاه کلمات ایجاد شده است. مدل درون‌سازی از کلمات دو کلمه‌ای ایجاد شده است. ایجاد مدل از کلمات دو کلمه‌ای، به این دلیل است که اغلب کلمات کلیدی از دو کلمه به بالا تشکیل شده‌اند. جهت نمایش بهتر ارزیابی‌ها، مدل آی‌کی (مدل پیشنهادی) با روش‌های آماری و روش‌های مبتنی بر گراف مقایسه شده است که نتایج به‌دست‌آمده نشان می‌دهد ترکیب مدل آی‌کی دو-گرم عملکرد بهتری در استخراج کلمات کلیدی نسبت به سایر روش‌ها به وجود آورده است.

: 10.22034/ABMIR.2024.20879.1041

E-ISSN: 2821-2037

۱- مقدمه

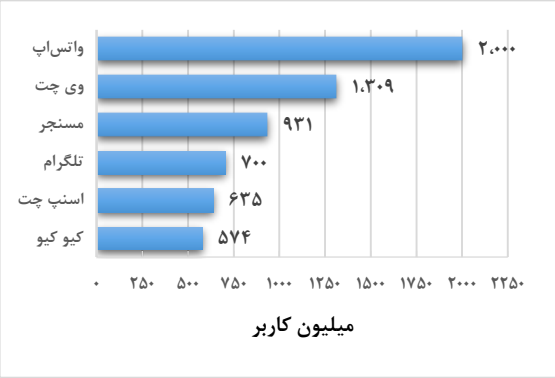
در عصر جدید استفاده از روش‌های نوین، برای بازیابی اطلاعات امری مهم محسوب می‌شود. کلمات کلیدی، مجموعه‌ای از کلمات مهم در سند هستند که محتوای سند را توصیف می‌کنند. کلمات کلیدی اطلاعات نحوی مفیدی را برای پردازش متن فراهم می‌کنند. استخراج کلمات کلیدی، فرایندی خودکار برای تشخیص کلمات مهم به‌کاررفته در سند است. این کلمات، معنا و مفهوم سند را برای خواننده آشکار می‌کنند و می‌توانند در بسیاری از کارهای زبان طبیعی مانند تجزیه‌وتحلیل متون و خلاصه‌سازی استفاده شوند [۱].

کلمات کلیدی را می‌توان به‌عنوان برچسب برای اسناد و دسته‌بندی محتوا و محصولات سایت استفاده کرد. کلمات کلیدی، برچسب‌هایی هستند که سند یا محصول را به بهترین شیوه ممکن برای کاربران معرفی می‌کنند.

یکی از موارد مؤثر در سئو، کلمات کلیدی و برچسب‌ها هستند. از نظر سئو، کلمات کلیدی عبارت‌هایی هستند که کاربران در موتورهای جست‌وجو وارد می‌کنند. انتخاب کلمات کلیدی مناسب برای هر داده، موجب رضایت کاربران و افزایش سئو سایت می‌شود [۲-٤].

اکثر پژوهش‌های موجود در زمینه استخراج کلمات کلیدی، برای رسانه اجتماعی توییتر، اسناد علمی و اسناد خبری صورت‌گرفته است و در حوزه آگهی‌های تلگرام، پژوهش‌های اندکی انجام شده است. تلگرام در ایران یکی از پرکاربردترین شبکه‌های اجتماعی است که روزانه حجم زیادی از اطلاعات را منتشر می‌کند. سایت ایده‌کاو سامانه‌ای هست که به جمع‌آوری پیام‌ها و آگهی‌های تلگرام می‌پردازد. با توجه به نیازمندی سامانه ایده‌کاو در استخراج کلمات کلیدی از آگهی‌های منتشر شده در تلگرام، در این پژوهش به دنبال ارائه روشی هستیم که بتواند کلمات کلیدی موجود در آگهی‌های تلگرام را شناسایی و استخراج کند. کیفیت کلمات کلیدی استخراج شده، نقش بسزایی در بهبود سئو و آمار بازید آگهی‌ها دارد.

پیام‌رسان تلگرام، برنامه‌ای رایگان و مبتنی بر رایانش ابری است. تلگرام پیام‌رسانی است که امکان ارسال پیام به‌صورت متن، صوت، تصویر و ویدئو را امکان‌پذیر می‌کند. تلگرام برای دستگاه‌های مبتنی بر اندروید، آی‌اواس، ویندوز، لینوکس و وب منتشر شده است. تلگرام با امکانات زیاد و کاربردی خود توانسته از سایر پیام‌رسان‌ها متمایز باشد و کاربران زیادی را به‌سوی خود جذب کند[۵]. در شکل ۱ به مقایسه محبوبیت برنامه‌های مختلف در حوزهٔ شبکه‌های اجتماعی می‌پردازیم [٦].



شکل (۱) : تعداد کابران فعال ماهانه در ژانویه ۲۰۲۳

در این پژوهش از روش‌های مختلف آماری و درون‌سازی جهت استخراج کلمات کلیدی استفاده شده است. ارزیابی‌ها نشان می‌دهد که ترکیب‌کردن روش‌های آماری به همراه مدل شخصی وُردتووِک که از واژگان کلیدی دو کلمه‌ای ایجاد شده است، از کیفیت بالاتری برخوردار است. در این پژوهش نام استفاده شده برای روش پیشنهادی آی‌کِی می‌باشد.

نوآوری استفاده شده در این پژوهش از ترکیب‌کردن روش‌های درون‌سازی واژه، بسامد کلمات و جایگاه کلمات ایجاد شده است. هدف ایجاد مدل از کلمات دو کلمه‌ای به این دلیل است که اغلب کلمات کلیدی از دو کلمه به بالا تشکیل شده‌اند. در آزمایش‌های انجام شده، مدل‌های ساخته شده از عبارت‌های دوکلمه‌ای، بهبود قابل‌توجهی نسبت به مدل‌های ساخته‌شده از عبارت‌های تک‌کلمه‌ای داشته‌اند .

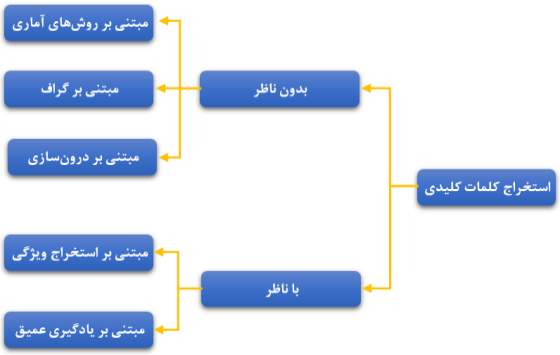
تمرکز اصلی این مقاله بر ساخت مدل‌های درون‌سازی از عبارت‌های دو کلمه‌ای و مقایسه با روش‌های آماری برای استخراج کلمات کلیدی است. به همین جهت در بخش دوم نگاهی به پژوهش‌های انجام شده در زمینه استخراج کلمات کلیدی می‌اندازیم. در فصل سوم چگونگی پیاده‌سازی و روش پیشنهادی

مطرح می‌شود. در بخش چهارم آزمایش‌های تجربی بر روی دادگان مطرح می‌شود و در فصل پنجم به جمع‌بندی مطالب و ذکر پیشنهادها برای پژوهش‌های آینده می‌پردازیم.

۲- پیشینه پژوهش

روش‌های متعددی برای استخراج واژگان‌کلیدی وجود دارد. طبق دسته‌بندی انجام شده، رویکردهای استخراج عبارات کلیدی به دو دستهٔ مبتنی بر یادگیری بدون ناظر و مبتنی بر یادگیری باناظر تقسیم می‌شود [۷].

تقسیم‌بندی روش‌های استخراج واژگان کلیدی در شکل ۲ آمده است:



شکل (۲) : دسته‌بندی روش‌های استخراج عبارات کلیدی

رویکردهای بدون ناظر اغلب مستقل از زبان و دامنه هستند و به مجموعه‌دادگان آموزشی برچسب‌دار نیازی ندارد. در مقابل، رویکردهای با ناظر دارای کیفیت بهتری نسبت به رویکرد بدون ناظر هستند؛ ولی به دادگان برچسب‌دار نیازمند می‌شوند.

۲-۱ رویکردهای آماری

در رویکردهای آماری، عبارت‌ها بر اساس اطلاعات آماری استخراج شده از متن به دست می‌آید. در این رویکرد کلمات کاندیدا با استفاده از شیوه‌های مختلف، رتبه‌بندی می‌شود. روند کلی استخراج عبارات کلیدی با استفاده از رویکردهای آماری، شامل موارد زیر می‌باشند:

- پیش‌پردازش متن
- استخراج ویژگی‌های آماری
- امتیازدهی عبارات
- مرتب‌سازی عبارات بر اساس امتیازات
- انتخاب تعداد مشخصی از عبارات کاندیدا که دارای بیشترین یا کمترین وزن هستند

روند کلی رویکردهای آماری در شکل ۳ نشان‌داده‌شده است.

شکل (۳) : روند کلی رویکردهای آماری

وون و همکاران برای استخراج کلمات کلیدی از ویژگی‌های فراوانی کلمه، TF-IDF، جایگاه کلمه و طول کلمه استفاده کرده‌اند. در ویژگی طول کلمه، برای عبارت‌های تک‌کلمه‌ای امتیاز یک و برای عبارت چندکلمه‌ای امتیاز دو را در نظر گرفته‌اند. امتیاز هر عبارت کاندیدا از حاصل‌ضرب چهار ویژگی ذکر شده به دست می‌آید. در مرحله آخر عبارت‌هایی که امتیاز بیشتری داشته باشد به‌عنوان کلمات کلیدی انتخاب می‌شوند. تعداد کلمات کلیدی انتخاب شده برای هر سند از رابطهٔ (۱) محاسبه می‌شود که عبارت بیانگر تعداد کلمات سند است [۸].

$$N = 2.5 \times log_{10}^{|d|} \qquad (۱)$$

فلورسکو[1] و همکاران، روشی مبتنی بر گراف و بدون ناظر، برای استخراج کلمات کلیدی پیشنهاد کرده‌اند. آنها گرافی از کلمات ایجاد کرده‌اند که گره‌های آن را کلمات و ارتباط بین کلمات را یال‌ها تشکیل داده‌اند. جهت امتیازدهی، از ویژگی‌های فراوانی کلمه، جایگاه کلمه و هم رخدادی کلمات استفاده کرده‌اند. با اعمال الگوریتم پیج‌رنک وزن‌دار بر گراف، گره‌هایی که دارای بیشترین وزن بوده‌اند به‌عنوان کلمات کلیدی معرفی شده‌اند. آنها در مقایسه با چند روش بدون ناظر مانند تکست‌رنک، اکسپندرنک، سینگل‌رنک و تاپیک‌رنک از عملکرد بهتری برخوردار بوده‌اند [۹].

[1] Florescu

۲-۲ رویکردهای مبتنی بر درون‌سازی

بنجیو[1] و همکارانش در سال ۲۰۰۳ عبارت درون‌سازی واژه را ابداع کرده‌اند. کلبرت[2] و وستون اولین کسانی بودند که توانایی درون‌سازی واژه‌های از پیش آموزش‌دیده را در سال ۲۰۰۸ نشان دادند. آنها نه‌تنها درون‌سازی واژه را به‌عنوان یکی از ابزارهای سودمند برای پردازش زبان طبیعی، تثبیت کردند، بلکه معماری شبکه عصبی‌ای معرفی کردند که اساس بسیاری از روش‌های امروزی است. پیشرفت درون‌سازی واژه را می‌توان به کار میکولوف[3] و همکارانش در سال ۲۰۱۳ که وُردتووک را ارائه کردند منتسب کرد. در سال ۲۰۱٤، پنینگتون[4] و همکارانش الگوریتم بردار سراسری را ارائه کردند و با این کار نشان دادند درون‌سازی واژه وارد استفاده روزمره شده است [۱۰].

روش درون‌سازی واژه، یک روش شبکه عصبی پیش‌خور است که با استفاده از یک لایه پنهان، توانایی تولید بردار برای کلمات را دارد. این روش با هدف کاهش ابعاد و افزایش کارایی در پردازش زبان طبیعی طراحی شده است. روش درون‌سازی با شبکه عصبی بازگشتی متفاوت است، زیرا در شبکه عصبی بازگشتی، خروجی هر گره، به ورودی گره‌های بعدی منتقل می‌شود و حافظه‌ای از داده‌های قبلی را نگه می‌دارد. درحالی‌که در شبکه عصبی پیش‌خور، اطلاعات فقط در یک جهت، از لایه ورودی به لایه خروجی حرکت می‌کنند و هر گره فقط به گره‌های لایه بعد خود متصل می‌شود.

یکی از مهم‌ترین وظایف در پردازش زبان طبیعی، تبدیل متن به حالتی قابل‌فهم برای پردازش است. برای نمایش معنا و مفهوم متن، روش‌های درون‌سازی واژه معرفی شده‌اند. در الگوریتم‌های پیش آموزش‌دیده، نمایش برداری کلمات شامل روابط معنایی، نحوی و منطقی بین کلمات است [۱۱]. در درون‌سازی، کلمات مشابه دارای بردارهای مشابه هستند و در فضای برداری در فاصلهٔ نزدیک به هم قرار می‌گیرند. برای مثال کلمات ملک، خانه و منزل هر سه در فضایی نزدیک به هم قرار می‌گیرند. در درون‌سازی واژگان، معنای واژه با استفاده از کلمات اطرافش به دست می‌آید [۱۲].

درون‌سازی واژه از مدل‌های مختلفی جهت ایجاد بردار و پردازش کلمات استفاده می‌کند. روش‌های مبتنی بر شبکه‌های عصبی عموماً پیش‌بینی محور هستند که در آنها مدل‌ها برای پیش‌بینی کلمات محتوایی با استفاده از کلمات وسطی، یا برعکس پیش‌بینی کلمه وسط توسط مجموعه‌ای از کلمات محتوایی ساخته می‌شوند. در این الگوریتم، دو معماری مدل کیسه کلمات پیوسته و مدل اسکیپ-گرام ارائه شده است:

- کیسه کلمات پیوسته: واژهٔ فعلی با توجه به زمینهٔ موردبحث متن، پیش‌بینی می‌شود و برای کلمات پر تکرار، بردارهای بهتری ارائه می‌دهد.
- اسکیپ-گرام: واژهٔ فعلی بر اساس کلمات اطرافش پیش‌بینی می‌شود و با داده‌های کوچک بهتر کار می‌کند. این مدل برای کلمات کمیاب، نمایش برداری بهتری نسبت به مدل کیسه کلمات پیوسته ارائه می‌دهد.

هارمان[5] و همکاران، برای استخراج کلمات کلیدی، روشی بر پایه امبدد رنک طراحی کرده‌اند. آنها علاوه بر محاسبه شباهت کسینوسی میان بردار سند درون‌سازی شده و بردار کلمه درون‌سازی شده، از عاملی به نام پاداش وزنی جهت امتیازدهی استفاده می‌کنند. برای محاسبه پاداش وزنی، برای کلماتی که در عنوان سند آمده‌اند امتیاز بیشتری نسبت به کلماتی که در محتوا سند آمده‌اند در نظر گرفته می‌شود [۱۳].

بنانی اسمیرس[6] و همکاران، روشی به نام امبدد رنک پیشنهاد کرده‌اند. آنها برای شناسایی کلمات کلیدی، الگوریتمی با استفاده از برچسب‌زن ادات سخن طراحی کرده‌اند. الگوریتم استخراج کلمات کلیدی، از یک یا چند اسم که به همراه یک یا چند صفت آمده است، تشکیل می‌شود. آنها از دو الگوریتم درون‌سازی داک‌تووک[7] و سنت‌تووک[8] برای نمایش برداری کلمات کلیدی استفاده کرده‌اند. در مرحله آخر شباهت کسینوسی بین بردار سند درون‌سازی شده و بردار کلمات کلیدی نامزد، محاسبه شده و مشابه‌ترین کلمات، استخراج می‌شوند [۱٤].

[1] Benjio
[2] Collobert
[3] Mikulov
[4] Jeffrey Pennington
[5] Haarman
[6] Bennani-Smires
[7] Doc2Vec
[8] Sent2vec

۳- روش پیشنهادی

هدف اصلی پژوهش استخراج کلمات کلیدی از آگهی می‌باشد. این واژگان از ترکیب‌های دوکلمه یا بیشتر تشکیل شده‌اند که در این مقاله به تشخیص واژگان دوکلمه‌ای پرداخته‌شده است.

نوآوری استفاده شده در این پژوهش، ساخت مدل اختصاصی وُردتووِک و ترکیب آن با روش‌های آماری مانند جایگاه کلمه و بسامد کلمه در سند می‌باشد. مدل اختصاصی روی کلمات دوتایی از آگهی‌های سایت ایده‌کاو، آموزش دیده است.

در این بخش ابتدا به معرفی دادگان و پیش‌پردازش می‌پردازیم. در بخش ۳-۳ مراحل استخراج کلمات کلیدی و در بخش ۳-٤ نحوه آموزش وُردتووِک شرح داده می‌شود. در بخش‌های ۳-۵ الی ۳-۷ روش‌های ترکیب شده جهت امتیازدهی کلمات توضیح داده خواهد شد و در انتها شبه کدی از کل روش‌های پیاده‌سازی شده بیان می‌شود.

۳-۱ مجموعه داده

در این پژوهش از مجموعه آگهی‌های سامانه ایده‌کاو به تعداد ۲۰۸،۰۳۸ عدد که از گروه‌ها و کانال‌های تلگرامی جمع‌آوری شده‌اند، استفاده شده است. برای بالابردن کیفیت ارزیابی نتایج، فقط از آگهی‌های فارسی و حوزه املاک استفاده کرده‌ایم. جهت پالایش، فقط آگهی‌های حوزه املاک انتخاب شده‌اند. پس از پالایش، تعداد آگهی‌های حوزه املاک به تعداد ۱۰،۵۳٤ عدد رسیده است.

۳-۲ پیش‌پردازش

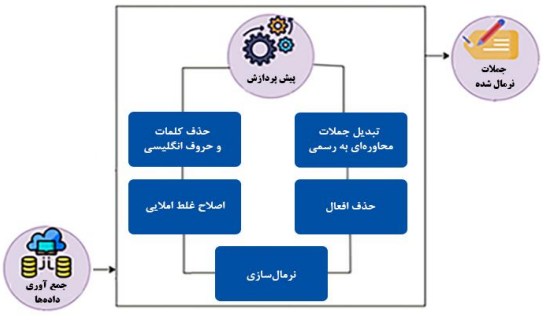
در پیام‌رسان تلگرام، افراد عموماً به سبک محاوره‌ای و مکالمه‌ای تبادل اطلاعات می‌کنند، ازاین‌رو پیش‌پردازش یکی از عوامل مؤثر در افزایش کیفیت داده‌ها محسوب می‌شود [۱۵].

پردازش زبان فارسی تا حدودی با پردازش زبان انگلیسی متفاوت است. در زبان انگلیسی هر حرف و هر کلمه، جدا و به شکلی خاص نوشته می‌شود، درحالی‌که در فارسی برخی از حروف به هم وصل و برخی جدا نوشته می‌شوند. اگر حروف، علائم نگارشی و کلمات فارسی به شکل‌های متفاوتی نوشته شوند، سیستم پردازشی قادر به تشخیص یکسان بودن کلمات نخواهد بود. برای مثال در صورت عدم پیش‌پردازش، سیستم دو کلمه «خانه» و «خونه» را به‌عنوان دو کلمه مجزا در نظر می‌گیرد که به طور قابل‌توجهی نتایج را تحت‌تأثیر قرار می‌دهد [۱۶].

جهت پیش‌پردازش متون، از کتابخانه متن‌کاوی استفاده‌شده است. در این پیش‌پردازش اقدامات زیر انجام شده‌اند:

- حذف کلمات و حروف انگلیسی، اعداد، آدرس ایمیل و شماره موبایل
- حذف علائم نگارشی و شکلک‌ها
- حذف فاصله‌های اضافی
- حذف لینک‌های تبلیغات
- تبدیل جمله‌های محاوره‌ای به رسمی
- اصلاح غلط‌های املایی
- حذف افعال

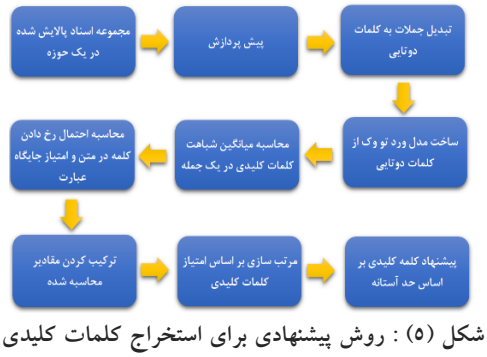
شکل ٤ نمای کلی از گردش کار در مرحله پیش‌پردازش را نشان می‌دهد.



شکل (٤) : گردش کار استخراج واژگان کلیدی

۳-۳ مراحل استخراج کلمات کلیدی

در این پژوهش، روشی برای استخراج کلمات کلیدی بر پایه وُردتووِک ارائه شده است. سامانه پیشنهادی شامل شش مرحله اصلی پیش‌پردازش، آموزش اختصاصی وُردتووِک روی کلمات دوتایی، محاسبه احتمال رخدادن کلمه در متن، محاسبه فاصله کلمات، محاسبه جایگاه کلمه در متن و سپس تعیین امتیاز برای هر کلمه هست که روال کلی آن در شکل ۵ نشان‌داده‌شده است.



شکل (۵) : روش پیشنهادی برای استخراج کلمات کلیدی

در روش پیشنهادی بعد از پیش‌پردازش مجموعه‌داده‌ها، مدل وُردتووِک روی کلمات دوتایی هر آگهی ساخته و میانگین شباهت کسینوسی هر کلمه با سایر کلمات محاسبه شده‌است. یکی دیگر از معیارهای مهم در شناسایی کلمات کلیدی، تعداد تکرار کلمات است اما این مورد به‌تنهایی نقشی در انتخاب کلمه کلیدی ندارد، زیرا ممکن است یک کلمه با بسامد کم، معنی اصلی متن را مشخص کند. به همین دلیل از مدل وُردتووِک برای ساخت بردار تعبیه کلمات استفاده شده است. این مدل، معنا و مفهوم کلمات را به همراه وابستگی بین کلمات نیز درک می‌کند. در استخراج کلمات کلیدی، کلمه‌ای که فاصله کمتری با سایر کلمات دارد و در اطراف آن تراکم کلمات بیشتر است، انتخاب می‌شود. بـه همـین علـت روش پیشنهادی روشی پویا است و قادر به استخراج هر نوع کلمه کلیدی است. در این پژوهش برای بهبود استخراج کلمات کلیدی، با روش‌های آماری نیز ترکیب شده است که در ادامه به توضیح هر کدام می‌پردازیم.

۳-۴ آموزش وُردتووِک روی مجموعه‌داده

در این قسمت به نحوهٔ آموزش مدل وُردتووِک روی کل آگهی‌های املاک می‌پردازیم. مدل‌های مرسوم ساخته شده همگی روی عبارت‌های تک‌کلمه‌ای آموزش‌دیده‌اند. مدلی که در این پژوهش آماده کرده‌ایم روی عبارت‌های دو کلمه‌ای آموزش‌دیده‌است. جهت ایجاد عبارت‌های دو کلمه‌ای از کتابخانه شمارش بردار[1] استفاده کرده‌ایم. بعد از تقسیم جملات به عبارت‌های دو کلمه‌ای، مدل وُردتووِک را با کمک کتـابخانه جنسیم[2] آموزش داده‌ایم. مدل ساخته شده بر اساس حالت اسکیپ-گرام آموزش دیده است.

اندازه بردار برای هر کلمه در این مدل برابر با ۴۰۰ می‌باشد. نمونه‌ای از مدل ساخته شده در شکل ٦ قابل مشاهده است.

شکل (٦) : مدل برداری وُردتووِک بر روی مجموعه داده

۳-۵ احتمال فراوانی کلمات در متن

کاربران معمولاً در نوشته‌های خود از کلمات تکراری استفاده می‌کنند. دلیل استفاده از فراوانی کلمات به این دلیل است که کاربران معمولاً برای راحتی و تسریع در نوشتار از کلمات خاصی برای نوشتن یا صحبت‌های خود استفاده می‌کنند. در این پژوهش فراوانی کلمات دوتایی را شمارش شده است. محاسبه فراوانی کلمات یکی از ویژگی‌های مؤثر در انتخاب کلمات کلیدی به شمار می‌رود.

احتمال فراوانی کلمه، به‌تنهایی نمی‌تواند در استخراج کلمات کلیدی تأثیرگذار باشد و باید با سایر روش‌های ذکر شده ترکیب شود تا بهترین نتیجه ایجاد شود.

جهت محاسبه احتمال فراوانی کلمه از رابطه (۲) استفاده می‌کنیم که عبارت $count(w)$ بیانگر تعداد تکرار عبارت و c بیانگر تعداد کل کلمات متن است.

$$f(w) = count(w)/c \qquad (۲)$$

۳-٦ امتیاز جایگاه کلمه در متن

اکثر افراد برای تأثیرگذاری بیشتر آگهی‌شان در ابتدای متن خود، کلمات مهم‌تر و پایه‌ای را می‌نویسند، به همین جهت ترکیب‌کردن امتیاز جایگاه کلمات با روش‌های ذکر شده باعث بهبود عملکرد می‌شود. به طور مثال به متن آگهی زیر توجه کنید:

«پراید دوگانه‌سوز فروشی. بدون رنگ. با شماره زیر تماس بگیرید.»

می‌بینیم که برچسب پراید دوگانه‌سوز در ابتدای جمله حاضر شده است و رفته‌رفته بر تأثیر کلمات دوتایی کاسته می‌شود.

[1] CountVectorizer

[2] Gensim

در این پژوهش از معکوس جایگاه هر عبارت دوکلمه‌ای، استفاده شده است. این عملکرد از رابطه (۳) محاسبه می‌شود.

$$p(w) = \frac{1}{position(w) + 1} \quad (۳)$$

در رابطه (۳) مقدار $position(w)$ به جایگاه عبارت دو کلمه‌ای در متن اشاره می‌کند. این مقدار از ۱ شروع می‌شود و به تعداد کلمات موجود در متن ادامه پیدا می‌کند.

۳-۷ محاسبه شباهت کلمات

پس از محاسبه جایگاه کلمات و احتمال فراوانی عبارت دو کلمه‌ای در متن، شباهت عبارت با تمامی عبارت‌های موجود در متن، با استفاده از بردار وُردتووِک و شباهت کسینوسی، محاسبه می‌شود. جهت محاسبه میانگین شباهت کسینوسی عبارت با تمام عبارت‌های موجود در متن، از معادله (٤) استفاده می‌شود.

$$sim\ (w) = \frac{\sum_i^n (1 - (w \cdot w_i)/|w| \cdot |w_i|)}{n} \quad (٤)$$

شباهت کسینوسی، برای تعیین شباهت اسناد استفاده می‌شود. از نظر ریاضی شباهت کسینوسی، کسینوس زاویه بین دو بردار پیش‌بینی‌شده در یک فضای چندبعدی است. در رابطه (٤) مقدار i به‌ازای تمام عبارت‌های متن ورودی تغییر می‌کند. مقدار w بیانگر کلمات دوتایی هست که قصد داریم امتیاز آن را محاسبه کنیم و w_i بیانگر سایر عبارت‌های منتخب است. مقدار n در این الگوریتم بیانگر تعداد کل عبارت‌های موجود در جمله است.

پس از محاسبه $sim(w)$ ،$f(w)$ و $p(w)$ امتیاز کلمه w مطابق رابطه (۵) محاسبه می‌شود.

$$score(w) = f(w) + p(w) + (2 * sim(w)) \quad (۵)$$

جهت تأثیرگذاری بیشتر مقدار وُردتووِک، مقدار $sim(w)$ را در ۲ ضرب کرده‌ایم. امتیاز تمامی عبارت‌ها محاسبه شده و بر اساس بیشترین امتیاز مرتب می‌شوند. در این پژوهش از حد آستانه‌های ۵، ۱۰ و ۱۵ برای انتخاب کلمات کلیدی استفاده شده است. برای بیان واضح‌تر مراحل ذکر شده، تمامی الگوریتم‌ها در شبه‌کد ۱ بیان شده‌اند.

Pseudo-code of Keyword Extraction

Input: $\boldsymbol{U}$ (sentence)
Output: $\boldsymbol{R}$ (list of keywords)
Variables:
var $\boldsymbol{u_i}$: each bigram word in $\boldsymbol{U}$
var $\boldsymbol{M}$: model of word2vec
var $\boldsymbol{F}$: frequency of bigram
var $\boldsymbol{P}$: position of bigram
var $\boldsymbol{D}$: dictionary of bigram with sum of score
var $\boldsymbol{S}$: similarity of vector
var $\boldsymbol{score}$: final score for bigram

Pseudo-code:
for each bigram $\boldsymbol{u_i}$ in $\boldsymbol{U}$:
  $\boldsymbol{P}$=1/(position+1) of $\boldsymbol{u_i}$ in $\boldsymbol{U}$;
  $\boldsymbol{F}$=count($\boldsymbol{u_i}$)/count of all word in $\boldsymbol{U}$;
  set $\boldsymbol{S}$=0;
  for each bigram $\boldsymbol{u_j}$ in $\boldsymbol{U}$:
    if $\boldsymbol{u_i}$ in $\boldsymbol{M}$ AND $\boldsymbol{u_j}$ in $\boldsymbol{M}$:
      set $\boldsymbol{S}$ += cosine similarity ($\boldsymbol{u_i}, \boldsymbol{u_j}$);
  set $\boldsymbol{score} = \boldsymbol{S} + \boldsymbol{F} + \boldsymbol{P}$;
  $\boldsymbol{D}$.append({$\boldsymbol{u_i}$: $\boldsymbol{Score}$});
set $\boldsymbol{R}$ = sort ($\boldsymbol{D}$);
return $\boldsymbol{R}$;

شبه‌کد(۱) : روش استخراج کلمات کلیدی

٤- آزمایش‌های تجربی

برای حفظ اطلاعات و مدیریت بهتر نتایج، وب‌سرویسی بر پایه جنگو طراحی شده‌است. جهت ارزیابی، از روش‌های آماری همچون TF-IDF، YAKE، عبارت اول، تکست‌رنک، پوزیشن‌رنک و تاپیک‌رنک و از روش‌های درون‌سازی همچون پارس‌برت، ویکی پارسی و مدل آی‌کِی استفاده شده‌است. جهت ارزیابی مدل آی‌کِی، جملات به عبارت‌های دوتایی و جهت ارزیابی سایر روش‌ها، جملات به تعداد کلمات تشکیل‌دهنده تقسیم شده‌اند. برای ساخت مدل آی‌کِی، به آموزش کلمات کلیدی دو کلمه‌ای و تک کلمه‌ای پرداخته شده است که نتایج مدل دو کلمه‌ای، بهبود قابل توجهی نسبت به مدل تک کلمه‌ای داشته است. جهت سهولت و دقت در ارزیابی، داده‌ها در پایگاه‌داده ذخیره شده و بعد از پیش‌پردازش، ارزیابی روی آن‌ها انجام شده‌است.

٤-١ معیارهای ارزیابی

در این قسمت به معرفی معیارهای ارزیابی می‌پردازیم. معیارهای ارزیابی در اغلب تحقیق‌ها شامل دقت، صحت، فراخوانی و امتیاز F1 هستند. این معیارها توسط چهار مؤلفه ماتریس اغتشاش محاسبه می‌شوند. هر یک از چهار مؤلفه عبارت‌اند از: مثبت‌های درست، منفی‌های درست، مثبت‌های نادرست و منفی‌های نادرست. با توجه به رابطه (٦)، جهت محاسبه مقدار دقت، نسبت عبارت‌هایی که به‌درستی به‌عنوان کلمه کلیدی تشخیص داده شده‌اند به همهٔ عبارت‌هایی که به‌عنوان کلمه کلیدی تشخیص داده شده‌اند، محاسبه می‌شود.

$$\frac{TP}{TP+FP} \qquad (٦)$$

نسبت عبارت‌هایی که به‌درستی به‌عنوان کلمه کلیدی تشخیص داده شده‌اند به کل کلمات کلیدی با عنوان مقدار فراخوانی، از رابطه (٧) محاسبه می‌شود.

$$\frac{TP}{TP+FN} \qquad (٧)$$

در نهایت جهت محاسبه امتیاز F_1 از رابطه (٨) استفاده می‌کنیم. این رابطه مشخص می‌کند دقت و فراخوانی در مجموع به چه صورت هستند.

$$\frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (٨)$$

علاوه بر معیار امتیاز F1، معیار افت همینگ نیز محاسبه شده است. معیار افت همینگ، معیار ارزیابی برای داده‌های چند برچسبی است. این معیار با این فرض که هر برچسب یک اهمیت مساوی دارد، نسبت برچسب‌هایی را که به اشتباه پیش‌بینی شده‌اند به کل تعداد برچسب‌ها محاسبه می‌کند. به عبارتی دیگر، نمونه‌ای که به برچسب نادرست نسبت‌داده‌شده باشد یا برچسب صحیحی که به یک نمونه متعلق باشد اما پیش‌بینی‌نشده باشد نسبت به کل برچسب‌های موجود محاسبه می‌شود.

این معیار برای هر نمونه ورودی تعداد برچسب‌های نادرست را شمارش کرده و سپس با تقسیم بر تعداد کل برچسب‌ها، خطای همینگ را به‌عنوان افت همینگ ارائه می‌دهد. فرمول این معیار از رابطه (٩) محاسبه می‌شود.

$$Hamming\ Loss(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \Delta\, \hat{y}_i|}{L} \qquad (٩)$$

در رابطه (٩) عملگر N تعداد کل نمونه‌ها و L تعداد کل برچسب‌ها را نمایش می‌دهد. y_i بیانگر برچسب واقعی و $\hat{y}$ بیانگر برچسب پیش‌بینی شده نمونه i ام است. $|y_i \Delta\, \hat{y}_i|$ تابع نشانگر است که اگر برچسب‌های واقعی و پیش‌بینی شده برای نمونه iم یکسان باشند ١ را برمی‌گرداند و در غیر این صورت ٠ را بر می‌گرداند.

٤-٢ معرفی روش‌های استفاده شده در ارزیابی

برای ارزیابی کیفیت برچسب‌زنی اسناد و ارزیابی هر یک از الگوریتم‌های درون‌سازی، ٤٠٠ سند به‌صورت تصادفی انتخاب و توسط فرد خبره برچسب‌گذاری شده است. برای هر آگهی، برچسب‌ها به‌صورت چند برچسبی انتخاب شده‌اند. در مدل چند برچسبی، برای هر آگهی می‌توان به تعداد برچسب‌های موردنیاز، برچسب اضافه کرد. به طور مثال برای آگهی «فروش املاک با نازل‌ترین قیمت. جهت هماهنگی، با املاکی سهند تماس حاصل فرمایید»، دو برچسب فروش املاک و املاکی سهند انتخاب می‌شود. جهت ارزیابی، روش آی‌کی با روش‌های بدون ناظر استخراج کلمات کلیدی همچون TF-IDF، YAKE، عبارت اول، تکست‌رنک، پوزیشن‌رنک، تاپیک‌رنک، برت و وُردتووِک مقایسه شده است که نتایج روش پیشنهادی عملکرد بهتری را ثبت کرده است.

٤-٣ ارزیابی روش‌های استخراج کلمات کلیدی

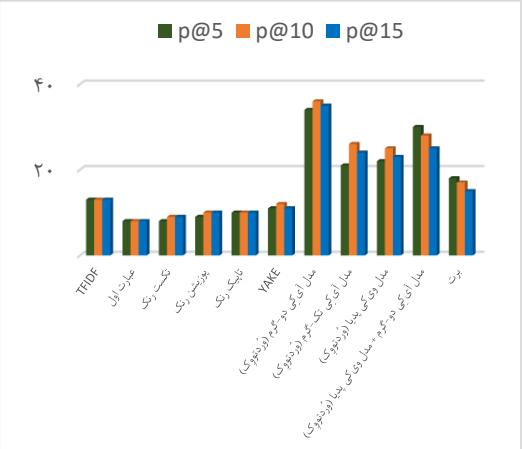
جهت بررسی عملکرد هر یک از روش‌های ذکر شده، آزماش‌هایی را روی آگهی‌های تلگرام انجام داده‌ایم. ارزیابی انجام شده با استفاده از دو معیار امتیاز F1 و افت همینگ انجام شده است.

جهت انجام آزمایش‌ها از سه آستانه ٥، ١٠ و ١٥ استفاده کرده‌ایم. در جدول ١ درصد امتیاز ماکرو-F1 روش‌های مختلف ثبت شده است. با توجه به نتایج، مدل آی‌کی دوکلمه‌ای (وُردتووِک) عملکرد بهتری در مقایسه با روش‌های درون‌سازی دیگر و روش‌های آماری داشته است.

جدول (۱) : مقایسه‌ی F1 روش‌های پیشنهادی

نام روش	ماکرو-F1		
	p@5	p@10	p@15
TF-IDF	۱۳	۱۳	۱۳
عبارت اول	۸	۸	۸
تکست‌رنک	۸	۹	۹
پوزیشن‌رنک	۹	۱۰	۱۰
تاپیک‌رنک	۱۰	۱۰	۱۰
YAKE	۱۱	۱۲	۱۱
آی‌کِی دو-گرم (وُردتووِک)	**۳۴**	**۳۶**	**۳۵**
آی‌کِی تک-گرم (وُردتووِک)	۲۱	۲۶	۲۴
مدل وی‌کی پدیا (وُردتووِک)	۲۲	۲۵	۲۳
مدل آی‌کِی دو-گرم + مدل وی‌کی پدیا (وُردتووِک)	۳۰	۲۸	۲۵
برت	۱۸	۱۷	۱۵

برای مقایسه و نمایش بهتر ارزیابی‌های انجام شده، اطلاعات را در قالب نمودار در سه آستانه ۵، ۱۰ و ۱۵ در شکل ۷ ترسیم کرده‌ایم.



شکل (۷) : نمودار ارزیابی روش‌های پیشنهادی

در جدول ۲ مقدار افت همینگ را برای تمامی معیارهای ارزیابی محاسبه کرده‌ایم که نشان می‌دهد بهترین مقدار افت همینگ به معیار مدل آی‌کِی دوکلمه‌ای با حد آستانه ۱۰ متعلق است.

جدول (۲) : مقایسه‌ی افت همینگ روش‌های پیشنهادی

نام روش	افت همینگ		
	p@5	p@10	p@15
TF-IDF	۰/۰۲۴۲	۰/۰۲۴۲	۰/۰۲۴۲
عبارت اول	۰/۰۲۷۱	۰/۰۲۷۱	۰/۰۲۷۱
تکست‌رنک	۰/۰۲۷۳	۰/۰۲۶۹	۰/۰۲۶۹
پوزیشن‌رنک	۰/۰۲۶۷	۰/۰۲۶۷	۰/۰۲۶۷
تاپیک‌رنک	۰/۰۲۶۸	۰/۰۲۶۸	۰/۰۲۶۸
YAKE	۰/۰۲۴۹	۰/۰۲۴۳	۰/۰۲۴۸
مدل آی‌کِی دو-گرم (وُردتووِک)	۰/۰۱۹۲	**۰/۰۱۸۴**	۰/۰۱۸۷
مدل آی‌کِی تک-گرم (وُردتووِک)	۰/۰۲۲۰	۰/۰۱۹۵	۰/۰۲۰۱
مدل وی‌کی پدیا (وُردتووِک)	۰/۰۲۱۵	۰/۰۲۰۲	۰/۰۲۰۵
مدل آی‌کِی دو-گرم + مدل وی‌کی پدیا (وُردتووِک)	۰/۰۱۹۳	۰/۰۱۹۶	۰/۰۲۰۴
برت	۰/۰۲۳۱	۰/۰۲۳۲	۰/۰۲۴۰

با توجه به آمارهای به‌دست‌آمده، الگوریتم وُردتووِک نسبت به سایر روش‌های استخراج کلمات کلیدی کیفیت بهتری داشته است. در میان روش‌های درون‌سازی، روش مدل آی‌کِی که با آموزش عبارت‌های دو کلمه‌ای آموزش‌دیده است کیفیت بالاتری داشته است. بهترین حد آستانه تعلق دارد به حد آستانه ۱۰، که از ماکرو-F1 و افت همینگ بهتری برخوردار است.

یکی دیگر از مواردی که در استخراج کلمات کلیدی مفید است، مدت زمانی هست که یک الگوریتم نیاز دارد تا یک کلمه کلیدی را استخراج کند. در جدول ۳ مقایسه زمان برای استخراج کلمات کلیدی انجام شده است. این میانگین زمان بر حسب میلی‌ثانیه بوده و مشخص‌کننده زمان طول‌کشیدن برای یافتن یک کلمه کلیدی است. در ارزیابی‌های انجام شده بهترین زمان ثبت شده برای مدل تکست‌رنک و در روش‌های درون‌سازی، بهترین زمان ثبت شده متعلق به مدل آی‌کِی دو کلمه‌ای است.

جدول (۳) : مدت زمان طول کشیدن استخراج یک کلمه

نام روش	مدت‌زمان (میلی‌ثانیه)
TF-IDF	۳۳۸
عبارت اول	۱۲۵
تکست‌رنک	**۱۲۴**
پوزیشن‌رنک	۱۳۳
تاپیک‌رنک	۱۴۳
YAKE	۲۵۷
مدل آی‌کِی دو-گرم (وُردتووِک)	۱۵۴
مدل آی‌کِی تک-گرم (وُردتووِک)	۱۵۶
مدل وی‌کی پدیا (وُردتووِک)	۲۶۰
مدل آی‌کِی دو-گرم + مدل وی‌کی پدیا (وُردتووِک)	۲۵۰
برت	۶۰۰

۵- نتیجه‌گیری و پژوهش‌های آتی

به‌طورکلی، استخراج کلمات کلیدی از داده‌های متنی، یکی از مهم ترین و پرکاربردترین مسائل در متن‌کاوی و بازیابی اطلاعات بوده است که در این مورد تشخیص کلمات کلیدی نقش تعیین‌کننده‌ای در پردازش اطلاعات دارد.

اغلب روش‌های استخراج کلمات کلیدی، روی عبارت‌های فارسی نتیجه مطلوبی ندارند. یکی دیگر از مشکلات، استفاده از افراد از کلمات عامیانه در آگهی‌ها هست که تشخیص کلمات کلیدی را با چالش همراه می‌کند.

روش پیشنهادی در این پژوهش، روشی ترکیبی از مدل‌های آماری و درون‌سازی متن جهت استخراج کلمات کلیدی است. این روش نیاز به دانش خارجی نداشته و با داده‌ای بدون برچسب به‌خوبی کار می‌کند. با توجه به اینکه روش‌های درون‌سازی معنا و مفهوم هر کلمه را تشخیص می‌دهند، امتیاز هر کلمه را با توجه به مکانش محاسبه می‌کنند. روش‌های درون‌سازی توانایی خوبی در تشخیص کلمات عامیانه دارند. از نکاتی که در روش‌های پیشنهادی به آن توجه شده است استخراج کلمات دوکلمه‌ای بوده است زیرا اکثر کلمات کلیدی از دو کلمه و بیشتر تشکیل می‌شوند. از نکات دیگری که در این پژوهش استفاده شده است، امتیاز دهی به واژگان ابتدای جمله است زیرا مفهوم اصلی در آگهی‌ها معمولا در ابتدا بیان می‌شود. و مورد آخر با توجه به اینکه اکثر آگهی‌ها به زبان عامیانه نوشته می‌شوند، ترکیب روش‌های درون‌سازی با روش‌های آماری کیفیت خوبی را در تشخیص کلمات کلیدی به وجود آورده‌است.

پژوهش‌های انجام شده در این پژوهش مربوط به آگهی‌های تلگرام در حوزه املاک بوده‌اند. در آینده با گسترش استخراج کلمات کلیدی در حوزه‌های دیگر مانند خودرو، کاریابی، لوازم‌خانگی، لوازم الکترونیکی و غیره می‌توانیم به مدل عمومی برای استخراج کلمات کلیدی دست پیدا کنیم تا مدل بتواند در تمام دسته‌بندی‌ها به‌خوبی عمل کند.

از دیگر اقداماتی که در آینده انجام می‌شود، ساخت مدل درون‌سازی از عبارت‌های سه و چهار کلمه‌ای می‌باشد تا سامانه بتواند عبارت‌های پویا و متناسب را استخراج کند.

هدف نهایی، رسیدن به یک مدل عمومی برای استخراج کلمات کلیدی از هر نوع و دسته می‌باشد. از اقداماتی که می‌توان انجام داد، تشخیص موضوع و دسته‌بندی آگهی و استخراج کلمه کلیدی متناسب با دسته است. به طور مثال آگهی حوزه خودرو توسط مدل خودرو و آگهی بدون دسته با مدل عمومی باید استخراج شود. این اقدام در بالابردن کیفیت کلمات استخراج شده مؤثر است.

سپاس‌گزاری

سپاس خداوند یکتای عزت‌مندی که رحمت و دانش او در سراسر گیتی گسترده شده، آسمان‌ها و زمین همه از آن اوست و علم و دانش حقیقی را بر هر که بخواهد موهبت می‌فرماید. رحمت و لطف او را بی‌نهایت سپاس می‌گویم چرا که فهم و درک مطالب این پژوهش را بر من ارزانی داشت و مرا به این اصل رساند که علم و ایمان دو بال یک پروازند. توفیق تلاش به من داد و هر بار که خطا کردم فرصتی دوباره، تا با امید، تلاشی تازه را آغاز می‌کنم و به خواست او به نتیجه مطلوب نائل آیم. به‌راستی که همه چیز از آن اوست و همه چیز به خواست اوست. همچنین به طور ویژه

از جناب آقای دکتر محمدعلی زارع چاهوکی بابت ارشادات ارزنده، ثمربخش و دلسوزانه کمال امتنان را دارم.

References

[1] A. Onan, S. Korukoğlu, and H. Bulut, “Ensemble of keyword extraction methods and classifiers in text classification,” Expert Syst. Appl., vol. 57, pp. 232–247, 2016.
[2] “Tagging posts properly for users and SEO,” Yoast, 11-Apr-2019. [Online]. Available: https://yoast.com/tagging-posts-properly-for-users-and-seo/. [Accessed: 23-Jan-2022].
[3] Williams, WordPress for beginners 2021: A visual step-by-step guide to mastering WordPress. Independently Published, 2020.
[4] Moz, “What are keywords?,” Moz, 28-Mar-2017. [Online]. Available:https://moz.com/learn/seo/what-are-keywords. [Accessed: 23-Jan-2022].
[5] “Telegram revenue and usage statistics (2022),” Business of Apps, 08-Aug-2017. [Online]. Available: https://www.businessofapps.com/data/telegram-statistics. [Accessed: 23-Jan-2022].
[6] H. Tankovska, “Most popular global mobile messenger apps as of January 2023, based on number of monthly active users,” 20022. [Online]. Available: https://www.statista.com/statistics/258749/most-popular global-mobilemessenger-apps/. [Accessed: 2- May- 2023].
[7] Y. Sun, H. Qiu, Y. Zheng, Z. Wang, and C. Zhang, “SIFRank: A new baseline for unsupervised keyphrase extraction based on pre-trained language model,” IEEE Access, vol. 8, pp. 10896–10906, 2020.
[8] Won, M., Martins, B. and Raimundo, F. (2019) Automatic extraction of relevant keyphrases for the study of issue competition. In Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, Berkeley, La Rochelle, France, April 7-13, 2019.
[9] C. Florescu and C. Caragea, “PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017.
[10] R.Ruder,"word-embeddings-1,"1142016.[Online].Available: https://www.ruder.io/word-embeddings-1/.
[11] Y. Goldberg, Neural network methods in natural language processing. San Rafael, CA: Morgan & Claypool, 2017.
[12] Kedia and M. Rasu, Hands-On Python Natural Language Processing: Explore tools and techniques to analyze and process text with a view to building real-world NLP applications. Birmingham, England: Packt Publishing, 2020.
[13] T. Haarman, B. Zijlema, and M. Wiering, “Unsupervised keyphrase extraction for web pages,” Multimodal technol. interact., vol. 3, no. 3, p. 58, 2019.
[14] K. Bennani-Smires, C. Musat, A. Hossmann, M. Baeriswyl, and M. Jaggi, “Simple Unsupervised Keyphrase Extraction using Sentence Embeddings,” in Proceedings of the 22nd Conference on Computational Natural Language Learning, 2018.
[15] M. Saraswathi and V. Balu, "Preprocessing Techniques for Effective Data Extraction and Computation," IUP Journal of Computer Sciences, Volume 7(3), p. 27, 2013.
[16] O. Hajipoor, et al., “Determine the Sentiment for Persian Words and Phrases Using Deep Learning,” Computer Society of Iran Conference, Volume 24, 2019.
[17] Harris Hawks’ Optimization algorithm," Knowledge-Based Systems,

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 2, No. 1, Spring & Summer 2024, pp. 39-50

Extracting keywords from Telegram ads based on the integration of word embedding and statistical features

Mohammad Ali Zare Chahooki*, Mohammad Mahdi Sedghian

Computer Engineering Department, Yazd University, Yazd, Iran

Article Information

Original Research Paper

Received:
2023 November 16

Accepted:
2024 April 8

Keywords:
Tag extraction, search engine optimization (SEO), deep learning, word embedding

Corresponding Author*:
chahooki@yazd.ac.ir

Abstract

In today's society, news and advertisements have a special place in people's communication with each other. By specifying the main words of the ad, you can understand the general concept. One of these uses of keywords is to use them in SEO. Keywords indicate the main purpose of the ad. Since keywords have many uses in recognizing the meaning of text, identifying automatic and improved methods for extracting this category of words has always been of interest to researchers. In terms of SEO, keywords are queries searched by users, for this reason, keyword extraction will play an effective role in increasing site SEO.Ideakav site is a system that collects Telegram messages and ads. According to the requirement of the idekav system in extracting keywords from advertisements published in Telegram, in this research we seek to provide a method that can identify and extract keywords. The quality of extracted keywords plays a significant role in improving SEO and advertising statistics. The data used in this research is from the data collection of Ideakav system. The approach of the current research is an unsupervised integrated method for extracting keywords that uses a combination of statistical methods and word internalization to extract keywords. The word embedding algorithm is able to understand colloquial conversations and the semantic structure of the text, therefore, it is useful in identifying keywords in Telegram ads that are often published in popular form. The obtained results show that the combination of the two-word model word embedding method along with the statistical methods presented in this research has produced a better performance in extracting keywords than other proposed methods.

: 10.22034/ABMIR.2024.20879.1041

E-ISSN: 2821-2037