Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 11, Number 1, Spring 2022, pp. 1-17
Received: 06/19/2021, Revised: 10/30/2021, Accepted: 01/22/2022

A Novel Hybrid Approach to Finding Meaningful Basis Vectors For Explicit Representation of Word Vectors

Atefe Pakzad[1], Morteza Analoui[2*]
1- School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.
2*- School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.
[1]a_pakzad@comp.iust.ac.ir, [2*]analoui@iust.ac.ir

Corresponding author's address: Morteza Analoui, Faculty of Computer Engineering, Iran University of

Abstract- The main purpose of this study is to represent the semantic word vectors with low dimensions, explicitly. The problem of finding a limited number of meaningful basis vectors for producing explicit semantic word vectors must be solved in such a way that a large accuracy drop is not caused by reducing the dimensions. In this study, we represent a hybrid approach to finding meaningful basis vectors. First, we obtain N basis vectors using the proposed methods: 1- The criterion of word similarity-to-word frequency ratio, 2- Feature selection method based on comparison of distance matrices, 3- Binary weighting method based on PSO algorithm. Then, to take advantage of the expertise of methods 1 and 2 to the same extent, we obtain the first combined basis vectors by combining half of the basis vectors obtained by the criterion of word similarity-to-word frequency ratio with half of the basis vectors selected by the feature selection method. In the next step, we obtain the common context words that have a weight "1" as the common basis vectors produced by the binary weighting method. In the next step, we add the common context words with a weight "1" obtained using the BPSO method to the first combined basis vectors obtained from word similarity-to-word frequency ratio and the feature selection methods. Thus, the second combined basis vectors are obtained, which are meaningful, and each basis vector is equivalent to an informative context word. Therefore, the explicit word vectors produced by meaningful basis vectors can be interpreted. We train the proposed approach using the UkWaC corpus and evaluate it using the word similarity task. Both first and second combined basis vectors improve accuracy. The increase in accuracy is greater in the first combined basis vectors. The evaluation results of explicit word vectors obtained with the first basis vectors show that despite the reduction of word vector dimensions from 5000 to 1511, the Spearman correlation coefficient on MEN, RG-65, and SimLex-999 test sets is increased by 2.47%, 7.39%, and 0.52%, respectively.

Keywords- basis vectors, word vector representation, interpretable word vectors, binary weighting, feature selection, word similarity task.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۱، شماره ۱، بهار ۱۴۰۱ ، صفحه ۱-۱۷
دریافت: ۱۴۰۰/۰۳/۲۹، بازنگری: ۱۴۰۰/۰۸/۰۸، پذیرش: ۱۴۰۰/۱۱/۰۲

ارائه‌ی یک رویکرد ترکیبی جدید برای یافتن بردارهای پایه معنادار جهت بازنمایی صریح بردارهای کلمه

عاطفه پاکزاد[۱]، مرتضی آنالویی*[۲]
۱- دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران.
*۲- دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران.
[1]a_pakzad@comp.iust.ac.ir, [2*]analoui@iust.ac.ir

* نشانی نویسنده مسئول: مرتضی آنالویی، تهران، رسالت، خیابان هنگام، خیابان دانشگاه، دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر.

چکیده- هدف اصلی این پژوهش بازنمایی صریح بردارهای معنایی کلمه با ابعاد کم است. برای تولید بردارهای معنایی کلمه صریح، بایستی مسئله‌ی یافتن تعداد محدودی بردار پایه معنادار به‌گونه‌ای حل شود که با کاهش ابعاد بردارهای کلمه افت دقت زیادی ایجاد نشود. ما در این پژوهش یک رویکرد ترکیبی برای یافتن بردارهای پایه معنادار ارائه می‌کنیم. در ابتدا، N بردار پایه را با روش‌های پیشنهادی ۱-معیار نسبت شباهت کلمه به تکرار کلمه، ۲-انتخاب ویژگی مبتنی بر مقایسه ماتریس‌های فاصله، ۳-وزن‌دهی دودویی مبتنی بر الگوریتم BPSO بدست می‌آوریم. سپس، برای بهره‌گیری از خبرگی روش‌های ۱ و ۲ به میزان یکسان، نیمی از بردارهای پایه بدست آمده با روش معیار نسبت شباهت کلمه به تکرار کلمه را با نیمی از بردارهای پایه انتخاب شده با روش انتخاب ویژگی، ترکیب کرده و بردارهای پایه ترکیبی اول را بدست می‌آوریم. در مرحله بعدی، کلمات بافتار مشترک دارای وزن "۱" بدست آمده با استفاده از روش BPSO را به بردارهای پایه ترکیبی اول حاصل از روش‌های نسبت شباهت کلمه به تکرار کلمه و روش انتخاب ویژگی می‌افزاییم. بنابراین، بردارهای پایه ترکیبی دوم بدست می‌آیند که معنادار هستند و هر بردار پایه معادل یک کلمه بافتار آگاهی‌بخش است. لذا بردارهای کلمه صریح تولید شده با استفاده از بردارهای پایه معنادار، تفسیرپذیر هستند. ما رویکرد پیشنهادی را با استفاده از پیکره UkWaC آموزش داده و بر روی وظیفه شباهت کلمه مورد ارزیابی قرار می‌دهیم. هر دو بردارهای پایه ترکیبی اول و دوم سبب بهبود دقت می‌گردند. این افزایش دقت در بردارهای پایه ترکیبی اول بیشتر است. نتایج حاصل از ارزیابی بردارهای کلمه صریح بدست آمده با بردارهای پایه اول نشان می‌دهد که با وجود کاهش ابعاد بردارهای کلمه از ۵۰۰۰ به ۱۵۱۱، ضریب همبستگی اسپیرمن بر روی مجموعه‌های آزمون MEN، RG-65 و SimLex-999 به ترتیب به میزان ۲.۴۷٪، ۷.۳۹٪ و ۰.۵۲٪ افزایش می‌یابد.

واژه‌های کلیدی: بردارهای پایه، بازنمایی بردارهای کلمه، بردارهای کلمه تفسیرپذیر، وزن‌دهی دودویی، انتخاب ویژگی، وظیفه شباهت کلمه

۱- مقدمه

معناشناسی توزیعی[۱] یک مدل معنا مبتنی بر کاربرد است. این مدل، مبتنی بر این فرضیه است که توزیع آماری بخش‌های زبانی[۲] نقش کلیدی در مشخص کردن رفتار معنایی آن‌ها دارد. بنیان نظری معناشناسی توزیعی با این فرضیه توزیعی شناخته شده است که لغت‌های با بافتارهای[۳] زبانی مشابه دارای معناهای مشابهی هستند و شباهت معنایی بخش‌های زبانی تابعی از توزیع آن‌ها در بافتارهای زبانی است[۱]. تحقیقات سالتن و همکاران [۲] سهم اساسی در توسعه معناشناسی توزیعی داشته است. در این مرجع مجموعه‌ای از سندها با یک ماتریس نمایش داده شده‌اند که سطرها متناظر با بردارهای بخش‌های زبانی و ستون های ماتریس بردارهای متناظر با سندها هستند. هر ورودی ماتریس رخدادهای

یک بخش لغوی در یک سند را ثبت می‌کند. بازنمایی توزیعی یک بخش زبانی، معمولا یک بردار توزیعی است که هم‌رخدادی‌های آن بخش زبانی را با بافتارهای زبانی نمایش می‌دهد؛ از این رو معناشناسی فضای بردار نامیده می‌شود. به صورت عملی، در رویکرد معناشناسی توزیعی معنای یک کلمه در پیکره با جستجوی کلمه‌های دیگری که با آن کلمه در یک بافتار مشخص رخ می‌دهند یاد گرفته می‌شود. توزیع حاصل با یک بردار در فضای بردار معنایی نمایش داده می‌شود. نمایش کلمه‌ها با بردار، روش مناسبی است زیرا مجموعه غنی از روش‌های محاسبه فاصله برداری وجود دارد که امکان آزمایش معیارهای مختلف تشابه کلمه‌ها را ایجاد می‌کند. با توجه به ماهیت هندسی نمایش برداری، نه تنها می‌توان معنای کلمه‌ها را به صورت منفرد مقایسه کرد بلکه می‌توان با استفاده از روش‌های بازیابی اطلاعات برای گروه‌بندی مفاهیم بوسیله موضوع، تمایل، یا دیگر کلاس‌های معنایی نیز از آن بهره برد [۳].

مدل‌های معنایی توزیعی برای تخمین شباهت کلمه‌ها مورد استفاده قرار می‌گیرند. این روش‌ها در سه جنبه مهم با یکدیگر تفاوت دارند: ۱-نوع بافتار، ۲-معیار شباهت، ۳-روشی که مدل فضای کلمه ساخته می‌شود. مدل‌های معنایی توزیعی شامل دو دسته مدل‌های مبتنی بر شمارش[۴] و مدل‌های مبتنی بر پیش‌بینی[۵] هستند. در مدل‌های مبتنی بر شمارش، توزیع‌های کلمه بوسیله بردارهای با ابعاد بالا اما پراکنده تعریف می‌شوند. این مدل‌ها، مدل‌های معنایی توزیعی صریح[۶] نامیده می‌شوند، زیرا هر یک از مولفه‌های بردار کلمه، متناظر با یک کلمه زبان طبیعی است [۴].

مدل‌های مبتنی بر پیش‌بینی عمدتا از روش‌های شبکه عصبی بهره می‌برند. بردارهای حاصل از روش‌های مبتنی بر پیش‌بینی ضمنی[۷] هستند و تعبیه کلمه[۸] نامیده می‌شوند. بنابراین مولفه‌های بردارهای کلمه فاقد معنی بوده و تفسیرپذیر نیستند [۵]. هدف تعبیه‌های کلمه، نگاشت کلمه‌های یک داده متنی به فضای با ابعاد کوچکتر با مقدار پیوسته به منظور در نظر گرفتن اطلاعات معنایی و نحوی است. کاربردهای بسیار زیادی برای تعبیه‌های کلمه به ویژه در وظایف پردازش زبان طبیعی وجود دارد که تعبیه‌های کلمه به عنوان داده ورودی یا ویژگی های داده متنی به طور مستقیم مورد استفاده قرار می‌گیرند [۶]. تعبیه‌های کلمه در وظایفی نظیر تحلیل معنایی[۹] [۷ و ۸]، تحلیل نحوی[۱۰] [۹]، برچسب‌زنی اجزای سخن[۱۱] [۱۰]، تحلیل تمایل[۱۲] [۱۳-۱۱]، شناسایی موجودیت‌های اسمی[۱۳] [۱۴]، استخراج کلمه‌های کلیدی[۱۴] [۱۵]، سیستم‌های توصیه‌گر[۱۵] [۱۶ و ۱۷]، بازیابی اطلاعات[۱۶] [۱۸] و غیره مورد استفاده قرار می‌گیرند. بردارهای کلمه ضمنی حاصل از مدل‌های مبتنی بر پیش‌بینی در مقایسه با بردارهای کلمه صریح، دقت و کارایی بالایی دارند، اما چون مولفه‌های آن‌ها معنادار و تفسیر پذیر نیست، نمی‌توانند اطلاعات متن مورد نظر را منعکس کنند. همچنین در برخی موارد باعث ایجاد بیش‌برازش[۱۷] می‌شوند. بردارهای معنایی صریح حاصل از مدل‌های مبتنی بر شمارش تفسیر پذیر هستند و اطلاعات متن مورد نظر را منعکس می‌کنند. متاسفانه این بردارها دارای ابعاد بالایی هستند که سبب ایجاد محدودیت‌های محاسباتی می‌شوند. برای حل مشکل ابعاد بالا، معمولا پژوهشگران از روش‌های کاهش ابعاد نظیر تجزیه مقدار منفرد (SVD)، تحلیل مولفه‌های اصلی[۱۸] (PCA) و فاکتورگیری نامنفی ماتریس[۱۹] (NMF) بهره می‌برند. بردارهای کلمه انتقال‌یافته حاصل دارای ابعاد پایینی هستند ولیکن تفسیرپذیر نیستند؛ زیرا مولفه‌های بردار معادل معنایی ندارند [۱].

۱-۱- بدست آوردن شباهت معنایی کلمه‌ها

برای بسیاری از مسائل پردازش زبان طبیعی بدست آوردن شباهت معنایی[۲۰] یا ارتباط معنایی[۲۱] کلمه‌ها ضروری است؛ یعنی به چه میزان مفاهیم مشخص شده توسط کلمه‌ها به ترتیب به دیگر کلمه‌ها شبیه یا مرتبط هستند. ارتباط معنایی طیف وسیعی از روابط بین مفاهیم (شامل شباهت هم می‌شود) را در نظر می‌گیرد. شباهت معنایی این نکته را در نظر می‌گیرد که چقدر این دو مفهوم یکسان هستند. برای مثال، کلمه‌های پستچی و نامه دارای ارتباط معنایی زیادی هستند؛ چون معمولا نامه‌ها توسط پستچی تحویل داده می‌شوند. اما این دو واژه شبیه نیستند؛ زیرا نشان‌دهنده مفاهیم متفاوتی هستند [۱۹]. برای بدست آوردن شباهت معنایی و ارتباط معنایی بین کلمه‌ها ابتدا بردارهای معنایی کلمه‌ها محاسبه می‌شود. سپس، با استفاده از معیارهای اندازه‌گیری شباهت، شباهت یا ارتباط معنایی کلمه‌ها محاسبه می‌شود. معمولا معیار شباهت کسینوسی برای اندازه‌گیری میزان شباهت یا ارتباط معنایی واژه‌ها مورد استفاده قرار می‌گیرد. در ادامه، به طور اجمالی نحوه بدست آوردن بردارهای معنایی کلمه‌ها را با روش‌های مبتنی بر شمارش توضیح می‌دهیم.

۱-۲- ساخت بردارهای کلمه در مدل‌های مبتنی بر شمارش

در ابتدا، برای یک پیکره خاص، تعدادی از پرتکرارترین کلمه‌های پیکره را به عنوان کلمه‌های بافتار در نظر می‌گیریم. همچنین با استفاده از تعداد زیادی از کلمه‌های پرتکرار پیکره یک واژه‌نامه می‌سازیم. هر یک از کلمه‌های واژه‌نامه یک کلمه هدف[۲۲] است. برای ساخت بردار کلمه هدف، بایستی هم‌رخدادی کلمه هدف با

کلمه‌های بافتار را محاسبه کرد. پس از تعیین کلمه‌های بافتار و کلمه‌های هدف، بایستی با استفاده از ماتریس هم‌رخدادی M بردارهای کلمه‌های هدف را بسازیم. برای ساخت ماتریس هم‌رخدادی، کلمه‌های بافتار متناظر با ستون‌های ماتریس و کلمه‌های هدف متناظر با سطرهای ماتریس هستند. هر مولفه m_{ij} ماتریس، تعداد هم‌رخدادی کلمه هدف iام (t_i) و کلمه بافتار jام (c_j) را مشخص می‌کند. هر سطر ماتریس هم‌رخدادی متناظر با بردار کلمه هدف متناظر با آن سطر است [۲۰].

توزیع معنایی کلمه، برداری در فضای برداری است که کلمه‌های بافتار، بردارهای پایه[۲۳] آن را مشخص می‌کنند. فضاهای برداری مدل‌های معنایی توزیعی، دارای بردارهای پایه متعامد هستند. بردار معنایی یک کلمه با برهم‌نهی وزن‌دار بردارهای پایه به صورت زیر مشخص می‌شود:

$$\overrightarrow{word} = \sum_h w_h \overrightarrow{n_h} \qquad (۱)$$

که بردارهای واحد[۲۴] متعامد $\overrightarrow{n_h}$ ، پایه فضای برداری هستند که معنا در آن قرار دارد. w_h وزن مرتبط با بردار پایه $\overrightarrow{n_h}$ است. این بردارهای پایه با کلمه‌های بافتار نشان داده می‌شوند [۳]. ابتدایی‌ترین شکل w_h در واقع همان شمارش‌های هم‌رخدادی است، اما برای کاهش بایاس تکرار یک تابع وزن‌دهی بر روی اعداد هم‌رخدادی خام اعمال می‌شود [۲۱]؛ زیرا شمارش‌های هم‌رخدادی خام به خوبی کار نمی‌کنند و مدل‌های معنایی توزیعی با اعمال تغییر شکل بر روی بردارهای خام به کارایی بالاتری دست پیدا می‌کنند [۴]. در روش‌های مبتنی بر شمارش از یک ماتریس هم‌رخدادی تنک M با ابعاد بالا برای بازنمایی معنایی کلمه‌ها استفاده می‌شود. هر سطر معرف یک کلمه هدف t واژه‌نامه و هر ستون یک کلمه بافتار بالقوه است. مقدار هر سلول ماتریس m_{ij} نیز ارتباط بین کلمه هدف t_i و کلمه بافتار c_j را مشخص می‌کند. یک معیار رایج برای ارتباط[۲۵]، اطلاعات متقابل نقطه‌ای[۲۶](PMI) است. PMI به عنوان نسبت لگاریتمی بین احتمال توأم کلمه‌های t و c و حاصلضرب احتمالات حاشیه‌ای آن‌ها تعریف می‌شود و به صورت زیر تخمین زده می‌شود [۲۲]:

$$PMI(t,c) = log\frac{\hat{p}(t,c)}{\hat{p}(t)\hat{p}(c)} = log\frac{\#(t,c).|D|}{\#(t).\#(c)} \qquad (۲)$$

$\#(t,c)$ تعداد جملاتی از پیکره است که کلمه هدف t با کلمه هدف c هم‌رخداد است. $\#(t)$ تعداد جملات پیکره دارای کلمه هدف t است و $\#(c)$ نشان‌دهنده تعداد جملاتی از پیکره دارای کلمه بافتار c است. در سطرهای ماتریس PMI (M^{PMI}) مولفه‌های زیادی از زوج‌های کلمه هدف-کلمه بافتار (t,c) وجود دارد که هرگز در پیکره مشاهده نشده‌اند و در نتیجه:

$$PMI(t,c) = log0 = -\infty \qquad (۳)$$

یک رویکرد رایج، جایگزینی M^{PMI} با M_0^{PMI} در مواقعی است که کلمه هدف t با کلمه بافتار c هم‌رخداد نبوده است. یعنی عبارت منفی بینهایت در ماتریس هم‌رخدادی M^{PMI} با عدد صفر جایگزین می‌شود و ماتریس M_0^{PMI} ایجاد می‌شود. یک رویکرد سازگارتر استفاده از PMI مثبت (PPMI) است که مقادیر منفی با عدد صفر جایگزین می‌شوند:

$$PPMI(t,c) = \max(PMI(t,c),0) \qquad (۴)$$

بولیناریا و لوی [۲۳] نشان داده‌اند که M^{PPMI}، M_0^{PMI} را در وظیفه‌های شباهت معنایی بهبود می‌بخشد.

ما در این پژوهش، سه روش ۱-انتخاب بر اساس معیار جدید نسبت شباهت کلمه‌ها به تکرار کلمه‌ها، ۲-انتخاب ویژگی براساس مقایسه ماتریس‌های فاصله و ۳-وزن‌دهی دودویی مبتنی بر الگوریتم BPSO را برای بدست آوردن بردارهای پایه معنی‌دار اولیه پیشنهاد می‌کنیم. سپس بردارهای پایه حاصل از روش‌های ذکر شده را ترکیب کرده و مجموعه بردارهای پایه نهایی را بدست می‌آوریم. در ادامه با استفاده از بردارهای پایه نهایی، بردارهای کلمه با ابعاد پایین صریح را برای کلمه‌های واژه‌نامه تولید می‌کنیم. سپس، شباهت بردارهای کلمه را بر روی مجموعه‌های آزمون با ضریب همبستگی اسپیرمن مورد ارزیابی قرار می‌دهیم. نتایج بدست آمده نشان می‌دهد با کاهش تعداد بردارهای پایه از ۵۰۰۰ به ۱۵۰۰، نه تنها شاهد کاهش ضریب همبستگی اسپیرمن مجموعه‌های آزمون نبوده‌ایم بلکه افزایش قابل ملاحظه‌ای در ضریب همبستگی اسپیرمن ایجاد می‌شود.

ادامه مقاله به صورت زیر تنظیم شده است: در بخش ۲ مطالعات پیشین انجام شده روی مدل‌های توزیعی معنایی را مورد بررسی قرار می‌دهیم. در بخش ۳ رویکرد ترکیبی پیشنهادی برای بدست آوردن بردارهای پایه نهایی را شرح می‌دهیم. در بخش۴ تنظیمات مورد نیاز برای آزمایش‌های عملی را توضیح می‌دهیم. در بخش ۵ نتایج آزمایش‌ها را ارائه کرده و مورد بررسی و تجزیه و تحلیل قرار می‌دهیم. در بخش ۶ نتیجه گیری ارائه می‌گردد.

۲- تحقیقات پیشین

بردارهای کلمه صریح بدست آمده با روش های مبتنی بر شمارش بدلیل تعداد زیاد کلمه‌های بافتار انتخاب شده، دارای ابعاد بالایی هستند. چون ابعاد بالای بردارهای کلمه صریح سبب ایجاد

محدودیت‌های محاسباتی می‌شود، معمولا بردارهای کلمه با ابعاد بالا به فضایی با ابعاد نهفته[۲۷] کمتر که از داده اصلی استخراج شده‌اند، نگاشت می‌شوند. بنابراین، بردارهای ضمنی با ابعاد کوچکتر ظاهر می‌شود. به این فرآیند استخراج ویژگی[۲۸] گفته می‌شود؛ زیرا ابعاد فضای کاهش‌یافته ویژگی‌های جدیدی هستند که از داده اصلی استخراج شده‌اند. برای تولید بردارهای کلمه ضمنی با ابعاد کم به جای بردارهای صریح با ابعاد زیاد معمولا از روش‌های کاهش ابعاد بهره گرفته می‌شود. رایج‌ترین راه برای تولید بازنمایی‌های ضمنی، نگاشت ماتریس هم‌رخدادی M به فضای معنایی نهفته کاهش‌یافته بوسیله الگوریتم‌های کاهش ماتریس نظیر SVD، PCA و NMF است [۴,۱].

اخیرا، مدل‌های پیش‌بینی یا تعبیه کلمه با استفاده از روش‌های شبکه عصبی بردارهای متراکم[۲۹] ضمنی را استنتاج می‌کنند. در حالی که مدل‌های مبتنی بر شمارش از هم‌رخدادی‌های کلمه هدف و کلمه بافتار استفاده می‌کنند و بردارهای کلمه صریح را تولید می‌کنند. دو موضوع مهم در مورد مدل‌های معنایی مطرح است: ۱-کیفیت اطلاعات بافتاری و ۲-کارایی. کیفیت کلمه‌های بافتاری و انتخاب کلمه‌های بافتار آگاهی‌بخش در مدل‌های فضای معنایی حائز اهمیت است. فاکتور کارایی نیز در مدل‌های فضای معنایی مهم است. واضح است که محاسبات شباهت کلمه برای بردارهای متراکم ساده‌تر است؛ زیرا عملیات ماتریسی با ابعاد کوچکتر مورد نیاز است. معمولا شباهت کسینوسی برای محاسبه شباهت بین بردارهای کلمه مورد استفاده قرار می‌گیرد.

مدل های مبتنی بر شمارش، بردارهای کلمه صریح با ابعاد بالا تولید می‌کنند؛ اما آن‌ها می‌توانند محاسبات را به طور موثری با استفاده از توابع درهم‌سازی[۳۰] انجام دهند. در توابع درهم‌سازی، کلیدها[۳۱] زوج‌های کلمه هدف-کلمه بافتار هستند و مقادیر آن‌ها امتیازهای غیر صفر است. در مدل‌های معنایی صریح، روابط ناموجود بسیاری وجود دارد که مقدار آن‌ها در بردار کلمه، عدد صفر است. بیمن و ریدل [۲۴] استدلال می‌کنند که نیازی به مدل‌سازی صریح روابط ناموجود که در بازنمایی برداری صفر هستند، نیست. و تنها زمانی ارزش دارد تا کلمه‌های بافتار را ذخیره کنیم که آن کلمه‌های بافتار به طور صریح دارای مقدار غیرصفر در بازنمایی پراکنده باشند.

معمولا روش‌های مبتنی بر شمارش، بردارهای کلمه صریح تنک را با استفاده از مدل‌های کاهش ابعاد که قبلا ذکر شد به بردارهای کلمه ضمنی تبدیل می‌کنند. در نتیجه، قابلیت تفسیرپذیری بردارها از دست می‌رود. در این حوزه موضوعی، توجه کمی به موضوع کاهش کلمه‌های بافتار و ایجاد بردارهای معنایی صریح با ابعاد کم شده است. تحقیقات گامالو [۲۴] و گامالو و بورداگ [۲۵] از استراتژی‌های فیلتر برای کاهش ابعاد بردارهای صریح استفاده می‌کنند تا مرتبط‌ترین کلمه‌های بافتار برای هر کلمه را بیابند. همچنین، بیمن و ریدل [۴] و پادرو و همکارانش [۲۶] مدل مبتنی بر شمارشی را شرح می‌دهند که از ایده فیلتر کردن بافتارها برای کاهش مقادیر غیر صفر استفاده می‌کند. در روش ایجاد فیلتر تنها R کلمه بافتار مرتبط براساس امتیاز درستنمایی[۳۲] بیشینه در جدول درهم‌سازی نگهداری می‌شود. هافمن و جیکوبز [۲۷] در علوم اعصاب بر این باورند که مدل‌های معنایی برای نشان دادن بردارهای کلمه به متغیرهای نهفته بی‌معنی نیاز ندارند. بنابراین، هم‌رخدادی‌های کلمه بایستی جایگزین متغیرهای نهفته کاهش یافته شود. همانطور که می‌دانیم بردارهای کلمه صریح، تفسیرپذیر هستند و هر بعد بردار کلمه معادل یک واژه است. در عوض، ابعاد بردارهای کلمه ضمنی معادل لغوی ندارند.

مدل‌های معنایی مبتنی بر پیش‌بینی، هر کلمه را به عنوان یک بردار d بعدی اعداد حقیقی نمایش می‌دهند. این مدل‌ها علی‌رغم دقت بالایی که دارند، متاسفانه تفسیرپذیر نیستند و ابعاد بردار کلمه معادل لغوی ندارد. نشان داده شده است که بردارهای کلمه‌ای که به هم نزدیک‌ترند، از نظر معنایی با یکدیگر ارتباط دارند. در بحث مدل‌های مبتنی بر پیش بینی مقالات [۲۹,۲۸] در skip-gram با روش آموزش نمونه‌گیری منفی (SGNS) به اوج خود رسیدند و یک الگوریتم تعبیه کارا ارائه کردند که نتایج بروزی بر روی وظایف زبانی مختلف فراهم می‌سازد. این الگوریتم با نام word2vec به شهرت رسیده است و برنامه‌ای برای ایجاد تعبیه-های کلمه است [۲۲].

تقریبا تمامی روش‌های عصبی، بازنمایی‌های متراکمی برای کلمه‌ها تولید می‌کنند که مختصات آن‌ها معنادار نیست. در نتیجه، مقادیر عددی تعبیه کلمه تنها در مقایسه با دیگر بازنمایی‌های کلمه تفسیر دارد. طراحی تعبیه‌های کلمه تفسیرپذیر به گونه‌ای که مختصات آن‌ها معنای متمایزی برای انسان داشته باشد، امری ضروری است. چندین پژوهش تلاش کرده‌اند تا تعبیه‌های کلمه موجود را به نوع تفسیرپذیر آن تبدیل کنند [۳۰]. تحقیقات مورفی و همکاران [۳۱] از فاکتورگیری ماتریس غیرمنفی (NMF) ماتریس هم‌رخدادی کلمه هدف-کلمه بافتار، برای استخراج تعبیه-های کلمه تفسیرپذیر استفاده می‌کند. سان و همکاران [۳۲] از مدل کیسه پیوسته کلمه‌ها[۳۳] (CBOW) برای مطالعه استفاده می‌کنند و تنظیم‌کننده[۳۴] l_1 را به هدف یادگیری آن می‌افزایند تا بردارهای پراکنده تفسیرپذیر تولید کنند. تحقیقات فاروکی و همکاران [۳۳] بردارهای کلمه بیش از حد کامل

پراکنده[۳۵] را برای حل مسئله بهینه‌سازی در تنظیمات یادگیری واژه‌نامه پیشنهاد می‌کند و یک نگاشت نامنفی پراکنده با ابعاد بالا برای تعبیه‌های کلمه تولید می‌کند. سوبرامانی و همکارانش [۳۴] از یک خودرمزگذار حذف نویز k-پراکنده[۳۶] برای ساخت یک نگاشت نامنفی پراکنده با ابعاد بالا برای تعبیه‌های کلمه استفاده می‌کنند که تعبیه‌های عصبی تفسیرپذیر پراکنده (SPINE) نامیده می‌شود.

ایده اصلی مطالعات فوق برای بدست آوردن نگاشت تفسیرپذیر تعبیه‌های کلمه، ایجاد پراکندگی در ابعاد بردار کلمه است. آن‌ها تعبیه‌های کلمه عمیق را به نگاشت نامنفی پراکنده با ابعاد بالا تعبیه‌های کلمه تبدیل می‌کنند. سپس یک معادل معنایی برای هر بعد کلمه پیدا می‌کنند. هیچ کدام از روش‌های ذکر شده به طور مستقیم قادر به ارائه بردارهای کلمه دارای مختصات معنادار نیستند. روش پیشنهادی ما در این مقاله برای اولین بار از بین کلمه‌های بافتار پرتکرار پیکره تعداد محدودی کلمه را به عنوان کلمه‌های بافتار آگاهی‌بخش انتخاب می‌کند و به عنوان بردارهای پایه برای ایجاد بازنمایی‌های کلمه با ابعاد پایین معرفی می‌کند. نتایج بدست آمده نشان می‌دهد با کاهش کلمه‌های بافتار از ۵۰۰۰ به حدود ۱۵۰۰، ضریب همبستگی اسپیرمن مجموعه های داده آزمون MEN [۳۵]، RG-65 [۳۶] و SimLex-999 [۳۷] در وظیفه شباهت کلمه افزایش می‌یابد.

۳- رویکرد پیشنهادی

ما در این پژوهش در نظر داریم تا بردارهای معنایی توزیعی صریح با ابعاد کم تولید کنیم. هدف اصلی ما این است که این کاهش ابعاد بردارها سبب کاهش کارایی بردارهای کلمه حاصل نشود. در ابتدا یک ماتریس هم‌رخدادی پایه[۳۷] می‌سازیم. تعداد H کلمه از پرتکرارترین کلمه‌های پیکره را به عنوان کلمه بافتار در نظر می‌گیریم. در مدل‌های مبتنی بر شمارش، برای شمارش هم‌رخدادی‌های ماتریس از یک پنجره با طول ثابت استفاده می‌شود. ما یک تابع نمایی برای شمارش هم‌رخدادی‌های کلمه‌ی بافتار و کلمه‌ی هدف به جای درنظر گرفتن پنجره با طول ثابت پیشنهاد می‌کنیم. این تابع برای شمارش هم‌رخدادی هر کلمه‌ی بافتار پیرامون کلمه هدف یک ضریب نمایی $e^{-0.1\alpha}$ نسبت می‌دهد. پارامتر α قدرمطلق فاصله کلمه هدف و کلمه بافتار در جمله‌هایی است که کلمه هدف و کلمه بافتار هم‌رخداد هستند. هر چه فاصله α کمتر شود (کلمه هدف و کلمه بافتار در جمله به هم نزدیک‌تر باشند)، ضریب بزرگتری برای شمارش هم‌رخدادی نسبت داده می‌شود. در ادامه، برای انتخاب N کلمه بافتار آگاهی‌بخش از H کلمه بافتار اولیه سه روش ۱-استفاده از معیار نسبت شباهت کلمه به تکرار کلمه، ۲-انتخاب ویژگی براساس مقایسه ماتریس‌های فاصله و ۳-وزن‌دهی دودویی مبتنی بر PSO را پیشنهاد می‌دهیم. این روش‌ها سعی می‌کنند به گونه‌ای کلمه-های بافتار آگاهی‌بخش را انتخاب کنند که شاهد کمترین میزان افت دقت و کارایی باشیم.

۳-۱- معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه

تعبیه‌های کلمه بدست آمده با مدل‌های مبتنی بر پیش‌بینی نظیر word2vec در وظایف پردازش زبان طبیعی کارایی خوبی دارند. برای کاهش ابعاد بردارهای کلمه صریح یک معیار جدید به نام نسبت شباهت کلمه به تکرار کلمه (معیار SF) تعریف می‌کنیم. این معیار با تقسیم فاکتور شباهت کلمه (WS) بر فاکتور تکرار کلمه (WF) محاسبه می‌شود. فاکتور WS شباهت کلی یک کلمه بافتار اولیه با دیگر کلمه‌های بافتار اولیه است. براساس مطالعات ما، کلمه‌های بافتار دارای فاکتور WS بزرگتر، معنادار و آگاهی‌بخش هستند. فاکتور شباهت را برای کلیه کلمه‌های بافتار اولیه، محاسبه می‌کنیم. نحوه محاسبه فاکتور شباهت کلمه در الگوریتم ۱ توضیح داده شده است.

الگوریتم ۱ محاسبه فاکتور شباهت کلمه

۱-10K از پرتکرارترین کلمه‌های پیکره (اسم‌ها، فعل‌ها، صفت‌ها و قیدها) را به عنوان کلمه‌های بافتار کاندید انتخاب می‌کنیم.

۲-یک ماتریس مربعی با نام S ایجاد می‌کنیم که دارای 10K سطر و ستون است. هر سطر و ستون به یکی از کلمه‌های بافتار کاندید اشاره می‌کند.

۳- کلمه‌های متناظر با سطر iام (r_i) و ستون jام (c_j) را به عنوان زوج i و j در نظر می‌گیریم و $p_{i,j}$ می‌نامیم.

۴- شباهت کسینوسی هر زوج i و j ($p_{i,j}$) را بدست می‌آوریم و در مولفه $s_{i,j}$ ذخیره می‌کنیم. ما از بردارهای word2vec کلمه‌ها برای محاسبه شباهت زوج‌های کلمه استفاده می‌کنیم.

۵- $WS_i = \sum_{j=1}^{j=10K} s_{i,j}$ را محاسبه می‌کنیم. WS_i فاکتور WS کلمه متناظر با سطر iام (r_i) ماتریس S است.

۶- فاکتور WS را برای همه کلمه‌های بافتار کاندید که متناظر با سطرهای ماتریس S هستند، محاسبه می‌کنیم.

۷- کلمه‌های بافتار کاندید را براساس فاکتور WS رتبه-بندی می‌کنیم. WS_i بزرگتر یعنی کلمه iام آگاهی‌بخش‌تر است.

کلمه‌های بافتار کاندیدی که دارای فاکتور WS بزرگتری هستند، معمولا فراوانی بالایی دارند. برای رفع بایاس ناشی از فراوانی کلمه، فاکتور WS کلمه بافتار کاندید را بر فراوانی آن کلمه تقسیم می‌کنیم. در ادامه، فاکتور WS هر کلمه بافتار کاندید را با نرم بینهایت، نرمال‌سازی می‌کنیم (NWS_i). فراوانی هر کلمه در پیکره را با فاکتور WF نشان می‌دهیم. فاکتور WF را با نرم بینهایت، نرمال‌سازی می‌کنیم (NWF_i). سپس معیار نسبت شباهت کلمه به تکرار کلمه (معیار FS) را به صورت زیر محاسبه می‌کنیم:

$$FS_i = \frac{NWS_i}{NWF_i} \qquad (۵)$$

برای کاهش تاثیر کاهنده فاکتور WF، می‌توانیم معیار FS را به صورت زیر بنویسیم:

$$FS_i = \frac{NWS_i}{\sqrt{NWF_i}} \qquad (۶)$$

سپس N کلمه بافتار کاندید با معیار FS بزرگتر را به عنوان کلمه بافتار آگاهی‌بخش انتخاب می‌کنیم. سپس بردارهای کلمه‌های واژه‌نامه را براساس N کلمه بافتار انتخاب شده می‌سازیم و مورد ارزیابی قرار می‌دهیم.

۳-۲- الگوریتم انتخاب ویژگی مبتنی بر مقایسه ماتریس‌های فاصله

در ابتدا، H کلمه از پرتکرارترین کلمه‌های پیکره را به عنوان کلمه‌های بافتار در نظر می‌گیریم. سپس، بردارهای کلمه‌های مجموعه آموزشی را بوسیله ماتریس هم‌رخدادی M بدست می‌آوریم. برای شمارش هم‌رخدادی از ایده تابع نمایی بهره می‌بریم. سپس با اعمال معیار PPMI به ماتریس هم‌رخدادی، ماتریس M^{PPMI} را بدست‌می‌آوریم. ماتریس M^{PPMI}، $\aleph \times H$ است. $\aleph$ تعداد کلمه‌های هدف مجموعه آموزشی است. ماتریس D(M) ماتریس فاصله کامل است. این ماتریس نامنفی و متقارن است. فاصله‌ی بین کلمه‌های هدف t_i و t_j در ماتریس M با مولفه $D(M)_{ij}$ مشخص می‌شود. $D(M)_{ij}$ فاصله اقلیدسی بین بردارهای کلمه t_i ($\vec{t}_i$) و t_j ($\vec{t}_j$) است. میزان آگاهی‌بخشی کلمه بافتاری lام را با حذف ستون lام ماتریس M بررسی می‌کنیم. ماتریس M_{-l} همان ماتریس M است که ستون lام آن حذف شده است. در گام بعدی، ماتریس فاصله برای ماتریس M_{-l} را بدست می‌آوریم و آن را ماتریس فاصله محدود $D_{-l}(M)$ می‌نامیم. ما برای هر یک از H کلمه بافتار اولیه، ماتریس فاصله محدود را بدست می‌آوریم. ماتریس فاصله کامل (D(M)) و ماتریس فاصله محدود ($D_{-l}(M)$) هر دو $\aleph \times \aleph$ هستند. در ادامه یک روش حذفی ساده رتبه‌بندی برای انتخاب تعدادی کلمه بافتار به عنوان کلمه‌های بافتار آگاهی‌بخش پیشنهاد می‌کنیم. مراحل این روش در الگوریتم ۲ شرح داده شده است.

الگوریتم ۲ روش انتخاب کلمات بافتار آگاهی‌بخش
۱- ماتریس فاصله کامل D(M) را محاسبه کنید. ۲-برای هر کلمه بافتار اولیه l=1,…,H: a) ماتریس فاصله محدود را با حذف ویژگی lام ($D_{-l}(M)$) بدست آورید. b) ماتریس فاصله A را که حاصل تفاضل ماتریس‌های D(M) و $D_{-l}(M)$ است، بدست آورید. c) نرم فروبنیوس ماتریس فاصله A را محاسبه کنید. نرم فروبنیوس معیاری برای اهمیت کلمه بافتار lام است. ۳-کلمه‌های بافتار اولیه را براساس نرم فروبنیوس رتبه‌بندی کنید. به عدد نرم فروبنیوس بزرگتر رتبه بالاتری تعلق می‌گیرد.

نرم فروبنیوس ماتریس A که $\aleph \times \aleph$ است، به صورت زیر محاسبه می‌شود [۳۸]:

$$\|A\|_F = \left(\sum_{i=1}^{\mathcal{T}}\sum_{j=1}^{\mathcal{T}}(a_{ij})^2\right)^{1/2} = (\mathrm{trace}(A^T A))^{1/2} \qquad (۷)$$

یعنی برای هر کلمه بافتار کاندید l ماتریس $A = D(M) - D_{-l}(M)$ را بدست می‌آوریم. سپس، نرم فروبنیوس ماتریس A ($\|A\|_F$) را محاسبه می‌کنیم. در ادامه، نرم‌های فروبنیوس حاصل را طبق الگوریتم فوق رتبه‌بندی کرده و N کلمه بافتار با رتبه بزرگتر را به عنوان کلمه‌های بافتار آگاهی‌بخش انتخاب می‌کنیم. هر چه حاصل تفاضل ماتریس فاصله کامل و ماتریس فاصله محدود بیشتر باشد، به این معنی است که با حذف کلمه بافتار l، تفاوت زیادی در آرایش فضایی داده‌ها بوجود آمده است، لذا آن کلمه بافتار حائز اهمیت است. ما در این پژوهش با H=5K کلمه بافتار اولیه و $\aleph = 18K$ کلمه هدف در واژه‌نامه، با انجام سه آزمایش ۵۰۰، ۱۰۰۰ و ۱۵۰۰ کلمه بافتار آگاهی‌بخش را انتخاب می‌کنیم.

۳-۳- روش وزن‌دهی دودویی مبتنی بر BPSO

ما در این پژوهش، یک روش وزن‌دهی دودویی مبتنی بر الگوریتم بهینه‌سازی ازدحام ذرات (PSO) پیشنهاد می‌کنیم که N کلمه بافتار اولیه را به عنوان کلمه‌های بافتار آگاهی‌بخش انتخاب می‌کند. در ادامه، در بخش ۳-۳-۱ ابتدا خلاصه‌ای در مورد

الگوریتم بهینه‌سازی ازدحام ذرات بیان می‌کنیم. سپس در بخش ۳-۳-۲ روش وزن‌دهی دودویی را شرح می‌دهیم.

۳-۳-۱- الگوریتم بهینه‌سازی ازدحام ذرات

در سال ۱۹۹۵ ابرهارت و کندی [۳۹] الگوریتم ابتکاری[۳۸] بهینه‌سازی ازدحام ذرات (PSO) را ارائه کردند. آن‌ها ایده الگوریتم PSO را از ازدحام‌های موجود در طبیعت نظیر دسته پرندگان، ازدحام زنبورهای عسل و دسته ماهی‌ها الهام گرفته‌اند. هر راه حل که به آن ذره گفته می‌شود، در الگوریتم PSO معادل یک پرنده در حرکت جمعی پرندگان است. الگوریتم PSO جستجوهایش را با جمعیتی از ذره‌ها انجام می‌دهد. جمعیت ذره‌هایی که بوسیله دسته‌های طبیعی هدایت می‌شوند، از ارتباطاتی مبتنی بر محاسبات تکاملی بهره می‌برند. الگوریتم PSO از مجموعه‌ای از ذره‌های در حال پرواز در فضای مسئله استفاده می‌کند. ذرات در حال پرواز ذرات بهینه را دنبال می‌کنند و به سوی یک منطقه امیدوارکننده حرکت می‌کنند تا به یک بهینه سراسری دست پیدا کنند. سرعت هر ذره به عنوان یک راه حل بالقوه شناخته می‌شود. در هر لحظه، سرعت هر ذره براساس تجربه‌های خود ذره‌ها و تجربه‌های اجتماعی آن‌ها تغییر می‌کنند. مولفه‌های الگوریتم PSO موارد ۱-متغیرها، ۲-ذره‌ها، ۳-ازدحام‌ها، و ۴-فرایندها هستند. ذره‌ها راه‌حل‌های کاندید هستند و پروازشان را از موقعیت‌های تصادفی در فضای جستجو آغاز می کنند [۴۰].

موقعیت و سرعت ذره iام در تکرار kام به ترتیب با x_k^i و v_k^i نشان داده می‌شود. سرعت و موقعیت ذره iام در تکرار (k+1)ام براساس رابطه‌های زیر محاسبه می‌شود:

$$x_{k+1}^i = x_k^i + v_{k+1}^i \qquad (۸)$$

$$v_{k+1}^i = \mathrm{w}.\, v_k^i + c_1.r_1.\left(p_k^i - x_k^i\right) + c_2.r_2.\left(p_k^g - x_k^i\right) \qquad (۹)$$

پارامتر w نشان‌دهنده وزن اینرسی است. وزن اینرسی برای کنترل سرعت و تعادل قابلیت‌های اکتشاف و بهره‌برداری الگوریتم است. w بزرگ سرعت ذرات را بالا نگه می‌دارد و از گیرافتادن ذرات در بهینه محلی جلوگیری می‌کند. w کوچک سبب می‌شود تا از مزایای موقعیت جستجوی فعلی ذره بهره گرفته شود. پارامترهای c_1 و c_2 ثابت‌هایی هستند که به ترتیب نشانگر مولفه‌های شناختی و اجتماعی هستند. این پارامترها به ترتیب برای تغییر وزن‌دهی بین تجربه شخصی و اجتماعی مورد استفاده قرار می‌گیرند. پارامترهای r_1 و r_2 اعداد تصادفی در بازه [0,1] هستند. بهترین موقعیت ذره iام تا تکرار kام با $\mathrm{p_k^i}$ مشخص می‌شود. بهترین موقعیت سراسری در ازدحام تا تکرار kام با p_k^g نشان داده می‌شود [۴۰]. به دلیل طبیعت گسسته بسیاری از مسائل، کندی و ابرهارت در سال ۱۹۹۷ نسخه گسسته PSO را پیشنهاد کردند. نسخه گسسته PSO از متغیرهای دودویی گسسته استفاده می‌کند که پیاده‌سازی عملیات تصمیم‌گیری برای هر ذره با استفاده از تصمیم گسسته "true" یا "false" است. در PSO دودویی، حالت هر ذره با استفاده از اعداد دودویی ۰ و ۱ مشخص می‌شود. سرعت در الگوریتم PSO دودویی براساس تغییرات احتمالاتی تعریف می‌شود. یک بیت خاص راه حل ذره را به صفر یا یک تغییر می‌دهد. تابع سیگموئید برای نگاشت سرعت رابطه (۱۰) به بازه [0,1] استفاده می‌شود. تابع سیگموئید در رابطه (۱۱) نشان داده شده است [۴۱].

$$v_{k+1}^{ij} = w.\, v_k^{ij} + c_1 r_1\left(p_k^{ij} - x_k^{ij}\right) + c_2 r_2\left(p_k^{gj} - x_k^{ij}\right) \qquad (۱۰)$$

$$v_{k+1}^{\prime ij} = sig\left(v_{k+1}^{ij}\right) = \frac{1}{1 + e^{-v_{k+1}^{ij}}} \qquad (۱۱)$$

بالانویس i به ذره iام اشاره می‌کند. بالانویس j به بیت j سرعت آن ذره اشاره می‌کند. موقعیت یک ذره براساس رابطه (۱۲) به‌روزرسانی می‌شود [۴۲]:

$$x_{k+1}^{ij} = \begin{cases} 1, & \text{اگر } r^{ij} < v_{k+1}^{\prime ij} \\ 0, & \text{وگرنه} \end{cases} \qquad (۱۲)$$

که r^{ij} یک عدد تصادفی با توزیع یکنواخت در بازه [0,1] است. در بخش ۳-۳-۲ روش وزن‌دهی دودویی پیشنهادی برای انتخاب N کلمه بافتار آگاهی‌بخش را توضیح می‌دهیم.

۳-۳-۲- روش وزن‌دهی دودویی

ما H کلمه پرتکرار پیکره را به عنوان کلمه‌های بافتار اولیه در نظر می‌گیریم. بردار کلمه هدف t به صورت $\vec{t} = [v_1, v_2, \dots, v_H]$ است. ما در روش وزن‌دهی دودویی به هر مولفه بردار کلمه هدف t (v_j) یک وزن دودویی (bw_j) نسبت می‌دهیم. در واقع به کلمه‌های هدف واژه‌نامه، یک بردار وزن دودویی BW=$[bw_1, bw_2, \dots, bw_H]$ نسبت می‌دهیم. هدف ما کاهش ابعاد بردارهای کلمه بوسیله انتخاب N بردار پایه آگاهی‌بخش است. به همین دلیل، مسئله بهینه‌سازی ازدحام ذرات را با این محدودیت حل می‌کنیم که مجموع مولفه‌های بردار BW برابر N باشد، یعنی:

$$\sum_{j=1}^{H} bw_j = N \qquad (۱۳)$$

ابتدا، برای حل مسئله وزن‌دهی دودویی یک مجموعه آموزشی با f عضو با استفاده ازکلمه‌های هدف واژه‌نامه می‌سازیم. بردارهای کلمه‌های عضو مجموعه آموزشی را از ماتریس هم‌رخدادی واژه‌نامه استخراج می‌کنیم. $v_{i,j}$ مولفه jام بردار کلمه iام است. سپس، ماتریس هم‌رخدادی وزن‌دار TS را برای کلمه‌های مجموعه آموزشی می‌سازیم. ماتریس TS وزن‌دار دارای f سطر و H ستون است. مولفه jام ماتریس TS وزن‌دار با $v_{i,j}.bw_j$ مشخص می‌شود که متناظر با سلول سطر iام و ستون jام ماتریس TS است. ما بدلیل دقت و کارایی بالا بردارهای ضمنی بدست آمده با نرم‌افزار word2vec، سعی می‌کنیم تا در تابع هدف تفاوت بردارهای کلمه صریح وزن‌دار را با بردارهای کلمه ضمنی کاهش دهیم. بنابراین، بردارهای کلمه‌های مجموعه آموزشی را با استفاده از نرم‌افزار word2vec با ۱۰۰۰ بعد بدست می‌آوریم و در ماتریس هم-رخدادی WV ذخیره می‌کنیم. ماتریس هم‌رخدادی WV دارای f سطر و ۱۰۰۰ ستون است. هر مولفه ماتریس WV با $wv_{i,j}$ مشخص می‌شود که متناظر با سلول سطر iام و ستون jام ماتریس WV است.

سپس، به دلیل ابعاد نامساوی ماتریس‌های TS و WV، ماتریس‌های $\mathcal{A}$ و $\mathcal{B}$ را به صورت زیر بدست می‌آوریم:

$$\mathcal{A} = TS \times TS' \qquad (۱۴)$$

$$\mathcal{B} = WV \times WV' \qquad (۱۵)$$

ماتریس‌های TS' و WV' به ترتیب ترانهاده ماتریس‌های TS و WV هستند. ماتریس‌های $\mathcal{A}$ و $\mathcal{B}$، $f \times f$ هستند. هر مولفه ماتریس $\mathcal{A}$ با $a_{i,j}$ مشخص می‌شود که متناظر با سلول سطر iام و ستون jام ماتریس $\mathcal{A}$ است. همچنین، هر مولفه ماتریس $\mathcal{B}$ با $b_{i,j}$ مشخص می‌شود که متناظر با سلول سطر iام و ستون jام ماتریس $\mathcal{B}$ است. در واقع، هر مولفه $a_{i,j}$ ماتریس $\mathcal{A}$، حاصلضرب داخلی بردار کلمه هدف متناظر با سطر iام ($\overrightarrow{tw_i}$) و بردار کلمه هدف متناظر با سطر jام ماتریس وزن‌دار TS ($\overrightarrow{tw_j}$) است. همچنین، هر مولفه $b_{i,j}$ ماتریس $\mathcal{B}$، حاصلضرب داخلی بردار کلمه هدف متناظر با سطر iام ($\overrightarrow{\mathrm{tw2v_i}}$) و بردار کلمه هدف متناظر با سطر jام ($\overrightarrow{\mathrm{tw2v_j}}$) ماتریس WV است.

$$a_{ij} = \overrightarrow{tw_i}.\overrightarrow{tw_j} \qquad (۱۶)$$

$$b_{\mathrm{ij}} = \overrightarrow{\mathrm{tw2v_i}}.\overrightarrow{\mathrm{tw2v_j}} \qquad (۱۷)$$

روش وزن‌دهی دودویی براساس الگوریتم PSO تلاش می‌کند تا وزن‌های کلمه‌های بافتار کاندید را به گونه‌ای بدست آورد که تابع هدف مسئله کمینه شود. تابع هدف مسئله بهینه‌سازی برای بدست آوردن وزن‌های دودویی به صورت زیر تعریف می‌شود:

$$OF = \sum_{i=1}^{f}\sum_{j=1}^{f}(b_{i,j} - a_{i,j})^2 \qquad (۱۸)$$

برای حل مسئله وزن‌دهی دودویی براساس الگوریتم PSO، جمعیتی با NP عضو در نظر می‌گیریم. هر عضو این جمعیت یک ذره است. هر موقعیت ذره معادل یک بردار وزن برای کلمه‌های بافتار اولیه است. در واقع هر موقعیت ذره یک بردار با H مولفه است. هر مولفه موقعیت ذره وزن دودویی کلمه بافتار معادل را مشخص می‌کند. بهترین ذره پس از اجرای الگوریتم بهینه‌سازی انتخاب می‌شود. بهترین ذره حاصل، تابع هدف مسئله را کمینه می‌کند. ما موقعیت بهترین ذره را به عنوان بردار وزن دودویی نهایی در نظر می‌گیریم. روش وزن‌دهی دودویی مبتنی بر الگوریتم BPSO در الگوریتم ۳ شرح داده شده است.

الگوریتم ۳ روش وزن‌دهی دودویی مبتنی بر BPSO

۱- NP ذره در نظر بگیرید.

۲-برای هر ذره موقعیت ذره (x_k^i) و سرعت ذره (v_k^i) را مقداردهی اولیه کنید. موقعیت ذره بایستی با وزن‌های دودویی "۰" و "۱" به صورت تصادفی مقداردهی اولیه شود.

۳- k=1 ، k تعداد تکرارها را مشخص می‌کند.

۴-تابع هدف رابطه (۱۸) را به هر ذره جمعیت اعمال کنید.

۵- بهترین موقعیت شناخته شده ذره (p_k^i) بایستی با مقداردهی اولیه شود، یعنی $p_k^i = x_k^i$.

۶-انجام بده:

a) تابع هدف را مجددا برای همه ذره‌ها محاسبه کنید.

b) موقعیت بهترین ذره شناخته شده (p_k^i) را با x_k^i به-روزرسانی کنید.

c) بهترین موقعیت (p_k^i) هر ذره و بهترین موقعیت ازدحام (p_k^g) را به‌روزرسانی کنید.

d) سرعت ذره را براساس رابطه سرعت (۱۰) محاسبه کنید.

e) موقعیت ذره را براساس رابطه موقعیت (۱۲) به‌روزرسانی کنید.

۷-تا زمانی که بیشینه تعداد تکرار برآورده نشده باشد.

۸-بهترین موقعیت شناخته شده ازدحام (p_k^g) بردار وزن دودویی نهایی است.

ما به ثابت‌های c_1 و c_2، که به ترتیب به تجربه شخصی و اجتماعی ذره اشاره می‌کنند؛ مقدار ۰.۱۵ نسبت می‌دهیم. وزن اینرسی w را نیز ۰.۷ در نظر می‌گیریم. در این پژوهش، تعداد جمعیت (NP) و بیشینه تکرار (k)، را به ترتیب برابر ۳۰ و ۲۰ در نظر می‌گیریم. در

بخش ۳-۴ کلمه‌های بافتار بدست آمده با روش‌های مطرح شده در بخش‌های ۳-۲ و ۳-۳ را ترکیب می‌کنیم تا بردارهای پایه تفسیرپذیر کارامدتری تولید کنیم.

۳-۴- ترکیب روش‌های معیار نسبت شباهت کلمه به تکرار کلمه و انتخاب ویژگی

در این بخش پیشنهاد می‌کنیم کلمه‌های بافتار بدست آمده با روش‌های ۱- معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه و ۲- انتخاب کلمه‌های بافتار براساس یک الگوریتم انتخاب ویژگی مبتنی بر مقایسه ماتریس‌های فاصله را ترکیب کنیم تا بتوانیم افت حداقلی ایجاد شده در هر دو روش ذکر شده را که ناشی از حذف تعداد زیادی کلمه بافتار است، پوشش دهیم. هدف ما انتخاب بردارهای پایه معنادار به‌گونه‌ای است که بتوان بردارهای کلمه تفسیرپذیر کارا و موثر تولید کرد.

برای تحقق این هدف، ما سعی می‌کنیم تا از خبرگی هر دو روش به میزان یکسان استفاده کنیم؛ بنابراین پیشنهاد می‌کنیم که $\frac{N}{2}$ بردارهای پایه (کلمه‌های بافتار آگاهی‌بخش) را با استفاده از روش معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه انتخاب کنیم و $\frac{N}{2}$ باقی‌مانده بردارهای پایه (کلمه‌های بافتار آگاهی‌بخش) را با استفاده از روش انتخاب کلمه‌های بافتار براساس یک الگوریتم انتخاب ویژگی مبتنی بر مقایسه ماتریس‌های فاصله بدست آوریم. بدیهی است که تعدادی کلمه بافتار بین این دو گروه مشترک هستند. نتایج ارزیابی‌ها، رشد قابل ملاحظه‌ای در ضرایب همبستگی اسپیرمن مجموعه‌های داده آزمون نشان می‌دهد.

۳-۵- ترکیب روش‌های معیار نسبت شباهت کلمه به تکرار کلمه، انتخاب ویژگی و وزن‌دهی دودویی

در این بخش در نظر داریم تا به کلمه‌های بافتار ترکیبی بدست آمده در بخش ۳-۴، کلمه‌های بافتار آگاهی‌بخشی را که با استفاده از روش وزن‌دهی دودویی مبتنی بر BPSO بدست آورده‌ایم، بیفزاییم. ما در این مقاله، دو آزمایش برای یافتن N_1=1000 و N_2=500 کلمه بافتار آگاهی‌بخش انجام می‌دهیم و با استفاده از روش وزن‌دهی دودویی مبتنی بر BPSO کلمات بافتار آگاهی‌بخش را بدست می‌آوریم. سپس کلمه‌های مشترک بین کلمه‌های بافتار دارای وزن دودویی "۱" بدست آمده با پارامترهای N_1=1000 و N_2=500 را به عنوان بردار پایه انتخاب می‌کنیم و به بردارهای پایه ترکیبی بدست آمده در بخش ۳-۴ می‌افزاییم. نتایج ارزیابی، افزایش ضریب همبستگی اسپیرمن مجموعه‌های داده آزمون را نسبت به روش‌های ذکر شده در بخش‌های ۳-۱، ۳-۲ و ۳-۳ گزارش می‌کند.

۴- تنظیمات مربوط به آزمایش‌ها

ما در این بخش جزئیاتی در مورد پیکره مورد استفاده برای انجام آزمایش‌ها، جزئیات و پارامترهای به کار گرفته شده برای انجام آزمایش‌ها و نحوه ارزیابی آزمایش‌ها ارائه می‌کنیم.

۴-۱- پیکره

پیکره ukWaC[39] [۴۳] یک پیکره بسیار بزرگ برای زبان انگلیسی است که بیش از یک میلیارد کلمه دارد. این پیکره با خزیدن وب[40] ایجاد شده است. این پیکره به عنوان یک منبع عمومی برای زبان انگلیسی مورد استفاده قرار گرفته است. این پیکره شامل برچسب اجزای سخن (POS) و اندیس تجزیه وابستگی است. ما در این پژوهش به دلیل محدودیت‌های محاسباتی تنها از بخش ۱ پیکره ukWaC (ukwac_dep_parsed_01) استفاده می‌کنیم و سپس بردارهای کلمه حاصل را بر روی وظیفه شباهت کلمه با استفاده از مجموعه‌های آزمون MEN، RG-65 و SimLex-999 مورد ارزیابی قرار می‌دهیم.

۴-۲- نحوه استخراج بردارهای پایه

در ابتدا، واژه‌نامه را با استفاده از 20K از پرتکرارترین اسم‌ها، 10K از پرتکرارترین فعل‌ها، 10K از پرتکرارترین صفت‌ها و 5K از پرتکرارترین قیدها می‌سازیم. سپس، 5K از پرتکرارترین کلمه‌ها (اسم، فعل، صفت و قید) را به عنوان کلمه‌های بافتار اولیه در نظر می‌گیریم. در گام اول، معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه را برای 5K کلمه‌ی بافتار اولیه براساس توضیحات بیان شده در بخش ۳-۱ بدست می‌آوریم. سپس N_1=1000 و N_2=500 کلمه بافتار آگاهی‌بخش را که دارای بیشترین معیار $\frac{WS}{WF}$ و $\frac{WS}{\sqrt{WF}}$ هستند انتخاب می‌کنیم. سپس براساس کلمه‌های بافتار آگاهی‌بخش انتخاب شده بردارهای کلمه‌های هدف واژه‌نامه را می‌سازیم و ارزیابی می‌کنیم. در گام دوم، 18K از پرتکرارترین کلمه‌های هدف واژه‌نامه (8K اسم، 4k فعل، 4k صفت و 2k قید) را به عنوان مجموعه آموزشی برای یافتن کلمه‌های بافتار آگاهی‌بخش برای استفاده از روش انتخاب ویژگی با مقایسه ماتریس‌های فاصله در نظر می‌گیریم. سپس براساس روش انتخاب ویژگی که در بخش ۳-۲ شرح دادیم، کلمه‌های بافتار اولیه را رتبه‌بندی می‌کنیم و N_1=1000 و N_2=500 کلمه بافتار آگاهی-بخش را انتخاب می‌کنیم. سپس، براساس کلمه‌های بافتار آگاهی-بخش انتخاب شده، بردارهای کلمه‌های هدف واژه‌نامه را می‌سازیم و ارزیابی می‌کنیم.

در گام سوم، 2K کلمه بافتار را به عنوان مجموعه آموزشی برای یافتن کلمه‌های بافتار آگاهی‌بخش با استفاده از روش وزن‌دهی دودویی مبتنی بر PSO در نظر می‌گیریم. سپس، براساس الگوریتم ارائه شده در بخش ۳-۳ به N_1=1000 و N_2=500 کلمه بافتار آگاهی‌بخش وزن ۱ تعلق می‌گیرد. ما براساس کلمه‌های بافتار آگاهی‌بخش انتخاب‌شده، بردارهای کلمه‌های هدف واژه‌نامه را می‌سازیم و ارزیابی می‌کنیم. در گام چهارم، اجتماع ۵۰۰ کلمه بافتار بدست آمده با روش معیار نسبت شباهت کلمه به تکرار کلمه و ۵۰۰ کلمه بافتار بدست آمده بوسیله روش انتخاب ویژگی براساس مقایسه ماتریس‌های فاصله (مجموعه P) را بدست می‌آوریم. همچنین، اجتماع ۱۰۰۰ کلمه بافتار بدست آمده با روش معیار نسبت شباهت کلمه به تکرار کلمه و ۱۰۰۰ کلمه بافتار بدست آمده بوسیله روش انتخاب ویژگی براساس مقایسه ماتریس‌های فاصله (مجموعه O) را بدست می‌آوریم. سپس با استفاده از کلمه‌های بافتار آگاهی‌بخش در مجموعه‌های P و O، بردارهای کلمه‌های هدف واژه‌نامه را می‌سازیم و مورد ارزیابی قرار می‌دهیم. در گام پنجم، مسئله وزن‌دهی دودویی مبتنی بر BPSO را برای انتخاب N_1=1000 و N_2=500 کلمه بافتار آگاهی‌بخش حل می‌کنیم. سپس اشتراک مجموعه‌های کلمه‌های بافتار ۵۰۰ و ۱۰۰۰ تایی انتخاب شده را بدست می‌آوریم و به مجموعه‌های P و O می‌افزاییم. سپس بردارهای کلمه‌های واژه‌نامه را مبتنی بر کلمه‌های بافتار آگاهی‌بخش انتخاب شده، تولید کرده و ارزیابی می‌کنیم.

۴-۳- ارزیابی بردارهای پایه استخراج شده

ما در این پژوهش، از مجموعه‌های داده شباهت کلمه استاندارد که شامل زوج‌های کلمه و امتیاز استاندارد طلایی است، استفاده می‌کنیم. امتیاز استاندارد طلایی، قضاوت انسانی در مورد شباهت بین کلمه‌های هر زوج را مشخص می‌کند. ما شباهت کسینوسی را بر روی مجموعه‌های داده MEN، RG-65 و SimLex-999 اندازه‌گیری می‌کنیم. مجموعه داده MEN شامل نرخ‌های شباهت برای ۳۰۰۰ زوج کلمه است. مجموعه داده RG-65 دارای ۶۵ زوج کلمه و شباهت‌هایشان است. مجموعه داده SimLex-999 شامل نرخ‌های شباهت ۹۹۹ زوج کلمه است. ما شباهت زوج‌های کلمه و امتیاز استاندارد طلایی را از طریق ضریب همبستگی اسپیرمن ρ بدست می‌آوریم.

۵- نتایج و بررسی

در این بخش، نتایج ارزیابی روش‌های انتخاب بردارهای پایه آگاهی‌بخش ذکر شده در بخش ۳ را با جزئیات بیان می‌کنیم. در ابتدا، برای کلمه‌های هدف واژه‌نامه یک ماتریس هم‌رخدادی پایه می‌سازیم. برای ساخت ماتریس هم‌رخدادی از 5K کلمه پرتکرار پیکره به عنوان کلمه‌های بافتار اولیه استفاده می‌کنیم و برای شمارش هم‌رخدادی‌ها از پنجره با طول ثابت ۱۰ استفاده می‌کنیم. سپس از روش پیشنهادی تابع نمایی برای شمارش هم‌رخدادی‌ها استفاده می‌کنیم و ماتریس هم‌رخدادی را مجددا می‌سازیم. نتایج ارزیابی بردارهای کلمه بدست آمده با استفاده از پنجره ثابت و تابع نمایی بر روی مجموعه‌های آزمون در جدول ۱ آمده است. با شمارش هم‌رخدادی‌ها با تابع نمایی در مقایسه با پنجره با اندازه ثابت ۱۰، ضریب همبستگی اسپیرمن مجموعه‌های داده آزمون MEN، RG-65 و SimLex-999 به ترتیب به میزان ۱.۷۳٪، ۵.۷۴٪ و ۱.۴۴٪ افزایش یافته است. نتیجه حاصل نشان‌دهنده این نکته است که تاثیر معنایی کلمه‌های نزدیک‌تر به کلمه هدف بیشتر است. در این پژوهش، شمارش هم‌رخدادی‌ها برای ایجاد تمامی ماتریس‌های هم‌رخدادی با استفاده از تابع نمایی پیشنهادی انجام شده است. سپس، با استفاده از روش معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه با استفاده از دو معیار $\frac{WS}{WF}$ و $\frac{WS}{\sqrt{WF}}$، N=1000 کلمه بافتار را به عنوان بردار پایه انتخاب کرده و ماتریس‌های هم‌رخدادی کلمه‌های واژه‌نامه را با استفاده از تابع نمایی برای شمارش هم‌رخدادی‌ها می‌سازیم. نتایج ارزیابی ماتریس‌های هم‌رخدادی حاصل در جدول ۲ گزارش شده است. در بردارهای کلمه بدست آمده با استفاده از ۱۰۰۰ کلمه بافتار آگاهی‌بخش استخراج شده بوسیله معیار $\frac{WS}{WF}$ در مقایسه با ۵۰۰۰ کلمه بافتار پرتکرار، ضریب همبستگی اسپیرمن در مجموعه‌های آزمون MEN، RG-65 و SimLex-999 به میزان ۰.۹۷٪، ۰.۷۵٪ و ۲.۲۲٪ کاهش می‌یابد. در بردارهای کلمه بدست آمده با ۱۰۰۰ کلمه‌ی بافتار استخراج شده بوسیله معیار $\frac{WS}{\sqrt{WF}}$ در مقایسه با ۵۰۰۰ کلمه بافتار پرتکرار، ضریب همبستگی اسپیرمن به ترتیب به میزان ۰.۷۳٪ و ۲.۲۲٪ در مجموعه‌های داده آزمون MEN و SimLex-999 کاهش یافته است. همچنین ضریب همبستگی اسپیرمن در مجموعه RG-65 به میزان ۳.۵۹٪ افزایش یافته است. این کاهش دقت توجیه‌پذیر است، زیرا ابعاد بردارهای کلمه از ۵۰۰۰ به ۱۰۰۰ کاهش یافته است و حجم زیادی از اطلاعات مورد استفاده قرار نگرفته است. با بررسی جدول ۲ درمی‌یابیم که معیار $\frac{WS}{\sqrt{WF}}$ در مقایسه با معیار $\frac{WS}{WF}$ به‌ویژه در مجموعه RG-65 عملکرد بهتری دارد.

جدول ۱: ارزیابی ماتریس‌های هم‌رخدادی حاصل ازاندازه پنجره ثابت ۱۰ و پنجره پیشنهادی با ضریب نمایی $e^{-0.1\alpha}$.

مجموعه داده آزمون	ماتریس هم رخدادی با 5K کلمه پرتکرار پیکره به عنوان بردارهای پایه	
	پنجره با ضریب نمایی $e^{-0.1\alpha}$	اندازه پنجره=۱۰
MEN	۶۸.۶۲	۶۶.۸۹
RG-65	۶۲.۷۰	۵۶.۹۶
SimLex-999	۲۷.۶۶	۲۶.۲۲

جدول ۳ نتایج ارزیابی بردارهای کلمه بدست آمده با N_2=500 کلمه بافتار استخراج شده با روش معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه با استفاده از دو معیار $\frac{WS}{WF}$ و $\frac{WS}{\sqrt{WF}}$، را نشان می‌دهد. با کاهش ابعاد بردارهای کلمه از ۵۰۰۰ بعد به ۵۰۰ بعد با استفاده از معیار $\frac{WS}{WF}$، ضریب همبستگی اسپیرمن برای مجموعه‌های داده MEN، RG-65، SimLex-999 به ترتیب به میزان ۲.۶۱٪، ۲.۷۴٪ و ۴.۵۱٪ کاهش می‌یابد. همچنین با کاهش ابعاد بردارهای پایه از ۵۰۰۰ بعد به ۵۰۰ بعد با استفاده از معیار $\frac{WS}{\sqrt{WF}}$، ضریب همبستگی اسپیرمن به ترتیب به میزان ۲.۹۲٪، ۴.۲۸٪ و ۴.۶۲٪ کاهش می‌یابد. در بردارهای کلمه با ۵۰۰ بعد، معیار $\frac{WS}{\sqrt{WF}}$ اندکی بهتر عمل کرده است. کاهش شدید دقت ایجاد شده، ناشی از حذف زیادی از اطلاعات در عملیات کاهش بعد از ۵۰۰۰ به ۵۰۰ است.

جدول۲: انتخاب 1K بردار پایه از 5K بردار پایه پرتکرار براساس معیارهای $\frac{WS}{WF}$ و $\frac{WS}{\sqrt{WF}}$.

مجموعه داده آزمون	بردارهای پایه		
	1K کلمه بدست آمده با معیار $\frac{WS}{WF}$	1K کلمه بدست آمده با معیار $\frac{WS}{\sqrt{WF}}$	5K کلمه پرتکرار پیکره
MEN	۶۷.۶۵	۶۷.۸۹	۶۸.۶۲
RG-65	۶۱.۹۵	۶۶.۲۹	۶۲.۷۰
SimLex-999	۲۵.۴۴	۲۵.۳۹	۲۷.۶۶

جدول۳: انتخاب ۵۰۰ بردار پایه از 5K بردار پایه پرتکرار براساس معیارهای $\frac{WS}{WF}$ و $\frac{WS}{\sqrt{WF}}$.

مجموعه داده آزمون	بردارهای پایه		
	۵۰۰ کلمه بدست آمده با معیار $\frac{WS}{WF}$	۵۰۰ کلمه بدست آمده با معیار $\frac{WS}{\sqrt{WF}}$	5K کلمه پرتکرار پیکره
MEN	۶۶.۰۱	۶۵.۷۰	۶۸.۶۲
RG-65	۵۹.۹۶	۵۸.۴۲	۶۲.۷۰
SimLex-999	۲۳.۱۵	۲۳.۰۴	۲۷.۶۶

در گام بعدی، با استفاده از روش انتخاب ویژگی با مقایسه ماتریس‌های فاصله، N_1=1000 و N_2=500 کلمه بافتار آگاهی‌بخش را به عنوان بردار پایه انتخاب می‌کنیم. سپس با استفاده از بردارهای پایه بدست آمده با روش انتخاب ویژگی، ماتریس هم رخدادی با ابعاد ۵۰۰ و ۱۰۰۰ را تولید می‌کنیم. سپس بردارهای کلمه واژه‌نامه را با استفاده از مجموعه‌های آزمون ارزیابی می‌کنیم. نتایج ارزیابی‌ها در جدول ۴ نشان می‌دهد که با کاهش ابعاد از ۵۰۰۰ به ۱۰۰۰، ضریب همبستگی اسپیرمن برای مجموعه‌های MEN، RG-65 و SimLex-999 به ترتیب به میزان ۲.۱۹٪، ۳.۳٪ و ۰.۹۳٪ کاهش می‌یابد. با کاهش ابعاد بردارهای کلمه از ۵۰۰۰ به ۵۰۰ بوسیله روش انتخاب ویژگی مبتنی بر مقایسه ماتریس‌های فاصله، ضریب همبستگی اسپیرمن به ترتیب به میزان ۳.۳۷٪، ۴.۱٪ و ۲.۰۷٪ کاهش می‌یابد. با بررسی جدول ۲ و جدول ۴ متوجه می‌شویم که کاهش ابعاد با روش معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه در مجموعه‌های آزمون MEN و RG-65 که مبتنی بر ارتباط کلمه‌ها هستند، نسبت به روش انتخاب ویژگی با مقایسه ماتریس‌های فاصله عملکرد بهتری داشته است. در مورد مجموعه آزمون SimLex-999 که یک مجموعه آزمون مبتنی بر شباهت است، روش انتخاب ویژگی با مقایسه ماتریس-های فاصله عملکرد بهتری در مقایسه با روش معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه دارد.

جدول۴: انتخاب ۵۰۰ و ۱۰۰۰ بردار پایه از 5K بردار پایه پرتکرار براساس روش انتخاب ویژگی مبتنی بر ماتریس‌های فاصله.

مجموعه داده آزمون	بردارهای پایه		
	۵۰۰ کلمه	۱۰۰۰کلمه	5K کلمه پرتکرار پیکره
MEN	۶۵.۲۵	۶۶.۴۳	۶۸.۶۲
RG-65	۵۸.۶۰	۵۹.۴۰	۶۲.۷۰
SimLex-999	۲۵.۵۹	۲۶.۷۳	۲۷.۶۶

در ادامه، از روش وزن‌دهی دودویی مبتنی بر BPSO برای انتخاب N_1=1000 و N_2=500 کلمه بافتار آگاهی‌بخش استفاده می‌کنیم. سپس بردارهای کلمه واژه‌نامه را با استفاده از کلمه‌های بافتار آگاهی‌بخش انتخاب شده تولید می‌کنیم. نتایج ارزیابی‌ها در جدول ۵ گزارش شده است. با کاهش ابعاد بردارهای کلمه از ۵۰۰۰ بعد به ۵۰۰ بعد، ضریب همبستگی اسپیرمن مجموعه‌های MEN، RG-65 و SimLex-999 به ترتیب به میزان ۶.۰۷٪، ۷.۳۲٪ و ۲.۹۸٪ کاهش می‌یابد. همچنین با کاهش ابعاد بردارهای کلمه از ۵۰۰۰ بعد به ۱۰۰۰ بعد با استفاده از روش وزن‌دهی دودویی،

شاهد افت ضریب همبستگی اسپیرمن برای مجموعه‌های MEN، RG-65 و SimLex-999 به ترتیب به میزان ۳.۹۱٪، ۴.۶۸٪ و ۱.۰۵٪ هستیم.

جدول۵: انتخاب ۵۰۰ و ۱۰۰۰ بردار پایه از 5K بردار پایه پرتکرار براساس روش وزن‌دهی دودویی مبتنی بر PSO.

مجموعه داده آزمون	بردارهای پایه		
	۵۰۰ کلمه	۱۰۰۰کلمه	5K کلمه پرتکرار پیکره
MEN	۶۲.۵۵	۶۴.۷۱	۶۸.۶۲
RG-65	۵۵.۳۸	۵۸.۰۲	۶۲.۷۰
SimLex-999	۲۴.۶۸	۲۶.۶۱	۲۷.۶۶

با بررسی جدول‌های ۲-۵، متوجه می‌شویم که کاهش ابعاد با استفاده از هر سه روش، سبب افت ضریب همبستگی اسپیرمن مجموعه‌های آزمون می‌شود. این افت دقت در روش معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه در مجموعه‌های MEN و RG-65 که مبتنی بر ارتباط هستند، نسبت به دو روش دیگر کمتر است. اما کاهش دقت در مورد مجموعه آزمون SimLex-999 که مبتنی بر شباهت است، در روش انتخاب ویژگی با مقایسه ماتریس‌های فاصله نسبت به دو روش دیگر کمتر است. افت دقت در مورد مجموعه SimLex-999 در روش معیار نسبت شباهت کلمه به تکرار کلمه از همه بیشتر است. هدف ما در این پژوهش بدست آوردن بردارهای پایه با تعداد محدود به‌گونه‌ای است که ضرایب همبستگی اسپیرمن بردارهای کلمه با ابعاد کم حاصل بر روی مجموعه‌های آزمون دچار افت نشود. برای تحقق این هدف، با بهره‌گیری از خبرگی هر دو روش معیار نسبت شباهت کلمه به تکرار کلمه و روش انتخاب ویژگی به میزان یکسان و بدست آوردن دقت های بهتر در همه مجموعه‌های آزمون، نیمی از کلمه‌های بافتار آگاهی‌بخش را با روش معیار پیشنهادی نسبت شباهت کلمه به تکرار کلمه و نیم دیگر را با روش انتخاب ویژگی مبتنی بر مقایسه ماتریس‌های فاصله استخراج می‌کنیم. با ترکیب ۵۰۰ کلمه بافتار آگاهی‌بخش با استفاده از معیار $\frac{WS}{WF}$ و ۵۰۰ کلمه بافتار آگاهی‌بخش با استفاده از روش انتخاب ویژگی، ۸۳۳ کلمه بافتار به عنوان بردار پایه آگاهی‌بخش انتخاب می‌شود. سپس بردارهای کلمه‌های هدف را با کلمات بافتار ترکیبی بدست آمده تولید کرده و ارزیابی می‌کنیم. همانطور که در جدول ۶ مشاهده می‌کنید، ضریب همبستگی اسپیرمن مجموعه‌های آزمون MEN، RG-65 پس از کاهش بعد از ۵۰۰۰ به ۸۳۳ به ترتیب به میزان ۱.۳۹٪ و ۳.۶۶٪ افزایش داشته است. ضریب همبستگی اسپیرمن برای مجموعه آزمون SimLex-999 به میزان ۱.۹۵٪ کاهش یافته است. بردارهای کلمه‌ای که با کلمه‌های بافتار ترکیبی شامل ۵۰۰ کلمه بافتار انتخاب شده با معیار $\frac{WS}{\sqrt{WF}}$ و ۵۰۰ کلمه بافتار انتخاب شده با روش انتخاب ویژگی، تولید شده‌اند را ارزیابی می‌کنیم. نتایج ارزیابی‌ها در جدول ۶، نشان‌دهنده افزایش دقت قابل ملاحظه‌ای در مجموعه‌های داده آزمون MEN و RG-65 به ترتیب به میزان ۱.۲۸٪ و ۴.۳۸٪ است. در مجموعه آزمون مبتنی بر شباهت SimLex-999 افت دقت به میزان ۱.۸۶٪ رخ داده است.

ما ۱۴۹۷ بردار پایه ترکیبی را با ترکیب ۱۰۰۰ کلمه بافتار آگاهی‌بخش بدست آمده با معیار $\frac{WS}{WF}$ و ۱۰۰۰ کلمه بافتار آگاهی‌بخش بدست آمده با روش انتخاب ویژگی بدست می‌آوریم. بدیهی است که بین دو مجموعه کلمه بازاشتراکی وجود دارد. سپس بردارهای کلمه‌های هدف واژه‌نامه را مجددا تولید کرده و ارزیابی می‌کنیم. نتایج ارزیابی در جدول ۷ نشان داده شده است. همانطور که مشاهده می‌کنید، در بردارهای کلمه هدف پس از کاهش بعد از ۵۰۰۰ به ۱۴۹۷، ضریب همبستگی اسپیرمن در مجموعه‌های داده آزمون MEN، R-65 و SimLex-999 به ترتیب به میزان قابل ملاحظه ۲.۳۲٪، ۷.۰۶٪ و ۰.۵۵٪ افزایش یافته است. در ادامه با ترکیب ۱۰۰۰ کلمه بافتار بدست آمده با معیار $\frac{WS}{\sqrt{WF}}$ و ۱۰۰۰ کلمه بافتار بدست آمده با روش انتخاب ویژگی، ۱۵۱۱ بردار پایه ترکیبی بدست می‌آوریم (بین دو مجموعه کلمه‌های بافتار بازاشتراکی وجود دارد.). سپس بردارهای کلمه‌های هدف واژه‌نامه را با استفاده از بردارهای پایه ترکیبی تولید کرده و ارزیابی می‌کنیم. در جدول ۷، گزارش شده است که ضریب همبستگی اسپیرمن بردارهای کلمه پس از کاهش بعد از ۵۰۰۰ به ۱۵۱۱، به ترتیب به میزان ۲.۴۷٪، ۷.۳۹٪ و ۰.۵۲٪ افزایش یافته است. ما با استفاده از ترکیب بردارهای پایه انتخاب شده توسط دو روش ذکر شده، از خبرگی این دو روش کارامد بهره برده‌ایم. درنتیجه، با حذف حدود ۳۵۰۰ بردار پایه، نه تنها کاهش دقت نداشته‌ایم بلکه افزایش دقت هم رخ داده است.

در جدول ۵ مشاهده شد که کاهش بعد به ۵۰۰ و ۱۰۰۰ با استفاده از روش وزن‌دهی دودویی مبتنی بر BPSO در مجموعه آزمون مبتنی بر شباهت SimLex-999 سبب کاهش دقت کمتری نسبت به روش معیارهای $\frac{WS}{WF}$ و $\frac{WS}{\sqrt{WF}}$ می‌شود. لذا برای بهره‌گیری از خبرگی این روش در مورد مجموعه‌های داده مبتنی بر شباهت، کلمات مشترک بین N_1=1000 و N_2=500 کلمه بافتار آگاهی‌بخش انتخاب شده با این روش (۱۱۵ کلمه مشترک، این دو مجموعه کلمه بازاشتراکی دارند.) را به بردارهای پایه ترکیبی می‌افزاییم. سپس بردارهای کلمه‌ها را تولید کرده و مورد ارزیابی قرار می-

دهیم. نتایج ارزیابی در جدول ۸ گزارش شده است. کاهش بعد با ترکیب ۵۰۰ کلمه بافتار بدست آمده با معیار $\frac{WS}{\sqrt{WF}}$ و ۵۰۰ کلمه بافتار انتخاب شده با روش انتخاب ویژگی و ۱۱۵ واژه مشترک بدست آمده با روش وزن‌دهی دودویی (۹۳۸ کلمه)، سبب افزایش ضریب همبستگی اسپیرمن به میزان ۰.۷۷٪ و ۱.۲۱٪ در مجموعه‌های داده MEN و RG-65 می‌شود. بردارهای پایه ترکیبی حاصل سبب کاهش ضریب همبستگی اسپیرمن در مجموعه داده SimLex-999 به میزان ۱.۴۶٪ می‌شود. با افزودن ۱۱۵ کلمه مشترک بدست آمده با روش وزن‌دهی دودویی به ترکیب ۱۰۰۰ کلمه بافتار انتخاب شده با معیار $\frac{WS}{\sqrt{WF}}$ و ۱۰۰۰ کلمه بافتار انتخاب شده با روش انتخاب ویژگی (۱۶۰۴ کلمه)، سبب افزایش ضریب همبستگی اسپیرمن در مجموعه‌های داده MEN، RG-65 و SimLex-999 به میزان ۲.۰۴٪، ۵.۰۷٪ و ۰.۶۹٪ می‌شود.

جدول۶: ترکیب ۵۰۰ بردار پایه بدست آمده بوسیله معیارهای $\frac{WS}{WF}$ و $\frac{WS}{\sqrt{WF}}$ با ۵۰۰ بردار پایه بدست آمده با روش انتخاب ویژگی مبتنی بر ماتریس‌های فاصله.

مجموعه داده آزمون	بردارهای پایه		
	۸۳۳ کلمه	۸۳۴ کلمه	5K کلمه
	۵۰۰ کلمه بدست آمده با معیار $\frac{WS}{WF}$+۵۰۰ کلمه بدست آمده با روش انتخاب ویژگی	۵۰۰ کلمه بدست آمده با معیار $\frac{WS}{\sqrt{WF}}$+۵۰۰ کلمه بدست آمده با روش انتخاب ویژگی	کلمه-های پرتکرار
MEN	۷۰.۰۱	۶۹.۹۰	۶۸.۶۲
RG-65	۶۶.۳۶	۶۷.۰۸	۶۲.۷۰
SimLex-999	۲۵.۷۱	۲۵.۸۰	۲۷.۶۶

جدول۷: ترکیب ۱۰۰۰ بردار پایه بدست آمده بوسیله معیارهای $\frac{WS}{WF}$ و $\frac{WS}{\sqrt{WF}}$ با ۱۰۰۰ بردار پایه بدست آمده با روش انتخاب ویژگی مبتنی بر ماتریس‌های فاصله.

مجموعه داده آزمون	بردارهای پایه		
	۱۴۹۷ کلمه	۱۵۱۱ کلمه	5K کلمه
	1K کلمه بدست آمده با معیار $\frac{WS}{WF}$+ 1K کلمه بدست آمده با روش انتخاب ویژگی	1K کلمه بدست آمده با معیار $\frac{WS}{\sqrt{WF}}$+ 1K کلمه بدست آمده با روش انتخاب ویژگی	کلمه-های پرتکرار
MEN	۷۰.۹۴	۷۱.۰۹	۶۸.۶۲
RG-65	۶۹.۷۶	۷۰.۰۹	۶۲.۷۰
SimLex-999	۲۸.۲۱	۲۸.۱۸	۲۷.۶۶

جدول۸: ترکیب بردارهای پایه بدست آمده بوسیله معیار $\frac{WS}{\sqrt{WF}}$ با بردارهای پایه بدست آمده با روش انتخاب ویژگی و ۱۱۵ بردار پایه مشترک بدست آمده با روش وزن‌دهی دودویی.

مجموعه داده آزمون	بردارهای پایه		
	۹۳۸ کلمه	۱۶۰۴ کلمه	5K کلمه
	۵۰۰ کلمه معیار $\frac{WS}{\sqrt{WF}}$ + ۵۰۰ کلمه روش انتخاب ویژگی + ۱۱۵ کلمه روش وزن‌دهی دودویی	۱۰۰۰ کلمه معیار $\frac{WS}{\sqrt{WF}}$ + ۱۰۰۰ کلمه روش انتخاب ویژگی + ۱۱۵ کلمه روش وزن‌دهی دودویی	کلمه-های پرتکرار
MEN	۶۹.۳۹	۷۰.۶۶	۶۸.۶۲
RG-65	۶۳.۹۱	۶۷.۷۷	۶۲.۷۰
SimLex-999	۲۶.۲۰	۲۸.۳۵	۲۷.۶۶

جدول ۹ نتایج ارزیابی ترکیب‌های متفاوت بردارهای پایه با ابعاد کوچکتر مساوی ۱۰۰۰ را نشان می‌دهد. همانطور که مشاهده می‌کنید بردارهای پایه ترکیبی حاصل از سه روش در مجموعه داده SimLex-999 از بردارهای پایه ترکیبی حاصل از دو روش، به میزان ۰.۴٪ بهتر عمل کرده است. اما در مجموعه‌های داده MEN و SimLex-999، بردارهای پایه ترکیبی حاصل از دو روش به ترتیب به میزان ۰.۵۱٪ و ۳.۱۷٪ بهتر عمل کرده است. با مقایسه نتایج بردارهای پایه ترکیبی حاصل از دو روش با ۱۰۰۰ بردار پایه انتخاب شده توسط روش انتخاب ویژگی، شاهد افزایش دقت قابل ملاحظه ۳.۵۶٪ و ۷.۶۸٪ به ترتیب در مجموعه‌های داده MEN وRG-65 هستیم. در مجموعه SimLex-999 نیز ۰.۹۳٪ کاهش دقت رخ داده است. این نتایج حاکی از این است که بردارهای پایه ترکیبی به مراتب از روش انتخاب ویژگی با مقایسه ماتریس‌های فاصله که روش کارامدی است، عملکرد بهتری دارد.

جدول ۱۰ ارزیابی‌های ترکیب‌های متفاوت بردارهای پایه با ابعاد کوچکتر مساوی ۱۶۰۴ را نشان می‌دهد. همانطور که مشاهده می‌کنید، بردارهای ترکیبی حاصل از دو روش و بردارهای ترکیبی حاصل از سه روش (با حذف حدود ۳۵۰۰ بردار پایه) دقتی به مراتب بهتر از ۵۰۰۰ بردار پایه اولیه بدست آورده‌اند. بردارهای ترکیبی حاصل از دو روش در مجموعه داده RG-65 دقت را به میزان ۲.۳۲٪ بهبود بخشیده‌اند. با مقایسه عملکرد بردارهای ترکیبی حاصل از دو روش با ۱۵۰۰ بردار پایه انتخاب شده با روش انتخاب ویژگی بوسیله مقایسه ماتریس‌های فاصله، شاهد افزایش دقت مجموعه‌های داده آزمون MEN، RG-65 و SimLex-999 به ترتیب به میزان ۴.۱۳٪، ۹.۲۹٪ و ۱.۰۲٪ هستیم.

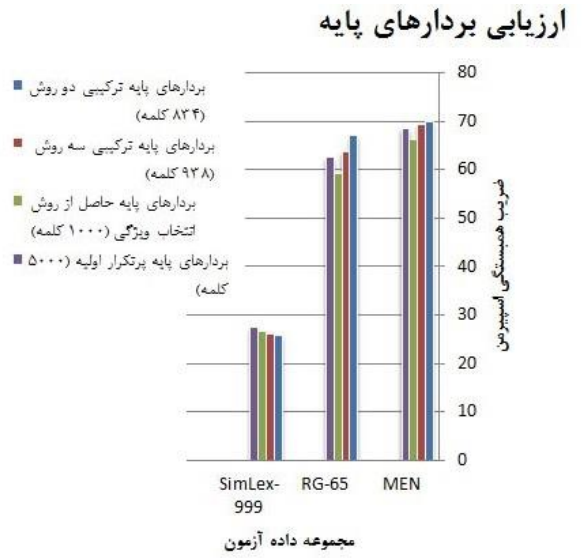
جدول۹: مقایسه ترکیب‌های متفاوت بردارهای پایه با ابعاد کوچکتر مساوی ۱۰۰۰.

مجموعه داده آزمون	بردارهای پایه		
	۹۳۸ کلمه	۸۳۴ کلمه	۱۰۰۰ کلمه
	۵۰۰ کلمه معیار $\frac{WS}{\sqrt{WF}}$ + ۵۰۰ کلمه روش انتخاب ویژگی + ۱۱۵ کلمه روش وزن‌دهی دودویی	۵۰۰ کلمه بدست آمده با معیار $\frac{WS}{\sqrt{WF}}$+۵۰۰ کلمه بدست آمده با روش انتخاب ویژگی	با روش انتخاب ویژگی
MEN	۶۹.۳۹	۶۹.۹۰	۶۶.۴۳
RG-65	۶۳.۹۱	۶۷.۰۸	۵۹.۴۰
SimLex-999	۲۶.۲۰	۲۵.۸۰	۲۶.۷۳

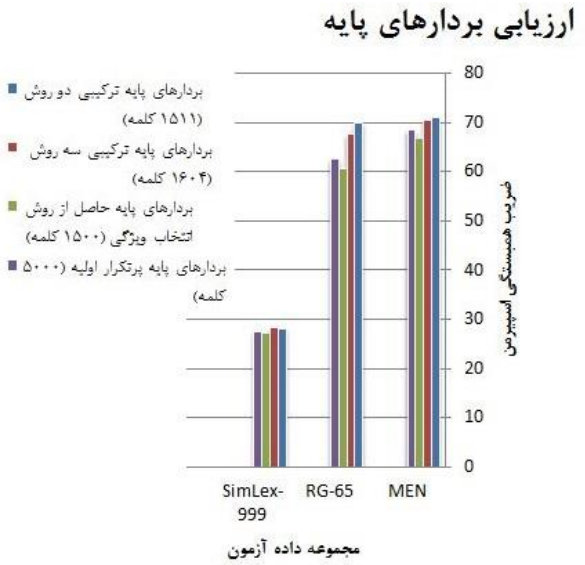
جدول۱۰: مقایسه ترکیب‌های متفاوت بردارهای پایه با ابعاد کوچکتر مساوی ۱۶۰۴.

مجموعه داده آزمون	بردارهای پایه		
	۱۶۰۴ کلمه	۱۵۱۱ کلمه	۱۵۰۰ کلمه
	۱۰۰۰ کلمه معیار $\frac{WS}{\sqrt{WF}}$ + ۱۰۰۰ کلمه روش انتخاب ویژگی + ۱۱۵ کلمه روش وزن‌دهی دودویی	1K کلمه بدست آمده با معیار $\frac{WS}{\sqrt{WF}}$+ 1K کلمه بدست آمده با روش انتخاب ویژگی	با روش انتخاب ویژگی
MEN	۷۰.۶۶	۷۱.۰۹	۶۶.۹۶
RG-65	۶۷.۷۷	۷۰.۰۹	۶۰.۸۰
SimLex-999	۲۸.۳۵	۲۸.۱۸	۲۷.۱۶

در شکل ۱ نتایج ارزیابی‌های ترکیب‌های متفاوت بردارهای پایه با ابعاد کوچکتر مساوی ۱۰۰۰ رسم شده است. همانطور که مشاهده کنید در مجموعه‌های داده مبتنی بر ارتباط MEN و RG-65، بردارهای پایه ترکیبی دو روش (۸۳۴ کلمه بافتار آگاهی‌بخش) نسبت به سایر بردارهای پایه انتخاب شده بهترین عملکرد را دارد. در مجموعه داده SimLex-999 در تمامی بردارهای پایه با ابعاد کم شاهد افت دقت هستیم. این افت دقت در بردارهای پایه حاصل از روش انتخاب ویژگی کوچکتر است. در شکل ۲، نتایج ارزیابی‌های ترکیب‌های متفاوت بردارهای پایه با ابعاد کوچکتر مساوی ۱۶۰۴ رسم شده است. نمودار نشان‌دهنده این نکته است که در دو مجموعه داده آزمون MEN و RG-65 عملکرد بردارهای پایه ترکیبی دو روش بهتر است. در مورد مجموعه داده آزمون SimLex-999، نتیجه اعمال بردارهای پایه ترکیبی سه روش اندکی بهتر از بردارهای پایه ترکیبی دو روش است. در هر دو بردارهای پایه ترکیبی دو روش (۱۵۱۱ بردار پایه) و بردارهای پایه ترکیبی سه روش (۱۶۰۴ بردار پایه) نسبت به بردارهای پایه اولیه (۵۰۰۰ بردار پایه)، ضریب همبستگی اسپیرمن در مجموعه‌های داده آزمون به میزان قابل ملاحظه‌ای بهبود یافته است. با وجود کاهش ابعاد بردارهای کلمه صریح از ۵۰۰۰ بعد به حدود ۱۵۰۰بعد، نه تنها دقت افت نکرده است، بلکه افزایش قابل ملاحظه‌ای در ضریب همبستگی اسپیرمن ایجاد شده است.



شکل ۱: ترکیب‌های متفاوت بردارهای پایه با ابعاد کوچکتر مساوی ۱۰۰۰.



شکل۲: ترکیب‌های متفاوت بردارهای پایه با ابعاد کوچکتر مساوی ۱۶۰۴.

۶- نتیجه‌گیری

هدف ما در این پژوهش انتخاب تعداد محدود بردارهای پایه تفسیرپذیر به‌گونه‌ای است که کاهش ابعاد بردارهای کلمه صریح، افت دقت زیادی در مجموعه‌های آزمون وظیفه شباهت کلمه ایجاد نکند. در این پژوهش دو رویکرد ترکیبی مبتنی بر روش‌های معیار نسبت شباهت کلمه به تکرار کلمه، انتخاب ویژگی مبتنی بر مقایسه ماتریس‌های فاصله و روش وزن‌دهی دودویی مبتنی بر BPSO برای انتخاب بردارهای پایه تفسیرپذیر ارائه کردیم.

در روش ترکیبی اول برای بهره‌گیری از خبرگی دو روش معیار نسبت شباهت کلمه به تکرار کلمه و روش انتخاب ویژگی مبتنی بر مقایسه ماتریس‌های فاصله، نیمی از بردارهای پایه را با استفاده از معیار نسبت شباهت کلمه به تکرار کلمه و نیم دیگر بردارهای پایه را با روش انتخاب ویژگی مبتنی بر مقایسه ماتریس‌های فاصله انتخاب کردیم. نتایج ارزیابی‌های این روش ترکیبی افزایش دقت ۰.۵-۷.۵٪ در اثر کاهش بعد از ۵۰۰۰ به حدود ۱۵۰۰ را نشان داده است. یعنی با کاهش ابعاد نه تنها کاهش دقت رخ نداده است، بلکه افزایش دقت هم ایجاد شده است.

در روش ترکیبی دوم، بردارهای مشترک بدست آمده با استفاده از روش وزن‌دهی دودویی را به بردارهای پایه روش ترکیبی اول افزودیم. نتایج ارزیابی روش ترکیبی دوم افزایش دقت به میزان ۰.۶-۵٪ در اثر کاهش بعد از ۵۰۰۰ به حدود ۱۶۰۰ را نشان داده است. بنابراین ما در این پژوهش با استفاده از روش‌های ترکیبی اول و دوم، حدود ۱۵۰۰ بردار پایه معنادار استخراج کرده‌ایم و با استفاده از آن‌ها، بردارهای کلمه صریح با ابعاد کم تولید کرده‌ایم. بردارهای کلمه حاصل بر روی مجموعه‌های آزمون وظیفه شباهت کلمه MEN، RG-65، SimLex-999 ارزیابی شده‌اند. نتایج رشد چشمگیر دقت بوسیله انتخاب صحیح بردارهای پایه معنادار را نشان داده است. یعنی با کاهش ابعاد از ۵۰۰۰ به حدود ۱۵۰۰ نه تنها کاهش دقت رخ نداده بلکه افزایش دقت نیز ایجاد شده است. هر کدام از بردارهای پایه معنادار معادل یک کلمه بافتار آگاهی‌بخش است. ما این کلمات بافتار را می‌توانیم در اختیار پژوهشگران علاقه‌مند قرار دهیم. محققین می‌توانند رویکرد ترکیبی پیشنهادی را بر روی پیکره مورد نظر خود اعمال کرده و بردارهای پایه معنادار انتخاب کنند. ما قصد داریم تا در آینده رویکرد عمومی ارائه شده در این پژوهش را برای کاربرد دسته‌بندی سوال، شخصی‌سازی کنیم. در وظیفه دسته‌بندی سوال، نیاز داریم تا اطلاعاتی در مورد کلمات جمله استخراج کنیم. ما نه تنها می‌توانیم رابطه بین هر کلمه جمله با مفاهیم ابعاد را استخراج کنیم، بلکه می‌توانیم میزان این رابطه را از ماتریس هم‌رخدادی تفسیرپذیر استخراج کنیم. بنابراین، می‌توان اطلاعاتی از ماتریس هم‌رخدادی برای هر کلمه سوال بدست آورد. با ترکیب اطلاعات استخراج شده و بردارهای کلمه صریح حاصل، می‌توانیم دقت وظیفه دسته‌بندی سوال را بهبود بخشیم.

جدول نمادها

WS	Word similarity	شباهت کلمه
WF	Word frequency	فراوانی کلمه
BPSO	Binary particle swarm optimization	بهنیه‌سازی ازدحام ذرات دودویی
SVD	singular value decomposition	تجزیه مقدار منفرد
NMF	non-negative matrix factorization	فاکتورگیری ماتریس غیرمنفی
PCA	principal component analysis	تحلیل مولفه اصلی
PPMI	Positive point-wise mutual information	اطلاعات متقابل نقطه‌ای مثبت
Skip-gram	Skip-gram negative sampling	نمونه‌گیری منفی skip-gram
CBOW	Continuous Bag of Words	کیسه پیوسته کلمه‌ها

مراجع

[1] A. Lenci, "Distributional models of word meaning", Annual review of Linguistics 4, pp.151-171, 2018.

[2] G. Salton, A. Wong, CS. Yang, "A vector space model for automatic indexing", Communications of the ACM 18.1, pp. 613-620, 1975.

[3] E. Grefenstette, "Category-theoretic quantitative compositional distributional models of natural language semantics", arXiv preprint arXiv, pp. 1311-1539, 2013.

[4] P. Gamallo, "Comparing explicit and predictive distributional semantic models endowed with syntactic contexts", Language Resources and Evaluation 51.3, pp.727-743, 2017

[5] M. Baroni, G. Dinu, G. Kruszewski , "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors", 52nd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, 2014.

[6] Y. Li, T. Yang, "Word embedding for understanding natural language: a survey", Guide to big data applications, Springer, Cham, pp. 83-104, 2018.

[7] C. Goddard , "Semantic analysis: A practical introduction", Oxford University Press, 2011.

[8] EH. Huang, R. Socher, CD. Manning, AY. Ng, "Improving word representations via global context and multiple word prototypes",

50th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, 2012.

[9] J. Andreas, D. Klein, "How much do word embeddings encode about syntax? ", 52nd Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, 2014.

[10] M. Pota, F. Marulli, M. Esposito, G. De Pietro, H. Fujita, "Multilingual POS tagging by a composite deep architecture based on character-level features and on-the-fly enriched word embeddings", Knowledge-Based Systems, 164, pp.309-323, 2019.

[11] SM. Rezaeinia, R. Rahmani, A. Ghodsi, H. Veisi , "Sentiment analysis based on improved pre-trained word embeddings", Expert Systems with Applications, 117, pp.139-147, 2019.

[12] F. Ali, D. Kwak, P. Khan, S. El-Sappagh, A. Ali, S. Ullah, K.H. Kim and K.S. Kwak, "Transportation sentiment analysis using word embedding and ontology-based topic modeling", Knowledge-Based Systems, 174, pp.27-42, 2019.

[13] B.O. Deho, A.W. Agangiba, L.F. Aryeh and A.J. Ansah, "Sentiment analysis with word embedding", In 2018 IEEE 7th International Conference on Adaptive Science & Technology, pp. 1-4, IEEE, 2018.

[14] D. Nozza, P. Manchanda, E. Fersini, M. Palmonari and E. Messina, "LearningToAdapt with word embeddings: Domain adaptation of Named Entity Recognition systems", Information Processing & Management, 58(3), 102537, 2021.

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

[16] C. Sundermann, J., Antunes, M. Domingues, and S. Rezende, "Exploration of word embedding model to improve context-aware recommender systems", In 2018 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 383-388, IEEE, 2018.

[17] D. Khattar, V. Kumar, V. Varma, M. Gupta, "Weave&rec: A word embedding based 3-d convolutional network for news recommendation", 27th ACM International Conference on Information and Knowledge Management, pp. 1855-1858, 2018.

[18] E. Bagheri, F. Ensan, F. Al-Obeidat, "Neural word and entity embeddings for ad hoc retrieval", Information Processing & Management, 54(4), pp.657-673, 2018.

[19] Dobó, J. Csirik, "A comprehensive study of the parameters in the creation and comparison of feature vectors in distributional semantic models", Journal of Quantitative Linguistics, 27.3, pp. 244-271, 2020.

[20] C. Heunen, M. Sadrzadeh, E. Grefenstette, "Quantum physics and linguistics: a compositional, diagrammatic discourse", Oxford University Press, 2013.

[21] D. Kartsaklis, "Compositional operators in distributional semantics", Springer Science Reviews, 2.1-2, pp. 161-177, 2014.

[22] O. Levy, Y. Goldberg, I. Dagan, "Improving distributional similarity with lessons learned from word embeddings", Transactions of the Association for Computational Linguistics, 3, pp. 211-225, 2015.

[23] JA. Bullinaria, JP. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study", Behavior research methods, 39.3, pp. 510-526, 2007.

[24] C. Biemann, M. Riedl , "Text: Now in 2D! a framework for lexical expansion with contextual similarity", Journal of Language Modelling, 1.1, pp. 55-95, 2013.

[25] M Padró, M Idiart, A Villavicencio, C. Ramisch, "Nothing like good old frequency: Studying context filters for distributional thesauri", Conference on Empirical Methods in Natural Language Processing, 2014.

[26] P. Gamallo, S. Bordag , "Is singular value decomposition useful for word similarity extraction?", Language resources and evaluation 45.2, pp. 95-119, 2011.

[27] MJ. Hofmann, AM. Jacobs, "Interactive activation and competition models and semantic context: from behavioral to brain data", Neuroscience & Biobehavioral Reviews, 46, pp. 85-104, 2014.

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

[29] T. Mikolov, I. Sutskever, K. Chen, GS. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality", arXiv preprint arXiv, 1310.4546, 2013.

[30] Panigrahi, HV. Simhadri, and C. Bhattacharyya, "Word2Sense: sparse interpretable word embeddings", 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[31] B. Murphy, P. Talukdar, T. Mitchell , "Learning effective and interpretable semantic models using non-negative sparse embedding", COLING, 2012.

[32] F. Sun, J. Guo, Y. Lan, J. Xu, X. Cheng, "Sparse word embeddings using l1 regularized online learning", Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016.

[33] M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, N. Smith, "Sparse overcomplete word vector representations", arXiv preprint arXiv, 1506.02004, 2015.

[34] Subramanian, D. Pruthi, H. Jhamtani, T. Berg-Kirkpatrick, E. Hovy, "Spine: Sparse interpretable neural embeddings", AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[35] E. Bruni, G. Boleda, M. Baroni, NK. Tran, "Distributional semantics in technicolor", 50th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, 2012.

[36] H. Rubenstein, JB. Goodenough, "Contextual correlates of synonymy", Communications of the ACM, 8.10, pp. 627-633, 1965.

[37] F. Hill, R. Reichart, A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation", Computational Linguistics, 41.4, pp. 665-695, 2015.

[38] K. Yang, C. Shahabi, "A PCA-based similarity measure for multivariate time series", 2nd ACM international workshop on Multimedia databases. 2004.

[39] J. Kennedy, R. Eberhart, "Particle swarm optimization", ICNN'95-international conference on neural networks, Vol. 4, IEEE, 1995.

[40] H. Garg, "A hybrid PSO-GA algorithm for constrained optimization problems", Applied Mathematics and Computation, 274, pp. 292-305, 2016.

[41] J. Kennedy, RC. Eberhart, "A discrete binary version of the particle swarm algorithm", Computational cybernetics and simulation, Vol. 5, IEEE, 1997.

[42] AH. El-Maleh, AT. Sheikh, SM. Sait, "Binary particle swarm optimization (BPSO) based state assignment for area minimization of sequential circuits", Applied soft computing, 13.12, pp. 4832-4840, 2013.

[43] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora", Language resources and evaluation, 43.3, pp. 209-226, 2009.

پاورقی‌ها:

[1] Distributional semantics
[2] Lexical items
[3] Contexts
[4] Count-based models
[5] Predictive models
[6] Explicit
[7] Implicit
[8] Word embedding
[9] Semantic analysis
[10] Syntax analysis
[11] POS tagging
[12] Sentiment analysis
[13] Named entity recognition
[14] Keyword extraction
[15] Recommendation systems
[16] Information retrieval
[17] Overfit
[18] Principal component analysis
[19] Nonnegative matrix factorization
[20] Semantic similarity
[21] Semantic relatedness
[22] Target word
[23] Basis vectors
[24] Unit
[25] Association
[26] Pointwise mutual information
[27] Latent
[28] Feature extraction
[29] Dense
[30] Hashing
[31] Keys
[32] Likelihood
[33] Continuous Bag of Words
[34] regularizer
[35] Sparse Overcomplete Word Vectors
[36] K-sparse denoising autoencoder
[37] baseline
[38] heuristic
[39] http://wacky.sslmit.unibo.it/
[40] Web crawling