Language and Linguistics
Spring & Summer 2023, 18 (35)
DOI: 10.30465/LSI.2023.8496

Original Research
Received: 2022-10-25
Accepted: 2023 - 3 – 26
ISSN: 23223847

Evaluating the Structure of the Inverted Pyramid in the Big Persian News Corpus: News Discourse Analysis based on the Correlation Coefficient between Title and News Content

Ghayoomi, Masood[1]

Assistant Professor of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran

Abstract

News discourse is a type of discourse analysis that deals with the analysis of news discourse. Due to the fact that in the formatting of news there are two hidden features of selection and prominence in the communication representation of news, the inverted pyramid of news is used to grade the importance of the discourse parts of the news. Although it is desirable to meet the structure of the inverted pyramid of news, sometimes this structure may change. In this article, we put an effort to analyse the discourse analysis of Persian news websites with the help of statistical analysis. To research the goal, data science can be used. This inter-discipline deals with data analysis from a scientific aspect, finding implicit concepts to be obtained from data analysis and extracting knowledge from the data. In the framework of data science, we examined the Persian news corpus and studied the existence of semantic correlation between the news title and the news content based on the structure of the news inverted pyramid. To achieve the goal, by using the crawling method, a relatively large news corpus with a volume of 14 billion words has been obtained from 24 news websites. After pre-processing and normalizing the corpus, in the framework of distributional semantics, the vector of title news and content have been created by using the Word2Vec tool for creating the vector model to have the vector representation of each news. After segmenting news content into three parts (lead, body and further explanation about the lead) according to the inverted pyramid, the Pearson correlation coefficient has been used to calculate the correlation between the title and each part of the news. Although Pearson's correlation coefficient was positive for a large number of news, zero value and no correlation was found for the news. On average, the correlation between the headline and the news lead and body was higher than the correlation between the headline and the lead development. This research can be used as a method to carefully select the title and content and filter the news according to the inverted pyramid structure.

Keywords: news corpus, Pearson correlation coefficient, distributional semantics, Word2Wok, news inverted pyramid

1. m.ghayoomi@ihcs.ac.ir

How to cite: Ghayoomi, M. (2023). Evaluating the Structure of the Inverted Pyramid in the Big Persian News Corpus: News Discourse Analysis based on the Correlation Coefficient between Title and News Content. *Language and Linguistics*, *18*(35), 21-45. doi: 10.30465/lsi.2023.8497

زبان و زبان‌شناسی
دوره ۱۸، شماره ۳۵، بهار و تابستان، سال ۱۴۰۱، ۲۱- ۴۵
DOI: 10.30465/LSI.2023.8497

مقاله پژوهشی
تاریخ دریافت: ۱۴۰۱/۰۸/۳۰؛ تاریخ پذیرش: ۱۴۰۲/۰۱/۰۶؛ تاریخ انتشار: بهار ۱۴۰۲
ISSN: 23223847

ارزیابی ساختار هرم وارونه در پیکرهٔ بزرگ خبری فارسی: تحلیل گفتمان خبری براساس همبستگی میان عنوان و محتوای خبر

قیومی، مسعود استادیار پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران

چکیده: گفتمان خبری گونه‌ای تحلیل گفتمان است که به تحلیل ساختار گفتمان خبری می‌پردازد. باتوجه‌به این که در قالب‌بندی اخبار دو ویژگی انتخاب و برجستگی در نمود ارتباطی خبر نهفته است، از ساختار هرم وارونهٔ خبر برای درجه‌بندی اهمیت بخش‌های گفتمانی خبر استفاده می‌شود. اگرچه رعایت ساختار هرم وارونهٔ خبر مطلوب است، گاهی ممکن است در گفتمان خبری این ساختار دچار تغییر شود که در این مقاله تلاش می‌شود با کمک تحلیل آماری، به تحلیل ساختار گفتمان وبگاه‌های خبری فارسی پرداخته شود. برای رسیدن به هدف می‌توان از علم داده استفاده کرد. این بین‌رشته‌ای از جنبهٔ علمی به تحلیل داده، یافتن مفاهیم ضمنیِ به‌دست‌آمده از تحلیل داده‌ها و استخراج دانش از داده‌ها می‌پردازد. در چارچوب علم داده به بررسی پیکره‌ای متون خبری فارسی پرداخته شده و وجود رابطهٔ همبستگی معنایی میان عنوان خبر و محتوای خبر در ساختار هرم وارونهٔ خبر مورد مطالعه قرار می‌گیرد. برای دستیابی به این هدف، با استفاده از روش خزش، یک پیکرهٔ خبری نسبتاً بزرگ با حجمی بالغ بر ۱٤میلیارد واژه از ۲٤ وبگاه خبری به‌دست آمده‌است. پس‌از پیش‌پردازش و اعمال یکدستی نسبی در این پیکره، در چارچوب معناشناسی توزیعی، بردار عنوان خبر و متن خبر با استفاده از مدل بردارسازی واژه وردt۲وک به‌دست آمده و براساس آن بردار هر خبر ساخته شده‌است. پس‌از بخش‌بندی محتوای هر خبر براساس هرم وارونهٔ خبر به سه قسمت سرنخ (لید)، بدنه و توسعهٔ سرنخ، با استفاده از

ضریب همبستگی پیرسون، میزان همبستگی میان عنوان و هریک از سه بخش خبر محاسبه شد. اگرچه ضریب همبستگی پیرسون برای حجم زیادی از خبرها مثبت بود، ارزش صفر و عدم وجود همبستگی برای خبرها یافت شد. به‌طور متوسط، همبستگی میان عنوان و بدنهٔ خبر بیش‌از همبستگی میان عنوان و توسعهٔ سرنخ بود. این پژوهش می‌تواند به‌عنوان روشی برای دقت در انتخاب عنوان و محتوا و پالایش خبر منطبق‌بر هرم وارونه استفاده شود.

کلیدواژه‌ها: پیکرهٔ خبری، ضریب همبستگی پیرسون، معناشناسی توزیعی، ورد۲وک، هرم وارونهٔ خبر.

۱ مقدمه

تحلیل گفتمان یکی از حوزه‌های زبان‌شناسی نظری است که علاوه‌بر نقش تعاملی خود در جامعه برای انتقال اطلاعات و توصیف روابط اجتماعی و رویکردهای شخصی، دارای نقش ارجاعی در توصیف محتوا و تحلیل کاربردی زبان نوشتاری و گفتاری یا نشانه‌ها در زبان است (براون[1] و یول[2]، ۱۹۸۳: ۱-۲). اگرچه هدف اولیهٔ زبان برای اهداف تعاملی است، می‌توان از آن برای اهداف ارجاعی و روابط اجتماعی نیز بهره برد. براون و یول (۱۹۸۳: ۶ و ۱۹۰) مفاهیم خاصی را برای متن تعریف کرده‌اند. بنا بر نظر آنان، هرگونه ثبت شفاهی عملکرد ارتباطی زبان، متن است؛ و هچنین متن چیزی جز پیوستگی معنایی میان جملات تشکیل‌دهندهٔ آن نیست. پر واضح است الزاماً همیشه عملکرد ارتباطی زبان به‌صورت شفاهی نیست و گونهٔ مکتوب، مانند متن خبری، کتاب و مقاله نیز وجود دارد تا ارتباط زبانی میان تولیدکنندهٔ محتوا و دریافت‌کنندهٔ محتوا حاصل شود. در نوع مکتوب متن که ازطریق خط میسر می‌شود، از زبان در نقش تعاملی خود برای ارتباط فرد با رویدادهای محیط پیرامون استفاده می‌شود. دوبوگراند[3] و درسلر[4] (۱۹۸۱: ۳) متن را نوعی نمود ارتباطی دانسته و هفت ویژگی را برای آن مطرح کرده‌اند. متن خبری کم‌وبیش حاوی ویژگی‌های ذکرشده است که مطالعهٔ در این حوزه در چارچوب تحلیل گفتمان، با اصطلاح «گفتمان خبری»[5] (بدنارک[6] و کپل[7]، ۲۰۱۲) شناخته می‌شود.

1. G. Brown
2. G. Yule
3. R. de Beaugrand
4. W. U. Dressler
5. News discourse
6. M. Bednarek
7. H. Caple

باتوجه‌به نقش تعاملی زبان در خبر، باید ارتباط متنی با مخاطب دارای ویژگی تأثیرگذاری باشد (دوبوگراند و درسلر، ۱۹۸۱: ۳۱). برای تحقق این ویژگی نیاز است عناوین خبری به‌گونه‌ای برای مخاطب جذاب باشد؛ به عبارتی دیگر، زبان خبر باید به‌گونه‌ای طراحی شود که برای مخاطب جاذبه داشته باشد که بل[1] (۱۹۹۱: ۱۲۱) از آن با اصطلاح «طراحی مخاطب»[2] یاد می‌کند.

برای بررسی ویژگی تعاملی زبان و رعایت تأثیرگذاری بر مخاطب ازطریق ایجاد جذابیت در عنوان خبر و ارتباط عنوان با محتوای خبر، در این پژوهش که اساساً ماهیت بین‌رشته‌ای دارد می‌کوشیم با کمک تحلیل آماری در چارچوب تحلیل گفتمان خبری و «علم داده»[3]، به بررسی همبستگی عنوان و بخش‌های تشکیل‌دهندهٔ خبر بپردازیم تا بتوان به این پرسش‌ها پاسخ داد که آیا به‌هنگام تهیه و انتشار خبر نکاتی که به‌صورت نظری مطرح است رعایت می‌شود؟ و برای جذب مخاطب، عناوین خبر متبلورکنندهٔ کدام بخش از خبر است؟

ساختار مقالهٔ حاضر به این صورت است که پس از مقدمه، در بخش ۲ مطالعات انجام‌شده در حوزهٔ تحلیل ساختار خبر مرور می‌شود. در بخش ۳، چارچوب نظری چندبعدی در حوزهٔ گفتمان خبری، معرفی هرم وارونهٔ خبر و همچنین علم داده ارائه می‌شود. شیوهٔ گردآوری داده‌ها، تهیهٔ پیکرهٔ زبانی و شیوهٔ انجام پژوهش در بخش ۴ توضیح داده می‌شود. در بخش ۵، به تحلیل داده‌ها و پاسخ به پرسش‌ها پرداخته شده و در نهایت، در بخش ۶، مقاله با نتیجه‌گیری خاتمه می‌یابد.

۲ پیشینهٔ مطالعاتی

ازآنجاکه تحلیل گفتمان خبری در چارچوب هرم وارونهٔ خبر و همبستگی معنایی در خبر محوریت موضوع این پژوهش را شکل می‌دهد، مطالعات انجام‌شدهٔ با موضوعات متمرکز بر تحلیل گفتمان با استفاده از هرم خبر و همبستگی محتوایی بررسی می‌شود.

ژانگ[4] و لیو[5] (۲۰۱۶) در پژوهش خود تلاش کرده‌اند که با استفاده از قیاس بین گفتمان و درخت تجزیه وابستگی جملات به طراحی چکیدهٔ متن بپردازند. برای این هدف، ۳۵۹

1. A. Bell
2. audience design
3. data science
4. H. Zhang
5. H. Liu

مقاله مجلهٔ وال استریت[1] که حاوی دادگان درختی گفتمان مبتنی‌بر نظریهٔ ساختار بلاغی شکل گرفته‌است را انتخاب کرده و هر درخت گفتمان را به سه درخت وابستگی در سطوح گفتمان، پاراگراف و جمله تبدیل کرده‌اند. آنها به‌طور تجربی ساختار «خلاصه+ جزئیات» یا «هرم وارونه»[2] گفتمان خبری را براساس داده‌های موجود مطالعه کرده‌اند.

یتریبرگ[3] (۲۰۱۰) به مقایسهٔ گزارش‌های خبری تلویزیونی اروپا و اسکاندیناوی در دهه‌های ۱۹۸۰ و ۱۹۹۰ با ساختار معیار هرم وارونه خبر پرداخته‌است. براساس یافته‌ها، در مقالات خبری تغییری صورت گرفته‌است به این صورت که از عینیت‌گرایی «پارادایم خبری» به سمت آرمان‌گرایی و «داستان‌سرایی» تغییر جهت صورت گرفته‌است و از ساختار هرم وارونه خبر خارج شده‌است.

امده[4]، کلیمت[5] و اشلوتس[6] (۲۰۱۶) درک خبر توسط نوجوانان ۱۲ تا ۱۷ ساله را مورد مطالعه قرار داده‌اند. در این پژوهش شیوهٔ سنتی پخش اخبار در رسانه که مبتنی‌بر ساختار هرم وارونهٔ خبر است را با شیوهٔ روایی خبر مقایسه کرده‌اند. براساس نتایج این پژوهش، به‌طورکلی خبر روایی نتوانسته‌است سبب افزایش درک خبر در نوجوانان شود.

نورامبونا[7]، هورنینگ[8] و میترا[9] (۲۰۲۰) به ارزیابی هرم وارونهٔ خبر و امتیازدهی به بخش‌های هرم به‌هنگام استخراج اطلاعات برای دو موضوع الف) شیوهٔ «۵ سؤال و ۱ چگونگی» در ساختار هرم وارونه خبر که با اصطلاح «5W1H» شناخته می‌شود و در برچسب‌گذاری بخش‌های مختلف خبر به‌کار رفته‌است و ب) خلاصه‌سازی متن پرداخته‌اند. در این پژوهش، از مقالات اسوشیتدپرس[10] مربوط به خبرهای فوری و غیرفوری در ژانر اقتصادی و سیاسی استفاده شده‌است. براساس نتایج به‌دست‌آمده، ساختار خبر در اخبار فوری و غیرفوری متفاوت است و اخبار فوری از ساختار هرم وارونهٔ اخبار بیشتر استفاده می‌کند.

مطالعات انجام‌شده در فارسی بیشتر در چارچوب تحلیل گفتمان تصویری بوده و به بررسی همبستگی معنایی عکس به‌کاررفته در متن و محتوای متن پرداخته‌اند. اقبالی (۱۳۹۰) به بررسی تصویرسازی کتاب‌های داستان کودک بین ۱۳۴۰ تا ۱۳۸۰ پرداخته و ۵۰ نمونه را بررسی کرده‌است. براساس یافته‌های وی، الزاماً همیشه همبستگی معنایی در کتب داستان کودکان رعایت نشده و گاهی تصویرسازی مغایر با متن داستان بوده‌است.

1. Wall Street Journal
2. inverted pyramid
3. E. Ytreberg
4. K. Emde
5. C. Klimmt
6. D. M. Schlütz
7. B. K. Norambuena
8. M. A. Horning
9. T. Mitra
10. Associated Press

عظیمی‌فرد و همکاران (۱۳۹۶) در چارچوب تحلیل گفتمان تصویری، به بررسی همبستگی معنایی عکس به‌کاررفته در خبر و محتوای خبری پرداخته‌اند. در این پژوهش، سه وبگاه خبری خبر، العالم و پرس‌تی‌وی با یکدیگر مقایسه شده‌است. در این مطالعه، به‌صورت تصادفی، ۳۰ خبر از مجموع ۶۶۴ خبر به‌عنوان داده‌های خبر انتخاب شده و به ارزیابی کیفی آنها پرداخته شده‌است. براساس تحلیل انجام‌شده، در ۱۶ خبر همبستگی عکس و متن خبر رعایت شده‌است، در ۶ خبر تاحدودی این همبستگی رعایت شده و در ۸ خبر عدم وجود همبستگی دیده شده‌است. شایان ذکر است در مقایسهٔ انجام‌شده، شبکهٔ پرس‌تی‌وی عملکرد بهتری در انتخاب تصویر و ارتباط آن با محتوای خبر در مقایسه با شبکهٔ العالم و خبر داشته‌است.

اردکانی‌فرد و همکاران (۱۴۰۰) به تحلیل محتوایی خبرگزاری‌های رسمی در سال ۱۳۹۹ با محوریت دین پرداخته‌اند و ویژگی‌های این دسته از اخبار را استخراج کرده‌اند. براساس نتایج به‌دست‌آمده از این پژوهش، ساختار معیار هرم وارونه خبر در این گونهٔ خبری به این صورت رعایت شده‌است که ابتدا تازگی خبر و سپس رویکرد تبلیغی در خبر حفظ شده‌است. همچنین، ضمن رعایت گفتمان سنتی نسبت به دین، به ابعاد سیاسی، اجتماعی و مناسکی دین توجه شده‌است.

۳ چارچوب نظری

در انجام پژوهش حاضر سه موضوع دخیل است که اساس نظری این مطالعه را شکل می‌دهد: الف) یک موضوع تحلیل گفتمان خبر و ویژگی‌های آن است؛ ب) موضوع دیگر علم داده برای پاسخ به سؤالات پژوهش در حجم زیاد داده با کمک رایانه، آمار و روش‌های پردازشی است؛ و ج) بازنمایی معنایی واژه‌ها در چارچوب «معناشناسی توزیعی»[1] است. در ادامه، نظریاتی که چارچوب این پژوهش را شکل می‌دهد توضیح داده می‌شود.

۳-۱ تحلیل گفتمان خبر

در زبان‌شناسی ساختگرایی، دوسوسور[2] (۱۹۱۶) دو سطح صورت و معنا را برای زبان قائل شده‌است که ازطریق آوا در شنیدار یا خط در نوشتار تجلی یافته و به مخاطب منتقل

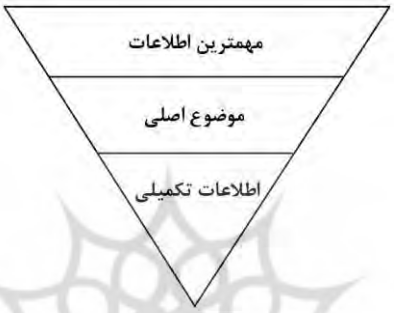
1. distributional semantics
2. F. de Saussure

می‌شود. خبر نوعی متن است که ازطریق نوشتار، پیامی که دارای معنا است را به مخاطب منتقل می‌کند. گفتمان خبری زیرمجموعهٔ تحلیل گفتمان است که به بررسی زبان در خبر می‌پردازد. خبر که گونه‌ای از متن است در نقش ارتباطی خود حاوی اطلاعات خارق‌العاده و خاصی است که برای جلب توجه مردم منتشر می‌شود. برای مثال، اگر گفته شود که «سگی فردی را گاز گرفته‌است» خبر نیست؛ چون اغلب اوقات این اتفاق می‌افتد و حاوی اطلاعات جدید نیست. درحالی‌که اگر «مردی سگی را گاز بگیرد» این رویداد یک خبر است (ایتول[1] و اندرسون[2]، ۲۰۰۶) و حاوی اطلاعات غیرقابل پیش‌بینی است. ژی[3] (۲۰۱۸) خبر را اطلاعاتی در مورد یک رویداد می‌داند و مطرح می‌کند که خبر به‌عنوان یک پنجره در جهت به‌دست‌آوردن دانش، ایفای نقش می‌کند. بنابراین، خبر نوعی نمود زبانی در ایجاد ارتباط میان فرد و اتفاقات محیط اطراف است و از نقش تعاملی زبان بهره برده می‌شود. همان‌طور که در مقدمه به آن اشاره شد، دوبوگراند و درسلر (۱۹۸۱: ۳) هفت ویژگی را برای نمود ارتباطی متن اشاره کرده‌اند که عبارت است از الف) وجود انسجام، ب) وجود پیوستگی معنایی، پ) وجود هدفمندی در نویسنده برای خلق متن، ت) وجود تبادل اطلاعات و انتظارات میان خواننده و شنونده در اطلاعات تبادل‌شده، ث) وجود عوامل مرتبط با تولید متن و ج) وجود ارتباط در تولید درک متن حاضر با تولید و درک متن پیشین. دو ویژگی انسجام و پیوستگی معنایی متن–محور و مابقی ویژگی‌ها کاربرد–محور است (آقاگلزاده، ۱۳۹۴: ۱۰۹). در متن خبری می‌توان این هفت ویژگی را یافت؛ ازاین‌رو، این نوع داده از درجهٔ اهمیت بالایی برخوردار است.

در چارچوب گفتمان خبری، یک ساختار کلی برای خبر معرفی شده‌است که به «هرم وارونه»[4] معروف است (بدنارک و کپل، ۲۰۱۲: ۱۰۰–۹۶). این چارچوب در قرن ۱۸ به‌عنوان ابزار مناسب انتشار اطلاعات معرفی شده‌است (پوتکر[5]، ۲۰۰۳). کاربرد عملی این چارچوب در اوایل قرن ۱۹ همزمان با اکتشاف تلگراف قابل مشاهده بود چراکه براساس اهمیت، موضوعات در ارتباطات مطرح می‌شد (کاناویلهاس[6]، ۲۰۰۷). در شکل ۱ ساختار هرم وارونهٔ خبر نمایش داده شده‌است. هر خبر متشکل از چهار بخش الف) عنوان یا سرخط خبر، ب) مقدمه یا سرنخ خبر، پ) بدنهٔ خبر و ت) اطلاعات تکمیلی جهت توسعهٔ سرنخ است. در سرخط، خلاصهٔ رویداد به‌گونه‌ای که برای خواننده ایجاد جذابیت کند

1. B. Itule
2. D. Anderson
3. Q. Xie
4. inverted pyramid
5. H. Pöttker
6. J. Canavilhas

بسیار کوتاه و موجز بیان می‌شود. ویژگی سرخط این است که معمولاً به‌صورت عبارت و به‌ندرت به‌صورت جمله بیان می‌شود. در بخش مقدمه یا سرنخ، جنبه‌های جذاب رویداد در اندازهٔ یک جملهٔ کوتاه یک سطری بیان می‌شود. به عبارتی دیگر، در قسمت سرنخ که تکمیل عنوان خبر است، مهمترین اطلاعات به‌صورت جمله بیان می‌شود. در بخش‌های بعدی که بدنه و توسعهٔ سرنخ است به بسط موضوع اصلی خبر و ارائهٔ اطلاعات تکمیلی می‌پردازد.



شکل ۱- هرم وارونهٔ خبر

اگرچه در هرم وارونه اطلاعات از نظر درجهٔ اهمیت مرتب شده‌است و رعایت این ساختار برای روزنامه‌نگاران مهم است؛ نگاه زبان‌شناسان در تحلیل گفتمان خبر به این صورت است که روابط منطقی یا بلاغی میان عنوان یا سرنخ و بقیهٔ اجزای خبر (بل، ۱۹۹۱: ۱۷۰) یا پیوند میان بخش‌های عنوان و سرنخ (فیز[1]، ادما[2] و وایت[3]، ۲۰۰۸) مورد توجه است. در پژوهش حاضر تلاش می‌شود پیوند و همبستگی معنایی میان عنوان و بخش‌های مختلف خبر در چارچوب علم داده بررسی شود.

۲-۳ علم داده

علم داده در دهه‌های ۱۹۶۰ و ۱۹۷۰ با مطرح‌شدن کشف اطلاعات و تحلیل محتوای معنایی شکل گرفت (کائو[4]، ۲۰۱۷). روش‌شناسی علم داده در چارچوب علم آمار و علم رایانه تکامل یافت؛ تا این که در سال ۱۹۹۸ مفهوم جدید علم داده توسط «چیکیو هایاشی»[5]

1. S. Feez
2. R. Edema
3. P. R. R. White
4. L. Cao
5. Chikio Hayashi

معرفی شد (مورتاگ[1] و دولین[2]، ۲۰۱۸). کلولند[3] (۲۰۰۱) مفهوم امروزی علم داده را این‌گونه معرفی می‌کند: علم داده یک بین‌رشته‌ای است که در آن داده‌های فاقد ساختار، ساختارمند شده و با بهره‌گیری از روش‌های الگوریتمی پردازش اطلاعات می‌توان به استخراج دانش و افزایش بینش پرداخت. بنابراین، داده، اطلاعات و دانش ارکان هرم دانش را در علم داده می‌سازد که در شکل ۲ نمایش داده شده‌است. در تعریف دالکیر[4] (۲۰۰۵)، «داده» محتوای قابل دیدن یا قابل تغییر، «اطلاعات» بازنمایی دادهٔ تحلیل‌شده و «دانش» اطلاعات نظری و مفید است. در پژوهش حاضر، داده همان متن خام خبری است؛ اطلاعات بخش‌های محتوای خبری و دانش یافتن همبستگی معنایی میان عنوان و بخش‌های خبری است.

شکل ۲- هرم دانش

در حوزهٔ علم داده، از آمار و روش‌های «یادگیری ماشین»[5] در هوش مصنوعی برای ساخت الگوریتم‌های پردازشی جدید بهره برده می‌شود؛ و براساس نیازهای تعریف‌شده در این حوزه، کار ذخیره‌سازی و استخراج اطلاعات از داده به‌صورت الگوریتمی انجام می‌پذیرد. همچنین، تحلیل داده برای حل مسائل پیچیده و نمایش بصری اطلاعات نیز به‌کار می‌رود تا به استخراج دانش و افزایش بینش از داده منجر شود. برای تحقق اهداف علم داده، به مدل‌سازی داده نیاز است. در مدل‌سازی داده، تمام فرایند پردازش داده به‌صورت روشمند و الگوریتمی تعریف و سازماندهی می‌شود. ازآنجاکه انواع داده‌های متنی، صوتی، تصویری و

1. F. Murtagh
2. K. Devlin
3. W. S. Cleveland
4. K. Dalkir
5. machine learning

عددی وجود دارد، نحوهٔ جمع‌آوری، ذخیره‌سازی، ساماندهی و استفاده از این اطلاعات متفاوت است.

۳-۳ معناشناسی توزیعی

مدل‌سازی داده بخش مهم علم داده است که روش‌شناسی انجام پژوهش را در دل خود دارد و بر فرایند تحلیل اثربخش است. برای بررسی وجود همبستگی معنایی میان عنوان خبر و سایر بخش‌های سازندهٔ خبر در پژوهش حاضر نیاز است این تحلیل در چارچوب معناشناسی توزیعی انجام پذیرد. هریس[1] (۱۹۵۴) معتقد است که واژه‌هایی که در یک بافت زبانی یکسان به‌کار می‌رود، این تمایل را دارند که از نظر معنایی به یکدیگر شبیه[2] باشد. بنا بر نظر وی، معنای هر واژه منعکس‌کنندهٔ بافتی است که آن واژه در آن بافت به کار می‌رود. نظر هریس در چارچوب معناشناسی توزیعی قرار می‌گیرد که منجر به معرفی‌شدن «فرضیهٔ توزیعی»[3] شده‌است. بر اساس این فرضیه، واژه‌هایی که تمایل به حضور در یک بافت یکسان دارد، از نظر معنایی مشابه یکدیگر است. بنا بر نظر هریس، معنای واژه تأثیرپذیر از «بافت جایگاهی»[4] است که آن واژه در آن بافت ظاهر شده‌است. در همین راستا، فرث[5] (۱۹۵۷) می‌افزاید که با توجه به واژه‌های اطراف یک واژه می‌توان معنای یک واژه را مشخص کرد. نتیجهٔ این نظرات آن است که بافت زبانی واژه، نقش بسیار مهمی در تعیین معنای یک واژه دارد.

در چارچوب معناشناسی توزیعی، میکولوو[6] و همکاران (۲۰۱۳) مدلی را معرفی کرده‌اند که می‌توان معنای واژه را به‌صورت بردار[7] بازنمایی کرد. در این شیوهٔ بازنمایی معنایی، به‌جای صورت‌واژه از برداروژه که براساس بافت جایگاهی واژه در یک پیکرهٔ زبانی به‌دست آمده‌است استفاده می‌شود.

۴ جمع‌آوری داده

انجام هر پژوهشی به داده نیاز دارد و هرقدر حجم داده‌های پژوهش بیشتر باشد، نتایج به‌دست‌آمده موثق‌تر است. هدف اصلی این پژوهش یافتن همبستگی میان عنوان و محتوای

1. Z. S. Harris
2. similar
3. distributional hypothesis
4. local context
5. J. R. Firth
6. T. Mikolov
7. vector

خبر است. ازاین‌رو، به یک پیکرهٔ خبری نیاز است که می‌توان به‌واسطهٔ خزش از وبگاه‌های خبری به تهیهٔ این پیکره اقدام کرد. دسترسی به بایگانی برخط وبگاه‌های خبری در رسیدن به هدف این پژوهش کمک شایانی می‌کند. وجود داده با حجم زیاد سبب می‌شود بتوان از روش‌های پیکره–محور[1] و نه پیکره–بنیان[2] (توگنینی–بونلی[3]، ۲۰۰۱: ۸۴–۸۵) در تحلیل داده‌ها بهره برد. در روش پیکره–محور از علم داده استفاده می‌شود و از بررسی داده به فرضیه می‌رسیم؛ درحالی‌که در روش پیکره–بنیان فرضیه اولیه‌ای وجود دارد و با بررسی داده به اثبات آن فرضیه می‌پردازیم. اکثر پژوهش‌های انجام‌شده فرضیه‌محور و از نوع پیکره–بنیان است. طرح حاضر می‌تواند زمینه را برای انجام طرح‌های پیکره–محور مبتنی‌بر علم داده فراهم کند.

۱-۴ اهمیت داده‌های خبری

امروز حجم زیادی از انواع داده‌ها در بستر وب وجود دارد؛ و ازمیان آنها دادهٔ متنی از حجم و تنوع بیشتری برخوردار است. بنابراین محتوای زیادی از این نوع داده تولید می‌شود. یکی‌از منابع این نوع داده که به‌طور روزانه به تولید حجم زیادی از محتوا می‌پردازد خبرهای منتشرشده در وبگاه‌های خبری است. این نوع داده که توجه پژوهشگران را به خود جلب کرده‌است از جنبهٔ زبان‌شناسی و کاربردشناسی اهمیت دارد که در ادامه توضیح داده می‌شود. این نوع داده از تنوع ژانر، مانند سیاسی، اجتماعی، اقتصادی و غیره، برخوردار است که می‌توان به‌صورت الگوریتمی ژانر خبر را تعیین کرد. تولید محتوا در وبگاه‌های خبری روزانه است. بنابراین با تحلیل هم‌زمانی می‌توان به بررسی سیر تحول موضوعات پرداخت. این داده‌ها معمولاً در وبگاه‌های خبری بایگانی می‌شود. بنابراین تا زمان حیات روزنامه، امکان دسترسی به سوابق اخبار وجود دارد. گزارش‌های روزانه از تحولات سیاسی، اجتماعی، اقتصادی و غیره سبب می‌شود بتوان زبان را رصد کرد و از نوواژه‌ها، تغییرات معنایی واژه‌ها و همچنین تغییرات زبانی باتوجه‌به مؤلفهٔ زمان آگاه شد. ازآنجاکه خبرها توسط افراد مختلف نگارش می‌شود، سبک نگارشی افراد در داده‌ها مستتر است که می‌تواند در مطالعات سبک‌شناسی نیز به‌کار رود. معمولاً دسترسی به اخبار این وبگاه‌ها رایگان است. درنتیجه، برای گردآوری داده‌ها نباید هزینه‌ای پرداخت شود. ازآنجاکه امروزه

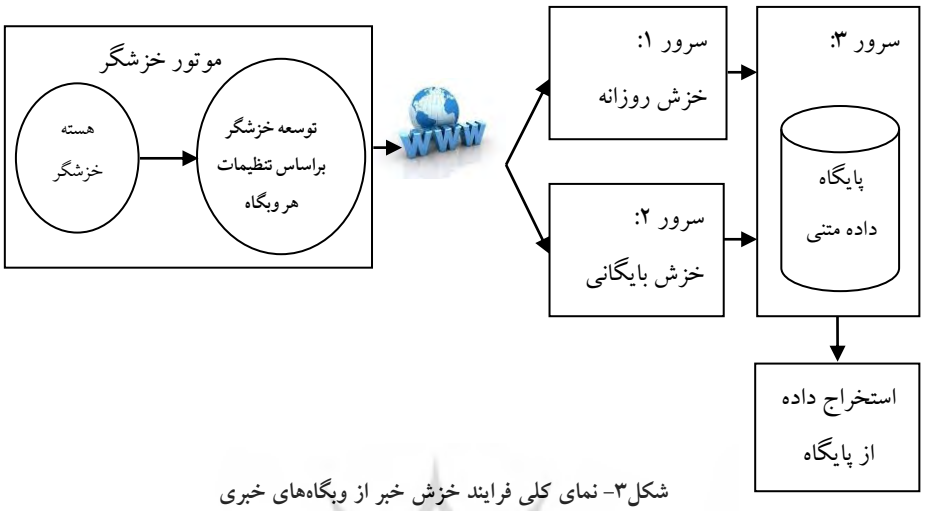
1. corpus-driven 2. corpus-based 3. E. Tognini-Bonelli

به‌منظور پردازش زبان طبیعی به حجم مناسب داده برای مدل‌سازی زبان نیاز است، این نوع داده می‌تواند خیلی سریع با کمترین هزینه نیاز الگوریتم‌های پردازشی را برآوَرَد. هر خبر متشکل از چند موضوع است که کنار یکدیگر قرار گرفته‌است. می‌توان موضوع‌های موجود در هر خبر را با کمک الگوریتم «مدل‌سازی موضوع»[1] (بلای[2] و همکاران، ۲۰۰۳) استخراج کرد و خبرهای دارای تشابه موضوعی را خوشه‌بندی کرد. علاوه‌بر آن می‌توان کلیدواژه‌های مرتبط با هر موضوع را نیز استخراج کرد که می‌تواند در تولید هشتگ[3] برای گروه‌بندی خبرها کاربرد داشته باشد. تحلیل احساسات نهفته در هر خبر و یافتن گرایش و رویکرد خبرگزاری در انتشار خبر موضوع دیگری است که کاربرد دیگر پیکرهٔ خبری را مشخص می‌کند. تمیزدادن خبر جعلی که با قصد فریب افراد منتشر می‌شود از خبر موثق رویکرد دیگری در تحلیل این نوع داده است. خلاصه‌سازی الگوریتمیِ خبر نیز موضوعی است که به داده‌های خبری نیاز دارد. علاوه‌بر موارد مطرح‌شدهٔ بالا که کارایی‌های مختلف پیکرهٔ خبری را مشخص می‌کند می‌توان از این داده در تحلیل‌های دیگری استفاده کرد. حال که اهمیت داده‌های خبری، چه از بعد پژوهشی و چه از بعد کاربردی، مشخص شد، در ادامه، شیوهٔ جمع‌آوری داده‌های مورد نیاز پژوهش حاضر توضیح داده می‌شود.

۲-۴ خزش و ذخیره‌سازی اطلاعات

مرحلهٔ اول، تهیهٔ «پایگاه دادهٔ متنی» است که به‌عنوان بایگانی کلی از اخبار بتواند برای سایر پژوهش‌های علوم انسانیِ مبتنی بر داده مورد استفاده قرار گیرد. در این مدل خزش، اخبار به‌صورت روزانه از وبگاه‌های خبری خزش شده و به داده‌های قبلی در این دو پایگاه داده افزوده می‌شود؛ بنابراین یک فرایند پیوستار و پویا در طراحی جمع‌آوری داده وجود دارد و مقطعی نیست. نتیجهٔ این عملکرد می‌تواند ضمن به‌روزرسانی روزانهٔ داده‌ها در پایگاه، کمک کند تا از این اطلاعات برای رصد تغییرات زبانی و یافتن نوواژه‌های واردشده به زبان کمک گرفت (قیومی، ۱۳۹۹). نمای کلی فرایند خزش خبر از وبگاه‌های خبری در شکل ۲ نمایش داده شده‌است.

1. Topic modeling
2. D. M. Blei
3. Hashtag



شکل۳- نمای کلی فرایند خزش خبر از وبگاه‌های خبری

همان‌گونه که در شکل ۳ مشخص است، برای جمع‌آوری داده به سه بخش نیاز است: الف) موتور خزشگر از وب، ب) تعریف فرایند استخراج اطلاعات از پایگاه‌های داده و ج) سرور برای خزش و ذخیره‌سازی داده در پایگاه داده نیاز است که این سه بخش در ادامه توضیح داده می‌شود.

به هنگام تهیهٔ موتور خزشگر از وب و تعریف فرایند استخراج اطلاعات، با بررسی‌های انجام‌شده از مجموعه وبگاه‌های خبری هدف، به این نتیجه دست یافته شد که یک موتور خزشگر دو بخشی طراحی شود. این دو بخش شامل هستهٔ خزشگر و توسعهٔ این هسته براساس تنظیمات مورد نیاز هر وبگاه خبری است:

الف) هستهٔ خزشگر: هستهٔ خزشگر حاوی مفاهیم پایه‌ای است که برای خزش هر وبگاه خبری نیاز است. بنابراین با ایجاد این هسته، اطلاعات کلیدیِ قابل استخراج از وبگاه‌ها مشخص می‌شود که این کار ضمن افزایش سرعت برای راه‌اندازی خزش یک وبگاه، در زمان صرفه‌جویی می‌کند.

ب) توسعه هستهٔ خزشگر و انطباق با نیاز هر وبگاه: بخش دوم موتور خزشگر حاوی تنظیمات هر وبگاه جهت استخراج اطلاعات مورد نظر از هر صفحه خبری با درنظرگرفتن مفاهیم پایهٔ مشخص‌شده در هستهٔ خزشگر است. در این بخش، ابتدا باید براساس ساختاری که در هر صفحه وب جهت نمایش خبر تنظیم شده‌است تجزیه شود و از هر صفحه فراداده‌هایی مانند نام وبگاه خبری، تاریخ و ساعت انتشار خبر، کد خبر، عنوان خبر،

خلاصه خبر که معمولاً توسط خبرگزاری‌ها درج می‌شود، متن خبر، کلیدواژه‌ها و پیوند دسترسی به خبر استخراج شود.

ساختار سامانه‌ای که برای خزش و ذخیره‌سازی داده‌های خزش‌شده طراحی شده‌است به این صورت است که به سه سرور نیاز است و هرکدام از سرورها نقش مشخصی را در جمع‌آوری و ذخیره داده دارد. سرور شماره (۱) فقط برای خزش روزانهٔ اخبار از وبگاه‌های خبری هدف تخصیص داده شده‌است. سرور شماره (۲) فقط برای خزش بایگانی اخبار از وبگاه‌های خبری هدف تخصیص داده شده‌است. وجود تعداد زیاد صفحات وب بایگانی سبب شد برای کاهش زمان در فرایند خزش از بایگانی وبگاه‌ها، از شیوهٔ خزش هم‌زمان و موازی استفاده شود. در سرور شماره (۳) پایگاه داده به‌نام «پایگاه دادهٔ متنی» وجود دارد که ساختار جداول آن براساس اطلاعات خزش‌شده طراحی شده‌است. ازآنجاکه حجم اخبار هر وبگاه نسبتاً زیاد بوده و خزش چندین وبگاه خبری موجب حجیم‌شدن و زمان‌برشدن استخراج داده و فراداده‌های مرتبط شد، داده‌های خزش‌شده از هر وبگاه در یک پایگاه مستقل سازماندهی شده‌است. بنابراین، در این پژوهش ۲۴ پایگاه داده تهیه شده‌است. استخراج اطلاعات پایگاه داده مربوط به هر وبگاه خبری و تجمیع آنها می‌تواند به یک پیکره بزرگ خبری منجر شود. برای کاربردی‌شدن این پیکره بزرگ که از منابع مختلف گردآوری شده‌است در پژوهش‌های مبتنی‌بر آمار و پردازش رایانشی نیاز است یکپارچگی نسبی در پیکره ایجاد گردد. ازاین‌رو، با درنظرداشتن اشکالات متداول در تهیه پیکره برای زبان فارسی (قیومی[1] و همکاران، ۲۰۱۰) می‌بایست کار پیش‌پردازش انجام پذیرد.

در فرایند خزش، ۲۴ وبگاه خبری را انتخاب کردیم. ویژگی این وبگاه‌ها دسترسی به «هم‌نشری بیش‌ساده»[2] در وبگاه است. ویژگی هم‌نشری بیش‌ساده این است که اجازهٔ دسترسی به به‌روزرسانی‌های وبگاه‌ها را در ساختار معیاری که قابلیت خوانش توسط رایانه دارد می‌دهد.

اطلاعات آماری موجود در «پایگاه دادهٔ متنی» از بایگانی ۲۴ وبگاه خبری خزش‌شده از تاریخ ۱۳۶۸/۴/۱ تا ۱۴۰۱/۲/۲۱ در جدول ۱ گزارش شده‌است. همانگونه که مشخص است، تعداد کل اسناد خزش‌شده، اعم از اخبار بایگانی و روزانه، بالغ‌بر ۴۴ میلیون خبر است که این حجم داده بیش‌از ۱۴ میلیارد واژه را شامل شده‌است. شایان ذکر است بعضی از وبگاه‌های خبری، علاوه‌بر ارائهٔ محتوای خبری به زبان فارسی، اخبار گزینش‌شده را به زبان‌های غیرفارسی نیز ارائه می‌کنند. حجم دادهٔ گزارش‌شده در جدول ۱ متعلق به بخش فارسی وبگاه‌های خبری منتخب است.

1. M. Ghayoomi
2. really simple syndication (RSS)

جدول ۱- اطلاعات آماری استخراج‌شده از بایگانی وبگاه‌های منتخب خزش‌شده

وبگاه خبری	تعداد اسناد خبری	تعداد واژه‌ها	تعداد زبان‌ها	تاریخ شروع
ایرنا	۷۴۸۹۸۳۷	۲۳۵۴۵۴۴۲۵۰	فارسی، انگلیسی، عربی، ترکی، اسپانیایی، اردو، روسی، آلمانی، فرانسه، چینی	۱۳۶۸/۴/۱
مهرنیوز	۵۸۰۷۷۰۴	۱۹۵۲۷۶۲۰۹۳	فارسی، انگلیسی، عربی، اردو، ترکی، کردی	۱۳۸۲/۴/۱۱
ایسنا	۶۲۳۶۱۶۳	۲۳۸۳۲۸۱۴۳۸	فارسی، انگلیسی، عربی، فرانسه	۱۳۷۸/۱/۲۲
خبرگزاری جوان	۶۲۹۶۱۱۵	۹۴۳۵۸۷۲۶۱	فارسی، انگلیسی، عربی	۱۳۹۰/۲/۴
تسنیم نیوز	۲۸۹۱۰۳۶	۱۲۱۲۱۵۸۷۵۷	فارسی، انگلیسی، عربی	۱۳۹۱/۸/۲۰
خبرگزاری صداوسیما	۲۲۴۷۵۷۳	۵۷۳۵۱۷۷۳۳	فارسی، انگلیسی، عربی	۱۳۹۰/۱۲/۹
مشرق نیوز	۱۲۰۰۱۴۷	۴۹۳۸۸۳۰۱۶	فارسی	۱۳۸۹/۱/۲۲
تابناک	۱۰۷۸۷۸۶	۴۶۴۷۶۲۵۶۸	فارسی، عربی	۱۳۸۳/۱/۱۱
خبرآنلاین	۱۰۱۷۰۲۲	۳۹۰۱۶۷۹۷۹	فارسی، انگلیسی، عربی	۱۳۸۷/۵/۲۵
شفقنا	۹۴۴۹۳۵	۱۰۳۰۰۱۲۲۱	فارسی، انگلیسی، عربی، ترکی، اسپانیایی، اردو، فرانسه	۱۳۹۴/۱۰/۲۴
همشهری آنلاین	۸۵۲۰۸۳	۳۶۰۶۹۶۰۹۱	فارسی	۱۳۸۵/۱/۱
عصرایران	۱۱۹۰۴۴۶	۴۶۶۱۳۴۷۴۱	فارسی، عربی	۱۳۸۵/۴/۱۳
فارس نیوز	۳۹۹۷۱۸۳	۱۴۰۰۳۳۵۴۴۵	فارسی، انگلیسی، عربی، ترکی	۱۳۸۱/۴/۱۱
نامه نیوز	۶۰۰۶۱۵	۵۶۹۱۵۹۲۳	فارسی	۱۳۹۱/۶/۵
شفاف	۵۲۹۶۱۷	۱۶۵۹۱۰۳۳۱	فارسی	۱۳۸۷/۱۱/۲۶
روزنو	۴۶۱۴۷۴	۱۹۶۱۴۱۳۵۳	فارسی	۱۳۹۱/۶/۱۴
ایران اکونومیست	۴۵۶۴۰۸	۱۴۵۳۲۳۲۰۹	فارسی	۱۳۹۰/۲/۲۷
رجانیوز	۲۵۲۷۰۱	۱۳۸۹۴۵۵۳۰	فارسی	۱۳۸۷/۲/۷
صبحانه آنلاین	۲۶۳۹۸۴	۱۱۰۴۲۸۵۶۵	فارسی	۱۳۹۲/۲/۱۷
شبکه خبر	۴۴۲۹۷۱	۷۵۵۸۸۴۲۴	فارسی	۱۳۹۰/۱۰/۲۰
اخبار بانک	۱۷۶۹۸۰	۶۴۵۲۲۳۲۲	فارسی	۱۳۹۰/۱۰/۱۴
تیترنیوز	۱۰۷۱۶۱	۲۸۴۸۶۹۳۳	فارسی	۱۳۹۲/۲/۲۳
وزارت بهداشت	۱۸۵۴۷	۱۲۳۸۴۵۷۷	فارسی، انگلیسی	۱۳۹۳/۸/۲۶
سازمان مدیریت بحران کشور	۹۹۶۹	۳۱۵۹۵۷۱	فارسی	۱۳۹۲/۳/۴
جمع کل	**۴۴۵۶۹۴۵۷**	**۱۴۰۹۶۶۳۹۳۳۱**	**۱۱ زبان**	از ۱۳۶۸/۴/۱ تا ۱۴۰۱/۲/۲۱

در جدول ۲ اطلاعات آماری استخراج‌شده از تاریخ ۱۳۶۸/۴/۱ تـا ۱۴۰۱/۲/۲۱ براسـاس تفکیک زبانی از پایگاه دادهٔ متنی گزارش شده‌است. شایان ذکر است پژوهش حاضر محـدود به بررسی خبرهای فارسی شده‌است.

جدول ۲- اطلاعات آماری از پیکرهٔ بزرگ متنی براساس تفکیک زبانی

زبان	تعداد وبگاه	تعداد اسناد خبری	تعداد واژه‌ها
فارسی	۲۴	۴۴۵۶۹۴۵۷	۱۴۰۹۶۶۳۹۳۳۱
عربی	۱۱	۷۰۱۴۰۷	۱۴۸۳۵۳۷۵۹
انگلیسی	۱۰	۸۸۳۰۳۱	۲۳۲۲۱۵۶۷۴
ترکی	۴	۸۱۶۹۱	۱۲۴۸۹۶۶۸
اردو	۳	۸۹۸۸۴	۲۱۷۰۵۶۹۶
فرانسه	۳	۳۸۱۱۳	۸۲۷۸۶۱۹
اسپانیایی	۲	۱۲۲۶	۳۳۱۵۸۶
کردی	۱	۶۰۲۹۰	۹۳۱۶۷۱۱
روسی	۱	۲۹۰۵۵	۳۸۷۹۸۰۶
آلمانی	۱	۱۱۸۱۴	۱۴۴۶۰۳۸
چینی	۱	۲۴۷۲	-

۵ تحلیل داده

در بخش ۳، گفتمان خبری و ساختار خبر معرفی شد. توضیح داده شد کـه خبـر متشـکل از عنوان و محتوا است و براساس هرم وارونه محتوا به سه قسمت سرنخ، بدنه و توسعهٔ سـرنخ تقسیم می‌شود. در این پژوهش می‌خواهیم میزان همبستگی معنایی میان عنوان و سـه بخـش محتوای خبری را مشخص کنیم تا از نظـر آمـاری ببینـیم ارتبـاط عنـوان بـا کـدام بخـش از محتوای خبر ارتباط بیشتر دارد. ازایـن‌رو، تـلاش مـی‌کنیم بـا کمـک ابـزار آمـاری ضـریب همبستگی پیرسون به پرسش‌های این پژوهش پاسخ دهیم.

همبستگی یک شیوهٔ آماری است که برای مقایسهٔ دو متغیر کمّی پیوسته استفاده می‌شـود. بـه عبـارتی دیگـر، در همبسـتگی، شـدت پیوسـتگی دو متغیـر سـنجیده می‌شـود. ضـریب همبستگی عددی بین ۱ تا ۱- است. اگر ضریب همبسـتگی پیرسـون عـدد ۱ باشـد، بیـانگر رابطهٔ مستقیم بین دو متغیر است به این مفهوم که هرگونه تغییر در یک متغیـر، متغیـر دیگـر نیز تغییر می‌کند. چنانچه این ضریب عدد ۱- باشد، رابطهٔ معکوس بین دو متغیر وجـود دارد

که با افزایش یک متغیر، متغیر دیگر کاهش می‌یابد. اگر این ضریب صفر باشد بیانگر عدم وجود رابطهٔ خطی بین این دو متغیر است (بادی[1] و اسمیت[2]، ۲۰۰۹: ۹۴-۹۲). از تساوی (۱) برای محاسبهٔ ضریب همبستگی پیرسون استفاده می‌شود:

(۱)

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

که در این تساوی، N تعداد جفت‌های مورد نظر برای مقایسه است، $\sum xy$ مجموع تعداد توالی جفت‌های مورد نظر است، $\sum x$ مجموع تعداد متغیر x و $\sum y$ مجموع تعداد متغیر y است، $\sum x^2$ مجموع مجذور تعداد متغیر x و $\sum y^2$ مجموع مجذور تعداد متغیر y است.

ازآنجاکه در این پژوهش به‌دنبال وجود رابطهٔ همبستگی بین عنوان و محتوای خبری هستیم، برای مقایسه‌پذیری این دو به‌عنوان دو متغیر در یک متن خبری، نیاز است ساختار متنی داده از شکل اولیهٔ خود خارج و در چارچوب معناشناسی توزیعی که براساس آرای هریس (۱۹۵۴) و فرث (۱۹۵۷) شکل گرفته‌است در قالب بردار بازنمایی شود. در این پژوهش تلاش کرده‌ایم با استفاده از ابزار ورد۲وک[3] که توسط میکولوو و همکاران (۲۰۱۳) تهیه شده‌است برای بردارسازی داده‌های خزش‌شده از وب استفاده کنیم.

یکی‌از متغیرهای این پژوهش عدم توازن بین عنوان خبر و متن خبر است؛ چراکه عنوان خبر به‌طور متوسط حدود ۱۰ واژه بوده و محتوای خبر از این تعداد بیشتر است. در محاسبهٔ همبستگی میان دو متغیر، تعداد دوبه‌دوی متغیرها یکسان است. بنابراین، بردارسازی داده در قالب معناشناسی توزیعی سبب یکسان‌شدن این متغیر شده و مقایسه‌پذیری داده را ساده می‌سازد؛ به این صورت که می‌توان براساس تعداد مشخص بُعد در بردارها، به مقایسهٔ کنترل‌شدهٔ داده پرداخت.

ازجمله دیگر متغیرهایی که در فرایند بازنمایی معنایی واژه‌ها باید تثبیت شود تعداد بُعد بردارها و همچنین تعیین تعداد واژه‌های همسایه در بافت جایگاهی برای محدودکردن بافت و تعیین معنای آن واژه در آن بافت است. برای این هدف، تعداد ۱۰۰ بُعد و ۸ واژه در بافت جایگاهی (۴ واژه قبل و ۴ واژه بعداز واژهٔ هدف) را در فرایند بردارسازی هر واژه مورد استفاده قرار می‌دهیم. حجم پیکرهٔ حاصل از تجمیع ۲۴ وبگاه خبری که بیش‌از

1. R. Boddy 2. G. Smith 3. Word2Vec

۱۴میلیارد واژه است توسط یک سرور که دارای دو پردازندهٔ ۱۲ هسته‌ای و ۲۵۶ گیگابایت حافظهٔ رم است پردازش شده و با استفاده از ورد۲وک بردار واژگان پیکره به‌دست آمده‌است. پس‌از بازنمایی معنایی واژه‌ها در قالب بردار، عنوان و محتوای هر خبر از صورت‌واژه به بردارهای ۱۰۰ بُعدی تبدیل شده‌است؛ به این صورت که بردار واژه‌های یک متن با یکدیگر جمع شده و سپس براساس تعداد واژگان متن، بردار میانگین واژه‌ها به‌دست آمده‌است. بردار میانگین سبب حذف اثرگذاری متغیر تعداد واژه‌های متن بر بردار می‌شود.

ازآنجاکه در این پژوهش قصد بررسی رابطهٔ همبستگی معنایی میان عنوان و بخش‌های سرنخ، بدنه و توسعهٔ سرنخ را داریم، هر متن خبری خزش‌شده را به سه بخش تقطیع کرده‌ایم؛ به این صورت که جملهٔ اول متن خبر به‌عنوان سرنخ خبر، جملهٔ آخر به‌عنوان توسعهٔ سرنخ و جملات بینابین سرنخ و توسعهٔ سرنخ به‌عنوان بدنهٔ خبر تلقی شده‌است. بر اساس این تقسیم‌بندی، متن‌هایی که حداقل حاوی سه جمله است را پالایش کرده و در این پژوهش استفاده کرده‌ایم. فرایند بازنمایی برداری را علاوه‌بر عنوان، به‌طور مجزا برای این بخش‌ها تکرار می‌کنیم.

در جدول ۳ رابطهٔ همبستگی پیرسونِ میانگین بین عنوان و سرنخ، عنوان و بدنه و همچنین عنوان و توسعهٔ سرنخ برای هر روزنامه گزارش شده‌است. براساس امتیاز ضریب همبستگی پیرسون میانگین در وبگاه‌های خبری، همبستگی عنوان–سرنخ و همچنین عنوان–بدنه تقریباً یکسان است و براساس آزمون تی[1] دو دنباله‌ای، تفاوت معناداری بین این دو بخش خبری وجود ندارد. ولی همبستگی عنوان–توسعهٔ سرنخ نسبت‌به دو بخش دیگر خبری با تفاوت معناداری کمتر است ($p < 0/01$). نتایج به‌دست‌آمده حاکی از آن است که اگرچه عنوان خبر باید بیشترین ارتباط معنایی را با محتوای خبر داشته باشد و براساس هرم وارونه این نقش به بخش سرنخ خبر واگذار شده‌است، در بررسی وبگاه‌های خبری می‌بینیم که همبستگی عنوان–سرنخ و همچنین عنوان–بدنه حدود 0/58 است. این عدد بیانگر این است که سرنخ و بدنهٔ خبر هر دو به یک اندازه با عنوان خبر همبستگی دارد که با ساختار هرم وارونه کمی متفاوت است؛ چراکه بدنه باید ارتباط معنایی کمتری را با عنوان داشته باشد. شایان ذکر است امتیاز به‌دست‌آمدهٔ ضریب همبستگی خیلی بالا نیست؛ چراکه با ۱ فاصله معناداری دارد. به‌نظر می‌رسد برای رفع این مشکل و انتقال اطلاعات بیشترِ همسو بین عنوان و سرنخ باید دقت بیشتری توسط خبرنگاران صورت پذیرد.

1. *t*-test

جدول ۳- رابطهٔ همبستگی میانگین پیرسون بین عنوان-سرنخ، عنوان-بدنه و عنوان-توسعهٔ سرنخ

وبگاه خبری	همبستگی پیرسون		
	عنوان-سرنخ	عنوان-بدنه	عنوان-توسعهٔ سرنخ
ایرنا	0/618047	0/566799	0/131376
مهرنیوز	0/56198	0/608299	0/498211
ایسنا	0/681766	0/62791	0/180871
خبرگزاری جوان	0/227679	0/234171	0/0597137
تسنیم نیوز	0/537342	0/627867	0/17796
خبرگزاری صداوسیما	0/54448	0/48171	0/350428
مشرق‌نیوز	0/557071	0/573402	0/453529
تابناک	0/602778	0/580151	0/460634
خبرآنلاین	0/585131	0/603539	0/158178
شفقنا	**0/704899**	**0/64996**	0/408475
همشهری آنلاین	0/536267	0/568211	0/467625
عصرایران	0/630682	0/583441	0/44618
فارس‌نیوز	0/572369	0/624237	0/144334
نامه‌نیوز	0/585108	0/612639	0/498552
شفاف	0/569855	0/569097	0/420367
روز نو	0/623696	0/613549	0/476263
ایران اکونومیست	0/594269	0/6193358	0/363095
رجانیوز	0/593515	0/561825	0/441498
صبحانه آنلاین	0/568287	0/578347	0/317141
شبکه خبر	0/551843	0/563152	0/394299
اخبار بانک	0/596059	0/600346	0/490011
تیترنیوز	0/580048	0/604544	**0/501147**
وزارت بهداشت	0/637157	0/589484	0/24137
سازمان مدیریت بحران کشور	0/561328	0/550338	0/452316
میانگین وبگاه‌های خبری	**0/575902**	**0/574682**	**0/355566**

ازمیان ۲۴ وبگاه خبری و براساس محاسبات انجام‌شده در جدول ۳، ۲۳ بگاه خبری (۹۶٪ وبگاه‌ها) ضریب همبستگی نسبتاً بالایی را در مقایسهٔ بین عنوان و سرنخ به‌دست آورده‌است. ازمیان این وبگاه‌های خبری، وبگاه «شفقنا» بالاترین و وبگاه «خبرگزاری جوان» پایین‌ترین ضریب همبستگی پیرسون را در عنوان-سرنخ و همچنین عنوان-بدنه به‌دست

آورده‌است. وبگاه‌های خبری «ایسنا»، «وزارت بهداشت» و «ایرنا»، به‌ترتیب، جایگاه‌های دوم تا چهارم در مقایسهٔ عنوان- سرنخ به‌دست آورده‌است.

۲۲ بگاه خبری (۹۲٪ وبگاه‌ها) ضریب همبستگی نسبتاً بالایی را در مقایسهٔ بین عنوان و بدنه به‌دست آورده‌است و مابقی کمتر بوده‌است. این نکته بیان می‌کند در این وبگاه‌ها مفهوم عنوان در بدنهٔ خبر متبلور است. چنین می‌توان نتیجه گرفت که رابطهٔ عنوان-سرنخ و عنوان-بدنه که بخش‌های اصلی خبر در هرم وارونه را شامل می‌شود در اکثر خبرگزاری‌ها رعایت می‌شود.

ازمیان ۲۴ وبگاه خبری، در ۱۳ وبگاه (۵۴٪ وبگاه‌ها) همبستگی عنوان-بدنه از عنوان-سرنخ بیشتر بود. این نتیجه بیان می‌کند که در این وبگاه‌های خبری، بدنهٔ خبر از اهمیت بیشتری نسبت‌به سرنخ خبر برخوردار است که این یافته با هرم وارونه در تضاد است. شایان ذکر است درمورد همبستگی عنوان-توسعهٔ سرنخ به این نتیجه رسیدیم که در همهٔ وبگاه‌های خبری این همبستگی در مقایسه با عنوان-سرنخ و عنوان-بدنه کاهش یافته‌است که منطبق با هرم وارونهٔ خبر است.

در جدول ۴، فراوانی نسبی تعداد خبرهایی که براساس ضریب همبستگی پیرسون، امتیاز بالاتر یا پایین‌تر از امتیاز ۰/۵ را به‌دست آورده، یا دارای امتیاز ضریب همبستگی صفر یا ۱- است گزارش شده‌است. همان‌طور که مشخص است، باتوجه‌به خبرگزاری، نتایج متفاوتی از نظر تعداد اسناد خبری برای تقسیم‌بندی چهارگانه مبتنی‌بر ضریب همبستگی پیرسون به‌دست آمده‌است. ازآنجاکه در پژوهش حاضر از داده‌های خبری استفاده می‌شود، ضریب همبستگی صفر یا ۱- به یک صورت تعبیر می‌شود و به مفهوم عدم وجود رابطهٔ همبستگی است. به‌طور کلی، 70/61٪ از تمامی مستندهای خبری رابطهٔ همبستگی نسبتاً بالا ($r \geq 0/5$) بین عنوان و سرنخ وجود دارد؛ 72/59٪ از خبرها رابطهٔ همبستگی نسبتاً بالا بین عنوان و بدنه وجود دارد؛ و 42/53٪ از خبرها رابطهٔ همبستگی نسبتاً بالا بین عنوان و توسعهٔ سرنخ وجود دارد. همچنین، 26/30٪ از متون خبری رابطهٔ همبستگی نسبتاً پایینی ($r < 0/5$) بین عنوان و سرنخ وجود دارد؛ 23/91٪ از خبرها همبستگی نسبتاً پایینی بین عنوان-بدنه وجود دارد؛ و همبستگی نسبتاً پایین بین عنوان و توسعهٔ سرنخ در 44/30٪ از خبرها اتفاق افتاده‌است.

براساس نتایج جدول ۴، بعضی از خبرهای وبگاه‌های خبری همبستگی بین عنوان با سرنخ، بدنه و توسعهٔ سرنخ وجود ندارد. به‌طور متوسط، در 1/61٪ و 1/60٪ از خبرهای وبگاه‌ها، به ترتیب، هیچ رابطهٔ همبستگی در عنوان-سرنخ و عنوان- بدنه ندارد و در 12/47٪ از خبرها، هیچ رابطهٔ همبستگی در عنوان-توسعهٔ سرنخ خبرها ندارد.

بنا بر نتایج گزارش‌شده در جدول ۴، در اکثر وبگاه‌های خبری، کمتر از یک درصد از اخبار، همبستگی در عنوان–سرنخ و همچنین عنوان–بدنه وجود ندارد؛ با این وجود، در بیش‌از ۳۰٪ از خبرهای وبگاه «خبرگزاری جوان»، همبستگی در عنوان–سرنخ و عنوان–بدنه وجود ندارد که بیانگر عدم دقت در انتخاب عنوان یا عدم رعایت مدل وارونه خبر به‌هنگام تنظیم خبر است. عدم همبستگی در عنوان–توسعهٔ سرنخ، در بیش‌از ۳۲٪ از خبرهای این خبرگزاری وجود دارد که بالاترین تعداد خبر در مقایسه با سایر خبرگزاری‌ها است.

در وبگاه‌های «شفقنا» و «ایسنا»، همبستگی عنوان–سرنخ و عنوان–بدنه در بیش‌از ۸۵٪ از خبرهای منتشرشده وجود دارد. اگرچه همبستگی عنوان–سرنخ در خبرهای وبگاه «ایران اکونومیست» در 74/56٪ از خبرها مشاهده شده‌است، در بیش‌از ۸۰٪ از خبرهای این خبرگزاری، همبستگی در عنوان–بدنه وجود دارد. نکتهٔ قابل توجه در بررسی همبستگی عنوان–توسعهٔ سرنخ این است که در بعضی از وبگاه‌ها، این رابطه در حجم بسیار کمی از خبرها، کمتر از ۱۰٪، اتفاق افتاده‌است؛ مانند «ایرنا»، «خبرگزاری جوان»، «تسنیم‌نیوز» و «فارس‌نیوز». ولی در خبرگزاری‌هایی مانند «مهرنیوز»، «نامه‌نیوز»، «روز نو»، «اخبار بانک» و «تیترنیوز» همبستگی عنوان–توسعهٔ سرنخ در بیش‌از ۵۰٪ از خبرها وجود دارد.

جدول ۴- فراوانی نسبی خبرهای دارای ضریب همبستگی پیرسون بالاتر و پایین‌تر از 0/5، صفر و منفی یک

	عنوان–سرنخ (٪)			عنوان–بدنه (٪)			عنوان–توسعهٔ سرنخ (٪)		
ضریب همبستگی	r ≥ 0/5	r < 0/5	r ≤ 0	r ≥ 0/5	r < 0/5	r ≤ 0	r ≥ 0/5	r < 0/5	r ≤ 0
ایرنا	72/41	18/60	2/30	70/08	26/33	0/90	1/40	53/71	11/74
مهرنیوز	68/54	30/10	0/23	78/43	20/93	0/10	54/46	41/78	0/80
ایسنا	85/21	13/60	0/27	81/88	17/38	0/10	13/87	39/71	8/93
خبرگزاری جوان	28/26	10/17	30/65	29/95	8/56	30/65	1/51	23/68	32/09
تسنیم نیوز	64/01	34/62	0/19	81/58	17/98	0/06	625	63/37	3/92

خبرگزاری صداوسیما	مشرق‌نیوز	تابناک	خبرآنلاین	شفقنا	همشهری آنلاین	عصرایران	فارس‌نیوز	نامه‌نیوز	شفاف	روز نو	ایران اکونومیست	رجانیوز	صبحانه آنلاین
5/09	2/23	1/71	16/13	1/44	1/19	1/75	3/31	1/20	2/23	1/78	8/46	1/93	6/08
40/08	44/94	45/63	38/00	51/52	46/65	47/31	64/84	39/08	48/46	41/57	33/37	49/33	45/00
35/32	47/07	47/67	10/88	38/66	48/12	44/93	1/37	55/69	41/29	51/51	40/18	43/23	29/49
2/26	0/53	0/65	0/24	0/06	0/36	0/66	0/10	0/12	0/27	0/09	0/12	0/30	0/12
33/38	27/13	24/80	21/94	14/02	28/27	23/12	17/96	21/23	28/74	20/92	18/66	30/34	26/43
56/23	71/09	73/09	77/07	85/64	70/16	74/09	81/50	78/00	69/98	78/54	80/71	68/15	72/91
0/86	0/25	0/90	0/18	0/07	0/30	0/50	0/19	0/17	0/32	0/12	0/15	0/32	0/13
28/65	31/91	20/99	26/50	10/07	35/49	17/98	28/67	26/40	29/31	20/57	24/54	25/42	29/92
66/62	66/64	75/50	72/36	89/59	62/71	79/54	70/15	72/56	69/17	78/75	74/56	73/03	69/14

شبکه خبر	65/54	32/67	0/33	69/37	29/05	0/30	40/42	40/61	6/67
اخبار بانک	74/66	24/44	0/15	77/06	21/84	0/19	53/78	40/58	1/28
تیترنیوز	71/18	27/30	0/24	76/84	22/14	0/19	55/20	40/71	0/87
وزارت بهداشت	77/71	19/03	0/23	72/50	25/99	0/27	21/94	36/08	10/26
سازمان مدیریت بحران کشور	66/83	31/97	0/19	66/57	31/40	0/34	45/26	47/15	1/65
متوسط کل	70/61	26/30	1/61	72/56	23/91	1/60	42/53	44/30	12/47

۶ جمع‌بندی و نتیجه‌گیری

آنچه در این مقاله توضیح داده شد، تحلیل گفتمان خبری و ارزیابی همبستگی معنایی میان عنوان خبر و قسمت‌های تشکیل‌دهندهٔ خبر براساس ساختار هرم وارونه با استفاده از روش آماری بود. برای رسیدن به هدف، در چارچوب علم داده ابتدا پیکرهٔ نسبتاً بزرگی در بازه زمانی ۱۳۶۸ تا ۱۴۰۱ با حجم بیش‌از ۱۴ میلیارد واژه به‌واسطهٔ خزش ۲۴ وبگاه خبری به‌دست آمد و برای این پژوهش مورد استفاده قرار گرفت. از معیار آماری همبستگی پیرسون برای یافتن همبستگی معنایی بین عنوان و بخش‌های خبر استفاده شد. برای یافتن همبستگی معنایی عنوان و بخش‌های خبر، در چارچوب معناشناسی توزیعی، بازنمایی معنایی عنوان و بخش‌های خبر با استفاده از ابزار ورد۲وک به بردار تبدیل شد و کار تحلیل انجام شد.

براساس نتایج به‌دست‌آمده، در ۹۶٪ از خبرهای وبگاه‌ها، همبستگی معنایی عنوان و سرنخ بیش از ۰/۵ بود. همبستگی نسبتاً بالای معنایی بین عنوان و بدنه خبر در ۹۲٪ از وبگاه‌ها اتفاق افتاده بود؛ و در ۸٪ از وبگاه‌ها، همبستگی نسبتاً بالای معنایی ($r \geq 0/5$) بین عنوان و توسعهٔ سرنخ خبر وجود داشت. این نتیجه بیانگر این نکته است که اکثر خبرگزاری‌ها ساختار وارونه خبر را رعایت می‌کند؛ ولی می‌توان نمونه‌هایی از وبگاه‌های خبری یافت که این ساختار را رعایت نمی‌کند. همچنین، علاوه‌بر وجود همبستگی معنایی

بین عنوان و سرنخ، تقریباً به یک میزان این همبستگی میان عنوان و بدنه نیز وجود دارد. دستاورد کاربردی این پژوهش این است که می‌توان با کمک روش‌شناسی معرفی‌شده، یک ارزیابی اولیه بر ساختار محتوایی انتشار خبر نمود و آن را با ساختار معیار هرم وارونه خبر مقایسه کرد.

منابع

آقاگلزاده، فردوس (۱۳۹۴). *تحلیل گفتمان انتقادی*. تهران: انتشارات علمی و فرهنگی.

اردکانی‌فرد، زهرا، محمدمهدی فرقانی و منا سلگی (۱۴۰۰). «روزنامه‌نگاری دین در ایران؛ تحلیل محتوای خبرگزاری‌های رسمی در سال ۱۳۹۹». *فصلنامه مطالعات فرهنگ – ارتباطات*. س۲۳، ش۵۹، ۱۵۰-۱۲۳.

اقبالی، پرویز (۱۳۹۰). *تبیین رابطه تصویر و متن در تصویرسازی کتاب‌های داستانی کودک در ایران از سال ۱۳۴۰ تا ۱۳۸۰*. رسالهٔ دکتری. تهران: دانشکدهٔ هنر دانشگاه شاهد.

عظیمی‌فرد، فاطمه، سیاوش صلواتیان و علی‌رضا عمادالدین (۱۳۹۶). «مقایسه کاربرد عکس خبری در وبگاه شبکه‌های خبر، العالم و پرس تی.وی». *فصلنامه مطالعات رسانه‌های نوین*. س۳، ش۱۱، ۱۳۶-۹۵.

قیومی، مسعود (۱۳۹۹). «چالش سامانه‌های پردازش طبیعی در مواجه با نوواژه‌های «کرونایی»». *همایش ملی ابعاد انسانی-اجتماعی کرونا در ایران*. تهران: پژوهشگاه علوم انسانی ومطالعات فرهنگی.

Bednarek, M., & H. Caple (2012). *News Discourse*. London: Continuum International Publishing Group.

Bell, A. (1991). *The Language of News Media*. Oxford: Blackwell.

Blei, D. M., et al. (2003). "Latent Dirichlet allocation". *Journal of Machine Learning Research*. 3: 993–1022.

Boddy, R., & G. Smith (2009). *Statistical Methods in Practice: For scientists and Technologists*. Chichester, U.K.: Wiley.

Brown, G., & G. Yule (1983). *Discourse Analysis*. Cambridge: Cambridge University Press.

Cao, L. (2017). "Data science: A comprehensive overview". *ACM Computing Surveys*. 50.

De Beaugrande, R., & W. U. Dressler (1981). *Introduction to Text Linguistics*. London: Longman.

De Saussure, F. (1916). *Cours de linguistique générale*. Lausanne, Paris: Payot.

Canavilhas, J. (2007). "Web journalism: From the inverted pyramid to the tumbled pyramid". *Biblioteca on-line de ciências da comunicação*.

Cleveland, W. S. (2001). "Data science: an action plan for expanding the technical areas of the field of statistics". *International Statistical Review*. 69, 21–26.

Dalkir, K. (2005). *Knowledge Management in Theory and Practice*. Amsterdam: Elsevier Science Ltd.

Emde, K., & C. Klimmt, & D. M. Schlütz (2016). "Does storytelling help adolescents to process the news?". *Journalism Studies*. 17(5): 608-627.

Feez, S., & R. Iedema, & P. R. R. White (2008) *Media Literacy*. Surry Hills, NSW: NSW Adult Migrant Education Service.

Firth, J .R. (1957). "A synopsis of linguistic theory 1930-1955," *Studies in Linguistic Analysis (special volume of the Philological Society),* chapter 1, pp: 1-32, Oxford: Blackwell.

Ghayoomi, M., S. Momtazi, and M. Bijankhan (2010). "A study of corpus development for Persian". *International Journal on Asian Language Processing*. 20(1): 17-33.

Harris, Z. S. (1954). "Distributional structure". *Word*. 23 (10): 146–162.

Itule, B., & D. Anderson (2006). *News Writing and Reporting for Today's Media*. 7th Edition, McGraw-Hill.

Mikolov, T., et al. (2013). "Distributed representations of words and phrases and their compositionality". *Advances in Neural Information Processing Systems.* C. J. C. Burges, et al. (eds.), 26: 3111–3119.

Murtagh, F., & K. Devlin (2018). "The development of data science: Implications for education, employment, research, and the data revolution for sustainable development". *Big Data and Cognitive Computing*. 2.

Norambuena, B. K., & M. A. Horning, & T. Mitra (2020). "Evaluating the inverted pyramid structure through automatic 5W1H extraction and summarization". *Computational Journalism Symposium*. Boston, MA, USA, March 20 – 21.

Pöttker, H. (2003). "News and its communicative quality: the inverted pyramid—when and why did it appear?". *Journalism Studies*. 4(4): 501-511.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work, Studies in Corpus Linguistics* . Amsterdam, The Netherlands: John Benjamins Publishing Company.

Xie, Q. (2018). "Critical discourse analysis of news discourse". *Theory and Practice in Language Studies*. 8 (4): 399-403.

Zhang, H., & H. Liu (2016). "Visualizing structural "inverted pyramids in English news discourse across levels". *Text & Talk*. 36: 110-89.

Ytreberg, E. (2010). "Moving out of the inverted pyramid: Narratives and descriptions in television news". *Journalism Studies*. 2(3): 357-371.

استناد به این مقاله: قیومی، مسعود (۱۴۰۱). ارزیابی ساختار هرم وارونه در پیکرهٔ بزرگ خبری فارسی: تحلیل گفتمان خبری براساس همبستگی میان عنوان و محتوای خبر. *زبان و زبان‌شناسی*، ۱۸ (۳۵)، ۲۱_۴۵. doi: 10.30465/lsi.2023.849