اولین کنفرانس بین المللی و ششمین کنفرانس ملی
کامپیوتر، فناوری اطلاعات و کاربردهای هوش مصنوعی
۳ اسفندماه ۱۴۰۱

استخراج کلمات کلیدی از متون فارسی با استفاده از مکانیسم توجه و شبکه‌های یادگیری عمیق

مصطفی قربان زاده، دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع)، ghorbanzade@ihu.ac.ir

دکتر محمدرضا حسنی آهنگر، هیئت علمی دانشکده کامپیوتر دانشگاه جامع امام حسین (ع)، Mrhasani@ihu.ac.ir

دکتر محسن نوروزی، پژوهشگر دانشکده کامپیوتر دانشگاه جامع امام حسین (ع)، m.norouzi@ihu.ac.ir

چکیده

استخراج عبارت‌های کلیدی همواره یکی از وظیفه‌های پراهمیت و دشوار در پردازش زبان طبیعی بوده که هم به‌خودی‌خود و هم به‌عنوان یک وظیفهٔ بالادست در وظایفی چون بازیابی اطلاعات، خلاصه‌سازی متن، دسته‌بندی متن و دیگر وظیفه‌ها کارایی دارد. بررسی مقالات کار شده روی زبان فارسی در این وظیفه نشان می‌دهد که توجه کمتری به روش‌های نوین هوش مصنوعی و شبکه عصبی در حل این مسئله شده است بنابراین در این پژوهش نشان داده شد چگونه با استفاده از شبکه‌های عصبی عمیق و استفاده از مدل های زبانی می‌توان به درک عمیقی از کلمات رسید و با دقت خوبی کلمات کلیدی متن را استخراج کرد. همچنین اثبات کردیم سازوکار توجه چند سر که نوع تعمیم یافته سازوکار توجه است تا چه اندازه توانسته به حل مسائل کلاس های نامتعادل کمک کند. پارس‌برت که نوع تعمیم یافته برت آموزش دیده شده روی متن فارسی می‌باشد، در درک کلمات فارسی کارآیی و دقت خوبی نسبت به دیگر مدل های زبانی چون word2vec و ... دارد و در این پژوهش مورد استفاده قرار گرفت. برای آموزش شبکه‌های عصبی نیاز به داده فراوان داریم. پس از بررسی تنها و غنی ترین مجموعه‌داده، PerKey مجموعه‌داده جمع آوری شده توسط آقای دوست محمدی بود که به‌صورت عمومی در دسترس قرار دارد. پس از آموزش روش پیشنهادی و ارزیابی نتایج روی سه معیار صحت و بازخوانی و امتیاز اف-۱ در مقایسه با روش‌های پایه آماری، مبتنی بر گراف و همچنین روش یادگیری عمیق دنباله به دنباله مقاله دوست محمدی و همکاران ۲۰۱۹ روش پیشنهادی ما روی ۵ و ۱۰ کلمه کلیدی برتر به ترتیب ۱۶.۰۴ درصد و ۲۲.۳۵ درصد رشد در معیار اف-۱، ۱۴.۰۹ درصد و ۱۳.۹۸ درصد رشد در معیار بازخوانی و همچنین ۲۴.۱۵ درصد و ۳۱.۶۴ درصد رشد در معیار صحت نسبت به شبکه عمیق دنباله به دنباله مقاله دوست محمدی و همکاران ۲۰۱۹ داشته است.

واژه‌های کلیدی: استخراج عبارت‌های کلیدی، مدل زبانی Bert، مکانیسم توجه، یادگیری عمیق

۱- مقدمه

واژه اطلاعات دیگر تنها به نوشته‌هایی چاپی گفته نمی‌شود. این واژه مواد دیگری چون اطلاعات روی فیلم، اسلاید، نوار تلویزیونی و ادوات ذخیره‌سازی کامپیوتری را نیز در بر می‌گیرد. چالش امروزه بشر در راستای کشف راه‌های بهتر برای رساندن سریع اطلاعات به کسانی است که بدان نیاز دارند. آنچه در این میان از اهمیت بسیار بالایی برخوردار است، این موضوع است که با گسترش روزافزون رسانه‌های ذخیره‌سازی الکترونیکی و رسانه‌های ارتباطی، اطلاعات زیاد و جامعی در دسترس هستند. این اطلاعات می‌توانند به‌صورت فایل‌های تصویری، فایل‌های صوتی، اسناد الکترونیکی یا اخبار ارسال شده از یک گروه خبری باشند. با گسترش اسناد الکترونیکی، پیداکردن اسناد موردنظر از بین حجم عظیمی از اطلاعات متنی به‌صورت دستی کاری دشوار و در عمل غیرممکن خواهد بود. حجم کثیری از اطلاعات متنی و اسناد را در نظر بگیرید که می‌خواهیم آن‌ها را دسته‌بندی کرده و هر دسته را در اختیار طالبان آن حوزه قرار دهیم. با این حجم از اطلاعات و رشد روز افزونی که در حوزه تولید و نشر اطلاعات و اخبار وجود دارد در عمل خواندن و دسته‌بندی تمامی محتواهای تولیدی ممکن نیست. یک راه‌حل برای این مشکل استفاده از کلمات کلیدی می‌باشد. به عبارت دیگر کلمات کلیدی نشانگر خلاصه‌ای کوتاه از محتویات داخل سند می‌باشند. استخراج عبارات کلیدی یک کار کاملاً شناخته شده در پردازش زبان طبیعی است که هدف آن استخراج کلمات مربوط به موضوعات اصلی موردبحث در قطعه ورودی متن است. هدف از کار استخراج عبارت کلیدی استخراج عبارات کلیدی صریح از یک متن مشخص است. در این مقاله ما رویکردی ترکیبی از یادگیری عمیق را در نظر می‌گیریم. در بخش بعدی رویکردهای مختلف استخراج و تولید عبارت کلیدی در انگلیسی و فارسی را موردبحث قرار می‌دهیم. در ادامه، ما مشکل خود را تعریف می‌کنیم و رویکرد خاصی را که در این مقاله اتخاذ می‌کنیم، توضیح می‌دهیم. در بخش چهارم، داده‌های آموزش و آزمایش، جزئیات پیاده‌سازی، مدل‌های پایه و معیارهای ارزیابی خود را موردبحث قرار می‌دهیم. سپس، نتایج را موردبحث قرار می‌دهیم. در آخر کار آینده را موردبحث قرار می‌دهیم.

۲- پیشینه

در ابتدا به کارهای کلی که بیشتر برای استخراج کلمات کلیدی و غالباً برای زبان انگلیسی کار شده‌اند، اشاره می‌کنیم و در نهایت پژوهش‌های صورت‌گرفته در زبان فارسی و چالش‌های مربوطه را بررسی خواهیم کرد.

۲-۱- رویکردهای بدون نظارت

مراحل اساسی یک سیستم استخراج عبارات کلیدی بدون نظارت به شرح زیر است.

۱- انتخاب واحدهای واژگانی کاندید بر اساس برخی روش‌های ابتکاری. نمونه‌هایی از این روش‌ها حذف واژگان توقف[1] و انتخاب کلماتی است که به یک بخش خاص از گفتار[2] تعلق دارند.
۲- رتبه‌بندی واحدهای واژگانی کاندید.
۳- تشکیل عبارات کلیدی با انتخاب کلمات از رتبه‌های برتر[3] یا با انتخاب عبارتی با رتبه بالا یا قسمتی از آن‌هایی که دارای امتیاز بالا هستند.

[1] Stop Words
[2] Pos (Part Of Speech)
[3] Top-Ranked

۱-۱-۲- روش‌های مبتنی بر آمار

Tf-Idf مبنای مشترک در این روش‌ها است. در این روش از پیکره متن استفاده می‌شود. اشکال روش‌هایی که از پیکره متن استفاده می‌کنند دشوار بودن جمع‌آوری پیکره مرتبط می‌باشد. ممکن است اصلاً پیکره‌ای با موضوع مورد بررسی وجود نداشته باشد و یا بسیار محدود باشد.

KP-Miner معرفی شده در [۱] یک سیستم استخراج عبارت کلیدی است که از انواع مختلف اطلاعات آماری فراتر از امتیازات Tf و Idf بهره برداری می‌کند. این سیستم یک فرآیند فیلتر کردن کاملاً مؤثر عبارات کاندید را دنبال می‌کند و از یک تابع امتیازدهی مشابه Tf-Idf استفاده می‌کند. به‌ویژه، سیستم آن‌هایی را که با علامت‌های دستوری/واژگان توقف از هم جدا نمی‌شوند، به‌عنوان کاندید نگه می‌دارد، و در عین حال حداقل ضریب فرکانس دیده شده مجاز[۴] و یک ثابت برش[۵] را در نظر می‌گیرد که برحسب تعدادی از کلمات تعریف می‌شود که پس از آن برای اولین بار یک عبارت ظاهر می‌شود. سپس، سیستم عبارات کاندید را با در نظر گرفتن امتیازهای Tf و Idf و همچنین موقعیت کلمه و یک عامل تقویت[۶] برای واژه مرکب بر تک واژه‌ها رتبه‌بندی می‌کند.

اهمیت زیاد استفاده از آمار و اطلاعات زمینه توسط روش‌های اخیر مانند YAKE که توسط [۲] ارائه شده است و روش پیشنهاد شده توسط [۳] تأیید شده است. YAKE، علاوه بر موقعیت/فرکانس عبارت، از معیارهای آماری جدیدی نیز استفاده می‌کند که اطلاعات زمینه و گسترش اصطلاحات را در سند ثبت می‌کند. ابتدا، YAKE متن را با تقسیم کردن آن به عبارات فردی پیش‌پردازش می‌کند. دوم، مجموعه‌ای از پنج ویژگی برای هر عبارت جداگانه محاسبه می‌شود: پوشش[۷] (جنبه حروف یک کلمه را منعکس می‌کند)، موقعیت کلمات[۸] (ارزش بیشتر برای کلماتی که در ابتدای یک سند وجود دارند)، فراوانی کلمه[۹] ، ارتباط کلمه با متن (تعداد اصطلاحات مختلفی را که در سمت چپ/راست کلمه کاندید رخ می‌دهد محاسبه می‌کند) و کلمه درجمله متفاوت[۱۰] (مشخص می‌کند که یک کلمه کاندید چند بار در جملات مختلف ظاهر می‌شود). سپس، تمام این ویژگی‌ها برای محاسبه امتیاز برای هر عبارت استفاده می‌شود. در نهایت، با استفاده از یک پنجره کشویی ۱ ۲ و ۳ تایی، یک دنباله پیوسته از کلمات کلیدی کاندید ایجاد می‌شود. برای هر کلمه کلیدی کاندید امتیازی تخصیص داده می‌شود. علاوه بر این، روش [۳] نشان می‌دهد که با استفاده از ترکیبی از ویژگی‌های آماری متنی ساده می توان به نتایجی دست یافت که با روش‌های پیشرفته رقابت می‌کند. اولین گام این روش، انتخاب عبارات کاندید با استفاده از الگوهای صرف و نحو[۱۱] است. سپس برای هر کاندید ویژگی های زیر محاسبه می‌شود: فرکانس کلمه، یعنی مجموع فراوانی هر کلمه عبارت کاندید، فراوانی سند معکوس، وقوع نسبی اول[۱۲]. امتیاز نهایی هر کاندید حاصل ضرب این چهار ویژگی است.

۲-۱-۲- روش‌های رتبه‌بندی مبتنی بر گراف

ایده اصلی در رتبه‌بندی مبتنی بر گراف، ایجاد یک گراف از یک سند است که عبارات کاندید سند را به‌عنوان گره دارد و هر یال این گراف عبارت‌های کلیدی کاندید مرتبط را به هم متصل می‌کند. هدف نهایی، رتبه‌بندی گره‌ها با استفاده از روش رتبه‌بندی مبتنی بر گراف، مانند رتبه صفحه[۱۳] گوگل [۴]، تابع موقعیتی[۱۴] [۵] و HITS [۶] یا به طور کلی حل یک مسئله بهینه‌سازی پیشنهادی در گراف است. بر اساس روش هایی که در زیر توضیح داده شده است، PageRank با موفقیت برای استخراج عبارت کلیدی مبتنی بر گراف استفاده شده است. رتبه صفحه بر اساس مرکزیت بردار ویژه است و به‌صورت بازگشتی وزن یک راس را به‌عنوان معیاری از تأثیر آن در گراف کلمات، بدون توجه به اینکه همسایگی آن چقدر منسجم است، تعریف می‌کند. با این حال، در [۷] رئوس هسته اصلی را به‌عنوان مجموعه ای از کلمات کلیدی

---

[4] Least Allowable Seen Frequency (Lasf)
[5] Cutoff Constant (Cutoff)
[6] Boosting Factor
[7] Casing
[8] Word Positional
[9] Word Frequency
[10] Word Difsentence
[11] Morphosyntactic
[12] Relative First Occurrence
[13] Pagerank
[14] Positional Function

برای استخراج از سند در نظر می‌گیرند، زیرا با منسجم ترین جزء متصل گراف مطابقت دارد. به همین دلیل، رئوس به طور شهودی کاندیدای مناسبی هستند.

TextRank اولین روش استخراج کلمات کلیدی مبتنی بر گراف بود که توسط [۸] پیشنهاد شد و الهام بخش محققان برای ساخت آن شد، که منجر به روش‌های شناخته شده شد. آنها ایده «رای دادن[15]» را اجرا می کنند: راسی که یک کلمه یا عبارت (واحد واژگانی) را نشان می‌دهد به یک کلمه دیگر پیوند می زند و سپس رای می‌دهد. تعداد رای بیشتر به برخی کلمات یا عبارات نشان می‌دهد که آنها اهمیت بیشتری دارند. همه واحدهای واژگانی متن مبدأ به این ترتیب رتبه‌بندی می‌شوند. عبارات کلیدی بازگشتی از N کلمه برتر ساخته می‌شوند.

SingleRank [۹] توسعه‌یافته TextRank است که وزن ها را به لبه ها اضافه می‌کند. به طور مشابه، برای روش‌های مبتنی بر آمار، آمارهای هم‌رخداد اطلاعات مهمی در مورد زمینه‌ها هستند. از این رو، وزن هر یال برابر است با تعداد همزمانی دو کلمه متناظر. سپس، تابع امتیاز برای یک گره به روشی مشابه محاسبه می‌شود. در مرحله پس پردازش، برای هر دنباله پیوسته از اسم ها و صفت ها در سند متن، نمرات کلمات تشکیل دهنده خلاصه می‌شود و کاندیدهای رتبه برتر T به‌عنوان عبارات کلیدی برگردانده می‌شوند.

TopicRank ارائه شده توسط [۱۰] ابتدا متن را برای استخراج عبارات کاندید پیش‌پردازش می‌کند. سپس، عبارات کاندید با استفاده از خوشه بندی انبوه سلسله مراتبی به موضوعات جداگانه‌ای دسته‌بندی می‌شوند. در مرحله بعد، گرافی از موضوعات ساخته می‌شود که یال های آن بر اساس معیاری که عبارات را موقعیت های آفست در متن در نظر می‌گیرد، وزن دهی می‌شود. سپس، از TextRank برای رتبه‌بندی موضوعات استفاده می‌شود و از هر یک از N موضوع مهم، یک کاندید کلمات کلیدی انتخاب می‌شود.

یک روش جدیدتر بسیار شبیه به TopicRank اما پیشرفته‌تر، MultipartiteRank (MR) ارائه شده توسط [۱۱] است که که یک مرحله میانی را معرفی می‌کند که در آن وزن لبه‌ها برای گرفتن اطلاعات موقعیت تنظیم می‌شوند که سوگیری نسبت به کاندیداهای عبارت کلیدی را که قبلاً در سند اتفاق می‌افتند، تنظیم می‌کنند. توجه داشته باشید که ابتکار برای معرفی یک گروه خاص از کاندیدها، به‌عنوان مثال، آن‌هایی که قبلا در متن ظاهر شده‌اند، می‌تواند برای برآوردن شرایط/نیازهای دیگر تطبیق داده شود.

۲-۲- رویکردهای نظارت شده

یادگیری نظارت شده که به‌عنوان یادگیری ماشین نظارت شده نیز شناخته می‌شود، زیرمجموعه‌ای از یادگیری ماشین و هوش مصنوعی است. با استفاده از مجموعه داده‌های برچسب‌گذاری شده برای آموزش الگوریتم‌هایی که برای طبقه‌بندی داده‌ها یا پیش‌بینی دقیق نتایج تعریف شده است.

۲-۲-۱- استخراج کلمات کلیدی با استفاده از نظارت

GenEx و KEA دو سیستم بسیار خوب با روش با ناظر هستند که در سال ۱۹۹۹ توسط [۱۲] مطرح شدند. ویژگی‌های مهم کلمات در این دو روش برای دسته‌بندی به دو گروه کاندید و غیرکاندید، فرکانس تکرار آن‌ها و مکان قرارگیری آن‌ها در سند می‌باشد. ترنی از درخت تصمیم C4.5 برای یادگیری استفاده کرده است و آن را با الگوریتم GenEx که از الگوریتم ژنتیک استفاده کرده است، مقایسه می‌کند. آزمایشات او نشان میدهد که GenEx نتیجه بهتری نسبت به درخت تصمیم C4.5 دارد. در KEA از الگوریتم یادگیری ماشین بیز استفاده شده است. این دو روش مبنای مهمی برای توسعه سایر روش‌ها هستند.

[15] Voting

۲-۲-۲- روش‌های یادگیری عمیق

پژوهش [۱۳] وظیفه استخراج کلمات کلیدی از توییت ها را بررسی می‌کنند. آن‌ها یک مدل شبکه عصبی بازگشتی عمیق (RNN) با دو لایه پنهان پیشنهاد می‌کنند. در لایه پنهان اول، آن‌ها اطلاعات کلید واژه را ثبت می‌کنند، در حالی که در لایه دوم، آن‌ها عبارت کلیدی را بر اساس اطلاعات کلید واژه با استفاده از روش برچسب گذاری توالی استخراج می‌کنند.

پژوهشگران [۱۴] seq2seq یک مدل مولد برای پیش بینی کلمات کلیدی با استفاده از یک چارچوب رمزگذار-رمزگشا[16] را پیشنهاد کرد که معنای مفهومی محتوا را از طریق یک روش یادگیری عمیق (CopyRNN) به دست می‌آورد. ابتدا، داده ها به جفت متن - عبارت تبدیل می‌شوند و سپس یک مدل رمزگذار-رمزگشا RNN برای یادگیری نگاشت از توالی منبع به توالی هدف به کار می‌رود. ایده کلیدی در اینجا فشرده‌سازی متن منبع در یک نمایش لایه پنهان با یک رمزگذار و تولید کلمه کلیدی با یک رمزگشا، بر اساس نمایش لایه پنهان قبلی است. با این حال، اشکالات اصلی این رویکرد این است که کلمات کلیدی تولید شده شامل تکرار هستند (حداقل دو عبارت بیان کننده یک معنی) و پوشش (موضوعات مهم در سند مسائل کلیدی را پوشش نمی دهند)، زیرا همبستگی میان کلمات کلیدی هدف را نادیده می‌گیرد.

۲-۳- زبان فارسی در استخراج کلمات کلیدی

در مقاله [۱۵] یک روش آماری، برای استخراج کلمات کلیدی اسناد فارسی، پیشنهاد شده است. روش پیشنهادی مبتنی برپیکره متنی می‌باشد. ابتدا عمل ریشه‌یابی و حذف کلمات عمومی انجام می‌گیرد. سپس ویژگی‌های آماری برای کلمات مختلف محاسبه شده و با استفاده از فازی‌سازی واعمال قواعد فازی، کلمات کلیدی محتمل، انتخاب می‌شوند. گام بعدی محاسبه رخداد همزمان پیشین و پسین کلمات کلیدی محتمل، با کلمات تکرار شونده، در جملات سند است. با اعمال یک آستانه وفقی[17] روی رخداد همزمان کلمات، کلمات کلیدی دو کلمه‌ای مشخص می‌شوند. بر خلاف اکثر روش‌های آماری که فقط کلمات کلیدی یک کلمه‌ای را استخراج می‌کنند، با استفاده از این روش کلمات کلیدی دو کلمه‌ای نیز استخراج می‌شوند.

در مقاله [۱۶] از شبکه رقابتی LVQ و شبکه عصبی MLP برای استخراج کلمات کلیدی استفاده شده است. این الگوریتم ها در واقع کلمات موجود در یک متن را به دو خوشه کلمات کلیدی وغیر کلیدی دسته‌بندی می‌کنند. برای این منظوراز ۸۰ متن جهت آموزش شبکه‌استفاده شده است. نتایج حاکی از آن است که شبکه عصبی MLP نتیجه بهتری نسبت به LVQ دارد.

در مقاله [۱۷]، نتایج چند روش نظارت‌نشده و نظارت‌شده بر روی وظایف استخراج و تولید عبارت‌های کلیدی در زبان فارسی مقایسه شده است و به این نتیجه رسیده که RNN با سلول‌های GRU می‌تواند به طور قابل‌توجهی از روش‌های دیگر در هر دو کار بهتر عمل کند.

۲-۴- مسائل و چالش‌های موجود در زبان فارسی

در تلاش برای ساخت یک سیستم پردازش و درک متون فارسی با مسائل و مشکلاتی مواجه می‌شویم که برخی در بیشتر زبان‌ها بروز کرده و برخی خاص زبان فارسی می‌باشند. زبان فارسی به دلیل همسایگی با کشورهای عربی زبان دارای کلمات قرضی بسیار و حتی برخی قواعد مشابه با آن می‌باشد. برخی از مهم‌ترین مشکلات فعلی پردازش متون فارسی در زیر آورده شده است[۱۸].

- عدم وجود منابع زبانی مناسب و کافی برای زبان فارسی (مانند هستان‌شناسی‌های لغوی، هستان‌شناسی‌های جامع عمومی و تخصصی، پیکره‌های عمومی و تخصصی ساده یا برچسب خورده)، عدم وجود استانداردهایی برای شیوه نگارش، فاصله‌گذاری و رمزنگاری حروف و علائم.
- تشخیص مرز کلمات به دلیل شیوه‌های نگارشی متفاوت (مانند «کتابخانه و کتاب خانه»، «می‌نویسم و می‌نویسم») دشوار می‌باشد. در زبان انگلیسی این مشکل بیشتر برای اسامی مرکب ایجاد می‌شود ولی در زبان فارسی به دلیل وجود شکل‌های مختلف حروف

[16] Encoder-Decoder
[17] Adaptive

(اول – وسط – آخر و چسبان و غیر چسبان) بیش از زبان انگلیسی مشکل‌ساز است. در زبان فارسی علاوه بر کلمات مرکب که مشکلی مشابه با انگلیسی ایجاد می‌کنند، مرز کلمات غیرمرکب نیز ممکن است به‌درستی تشخیص داده نشود.

- مشکلات رایانه‌ای خط فارسی از قبیل وجود چند مقدار برای حروف «ک» و «ی» که گاهی به فرم «ک» و «ی» عربی نوشته می‌شوند.
- در زبان‌های مختلف افعال مرکب، اصطلاحات، ضرب‌المثل‌ها و استعارات موارد مشکل‌آفرین در پردازش متون هستند. چرا که معمولاً معنایی کاملاً متفاوت با معنای ظاهریشان (ترکیب معانی اجزایشان) دارند. به‌علاوه اجزا چنین ترکیبی لزوماً در جمله به دنبال هم ظاهر نمی‌شوند و ممکن است بین آن‌ها کلمه یا حتی جمله دیگری واقع شود و این یافتن و پردازش سازه مرکب در کل جمله را مشکل می‌کند.
- هم‌نگاره‌ها و تحت آن مسئله حذف مصوت‌های کوتاه از نوشتار (مانند "اشکال" که هم می‌تواند معنی شکل‌ها و هم می‌تواند به معنی مشکل و یا مسئله باشد).
- وجود ابهام معنایی کلمات، مانند کلمه "شیر" که می‌تواند چندین معنا داشته باشد.
- بی ترتیب بودن زبان: اگرچه فارسی دارای ترتیب مرکزی فاعل - مفعول - فعل است ولی استثنائات فراوان و مکرر در ترتیب کلمات وجود دارد. این مسئله باعث می‌شود ساخت دستور زبان مدون و محاسباتی برای زبان و در نتیجه تجزیه و تحلیل نحوی جملات مشکل شود.

۳- مجموعه‌داده

همان‌طور که پیشتر بیان کردیم روش‌های مختلفی برای استخراج کلمات کلیدی وجود دارد. پژوهش حاضر یک رویکرد جدید برای استخراج کلمات کلیدی از متون فارسی می‌باشد. برای این منظور از یک روش مبتنی بر یادگیری عمیق با استفاده از مدل زبانی BERT [۱۹] برای حل این مسئله استفاده کرده‌ایم. مجموعه‌داده مورد استفاده در این پژوهش توسط آقای دوست محمدی و همکاران به نام PerKey [۲۰] جمع آوری شده است. PerKey شامل ۵۵۳ هزار مقاله خبری است که از ۶ وب سایت معتبر خبری جمع آوری شده است. و با نسبت ۸، ۱، ۱ به مجموعه‌داده آموزشی، آزمون و اعتبارسنجی تقسیم می‌شود. تعداد خبرگزاری و وب‌سایت خبری را در جدول ۱ می‌توانید ببینید.

جدول ۱. تعداد خبر بر اساس سایت خبری/خبرگزاری

سایت خبری	تعداد خبر	درصد از کل
خبرآنلاین	۱۶۵/۸۳۹	٪ ۲۹/۹۸
فارو	۱۶۱/۵۸۸	٪ ۲۹/۲۱
ایلنا	۱۴۵/۴۲۰	٪ ۲۶/۲۹
مشرق	۳۳/۶۲۱	٪ ۶/۰۷
الف	۳۳/۶۲۱	٪ ۶/۰۷
آفتاب	۱/۵۱۷	٪ ۰/۲۷
کل	**۵۵۳/۱۱۱**	**٪ ۱۰۰**

این خبرها بین ۴۰ تا ۵۰۰ کلمه طول دارند. اطلاعات بیشتر راجع به طول خبرها را می‌توانید در جدول ۲ مشاهده کنید.

جدول ۲. تعداد خبرها بر اساس تعداد واژگانشان

تعداد واژگان	تعداد مقالات خبری	درصد از کل
۴۰ تا ۱۰۰	۱۲۳/۹۱۳	٪ ۲۲/۴۰
۱۰۰ تا ۱۵۰	۹۶/۹۱۹	٪ ۱۷/۵۲
۱۵۰ تا ۲۰۰	۸۱/۸۵۷	٪ ۱۴/۷۹
۲۰۰ تا ۲۵۰	۶۶/۷۳۸	٪ ۱۲/۰۶
۲۵۰ تا ۳۰۰	۵۴/۱۹۹	٪ ۹/۷۸
۳۰۰ تا ۳۵۰	۴۲/۸۲۶	٪ ۷/۷۴
۳۵۰ تا ۴۰۰	۳۴/۴۵۷	٪ ۶/۲۳
۴۰۰ تا ۴۵۰	۲۸/۱۹۴	٪ ۵/۰۹
۴۵۰ تا ۵۰۰	۲۴/۰۸۸	٪ ۲۴/۳۵
کل	**۵۵۳/۱۱۱**	**٪ ۱۰۰**

جدول زیر تعداد خبرها بر اساس تعداد کلمات کلیدی را نشان می‌دهد. بر اساس جدول ۳ بیشتر خبرها دو عبارت کلیدی و پس از آن بیشتر خبرها دو عبارت کلیدی دارند.

جدول ۳. تعداد خبرها بر اساس تعداد عبارت‌های کلیدی‌شان

تعداد عبارت‌های کلیدی	تعداد مقالات خبری	درصد از کل
۲	۱۵۷/۴۶۶	٪ ۲۸/۴۶
۳	۲۰۲/۷۴۸	٪ ۳۶/۶۵
۴	۸۱/۵۰۰	٪ ۱۴/۷۳
۵	۵۶/۲۷۸	٪ ۱۰/۱۷
۶	۷/۰۶۰	٪ ۱/۲۷
۷	۲/۸۸۳	٪ ۰/۵۲
۸	۱/۳۸۲	٪ ۰/۲۵
۹	۱/۰۳۲	٪ ۰/۱۸
بیشتر از ۹	۴۲/۷۶۲	٪ ۷/۷۳
کل	**۵۵۳/۱۱۱**	**٪ ۱۰۰**

همچنین طبق مقاله [۱۷] مجموعه‌داده تعداد واژگان موجود در هر عبارت کلیدی تا ۷ واژه محدود شده‌اند. از آنجایی که عبارت‌های کلیدی نماینده موضوعات موردبحث در متن هستند، بدیهی است که درمورد برخی از موضوعات به صراحت در متن سخن نمی رود، بلکه اشاره های ضمنی ای به آن ها می‌شود. داده حاضر نیز از قاعده مستثنی نیست. جدول ۴ نشان می‌دهد که حدود ۳۵ درصد از عبارت‌های کلیدی، در متن حاضر نیستند. حاضر نبودن در متن به این معناست که عین این عبارت‌های کلیدی در متن نیامده اند. به این نوع از عبارت‌های کلیدی عبارت‌های کلیدی غایب می‌گویند.

جدول ۴. تعداد عبارت‌های کلیدی حاضر و غایب

	تعداد عبارت‌های کلیدی	درصد از کل
حاضر	۱/۳۲۴/۳۰۵	٪ ۶۴/۸۶
غایب	۷۱۷/۳۲۱	٪ ۳۵/۱۳
کل	**۲/۰۴۱/۶۲۶**	**٪ ۱۰۰**

در پایان این مجموعه‌داده با نسبت ۸۰، ۱۰ و ۱۰ درصد به سه زیرمجموعه آموزش، اعتبارسنجی و آزمون تقسیم شد.

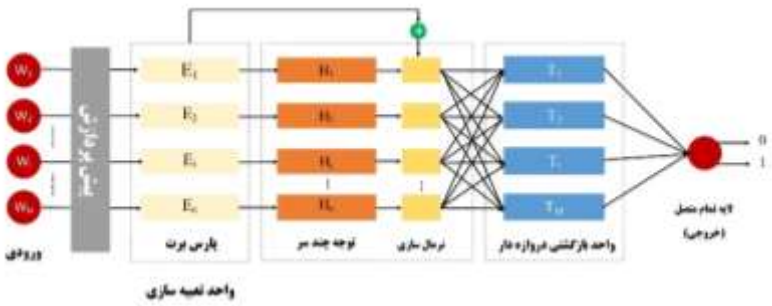
۴- روش پیشنهادی

برای حل مسئله استخراج کلمات کلیدی با استفاده از شبکه‌های عصبی مصنوعی، راه حلی که به ذهن می‌رسد این است که مسئله را به‌صورت یک مسئله برچسب‌زنی دنباله ببینیم، به این صورت که واژگان خارج از کلمات کلیدی با برچسب صفر معرفی شوند و واژگان درون عبارت‌های کلیدی با برچسب یک مشخص شوند. روش برچسب‌زنی دنباله با استفاده از شبکه‌های عصبی بازگشتی نخستین راهکاری بود که در این پژوهش استفاده شد و به‌خاطر مشکلات و دقت پایینش پی گرفته نشد. از مشکلات این روش برای این وظیفه می‌توان به پراکنده[18] بودن برچسب‌ها اشاره کرد، به‌طوری‌که برچسب‌های صفر چند ده برابر برچسب یک هستند. این موضوع آموزش شبکه را دشوار می‌کند. یکی از راه‌های برخورد با کلاس‌های نامتعادل[19] در این مسائل استفاده از تابع زیان مناسب است، به‌طوری‌که این تابع زیان[20] بیشتری برای کلاس یک قائل شود تا این عدم تعادل تا حدی جبران شود. علاوه بر این به بر اساس رابطه (۱) سوگیری[21] لایه آخر شبکه را مقداردهی اولیه کرد تا خروجی را به سمت نتیجه دلخواه ما منحرف کند [۲۱].

$$p_0 = \frac{pos}{pos+neg} = \frac{1}{1+e^{-b_0}} \quad (۱)$$

$$b_0 = -log_e\left(\frac{1}{p_0-1}\right) = log_e(pos/neg) \quad (۲)$$

این روش نیز راه به جایی نبرد و رفتار شبکه عصبی بهبود نیافت. پس از این آزمایش‌ها ما به روش اصلی پژوهش رسیدیم، یعنی استفاده از مکانیسم توجه[22] که با استفاده از اتصال پرشی و لایه نرمال‌سازی به شبکه کمک می‌کند تا میزان اهمیت هر واژه از متن ورودی را برای تولید هر واژه از متن خروجی محاسبه کند و بر چالش کلاس‌های نامتعادل غلبه کند. شکل (۱) معماری روش پیشنهادی را نشان می‌دهد که در ادامه به‌تفصیل مراحل مختلف آن را شرح خواهیم داد.



شکل ۱. معماری روش پیشنهادی

در این معماری روند کلی به سه فاز **پیش‌پردازش، پردازش، دسته‌بندی** تقسیم می‌شوند. فاز اول تحت عنوان فاز پیش‌پردازش، شامل مجموعه فعالیت‌هایی است که از آنها تحت عنوان فعالیت‌های پیش‌پردازشی یاد می‌شود. این فعالیت‌ها که در غالب وظایف پردازش زبان

[18] Sparse
[19] Imbalanced Classes
[20] Loss Function
[21] Biase
[22] Transformers

طبیعی وجود دارند، به‌منظور آماده‌سازی همه زیرساخت‌های لازم جهت انجام اهداف محاسباتی و پردازشی موردنظر در مراحل آتی انجام می‌پذیرد. ورودی این فاز اسناد متنی خام هستند که از یک رشته از کاراکترها تشکیل می‌شوند.

۴-۱- فاز اول: پیش‌پردازش

پیش‌پردازش یکی از اجزای کلیدی در بسیاری از الگوریتم‌های کلمات کلیدی می‌باشد. این مرحله نقش بسیار مهمی در داده‌کاوی متون ایفا می‌کند و اولین گام در فرایند استخراج از متن می‌باشد. این گام می‌تواند تأثیر قابل‌توجهی بر روی نتایج نهایی داشته باشد. همان‌طور که پیش‌تر اشاره شد، پیش‌پردازش شامل مجموعه‌ای از فعالیت‌های ابتدایی بر روی متون ورودی است که زیرساخت‌های لازم را برای انجام هدف اصلی آماده می‌کند و هدف آن آماده‌سازی متون ورودی به‌منظور انجام محاسبات موردنظر پژوهشگر در راستای اهداف پژوهش است. همان‌طور که پیش‌تر بیان شد، متون استفاده شده در این پژوهش به زبان فارسی می‌باشد. ما برای آماده‌سازی متن خود نرمال‌سازی متن و واحدسازی متن را به‌عنوان فاز پیش‌پردازش انجام دادیم.

۴-۲- فاز دوم: پردازش

این فاز از چهار مرحله اصلی تشکیل شده است:

۱. تعبیه کلمات[۲۳]
۲. مکانیسم توجه[۲۴]
۳. شبکه واحد دروازه‌دار بازگشتی[۲۵]
۴. لایه تماماً متصل[۲۶]

برای تعبیه‌کردن و بازنمایی کلمات از مدل زبانی پیش آموزش‌دیده شده پارس برت [۲۲] استفاده کردیم که مبتنی بر BERT روی داده فارسی آموزش داده شده است. متن کدگزاری و واحدسازی شده با حداکثر طول ۵۰۰ و حاشیه[۲۷] صفر به‌عنوان ورودی در اختیار مدل قرار می‌گیرد.

بردارهای تعبیه شده در این لایه در اختیار یک لایه توجه چند سر قرار می‌گیرد تا ارتباطات معنایی بین کلمات را کشف کند. این لایه دو نوع خروجی را می‌تواند در اختیار ما قرار دهد. ماتریسی سه‌بعدی به شکل (۵۰۰،۵۰۰،۸) که اولی تعداد سرهای سازوکار توجه است و سپس یک ماتریس ۵۰۰×۵۰۰ که وزن هر کلمه نسبت به‌تمامی کلمات متن را نشان می‌دهد. خروجی لایه با همان ابعاد ۷۶۸×۵۰۰ می‌باشد که تأثیر وزن‌ها بر روی بردار کلمات اعمال شده است. در مرحله بعد خروجی لایه قبلی (خروجی نوع دوم) با خروجی BERT جمع می‌شود تا کلمات مؤثر مشخص شوند. در واقع این کار باعث می‌شود خروجی لایه توجه که بردارهای تغییریافته است به نسبت کلیدی‌تر و مرتبط‌تر بودن که لزوماً دیگر بردار معنایی آن کلمه نمی‌باشد با بردارهای معنایی خروجی BERT جمع شود و کلمات مشخص شود. سپس در یک لایه نرمال‌سازی قرار می‌گیرد تا بردارهای کلمات نرمال شوند (بین ۰ تا ۱ قرار گیرند).

سپس از شبکه عصبی واحد بازگشتی دروازه‌دار دوطرفه استفاده می‌کنیم. برای درک عمیق‌تر کلمات این بلوک به تعداد سه بار متوالی در شبکه تکرار می‌شود. البته تعداد بالاتر امتحان شده دقت تا حدودی افزایش می‌یابد منتهی شبکه پیچیده‌تر شده و سرعت آموزش به‌مراتب پایین می‌آید.

---

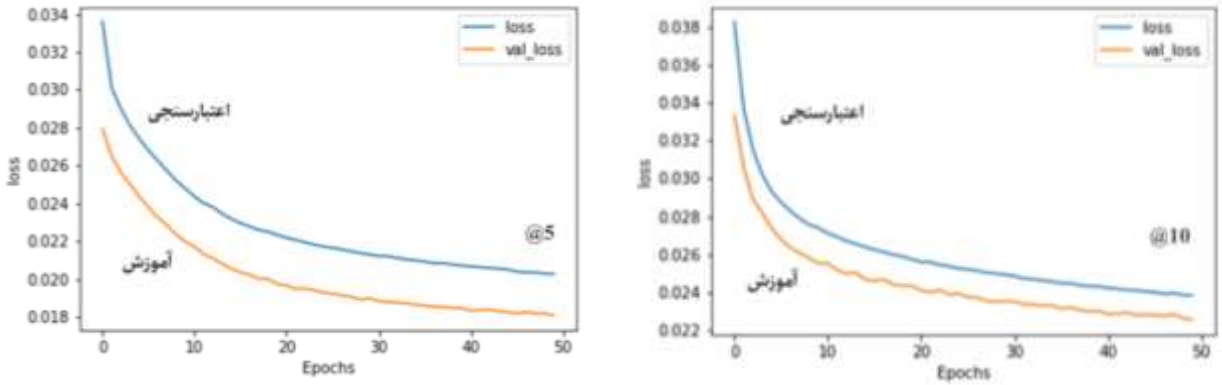
[23] Word Embedding
[24] Attention Mechanism
[25] Gate Recurrent Unit
[26] Fully-Connected Layer
[27] Padding

۳-۴- فاز سوم: دسته‌بندی

در این مرحله خروجی شبکه تبدیل‌کننده را به یک لایه تمام متصل با تعداد نرون ۲ می‌دهیم تا عمل دسته‌بندی را روی ورودی ۷۶۸×۵۰۰ انجام دهد. خروجی این لایه ۲×۵۰۰ می‌باشد که شامل دو احتمال برای کلاس‌های ۰ و ۱ است.

۴-۴- آموزش شبکه

برای آموزش شبکه از زبان برنامه‌نویسی پایتون به همراه کتابخانه‌های کراس[۲۸] و تنسورفلو[۲۹] استفاده کرده‌ایم. همچنین برای تعداد دوره[۳۰] برای آموزش شبکه را ۵۰ مناسب دیدیم و نرخ یادگیری برای این شبکه عمیق ۰.۰۰۱ با بهبوددهنده[۳۱] SGD [۳۲] قرار دادیم. تابع زیان شبکه به‌عنوان ایده این مقاله تابع خطای Focal انتخاب شد که تأثیر مثبتی در آموزش شبکه داشت. شکل ۲ نمودار خطای آموزش و اعتبارسنجی را در ۵۰ دوره برای ۵ و ۱۰ کلمه کلیدی برتر نشان می‌دهد.



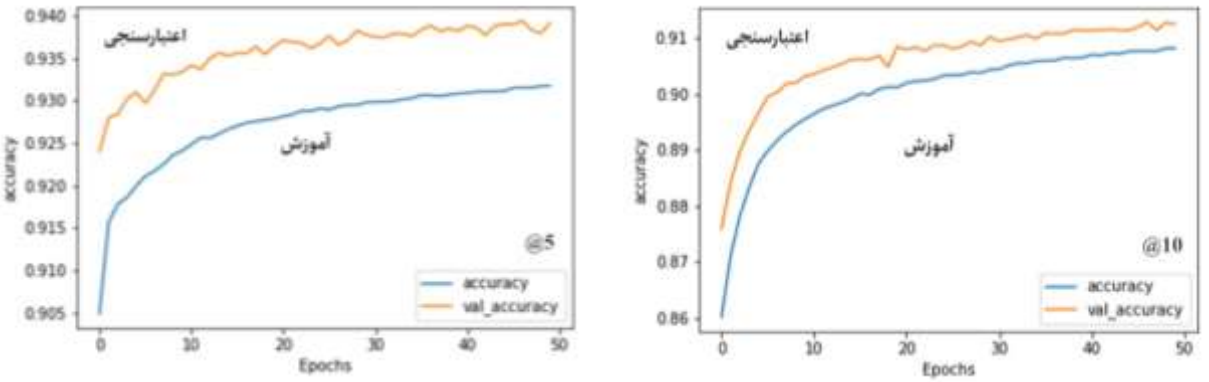
شکل ۲. نمودار خطای آموزش و اعتبارسنجی برای ۵ و ۱۰ کلمه کلیدی برتر

نمودار خطا باتوجه‌به شیب و سرعت کاهش خطا و همچنین نسبت خطای آموزش به اعتبارسنجی جزو دسته شبکه‌های خوش آموزش قرار می‌گیرد.

همچنین با بررسی نمودار دقت طبق شکل ۳ برای حد ۵ و ۱۰ کلمه کلیدی برتر رشد خوب و دقت بالای آموزش شبکه از ۸۶ تا ۹۲ درصد دقت را استدلال کرد.

---

[28] Keras
[29] Tensorflow
[30] Epochs
[31] Optimizer
[32] Stochastic gradient descent



شکل ۳. نمودار دقت آموزش و اعتبارسنجی برای ۵ و ۱۰ کلمه کلیدی برتر

در این پژوهش نخست دقت روش‌های پایه را در وظیفهٔ استخراج عبارت‌های کلیدی از متون خبری زبان فارسی به دست آوردیم. این کار به این دلیل انجام شد تا دقت‌های پایه‌ای در اختیار داشته باشیم تا بتوانیم عملکرد روش پیشنهادی را به‌درستی بسنجیم. این روش‌های پایه عبارت بودند از روش‌های نظارت نشده آماری و مبتنی بر گراف و همچنین دو روش نظارت شده که همگی از روش‌های مهم و اصلی این وظیفه به‌حساب می‌آیند. این نتایج بر روی مجموعه‌داده Perkey به‌دست‌آمده است. سپس شبکه پیشنهادی خود را که از یک مدل زبانی مبتنی بر BERT برای بازنمایی عددی کلمات، سپس استفاده از سازوکار توجه چند سر که میزان اهمیت هر کلمه را کشف می‌کند و سپس سه بلوک شبکه عصبی بازگشتی دروازه‌دار دوطرفه و در نهایت یک لایه تماماً متصل برای دسته‌بندی کلمات است را آموزش دادیم و نتایج را در جدول ۵ با روش‌های پایه مقایسه کردیم.

جدول ۵. نتایج همه روش‌ها بر روی عبارت‌های کلیدی حاضر داده زیرمجموعه

	P@5	R@5	F_1@5	P@10	R@10	F_1@10
TFIDF	.1726	.2713	.2110	.1218	.3792	.1843
KPMiner	*.1902*	.2523	.2169	*.1634*	.3240	.2173
YAKE	.0733	.1089	.0876	.0664	.1974	.0994
S.Rank	.0733	.0996	.0694	.0624	.2180	.0970
T.Rank	.0987	.1634	.1230	.0665	.2136	.1015
M.Rank	.1093	.1761	.1349	.0771	.2468	.1176
KEA	.1839	*.2937*	*.2262*	.1301	*.4124*	*.1979*
RNN	.5249	.4794	.5012	.4621	.5588	.5059
Proposal method	**.7664**	**.6203**	**.6616**	**.7785**	**.6986**	**.7294**

۵- نتیجه‌گیری

هدف کلی این مقاله ارتقای دقت استخراج کلمات کلیدی اخبار فارسی، با استفاده از نوع جدیدی از الگوریتم های هوش مصنوعی با عنوان الگوریتم های یادگیری عمیق است. پیش‌تر در خصوص اهمیت استخراج کلمات کلیدی توضیحات اجمالی ارائه شده است و کاربرد های آن نیز آورده شده است. هر چه دقت استخراج بالاتری حاصل آید می تواند بر روند استفاده آن نیز تأثیر گذار باشد. مجموعه‌داده مورد استفاده در این پژوهش PerKey جمع آوری شده توسط دوست محمدی و همکاران می‌باشد که به نسبت حجم داده‌ای که دارد مناسب آموزش شبکه‌های عصبی می‌باشد. مدل پیشنهادی ما یک شبکه یادگیری عمیق بر اساس مدل زبانی برت بود که در ارزیابی ها نسبت به شبکه عصبی عمیق دنباله به دنباله مقاله [۱۷] ، روش پیشنهادی ما روی ۵ و ۱۰ کلمه کلیدی برتر به ترتیب ۱۳.۲ درصد و ۲۳.۲۴ درصد رشد

در معیار اف-۱، ۱۱.۴۸ درصد و ۱۴.۹۴ درصد رشد در معیار بازخوانی و همچنین ۲۵.۵۲ درصد و ۳۲.۲۵ درصد رشد در معیار صحت نسبت به شبکه عمیق دنباله به دنباله مقاله [۱۷] داشته است. مجموعه‌داده خوب در تشخیص و استخراج کلمات کلیدی نقش بسزایی دارد؛ بدین‌منظور پیشنهاد می‌شود به‌عنوان کارآتی یک مجموعه‌داده مناسب تهیه شود و برچسب کلمات کلیدی با دقت بیشتری و با معیارهای مناسب توسط جمعی خبره تهیه گردد.

مراجع

[۱] E. Doostmohammadi, M. H. Bokaei, and H. Sameti, "Persian keyphrase generation using sequence-to-sequence models," in *2019 27th Iranian Conference on Electrical Engineering (ICEE)*, 2019: IEEE, pp. 2010-2015 .

[۲] S. R. El-Beltagy and A. Rafea, "KP-Miner: A keyphrase extraction system for English and Arabic documents," *Information systems*, vol. 34, no. 1, pp. 132-144, 2009.

[۳] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, "A text feature based automatic keyword extraction method for single documents," in *European conference on information retrieval*, 2018: Springer, pp. 684-691 .

[۴] M. Won, B. Martins, and F. Raimundo, "Automatic extraction of relevant keyphrases for the study of issue competition," in *Proceedings of the۲۰ th international conference on computational linguistics and intelligent text processing*, 2019, pp. 7-13 .

[۵] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1, pp. 1.۱۹۹۸ /۰۱/۰۴/۱۹۹۸ ,۰۷-۱۱۷

[۶] P. J.-J. Herings, G. v. d. Laan, and D. Talman, "The positional power of nodes in digraphs," *Social Choice and Welfare*, vol. 24, no. 3, pp. 439-454, 2005/06/01 2005.

[۷] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604-632, 1999.

[۸] F. Rousseau and M. Vazirgiannis, "Main Core Retention on Graph-of-Words for Single-Document Keyword Extraction," Cham, 2015: Springer International Publishing, in Advances in Information Retrieval, pp. 382-393 .

[۹] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," Barcelona, Spain, July 2004: Association for Computational Linguistics, in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404-411.

[۱۰] X. Wan and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge," in *AAAI*, 2008, vol. 8 ,pp. 855-860 .

[۱۱] A. Bougouin, F. Boudin, and B. Daille, "Topicrank: Graph-based topic ranking for keyphrase extraction," in *International joint conference on natural language processing (IJCNLP)*, 2013, pp. 543-551 .

[۱۲] F. Boudin, "Unsupervised keyphrase extraction with multipartite graphs," *arXiv preprint arXiv:1803.08721*, 2018.

[۱۳] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical automatic keyphrase extraction," in *Proceedings of the fourth ACM conference on Digital libraries*, 1999, pp. 254-255 .

[۱۴] Q. Zhang, Y. Wang, Y. Gong, and X.-J. Huang, "Keyphrase extraction using deep recurrent neural networks on twitter," in *Proceedings of the 2016 conference on empirical methods in natural language processing* ,۲۰۱۶ ,pp. 836-845 .

[۱۵] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, and Y. Chi, "Deep keyphrase generation," *arXiv preprint arXiv:1704.06879*, 2017.

[۱۶] م. محمدی و م. آنالویی، "استخراج کلمات کلیدی اسناد فارسی"، سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، ۱۳۸۶.

[۱۷] ع. احمدی و ط. حسینی خواه، "استخراج کلمات کلیدی یک متن با استفاده از شبکه‌های عصبی"، دهمین کنفرانس بین‌المللی مهندسی صنایع، ۱۳۹۲.

[۱۸] م. رسولی و ب. مینایی بیدگلی، "روشی جدید در خطایابی املایی در زبان فارسی"، دومین کنفرانس داده‌کاوی ایران، ۱۳۸۷.

[۱۹] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[۲۰] E. Doostmohammadi, M. H. Bokaei, and H. Sameti, "Perkey: A persian news corpus for keyphrase extraction and generation," in *2018 9th International Symposium on Telecommunications (IST)*, 2018: IEEE, pp. 460-465 .

[۲۱] "Classification on imbalanced data." https://www.tensorflow.org/tutorials/structured_data/imbalanced_data

[۲۲] M. Farahani ,M. Gharachorloo, M. Farahani, and M. Manthouri, "Parsbert: Transformer-based model for persian language understanding," *Neural Processing Letters*, vol. 53, no. 6, pp. 3831-3847, 2021.