Biannual scientific-research journal of language and linguistics
Fall & Winter 2024, 19 (37), 193-220
https://lsi-linguistics.ihcs.ac.ir/
DOI: 10.30465/LSI.2024.47545.1725

Original Research
Received: 28/Oct/2023 *** Accepted: 19/Jan/2024
ISSN: 2423-3847

A Corpus-based Study of Using Function and Content Words in Persian Authorship Attribution

Soltanzadeh, Fatemeh[1] PhD of Linguistics, Allameh Tabataba'i University, Tehran, Iran

Mirzaei, Azadeh[2] Associate Professor of Linguistics, Allameh Tabataba'i University, Tehran, Iran

Bahrani, Mohammad[3] Department of Computer Science, Faculty of Statistics, Mathematics and Computer, Allameh Tabataba'i University, Tehran, Iran

Modarres Khiabani, Shahram[4] Department of English Language and Translation, Islamic Azad University, Karaj, Alborz, Iran

Abstract

Nowadays, corpora are widely used in authorship attribution. In this research, a corpus of persian contemporary texts was applied to identify the authorship of texts and the effectiveness of function and content words in this task was compared. In order to reach this goal, seven contemporary writers named Hoshang Golshiri, Bozor Alavi, Ahmad Mahmoud, Mahmoud Dolatabadi, Nader Ebrahimi, Jalal Al Ahmad and Gholamhossein Saedi were selected and their books were collected. Then by using this corpus and deep learning algorithms like multilayer perceptron and Long Short Term Memory, effectiveness of function and content words was evaluated. The results of the research indicated that function words based method was superior to content words one in authorship attribution. In addition, pronouns, especially demonstrative and personal pronouns, showed the highest efficiency among the types of function words to determine the author of a text. Moreover, features based on conjunctions and auxiliary verbs were valuable to recognize persian writers.

Keywords: Function words, Content words, Corpus, Authorship Attribution.

1. fatemeh.slt@gmail.com (Corresponding Author)
2. azadeh.mirzaei@atu.ac.ir
3. bahrani@atu.ac.ir
4. shmodarress@yahoo.com

How to cite: Soltanzadeh, F., Mirzaei, A., Bahrani, M., & Modarres Khiabani, S. (2024). A Corpus-based Study of Using Function and Content Words in Persian Authorship Attribution. *Language and Linguistics*, *19*(37), 193 - 220. doi: 10.30465/lsi.2024.47545.1725

مقاله پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۹/۰۶ *** تاریخ پذیرش: ۱۴۰۲/۱۰/۳۰

ISSN: ۲۳۲۲-۳۸۴۷

دوفصلنامه علمی- پژوهشی زبان و زبان‌شناسی
دوره ۱۹، شماره ۳۷، پاییز و زمستان ۱۴۰۲، ۱۹۳- ۲۲۰
lsi-linguistics.ihcs.ac.ir
DOI: 10.30465/LSI.2024.47545.1725

بررسی پیکره‌بنیاد تعیین سبک نگارش متون فارسی با واژه‌های دستوری و محتوایی

سلطان‌زاده، فاطمه — گروه زبان‌شناسی، دانشکده ادبیات فارسی و زبان‌های خارجه، دانشگاه علامه طباطبائی، تهران، ایران

میرزائی، آزاده — گروه زبان‌شناسی، دانشکده ادبیات فارسی و زبان‌های خارجه، دانشگاه علامه طباطبائی، تهران، ایران

بحرانی، محمد — گروه رایانه، دانشگاه علامه طباطبائی، دانشکده آمار، ریاضی و رایانه، تهران، ایران

مدرس خیابانی، شهرام — عضو هیئت علمی دانشگاه آزاد اسلامی کرج، البرز، ایران

چکیده

امروزه پیکره‌های زبانی در تعیین سبک نگارش کاربرد فراوان دارند. در این پژوهش از پیکره‌ای به زبان فارسی برای تعیین سبک نگارش متون معاصر استفاده و کارآیی واژه‌های دستوری با واژه‌های محتوایی در راستای این هدف مقایسه شد. برای این منظور، پیکره‌ای از آثار هفت نویسندهٔ معاصر به نام‌های هوشنگ گلشیری، بزرگ علوی، احمد محمود، محمود دولت‌آبادی، نادر ابراهیمی، جلال آل احمد و غلامحسین ساعدی انتخاب و گردآوری شد. سپس با استفاده از این پیکره و الگوریتم‌های یادگیری عمیق چون پرسپترون چندلایه و حافظهٔ طولانی کوتاه‌مدت، کارآیی واژه‌های محتوایی و انواع واژه‌های دستوری در تشخیص سبک نگارش متن سنجیده شد. نتایج ارزیابی پژوهش نشان داد روش استفاده از واژه‌های دستوری عملکرد بهتری نسبت به روش واژه‌های محتوایی در تعیین سبک نگارش متن دارد. همچنین در میان انواع واژه‌های دستوری ضمایر خصوصاً ضمایر شخصی و اشاره بیشترین نقش را در تفکیک سبک نویسندگان پیکره داشتند. به علاوه، حروف ربط و افعال کمکی در تعیین نویسندگان فارسی بسیار مؤثر بودند.

کلیدواژه: واژه‌های دستوری، واژه‌های محتوایی، پیکرهٔ زبانی، تعیین سبک نگارش

۱ مقدمه

امروزه پیکره‌های زبانی در تعیین سبک نگارش یک متن به‌صورت خودکار کاربرد فراوان دارند. این متن می‌تواند متن یک نامه یا ایمیل تهدیدآمیز، یک نامهٔ تروریستی، یک اثر ادبی و یا یک کتاب یا مقالهٔ علمی باشد. یکی از رویکردهای مهم در زبانشناسی رایانشی رویکرد آماری است که با استفاده از داده‌های بزرگ، الگوهای پرتکرار به کارگرفته‌شده در متن کشف می‌شود، سبک نگارش نویسنده متن هویدا می‌گردد و بدین ترتیب سبک نگارش نویسندگان مختلف تشخیص داده می‌شود. به این کار اصطلاحاً سبک‌سنجی رایانشی[۱] می‌گویند. سبک‌سنجی رایانشی بر این فرض استوار است که هر فرد یک گویش فردی[۲] و یا سبک خاص و منحصربه‌فرد در گفتار یا نوشتار دارد. در حقیقت هنگامی‌که مردم از زبان استفاده می‌کنند، ساخت‌های خاصی از واژگان–دستور[۳] را برمی‌گزینند که سایر افراد ممکن است از آنها کمتر استفاده نمایند. در مرحله بعد هر فرد این ساخت‌ها را به شکل متفاوتی از دیگر افراد ترکیب می‌کند تا پیام خود را منتقل کند و این به معنای منحصربه‌فرد بودن گویش فردی و سبک گفتار یا نوشتار افراد بشر است (کلتارد[٤]، ۲۰۰٤).

تعیین سبک نگارش یا انتساب نویسندگی[۵] عبارت است از تشخیص خودکار هویت نویسندهٔ[٦] یک متن که فاقد نام نویسنده است بر اساس مشابهت‌های سبکی میان آثار یک نویسندهٔ شناخته‌شده و متن فاقد نام نویسنده (سانگ[۷] و بروک[۸]، ۲۰۰۹). پژوهش حاضر تلاش دارد به بررسی تعیین سبک نگارش متون داستانی زبان فارسی بپردازد. برای این منظور پیکره‌ای از آثار هفت نویسندهٔ مرد هم‌عصر به نام‌های هوشنگ گلشیری، بزرگ علوی، احمد محمود، محمود دولت‌آبادی، نادر ابراهیمی، جلال آل احمد و غلامحسین ساعدی تهیه و سپس نقش واژه‌های دستوری و محتوایی در سبک‌سنجی رایانشی نثر فارسی معاصر بررسی و مقایسه شده است.

۲. مبانی نظری

واژه‌های درون یک متن به طور تصادفی دور هم جمع نشده‌اند. متون ژانرهای مختلف به لحاظ نوع واژه‌هایی که در آنها به کار می‌روند، با هم متفاوتند. در هر متن، تعدادی واژهٔ مشخص قرار دارند که نمایندهٔ آن متن هستند. این واژه‌ها که واژه‌های محتوایی، واژه‌های کلیدی یا کلیدواژه

1. computational stylometry
2. ideolect
3. lexicogrammar
4. Coulthard
5. authorship attribution
6. Author identification
7. M. Song
8. Y. Brook

نامیده می‌شوند، گویای محتوای درونی متن و همچنین اندیشهٔ صاحبان آن اثر است (میرزایی، ۱۳۹۷). بنا بر تعریف، واژه‌های محتوایی واژه‌هایی هستند که معنای مستقل دارند و عمدتاً شامل اسم، صفت و فعل می‌شوند (میرزایی و صفری، ۱۳۹٤).

در مقابل واژه‌های محتوایی، واژه‌های دستوری قرار دارند. واژه‌های دستوری واژه‌هایی هستند که معنای واژگانی محدود دارند و روابط دستوری میان اجزای جمله را بیان می‌کنند (کلامر[1] ودیگران، ۲۰۰۹). تعیین مرز دقیق میان واژه‌های محتوایی و دستوری عملاً امکان‌پذیر نیست. در واقع واژه‌های دستوری و محتوایی در دو سوی یک پیوستار قرار دارند و در میانهٔ این پیوستار واژه‌هایی وجود دارند که نمی‌توان به طور قطعی آنها را در یک طبقهٔ مشخص قرار دارد. بنابر تعریف ردفورد[2] (۲۰۰٤)، واژه‌های دستوری حاوی اطلاعاتی در خصوص نقش‌های دستوری منتسب به برخی مفاهیم (چون شخص، شمار، جنس، حالت و مانند آن) هستند و معنای قائم‌به‌ذات ندارند (میرزایی و صفری، ۱۳۹٤).

در این پژوهش در تعریف واژه‌های دستوری ملاک پربسامدبودن واژه در نظر گرفته شده است. درحقیقت، منظور از واژه‌های دستوری در پژوهش حاضر، فهرستی از واژه‌های پربسامد در پیکره‌های زبانی است که در پژوهش‌های حوزهٔ زبانشناسی رایانشی[3] تحت عنوان ایست‌واژه[4] از آنها یاد می‌کنند. در مجموع، در پژوهش حاضر حرف اضافه، حرف ربط، کمّیت‌نما، ضمیر، فعل کمکی، همکرد فعل، متمم‌نمای «که»، حرف تعریف «یه» و نقش نمای مفعولی «را» کلمات دستوری به شمار می‌آیند.

محاسبه بسامد واژه‌های دستوری یک روش معتبر در سبک‌سنجی رایانشی است. واژه‌های دستوری به این دلیل که توسط نویسنده به طور ناخودآگاه به کار گرفته می‌شوند و به طور مکرر در متن حضور دارند، در سبک‌سنجی رایانشی بسیار مورد توجه هستند (گلشائی، ۱۳۹۸). پژوهش حاضر تلاش دارد روش محاسبهٔ بسامد واژه‌های دستوری را مبنای مقایسه قرار دهد و کارآیی واژه‌های محتوایی را در تعیین سبک نگارش متون فارسی محک بزند.

۳. پژوهش‌های پیشین

تحلیل مبتنی بر پیکره روش سودمندی است که به واسطه آن می‌توان به بررسی بازنمود ایدئولوژی در صورت‌های زبانی در حوزه‌ای چون زبان‌شناسی پرداخت. متن‌های ادبی نیز اعم از رمان‌ها و سفرنامه‌هایی که توسط نویسندگان شکل می‌گیرد، نشأت گرفته از ایدئولوژی نهفته

1. T. Klammer
2. A. Radford
3. computational linguistics
4. Stop word

ویژه‌ای است تا مخاطب را به سمت آنچه مدنظر نویسنده است سوق دهد (کمال‌پور و دیگران، ۱٤۰۱). همچنین باورها و ایدئولوژی‌های مختلف در زبان و در متون با استفاده از واژه‌های محتوایی یا کلیدواژه‌ها بازنمایی می‌شوند(میرزایی، ۱۳۹۷). به همین سبب، برخی پژوهش‌های پیکره‌بنیاد از واژه‌های محتوایی در حوزه‌های نقد ادبی، تحلیل گفتمان انتقادی[1] و زبان‌شناسی اجتماعی استفاده کرده‌اند. برای مثال، پژوهش کمال‌پور و دیگران (۱٤۰۱) به بررسی سفرنامه و رمان‌های «غرب‌زدگی، خسی در میقات، مدیر مدرسه، اورازان، تات‌نشین‌های بلوک زهرا» از جلال آل احمد و «مدار صفر درجه، غریبه‌ها، پسرک بومی و همسایه‌ها» از احمد محمود پرداخته است. این پژوهش درصدد بوده است تا با تکیه بر کلیدواژه‌های موجود در آثار این دو نویسنده و با بهره‌گیری از رویکرد نظری بیکر[2] (۲۰۰٦)، به بررسی فضای سیاسی و اجتماعی حاکم در آثار آنها بپردازد. این پژوهش با استفاده از ۳ کلیدواژهٔ واژگانی که از فراوانی زیادی در هر اثر برخوردار بودند، مشخص کرد که رویکرد نویسندگان بیانگر چه نوع ویژگی‌های ایدئولوژیک بوده است. نتایج حاصل نشان دهنده نگرش و ایدئولوژی آسیب شناختی جلال آل احمد و احمد محمود در جوامع انسانی بود.

پژوهشی دیگر (امیری و دیگران، ۱۳۹٦) در حوزهٔ تحلیل پیکره‌بنیاد ترانه‌های فارسی انجام شده‌است. هدف این پژوهش تحلیل و مقایسهٔ ترانه‌های فارسی موسیقی پاپ بوده و تحلیل آن بر اساس الگوی بیکر (۲۰۰٦) صورت گرفته است. جامعهٔ آماری آن ترانه‌های فارسی موسیقی پاپ در دو ژانر اجتماعی و عاشقانه است. نتایج پژوهش نشان می‌دهند که ترانه‌های عاشقانه و اجتماعی به لحاظ واژه‌های پربسامد، کلیدواژه‌ها و باهم‌آیی‌ها تفاوت معناداری با یکدیگر دارند. جنسیت ترانه‌سرا بر گزینش‌های واژگانی تأثیر معناداری دارد. بررسی واژگانی ترانه‌های پیش و پس از انقلاب حکایت از معناداری تأثیر متغیر زمان ترانه بر واژه‌های دو پیکره دارد.

هوور[3] (۲۰۰۷) و تولان[4] (۲۰۰۸) اشاره می‌کنند که رویکردهای کمی پیکره‌بنیاد طبیعتاً با مسألهٔ سبک و نویسندگی در ارتباط است. باروز[5] (۱۹۹۲: ۲۰٤–۱٦۷؛ ۲۰۰۳: ۵–۳۲؛ ۲۰۰۷: ۲۷–٤۷) واژه‌های دستوری را بهترین شاخص سبک نویسندگی می‌داند. هوور (۱۹۹۹) مشخصه‌های سبکی و ادبی را در سه بخش از «وارثان» اثر گلدلینگ بررسی کرده و قابلیت خوانش، پیچیدگی جمله و غنای واژه‌های آن را به لحاظ آماری بررسی می‌کند. با توجه به سهم نویسندگی و با فرض باروز (۱۹۸۷) که واژه‌های دستوری با فراوانی زیاد نسبت به واژه‌های محتوایی، اغلب

1. Critical discourse analysis
2. P. Baker
3. D.L. Hoover,
4. M. J. Toolan
5. J. F. Burrows

شاخص‌های بهتری از تمایزات سبکی و ادبی است، هوور (۲۰۰۳) تمایزات ادبی و سبکی میان رمان‌های نوشته‌شده و ویژگی‌های سبک نویسندگی آنها را تشخیص می‌دهد. آرگامون[۱] و لویتان[۲] (۲۰۰۵) یک تحلیل آماری از واژه‌های دستوری در رمان‌های مشابهی که توسط هوور (۲۰۰٤) بررسی شده‌اند، انجام می‌دهند. نتایج پژوهش آنها نشان می‌دهد که واژه‌های دستوری نسبت به باهم‌آیندها[۳] و جفت‌واژه‌ها شاخص‌های بهتری از سبک نویسنده است (کمال‌پور و دیگران، ۱۳۹۹).

تحقیقات بسیاری از محاسبهٔ بسامد واژه‌های دستوری در تشخیص خودکار نویسندهٔ متن استفاده کرده‌اند (آرگامون و لویتان، ۲۰۰۵؛ کستمونت[٤]، ۲۰۱٤؛ سگارا[٥] و دیگران، ۲۰۱۵؛ زائو[٦] و زوبل[۷]، ۲۰۰۵).

در چندین پژوهش از تعداد تکرار باهم‌آیی‌های دوتایی[۸]، سه‌تایی[۹] و چهارتایی[۱۰] کلمات در متن استفاده شده است (هاواردز[۱۱] و استاماتاتوس[۱۲]، ۲۰۰٦؛ کشلج[۱۳] و دیگران، ۲۰۰۳؛ ساری[۱٤] و دیگران، ۲۰۱۷).

بایرامی (۲۰۲۱)، با استفاده از معیار بسامد کلمه – معکوس بسامد سند[۱۵] (شوتز[۱٦] و دیگران، ۲۰۰۸)، واژه‌های محتوایی هر متن را استخراج کرده و با مقایسه اسناد مشابه به لحاظ کلمات محتوایی، به کمک روش یادگیری ماشینی و الگوریتم جنگل تصادفی[۱۷] به کارایی خوبی در تشخیص نگارش متن دست می‌یابد. پژوهش الحقیل[۱۸] (۲۰۲۱)، دو روش مدل بسته کلمات[۱۹] و تحلیل معنایی پنهان[۲۰] را برای سنجش مشابهت بین اسناد، با یکدیگر مقایسه کرده است. در روش مدل بسته کلمات هر سند به‌صورت یک ماتریس نمایش داده می‌شود که بیانگر تعداد

1. S. Argamon
2. S. Levitan
3. collocate
4. A. Kestemont
5. S. Segarra
6. Y. Zhao
7. J. Zobel
8. bigram
9. trigram
10. quadgram
11. J. Houvardas
12. E. Stamatatos
13. V. Kešelj
14. Y. Sari
15. Term frequency-Inverse document frequency (tf-idf)
16. H. Schütze
17. Random forest
18. N. Alhuqail
19. Bag-of-words
20. Latent semantic analysis

تکرار هر کلمه از مجموعه واژگان در یک سند است. تحلیل معنایی پنهان نیز مشابهت معنایی بین اسناد و کلمات را بررسی می‌کند. تحلیل معنایی پنهان بر این فرض استوار است که کلمات به لحاظ معنایی مشابه در اسناد مشابهی یافت می‌شوند. نتایج پژوهش حاکی از این است که روش مدل بسته کلمات در قیاس با روش تحلیل معنایی پنهان عملکرد بهتری در سبک‌سنجی رایانشی داشته است. پژوهش کیان[1] و دیگران (۲۰۱۷) با کمک روش‌های مبتنی بر شبکهٔ عصبی بازگشتی[2] همچون شبکهٔ عصبی واحد بازگشتی گیتی[3] و حافظهٔ طولانی کوتاه‌مدت، میزان مشابهت بین اسناد را بررسی و نویسندهٔ متن را تشخیص می‌دهد و به کارایی مناسبی در تشخیص هویت نویسنده دست می‌یابد.

در پژوهش رمضانی (۲۰۲۱)، یک روش مستقل از زبان در تشخیص سبک نگارش متون پیشنهاد شده است. در این روش از مجموعه از ویژگی‌های زبانی چون بسامد واژه‌ها، بسامد برچسب اجزای سخن[4]، باهم‌آیی‌های کلمات، طول کلمات، طول جملات، باهم‌آیی‌های حروف و غیره برای تعیین سبک نگارش متن استفاده می‌شود. در مورد بسامد واژه‌ها، واژه‌های دستوری کنار گذاشته شده و با استفاده از معیار بسامد کلمه – معکوس بسامد سند، کلمات محتوایی متن شناسایی می‌شود و میزان مشابهت بین اسناد با کمک کلمات به دست می‌آید و نویسندهٔ متن تشخیص داده می‌شود. نتایج ارزیابی روش پیشنهادی حاکی از این است که بسامد واژه‌های محتوایی، بسامد برچسب اجزای سخن و باهم‌آیی‌های دوتایی بیشترین کارایی را در میان ویژگی‌های زبانی داشته‌اند. این پژوهش، کارایی بالایی از روش پیشنهادی را در مورد زبان فارسی و انگلیسی گزارش کرده است.

از ویژگی‌های نحوی بسیار مهم و پرکابرد در تعیین سبک نگارش می‌توان به توالی کلمات با برچسب اجزای سخن خاص اشاره کرد. برای تعریف این ویژگی از توالی برچسب‌های مختلف اجزای کلمات به‌عنوان معیاری برای متمایز ساختن سبک نویسندگان مختلف بهره گرفته می‌شود. همچنین بسامد هر یک از مقوله‌های اسم، صفت، فعل، حرف‌اضافه و غیره نوع دیگری از ویژگی‌هاست که در پژوهش‌های چندی برای تعیین خودکار نویسنده به کار گرفته شده است (گامون[5]، ۲۰۰٤؛ کوکشکینا[6] و دیگران، ۲۰۰۱؛ زائو و زوبل، ۲۰۰۷). در این پژوهش‌ها با استفاده از یک تجزیه‌گر نحوی مبتنی بر دستور ساخت سازه‌ای می‌توان جملات یک متن را تجزیه کرده

1. C. Qian
2. Recurrent neural network
3. Gated Recurrent Unit
4. Parts-of-Speech gram
5. M. Gamon
6. O. Kukushkina

و بسامد قواعد بازآرایی را محاسبه کرده و به‌عنوان یک ویژگی در نظر گرفت. همچنین از قطعه‌بند[1] استفاده می‌شود تا مرز گروه‌های نحوی را تشخیص داده شوند و بدین ترتیب بسامد گروه‌های نحوی مختلف و میانگین طول گروه‌های اسمی، فعلی، صفتی و غیره را به‌عنوان ویژگی لحاظ کند (گامون، ۲۰۰٤). استفاده از روابط وابستگی میان کلمات بر اساس دستور وابستگی از دیگر ویژگی‌های نحوی در تعیین خودکار نویسنده محسوب می‌شود. در این روش ابتدا درخت وابستگی برای هریک از جملات متن توسط تجزیه‌گر وابستگی استخراج شده و سپس بسامد هر نوع رابطه وابستگی خاص در یک جمله محاسبه می‌شود. برای مثال نسبت فعل‌های وجهی به کل فعل‌های موجود در متن، درصد رخداد روابط قیدی و توصیفگر[2] به‌عنوان ویژگی‌های نحوی در نظر گرفته شده است (وانر[3]، ۲۰۱۷).

برخی مطالعات از معناشناسی قالب‌بنیاد[4] در شناسایی خودکار نویسنده استفاده کرده‌اند. در پژوهش هدگارد[5] و سیمونسن[6] (۲۰۱۱)، از روابط معنایی تعریف‌شده بر اساس شبکهٔ قاب‌ها در تشخیص نویسنده متون اصلی و متون ترجمه‌شده استفاده شده است اما به‌دلیل پیچیدگی روابط معنایی و خطای زیاد تجزیه‌گر معنایی نتایج پژوهش در حد مطلوب نبوده و کارایی ویژگی‌های معنایی از ویژگی‌های واژگانی و نحوی کمتر بوده است. پژوهش (گامون،۲۰۰٤) از گراف وابستگی معنایی جملات استفاده کرده و اطلاعاتی از قبیل شخص، شمار، زمان و نمود افعال را در تشخیص خودکار هویت نویسنده استفاده کرده است. همچنین از روابط معنایی بین کلمات در گراف استفاده کرده است. نتایج پژوهش حاکی از این است که هنگامی که ویژگی‌های معنایی، واژگانی و نحوی با یکدیگر ترکیب می‌شوند دقت سامانه تشخیص سبک نویسنده ارتقا یافته است. پژوهش (مک کارتی[7] و دیگران، ۲۰۰٦) از شبکهٔ واژگان[8] و اطلاعاتی نظیر ترادف استفاده کرده‌اند. علاوه بر آن آنها از تحلیل معنایی پنهان استفاده کرده‌اند که مشابهت معنایی بین کلمات را به‌صورت خودکار محاسبه کرده و به‌عنوان ویژگی لحاظ می‌کند.

از ساخت گفتمانی متن نیز برای تعیین سبک نویسنده استفاده می‌شود. برای این منظور متن توسط تجزیه‌گر گفتمانی به واحدهای گفتمانی پایه[9] تقسیم شده و از طریق روابط گفتمانی بین

1. Chunker
2. Modifier
3. L. Wanner
4. frame semantics
5. S. Hedegaard
6. J. Simonsen
7. P. McCarthy
8. WordNet
9. elementary discourse unit

آنها پیوند برقرار می‌شود. درنهایت بسامد هر رابطهٔ گفتمانی به کل واحدهای گفتمانی پایه محاسبه می‌شود (وانر، ۲۰۱۷).

در حوزهٔ تشخیص سبک نگارش در حوزهٔ زبان فارسی چندین پژوهش انجام شده است. دباغ (۲۰۰۷) بر اساس واژه‌های دستوری پربسامد و با استفاده از روش‌های آماری به تفکیک سبک نگارش نظامی گنجوی/ شهریار و عبدالحسین زرین‌کوب/ سیمین دانشور پرداخته است. گلشائی (۱۳۹۸) با تکیه بر گویش فردی و با استفاده از واژه‌های دستوری زبان فارسی تشخیص هویت نویسنده متن را برای پنج نویسنده معاصر بررسی کرده است. نتایج پژوهش وی نشان می‌دهد که نویسنده‌های مختلف واژه‌های دستوری را به طور مشابه به کار نمی‌گیرند. درواقع با وجود اینکه برخی واژه‌های دستوری پربسامد توسط همه نویسنده‌ها به کارگرفته می‌شود اما اولویت به‌کارگیری آنها توسط نویسندگان مختلف متفاوت است. در پژوهش دیگر ویژگی‌های مؤثر در شناسایی نویسنده در متون برخط فارسی بررسی شده است. یافته‌های پژوهش حاکی از آن است که ویژگی‌های نگارشی (بسامد نسبی علائم نگارشی) بیشترین تأثیر را در شناسایی متون کوتاه داشته است (عارفی و دیگران، ۱٤۰۰).

در پژوهش آذین و بحرانی (۱۳۹۳)، هدف اصلی تعیین کارامدترین ویژگی‌های سبکی در متون فارسی و کمّی‌سازی آن‌ها برای استفاده در سامانه‌های شناسایی شاعر بوده است. به این منظور، تأثیر ویژگی‌های سبکی آثار چهار شاعر شعر نو (مهدی اخوان ثالث، نیما یوشیج، احمد شاملو و سهراب سپهری) در سه سطح واژگانی، نحوی و حرفی بررسی شده است. با بررسی نتایج دسته‌بندی‌ها در هر سطح مشخص شد که ویژگی‌های نحوی از کارایی بالاتری نسبت به ویژگی‌های واژگانی و حرفی در دسته‌بندی شاعران برخوردارند. در پژوهش دیگر (جوانمردی و اکبری، ۱۳۹۷)، هدف اصلی استخراج ویژگی‌های صوری متن و دسته‌بندی اشعار مربوط به دو شاعر حوزهٔ دفاع مقدس (قیصر امین‌پور و محمدرضا عبدالملکیان) بوده است. نتایج ارزیابی‌ها نشان داده است که ویژگی‌های واژگانی بدون حذف واژه‌های دستوری در میان انواع ویژگی‌ها از بالاترین دقت برخوردار بوده است. این نتیجه، نشان‌دهنده کارآیی قابل‌ملاحظهٔ این نوع ویژگی در شناسایی سبک نویسنده بوده است.

رهگذر[1] در رسالهٔ دکتری خود (رهگذر، ۲۰۲۰) پژوهشی در خور توجه و متفاوت در حوزهٔ تحلیل شعر حافظ انجام داده‌است. موضوع پژوهش وی طبقه‌بندی خودکار اشعار حافظ و تحلیل معنایی درزمانی[2] آنها بوده است. هدف نهایی این پژوهش طبقه‌بندی خودکار اشعار دیوان حافظ جهت تصحیح اشعار حافظ در نسخ مختلف و تأیید هویت شاعر اشعاری است که به حافظ

1. A. Rahgozar
2. Chronological

منتسب شده است. رهگذر پژوهش خود را براساس تقسیم‌بندی شش‌گانهٔ محمود هومن از دوران زندگی حافظ (هومن، ۱۳۵۷) که شامل دوران جوانی، پس از جوانی، بلوغ فکری، میانسالی، پیش از سالمندی و سالمندی است، انجام داده است. وی پس از پیش‌پردازش اشعار، از روش وزن‌دهی بسامد کلمه–معکوس بسامد سند استفاده و کلمات را امتیازدهی کرده است. سپس با استفاده از روش تحلیل معنایی پنهان مشابهت معنایی بین اسناد و کلمات را پیدا و با روش ماشین بردار پشتیبان اسناد را طبقه‌بندی کرده است. تحلیل‌های معنایی پنهان در حوزهٔ نظم در مقایسه با نثر به دلیل عنصر خیال، ایجاز و استعاره در شعر پیچیده‌تر و چالش‌برانگیزتر است. تحلیل معنایی پنهان بر این فرض استوار است که کلمات به لحاظ معنایی مشابه در اسناد مشابهی یافت می‌شوند. برای مثال کلماتی مانند «دل»، «عشق»، «مرا»، «چشم»، «غم»، «نظر»، «شمع»، «آتش»، «خرابات»، «حافظ» و غیره، کلمات مرتبط با یکدیگر هستند که در دورهٔ میانسالی حافظ پرکاربرد بودند. پژوهش رهگذر نشان می‌دهد که اشعار موجود در یک طبقه (یک بازهٔ زمانی از دوران زندگی حافظ) به لحاظ موضوع و مضمون شعر به یکدیگر نزدیک‌تر هستند تا به شعری در طبقهٔ دیگر (رهگذر، ۲۰۲۰).

با توجه به فقدان تمرکز پژوهش‌های تشخیص هویت نویسنده در نثر فارسی بر واژه‌های محتوایی (گلشائی، ۱۳۹۸؛ عارفی و دیگران، ۱۴۰۰؛ دباغ[1]، ۲۰۰۷)، در پژوهش حاضر تلاش بر آن است که کارایی روش واژه‌های محتوایی در مقایسه با روش واژه‌های دستوری سنجیده شود. به‌علاوه، پژوهش‌های انجام‌شده در حوزهٔ تشخیص خودکار نویسنده در زبان فارسی، برای واژه‌های دستوری دسته‌بندی معنایی ارائه ننموده‌اند. در این پژوهش سعی شده است که این خلأ پر شود.

پرسش این پژوهش عبارت است از: در تشخیص سبک نگارش متون داستانی معاصر، واژه‌های دستوری مؤثرتر هستند یا واژه‌های محتوایی؟ در راستای پاسخ به این پرسش، در پژوهش حاضر، هدف طراحی سامانه‌ای است که از میان هفت نویسندهٔ داستانی معاصر، یک نویسنده را به عنوان نویسندهٔ متن پیشنهاد دهد. به منظور دست‌یابی بدین هدف، پیکره‌ای از آثار نویسندگان معاصر ایرانی گردآوری شد. برای انتخاب نویسندگان، ابتدا تعداد آثار هر نویسنده در پیکرهٔ بیجن‌خان (بیجن‌خان و دیگران، ۲۰۱۱) جستجو شد و نویسندگانی با بیشترین تعداد آثار در پیکره که هم‌عصر و با جنسیت یکسان بودند، یافت شد؛ به این ترتیب هفت نویسندهٔ مرد هم‌عصر به نام‌های هوشنگ گلشیری، بزرگ علوی، احمد محمود، محمود دولت‌آبادی، نادر ابراهیمی، جلال آل احمد و غلامحسین ساعدی انتخاب شد. سپس دیگر آثار این نویسندگان در

1. R. Dabagh

پایگاه داده‌های زبان فارسی (عاصی، ۱۹۹۷) و وب مورد جستجو قرار گرفت و به متون پیکره اضافه شد.

در این پژوهش برای واژه‌های دستوری، دسته‌بندی معنایی ارائه شد و فهرست هر کدام از انواع واژه‌های دستوری گردآوری شد. سپس از روش شبکهٔ عصبی مصنوعی برای مرحلهٔ یادگیری ماشینی استفاده و سبک نگارش متن به طور خودکار تشخیص داده شد. برای واژه‌های محتوایی نیز از الگوریتم حافظهٔ طولانی کوتاه‌مدت[1] استفاده شد.

٤. روش پژوهش

نقش واژه‌های دستوری و محتوایی در متمایز ساختن سبک نگارش نویسندگان پیکرهٔ گردآوری شده این پژوهش در دو مرحلهٔ مجزا موردبررسی قرار گرفت.

٤-١. واژه‌های دستوری

فهرست واژه‌های دستوری و محتوایی وابسته به حوزهٔ تخصصی یا ژانر هر متن است (میرزایی و صفری، ١٣٩٤)، به همین سبب برای تهیهٔ فهرست واژه‌های دستوری، از پیکرهٔ گردآوری شده در پژوهش حاضر استفاده شد. یکی از راهکارهای مطرح در استخراج واژه‌های دستوری، روش وزن‌دهی بسامد کلمه – معکوس بسامد سند است. در این روش میزان تکرار یک کلمه در یک سند در مقابل تعداد تکرار آن در مجموعه کلیه سندها سنجیده می‌شود. در این روش به هر کلمه در یک متن عددی تخصیص می‌دهد. این عدد برای کلماتی دارای مقادیر بیشینه است که در مجموعه متون پیکره در تعداد اسناد کمتری اما به‌وفور یافت شود. این کلمات، واژه‌های کلیدی یا محتوایی متن هستند. از سوی دیگر، این معیار برای کلماتی که تقریباً در تمامی اسناد پیکره یافت شوند، دارای مقادیر کمینه است. این کلمات همان واژه‌های دستوری تلقی می‌شوند (شوتز و دیگران، ۲۰۰۸). در پژوهش حاضر، بر اساس معیار بسامد کلمه – معکوس بسامد سند، ۲۲۷ واژه‌ای که دارای مقادیر کوچک‌تری بودند، به‌عنوان واژهٔ دستوری انتخاب شد.

بر اساس هر یک از واژه‌های دستوری، یک ویژگی در بردار ویژگی برای مرحلهٔ یادگیری ماشینی استخراج شد. این ویژگی بیانگر بسامد آن واژهٔ دستوری به کل واژه‌های دستوری در آن متن است. برای محاسبهٔ بسامد نسبی واژه‌های دستوری از این فرمول استفاده شده است:

1. Long Short Term Memory (LSTM)

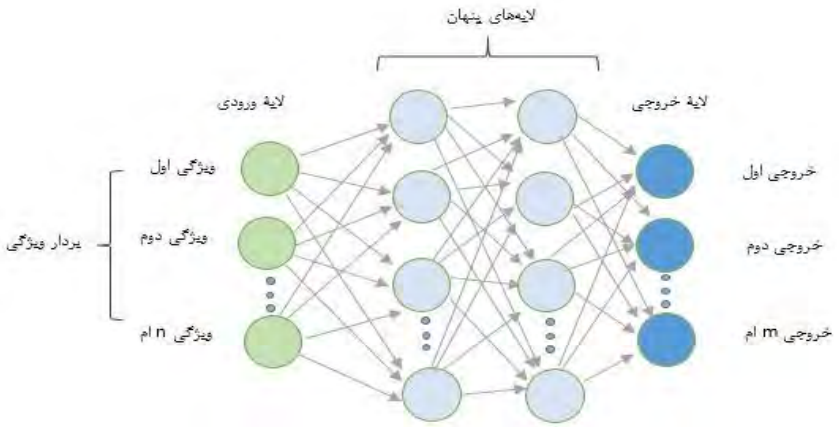
فرمول (۱)

$$\frac{\text{count}(w)}{\sum_{w' \in FW} \text{count}(w')}$$

در این فرمول w هر یک از واژه‌های دستوری است که بسامد آن نسبت به بسامد کل واژه‌های دستوری در یک متن خاص محاسبه می‌شود (آرگامون و دیگران، ۲۰۰۷). در صورت کسر تعداد رخداد یک واژهٔ دستوری خاص و در مخرج آن تعداد رخداد کل واژه‌های دستوری در متن قرار دارد.

برای مرحله یادگیری ماشینی در استفاده از واژه‌های دستوری از روش یادگیری با ناظر[1] استفاده می‌شود. در این روش به یک سیستم، مجموعه‌ای از جفت‌های ورودی – خروجی ارائه شده و سیستم تلاش می‌کند تا تابعی از ورودی به خروجی را فراگیرد. یادگیری باناظر نیازمند تعدادی داده ورودی به‌منظور آموزش سیستم است. یکی از انواع طبقه‌بندها[2]، طبقه‌بند پرسپترون چندلایه[3] است که از آن در این پژوهش بهره گرفته شده است. شبکه عصبی پرسپترون چندلایه، دسته‌ای از شبکه‌های عصبی مصنوعی پیش‌خور[4] محسوب می‌شود. در یک شبکه عصبی پرسپترون چندلایه، حداقل سه لایه از گره‌ها وجود دارد که این لایه‌ها عبارتند از: یک لایه ورودی، یک لایه پنهان و یک لایه خروجی. لایهٔ اول (ورودی) حاوی بردار ویژگی است که در این پژوهش همان بسامد نسبی هر یک از واژه‌های دستوری است. در اینجا بسامد نسبی ۲۲۷ واژه‌ای که واژهٔ دستوری محسوب می‌شوند، به عنوان ورودی دریافت می‌شود. در این نوع از شبکه عصبی مصنوعی، از خروجی‌های لایه اول (ورودی)، به عنوان ورودی‌های لایه بعدی (پنهان) استفاده می‌شود؛ این کار به همین شکل ادامه پیدا می‌کند تا زمانی که پس از تعداد خاصی از لایه‌ها، خروجی‌های آخرین لایه پنهان به عنوان ورودی‌های لایه خروجی مورد استفاده قرار می‌گیرد. همه لایه‌هایی که بین لایه ورودی و لایه خروجی قرار می‌گیرند، «لایه‌های پنهان» نامیده می‌شوند. در شکل ۱ تصویری از یک شبکهٔ عصبی پرسپترون چندلایه نمایش داده شده است.

1. supervised learning
2. classifier
3. multilayer perceptron
4. Feedforward Neural Networks



شکل ۱. نمایی از یک شبکهٔ عصبی پرسپترون چندلایه

در این پژوهش یک لایهٔ پنهان با هزار نورون[1] (سلول عصبی) در نظر گرفته شده است. شبکه‌های پرسپترون چند لایه حاوی مجموعه‌ای از وزن‌ها نیز هستند که باید برای آموزش و یادگیری شبکه عصبی تنظیم شوند. این طبقه‌بند در نهایت از میان هفت نویسنده، یک نویسنده را به عنوان خروجی برمی‌گرداند.

در این پژوهش برای استفاده از طبقه‌بند پرسپترون چند لایه از یک ابزار[2] یادگیری ماشین به زبان پایتون[3] استفاده شده است. همچنین از تابع فعالسازی[4] واحد یکسو شده‌ی خطی[5] و الگوریتم بهینه‌سازی آدام[6] استفاده شده است. تابع فعالسازی در مورد فعالسازی هر نورون تصمیم‌گیری می‌کند. الگوریتم آدام یک الگوریتم بهینه‌سازی است که برای به‌روزرسانی وزن‌های شبکه بر اساس تکرار در داده‌های آموزشی استفاده می‌شود. همچنین بیشینه تعداد تکرارها[7] در این پژوهش برابر با یک میلیون تکرار در نظر گرفته شده است.

٤-۲. واژه‌های محتوایی

در پژوهش حاضر از شبکهٔ حافظهٔ طولانی کوتاه‌مدت دوطرفه[8] برای بررسی تأثیر واژه‌های محتوایی در تشخیص سبک نگارش متن استفاده شده است. این نوع شبکه، یک نسخه بهبود

1. neuron
2. https://scikit-learn.org/
3. python
4. Activation Function
5. Rectified Linear Unit (ReLU)
6. adam
7. maximum iteration (max_iter)
8. Bidirectional Long short-term memory (BLSTM)

یافته از شبکه‌های عصبی بازگشتی است که باعث می‌شوند به خاطر سپردن داده‌های گذشته در حافظه، آسان‌تر شود. ازآنجاکه توالی واژه‌های محتوایی می‌تواند در تشخیص سبک نگارش مؤثر باشد، از این نوع شبکهٔ عصبی حافظه طولانی کوتاه‌مدت استفاده شده است. شبکهٔ حافظهٔ طولانی کوتاه‌مدت دوطرفه مدل گسترش یافتهٔ شبکهٔ حافظهٔ طولانی کوتاه‌مدت است که برخلاف شبکهٔ حافظهٔ طولانی کوتاه‌مدت استاندارد، ورودی در هر دو جهت جریان دارد و می‌تواند از اطلاعات هر دو طرف استفاده کند. همچنین ابزار قدرتمندی برای مدل سازی وابستگی‌های متوالی بین ورودی‌ها در هر دو جهت دنباله است. در شبکهٔ حافظهٔ طولانی کوتاه‌مدت دوطرفه به جای آموزش یک مدل، دو مدل وجود دارد. مدل اول دنباله ورودی ارائه شده را می‌آموزد و مدل دوم عکس آن دنباله را می‌آموزد. شبکهٔ حافظهٔ طولانی کوتاه‌مدت دوطرفه نوعی شبکهٔ حافظهٔ طولانی کوتاه‌مدت دو لایه است، اما جهت لایه‌ها کاملاً مخالف یکدیگر هستند. به این معنا که دنبالهٔ ورودی از دو جهت روبه جلو[1] و روبه عقب[2]، در یک زمان یکسان عبور داده می‌شوند (بابانژاد باقری و دیگران ، ۱٤۰۲).

برای به‌کارگیری الگوریتم شبکهٔ عصبی حافظه طولانی کوتاه‌مدت دوطرفه از کتابخانهٔ یادگیری عمیق کرس[3]، از ابزار تنسورفلو[4] و زبان برنامه‌نویسی پایتون استفاده شده است. برای این کار از هر متن، علائم نگارشی و واژه‌های دستوری حذف می‌شود و باقی متن به عنوان ورودی به شبکه داده می‌شود. این ابزار قابلیت این را دارد که واژه‌های محتوایی پرکاربرد در این پیکره را به صورت خودکار استخراج کرده و از آن در جهت یادگیری سامانه استفاده نماید. حجم مجموعه واژگان محتوایی در این سامانه برابر با پنج هزار کلمه تعریف شده است. همچنین تابع فعالسازی بیشینه هموار[5]، الگوریتم بهینه‌سازی آدام و روش پس‌کوتاه‌سازی[6] و پس‌لایه‌گذاری[7] برای این سامانه در نظر گرفته شده است. الگوریتم درنهایت بر اساس مجموعه واژه‌های محتوایی یافت شده در هر متن و با توجه به توالی رخداد آنها در متن، اسناد مشابه به یکدیگر را می‌یابد و نویسندهٔ متن را پیش‌بینی می‌کند.

بیان یک نکته در اینجا حائز اهمیت است. در فرآیند یادگیری ماشینی گاه بیش‌برازش[8] اتفاق می‌افتد. بیش‌برازش به این معناست که الگوریتم داده‌های آموزشی را بسیار خوب یادگرفته است

1. Forward
2. Backward
3. Keras
4. tensorflow
5. softmax
6. Post truncation
7. Post padding
8. overfitting

ولی قابلیت پیش‌بینی خوب داده‌های جدید را ندارد. یکی از دلایل این مشکل می‌تواند این باشد که حجم داده برای یادگیری ماشینی کافی نیست. خوشبختانه، پیکرهٔ تهیه شده در این پژوهش از حجم بسیار خوبی برای مرحلهٔ یادگیری ماشینی برخوردار بوده است. در هر حال، برای پیش‌گیری از بیش‌برازش احتمالی از روش حذف تصادفی[1] استفاده شده است. در حقیقت این روش، روشی برای منظم‌سازی[2] است. در این روش در هر دور آموزشی، به جای استفاده از همه نورون‌ها، تنها برخی از گره‌ها با احتمال خاصی فعال می‌شوند. این مدل، همه لایه‌های یک شبکه عصبی را می‌پیماید و احتمالی برای حفظ یا حذف گره‌ها در نظر می‌گیرد. البته لایه ورودی و خروجی بطور مشابه حالت اولیه حفظ می‌شوند. احتمال حفظ هر گره به صورت تصادفی تعیین و فقط آستانه‌ای برای آن درنظر گرفته می‌شود. در این پژوهش، آستانه ۰/۲ برای حذف تصادفی در دو لایهٔ شبکه عصبی در نظر گرفته شده است.

۵. ارزیابی پژوهش

به منظور ارزیابی کیفیت سامانهٔ طراحی‌شده از روش ارزیابی متقاطع K تایی[3] استفاده می‌شود. اگر مجموعه داده‌های آموزشی را به طور تصادفی به K زیرنمونه تفکیک کنیم، می‌توان در هر مرحله از فرایند، تعداد K-1 از این لایه‌ها را به عنوان مجموعه داده آموزشی و یکی را به عنوان مجموعه داده آزمون در نظر گرفت. مشخص است که تعداد تکرارهای فرآیند برابر با K خواهد بود (رفائیل‌زاده و دیگران، ۲۰۰۹). در این پژوهش، K=۵ در نظر گرفته شده است. حجم کل پیکره برابر با ۲۰٦۹۲٤۳ کلمه است که چهار قسمت از آن برای آموزش و یک قسمت از آن به عنوان دادهٔ آزمون در نظر گرفته شده است. در جدول ۱ آماره‌های پیکرهٔ داستانی معاصر آمده است. در پیوست ۱ فهرست آثار متعلق به این نویسندگان ذکر شده است.

علاوه بر پیکرهٔ بیجن‌خان (بیجن‌خان و دیگران، ۲۰۱۱) و پایگاه داده‌های زبان فارسی (عاصی، ۱۹۹۷)، دیگر آثار هفت نویسنده در وب جستجو شد و کتب متعلق به آنها جمع‌آوری گشت. در انتخاب آثار نویسندگان ژانر رمان و داستان (کوتاه و بلند) مورد توجه بود. از آنجا که برخی از کتب، شامل مجموعه داستان‌های کوتاه بود، داستان‌های کوتاه یک مجموعه نیز تفکیک شد. در مرحلهٔ بعد کتب متعلق به نویسندگان که به قالب پی‌دی‌اف بودند، توسط نرم‌افزار ایبو[4] به قالب متنی در آمدند. این نرم‌افزار از کارایی مطلوبی در تبدیل پی‌دی‌اف به متن زبان فارسی برخوردار است. از آنجا که زبان مورد استفاده در جملات پیکره به گونهٔ محاوره تعلق دارد و

1. Dropout
2. Regularization
3. K-Fold Cross Validation
4. https://www.eboo.ir/

ابزارهای پیش‌پردازش در دسترس زبان فارسی در مورد زبان محاوره از کارآیی مطلوبی برخوردار نیستند، پیکره برچسب‌گذاری نشده است. در نهایت کل پیکره هنجارسازی[1] شد و به‌صورت هماهنگ در آمد و از آن برای آموزش سامانه استفاده شد.

جدول ۱. آماره‌های پیکره داستانی

ردیف	نام نویسنده	تعداد آثار	مجموع نمونه‌ها
۱	نادر ابراهیمی	۸	۲۶۱۸۶۶
۲	احمد محمود	۹	۳۲۷۲۱۷
۳	جلال آل احمد	۹	۲۵۴۲۷۶
٤	بزرگ علوی	۷	۳۲۶۸۰۲
٥	هوشنگ گلشیری	٦	۲٤۷۷۹۱
٦	محمود دولت‌آبادی	۸	۳٤٦۱۳۲
۷	غلامحسین ساعدی	٦	۳۰٥۱٥۹
	مجموع	٥۳	۲۰٦۹۲٤۳

برای محاسبهٔ کارایی سامانه از معیار دقت[2] استفاده شد. دقت در تشخیص خودکار نویسنده عبارت است از نسبت نتایج درست به دست آمده توسط سامانه به کل تعداد نمونه‌های موجود در داده آزمون. از آنجا که فرآیند ارزیابی پنج بار تکرار شد، در نهایت از نتایج به دست آمده در پنج مرحله برای کل سامانه و همینطور برای هر نویسنده میانگین وزن‌دار گرفته شد. علاوه بر این میزان کارایی تمامی ویژگی‌های تعریف‌شده در سامانه بررسی و بهترین ویژگی‌ها انتخاب شد. این کار از طریق آزمون F تحلیل واریانس یک طرفه[3] انجام شد. آزمون تحلیل واریانس یک طرفه برای آزمون مقایسه میانگین یک متغیر کمی در بین بیش از دو گروه مستقل استفاده می‌شود. بر اساس آزمون F تحلیل واریانس یک طرفه، ویژگی‌های برتر به دست آمد. ۱۰ بهترین ویژگی متعلق به کل واژه‌های دستوری و انواع واژه‌های دستوری در میان ۱٤۷ ویژگی برتر در جدول زیر گزارش شده است.

با مشاهدهٔ این جدول می‌توان دریافت که سه بهترین کلمه در میان کل کلمات دستوری عبارتند از: «آن»، «و»، و «این». در میان حروف ربط «و»، «اما» و «بعد»، در میان افعال کمکی «بودن»، «شدن» و «کردن»، در میان ضمایر، ضمایر اشاره «آن»، «این» و «همان»، ضمیر مشترک

1. normalization
2. Accuracy
3. ANOVA

«خود» و ضمیر شخصی «من»، در میان حروف اضافه «توی»، «برای» و «برا» و در میان کمّیت‌نماها، «همه»، «هر» و «هیچ» تأثیر بسزایی در تعیین سبک نویسنده داشته‌اند.

جدول ۲. بهترین کلمات دستوری و انواع آن

ردیف	همهٔ کلمات دستوری	حروف ربط	افعال کمکی	حروف اضافه	کمّیت‌نما	ضمایر	سایر
۱	آن	و	بودن	توی	همه	آن	مرا
۲	و	اما	شدن	برای	هر	این	یه
۳	این	بعد	کردن	برا	هیچ	همان	رو
٤	همان	بل	هستن	در	بعضی	خود	که
٥	بودن	پس	داشتن	بین	چندین	من	بهت
٦	اما	اگر	نیستن	بی	قدری	آنها	برات
۷	توی	وقتی	نبودن	میان	–	وی	بهش
۸	شدن	ولی	نشدن	لای	–	او	اینو
۹	برای	چون	خواستن	دربارهٔ	–	ما	آنکه
۱۰	خود	گویی	–	بدون	–	اون	آنرا

در مرحلهٔ ارزیابی کارآیی سامانه، کل کتابهای موجود در پیکره به اسنادی حاوی ۵۰۰ کلمه تقسیم شد. سپس برای بررسی و مقایسهٔ بهتر عملکرد واژه‌های دستوری و محتوایی، این اسناد به دو طریق تقسیم شد و سامانه به دو شکل مورد ارزیابی قرار گرفت. لازم به ذکر است که بسیاری از پژوهش‌ها از نوع اول ارزیابی استفاده می‌کنند. باتوجه به این‌که در ارزیابی سامانه، بررسی واژه‌های محتوایی نیز مورد نظر بود و در تقسیم‌بندی به روش اول، واژه‌های محتوایی مربوط به یک کتاب، در دو بخش دادهٔ آموزش و آزمون قرار می‌گیرند، برای بررسی دقیق‌تر سبک نگارش ارزیابی به روش دوم نیز انجام گرفت.

در نوع اول، کل اسناد پیکره که هر یک حاوی ۵۰۰ کلمه هستند، مانند کارت‌هایی که در بازی‌ها بُرزده می‌شوند، به صورت تصادفی با یکدیگر جابه‌جا شدند؛ به این کار اصطلاحاً بُرزدن تصادفی[1] گویند. سپس کل داده، به پنج قسمت تقسیم شد: چهار قسمت برای آموزش و یک قسمت برای آزمون. در نوع دوم، تلاش بر این بود که یک داستان خاص بین دادهٔ آموزش و آزمون تقسیم نشود. کل آثار هر نویسنده به پنج قسمت تقسیم شد: چهار قسمت برای آموزش و یک قسمت برای آزمون. این تقسیم‌بندی به گونه‌ای بود که داستان‌های موجود قسمت آموزش از

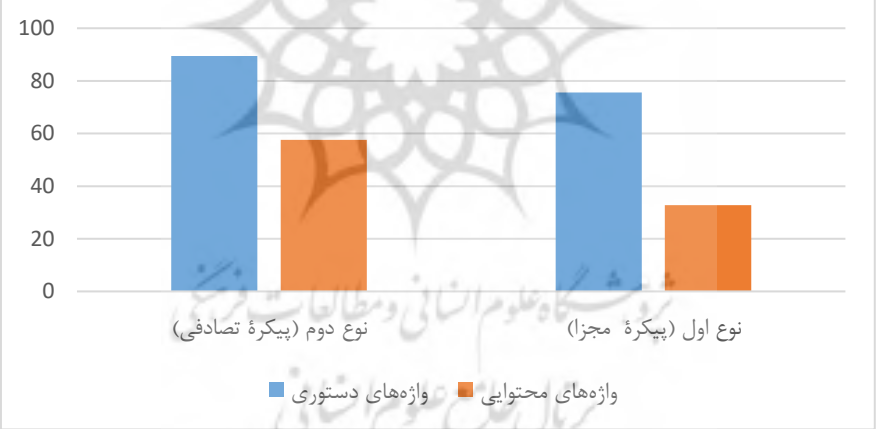
1. random shuffle

داستان‌های موجود قسمت آزمون متمایزند. در مرحلهٔ بعد ارزیابی سامانه به روش به ارزیابی متقاطع ۵ تایی برای هر دو نوع داده انجام شد. در جدول ۳ نتایج ارزیابی سامانه گزارش شده است.

جدول ۳. نتایج ارزیابی سامانه با واژه‌های دستوری و محتوایی

نوع واژه / پیکره	واژه‌های دستوری	واژه‌های محتوایی
نوع اول (تصادفی)	۸۹/٤۱	۵۷/٦
نوع دوم (مجزا)	۷۵/۵٤	۳۲/۸

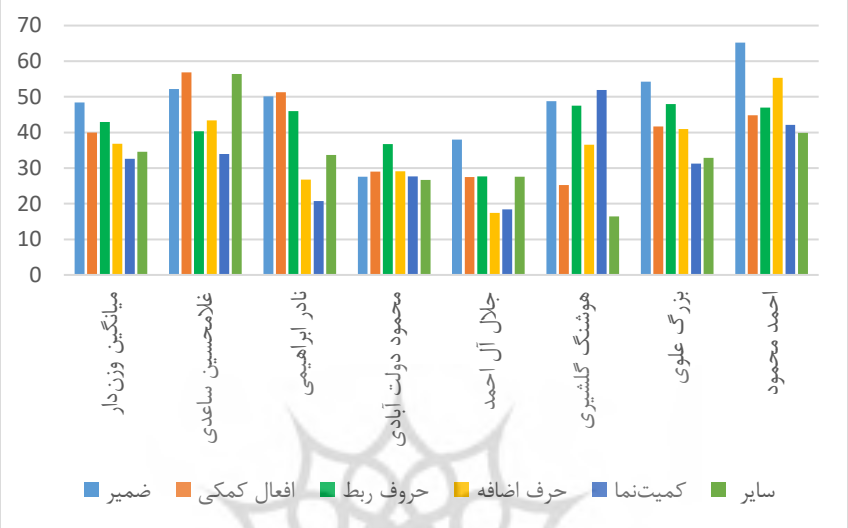
با توجه به جدول ۳، میانگین وزن‌دار برای ۱٤۷ بهترین واژهٔ دستوری با استفاده از الگوریتم پرسپترون چندلایه در دو حالت تقسیم‌بندی به روش اول و دوم (به‌ترتیب: تصادفی و مجزا) برابر با ۸۹/٤۱ و ۷۵/۵٤ درصد و برای واژه‌های محتوایی با الگوریتم حافظهٔ طولانی کوتاه‌مدت به‌ترتیب برابر با ۵۷/٦ و ۵۲/۸ درصد گزارش شده است. در نمودار شکل ۲ نیز نتایج جدول نمایش داده شده است. با مشاهدهٔ نمودار می‌توان دریافت در هر دو نوع داده روش واژه‌های دستوری بر روش واژه‌های محتوایی برتری دارد.



شکل ۲. نمودار مقایسه‌ای ارزیابی سامانه بر اساس معیار دقت

برای بررسی این که بسامد نسبی کدامیک از انواع واژه‌های دستوری بیشترین نقش را در تعیین سبک نگارش متن ایفا می‌کند، فهرست کلمات دستوری به «ضمایر»، « کمّیت‌نماها»، «افعال کمکی»، «حروف اضافه»، «حروف ربط» و «سایر» تفکیک شد. منظور از «سایر»، «حرف تعریف»، «متمم‌نما»، «را» و ترکیب انواع کلمات دستوری (مانند «بهت»، «برایش») و غیره است. سپس کارایی سامانه با هر یک از این گروه‌ها سنجیده شد که در نمودار شکل ۳ آمده است. با مشاهده

شکل ۳ می‌توان دریافت که به ترتیب «ضمایر»، «حروف ربط»، «افعال کمکی»، «حروف اضافه»، «سایر» و « کمّیت‌نماها» در سبک‌سنجی رایانشی متون پیکره این پژوهش مؤثر هستند.



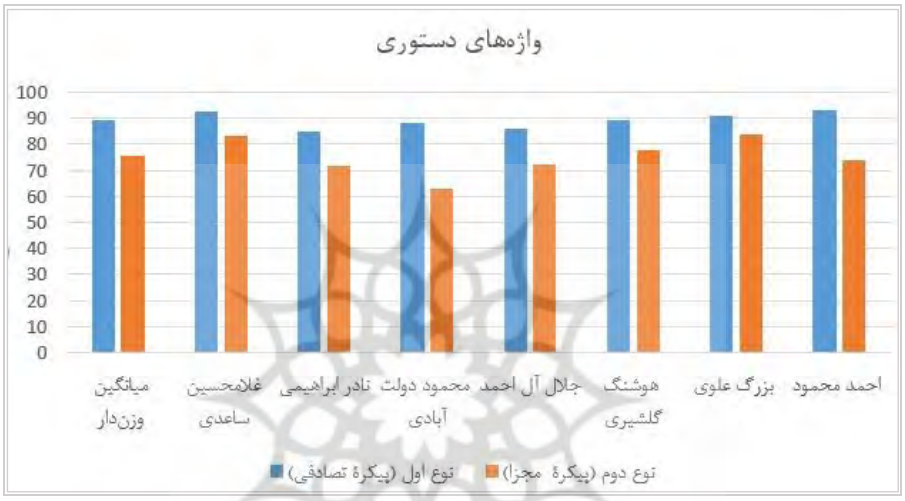
شکل ۳. نمودار مقایسه‌ای کارایی سامانه با انواع مختلف واژه‌های دستوری

در شکل ٤ نیز نمودار مقایسه‌ای نتایج ارزیابی برای نویسندگان مختلف بر اساس ۱٤۷ واژهٔ دستوری برتر، با هر دو نوع داده نمایش داده شده است. با مشاهدهٔ این نمودار می‌توان دریافت که در روش اول ارزیابی، سامانه سبک نگارش احمد محمود، غلامحسین ساعدی و بزرگ علوی را بهتر از سایر نویسندگان و به‌ترتیب با دقتی برابر با ۹۳/۱٤، ۹۲/٦۹ و ۹۰/۹۹ درصد تشخیص داده است و سبک نادر ابراهیمی را با کمترین دقت و برابر با ۸٤/۹٦ درصد از سایرین متمایز کرده است. در روش دوم ارزیابی نیز به‌ترتیب سبک بزرگ علوی، غلامحسین ساعدی و احمد محمود با بالاترین دقت و محمود دولت‌آبادی با کمترین دقت تعیین شده است.

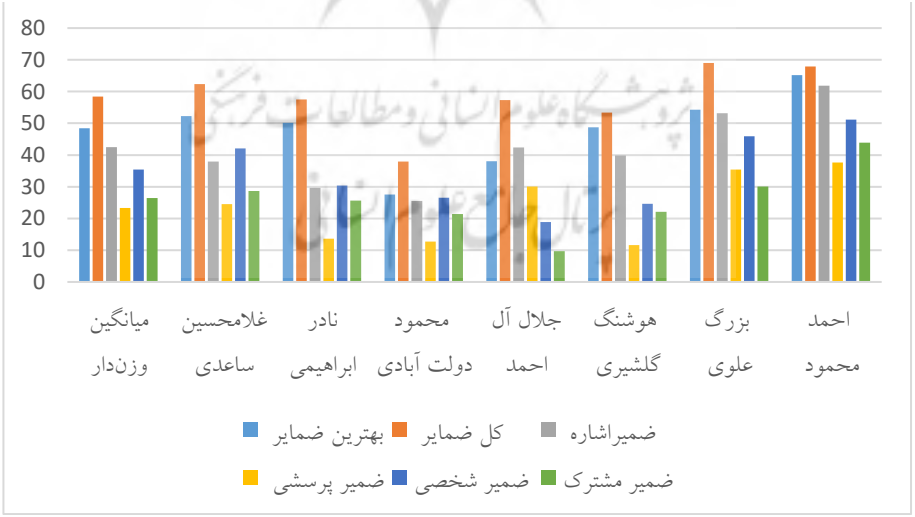
همان‌طور که پیشتر اشاره شد، در این پژوهش برای واژه‌های دستوری دسته‌بندی معنایی ارائه شد و کارایی آنها به طور مجزا در سبک‌سنجی رایانشی بررسی شد. همانطور که در نمودار شکل ۳ آمده است، ضمایر، حروف ربط و افعال کمکی بهترین نتایج را در میان انواع کلمات دستوری دارند. علاوه بر آن، حروف اضافه و «سایر» انواع کلمات دستوری نیز تأثیر خاص خود را دارند و در میان فهرست ویژگی‌های منتخب جای دارند.

برای تحلیل بهتر نقش ضمایر در تشخیص خودکار نویسنده، فهرست ضمایر با ۷۱ کلمه به گروه‌های «ضمایر شخصی»، «ضمایر مشترک»، «ضمایر اشاره»، «ضمایر پرسشی» و همچنین

فهرست بهترین ضمایر (۳٤ کلمه) با انواع مختلف تفکیک و کارایی آنها مقایسه شد. در نمودار شکل ۵ این مقایسه آمده است. با مشاهدهٔ نمودار می‌توان دریافت که ضمایر شخصی و اشاره بهترین نتایج را دارند. بررسی ویژگی‌های منتخب حاکی از این است که ضمیر اشارهٔ «آن» بالاترین امتیاز را در میان کل واژه‌های دستوری دارد. در میان ضمایر با بهترین امتیاز ضمایر اشاره «این»، «همان» و «آنها»، ضمیر شخصی «من» و ضمیر مشترک «خود» جای دارند که نشان‌دهندهٔ تأثیر بسزای این ضمایر در تعیین سبک نویسنده است.

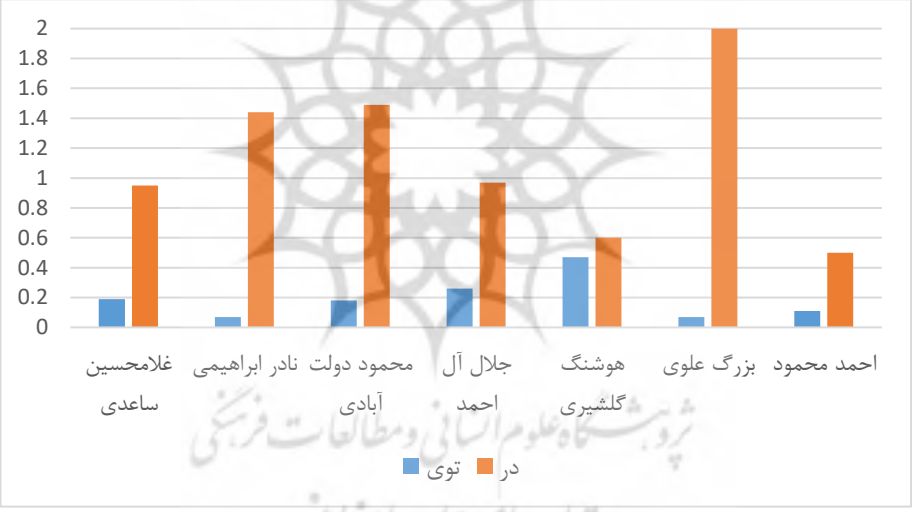


شکل ٤. نمودار مقایسه‌ای ارزیابی سامانه با روش واژه‌های دستوری برای نویسندگان مختلف



شکل ۵. نمودار مقایسه‌ای کارایی سامانه با انواع مختلف ضمیر

در مورد حروف اضافه، بیان یک نکته حائز اهمیت است. در میان حروف اضافه، حروف اضافه‌ای مانند «توی»، «برای» و «برا»، «در»، «بین»، «میان» و غیره می‌توانند ممیز سبک نویسندگان باشند. برای مثال دو حرف اضافهٔ «در» و «توی»، گاه می‌توانند به جای یکدیگر به کار روند. در شکل ٦ تفاوت درصد حضور این دو حرف اضافه در آثار هر یک از نویسندگان نمایش داده شده است. همانطور که قابل ملاحظه است گلشیری نسبت به سایر نویسندگان تمایل بیشتری نسبت به استفاده از «توی» به جای «در» نسبت به سایر نویسندگان دارد و بزرگ علوی و نادر ابراهیمی کمترین تمایل را دارند. علاوه‌براین، احمد محمود از سایر نویسندگان از «در» کمترین بهره را گرفته است. همچنین با حذف حروف اضافهٔ کارایی کل سامانهٔ تشخیص سبک نویسنده حدود یک درصد کاهش پیدا می‌کند. با این توصیف حروف اضافه در تشخیص سبک نگارش تا حدودی می‌توانند کارا باشند؛ چراکه اگر بی تأثیر بوند، نمی‌بایست در کارایی سامانه کاهش صورت می‌گرفت.

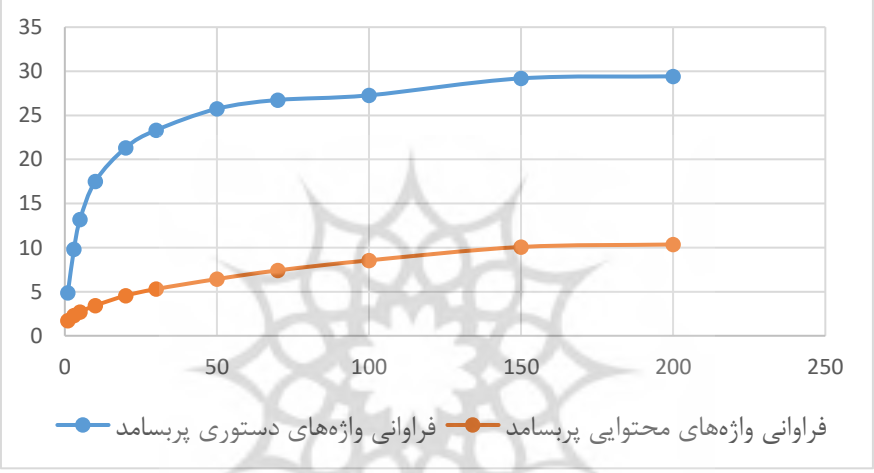


شکل ٦- درصد حضور دو حرف اضافهٔ «در» و «توی» در آثار نویسندگان پیکره

٦. بحث و نتیجه گیری

در پژوهش حاضر هدف بر آن بود که ویژگی‌های زبانی مؤثر در سبک‌سنجی رایانشی متون داستانی زبان فارسی بررسی گردد. پرسش مطرح شده در پژوهش حاضر عبارتست از: در تشخیص سبک نگارش متون داستانی معاصر، واژه‌های دستوری مؤثرتر هستند یا واژه‌های محتوایی؟ همانطور که در نمودار شکل ۲ نمایش داده شده است، روش بسامد واژه‌های دستوری بر روش واژه‌های محتوایی برتری دارد.

از دلایل برتری روش واژه‌های دستوری بر روش محتوایی، می‌تواند وفور واژه‌های دستوری در مقایسه با واژه‌های محتوایی در تمامی متون باشد. در شکل ۷ مقایسه‌ای برای فراوانی واژه‌های دستوری و محتوایی پرکاربرد در پیکره آمده است. در این شکل به ترتیب درصد فراوانی واژهٔ دستوری و محتوایی نخست، سه واژهٔ نخست، ۵ واژهٔ نخست، ۱۰ واژهٔ نخست و غیره آمده است. همانطور که در این شکل قابل ملاحظه است، در پیکرهٔ داستانی بین فراوانی واژه‌های دستوری و محتوایی پرکاربرد تفاوت قابل توجهی وجود دارد و واژه‌های دستوری با اختلاف زیاد از واژه‌های محتوایی در پیکره بیشتر حضور دارند.



شکل ۷- مقایسه فراوانی واژه‌های دستوری و محتوایی در پیکره

نتایج این پژوهش در تأثیر قابل توجه واژه‌های دستوری در تشخیص هویت نویسندهٔ متن در راستای یافته‌های برخی از پژوهش‌های پیشین (باروز، ۱۹۸۷، ۱۹۹۲: ۱٦۷-۲۰٤؛ ۲۰۰۳: ۵-۳۲؛ ۲۰۰۷: ۲۷-٤۷)، (گلشائی، ۱۳۹۸، دباغ، ۲۰۰۷)، هوور (۲۰۰۳) و آرگامون و لویتان (۲۰۰۵) است. واژه‌های دستوری با فراوانی زیاد نسبت به واژه‌های محتوایی، اغلب شاخص‌های بهتری از تمایزات سبکی و ادبی است (باروز، ۱۹۸۷).

سامانه سبک نگارش احمد محمود، غلامحسین ساعدی و بزرگ علوی را با دقت بالاتری نسبت به هوشنگ گلشیری، نادر ابراهیمی، جلال آل احمد تشخیص می‌دهد. دلیل این امر می‌تواند به تفاوت در حجم داده این نویسندگان در پیکره مربوط باشد. گرچه تعداد آثار هوشنگ گلشیری، نادر ابراهیمی و جلال آل احمد در مقایسه با سایر نویسندگان کم نیست اما حجم کتب آنها زیاد نیست؛ در نتیجه تعداد کل کلمات آثار این نویسندگان در قیاس با سایرین کمتر است و این امر می‌تواند عاملی باشد که سامانه با دقت پایین‌تری سبک آنها را تشخیص دهد. دقت

سامانه در مورد محمود دولت‌آبادی در حد انتظار نبوده است. گرچه حجم دادهٔ این نویسنده در پیکره بسیار مطلوب است اما اکثر آثار به رمان‌ها و داستان‌های بلند تعلق دارند و داستان‌های کوتاه در آثار ایشان کمتر دیده شده است؛ بنابراین تنوع داستان‌ها کمتر است و احتمالاً همین مسئله بر کارآیی سامانه تأثیر منفی گذاشته است. با این توصیف دو عامل حجم داده و تنوع در آثار نویسندگان می‌تواند در کیفیت سامانه مؤثر باشد.

به عنوان کارهای آتی، می‌توان مجموعه ویژگی‌های تعریف شده در این پژوهش را برای تعیین جنسیت، ملیت، ردهٔ سنی و نوع شخصیت نویسندهٔ متن به کار برد. همچنین می‌توان پیکرهٔ پژوهش حاضر را به منظور سبک‌شناسی برای اهداف زبانی، نقد ادبی و تحلیل گفتمان انتقادی به کار گرفت. برای این منظور نیاز است که پیکرهٔ این پژوهش توسط خطایاب املایی پیش‌پردازش شود. گرچه کیفیت نرم‌افزار تبدیل قالب پی‌دی‌اف به قالب متنی رضایت‌بخش است و حتی برای نسخه‌هایی از کتب که کیفیت تصویر آنها پایین است، نیز کیفیت آن قابل قبول است ولی این سامانه قطعاً درصدی خطا را برای این نوع تبدیل دارد. این مشکل برای واژه‌های محتوایی نسبت به واژه‌های دستوری دو چندان است. واژه‌های دستوری فهرست بسته و صورت نگارش ساده‌تر و کوتاه‌تری دارند. در مقابل واژه‌های محتوایی به فهرست بازی تعلق دارند و صورت نگارش آنها عموماً پیچیده‌تر است و همین مسئله می‌تواند بررسی آنها را قدری دچار چالش کند. بدیهی است که این مسئله می‌توانسته کیفیت بخشی از پیکره را که حاصل استفاده از این نرم‌افزار است، تحت تأثیر قرار دهد. در صورتی که هدف بررسی دقیق‌تر واژه‌های محتوایی باشد، به‌کارگیری یک خطایاب املایی می‌تواند در کیفیت پیکره مؤثر باشد. لازم به ذکر است که در این پژوهش ازآنجاکه سامانه با واژه‌های محتوایی پرکاربرد سروکار داشته است و صورت‌های نگارشی غلط در میان این مجموعه واژه جایی ندارند، این مشکل قابل توجه نیست اما ممکن است درصدی خطای جزئی به سامانه تحمیل کند.

منابع

آذین، زهرا و بحرانی، محمد. (۱۳۹۳). «شناسایی خودکار شاعران شعر نو با استفاده از ویژگی های سبکی». ***مجموعه مقالات نهمین همایش زبانشناسی ایران***، تهران: دانشگاه علامه طباطبایی.

امیری، محمد عارف، رستم بیک تفرشی، آتوسا و مدرسی، یحیی. (۱۳۹٦). تحلیل گفتمان پیکره–بنیاد ترانه‌های فارسی: *زبان‌شناخت*. ۸(۱٦)، ۱–۲۵.

بابانژاد باقری، سیده مریم، پورآقاجان، عباسعلی و عباسیان، محمد مهدی. (۱٤۰۲). «پیش‌بینی ارزش شرکت مبتنی بر روش‌های یادگیری عمیق». *اقتصاد مالی*. ۱۷(٦٤). ۲۹۱–۳۱۸. doi: 10.30495/fed.2023.705603

جوانمردی، کامیار و اکبری، منوچهر. (۱۳۹۷). «روش‌های یادگیری ماشین در بررسی ویژگی‌های زبان شعری در اشعار شاعران دفاع مقدس (مطالعهٔ موردی: اشعار دو شاعر دفاع مقدس؛ قیصر امین‌پور و محمدرضا عبدالملکیان)». *مطالعات دفاع مقدس*، ۱۵(٤)، ١٤٤-١٢١.

عارفی، سمیه.، بصیری، محمد احسان، و روزمند، امید. (١٤٠٠). «انتخاب ویژگی برای شناسایی نویسنده در متون کوتاه برخط فارسی». *فناوری اطلاعات و ارتباطات ایران*، ۳۵-۵٦.

کمال پور، مهسا، مدرس خیابانی، شهرام و حجازی، محمد جواد. (١٤٠١). «بررسی و مقایسهٔ پیکره‌بنیاد گفتمان آثار جلال آل احمد و احمد محمود». *تفسیر و تحلیل متون زبان و ادبیات فارسی (دهخدا)*، ١٤(٥٢)، ٣٣٢-٣٥٥.

کمال پور، مهسا، مدرس خیابانی، شهرام و حجازی، محمد جواد. (۱۳۹۹). «نقش کلیدواژه‌ها در تحلیل گفتمان مطالعهٔ موردی: «خسی در میقات» و «غرب‌زدگی» دو اثر از جلال آل احمد». *فصلنامه علمی - پژوهشی زبان‌شناسی اجتماعی*، ۳(۳)، ۵۵-۷٦.

گلشائی، رامین. (۱۳۹۸). «واژه‌های دستوری به‌مثابه نشانگرهای گویش فردی: رویکردی پیکره‌ای به شناسایی هویت نویسنده در زبان فارسی»، *جستارهای زبانی*، ۵۱(۱۰)، ۳۱۷-۲۹۳.

میرزائی، آزاده. و صفری، پگاه. (١٣٩٤). «ساخت واژه - متن‌های تخصصی و عمومی زبان فارسی بر اساس بسامدگیری واژه‌های نقشی و محتوایی». *مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای*، تهران، ۱۷۵-۱۹۱.

میرزایی، آزاده. (۱۳۹۷). *آشنایی با زبانشناسی پیکره‌ای*. تهران: انتشارات دانشگاه علامه طباطبایی.

هومن، محمود. (۱۳۵۷). *حافظ*، به کوشش اسماعیل خویی. تهران: طهوری.

Alhuqail, N. (2021). "Author Identification Based on NLP". *European Journal of Computer Science and Information Technology*, Vol.9, No.1, pp.1-26, 2021, Available at SSRN: https://ssrn.com/abstract=3820262

Argamon, S & ,Levitan, S. (2005). "Measuring the usefulness of function words for authorship attribution". *Proceeding of the Joint Conference on Association for Literary and Linguistic Computing/Association Computer Humanities.*

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Levitan, S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, *58*(6), 802-822. https://doi.org/10.1002/asi.20553.

Baker, P. (2006). *Using Corpora in Discourse Analysis.* London Continuum International Publishing Group.

Bayrami, P.; Rice, J.E. (2021). "Code authorship attribution using content-based and non-content-based features". *In Proceedings of the 2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Canada, pp. 1–6.

Burrows, J. F. (1987). *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press.

Burrows, J. F. (1992). *Computers and the study of literature*. In C.S. Butler (Ed.), *Computers and written texts: An applied perspective*, (167–204). Oxford: Blackwell.

Burrows, J.F. (2003). *Questions of authorship: Attribution and beyond*. *Computers and the humanities*, *37* (1), 5-32.

Burrows, J. F. (2007). *All the way through: Testing for authporship in different frequency strata*. Literary and linguistics computing, *22*(1), 27-47.

Coulthard, M. (2004). "Author identification, idiolect, and linguistic uniqueness". *Applied linguistics*, *25*(4), 431-447 .

Dabagh, R. M. (2007). "Authorship attribution and statistical text analysis." *Metodoloski zvezki*, *4*(2), 149 .

Gamon, M. (2004). "Linguistic correlates of style: authorship classification with deep linguistic analysis features". *International Conference on Computational Linguistics.*

Hedegaard, S., & Simonsen, J. (2011). "Lost in translation: authorship attribution using frame semantics". *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *Portland*, *USA*, 65-70.

Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, 37(2), 151–178.

Hoover, D. L. (2004). Frequent collocations and authorial style. *Literary and Linguistic Computing* 18(3), 261–268.

Hoover, D. L. (2007). Quantitative analysis and literary studies. In R. Siemens & S. Schreibman (Eds). *A companion to digital literary studies* (pp. 517–533). Oxford: Blackwell.

Houvardas, J., & Stamatatos, E. (2006). "N-gram feature selection for authorship identification". *AIMSA. 4183.* 77-86. 10.1007/11861461_10.

Kešelj, V., Peng, F., Cercone, N., & Thomas, C.(2003). "N-gram-based author profiles for authorship attribution". *Proceedings of the Conference Pacific Association for Computational Linguistics PACLING'03.*

Kestemont, M. (2014). "Function words in authorship attribution. From black magic to theory?" *CLFL@EACL*. 59-66. 10.3115/v1/W14-0908.

Klammer, T., Schulz, M. & Della, A. (2009). *Analyzing English Grammar* .Longman.

Kukushkina, O. V., Polikarpov, A. A., & Khmelev, D. V. (2001). "Using literal and grammatical statistics for authorship attribution". *Problems of Information Transmission*, *37*(2), 172-184 .

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). "Analyzing Writing Styles with Coh-Metrix". *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference.*

Qian, C., He, T., & Zhang, R. (2017). Deep Learning based Authorship Identification.

Radford, A. (2004). *Minimalist syntax: Exploring the structure of English*: Cambridge University Press.

Rahgozar, A. (2020). Automatic Poetry Classification and Chronological Semantic Analysis. (PhD degree in E-Business, University of Ottawa, Ottawa, Canada).

Ramezani, R. (2021). "A Language-independent author attribution approach for author identification of text documents". *Expert Systems with application*, vol. 180.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). "Cross-Validation". *Encyclopedia of database systems*, *5*, 532-538 .

Sari, Y., Vlachos, A., & Stevenson, M. (2017). "Continuous n-gram representations for authorship attribution". *European Chapter of the Association for Computational Linguistics (EACL 2017).*

Segarra, S., Eisen, M., & Ribeiro, A. (2015). "Authorship attribution through function word adjacency networks". *IEEE Transactions on Signal Processing, 63*(20), 5464-5478 .

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

Song, M. and Yi-Fang Brook W. (2009). *Handbook of Research on Text and Web Mining Technologies.* IGI Global.

Toolan, M.J. (2008). Narrative progression in short story: First steps in a corpus stylistic approach. *Narrative, 16*(2), 105–120.

Wanner, L. (2017). "On the relevance of syntactic and discourse features for author profiling and identification". *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.*

Zhao, Y., & Zobel, J. (2005). "Effective and scalable authorship attribution using function words". *Information Retrieval Technology. AIRS 2005. Lecture Notes in Computer Science, vol 3689.* Springer, Berlin, Heidelberg.

Zhao, Y., & Zobel, J. (2007). "Searching with style: Authorship attribution in classic literature". *Proceedings of the Thirtieth Australasian Conference on Computer Science. 62. 59-68.*

پیوست ۱. فهرست کتاب‌های پیکره

نام نویسنده	نام کتاب	ژانر
نادر ابراهیمی	یک عاشقانه آرام	رمان
	بار دیگر شهری که دوست می‌داشتم	رمان
	مصابا و رؤیای گاجرات	داستان کوتاه
	بر جاده‌های آبی سرخ	رمان
	رونوشت بدون اصل	داستان کوتاه
	چهل نامه کوتاه	قطعه ادبی (نامه‌نگاری)
	تضادهای درونی	داستان کوتاه
	آتش بدون دود	رمان
احمد محمود	همسایه‌ها	رمان
	مدار صفر درجه	رمان
	غریبه‌ها و پسرک بومی	داستان کوتاه
	زائری زیر باران	داستان کوتاه
	بیهودگی	داستان کوتاه
	دریا هنوز آرام است	داستان کوتاه
	قصهٔ آشنا	داستان کوتاه
	مول	داستان کوتاه
	درخت انجیر معابد	رمان
جلال آل احمد	نفرین زمین	رمان
	نون و القلم	رمان
	سرگذشت کندوها	رمان
	داستان بچه مردم	داستان کوتاه
	مدیر مدرسه	رمان
	پنج داستان	داستان کوتاه
	هر آدمی سنگی است بر گور پدر خویش	رمان (خودزندگی‌نامه)
	از رنجی که می‌بریم	داستان کوتاه
	زن زیادی	داستان کوتاه

نام نویسنده	نام کتاب	ژانر
بزرگ علوی	گیله‌مرد	داستان کوتاه
	چمدان	داستان کوتاه
	میرزا	داستان کوتاه
	چشم‌هایش	رمان
	سالاری‌ها	رمان
	ورق‌پاره‌های زندان	داستان کوتاه
	موریانه	رمان
هوشنگ گلشیری	نیمه تاریک ماه	داستان کوتاه
	شاه سیاه‌پوشان	داستان بلند
	شاهزاده احتجاب	رمان
	در ولایت هوا	رمان
	بره گمشده راعی	رمان
	کریستین و کید	رمان
محمود دولت‌آبادی	کلیدر	رمان
	روزگار سپری‌شدهٔ مردم سالخورده	رمان
	آوسنه بابا سبحان	داستان بلند
	روز و شب یوسف	داستان بلند
	سلوک	رمان
	سفر	داستان بلند
	لایه‌های بیابانی	داستان کوتاه
	داستان آینه	داستان کوتاه
غلامحسین ساعدی	گدا	داستان کوتاه
	آشغالدونی	داستان بلند
	آشفته‌حالان بیداربخت	داستان کوتاه
	تاتار خندان	رمان
	غریبه در شهر	رمان
	عزاداران بیل	داستان کوتاه

استناد به این مقاله: سلطان‌زاده، فاطمه، میرزائی، آزاده، بحرانی، محمد و مدرس خیابانی، شهرام. (۱۴۰۲). بررسی پیکره‌بنیاد تعیین سبک نگارش متون فارسی با واژه‌های دستوری و محتوایی. *زبان و زبان‌شناسی*، **۱۹** (۳۷)، ۱۹۳- ۲۲۰.
doi: 10.30465/lsi.2024.47545.1725