University of Qom

Sciences and Techniques of Information Management

Volume 9, Issue 3, 2023

Presenting a Scale-Free Complex Network with a Persian Language Layered Composition Pattern

Ali Sarabadani
P.hD., Student, Department of Computer and Information Technology, Technical and Engineering Faculty, Qom University, Qom, Iran. alisarabadani14@gmail.com

Khairale Rahsafarfard
Assistant Professor, Department of Computer Engineering and Information Technology, Faculty of Technology and Engineering, University of Qom, Qom, Iran **(Corresponding author)**.
rahsepar@qom.ac.ir

Sepideh Chehreh
P.hD., Student, Department of Computer and Information Technology, Technical and Engineering Faculty, Qom University, Qom, Iran. s.chehreh@stu.qom.ac.ir

Abstract

Purpose: This article proposes a method for investigating the patterns of composition and topological structure of the Persian language. The enhanced method analyzes Persian text by representing it as a simultaneous network graph within the framework of complex network theory.

Method: A null model of the same size is generated using the Erdos-Renyi random graph for comparison with the Persian network. The comparison is based on the average path length, clustering coefficient, and hierarchy of both networks. From the analysis of these key features, it can be seen that the Persian network graph differs from the random network. The smaller average path length and high clustering coefficient also confirm the influence of the small-world model in the Persian language.

Findings: For the first time, the Persian text was successfully converted into a complex network. An open, unbounded set of over two million words is created using a random forest approach.

Conclusion: The resulting network designed using the Bygram bag model contains 3256 nodes and 79705 edges. In addition, unlike the random network where there is only one community, 12 communities have been identified in the Persian network. Statistical evidence indicates that the Persian network is a scale-free network with a layered composition pattern.

Keywords: Persian Language, Natural Language Processing, Complex Network, Small World Model, Layered Composition Model.

Cite this article: Sarabadani, A., Rahsafarfard, K. & Chehreh, S. (2023). Presenting a Scale-Free Complex Network with a Persian Language Layered Composition Pattern. *Sciences and Techniques of Information Management*, 9(3): 215-240. https://doi.org/10.22091/STIM.2022.8590.1858

Received: 2023-06-25 ; **Revised:** 2023-07-14 ; **Accepted:** 2023-07-24 ; **Published online:** 2023-07-27

Article type: Research **Published by:** University of Qom.

علوم و فنون مدیریت اطلاعات
دوره ۹، شماره ۳، ۱۴۰۲

ارائه شبکه پیچیده بدون مقیاس با الگوی ترکیب لایه‌ای زبان فارسی

علی سرآبادانی
دانشجوی دکتری، گروه کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران.
alisarabadani14@gmail.com

خیراله رهسپارفرد
استادیار، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران (**نویسنده مسئول**).
rahsepar@qom.ac.ir

سپیده چهره
دانشجوی دکتری، گروه کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران.
s.chehreh@stu.qom.ac.ir

چکیده

هدف: پژوهش حاضر روشی را برای بررسی الگوهای ترکیب و ساختار توپولوژیکی زبان فارسی پیشنهاد کرده، و روش بهبودیافته متن فارسی را در قالب گراف شبکه همزمان در چارچوب نظریه شبکه پیچیده بررسی می‌کند.

روش: یک مدل تهی با اندازه مشابه، با توجه به گراف تصادفی اردوش- رینی، برای مقایسه با شبکه فارسی تولید می‌شود. مقایسه براساس طول مسیر متوسط، ضریب خوشه‌بندی و سلسله مراتب هر دو شبکه است. از تجزیه و تحلیل این ویژگی‌های کلیدی، مشاهده می‌شود که گراف شبکه فارسی با شبکه تصادفی متفاوت است. طول مسیر متوسط کوچک‌تر و ضریب خوشه‌بندی بالا نیز تأثیر مدل جهانی کوچک را در زبان فارسی تأیید می‌کند.

یافته‌ها: برای اولین بار، متن فارسی با موفقیت به شبکه پیچیده تبدیل شد. یک مجموعه باز و بدون حاشیه بیش از دو میلیون کلمه، با استفاده از رویکرد جنگل تصادفی ساخته شده است.

نتیجه‌گیری: شبکه حاصل طراحی شده، با مدل کیسه بایگرام شامل ۳۲۵۶ گره و ۷۹۷۰۵ لبه می‌باشد. علاوه‌بر این، برخلاف شبکه تصادفی که تنها یک جامعه وجود دارد، ۱۲ اجتماع در شبکه فارسی شناسایی شده است. واقعیت‌های آماری نشان می‌دهد که شبکه فارسی یک شبکه بدون مقیاس با الگوی ترکیب لایه‌ای است.

کلیدواژه‌ها: زبان فارسی، پردازش زبان طبیعی، شبکه پیچیده، مدل جهان کوچک، الگوی ترکیب لایه‌ای.

استناد به این مقاله: سرآبادانی، ع.، رهسپارفرد، خ.، چهره، س. (۱۴۰۲). ارائه شبکه پیچیده بدون مقیاس با الگوی ترکیب لایه‌ای زبان فارسی. *علوم و فنون مدیریت اطلاعات*، ۹(۳): ۲۱۵-۲۴۰. https://doi.org/10.22091/STIM.2022.8590.1858

تاریخ دریافت: ۱۴۰۲/۰۴/۰۴ ؛ **تاریخ اصلاح:** ۱۴۰۲/۰۴/۲۳ ؛ **تاریخ پذیرش:** ۱۴۰۲/۰۵/۰۲ ؛ **تاریخ انتشار آنلاین:** ۱۴۰۲/۰۵/۰۵

نوع مقاله: پژوهشی
ناشر: دانشگاه قم

۱. مقدمه

زبان راه ارتباطی انسان‌ها برای نشان دادن عبارت‌های مختلف، به اشتراک گذاشتن تجربه و گسترش دانش است. با این وجود، پژوهشگران زبان‌شناس، زبان را سیستم پیچیده‌ای می‌دانند که از قوانین مشخصی پیروی می‌کند (فرامکین، رادمن و هایماس[۱]، ۲۰۱۸). به عقیده پژوهشگران، زبان‌ها به شیوه‌ای نظام‌مند تکامل می‌یابند و ساختاری کاملاً مشخص با شماری از الگوهای پنهان دارند. تجزیه و تحلیل آماری و درک عمیق این الگوها این توانایی را دارد تا ساختار هر زبانی را آشکار نماید (راسل و نوروینگ[۲]، ۲۰۱۶؛ لیچن، بنگیو و هیلتن[۳]، ۲۰۱۵؛ رابرت[۴]، ۲۰۱۴). پردازش زبان طبیعی[۵] یک زمینه پژوهشی برای پیدا کردن الگوهای زبان‌ها بوده و به گونه‌ای است که رایانه‌ها بتوانند آن‌ها را درک کنند. به این منظور، دانشمندان قصد دارند رایانه‌ها را چنان مستقل و قابل اعتماد بسازند که بدون تعامل انسانی، ماشین‌ها بتوانند وظیفه‌های چندزبانی را به روشی هوشمندانه انجام دهند. بنابراین، تحقیق‌هایی برای دانستن روند دقیق یادگیری زبان در انسان در حال انجام است. ایده شبکه جهانی کوچک اساساً متعلق به حوزه روانشناسی است. هدف اصلی این مفهوم بررسی فرآیند یادگیری زبان در انسان و ساختار توپولوژیکی زبان ذخیره شده در مغز انسان است. تئوری‌های پذیرفته شده (کانچو و سوله[۶]، ۲۰۰۱) بیان می‌کنند که کلمه‌ها در مغز انسان به صورت اجزای زبانی وابسته به هم ذخیره می‌شوند. کلمه‌ها از هزاران واژه تشکیل شده‌اند و طبق ترتیب مشخصی همراه با مفهوم‌ها، معنی‌ها و وابستگی دوسویه، حفظ می‌شوند و دلیل مدیریت کارآمد زبان توسط مغز انسان، نیز ویژگی‌های شبکه‌های کوچک جهانی است. از نظر مفهومی، عناصر زبان در قالب گره‌ها و پیوندهای متصل ذخیره می‌شوند که این شیوه در زمینه پردازش زبان طبیعی برای یادگیری ماشین[۷] و هوش مصنوعی[۸] نیز پیاده‌سازی شده است. با کمک مدل‌های زبانی مختلف که در مغز انسان ذخیره شده است، مدل‌های زبانی جدیدی مانند هوش مصنوعی، علوم کامپیوتر، فناوری اطلاعات، پردازش اطلاعات،

1. Fromkin, Rodman & Hyams
2. Russell & Norvig
3. LeCun, Bengio & Hinton
4. Robert
5. Natural Language Processing
6. Cancho & Solé
7. Machine learning
8. Artificial Intelligence

تغییر داده‌ها، پردازش متن و گفتار، رباتیک، ریاضیات، تاریخ، روانشناسی و زبان‌شناسی در حال توسعه هستند. بر این اساس، برخی از پیاده‌سازی‌های مهم این پژوهش از پردازش متن، ترجمه، تشخیص گفتار، ترجمه گفتار، تحلیل محتوا، خلاصه‌سازی متن و گفتار، چندزبانی، وظیفه‌های دوسویه زبانی، ابزارهای بازیابی، ترجمه ماشینی، و رابط کاربری هوشمند برای سیستم‌های بازیابی اطلاعات چندزبانه و سیستم‌های بازیابی اطلاعات متقابل زبانی مستقل[1] تشکیل شده است (گارهام[2]، ۲۰۱۷). این سیستم‌ها با حذف بازدارنده‌های زبانی و شرکت‌کننده‌های اصلی در سازمان‌دهی مرتبط هستند. در عصر رایانه‌ها، بهبود ویژگی‌های چندزبانی سیستم‌ها و ابزارهای هوشمند کنونی برای سازگار کردن آن‌ها با زبان‌های دنیا، به عنوان یك عنصر حیاتی شناخته شده‌اند. بنابراین، مدل‌های زبان‌شناختی مصنوعی را می‌توان از سیستم‌های زبان واقعی، پس از آشکار ساختن الگوهای پنهان در زبان‌های انسانی تولید کرد و در این مدل‌ها، ماشین‌ها با ویژگی‌های چندزبانه، این توانایی را دارند که به عنوان یك عنصر سودمند، مورد استفاده قرار گیرند (جن و لئو[3]، ۲۰۱۹؛ سایگل و همکاران[4]، ۲۰۱۸).

پژوهش حاضر روشی را برای بررسی الگوهای ترکیب و ساختار توپولوژیکی زبان فارسی پیشنهاد می‌کند. همچنین روش بهبودیافته متن فارسی را در قالب گراف شبکه همزمان، در چارچوب نظریه شبکه پیچیده بررسی می‌نماید. برای اولین بار، متن فارسی با موفقیت به شبکه پیچیده تبدیل شد. یک مجموعه باز و بدون حاشیه بیش از دو میلیون کلمه، با استفاده از رویکرد جنگل تصادفی ساخته است.

۲. پیشینه پژوهش

در این پژوهش مقاله‌های مختلف در دامنه موضوعی مورد مطالعه، بررسی شده و پس از بررسی مجموعه داده‌ها و شاخص‌های ارزیابی مورد استفاده در آن‌ها، به یک جمع‌بندی برای مدل پیشنهادی رسیده است.

جدول (۱) تحقیقات پیشین مورد بررسی را معرفی می‌کند.

1. CLIR
2. Garnham
3. Chen & Lou
4. Siegel

جدول ۱- مروری بر تحقیقات پیشین

منابع	نقاط ضعف	نقاط قوت	انواع راهکارها	دسته اصلی
(پریونسی و همکاران[۳]، ۲۰۲۰)	عدم دسترسی، به شرط مستقل بودن داده در دنیای واقعی	دسته‌بندی سریع و آسان، عملکرد بهتر در صورت مستقل بودن داده،	نظریه بیز[۲]	کیسربا ناظر[۱]
(کیسرو، اندرو و همبرگ[۵]، ۲۰۱۹؛ باران گال و همکاران[۶]، ۲۰۲۰)	محاسبات سنگین، نیاز به حافظه بالا، ذخیره داده اولیه، حساس به ویژگی نامناسب	عدم نیاز به داده پیش‌فرض، دقت بالا، چند منظوره، الگوریتم ساده	کی- نزدیک‌ترین همسایه[۴]	
(پاول و همکاران[۸]، ۲۰۱۸؛ ژنگ و همکاران[۹]، ۲۰۲۰)	تعریف مناسب پارامترها	گیر نیافتادن در ماکزیمم محلی، تخمین داده با ابعاد بالا، کنترل میزان خطا	ماشین بردار پشتیبان[۷]	
(لین و همکاران[۱۲]، ۲۰۱۹؛ برور و همکاران[۱۳]، ۲۰۱۹)	زمان محاسبه زیاد برای داده با ابعاد بالا، افزایش محاسبات	سادگی و قابلیت درک	خوشه‌بندی سلسله مراتبی[۱۱]	بدون ناظر[۱۰]
(هووارد و همکاران[۱۵]، ۲۰۱۸؛ سان و همکاران[۱۶]، ۲۰۱۹)	عدم تطبیق با عنواع داده‌ها	کشف قوانین موجود در بین حجم عظیم داده‌ها	قوانین انجمنی[۱۴]	
(زی و همکاران[۱۸]، ۲۰۲۰)	نیازمند تعریف دسته‌ها	امکان انجام محاسبات زیاد برای داده با ابعاد بالا، کاهش محاسبات	کی- میانگین[۱۷]	

1. Supervised
2. Naive Bayes
3. Piryonesi.
4. K-Nearest Neighbor
5. Kiselev, Andrews & Hemberg
6. Baran-Gale
7. Support Vector Machine
8. Paul
9. Zhang
10. Unsupervised
11. Hierarchical Cluster Analysis
12. Lin
13. Breuer.
14. Association Rule Learning
15. Howard
16. Sun.
17. K-Means
18. Xie

بنابراین، با گسترش سیستم‌های پایگاهی و حجم داده‌های ذخیره شده در این سیستم‌ها، به ابزاری نیاز است تا بتوان این داده‌ها را پردازش کرد و اطلاعات حاصل از آن را در اختیار کاربران قرار داد. امروزه به روش‌هایی نیاز داریم که به اصطلاح به کشف دانش بپردازند. یعنی روش‌هایی که با کمترین دخالت کاربر و به صورت خودکار، الگوها و رابطه‌های منطقی را بیان نمایند. یکی از روش‌های مهم که با آن می‌توان الگوهای مفیدی را در میان داده‌ها تشخیص داد، داده‌کاوی است. این روش از فرآیند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های بزرگ تشکیل شده است که از آن در تصمیم‌گیری‌ها در پژوهش‌های مختلف و در فعالیت‌های تجاری مهم استفاده می‌شود. پردازش زبان طبیعی یک زمینه پژوهشی برای پیدا کردن الگوهای زبان‌ها بوده و به گونه‌ای است که رایانه‌ها بتوانند آن‌ها را درک کنند. زبان فارسی به دلیل متن دوسویه، خط دشوار، مرزهای نامشخص کلمه‌ها، انواع مختلف کلمه‌ها و مشکل‌های موجود در ریشه‌یابی، به عنوان زبانی سخت برای پردازش معرفی شده است. بنابراین، زبان فارسی بین پژوهشگران و توسعه‌دهندگان ابزارهای زبان‌شناسی، زبانی شناخته نشده باقی مانده است. رمزگذاری یونی‌کد و عدم پژوهش در مورد زبان فارسی، موجب معرفی آن به عنوان یک زبان منبع ضعیف شده است (دائود، خان و چی[1]، ۲۰۱۷). پیش از این، برخی از پژوهش‌های سنّتی در مورد زبان فارسی شامل ساخت پیکره[2]، توسعه مترجمان و ابزارهای تایپ مانند صفحه کلید فارسی و موارد مشابهی از این قبیل، مورد استفاده قرار گرفته بود (خان، باخت و واگن[3]، ۲۰۱۹). ساختار دستوری زبان فارسی تا حدودی با استثناءها و بهانه‌های زیادی مورد مطالعه قرار گرفته است. اما هنوز پژوهش‌های عمیق آماری در مورد ساختار این زبان در مقایسه با سایر زبان‌ها، به درستی انجام نشده است (یول[4]، ۲۰۱۴). بنابراین، با استفاده از منابع محدود زبانی، برای اولین بار با حداقل پردازش دستی، زبان فارسی با موفقیت به شبکه تبدیل شده و ایده شبکه زبان از نظر تبدیل زبان به گراف، بررسی ساختار و اجرای خصوصیت‌های شبکه پیچیده بر روی آن، مورد توجه قرار گرفته است. از طرفی می‌توان به معتبرترین پایگاه داده زبان فارسی به نام فارس بیس اشاره کرد. فارس‌بیس مجموعه بزرگی از موجودیت‌ها و ارتباط بین آن‌ها است که یک پایگاه دانش پویا را تشکیل می‌دهد. گراف دانش فارسی با بیش از ۵۰۰ هزار موجودیت زبان فارسی و

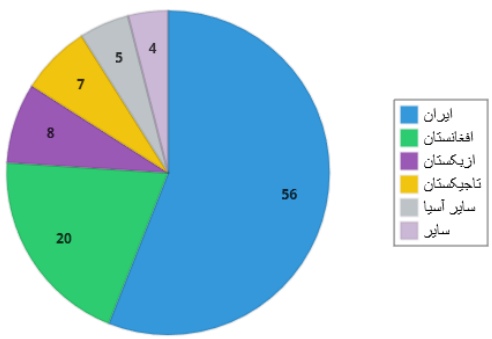
1. Daud, Khan & Che
2. Corpus
3. Khan, Bakht & Wagan
4. Yule

۷ میلیون رابطه میان آن‌ها ایجاد شده است. با توجه به اینکه این گراف بین دامنه‌ای است، در حوزه‌های متنوع اشخاص مشهور، مکان‌های مهم، سازمان‌ها و شرکت‌ها، آثار ادبی و هنری، گونه‌های زیستی شامل گیاهان و حیوانات، رویدادها، زیست‌شناسی، اخترشناسی دارای اطلاعات غنی است. از مهم‌ترین مولفه‌های این محصول می‌توان به سامانه جستجو روی گراف دانش اشاره نمود. با استفاده از این مولفه می‌توان به موتورهای جستجو خدمت‌های ارزنده‌ای ارائه نمود، به طوری‌که قابلیت جستجو روی پرسش‌های پر کاربرد دنیای وب وجود دارد. در برخی از پرس‌وجوها، پاسخ‌های گراف دانش فارسی، از گراف دانش گوگل در زبان فارسی بهتر عمل می‌نماید (سجادی و مینایی بیدگلی، ۱۳۹۸).

۳. ادبیات موضوع

۳-۱. زبان فارسی

فارسی ایرانی[1] یا فارسی غربی یا پارسی باختری، یکی از گونه‌های زبان فارسی[2] است. این زبان، زبان رسمی ایران بوده و توسط اقلیّت‌های بزرگی در عراق و کشورهای عربی خلیج فارس، به ویژه بحرین، گفتگو می‌شود. فارسی ایرانی در کنار فارسی افغانستانی و فارسی تاجیکی، یکی از سه گویش عمده و رسمی فارسی در جهان است. ایران بزرگ‌ترین کشور فارسی زبان جهان بوده و تهران بزرگ‌ترین شهر فارسی زبان جهان است و ۹۸ درصد تهرانی‌ها به زبان فارسی تسلّط دارند. شکل (۱) پراکندگی گویش فارسی در جهان را برحسب درصد نشان می‌دهد.



شکل ۱- پراکندگی گویش فارسی در جهان

1. Iranian Persian
2. Persian Language

۳-۲. شبکه پیچیده

شبکه‌های پیچیده در مطالعه علوم مدرن مانند مطالعه شبکه‌های بیولوژیکی، شبکه‌های قدرتی، اقتصاد کلان و مسائل استنتاجی بر روی گراف‌ها، بسیار رایج هستند. زبان، شبکه‌ای از عناصر به هم پیوسته است که در درون آن قوانین، ساختار و الگوهایی وجود دارد و می‌توان آن را به عنوان یک سیستم پیچیده از اشیاء در نظر گرفت و به طور تصادفی بررسی کرد. شبکه‌های پیچیده در مطالعه علوم مدرن مانند مطالعه شبکه‌های بیولوژیکی، شبکه‌های قدرتی، اقتصاد کلان و مسائل استنتاجی بر روی گراف‌ها بسیار رایج هستند. در بسیاری از سیستم‌های پیچیده، به ویژه با آن‌هایی که در طبیعت مواجه می‌شویم، مانند شکل‌های پرواز پرندگان و حرکت دسته‌جمعی ماهی‌ها، رفتار کلی آن‌ها از رفتار جزئی و تک‌به‌تک نشأت می‌گیرد (ویلهلمن[1]، ۲۰۰۸). در حالی‌که هر گره عامل در این شبکه‌های بیولوژیکی، قابلیت رفتار پیچیده را ندارد و ترکیب بین چندین گره باعث به وجود آمدن رفتارهای بسیار پیچیده در سطح شبکه می‌گردد. تلاش‌های پژوهشی سعی بر کشف پیچیدگی چنین شبکه‌های پیچیده‌ای مانند پردازش سیگنال، یادگیری ماشین، بهینه‌سازی، کنترل، آمار، علوم کامپیوتری و اجتماعی را دارند (باراباسی و بونابو[2]، ۲۰۰۳). در تمامی این زمینه‌ها یک علاقه روبه رشد در مسائل اجتماعی و یادگیری بر روی گراف‌ها مانند استنباط روابط از ارتباطات درونی بر روی شبکه‌های اجتماعی، مدل کردن واکنش‌های بین عامل‌ها در شبکه‌های بیولوژیکی، انتشار اطلاعات بین عامل‌های توزیع‌یافته و بهینه‌سازی تابع‌های کاربردی دیده می‌شود. درواقع شبکه‌های واقعی با خاصیت دنیای کوچک، خاصیت کوتاهی طول مسیر را از شبکه‌های تصادفی، و معقول بودن ضریب خوشگی را از شبکه‌های منظم، به دلیل ویژگی‌های شبکه پیچیده به‌دست آورده‌اند (استراگاتز و وات[3]، ۱۹۹۸؛ استنلی و همکاران[4]، ۲۰۰۰).

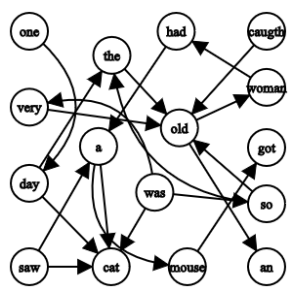
در زمینه تئوری شبکه[5]، یک شبکه پیچیده[6] یک گراف با ویژگی‌های توپولوژیکی است که این ویژگی‌ها در شبکه‌های ساده مانند شبکه‌ها و گراف‌های تصادفی رخ نمی‌دهند، اما اغلب در شبکه‌هایی که سیستم‌های واقعی را نشان می‌دهند، رخ می‌دهند. مطالعه شبکه‌های پیچیده، یک حوزه

1. Wilhelm
2. Barabasi & Bonabeau
3. Strogatz & Watts
4. Stanley
5. Network Theory
6. Complex Network

جوان و فعال از تحقیق‌های علمی است (آلبرت و باربسی[۱]، ۲۰۰۲؛ نیومن[۲]، ۲۰۱۰). شبکه‌های پیچیده عمدتاً از یافته‌های تجربی شبکه‌های دنیای واقعی مانند شبکه‌های کامپیوتری، شبکه‌های بیولوژیکی، شبکه‌های فناوری، شبکه‌های مغزی، الهام گرفته شده است (باسط و اسپرنز[۳]، ۲۰۱۷؛ فورنیتو[۴]، ۲۰۲۰؛ صابری و همکاران[۵]، ۲۰۲۱). بیشتر شبکه‌های اجتماعی، زیست‌شناختی و فناوری، ویژگی‌های توپولوژیکی غیر پیش پا افتاده را از طریق الگوهای ارتباطی بین عناصر خود نشان می‌دهند که این الگوها کاملاً منظم و تصادفی نیستند.

۳-۳. شبکه‌های زبان

زبان راه ارتباطی انسان‌ها برای نشان دادن عبارت‌های مختلف، به اشتراک گذاشتن تجربه و گسترش دانش است. با این وجود، پژوهشگران زبان‌شناس، زبان را سیستم پیچیده‌ای می‌دانند که از قوانین مشخصی پیروی می‌کند. زبان شبکه‌ای از عناصر به هم پیوسته است که در درون آن قوانین، ساختار و الگوهایی وجود دارد و می‌توان آن را به عنوان یک سیستم پیچیده از اشیاء در نظر گرفت و به طور تصادفی بررسی کرد (جائو و همکاران[۶]، ۲۰۱۷). برای ساخت شبکه زبانی، شبکه نحوی، شبکه معنایی یا شبکه‌های هم‌رخداد بسته به پیکره و هدف تحقیق در نظر گرفته می‌شود (جائو، لوك و چنگ[۷]، ۲۰۱۸). شکل (۲)، شبکه ساده‌ای از متن زبان انگلیسی را نشان می‌دهد که با کلمه‌ها به عنوان گره‌های متصل، از طریق پیوندها به گراف شبکه تبدیل شده است.

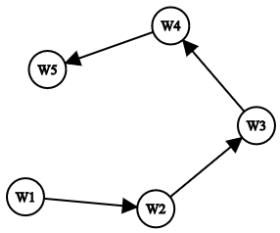


شکل ۲- شبکه ساده‌ای از متن زبان انگلیسی

1. Albert & Barabási
2. Newman
3. Bassett & Sporns
4. Fornito
5. Saberi
6. Gao & Kurths
7. Goh, Luke & Cheong

۳-۴. شبکه همزمانی[1]

تکنیک‌های مختلفی برای ساخت شبکه زبانی تعریف شده‌اند، اما پذیرفته‌شده‌ترین آن‌ها شبکه همزمانی است که کمك می‌کند رابطه بین دو کلمه در یک متن شناخته شود (چن و همکاران[2]، ۲۰۱۸). در این ساختار کلمه‌هایی که در یک جمله با هم ظاهر می‌شوند، براساس همزمانی به هم متصل می‌گردند. اگر به عنوان مثال یک جمله کوتاه از پنج کلمه تشکیل شده باشد، آنگاه می‌توان هر کلمه را با w و مکان آن‌ها را با اعداد متوالی نشان داد، بنابراین، جمله به w1, w2, w3, w4 و w5 تبدیل می‌شود. شکل (۳) ساده‌ترین شبکه همزمان را نشان می‌دهد.



شکل ۳- ساده‌ترین شبکه همزمان

در شبکه‌های هم زمان، شبکه‌ها با دو رویکرد متفاوت ساخته می‌شوند. رویکرد اول، از طریق تکنیک انتخاب کلمه‌های منحصربه‌فرد از متن، با انتخاب یک جمله، پاراگراف یا صفحه صورت می‌پذیرد. رویکرد دوم از طریق پیوند دادن عناصر با توجه به متن پژوهش انجام می‌گیرد و این رویکرد اطلاعات دقیق‌تری در مورد ساختار می‌دهد (چن و همکاران، ۲۰۱۸). شبکه‌های همزمان با توجه به لینک‌ها به عنوان شبکه‌های جهت‌دار، شبکه‌های هدایت نشده، شبکه‌های وزن‌دار و شبکه‌های بدون وزن، طبقه‌بندی می‌شوند.

۳-۵. شبکه‌های جهت‌دار[3]

در شبکه‌های جهت‌دار، جهت پیوندها در بین گره‌ها، به‌عنوان درجه داخلی و درجه خارجی در نظر گرفته می‌شود.

۳-۶. شبکه‌های بدون جهت[4]

در صورت نادیده گرفتن جهت‌ها در گراف مربوط به کلمه‌ها، تمامی کلمه‌ها بدون در نظر گرفتن

1. Co-Occurrence Network
2. Chen
3. Directed Networks
4. Undirected Networks

ترتیب کلمات، پیوند داده می‌شوند. به این نوع شبکه، شبکه بدون جهت گفته می‌شود. تفاوت اصلی بین شبکه‌های جهت‌دار و بدون جهت، تعداد پیوندهای متفاوت برای تعداد مساوی گره است. وزن‌ها معمولاً مقدار مثبت تعداد پیوندهای بین دو گره هستند. وزن لبه در شبکه، وزن‌دار در نظر گرفته می‌شود و در شبکه بدون وزن، وزن به یال‌ها اختصاص داده نمی‌شود. یک شبکه همزمان که به شکل گراف مدل‌سازی شده است، رفتار، ترکیب و ویژگی‌های شبکه متنی را نشان می‌دهد.

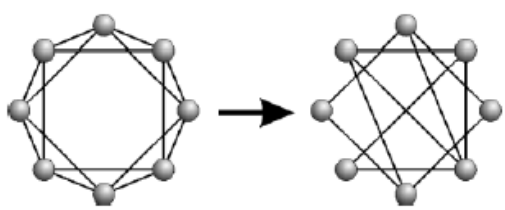
۳-۷. مدل جهانی کوچک

در سال ۱۹۹۸، Strogatz و Watts به معرفی مدلی جدید پرداختند (نیومن و وات[1]، ۲۰۱۰؛ باراباسی و آلبرت[2]، ۲۰۱۷). این مدل به صورت همزمان دارای خاصیت جهان کوچک و ضریب خوشگی بالا می‌باشد.

مراحل ساخت این مدل به این شرح است:

- ابتدا گره‌ها به‌صورت حلقه‌ای چیده می‌شوند و هر گره به k گره بعدی و k گره قبلی وصل می‌گردد.
- سپس با حرکت روی گره‌ها در یک جهت، مثلاً ساعتگرد، یال‌های بین گره‌ها یکی‌یکی با احتمال ثابت p اتصال دوباره می‌یابند.
- سر دیگر یال از رأس جدا می‌شود و به صورت تصادفی به یکی دیگر از گره‌های شبکه متصل می‌گردد.

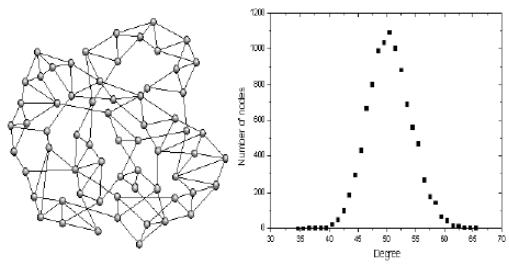
این روند در شکل (۴) نشان داده شده است. در حالت‌های نهایی، در صورتی که p = ۰ باشد، گراف حاصل حلقه و در صورتی‌که p =۱ گردد، گراف حاصل، گرافی تصادفی خواهد بود.



شکل ٤- مراحل ساخت گراف جهان کوچک (چیترادراگو و هلمی[3]، ۲۰۱٤)

1. Newman & Watts
2. Barabási & Albert
3. Chitradurga & Helmy

در شکل (۵) نمونه‌ای از مدل جهان کوچک و نمونه‌ای از توزیع درجه‌های گره‌ها در آن آمده است.



شکل ۵- گراف حاصل از مدل جهان کوچک، (a) مدل جهان کوچک، (b) توزیع درجات ۱۰ گراف با ۱۰۰۰ گره و p= ۰,۳ (هلمی[1]، ۲۰۱۸)

۳-۸. ویژگی‌های آماری یک شبکه پیچیده

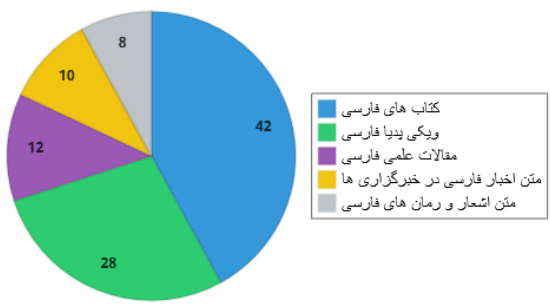
شبکه‌های پیچیده را می‌توان از طریق برخی ویژگی‌های آماری، از شبکه‌های تصادفی متمایز کرد که دارای خصوصیات مهمی نظیر اثر مدل جهانی کوچك، آزادی مقیاس و سلسله مراتب در شبکه است (فرتناتو[2]، ۲۰۱۸). شبکه‌های جهانی کوچک تمایل دارند از شش درجه قانون جدایی پیروی کنند که متوسط طول مسیر کوچک باید کمتر از شش باشد و در این بین ضریب خوشه‌بندی بالا، جهانی بودن کوچک یک شبکه را تایید می‌کند. فراوانی مثلث‌های متصل، به صورت آماری از طریق ضریب خوشه‌بندی در شبکه‌ها اندازه‌گیری می‌شود و مقدار ضریب خوشه‌بندی بین ۰ و ۱ است. در شبکه‌های پیچیده، مقدار ضریب خوشه‌بندی همیشه بالاتر از شبکه‌های تصادفی است. وجه تمایز مهم دیگر شبکه‌های پیچیده، وجود خوشه‌ها و لایه‌ها است. برخلاف شبکه‌های تصادفی[3]، شبکه‌های پیچیده به صورت سلسله مراتبی[4] و پارادایمیک[5] ساختار یافته‌اند. شبکه پیچیده زبان نه‌تنها منبع درک ساختار زبان است، بلکه روش‌های علمی قوی مدل‌سازی زبان را برای تغییر این جهان با ظهور چندزبانی ارائه می‌کند. پژوهش در زبان فارسی می‌تواند از نظر تئوری گراف حقایق ساختاری ریاضی کافی را در مورد زبان، همراه با بینش عمیق در روند تکامل و الگوهای ترکیب ارائه دهد. تبدیل زبان

1. Helmy
2. Fortunato
3. Random Networks
4. Hierarchical
5. Paradigmatic

فارسی به گراف شبکه، همزمان می‌تواند با کاوش در زبان فارسی به عنوان یک شبکه پیچیده به اهداف یادگیری عمیق در زبان فارسی نیز دست یابد. این کار بستری قوی برای انجام تحقیقات پیشرفته در پردازش زبان طبیعی فارسی فراهم می‌کند. اطلاعات به‌دست آمده را می‌توان در مدل‌سازی زبان فارسی، توسعه برنامه کاربردی مبتنی بر هوش مصنوعی، طبقه‌بندی کلمات، یادگیری ماشینی، تولید متن خودکار، مقایسه فارسی با زبان‌های دیگر برای بررسی شباهت‌ها و تفاوت‌ها، تشخیص نقش‌ها و مواردی از این قبیل مورد استفاده قرار داد (بلسن، گلیچ و لسکوك[1]، ۲۰۱۶؛ بافنا، پرامد و ویدیا[2]، ۲۰۱۸).

۴. مجموعه دادگان

در این پژوهش به یك مجموعه استاندارد باز و بدون حاشیه از زبان فارسی و مقدار زیادی متن در مورد موضوعات متعدد از زندگی روزمره در قالب استاندارد UTF-8 نیاز بود. بسیاری از کتاب‌های ادبیات فارسی، در مجموعه گنجانده شده است. نوع کتاب‌های انتخابی الکترونیکی و محل انتخاب نیز کتابخانه عمومی است و نحوه انتخاب کتاب نیز بر اساس این است که زبان آن‌ها دارای پیچیدگی زبانی خاصی نباشند. برای جمع‌آوری متن کافی برای اهداف پژوهشی، از وبلاگ‌های خبری محلی فارسی، متن، شعرها و رمان‌های فارسی استفاده شده است. مقاله‌های ویکی پدیای فارسی نیز بخشی از مجموعه داده‌های مورد استفاده هستند. برای این پژوهش ابتدا متن خام به زبان فارسی از منبع‌های متفاوت در اندازه‌های مختلف جمع‌آوری شد. ابتدا، اندازه کل متن ۱۱۹۶۴۷۲۳ کلمه بود. شکل (۶) مشارکت دقیق همه منبع‌های موجود در مجموعه داده‌ها را برحسب درصد نشان می‌دهد.



شکل ٦- مشارکت دقیق همه منبع‌های موجود در مجموعه داده‌ها

1. Benson, Gleich & Leskovec
2. Bafna, Pramod & Vaidya

اندازه مجموعه نهایی تمیز شده ۲۰۴۱۵۸۴ و اندازه واژگان ۷۴۲۳۹ بود. اندازه مجموعه به معنای تعداد کل کلمه‌ها می‌باشد، در حالی‌که اندازه واژگان نشان‌دهنده تعداد انواع کلمه‌های منحصربه‌فرد در متن تمیز است. اعداد، نمادها، علائم نگارشی، و علائم ریاضی در شبکه نهایی گنجانده نشده است.

۵. روش پیشنهادی و پیاده‌سازی

در این مرحله، ساخت شبکه متن مربوط به زبان فارسی با کمك منبع‌های متعدد انجام می‌گیرد. متن فارسی بدون در نظر گرفتن دستور زبان و معنی، به شبکه همزمان تبدیل می‌شود. برای ایجاد این شبکه نیاز به انجام پیش‌پردازش متن فارسی بوده که در ادامه به این موضوع پرداخته شده است.

۵-۱. پیش‌پردازش متن فارسی

برای پردازش زبان طبیعی و انجام عملیات خودکار بر روی متن مانند ترجمه، خلاصه‌سازی، تصحیح املاء، استخراج کلمات کلیدی، خوشه‌بندی، طبقه‌بندی و غیره، پژوهشگران نیازمند ابزارهایی جهت پیش‌پردازش و آماده‌سازی متن‌ها هستند. پیش‌پردازش داده‌ها، مهم‌ترین مرحله در فرایند کشف دانش از داده‌های متنی می‌باشد. پردازش متن به صورت خام امکان‌پذیر نیست و لازم است با انجام چند مرحله پیش‌پردازش، متن را برای انجام پردازش‌های لازم آماده کرد. پیچیدگی پیش‌پردازش داده‌ها، به منابع داده مورد استفاده بستگی دارد. اگر داده‌های وارد شده نویزدار و غیرقابل اطمینان باشند، کشف دانش از آن‌ها بسیار مشکل می‌شود. مراحل آماده‌سازی و فیلتر کردن داده‌ها، زمان قابل توجهی از زمان پردازش را به خود اختصاص خواهد داد. پیش‌پردازش داده‌ها شامل تمیز کردن، انتخاب نمونه، نرمال‌سازی، تبدیل، استخراج ویژگی‌ها، انتخاب و غیره است. خروجی به دست آمده پیش‌پردازش داده‌ها، یک مجموعه داده پالایش شده است که می‌تواند برای آموزش الگوریتم‌های متن‌کاوی استفاده شود. در این فرآیند حذف، کلمه‌های توقف[1] بر میزان قوانین استخراج شده تأثیر قابل توجهی دارند (بائور، هدرو و اسچیندر[2]، ۲۰۱۵). روش‌های پیش‌پردازش داده‌های متنی به دو صورت انجام می‌شود. دسته اول روش‌های وابسته به زبان هستند که براساس برخی قوانین نحوی و ساختاری زبان انجام می‌شوند. روش‌های دیگر، مستقل از زبان هستند و بیشتر بر مبنای پیکره‌های زبانی و با استفاده از روش‌های یادگیری ماشین صورت می‌گیرند که به تحلیل معنایی

1. Stop Word
2. Bauer, Hoedoro & Schneider

شهرت دارند. البته در برخی از موارد ترکیبی از هر دو روش مورد استفاده قرار می‌گیرد. از این‌رو طراحی و پیاده‌سازی این ابزارها برای زبان‌های مختلف، به روش‌های مختلف و مخصوص زبان مربوطه صورت می‌گیرد. این فرآیند شامل بخش‌های زیر می‌باشد (لوکاس و همکاران[1]، ۲۰۱۹).

۵-۱-۱. تشخیص زبان

تشخیص زبان اولین گام در پردازش زبان طبیعی می‌باشد. تشخیص زبان می‌تواند به عنوان یک تکنیک فیلترسازی در بازیابی اطلاعات برای کمک به کاربران علاقه‌مند به اطلاعات نوشته شده با یک زبان خاص به‌کار رود. علاوه‌بر این، از آنجا که بسیاری از تکنیک‌های پیش‌پردازش زبان نیازمند شناسایی زبان سند می‌باشند، تشخیص زبانی گام مهم پیش‌پردازش برای تکنیک‌های پردازش زبان دیگر مانند بن واژه‌یابی و با ترجمه ماشینی است. روش کار تشخیص زبان به این صورت است که ابتدا یک مدل برای سند و یك مدل برای هر یک از زبان‌های مورد نظر تهیه می‌شود و پس از آن با مقایسه مدل سند و مدل زبان‌ها، زبان سند تعیین می‌گردد. ابزارهای زیادی برای تشخیص زبان سندها ارائه شده‌اند که شامل Java Text Categorization Library، Apache Tika و J LangDetect هستند (ژنگ و همکاران[2]، ۲۰۱۷).

۵-۱-۲. جداسازی جملات

مرحله دیگری که بسته به نیاز در NLP انجام می‌شود، جداسازی جمله‌ها می‌باشد. یک جمله زمانی تمام می‌شود که یک کاراکتر پایانی جمله مانند نقطه یا علامت هجاوندی مشاهده شود. این ابزار با توجه به کاراکترهای جداکننده جمله، توانایی تشخیص جمله‌ها را در متن ورودی دارد. برای ایجاد این ابزار باید ابتدا تمامی کاراکترها، نمادها و احیاناً قواعد دستوری که باعث شکسته شدن جمله‌ها می‌شوند، شناسایی گردند. با توجه به پایه بودن جمله در بسیاری از پردازش‌های زبانی، خروجی دقیق این ابزار از درجه اهمیت بالایی برخوردار است و از نمونه‌های زبان انگلیسی آن می‌توان به OpenNLP، Stanford NLP، NLTK و Freeling اشاره کرد.

۵-۱-۳. هنجارساز

هنجارساز[3] متن نیز فرآیندی پردازشی است که متن را به یک حالت استاندارد تبدیل می‌کند و

1. Lucas
2. Zhang
3. Normalizer

شامل مراحل متعددی برای اصلاح اشتباه‌های سهوی کاربران و چندگانگی نوشتاری در زبان فارسی است. در هنجارسازی و استاندارد کردن یک متن باید به نوع متن، زبان، موضوع و پردازش‌هایی که پس از آن قرار است بر روی آن انجام شود، توجه داشت. هنجارسازی واحدهای متنی، به طوری که برای استفاده در پردازش‌های بعدی توسط ماشین قابل استفاده باشند، امری بدیهی و لازم است. هنجارساز متن شامل طبقه‌بندی نهادهای متنی مانند تاریخ، زمان، عدد، مبلغ ارز و غیره است. هنجارسازی عمدتاً شامل حذف علائم، نقطه‌گذاری، تبدیل کل متن به حروف کوچك یا بزرگ، تبدیل عددها به کلمه‌ها، گسترش اختصارها و غیره می‌باشد. البته برای هنجارسازی روشی جامع وجود ندارد. در این فرآیند مشکل‌هایی مانند وجود encodingهای مختلف برای بعضی از کاراکترها مانند حرف «ی» و«ك»، روش‌های مختلف چسبیدن حروف اضافه به کلمه‌های اصلی، روش‌های مختلف اتصال اجزای مرکب و کلمه‌های چند املایی مطرح هستند.

۵-۱-۴. واحدساز

تکه‌تکه کردن سند به قسمت‌های کوچک به نام واحد را واحدساز[1] گویند. برای شکستن یک متن براساس واحدهای با معنی مانند کلمه، پاراگراف، جمله و نمادهای معنادار، از واحدساز استفاده می‌شود. واحدسازی در سطح کلمه‌ها رخ می‌دهد و واحدهای استخراج شده می‌توانند به عنوان ورودی ماژول‌های دیگر مانند ریشه‌یاب و برچسب‌گذار استفاده شوند. عموماً بعد از این مرحله، حذف کلمه‌های توقف انجام می‌شود و متن براساس انتخاب هر کدام از این واحدها و با استفاده از tab یا Space شکسته خواهد شد.

۵-۱-۵. کلمه‌های توقف

پس از هنجارسازی متن، بایستی فهرست واژه‌ها را نیز برای برخی کاربردها حذف کنیم. کلمه‌های توقف تعدادی کلمه پر تکرار هستند که شامل عمومی‌ترین فعل‌ها، ضمیرها، قیدها، حرف‌های ربط و حرف‌های اضافه می‌باشند. کلمه‌های توقف مفهوم خاصی ندارند و از لحاظ معنایی با اهمیت نیستند، ولی در جمله‌ها و متن‌ها بسیار تکرار می‌شوند. «اگر»، «ولی»، «و»، «که» در زبان فارسی و to, for, about در زبان انگلیسی ازجمله کلمه‌های توقفی هستند که باید در مراحل پیش‌پردازش حذف شوند. در اغلب کاربردهای متن، حذف این کلمه‌ها، نتیجه‌های پردازش را به شدت بهبود می‌دهد و سبب کاهش بار محاسبه‌ها و افزایش سرعت پردازش خواهد شد. به همین

1. Tokenizer

دلیل این کلمه‌ها را در اغلب موارد در فاز پیش‌پردازش حذف می‌کنند. برای زبان فارسی چندین فهرست از این کلمه‌ها منتشر شده است که به طور میانگین شامل ۸۰۰ کلمه می‌باشند. برای حذف این کلمه‌ها عموماً فهرستی از این کلمه‌ها از پیش تهیه می‌شود و سپس در صورت رخداد این کلمه‌ها در متن، از سند حذف می‌شوند.

۵-۱-۶. ریشه‌یاب و بن واژه‌یاب

ریشه‌یابی و بن واژه‌یابی از اقدام‌هایی است که بنا به نیاز، در زمره تحلیل صرفی دسته‌بندی می‌شود. ریشه‌یابی اصطلاحی است که برای تعریف فرآیند کاهش دادن تعداد حرف‌های یك کلمه و رسیدن به ریشه آن به‌کار می‌رود. منظور از ریشه در این تعریف، ریشه زبانی نیست و هدف این است که فرمت‌های گوناگون یک کلمه دارای ریشه‌های یکسان باشند. معمولاً ریشه‌یابی لغت‌ها براساس قاعده‌های ساخت واژه‌ای و سپس حذف پسوندها می‌باشد و پس از واحدسازی متن انجام می‌شود. یکی از اقدام‌های مؤثر برای استخراج ریشه کلمه، حذف پیشوندها است. برای حذف پسوندها از آنالیزهای آماری و داده‌کاوی استفاده می‌شود، که این روش هم می‌تواند راهی برای تشخیص ریشه باشد. این روش برای ریشه‌یابی لغت‌ها و تشخیص نوع کلمه ساخته شده از آن ریشه استفاده می‌شود. معروف‌ترین الگوریتم ریشه‌یابی در انگلیسی porter است.

در فرهنگ‌های لغت، بن واژه[1]، معنایی انتزاعی از صورت‌های مختلف یک واژه است. بن واژه‌یابی نیز تحت عنوان تقلیل و ساده‌سازی صورت‌های گوناگون صرفی یك واژه و گزینش یک صورت پایه از کلمه‌های تعریف شده است. بن‌واژه معنایی انتزاعی از صورت‌های مختلف یک واژه است. در زبان‌شناسی محاسباتی، بن‌واژه‌یابی کردن، فرایندی الگوریتمی است که یک کلمه براساس معنای آن تحلیل می‌شود. برخلاف ریشه‌یابی، بن‌واژه‌یابی، شناسایی صحیح بخشی از گفتار و معنای یک کلمه در یک جمله و همچنین در محدوده اطراف آن جمله، جمله‌های همسایه یا حتی کل سند می‌باشد. در بسیاری از زبان‌ها، کلمه‌ها در چندین شکل نحوی ظاهر می‌شوند. بنابراین، هدف هر دو ابزار بن‌واژه‌یابی و ریشه‌یابی کاهش شکل‌های فراوان و بعضی از شکل‌های متداول مشتق شده از یک کلمه به یک شکل پایه مشترک، و اغلب شامل حذف عبارت‌های مشتق شده است.

۵-۱-۷. تشخیص موجودیت‌های اسمی

تشخیص موجودیت‌های اسمی فرآیندی است که هدف از آن تشخیص و شناسایی کلمه‌ها یا

1. Lemma

عبارت‌هایی است که نمایانگر یک موجودیت می‌باشند. این فرآیند برای تشخیص اسامی و نوع آن‌ها از قبیل نام افراد، سازمان‌ها، مکان‌ها و غیره به‌کار می‌رود. تشخیص موجودیت‌های اسمی به طور خاص می‌تواند در حل مسئله‌هایی مانند رفع ابهام و تشخیص هویت اصلی اشخاص با اسامی مشترک، از روی موضوع متن و با کمک ابزارهای جانبی، پژوهشگران را یاری دهد.

برای تشخیص اینکه یک کلمه اسم است، راه‌های مختلفی وجود دارد، می‌توان به لغت‌نامه، استفاده از شبکه واژگان، در نظر گرفتن ریشه کلمه، استفاده از قواعد نحوی ساخت واژه و مواردی از این قبیل اشاره کرد. در این فرآیند پس از تشخیص اسم‌ها با استفاده از یک لغت‌نامه، اسامی افراد، مکان‌ها، مقادیر عددی و موارد مشابه، نوع اسم تشخیص داده می‌شود. تشخیص درست واحدهای اسمی، یک نیاز مهم در حل مسائلی مانند پاسخگویی به سؤالات، سیستم‌های خلاصه‌سازی، بازیابی اطلاعات، استخراج اطلاعات، ترجمه ماشینی، تفسیر ویدیوئی و جستجوی معنایی در وب است و از نمونه‌های زبان انگلیسی این ابزار می‌توان به Ilinois NER, Stanford NER اشاره کرد.

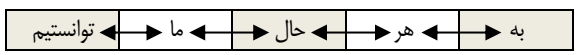
۵-۱-۸. استخراج کلمات کلیدی

استخراج کلمات کلیدی، فرآیند شناسایی خودکار اصطلاحات به‌کار رفته در یک سند است. یکی از عملیات‌های مهم در فرآیندهای خوشه‌بندی، طبقه‌بندی، استخراج اطلاعات و مشخص کردن موضوع مورد بحث در یک سند، استخراج کلمات کلیدی متن است. عبارات کلیدی، اصطلاحات کلیدی و کلمات کلیدی عبارتند از اصطلاحاتی که بر شرایطی که بیشترین اطلاعات مرتبط در سند را ارائه می‌دهند. اگرچه این اصطلاحات با هم متفاوت هستند، ولی عملکرد آن‌ها یکسان است. با یافتن کلمات کلیدی می‌توان راحت‌تر و در زمانی کوتاه‌تر به مفهوم یک متن، خبر یا مقاله پی برد. برای انتخاب کلمات کاندید به عنوان کلمات کلیدی، بایستی تمام کلمات، عبارات، اصطلاحات و مفاهیمی که به طور بالقوه کلمات کلیدی باشند را استخراج کرد. سپس با استفاده از تکنیک‌های پردازش متن و یادگیری ماشین، خواص هر کاندید محاسبه و یک نمره یا آستانه احتمالی به آن اختصاص می‌یابد. سپس تمام کاندیداها را می‌توان به وسیله ترکیب خاص، برای انتخاب مجموعه نهایی کلمات کلیدی یک سند ارزیابی کرد. به طور کلی سه روش متداول برای استخراج کلمات کلیدی وجود دارد:

- روش TF-IDF
- روش مبتنی بر یادگیری ماشین
- ترکیب روش‌های تحلیل آماری و زبانشناختی.

استخراج کلمات کلیدی غالباً با استفاده از تکنیک‌های یادگیری ماشین، دارای نتایج بهتری است. لیکن استفاده از روش TF-IDF به دلیل سهولت و استفاده از منابع سیستمی، کمتر متداول می‌باشد.

با توجه به ۷ مرحله ارائه شده پس از تمیز کردن مناسب متن، کلمه‌های مناسب به عنوان گره انتخاب می‌شوند. رویکرد جنگل تصادفی، متن خام را به کلمات صحیح، کلمات شکسته نادرست قابل حذف و نمادهای قابل حذف، طبقه‌بندی می‌کند (جولین و همکاران[1]، ۲۰۱۶؛ یانگ و همکاران[2]، ۲۰۲۱). روش TF-IDF برای تولید فهرست کلمه‌های توقف استفاده می‌شود. کلمه‌های توقف، کلمه‌های کم‌معنی‌تری هستند که اغلب در یک زبان ظاهر می‌شوند، اما نمی‌توانند به‌عنوان کلیدواژه برای جستجوی داده‌ها استفاده شوند. لیست کلمه‌های توقف به صورت دستی از فهرست کلمه‌ها با فرکانس بالا ایجاد می‌شود. در ادامه، فهرست کلمه‌های توقف به عنوان مجموعه آموزشی، برای حذف کلمه‌های توقف در الگوریتم یادگیری نظارت شده[3] استفاده شده است (سینر، کوکل و تاسدمیر[4]، ۲۰۲۲). هر کلمه منحصربه‌فرد بدون توجه به شکل آن، به عنوان یک گره در نظر گرفته می‌شود. کلمه‌های فارسی برای ساخت شبکه همزمان، از طریق مدل بای گرام[5] (ویجی مینا و ۲۰۱۶) مطابق شکل (۷) به هم متصل می‌شوند.



شکل۷- کلمات فارسی برای ساخت شبکه همزمان از طریق مدل بای گرام

نمونه‌ای از فرآیند تولید بای گرام از کلمات همزمان در جدول (۲) آمده است.

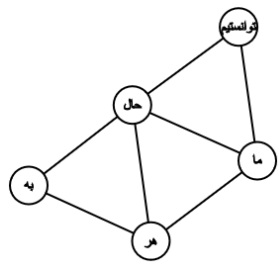
جدول ۲- نمونه‌ای از فرآیند تولید بای گرام از کلمه‌های همزمان

Bigrams	Word Indices	
به هر	۲	۱
به حال	۳	۱
هر حال	۳	۲
هر ما	۴	۲

1. Joulin
2. Yang
3. Supervised Learning Algorithm
4. Cinar, Koklu & Tasdemir
5. Bigrams
6. Vijaymeena & Kavitha

Bigrams		Word Indices	
حال	ما	۴	۳
حال	توانستیم	۵	۳
ما	توانستیم	۵	۴

هر کلمه در شبکه به عنوان یک گره در نظر گرفته می‌شود و پیوندهایی بین گره‌ها ایجاد می‌گردد. بسته به پنجره، هر کلمه به تعداد کلماتی که در کنار آن در متن ظاهر می‌شود، پیوند داده خواهد شد. در ایجاد گراف مربوط به متن، از حلقه‌های طوقه[1] اجتناب می‌شود، بنابراین، یک کلمه نمی‌تواند به خودش متصل شود. همانطور که در شکل (۸) نشان داده شده است، شبکه پس از اتصال گره‌های جفت شده، به گراف کوچکی از پنج گره و هفت پیوند تبدیل شد.



شکل ۸- ایجاد شبکه پس از اتصال گره‌های جفت شده

برای ایجاد گراف مربوط به متن، هر کلمه صحیح به عنوان یک گره در نظر گرفته می‌شود که کلمه‌های توقف به عنوان گره برای جلوگیری از نویز در شبکه گنجانده نشده است. همه هموگراف‌ها یعنی کلمه‌هایی که مجموعه کاراکترهای یکسان دارند، در گره‌های منفرد ادغام می‌شوند و فرم‌های چند کلمه‌ای به عنوان گره‌های غیریکسان معرفی شده‌اند. بنابراین، یکی از مشکل‌هایی که در ریشه‌یابی کلمات فارسی وجود دارد، این است که همه کلمه‌ها بدون توجه به صورت و بار معنایی آن‌ها، برای گره‌ها در نظر گرفته می‌شوند. تمامی پردازش‌ها و پیش‌پردازش‌ها در این پژوهش با پایتون Python 3.10.6 انجام شده است.

۶. ارزیابی

برای تجزیه و تحلیل ویژگی‌های ساختاری بصری‌سازی[2]، گراف حاصل به Gephi 0.9.7 فرستاده شده است. شبکه تصادفی با اندازه مشابه و تعداد مساوی گره و پیوند، توسط گراف تصادفی

1. Self
2. Visualization

Erdos-Renyi تولید می‌شود. این شبکه تصادفی به عنوان حالت تهی استفاده شد (فونترا[1]، ۲۰۲۱). برای احراز هویت برخی از ویژگی‌های مهم محلی و جهانی شبکه پیچیده در شبکه زبان فارسی[2]، با مدل Null شبکه تصادفی[3]، مقایسه‌هایی انجام گرفته است که ابزار تجسّم گراف به طرز شگفت‌انگیزی برای تجسّم شبکه متن فارسی و ویژگی‌های استخراج هر دو شبکه پشتیبانی می‌شود. نتایج همه گره‌ها به شدت با پیوندهایی در شبکه زبان فارسی مرتبط هستند و پیوند از یک کلمه به کلمه دیگر، به این معنی است که دو کلمه همزمان در متن وجود دارند.

جدول ۳- مقایسه شبکه تصادفی و شبکه زبان فارسی

Properties	شبکه زبان فارسی (FLN)	شبکه تصادفی (NRN)
Number of nodes	۳۲۵۶	۳۲۵۶
Number of links	۷۹۷۰۵	۷۹۷۰۵
Diameter	۷	۳
Radius	۴	۴
Average path length	۳/۸۴۵۱۹۵۶	۳/۹۹۳۲۵۵۱
Clustering coefficient	۰/۵۶۹	۰/۰۰۹
Communities	۱۲	۱

پس از انجام محاسبات در هر دو شبکه، مشاهده‌ها نشان می‌دهد که میانگین مسیر شبکه تصادفی ۳.۹ است، در حالی‌که میانگین طول مسیر شبکه زبان فارسی ۳.۸ است. به طور مشابه، متوسط ضریب خوشه‌بندی شبکه تصادفی ۰.۰۰۹ و برای شبکه زبان فارسی مقدار آن ۰.۵۶۹ است. همچنین، حداکثر تعداد جوامع شناسایی شده در شبکه زبان فارسی ۱۲ بوده و هیچ انجمن و لایه‌ای در شبکه تصادفی شناسایی نشده است. ازجمله مشاهده‌های مهم می‌توان به این موضوع اشاره کرد که:

- فارسی قبلاً به عنوان یک زبان سفارش آزاد اعلام شده بود.
- کلمه‌ها در زبان فارسی می‌توانند به روش‌های مختلف با هم وجود داشته باشند.
- در بیشتر موارد، ترتیب کلمه‌ها بر معنا یا زمینه تأثیر نمی‌گذارد.

بنابراین می‌توان اینگونه بیان نمود که ترتیب سفارش رایگان، فقط در جملات کوتاه یافت می‌شود و اکثر جمله‌های کوتاهی که از سه، چهار یا پنج کلمه تشکیل شده‌اند، تا حدی بدون ترتیب ظاهر

1. Fontoura
2. Farsi Language Network (FLN)
3. Null model of Random Network (NRN)

می‌شوند و پیام کوتاه را می‌توان به درستی در زبان فارسی با کمک چند کلمه خودانگیخته، بدون ترتیب کلمات خاص منتقل کرد. این مشاهدات برای توسعه مترجم‌ها و ابزارهای تولید خودکار متن، با ویژگی‌های قابل اعتماد پشتیبانی از زبان فارسی بسیار مفید خواهد بود. اما تحقیقات عمیق بیشتری برای یافتن ویژگی دقیق زبان فارسی با مقایسه شبکه‌های هم‌وقوع جهت‌دار و وزن‌دار متن نرمال‌شده و درهم‌آمیخته مورد نیاز است. بنابراین، تجسّم مثلث سه‌گانه و N گرام در شبکه می‌تواند برای تولید جمله‌های کوتاه مفید باشد که در پژوهش‌های بعدی می‌توان این ویژگی را برای تأثیر ترتیب کلمه‌ها در ساختار فارسی مورد بررسی قرار داد.

۷. نتیجه‌گیری

در مقایسه ویژگی‌های کلیدی کشف شده از هر دو شبکه، شبکه فارسی با شبکه تصادفی متفاوت است که به وضوح اثر جهانی کوچک و آزاد بودن مقیاس را نشان می‌دهد. میانگین طول مسیر شبکه زبان فارسی، کوچک‌تر از طول مسیر متوسط شبکه تصادفی است. ضریب خوشه‌بندی شبکه زبان فارسی در مقایسه با شبکه تصادفی بسیار بالا است. اتصال گره‌ها در شبکه تصادفی یکنواخت بوده و هیچ‌گونه خوشه‌بندی یافت نمی‌شود. در سیستم فارسی حداکثر ۱۲ اجتماع شناسایی شده و براساس این تحلیل، زبان فارسی یک شبکه پیچیده معرفی می‌شود. بنابراین، می‌توان آن را به عنوان یک شبکه پیچیده با کاربرد تکنیک‌ها و فناوری‌های پیچیده شبکه مورد مطالعه قرار داد. تحقیقات پیش‌رو در شبکه‌های پیچیده فارسی برای یادگیری عمیق، کاوش ساختار، تشخیص ویژگی‌ها، مدل‌سازی زبان، طبقه‌بندی متن، طبقه‌بندی زبان و موارد دیگر در پردازش زبان طبیعی، زبان‌شناسی، یادگیری ماشین و هوش مصنوعی مفید خواهند بود.

منـابـع

سجادی، م.ب.، مینایی بیدگلی، ب. (۱۳۹۸). معماری سامانهٔ گراف دانش زبان فارسی. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۵(۲).

References

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1): 47–49. https://doi.org/10.1103/RevModPhys.74.47

Bafna, P., Pramod, D. & Vaidya, A. (2018). Document clustering: TF-IDF approach. In: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). *Neurocomputing*, 300: 70-79.

Barabasi, A. & Bonabeau, E. (2003). Scale-Free Networks. *Scientific American*, 288(5): 50–59. https://doi.org/10.1038/scientificamerican0503-60. PMID: 12701331

Barabási, A.L. & Albert, R. (2017). Emergence of scaling in random networks. *Science*, 286(15): 509-512.

Baran-Gale, J. & et al. (2020). Ageing compromises mouse thymus function and remodels epithelial cell differentiation. *eLife*, 9: e56221

Bassett, D.S. & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, 20(3): 353–364. https://doi.org/10.1038/nn.4502

Bauer, A., Hoedoro, N. & Schneider, A. (2015). *Rule-based Approach to Text Generation in Natural Language-Automated Text Markup Language (ATML3).* In: Challenge+DC@RuleML.

Benson, A.R., Gleich, D.F. & Leskovec, J. (2016). Higher-order organization of complex networks. *Science*, 353(6295): 163-166.

Breuer, A., Elflein, S., Joseph, T., Termöhlen, J., Homoceanu, S. & Fingscheidt, T. (2019). *Analysis of the effect of various input representations for LSTM-based trajectory prediction*. IEEE Intell Transp Syst Conf (ITSC): 2728–2735.

Cancho, R.F.I. & Solé, R.V. (2001). The small world of human language. Proceedings of the Royal Society of London. *Series B: Biological Sciences*, 268(1482): 2261-2265.

Chen, G. & Lou, Y. (2019). *Multi-Language Naming Game*. In: Naming Game. Springer: 135-154.

Chen, H., Chen, X. & Liu, H. (2018). How does language change as a lexical network? An investigation based on written Chinese word co-occurrence networks. *PloS one*, 13(2): e0192545.

Chitradurga, R. & Helmy, A. (2014). *Analysis of wired short cuts in wireless sensor networks*. IEEE/ACS International Conference on, Pervasive Services: 167-176.

Cinar, I., Koklu, M. & Tasdemir, S. (2020). *Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods.* https://doi.org/10.30855/gmbd.2020.03.03.

da Fontoura Costa, L. (2021). *A caleidoscope of datasets represented as networks by the coincidence methodology*. URL= https://researchgate.net/publication/356392287_A_Caleidoscope_of_Datasets_Represented_as_Networks_by_the_ Coincidence Methodology.

Fornito, A. (2020). *An Introduction to Network Neuroscience: How to build, model, and analyse connectomes* - 0800-10:00 | OHBM". URL= https://www.pathlms.com/ohbm/courses/12238/sections/15846/video_presentations/13753

Sajjadi, M.B. & Minaei Bidgoli, B. (2018). Persian language knowledge graph system architecture. *Journal of Information Processing and Management*, 35(2). [in persian]

Daud, A., Khan, W. & Che, D. (2017). Urdu language processing: a survey. *Artificial Intelligence Review*, 47(3): 279-311.

Fortunato, S. (2018). *Community structure in complex networks*. in EGC.

Fromkin, V., Rodman, R. & Hyam, N. (2018). *An introduction to language Cengage Learning.* Michael Rosenberg.

Gao, Z.-K., Small, M. & Kurths, J. (2017). Complex network analysis of time series. *EPL*, 116(5): 50001.

Garnham, A. (2017). *Artificial intelligence an introduction*. Routledge.

Goh, W.P., Luke, K.-K. & Cheong, S.A. (2018). Functional shortcuts in language co-occurrence networks. *PloS one*, 13(9): e0203025.

Helmy, A. (2018). Small worlds in wireless networks. *IEEE Commun Lett*, 7(10): 490-492.

Howard., J. & Ruder, S. (2018). *Universal language model fine-tuning for text classification*. In: Annual Meeting of the Association for Computational Linguistics: 328–339.

Joulin, A. & et al. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv*, 1607.01759.

Khan, N., Bakht, M.P. &d Waga, R.A. (2019). *Corpus Construction and Structure Study of Urdu Language using Empirical Laws.* Urdu News Headline, Text Classification by Using Different Machine Learning Algorithms.

Kiselev, V.Y., Andrews, T.S. & Hemberg, M. (2019). *Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet.*, 20: 273–282.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). *Deep learning.* Nature, 521(7553): 436. https://doi.org/10.1038/nature14539

Lin, C., King, J., Bharadwaj, P., Chen, C., Gupta, A., Ding, W. & Prasad, M. (2019). EOG-based eye movement classification and application on HCI baseball game. *IEEE Access*, 7: 96166–96176.

Lucas, J., Tucker, G., Grosse, R. & Norouzi, M. (2019). *Understanding posterior collapse in generative latent variable models.* URL= https://openreview.net/pdf?id=r1xaVLUYuE

Newma, M. (2010). *Networks: An Introduction*. Oxford University Press.

Newman, M.E.J. & Watts, D.J. (2010). Renormalization group analysis of the small-world network model. *Physics Letter A*, vol. 263: 341-346.

Paul, G., Cao, F., Huang, Q.T., Wang, H.S., Gu, Q., Zhang, K., Shao, M. & Li., Y. (2018). An EOG-based human-machine interface for wheelchair control. *IEEE Trans Biomed Eng*, 65: 2023–2032.

Piryonesi, S.M., & Tamer, E.D. (2020). Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation Engineering, Part B Pavements*,146(2).

Robert, C. (2014). *Machine learning, a probabilistic perspective.* Taylor & Francis.

Russell, S.J. & Norvig, P. (2016). *Artificial intelligence: a modern approach.* Malaysia: Pearson Education Limited.

Saberi, M., Khosrowabadi, R., Khatibi, A., Misic, B. & Jafari, G. (2021). Topological impact of negative links on the stability of resting-state brain network. *Scientific Reports*, 11(1): 2176. https://doi.org/10.1038/s41598-021-81767-7

Siegel, J.S. & et al. (2018). Re-emergence of modular brain networks in stroke recovery. *Cortex*, 101: 44-59.

Stanley, H.E., Amaral, L.A.N., Scala, A. & Barthelemy, M. (2000). Classes of small-world networks. *PNAS*, 97(21): 11149–52. https://doi.org/10.1073/pnas.200327197.

Strogatz, S. & Watts, D.J. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684): 440–442. https://doi.org/10.1038/30918.

Sun, C. Qiu, X., Xu, Y. & Huang, X. (2019). *How to fine-tune BERT for text classification?* in: China National Conference on Chinese Computational Linguistics: 194–206.

Vijaymeena., M.K. & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(1): 19–28.

Wilhelm, T. & Kim, J. (2008). *What is a complex graph? Physica A: Statistical Mechanics and its Applications*, 387(11): 2637–2652. https://doi.org/10.1016/j.physa.2008.01.015.

Xie, Q., Dai, Z., Hovy, E., Luong, M.T. & Le, Q.V. (2020). *Unsupervised data augmentation for consistency training*. in: Annual Conference on Neural Information Processing Systems.

Yang, H., Cheng, J., Yang, Z., Zhang, H., Zhang, W., Yang, K. & Chen, X. (2021). A node similarity and community link strength-based community discovery algorithm Complexity. *Complexity*, 22:1-17. https://doi.org/10.1155/2021/8848566

Yule, C.U. (2014). *The statistical study of literary vocabulary*. Cambridge University Press.

Zhang, B., Zhou, W., CaiH, S., Wang, J., Zhang, Z. & Lei, T. (2020). *Ubiquitous depression detection of sleep physiological data by using combination learning and functional networks.* IEEE. https://doi.org/10.1109/ACCESS.2020.2994985

Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D. & et al. (2017). *Adversarial feature matching for text generation*. URL= https://arxiv.org/pdf/1706.03850.pdf