بهبود دقت واژگان کلیدی استخراج‌شده از متن فارسی با استفاده از الگوریتم Word2Vec

محمدرضا حسنی آهنگر* و علی امیری‌جزه

آزمایشگاه داده‌های حجیم، مرکز داده‌های حجیم و جنگ‌شناختی سایبری، دانشکدهٔ رایانه و قدرت سایبری، دانشگاه جامع امام حسین(ع)، تهران، ایران

چکیده

واژگان کلیدی لغات مهمی از سند هستند که بیان‌گر توصیفی از متن هستند و نقش بسیار مهمی در فهم دقیق و سریع از محتوا دارند. شناسایی واژگان کلیدی از متن با روش‌های معمول کاری زمان‌بر و پرهزینه است. در این مقاله ابتدا با استفاده از شبکهٔ عصبی پیشرو و از طریق الگوریتم Word2Vec ماتریس همبستگی واژگان را به‌ازای یک سند محاسبه و سپس با استفاده از ماتریس همبستگی و یک فهرست اولیهٔ محدود از واژگان کلیدی، نزدیک‌ترین واژگان را از نظر شباهت در قالب فهرست نزدیک‌ترین همسایگی‌ها استخراج می‌کنیم. فهرست به‌دست‌آمده را به‌صورت نزولی مرتب و از ابتدای فهرست، درصدهای مختلفی از واژگان را انتخاب و به‌ازای هر درصد، ده مرتبه فرایند آموزش شبکهٔ عصبی و ساخت ماتریس همبستگی و استخراج فهرست نزدیک‌ترین همسایگی‌ها را تکرار و درنهایت میانگین دقت، فراخوانی و معیارF را محاسبه می‌کنیم. این کار را تا جایی ادامه می‌دهیم که به بهترین نتایج در ارزیابی دست یابیم؛ نتایج نشان می‌دهند که به‌ازای انتخاب حداکثر چهل درصدِ واژگان از ابتدای فهرستِ نزدیک‌ترین همسایگی‌ها، نتایج مورد قبولی به‌دست می‌آید. الگوریتم بر روی پیکره‌ای با هشتصد خبر که به‌صورت دستی واژگان کلیدی آن‌ها را استخراج کرده‌ایم، آزمایش‌شده است و نتایج آزمایش‌ها نشان می‌دهد که دقت روش پیشنهادی ۷۸ درصد خواهد بود.

واژگان کلیدی: واژگان کلیدی، الگوریتم word2Vec، شبکه عصبی، وزن دهی ویژگی

Improving Precision of Keywords Extracted From Persian Text Using Word2Vec Algorithm

Mohammad Reza Hasani Ahangar* & Ali Amiri jezeh

Big Data Laboratory, Computer Science and Technology Center, Faculty of Information and Communication Technology, Imam Hossein University, Tehran, Iran

Abstract

Keywords can present the main concepts of the text without human intervention according to the model. Keywords are important vocabulary words that describe the text and play a very important role in accurate and fast understanding of the content. The purpose of extracting keywords is to identify the subject of the text and the main content of the text in the shortest time. Keyword extraction plays an important role in the fields of text summarization, document labeling, information retrieval, and subject extraction from text. For example, summarizing the contents of large texts into smaller texts is difficult, but having keywords in the text can make you aware of the topics in the text. Identifying keywords from the text with common methods is time-consuming and costly. Keyword extraction methods can be classified into two types with observer and without observer. In general, the process of extracting keywords can be explained in such a way that first the text is converted into smaller units called the word, then the redundant words are removed and the remaining words are weighted, then the keywords are selected from these words. Our proposed method in this paper for identifying keywords is a method with observer. In this paper, we first calculate the word correlation matrix per document using a feed forward neural network and Word2Vec algorithm. Then, using the correlation matrix and a limited

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding

•تاریخ ارسال مقاله: ۱۳۹۷/۰۲/۰۲ •تاریخ پذیرش: ۱۳۹۹/۱۲/۰۹ •تاریخ انتشار: ۱۴۰۰/۰۳/۰۱ •نوع مطالعه: کاربردی

سال ۱۴۰۰ شمارهٔ ۱ پیاپی ۴۷

[DOI: 10.52547/jsdp.18.1.60]

initial list of keywords, we extract the closest words in terms of similarity in the form of the list of nearest neighbors. Next we sort the last list in descending format, and select different percentages of words from the beginning of the list, and repeat the process of learning the neural network 10 times for each percentage and creating a correlation matrix and extracting the list of closest neighbors. Finally, we calculate the average accuracy, recall, and F-measure. We continue to do this until we get the best results in the evaluation, the results show that for the largest selection of 40% of the words from the beginning of the list of closest neighbors, the acceptable results are obtained. The algorithm has been tested on corpus with 800 news items that have been manually extracted by keywords, and laboratory results show that the accuracy of the suggested method will be 78%.

Keywords: keywords, word2vec algorithm, neural network, giving weight features

۱- مقدمه

هدف از استخراج واژگان کلیدی تشخیص موضوع متن و محتوای اصلی متن در کمترین زمان است. استخراج واژگان کلیدی نقش مهمی در زمینه‌های خلاصه‌سازی متون، برچسب‌گذاری اسناد، بازیابی اطلاعات و استخراج موضوع از متن دارد [1]. یکی از مشکلات مهم در تشخیص واژگان کلیدی، شناسایی واژگان غیرکلیدی است [2] به‌گونه‌ای که بیان‌گر موضوع اصلی متن نیستند. واژگان کلیدی، عناصر بسیار مهمی در جست‌وجو و دسترسی اطلاعات هستند. آن‌ها می‌توانند به‌عنوان مجموعهٔ واژگان تشریح‌کنندهٔ سند در طی عملیات جست‌وجو مدنظر قرار بگیرند. به‌عبارت‌دیگر، به هر عبارت مهمی که محتویات داخل سند را تشریح کند، واژه کلیدی گفته می‌شود [3]. واژگان کلیدی در دو گروه طبقه‌بندی می‌شوند [4]، واژگان کلیدی تابعی[1] و واژگان کلیدی آموزنده[2].

واژگان کلیدی تابعی یا غیرمفید برای واژگان دستوری یا مرتبط با زبان استفاده می‌شوند [4] که منظور از واژگان دستوری، واژگان معناداری هستند که به اشیا و مفاهیم اشاره می‌کنند، مانند "کتاب، قلم و نوشتن"؛ این واژگان ارتباط کمی با محتویات سند دارند و باید حذف شوند و در نظر گرفته نشوند.

واژگان کلیدی آموزنده ارتباط خیلی قوی با محتویات و معنای متن سند دارند. تشخیص مرز بین واژگان کلیدی تابعی و آموزنده خیلی سخت نیست و می‌توان یک مرز فازی برای آن‌ها در نظر گرفت [4, 5]. برای مثال سندی را که اثرات افزایش گازهای گلخانه‌ای را در سوراخ‌شدن لایه ازون بررسی کرده است در نظر بگیرید؛ آن دسته از واژگان کلیدی که ارتباط کمتری نسبت به محتویات سند دارند، به‌عنوان واژگان کلیدی غیرمفید در نظر گرفته می‌شوند، مانند واژگانی از قبیل "سال، جهان، جبران، صدمات" البته ممکن است این واژگان در سند دیگری به‌عنوان واژگان کلیدی آموزنده در نظر گرفته شوند. در سند یادشده همان‌طور که بیان شد، واژگانی به‌عنوان واژگان کلیدی آموزنده در نظر گرفته می‌شوند که ارتباط خیلی قوی با محتویات و معنی متن داشته باشند، واژگانی از قبیل " گاز، کارخانه، گرم، زمین، ازون، آب، یخ، قطب" البته این واژگان نیز ممکن است در سند دیگری به عنوان واژگان کلیدی غیرمفید لحاظ شوند. در این مقاله استخراج واژگان کلیدی آموزنده مدنظر است. تاکنون روش‌های بسیاری برای استخراج واژگان کلیدی از یک سند ارائه شده است، که دارای معایب و مزایایی هستند، روشی که در این مقاله ارائه می‌شود یک روش مبتنی بر شبکه عصبی با درنظرگرفتن همبستگی واژگان از لحاظ معنایی و شباهت با یکدیگر است. در این مقاله ابتدا مروری بر کارهای پیشین و مشابه صورت پذیرفته است؛ سپس روش رایج استخراج واژگان کلیدی بیان و در ادامه راه‌حل پیشنهادی با بیان جزئیات مطرح و در پایان نتایج کار ارائه شده است.

۲- مروری بر کارهای پیشین

واژگان و عبارات کلیدی می‌توانند مفاهیم اصلی متن را بدون دخالت انسان بسته به مدل ارائه دهند [6]. خلاصه‌سازی محتویات متن‌های بزرگ به متن‌های کوچک‌تر برای آنکه بهتر فهمیده شوند کاری سخت است اما با داشتن واژگان کلیدی موجود در متن می‌توان از موضوعات داخل متن آگاه شد. ازاین‌روی کارهای بسیاری در این حوزه صورت پذیرفته است.

در [7] ابتدا اسناد خوشه‌بندی می‌شوند، سپس با استفاده از ویژگی‌های به‌دست‌آمده از نزدیک‌ترین همسایگان یک سند، K سندی که در واقع بیشترین شباهت را به سند مشخص شده دارند، به سند اضافه و سند بسط داده می‌شود، درنهایت الگوریتم رتبه‌بندی گراف بر روی سندی که گسترش یافته است، اعمال می‌شود و بر اساس ویژگی‌های سند و همسایگان سند، واژگان کلیدی استخراج‌شده از سند، بهبود می‌یابند. در [8] ابتدا عبارات

[1] Functional
[2] Informative

[DOI: 10.52547/jsdp.18.1.60]

اسمی از سند استخراج می‌شوند، سپس عباراتی که دست‌کم یک واژه مشابه دارند در داخل یک خوشه قرار می‌گیرند؛ با ویژگی تعداد تکرار عبارات اسمی و واژگان داخل این عبارات، خوشه‌ها رتبه‌بندی می‌شوند؛ درنهایت عباراتی به‌عنوان عبارات کلیدی انتخاب می‌شوند که خوشهٔ مربوطه بالاترین رتبه را داشته باشد. در [9] برای استخراج واژگان کلیدی از الگوریتم‌های یادگیری با ناظر استفاده شده است، به‌گونه‌ای که ابتدا عبارات کلیدی در یک سند را به‌صورت فهرستی از عبارات کلیدی مشخص می‌کنند، سپس الگوریتم باید یاد بگیرد که اسناد را در دو دسته منفی و مثبت طبقه‌بندی کند. در [10] ابتدا واژگان پرتکرار از متن استخراج می‌شوند، سپس هم‌رخدادی این واژگان با تمام واژگان موجود در متن در یک ماتریس محاسبه می‌شود، در این روش ادعا شده است که هر واژه‌ای که با واژگان پرتکرار هم‌رخدادی بیشتری داشته باشد، بهترین نامزد برای واژهٔ کلیدی است. در [11] ابتدا نامزدهای واژگان کلیدی از متن استخراج می‌شوند، سپس هم‌رخدادی بین این نامزدها محاسبه و به‌ازای فراوانی به هر یک امتیازی نسبت داده می‌شود، نامزدهایی به‌عنوان عبارات کلیدی انتخاب می‌شوند که بیشترین امتیاز را داشته باشند. در [12] از شبکه عصبی برای استخراج عبارات کلیدی از سند استفاده می‌شود، ابتدا شبکهٔ عصبی با استفاده از ویژگی‌هایی همچون فراوانی واژه و موقعیت واژه در سند آموزش داده می‌شود، سپس عبارات کلیدی نامزدشده با درنظرگرفتن یک حد آستانه تعریف‌شده در دو دستهٔ عبارات کلیدی و عبارات غیرکلیدی دسته‌بندی می‌شوند. در [13] از شبکه رقابتی LVQ[1] و شبکه عصبی MLP[2] برای استخراج واژگان کلیدی متون استفاده شده است، این الگوریتم‌ها درواقع واژگان موجود در یک متن را به دو خوشه واژگان کلیدی و غیرکلیدی دسته‌بندی می کنند، در این روش ادعا شده که نتیجه‌ها بر روی هشتاد سند متنی نشان داده است که روش MLP نسبت به روش دیگر نتایج بهتری را به دست می‌دهد. در [7] واژگان کلیدی که استخراج می‌شوند، نمی‌توانند نمایندگان مناسبی برای اسنادی باشند که بیشترین فاصله را از مرکز خوشه دارا هستند. روش [8] وابسته به زبان است و شناسایی وابسته‌های پسین و پیشین به‌ازای زبان‌های مختلف تفاوت دارند، همچنین استخراج عبارات اسمی، بازنمایی مناسبی برای شناسایی تمام واژگان کلیدی یک سند نیستند. روش [9] نیاز به دادهٔ قابل توجهی برای یادگیری و آموزش دارد و فقط اسنادی که در یک دسته قرار می‌گیرند، واژگان کلیدی مشابه‌ای دارند و شناسایی واژگان کلیدی جدید در این اسناد ممکن نیست. در [10] هم‌رخدادی صریح واژگان مدنظر قرار گرفته‌است و واژگانی که به‌منظور جلوگیری از تکرار با ضمائر و واژگان اختصار بیان شده‌اند جزو واژگان پرتکرار قرار نمی‌گیرند. در [11] واژگانی که فقط با یکدیگر تکرار می‌شوند، تشکیل یک عبارت کلیدی را نمی‌دهند. در [12, 13] عبارات کلیدی تابعی به مراتب بیشتر از دیگر روش‌ها وجود دارد، چون از ویژگی‌های روابط معنایی بین واژگان برای شناسایی عبارات کلیدی استفاده نشده است.

۳- روش استخراج واژگان کلیدی

روش‌های استخراج واژگان و عبارات کلیدی را می‌توان در دو نوع با ناظر[3] و بدون ناظر[4] دسته‌بندی کرد [7]. اگر در مجموعه‌ای از اسناد واژگان کلیدی مشخص‌شده باشند، فرایند استخراج واژگان کلیدی با ناظر و در غیر این صورت بدون ناظر خواهد بود؛ اما به‌طورکلی می‌توان فرایند استخراج واژگان کلیدی از متن را این‌گونه توضیح داد که در ابتدا متن به واحدهای کوچک‌تری به نام واژه تبدیل می‌شود، بعدازآن واژگان زائد حذف و واژگان باقیمانده وزن‌دهی می‌شوند، سپس واژگان کلیدی از بین این واژگان انتخاب می‌شوند.

۴- بردار ویژگی‌ها

واژه‌ها و واژگان برای چندین دهه به‌عنوان واحد اصلی در زبان طبیعی مورد مطالعه قرارگرفته‌اند [14]. در پردازش زبان طبیعی یکی از مهم‌ترین پارامترها در تمام وظایف، چگونگی ارائه واژگان به‌عنوان ورودی به هر یک از مدل‌هاست، به بیانی ساده‌تر ابتدا باید مفهوم شباهت و تفاوت بین واژگان را داشته باشیم و این در قالب بردار ویژگی‌های متن قابل بیان است [15]. آنچه در انتخاب واژگان کلیدی حائز اهمیت است، وزن‌دهی واژگان نامزد است. به‌منظور وزن‌دهی واژگان برای استخراج واژگان کلیدی [16]، می‌توان از روش‌هایی اعم از تعداد تکرار واژه (TF) در سند و اسناد مختلف (IDF) [17]، روش‌های یادگیری ماشین [18, 19] و ترکیب روش‌های آماری و زبان‌شناختی [20] استفاده کرد. ازآن‌جایی‌که تعداد قابل‌توجهی از روش‌های قدیمی از معیار TF-IDF استفاده می‌کنند [21] و به‌منظور مقایسه با روش پیشنهادی، در ادامه این روش را توضیح خواهیم داد.

[1] Learning Vector Quantization
[2] Multilayer Perceptron
[3] Superwise
[4] Unsuperwise

[DOI: 10.52547/jsdp.18.1.60]

$$p(w_o|w_I)=\frac{\exp(v'^{\mathrm{T}}_{wo}v_{wI})}{\sum_{w=1}^{W}\exp(v'^{\mathrm{T}}_{w}v_{wI})} \quad (۱)$$

یکی از جدیدترین و پیشرفته‌ترین روش‌ها برای بازنمایی[1] واژگان و استخراج بردار ویژگی‌ها از متن، الگوریتم Word2Vec است. درواقع خروجی Word2Vec بازنمایی بردار چگال واژگان است که می‌تواند مفاهیم معنایی و همبستگی موضوعی را نشان دهد [22]. سطرهای ماتریس بردار ویژگی‌ها در این روش، بیان‌گر روابط معنایی و روابط موضوعی واژگان با یکدیگر هستند [15 ,14].

۴-۱- وزن‌دهی به روش TF-IDF[2]

در این روش تعداد تکرار یک کلمه در یک سند را در مقابل تعداد تکرار آن کلمه در مجموعهٔ تمام اسناد در نظر می‌گیریم. در این روش اگر واژه‌ای در اسناد زیادی ظاهر شود، وزن واژه کمتر خواهد شد و واژه اهمیت کمتری پیدا خواهد کرد.

$$a_{td}=tf_{td}*\log\frac{N}{n_a} \quad (۲)$$

در فرمول (۱) tf_{td} نمایانگر تعداد دفعات ظاهرشدن واژه t در سند d و N بیان‌گر تعداد تمامی اسناد موجود در مجموعه‌داده است و اما n_d نشان‌دهندهٔ تعداد اسنادی است که واژه t در آن وجود دارد [23]. پس از محاسبهٔ TF-IDF آن دسته از واژگانی که TF-IDF بیشتری نسبت به سایر واژگان دارند، به‌عنوان واژگان کلیدی انتخاب می‌شوند [21].

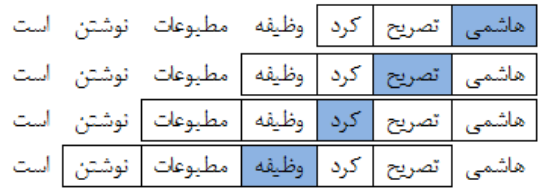
۴-۲- وزن‌دهی به روش skip-gram

در الگوریتم Word2Vec هدف از آموزش مدل skip-gram این است که واژه‌هایی را پیدا کنیم که برای پیش‌بینی واژگان همسایه در یک جمله یا سند مفید باشند[24]. همان‌طور که در فرمول (۳) مشاهده می‌شود هدف از روش skip-gram بیشینه‌کردن میانگین لگاریتم احتمال آمدن واژگان در اطراف یک واژه است.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c,\, j\neq 0}\log p\left(w_{t+j}|w_t\right) \quad (۳)$$

در فرمول (۳) c تعداد مجموعه‌دادهٔ ورودی را برای آموزش نشان می‌دهد که با افزایش این مقدار، دقت و هزینه بالاتر می‌رود و $p(w_{t+j}|w_t)$ بر اساس فرمول (۱) به‌دست می‌آید که v_w و v'_w به‌ترتیب بردارهای ورودی و خروجی واژگان هستند و W تعداد تمام واژگان است.

شکل (۲) معماری مدل skip-gram را نشان می‌دهد، در این مدل با داشتن یک واژه باید واژگان بعدی را پیش‌بینی کنیم[24]. ازآنجایی‌که واژگان دورتر نسبت به واژگان نزدیک‌تر رابطه موضوعی یا مفهومی کمتری دارند[25]، پس لازم است که در ابتدا بیشینه فاصلهٔ بین واژهٔ فعلی و پیش‌بینی‌شده در یک جمله را مشخص کنیم، ما این فاصله را با نام اندازه پنجره[3] می‌شناسیم. در شکل (۱) این فاصله برابر دو در نظر گرفته شده است، یعنی با داشتن یک واژه باید دو لغت قبل‌تر و دو لغت بعدتر را پیش‌بینی کنیم.



(شکل-۱): اندازهٔ پنجره در مدل Skip-gram
(Figure-1): Window Size in Skip-gram Model

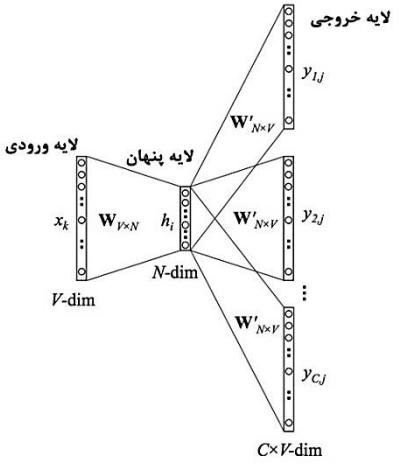
در شکل (۲) و در لایه ورودی، x نشان‌دهندهٔ واژهٔ ورودی در نمونهٔ آموزشی است که به‌صورت یک بردار $1\times V$ به شبکهٔ عصبی داده می‌شود و $[y_1,\dots,y_c]$ یک بردار به طول $1\times V$ است که مربوط به واژگان خروجی در نمونه آموزشی با هدف کمینه‌سازی مجموع خطای پیش‌بینی همه واژگان همسایه در لایه خروجی است، ماتریس $W_{V\times N}$، ماتریس وزن بین لایهٔ ورودی و لایهٔ پنهان است که سطر i^{th} نشان‌دهندهٔ وزن‌های مربوط به کلمهٔ i^{th} در بین مجموعه لغات است. ماتریس وزن W همان چیزی است که ما علاقه‌مند به یادگیری آن هستیم؛ زیرا پس از پایان آموزش شبکهٔ عصبی شامل بردار تمام واژگان موجود در مجموعه‌واژگان است. هر بردار کلمهٔ خروجی دارای یک ارتباط $N\times V$ با ماتریس خروجی $\acute{W}$ دارد.

برخی از پارامترهایی که در این روش مورداستفاده قرارگرفته‌اند بدین شرح است:

۱. به‌صورت پیش‌فرض بیشینه فاصلهٔ بین واژهٔ فعلی و پیش‌بینی‌شده در یک جمله در الگوریتم مقدار سه در نظر گرفته شده است؛ اما با توجه به این که در زبان فارسی عباراتی با طول چهار واژه داریم و فروانی عباراتی با بیشتر از چهار واژه به کمینه می‌رسد، بنابراین بیشینه فاصلهٔ بین واژهٔ فعلی و پیش‌بینی‌شده در یک جمله چهار در نظر گرفته شده است.
۲. کمینه فرکانس واژگان به‌منظور نادیده‌گرفته‌شدن در فرایند آموزش عدد یک در نظر گرفته شده است.
۳. کمینه فرکانس واژگان به‌منظور نادیده‌گرفته‌شدن در فرایند آموزش عدد یک در نظر گرفته شده است.
۴. طول ابعاد بردار ویژگی به اندازهٔ تعداد واژگان موجود در متن در نظر گرفته شده است.

[1] Representation
[2] Term Frequency inverse Document Frequency
[3] Window Size

[DOI: 10.52547/jsdp.18.1.60]



(شکل-۲): معماری مدل skip-gram [26]

(Figure-2): Skip-gram Model Architecture [26]

۵. کمینه فرکانس واژگان به‌منظور نادیده‌گرفته‌شدن در فرایند آموزش عدد یک در نظر گرفته شده است.

۶. طول ابعاد بردار ویژگی به اندازهٔ تعداد واژگان موجود در متن در نظر گرفته شده است.

۵- راه‌حل پیشنهادی

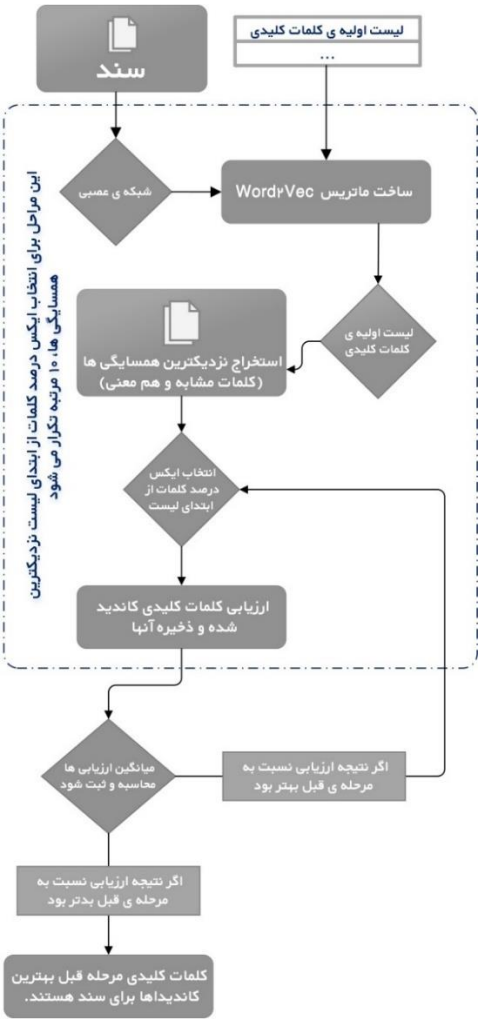
در روش پیشنهادی، واژگان کلیدی از بردار ویژگی‌های یک سند و همبستگی معنایی و موضوعی واژگان به‌دست می‌آیند. تمام ویژگی‌های یک سند تحت بردار ویژگی‌ها با استفاده از شبکهٔ عصبی پیشرو ۱ و از طریق الگوریتم Word2Vec در قالب یک ماتریس نمایش داده می‌شوند. با توجه به نتایجی که در [25] آمده است، در این مقاله برای ساخت بردار واژگان از روش Skip-gram استفاده می‌کنیم. از مزایای این روش برای ساخت بردار ویژگی‌ها نسبت به روش TF-IDF این است که از ماتریس نهایی ایجادشده با این روش می‌توان میزان ارتباط بین واژگان را محاسبه کرد. محاسبه میزان ارتباط بین واژگان با ترکیب بردارهای آنها با یکدیگر امکان‌پذیر است. برای مثال سه واژه "فرانسه"، "ایران" و "تهران" را در نظر بگیرید با توجه به رابطه "بردار(فرانسه) ـ بردار(ایران) + بردار(تهران) =؟" به جواب "بردار(پاریس)" خواهیم رسید.

معماری روش پیشنهادی در شکل (۱) بیان شده است. در روش پیشنهادی فهرست اولیه واژگان کلیدی به شرط داشتن ارتباط قوی با محتوای سند که همان واژگان کلیدی آموزنده هستند، توسط خبره مشخص می‌شود. مجموعه اسنادی که توسط خبره واژگان کلیدی آنها مشخص شده حدود هشتصد سند است. به‌منظور ارزیابی واژگان کلیدی نامزدشده بعد از استخراج واژگان کلیدی با استفاده از معیار F ارزیابی صورت می‌پذیرد.

[1] Feed forward Neural Network

۶- آموزش شبکهٔ عصبی

در آموزش شبکهٔ عصبی نمی‌توانیم یک واژه را به‌عنوان یک رشته متنی به شبکهٔ عصبی بدهیم، بدین منظور لازم است ابتدا لغت‌نامهٔ واژگان را به‌دست آوریم، برای مثال V واژهٔ منحصربه‌فرد داریم که می‌خواهیم از بین آن‌ها واژهٔ "ایران" را به‌عنوان ورودی به شبکهٔ عصبی بدهیم، تنها کافی است که یک بردار را به طول $1\times V$ که تمام خانه‌های آن با عدد صفر پر شده‌اند، به‌جز خانه‌ای را که مربوط به واژهٔ "ایران" است، به‌عنوان ورودی به شبکهٔ عصبی بدهیم.



(شکل-۳): معماری روش پیشنهادی

(Figure 3): Suggested Architecture for Improving Keywords

$$[0\ \ 0\ \ 0\ \ 1\ \ 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10\ \ 12\ \ 19]$$

(شکل-۴): بردار ویژگی‌ها در آموزش شبکه عصبی

(Figure-4): Features Vector in Neural Network Training

برای مثال در جملهٔ "بزرگ‌ترین دشمن اسرائیل ایران است"، واژهٔ "ایران" چهارمین واژه است، در شکل

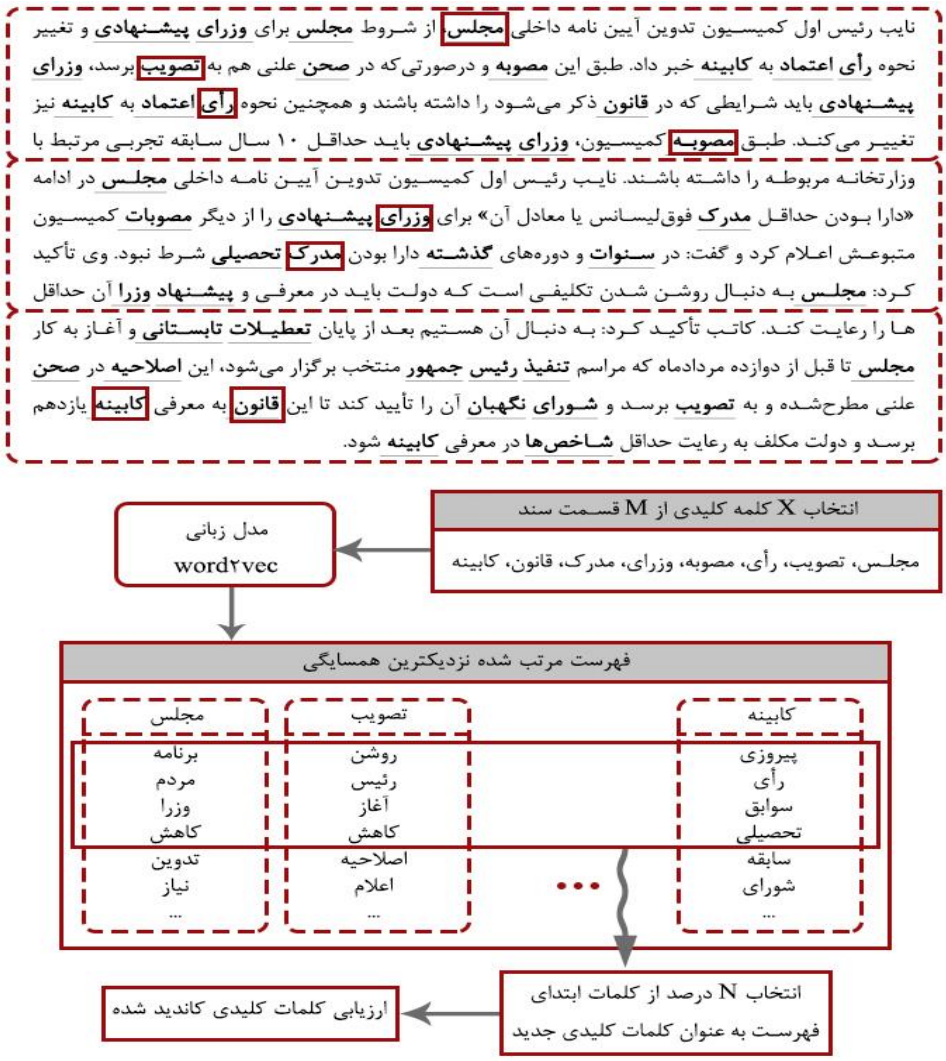
[DOI: 10.52547/jsdp.18.1.60]

(۴) نیز در چهارمین خانه از ماتریس 1×5 عدد یک بیان‌گر این موضوع است. تعداد ویژگی‌ها، پارامتری هستند که می‌توان برحسب نیاز کم و زیاد کرد؛ در این مثال ما سه ویژگی را برای ماتریس دوم در نظر گرفته‌ایم، درواقع این ماتریس همان وزن‌های لایه نخست است که تعداد سطرهای آن بیان‌گر تعداد واژگان لغت‌نامه است که سطر چهارم از این ماتریس درواقع ویژگی‌های استخراج‌شده برای واژه "ایران" است. ماتریس نخست به‌صورت یک شاخص عمل می‌کند و درواقع بیان‌گر شماره سطر واژه مورد نظر در ماتریس وزن است که ویژگی‌های متناظر با آن در خروجی نمایش داده شده است.

(شکل-۵): نحوهٔ تولید بردار خروجی در فرایند یادگیری شبکه عصبی

(Figure-5): How to generate an output vector in the process of learning the neural network

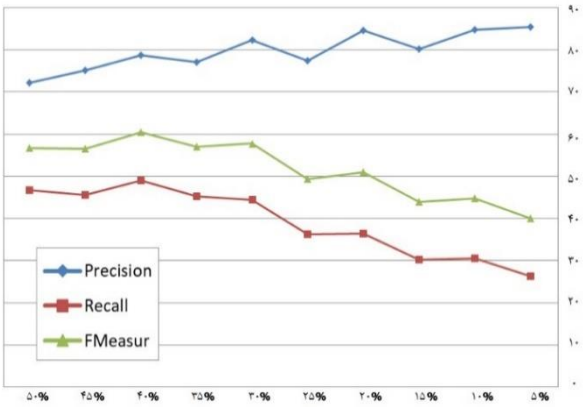
در شکل (۵) از چپ به راست به‌ترتیب بردار ورودی ($1\times V$)، ماتریس وزن لایه ورودی به لایه پنهان و همچنین تولید بردارهای نهایی متن ($V\times D$)، لایه پنهان ($1\times D$)، ماتریس وزن لایه پنهان به لایه خروجی ($D\times V$) و بردار خروجی ($1\times V$) در فرایند یادگیری شبکه عصبی نشان داده شده است. این شکل به‌خوبی نشان می‌دهد که چگونه ابعاد ماتریس ورودی در لایه‌ی پنهان کمتر شده است، D تعداد ابعاد بردارهای واژه را نشان می‌دهد، در واقع در این مثال در لایهٔ پنهان یک بردار به طول V به یک بردار به طول تعداد ویژگی‌ها تبدیل شده است. در مرحلهٔ بعد به‌منظور تولید بردار خروجی لازم است که ماتریس $1 \times D$ را در ترانهادهٔ ماتریس ویژگی‌ها ضرب کنیم. نکته‌ای که لازم است، گفته شود این است که مجموع احتمالاتی که در بردار خروجی می‌آیند در هر سطر برابر یک است. در پایان کار بردارهای نهایی واژگان، همان بردار موردنظر ما است.



(شکل-۶): نحوه شناسایی واژگان کلیدی جدید با استفاده از روش پیشنهادی

(Figure-6): How to identify new keywords using the suggested method

[DOI: 10.52547/jsdp.18.1.60]



(نمودار-۱): ارزیابی روش پیشنهادی(محور افقی درصد واژگان انتخاب شده از نزدیک‌ترین فهرست همسایگی و محور عمودی درصد ارزیابی)

(Chart-1): Assessment of the suggested method (horizontal axis The percentage of words selected from the list of nearest neighbors and the vertical axis is the percentage of evaluation)

(جدول-۱): ارزیابی روش پیشنهادی

(Table-1): Assessment of the suggested method

	درصد انتخاب	ارزیابی به ازای تعداد دفعات تکرار فرایند انتخاب واژگان کاندید										میانگین
		1	2	3	4	5	6	7	8	9	10	
دقت	5%	93.33	90.91	91.67	80.00	81.82	91.67	76.92	91.67	78.57	76.92	85.35
	10%	78.57	84.62	92.31	88.24	85.71	78.57	80.00	87.50	92.31	80.00	84.78
	15%	93.75	73.33	86.67	80.00	78.57	84.62	81.25	75.00	86.67	62.50	80.24
	20%	94.12	80.00	88.24	88.24	73.68	83.33	84.21	87.50	77.78	88.24	84.53
	25%	73.68	83.33	88.24	80.95	72.22	70.00	73.68	73.68	68.42	88.89	77.31
	30%	85.71	80.95	91.30	83.33	85.71	85.71	75.00	71.43	81.82	81.82	82.28
	35%	84.00	75.00	75.00	79.17	72.73	80.00	69.57	86.36	81.82	66.67	77.03
	40%	79.16	76.92	76.00	73.08	79.17	82.61	75.00	81.48	83.33	80.77	78.75
	45%	75.00	72.00	83.33	70.00	74.07	77.27	80.00	75.00	66.67	77.78	75.11
	50%	78.26	74.07	70.37	69.23	73.91	70.37	76.92	70.37	76.92	60.71	72.11
بازخوانی	5%	35.00	25.00	27.50	20.00	22.50	27.50	25.00	27.50	27.50	25.00	26.25
	10%	27.50	27.50	30.00	37.50	30.00	27.50	30.00	35.00	30.00	30.00	30.50
	15%	37.50	27.50	32.50	30.00	27.50	27.50	32.50	30.00	32.50	25.00	30.25
	20%	40.00	30.00	37.50	37.50	35.00	37.50	40.00	35.00	35.00	37.50	36.50
	25%	35.00	37.50	37.50	42.50	32.50	35.00	35.00	35.00	32.50	40.00	36.25
	30%	45.00	42.50	52.50	50.00	45.00	45.00	37.50	37.50	45.00	45.00	44.50
	35%	52.50	45.00	45.00	47.50	40.00	50.00	40.00	47.50	45.00	40.00	45.25
	40%	47.50	47.50	47.50	47.50	47.50	45.00	55.00	50.00	52.50	49.00	
	45%	45.00	45.00	50.00	35.00	50.00	42.50	50.00	45.00	40.00	52.50	45.50
	50%	47.50	45.00	47.50	45.00	42.50	47.50	50.00	47.50	50.00	46.75	
معیارF	5%	50.90	39.22	42.31	32.00	35.29	42.31	37.74	42.31	40.74	37.74	40.05
	10%	40.74	41.51	45.28	52.63	44.44	40.74	43.64	50.00	45.28	43.64	44.79
	15%	53.57	40.00	47.27	43.64	40.74	41.51	46.43	42.86	47.27	35.71	43.90
	20%	56.14	43.64	52.63	52.63	47.46	51.72	54.24	50.00	48.28	52.63	50.94
	25%	47.45	51.72	52.63	55.74	44.83	46.67	47.46	47.46	44.07	55.17	49.32
	30%	59.02	55.74	66.67	62.50	59.02	59.02	50.00	49.18	58.06	58.06	57.73
	35%	64.61	56.25	56.25	59.37	51.61	61.54	50.79	61.29	58.06	50.00	56.98
	40%	59.37	60.61	58.46	57.58	59.37	60.32	56.25	65.67	62.50	63.64	60.38
	45%	56.25	55.38	62.50	46.67	59.70	54.84	61.54	56.25	50.00	62.69	56.58
	50%	57.14	59.70	56.72	54.55	53.97	56.72	60.61	56.72	60.61	50.00	56.67

[DOI: 10.52547/jsdp.18.1.60]

۷- نزدیک‌ترین همسایگی

با استفاده از مدل word2Vec که خروجی یادگیری شبکه‌ی عصبی است، می‌توان واژگانی را که رابطهٔ معنایی و همبستگی موضوعی با یکدیگر دارند، را پیدا کرد. در این مقاله یک پیکرهٔ هشتصدتایی را از اخبار که واژگان کلیدی آنها توسط خبره استخراج شده‌اند، آماده کردیم، سپس با استفاده از روش اعتبارسنجی متقابل[1] که یک روش برای ارزیابی مدل است و بیان‌گر میزان مستقل‌بودن و تعمیم‌پذیری نتایج یک تحلیل آماری بر روی یک مجموعه‌داده نسبت به داده‌های آموزشی است، X واژه کلیدی را از M قسمت سند شناسایی و انتخاب می‌کنیم. برای این منظور ابتدا متن سند به‌صورت تصادفی به M قسمت تقسیم می‌شود که با توجه به طول سندهای مورداستفاده در این مقاله، بیشینه مقدار M در روش پیشنهادی مقدار پنج در نظر گرفته شده است. انتخاب X واژه کلیدی این‌گونه است که به‌صورت تصادفی از هر قسمت تعدادی واژه کلیدی را که از قبل توسط خبره برچسب‌گذاری شده است، انتخاب می‌کنیم. اولویت انتخاب با واژگان کلیدی است که در قسمت‌های بیشتری از متن برچسب خورده وجود داشته باشند. حال با استفاده از مدل word2Vec که به‌ازای سند مربوطه ساخته شده است و این X واژهٔ کلیدی کاندیدشده، نزدیک‌ترین واژگان را به این واژه‌ها می‌یابیم. این فهرست جدید از واژگان را فهرست نزدیک‌ترین همسایگی‌ها می‌نامیم. این فهرست را بر اساس امتیاز واژگان که معرف میزان شباهت معنایی و موضوعی آن‌ها با واژگان نامزدشدهٔ اولیه هستند به‌صورت نزولی مرتب می‌کنیم. حال از بین نزدیک‌ترین همسایگی‌ها، پنج درصد از واژگان ابتدای فهرست را به‌عنوان واژگان کلیدی انتخاب، سپس دقت، بازخوانی و معیارF را محاسبه و این کار را به‌ازای پنج درصد واژگان ابتدای فهرست، ده بار تکرار و درنهایت به‌ازای دقت، بازخوانی و معیارF، میانگین این ده مرحله را ثبت می‌کنیم. این فرایند را به‌ازای انتخاب درصدهای مختلفی از فهرستِ نزدیک‌ترین همسایگی‌ها، تا جایی ادامه می‌دهیم که معیارF به بالاترین مقدار خود برسد. در شکل (۶) به‌خوبی مراحل کار برای انتخاب واژگان کلیدی نامزد و ارزیابی آنها نشان داده شده است، همچنین در نمودار (۱) و جدول (۱) به‌خوبی نتایج کار قابل مشاهده است.

۸- نتایج و بحث

در این مقاله با استفاده از مدل Skip-gram برای آموزش شبکهٔ عصبی و الگوریتم Word2Vec به‌منظور ساخت بردار ویژگی‌ها، توانستیم بردار ویژگی‌های تمام واژگان موجود در یک متن را استخراج کنیم، با این تفاوت که در روش به‌کار گرفته‌شده نسبت به سایر روش‌ها از ماتریس نهایی می‌توان رابطهٔ معنایی و همبستگی موضوعی بین لغات را محاسبه کرد. از این ویژگی استفاده کردیم و با بهره‌گیری از یک فهرست اولیه از واژگان کلیدی توانستیم واژگان مشابه با آن‌ها را در قالب فهرست نزدیک‌ترین همسایگی‌ها استخراج کنیم. فهرست به‌دست‌آمده را به‌صورت نزولی مرتب و از ابتدای فهرست، درصدهای مختلفی از واژگان را انتخاب و به‌ازای هر درصد، ۱۰ مرتبه فرایند آموزش شبکهٔ عصبی و ساخت ماتریس همبستگی و استخراج فهرست نزدیک‌ترین همسایگی‌ها را تکرار و درنهایت میانگین دقت، فراخوانی و معیارF را محاسبه می‌کنیم. این کار را تا جایی ادامه می‌دهیم که به بهترین نتایج در ارزیابی دست یابیم، نتایج نشان می‌دهند که به‌ازای انتخاب حداکثر چهل درصدِ واژگان از ابتدای فهرستِ نزدیک‌ترین همسایگی‌ها، نتایج مورد قبولی به دست می‌آید. نتایجِ آزمایش‌ها نشان می‌دهد که دقت روش پیشنهادی ۷۸ درصد خواهد بود.

نتایج به‌دست‌آمده در روش پیشنهادی با [27] مقایسه و در جدول (۲) نشان داده شده است.

(جدول-۲): مقایسه نتایج حاصل از روش پیشنهادی با حدآستانه ۵ و ۱۰ درصد

(Table-2): Comparison of the results of the proposed method with thresholds of 5% and 10%

معیار ارزیابی	روش پیشنهادی		روش [27]	
	٪۵	٪۱۰	٪۵	٪۱۰
دقت	۸۵/۳۵	۸۴/۷۸	۴۹/۲۳	۲۶/۱۹
فراخوانی	۲۶/۲۵	۳۰/۵۰	۶۰/۱۴	۶۷/۹۰
معیار F	۴۰/۰۵	۴۴/۷۹	۵۳/۹۲	۳۷/۸۰

۹- نتیجه‌گیری و کارهای آینده

در این مقاله سعی بر آن شد تا روشی جدید و مستقل از زبان برای بهبود واژگان کلیدی استخراج‌شده از متون ارائه شود که علاوه‌بر درنظرگرفتن فراوانی واژگان، از روابط و همبستگی بین واژگان نیز برای بهبود کیفیت استفاده می‌کند، همان چیزی که در روش‌های قبلی که در این مقاله و در بخش کارهای پیشین به آن‌ها اشاره شده است، مورد توجه قرار نگرفته است.

به‌عنوان پیشنهاد جهت پروژه‌های آینده می‌توان موارد زیر را مطرح کرد:

[1] Cross-validation

[DOI: 10.52547/jsdp.18.1.60]

۱. استفاده از روش‌های رتبه‌بندی واژگان در فهرست نزدیک‌ترین همسایگی‌ها؛
۲. استفاده از ویژگی هم رخدادی واژگان برای شناسایی عبارات کلیدی؛
۳. استفاده از ضریب وابستگی بین واژگان با توجه به ترادف و تضاد بین آنها.

۱۰- مراجع 10- References

[1] F. Liu, X. Huang, W. Huang, and S. X. Duan, "Performance Evaluation of Keyword Extraction Methods and Visualization for Student Online Comments," Symmetry, vol. 12, no. 11, p. 1923, 2020.

[2] H. Yan, Q. He, and W. Xie, "Crnn-Ctc Based Mandarin Keywords Spotting," *in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: IEEE*, pp. 7489-7493 .

[3] Y. Zhang, M. Tuo, Q. Yin, L. Qi, X. Wang, and T. Liu, "Keywords extraction with deep neural network model ", *Neurocomputing*, vol. 383, pp. 113-121, 2020.

[4] M. Mohammadi and M. Analouyi, "Keyword extraction in Persian documents," *presented at the 13th Conference of Iran Computer Association*, Kish, Iran, 2007.

[5] H. Veisi, N. Aflaki, and P. Parsafard, "Variance-based features for keyword extraction in Persian and English text documents," *Scientia Iranica*, vol. 27, no. 3, pp. 1301-1315, 2020.

[6] C. Zhang, "Automatic keyword extraction from documents using conditional random fields," *Journal of Computational Information Systems*, vol. 4, no. 3, pp. 1169-1180, 2008.

[7] X. Wan and J. Xiao, "Single Document Keyphrase Extraction Using Neighborhood Knowledge," *in AAAI*, 2008, vol. 8, pp. 855-860 .

[8] D. B. Bracewell, F. Ren, and S. Kuriowa, "Multilingual single document keyword extraction for information retrieval," *in 2005 International Conference on Natural Language Processing and Knowledge Engineering, 2005: IEEE*, pp. 517-522 .

[9] P. D. Turney, "Learning to extract keyphrases from text," arXiv preprint cs/021, 2002, 2013.

[10] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 01, pp. 157-169, 2004.

[11] S. Rose, D. Engel, N .Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text mining: applications and theory*, vol. 1, pp. 1-20, 2010.

[12] J. Wang, H. Peng, and J.-s. Hu, "Automatic keyphrases extraction from document using neural network," *in Advances in Machine Learning and Cybernetics: Springer*, 2006, pp. 633-641.

[13] A. Ahmadi and T. Hosseinkhah, "Extract keywords from a text using neural networks," presented at the 10th International Conference on Industrial Engineering, Tehran, Iran Industrial Engineering Association, Amirkabir University of Technology, 2013.

[14] S. De Deyne, S. Verheyen, and G. Storms, "Structure and organization of the mental lexicon: A network approach derived from syntactic dependency relations and word associations ",*in Towards a theoretical framework for analyzing complex linguistic networks: Springer*, 2016, pp. 47-79.

[15] E. L. Lin and G. L. Murphy, "Thematic relations in adults' concepts," *Journal of experimental psychology: General*, vol. 130, no. 1, p. 3, 2001.

[16] F. Liu, D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts," *in Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, 2009, pp. 620-628 .

[17] X. Ao, X. Yu, D. Liu, and H. Tian, "News keywords extraction algorithm based on TextRank and classified TF-IDF," in 2020 *International Wireless Communications and Mobile Computing (IWCMC), 2020: IEEE*, pp. 1364-1369 .

[18] F. Liu, X. Huang, and W. Huang, "Comparing Machine Learning Algorithms to Predict Topic Keywords of Student Comments," *in International Conference on Cooperative Design*, Visualization and Engineering, 2020: Springer, pp. 178-183 .

[19] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257-289, 2020.

[20] J. R. Thomas, S. K. Bharti, and K .S. Babu,

[DOI: 10.52547/jsdp.18.1.60]

"Automatic keyword extraction for text summarization in e-newspapers," *in Proceedings of the international conference on informatics and analytics*, 2016, pp. 1-8 .

[21] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," technometrics, vol. 16, no. 1, pp. 125-127, 1974.

[22] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp .111-133,1974.

[23] M. Stone, "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 44-47, 1977.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[25] C. Manning and R. Socher, "Natural language processing with deep learning," *Lecture Notes Stanford University School of Engineering*, 2017.

[26] J. Hu, S. Li, Y. Yao, L. Yu, G. Yang, and J. Hu, "Patent keyword extraction algorithm based on distributed representation for patent classification," *Entropy*, vol. 20, no. 2, pp. 104, 2018.

[27] H. Omid and S. Saeedeh, Sadidpour, "Automatic extraction of Persian short text keywords using word2vec," *Electronic and cyber defense*, vol. 8, 2, pp. 105-114, 2020.

محمدرضا حسنی‌آهنگر دکترای خود را در رشتهٔ هوش مصنوعی و رباتیک از دانشگاه علم و صنعت در سال ۱۳۹۰ اخذ کرد. تخصص ایشان هوش مصنوعی است و در حال حاضر به‌عنوان عضو هیأت علمی دانشکدهٔ مهندسی رایانه دانشگاه امام حسین(ع) به تدریس دروس هوش مصنوعی مشغول هستند. از سال ۱۳۹۱ ریاست دانشگاه جامع امام حسین(ع) به عهدهٔ ایشان است.

نشانی رایانامه ایشان عبارت است از:

mrhasani @ihu.ac.ir

علی امیری‌جزه کارشناسی‌ارشد خود را در رشتهٔ هوش مصنوعی از دانشگاه امام حسین(ع) در سال ۱۳۹۹ اخذ کرد. حوزهٔ مورد علاقهٔ وی پردازش زبان طبیعی، تحلیل‌های زبانی و داده‌های حجیم و پایان‌نامهٔ او در زمینهٔ شناسایی موجودیت‌های اسمی با استفاده از یادگیری عمیق و رویکرد تقویتی است.

نشانی رایانامه ایشان عبارت است از:

aamirij@ihu.ac.ir

[DOI: 10.52547/jsdp.18.1.60]