Razi University

Research in Western Iranian Languages and Dialects

ISSN Print: 2345-2579 ISSN Online: 2676-573X

Linguistics Society of Iran

Automatic Author Gender Identification From Persian Text Using ParseBert and Linguistic Features

Mohamad Bahrani[1], and **Seyedeh Sareh Sadeghi**[2]✉

1. Assistant Professor, Department of Mathematics, Faculty of Statistics, Mathematics and Computer Science, Allameh Tabatabaei University, Tehran, Iran. E-mail: Bahrani@atu.ac.ir
2. Corresponding Author, Ph.D. in Linguistics, Department of English Language and Literature, Faculty of Literature and Human Sciences, Bu Ali Sina University, Hamedan, Iran. E-mail: Sarehsadeghi36@yahoo.com

Article Info

Article type:
Research Article

Article history
Received: 07 Feb 2025
Received in revised form: 02 Aug 2025
Accepted: 16Aug 2025
Published online: 23 Sep 2026

Keywords:
automatic author gender identification,
gender tagged corpus,
ParseBert model,
speech gender differences,
text embedding

ABSTRACT

The sentences that people use during writing contain valuable information that can be used to identify the author's gender. Meanwhile, the use of deep learning algorithms in natural language processing helps identify hidden patterns in the text. In this research, an attempt is made to design a system for the Persian language that identifies the author's gender by fine-tuning the parameters of the ParsBERT model. For this purpose, first, a corpus of 5,000 documents labeled with gender tags is prepared, and then the author's gender identification system is designed and evaluated using 10-fold cross-validation. Experimental results show that the F-measure of the gender identification task is 76.5%. The proposed method is also compared with classic machine learning methods. It also achieves better results compared with the LSTM model. The results obtained from comparing the corpus prepared in this research with the corpus prepared in previous research for gender identification show an improvement in the system's performance. Thus, the use of new deep learning methods, such as the ParsBERT model, and appropriate data are among the main achievements of this research.

Cite this article: Bahrani, M., & Sadeghi, S. S. (2026). Automatic author gender identification from Persian text using ParseBert and linguistic features. *Research in Western Iranian Languages and Dialects*, *14*(3), 19–35. http//doi.org/10.22126/jlw.2025.11743.1814

© The Author(s).
DOI: https://doi.org/10.22126/jlw.2025.11743.1814

Publisher: Razi University

Introduction

Automatic gender classification has become a pivotal research area within computational linguistics due to its wide-ranging applications, including the enhancement of user-friendly digital environments and the improvement of security in online interactions. Since Koppel et al. (2002) pioneered automatic gender identification, technological advancements have increased interest in developing systems that infer demographic attributes, such as gender and age, as well as affective characteristics such as sentiment, from text. For Persian, a language with limited NLP resources, developing such systems is crucial for creating safer and more interactive online spaces, especially given that users can conceal their identities on social media platforms. This study stands out by introducing a gender-labeled emotional corpus and employing ParsBERT, a transformer-based model specifically designed for Persian, to capture subtle linguistic patterns that distinguish between male and female writing styles. The research examines whether ParsBERT can accurately identify the gender of authors in Persian texts and whether the use of an emotionally labeled corpus can enhance classification performance compared with other datasets. The hypotheses of this study are that linguistic features, such as emotional adjectives and adverbs, differ significantly between genders and that deep learning models, such as ParsBERT, outperform traditional machine learning methods.

Theoretical Background

The study is grounded in several decades of research on gender differences in language use. Lakoff (1975) argued that women tend to use specific linguistic patterns, such as confirmatory phrases (e.g., “isn’t it?”), which may reflect social dynamics. Tannen (1994) expanded on this perspective by identifying collaborative speech styles among women, emphasizing empathy, in contrast to competitive styles among men, which focus more on information transfer. Holmes (2013) highlighted the influence of social, cultural, and situational factors, cautioning against overgeneralizing gender differences. Brown and Levinson (1987) associated linguistic politeness strategies with social power dynamics. Recent studies, such as Brody (2013), have explored multifaceted factors, including age and ethnicity, while Chai et al. (2016) have examined online communication. Women have been reported to use more descriptive and emotional vocabulary, such as adjectives (“beautiful,” “amazing”) and intensifiers (“very,” “truly”), while men have been reported to favor functional, direct language (Newman et al., 2008). These differences provide a basis for feature selection in automated gender identification, while acknowledging the complexity of syntactic, semantic, and contextual linguistic patterns.

Related Work

Previous research on gender identification has covered various languages and methodological approaches. Cheng et al. (2011) achieved 85.1% accuracy in English using statistical features and classifiers, including Naive Bayes, Support Vector Machines (SVM), and Decision Trees. In Arabic, Alsmearat et al. (2017) employed bag-of-words representations and morphological features, while Rangel et al. (2019) found that SVM and neural networks were effective. Sboev et al. (2018) used convolutional and LSTM networks for Russian texts, which outperformed traditional methods. In Persian, Moradi and Bahrani (2015) reported 73.8% accuracy using psycholinguistic and stylistic features with traditional classifiers. Sajedi and Taslimi (2019) achieved 89.5% accuracy using Bayesian Random Forests on blog data. Zarifi and Naghavi (2022) reported an accuracy of 81.09% using conceptual vectorization, while Hekmatianzadehpour and Jalali Bidgoli (2019) obtained 85.7% accuracy by incorporating punctuation and n-grams. These studies highlight the potential of combining linguistic features with machine learning approaches, but they also

underscore the relatively limited attention given to Persian-specific systems, which is the gap addressed by this research.

Methodology

The methodology focuses on constructing a gender-labeled emotional corpus and applying ParsBERT in conjunction with traditional classifiers. The corpus, adapted from Sadeghi et al. (2021), includes 23,000 Persian sentences categorized into five emotion categories (anger, sadness, joy, fear, and surprise), of which 5,000 documents (2,500 for each gender) were manually labeled according to gender. Preprocessing involves normalization, stemming, and stop-word removal to address Persian-specific challenges, such as inconsistent spacing and orthographic variations (Ghayoomi et al., 2009). For text representation, Word2Vec and bag-of-words approaches are employed, with 100-dimensional vectors representing the average word embeddings of each document. Classification models include Naive Bayes, SVM, Decision Trees (J48), LSTM, and ParsBERT. ParsBERT, a BERT-based model developed for Persian, is fine-tuned with a batch size of 32, a learning rate of 10^{-4}, and a feed-forward layer with softmax activation for binary classification (female vs. male). It employs 12 hidden layers, 12 attention heads, and a 768-dimensional hidden representation, and is optimized using the Adam optimizer (Kingma & Ba, 2014). Performance is evaluated using 10-fold cross-validation with precision, recall, and F1-score as evaluation metrics.

Evaluation

ParsBERT achieved the highest performance, with an accuracy of 77%, a recall of 76%, and an F1-score of 76.5%, surpassing traditional classifiers and the LSTM model. Its superior performance can be attributed to its transformer-based architecture, which is capable of capturing deep semantic relationships and subtle stylistic differences, such as pronoun usage, sentence length, and emotional tone, that are not readily captured by TF-IDF or Word2Vec embeddings. A comparison with the corpus developed by Moradi and Bahrani (2015), which comprises conversational texts, reviews, and narratives, indicates that the model performs better on the emotional corpus. This improvement may be attributed to the corpus's emphasis on emotionally charged texts, which may provide more salient gender-related linguistic patterns.

The error analysis revealed three main types of errors: neutral texts that are difficult to distinguish based on gender, texts in which one gender adopts stylistic patterns associated with the other gender, and errors arising from model limitations resulting from the relatively small size of the dataset. Increasing the size and diversity of the corpus may help reduce these errors and consequently improve overall classification performance.

Conclusion

This study contributes to the advancement of automatic gender identification in Persian by introducing a gender-labeled emotional corpus and leveraging the deep-learning capabilities of ParsBERT. The achieved accuracy of 77% demonstrates the effectiveness of transformer-based models in capturing nuanced linguistic differences in comparison with traditional methods. The emotional corpus may further enhance classification performance by providing linguistically rich, emotionally charged texts that highlight gender-specific language patterns. The publicly available corpus provides a valuable resource for future research in Persian NLP. More broadly, this study contributes to the development of NLP resources for low-resource languages and may support more effective human–machine interactions in digital environments.

Future research could focus on developing larger and more diverse corpora and investigating more advanced transformer-based architectures to further improve classification accuracy and generalizability, while addressing the remaining challenges in Persian NLP.

Ethical Considerations
Not applicable

Funding
Not applicable

Conflict of interest
The authors declare no conflict of interest

مطالعات زبان‌ها و گویش‌های غرب ایران

شاپا چاپی: ۲۵۷۹ – ۲۳۴۵ شاپا الکترونیکی: ۵۷۳x – ۲۶۷۶

تشخیص خودکار جنسیت نویسندهٔ متن فارسی با استفاده از مدل پارس‌برت و ویژگی‌های زبانی

محمد بحرانی[۱] | سیده ساره صادقی[۲] ✉

۱. استادیار، گروه ریاضی، دانشکدهٔ آمار، ریاضی و رایانه، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: Bahrani@atu.ac.ir

۲. نویسندهٔ مسئول، دکتری زبان‌شناسی همگانی، گروه زبان و ادبیات انگلیسی، دانشکدهٔ زبان و ادبیات، دانشگاه بوعلی‌سینا، همدان، ایران. رایانامه: Sarehsadeghi36@yahoo.com

اطلاعات مقاله

نوع مقاله: پژوهشی

تاریخ دریافت: ۱۴۰۳/۱۱/۱۹

تاریخ بازنگری: ۱۴۰۴/۵/۱۱

تاریخ پذیرش: ۱۴۰۴/۵/۲۵

تاریخ انتشار: ۱۴۰۵/۷/۱

کلیدواژه‌ها:
شناسایی خودکار جنسیت نویسنده، پیکرهٔ برچسب‌گذاری‌شدهٔ جنسیتی، مدل پارس‌برت
تفاوت‌های گفتاری جنسیت، تعبیه‌سازی متن

چکیده

متون نوشتاری دربردارندهٔ اطلاعات زبانی ارزشمندی هستند که می‌توان از آن‌ها برای تعیین جنسیت نویسنده بهره برد. دراین‌بین، استفاده از الگوریتم‌های یادگیری عمیق در پردازش متون زبان طبیعی، شناسایی الگوهای پنهان در داده‌ها را آسان کرده است. در این پژوهش با هدف شناسایی جنسیت نویسنده در زبان فارسی با تنظیم دقیق پارامترهای مدل پارس‌برت، سامانه‌ای برای این زبان طراحی شد. به این منظور، باتکیه‌بر ویژگی‌های زبانی متمایزکنندهٔ جنسیتی، ابتدا پیکره‌ای شاملِ ۵۰۰۰ سندِ برچسب‌گذاری‌شده تهیه شد. سپس سامانه‌ای مبتنی‌بر ویژگی‌های زبانی و تعبیه‌سازی برداری طراحی شد. این سامانه با روش اعتبارسنجی متقابل ۱۰-تایی ارزیابی شد. برپایهٔ یافته‌ها، معیار-اف در سامانهٔ طراحی‌شده، معادل ۷۶/۵ درصد تعیین شد. همچنین مشخص شد مدل پیشنهادی در مقایسه با دسته‌بندی‌های رایج یادگیری ماشین و در مقایسه با مدل حافظهٔ طولانی کوتاه‌مدت، عملکرد مطلوب‌تری دارد. افزون‌براین، مقایسهٔ عملکرد سامانه بر پیکرهٔ تهیه‌شده در این پژوهش با پیکره‌های پیشین، بهبود عملکرد سامانه را نشان داد. بنابراین، بهره‌گیری از روش‌های نوین همچون مدل پارس‌برت و استفاده از دادگان مناسب، از دستاوردهای اصلی این پژوهش محسوب می‌شود.

استناد: بحرانی، محمد؛ صادقی، سیده ساره (۱۴۰۵). تشخیص خودکار جنسیت نویسندهٔ متن فارسی با استفاده از مدل پارس‌برت و ویژگی‌های زبانی. *مطالعات زبان‌ها و گویش‌های غرب ایران*، ۱۴ (۳)، ۱۹-۳۵. http//doi.org/10.22126/jlw.2025.11743.1814

© نویسندگان

DOI: http//doi.org/10.22126/jlw.2025.11743.1814

ناشر: دانشگاه رازی

۱– مقدمه

طبقه‌بندی جنسیتی به‌دلیل کاربردهای گسترده و فزاینده‌اش، به حوزه‌ای پیشرو در پژوهش‌های پردازش متن تبدیل شده است. در دههٔ اخیر، طبقه‌بندی خودکار جنسیت توجه بسیاری را به خود جلب کرده است؛ زیرا تقویت این توانایی با برنامه‌های کاربردی مختص یک جنسیت خاص، امکان ایجاد محیط‌های کاربری تعاملی‌تر و انسان‌گونه را فراهم می‌آورد. از مباحث اساسی در حوزهٔ پردازش زبان طبیعی، «شناسایی خودکار جنسیت نویسنده»[1] است که از سال ۲۰۰۲ میلادی مطرح شده است (کاپل[2] و دیگران، ۲۰۰۲). پیشرفت‌های چشمگیر در حوزهٔ فناوری اطلاعات، درکنار کاربردهای کاربردی سامانه‌های شناسایی جنسیت در حوزه‌های تبلیغاتی و امنیتی، سبب شده است تا پژوهشگران با تمرکز بیشتر به طراحی سامانه‌هایی بپردازند که به تشخیص هم‌زمان جنسیت، سن و تحلیل احساسات نویسندهٔ متن می‌پردازند. اهمیت بررسی اطلاعات در شبکه‌های اجتماعی، وبگاه‌ها و رایانامه‌ها می‌تواند در ارتقای امنیت و سلامت فضای ارتباطی تأثیرگذار باشد؛ به‌ویژه اینکه در فضای مجازی امکان پنهان‌سازی هویتِ کاربران مانند نام، جنسیت، سن و مکان، زمینه را برای سوءاستفاده‌های امنیتی و اجتماعی یا فعالیت‌های غیرقانونی آن‌ها فراهم کرده است. دراین‌زمینه، پژوهشگران معمولاً با بررسی تفاوت سبک‌های زبانی بین جنسیت‌های مختلف در طبقات مختلف می‌کوشند به بهبود عملکرد سامانه‌های شناسایی جنسیت کمک کنند. بااین‌حال، مرور پیشینهٔ پژوهش‌ها در زبان فارسی، به‌منزلهٔ یکی از زبان‌های کم‌پردازش‌شده، نشان می‌دهد که شمار سامانه‌های طراحی‌شده برای شناسایی جنسیت نویسنده در این زبان بسیار محدود است. ازاین‌رو، طراحی سامانهٔ شناسایی جنسیت نویسنده برای زبان فارسی به‌منظور ایجاد بستری امن برای کاربران فارسی‌زبان گامی ضروری تلقی می‌شود. نوآوری پژوهش حاضر در مقایسه با مطالعات پیشین، در بهره‌گیری روش‌های نوین پردازش زبان فارسی و تهیهٔ پیکره‌ای تخصصی است که بر متغیرهای جنسیت و احساس متمرکز است.

در مطالعات پیشین درزمینهٔ تفاوت‌های جنسیتی، بیشتر بر تحلیل ویژگی‌های زبانی همچون بسامد کاربرد رنگ‌واژه‌ها، صفات، قیدها، دشنام‌ها و ضمایر موجود در گفتار زنان و مردان پرداخته شده است. بااین‌حال، بررسی و شناسایی این ویژگی‌ها، به‌صورت دستی و خودکار، به مطالعات بیشتر نیاز دارد؛ ضمن آنکه در حوزهٔ شناسایی خودکار جنسیت متن، ضرورت بهره‌گیری از ویژگی‌های دیگر برای افزایش دقت مدل‌ها، موضوعی است که باید متخصصان به آن توجه کنند. افزون‌براین، رایانه به‌صورت خودکار قادر نیست نوع تفاوت‌های موجود در نگارش زنان و مردان را در لایه‌های آشکار زبان شناسایی کند و تشخیص این تفاوت‌ها به‌شدت به بافتی وابسته است که گوینده در جمله‌های خود به کار می‌برد. یکی از اساسی‌ترین دلایل این پیچیدگی، وجود مؤلفه‌های نحوی، معنایی، واژگانی و بافتی پیچیده در زبان است. ازسوی دیگر، الگوریتم‌های یادگیری عمیق که در سال‌های اخیر به کار گرفته شده‌اند، به‌طور مطلوبی قادرند الگوهای موجود در این پیچیدگی‌های زبانی را به‌صورت خودکار شناسایی و استخراج کنند و در فرایند تحلیل و طبقه‌بندی متن به کار گیرند. برای نمونه، سامانه با استفاده از الگوریتم برت[3] می‌تواند کوچک‌ترین واحدهای زبان محاوره‌ای همچون حروف اضافه را تااندازه‌ای مشابه انسان درک کند و تفاوت‌های ظریف معنایی و بافتی واژه‌ها را بهتر تشخیص دهد. به‌این‌ترتیب، توانایی این الگوریتم در پردازش دقیق تک‌تک واژه‌های یک جمله و درک تفاوت‌های معنایی آن‌ها، امکان کشف روابط پنهان میان واژه‌ها را فراهم آورده است؛ ازاین‌رو، در این پژوهش به کار گرفته شده است.

به‌این‌ترتیب، پژوهش حاضر در حوزهٔ زبان‌شناسی محاسباتی قرار می‌گیرد. هدف از این مطالعه، ارائهٔ راهکارهای مبتنی‌بر مدل‌های پردازش زبان طبیعی برای شناسایی جنسیت نویسنده در زبان فارسی است. در این پژوهش، به‌طور مختصر، ابتدا پیشینهٔ پژوهش در حوزهٔ شناسایی جنسیت نویسنده و تفاوت‌های جنسیتی گفتار بررسی می‌شود؛ دراین‌زمینه، بر تحلیل ویژگی‌های انتخاب‌شده، ماهیت پیکره‌ها و میزان دقت روش‌های موجود تأکید شده است. سپس ضمن معرفی سامانهٔ پیشنهادی، به فرایند تهیهٔ پیکرهٔ مناسب، انتخاب روش مناسب برای «تعبیه‌سازی متن»[4] و «مدل پارس‌برت»[5] پرداخته می‌شود. در ادامه، نتایج مربوط به تنظیم دقیق مدل پارس‌برت با استفاده از داده‌های پژوهش ارزیابی می‌شود. همچنین با مقایسهٔ این روش با مدل‌های آماری همچون «بیز ساده»[6]، «درخت تصمیم

1. automatic author gender identification
2. M. Koppel
3. Bert algorithm
4. text embedding
5. ParsBert model
6. naïve Bayes

J48»[1]، «ماشین بردار پشتیبان»[2] و دیگر «شبکه‌های عصبی»[3] مانند «حافظهٔ طولانی کوتاه‌مدت»[4]، «میزان دقت»[5] و «بازخوانی»[6] سامانهٔ طراحی‌شده بررسی می‌شود. افزون‌براین، عملکرد سامانهٔ تشخیص جنسیت نویسنده، در این پژوهش، با مقایسهٔ پیکرهٔ حاضر با پیکره‌های مشابه در مطالعات پیشین بررسی می‌شود.

پرسش‌های اصلی پژوهش عبارت‌اند از:

۱. آیا مدل پارس‌برت قابلیت شناسایی جنسیت نویسنده در متون فارسی را با دقت بالا دارد؟

۲. آیا پیکرهٔ احساسی برچسب‌گذاری‌شدهٔ جنسیتی نسبت‌به پیکره‌های دیگر عملکرد بهتری در این وظیفه دارد؟

فرضیه‌های پژوهش بر دو اصل استوار است:

۱. تفاوت در توزیع ویژگی‌های زبانی مانند قیدها و صفت‌های احساسی در متون زنان و مردان امکان تشخیص جنسیت را فراهم می‌آورد.

۲. مدل‌های مبتنی‌بر یادگیری عمیق مانند پارس‌برت نسبت‌به روش‌های سنتی در شناسایی جنسیت نویسنده دقیق‌تر هستند.

۲- مبانی نظری

مطالعهٔ تفاوت‌های جنسیتی در گفتار از دههٔ ۱۹۷۰ میلادی آغاز شد. لیکاف[7] (۱۹۷۵)، از چهره‌های برجسته در این دوره، ادعا کرد که زنان از الگوهای زبانی خاصی استفاده می‌کنند که موقعیت اجتماعی پایین‌تر آن‌ها را نشان می‌دهد. برای نمونه، استفاده از عبارت‌های تأییدیِ «آیا این‌طور نیست؟» یا «درست است؟» قاطع‌نبودن زنان و تمایلشان به جلب تأیید دیگران را بازنمایی می‌کند. در دهه‌های ۱۹۸۰ و ۱۹۹۰، پژوهش‌ها در این حوزه گسترش یافت و نظریه‌های پیچیده‌تری مطرح شدند. دراین‌میان، تانن[8] (۱۹۹۹) با پرداختن به تفاوت‌های جنسیتی در گفتار استدلال کرد که زنان و مردان به‌دلیل تفاوت‌های فرهنگی، سبک‌های گفتاری متفاوت دارند. به‌باور او، زنان بیشتر از «سبک گفتاری همکارانه»[9] استفاده می‌کنند که بر ایجاد ارتباط و همدلی تأکید دارد؛ درحالی‌که مردان «سبک گفتاری رقابتی»[10] را به کار می‌برند که بر انتقال اطلاعات و اثبات توانایی‌ها متمرکز است. بااین‌حال، پژوهش‌ها دربارهٔ تفاوت‌های جنسیتی در گفتار با «رویکردهای انتقادی»[11] و با این استدلال همراه شد که تفاوت‌های جنسیتی در گفتار بیش‌ازحد ساده‌انگارانه هستند و نمی‌توان آن‌ها را به همهٔ زنان و مردان تعمیم داد. برای نمونه، هولمز[12] (۲۰۱۳) نشان داد که تفاوت‌های جنسیتی در گفتار از عوامل اجتماعی، فرهنگی و موقعیتی تأثیر می‌پذیرند و نمی‌توان آن‌ها را فقط به جنسیت نسبت داد. دراین‌زمینه، برخی پژوهشگران به نقش قدرت و نابرابری‌های اجتماعی در شکل‌گیری تفاوت‌های جنسیتی در گفتار اشاره کرده‌اند. برای نمونه، براون و لوینسون[13] (۱۹۸۷)، در چارچوب «نظریهٔ ادب زبانی»[14]، نشان دادند که راهبردهای زبانی متفاوت، بیش از آنکه محصول جنسیت باشند، از ضرورت‌های اجتماعی و فرهنگی ناشی می‌شوند. در سال‌های اخیر، پژوهش‌ها دربارهٔ تفاوت‌های جنسیتی در گفتار به‌سمت بررسی نقش عوامل چندگانه مانند سن، طبقه اجتماعی، قومیت و هویت جنسیتی سوق یافته‌اند. دراین‌زمینه، برادی[15] (۱۹۹۳) معتقد است که تفاوت‌های جنسیتی در گفتار ممکن است درمیانِ گروه‌های سنی مختلف متفاوت باشد یا از هنجارهای فرهنگی خاص تأثیر پذیرد. افزون‌براین، با ظهور رسانه‌های اجتماعی و فناوری‌های دیجیتال، پژوهش‌های جدیدی دربارهٔ تفاوت‌های جنسیتی در ارتباط‌های برخط انجام شده است (چای[16] و دیگران، ۲۰۱۶). شایان ذکر است که از دیرباز، تفاوت‌های جنسیتی در گفتار مرکز توجه پژوهشگران زبان‌شناسی بوده

1. decision tree
2. support vector machine (SVM)
3. neural networks
4. long short-term memory (LSTM)
5. precision
6. recall
7 . R. Lakoff
8. D. Tannen
9. collaborative speaking style
10. coaporative speaking style
11. critical approaches
12. J. Holmes
13. P. Brown & S. C. Levinson
14. theory of linguistic literature
15. L. R. Brody
16. C. Chai

است؛ این تفاوت‌ها در سطح واژگان، دستور زبان، لحن و کاربرد زبان در موقعیت‌های اجتماعی مختلف مشاهده می‌شود. برخی از این تفاوت‌ها در عوامل زیستی ریشه دارند؛ اما بسیاری از آن‌ها از هنجارهای اجتماعی و فرهنگی ناشی می‌شوند.

۲-۱ واژگان و انتخاب کلمات

برپایهٔ مطالعات موجود، زنان تمایل بیشتری دارند که از واژگان توصیفی و احساسی استفاده کنند (نیومان[1] و دیگران، ۲۰۰۸). نیومان و دیگران (۲۰۰۸) معتقدند این تمایز را می‌توان در سه سطح تحلیل کرد: نخست در حوزهٔ استفاده از صفت‌ها، زنان بیشتر از صفات کیفی و عاطفی مانند «زیبا»، «دل‌نشین»، «شگفت‌انگیز» و «وحشتناک» برای توصیف افراد، مکان‌ها یا تجربه‌های خود بهره می‌برند؛ درحالی‌که مردان تمایل دارند از صفات عملی و عینی‌تر مانند «سریع»، «قوی»، «کارآمد» و «مفید» استفاده کنند که بر ویژگی‌های عملی تأکید دارند. دوم در سطح قیدها، زنان بیشتر از قیدهای تأکیدی و احساسی مانند «خیلی»، «واقعاً»، «کاملاً» و «فوق‌العاده» برای تقویت بار عاطفی جمله‌های خود استفاده می‌کنند؛ درمقابل، مردان بیشتر از قیدهای مربوط به دقت یا شدت عمل مانند «به‌سرعت»، «به‌طور کامل» و «به‌طور مستقیم» استفاده می‌کنند که بر جنبه‌های عملی تمرکز دارند. درنهایت، ازمنظر واژگان محتوایی، زنان با استفاده از مفاهیم مرتبط با هیجان و روابط عاطفی مانند «عشق»، «ناراحت»، «خوشحال»، «ترس» و «هیجان‌زده» احساساتشان را بیان می‌کنند که این می‌تواند بازتاب‌دهندهٔ نقش‌های اجتماعی آن‌ها در توجه به روابط عاطفی باشد. درمقابل، مردان کمتر از واژه‌های احساسی استفاده می‌کنند؛ به‌گونه‌ای که دایرهٔ واژه‌های آن‌ها بیشتر حول محور مفاهیم عمل‌گرایانه و اطلاعاتی مانند «کار»، «مسئله»، «راه‌حل» و «هدف» می‌چرخد که بیشتر بر جنبه‌های عملی و منطقی تأکید دارند.

۲-۲ سبک گفتاری

زنان معمولاً از سبک گفتاری همکارانه، مشارکتی و تعاملی بهره می‌گیرند و با استفاده از پرسش‌های ترغیبی، درپی حفظ تداوم گفت‌وگو و تحکیم پیوندهای عاطفی هستند. درمقابل، مردان سبک گفتاریِ رقابتی و مستقیم را برمی‌گزینند و بیشتر بر انتقال اطلاعات متمرکز هستند. هانتر[2] و دیگران (۲۰۰۵) معتقدند شکل‌گیری این تفاوت‌های گفتمانی، به عواملی همچون «هنجارهای اجتماعی»، «فرهنگ و ارزش‌های جامعه»، «آموزش و جامعه‌پذیری»، «رسانه‌ها و کلیشه‌های جنسیتی» و «تغییرات فرهنگی و اجتماعی» وابسته است.

۲-۳ میزان صحبت‌کردن

براساس یافته‌های پژوهش، توزیع زمانیِ سخن‌گویی میان جنسیت‌ها بسته به بافتِ گفت‌وگو تغییر می‌کند؛ به‌گونه‌ای که زنان در موقعیت‌های غیررسمی حجم بیشتری از گفت‌وگو را به خود اختصاص می‌دهند؛ درحالی‌که در بافت‌های رسمی یا عمومی، مردان ممکن است بیشتر صحبت کنند (کندال[3] و تانن، ۱۹۹۷). این تفاوت می‌تواند از نقش‌های اجتماعی سنتی ناشی شود که حضور و نفوذ مردان را در عرصه‌های عمومی تقویت می‌کند.

۳- پیشینهٔ پژوهش

چنگ[4] و دیگران (۲۰۱۱) در یکی از نخستین پژوهش‌هایشان، با مطالعهٔ برخی ویژگی‌های زبانی همچون «کلمات کاربردی»، «ویژگی‌های مبتنی‌بر کلمه» و «ویژگی‌های ساختاری» و با بهره‌گیری از الگوریتم‌های دسته‌بندی‌کننده[5] مانند «بیز ساده»، «ماشین بردار پشتیبان» و «درخت تصمیم J48»، سامانه‌ای برای شناسایی جنسیت نویسنده طراحی کردند. «میزان صحت»[6] به‌دست‌آمده از سامانهٔ آن‌ها برابر با ۸۵/۱ درصد گزارش شد. بایست در نظر داشت که طبقهٔ اجتماعی افراد نیز بر الگوی گفتار زنان و مردان تأثیرگذار است. شایان توجه است که مطالعات تجربی و نظری انجام‌شده دراین‌زمینه به حوزهٔ زبان‌شناسی اجتماعی مربوط می‌شوند. برای نمونه، سیماکی[7] و دیگران (۲۰۱۶) با تمرکز بر ویژگی‌های زبان‌شناسی اجتماعی به شناسایی سن و جنسیت کاربران رسانه‌های اجتماعی

1. M. L. Newman
2. D. Hunter
3. S. Kendall
4. N. Cheng
5. classifier
6. accuracy
7. V. Simaki

پرداختند. در این مطالعه، شاخص‌های آماری متعددی همچون «بسامد واژه‌ها»، «باهم‌آیی واژه‌ها»[1]، «برچسب‌های نحوی» و «برچسب‌های جنسیتی و سن افراد» بررسی شدند و درنهایت، هریک از ویژگی‌ها براساس نوع کارکردشان تحلیل شدند. یک تحلیل از ویژگی‌های پربازده‌تر براساس ماهیت هر ویژگی انجام شد و اطلاعات به‌دست‌آمده پس از امتیازدهی اهمیت ویژگی‌ها بحث شد.

افزون‌براین، السمرات[2] و دیگران (۲۰۱۷) معتقدند که «کیسه‌واژه‌ها»[3] و تصریف کلمات از مؤلفه‌های به‌کاررفته در زبان عربی به شمار می‌روند. دراین‌زمینه، رنگل[4] و دیگران (۲۰۱۹) با مرور سامانه‌های تشخیص جنسیتِ زبان عربی در عناوین خبری و توییتر نشان دادند که دو روش «ماشین بردار پشتیبان» و «شبکه‌های عصبی»، همواره از پرکاربردترین و کارآمدترین دسته‌بندی‌ها در این زبان بوده‌اند. همچنین الوارز کارمونا[5] (۲۰۱۸) با تمرکز بر داده‌های نوشتاری زبان انگلیسی و تصویری توییتر و با به‌کارگیری روش «ماشین بردار پشتیبان» برای شناسایی جنسیت نویسنده نشان داد که این روش، از میزان دقت و صحت قابل‌اعتمادی برخوردار است.

اسبو[6] و دیگران (۲۰۱۸) نیز با تمرکز بر زبان روسی و با بهره‌گیری از «شبکه‌های عصبی پیچشی»[7] و «حافظهٔ طولانی کوتاه‌مدت»، جنسیت نگارنده را شناسایی کردند. آن‌ها نتایج پژوهش را با الگوریتم‌های دسته‌بندی‌کننده‌ای همچون «بیز ساده» و «ماشین بردار پشتیبان» مقایسه کردند؛ برپایهٔ این مقایسه، بهبود میزان صحت نتایج و کاهش میزان تابع هزینه مشاهده شد. علاوه‌براین، سفارا[8] و دیگران (۲۰۲۰) در حوزهٔ زبان انگلیسی، با استفاده از «شبکه‌های عصبی» و «الگوریتم بهینه‌ساز وِیل»[9] به شناسایی جنسیت نویسنده پرداختند. براساس یافته‌های آن‌ها بر مجموعه‌پیکره‌های انرون[10] مشخص شد که سامانهٔ طراحی‌شده نسبت‌به نسبت‌به ماشین‌های یادگیری همچون «بیز ساده»، «ماشین بردار پشتیبان» و «درخت تصمیم J48» عملکرد مطلوب‌تری داشته است؛ به‌طوری که دقت و بازخوانی آن به‌ترتیب ۹۷/۱۶ درصد و ۹۹ درصد گزارش شد.

در ادامه، درزمینهٔ شناسایی جنسیت نویسنده به پژوهش‌هایی در زبان فارسی پرداخته می‌شود. مرادی و بحرانی (۱۳۹۴) با بهره‌گیری از ویژگی‌های «روان‌شناختی»[11] و «سبک‌شناختی»[12] و به‌کارگیری ماشین‌های یادگیری «بیز ساده»، «درخت تصمیم J48» و «ماشین بردار پشتیبان» به شناسایی جنسیت نویسنده پرداختند. در پژوهش آن‌ها، بیشترین میزان صحت به‌دست‌آمده ۷۳/۸ درصد گزارش شد. ساجدی و تسلیمی (۱۳۹۸) نیز در پژوهشی دیگر با تمرکز بر بلاگ‌ها، از ویژگی‌های نحوی، مبتنی‌بر واژه، مبتنی‌بر حروف، واژگان گرامری و ویژگی‌های «ان–تایی»[13] حروف استفاده کردند. بنابر نتایج پژوهش آن‌ها، به‌کارگیری الگوریتم «جنگل تصادفی بیز»[14]، میزان صحت دسته‌بندی را تا ۸۹/۵ درصد افزایش داد. در پژوهشی دیگر، ظریفی و نقوی (۲۰۲۲) روشی برای بهبود تشخیص جنسیت نویسنده در متون کوتاه فارسی ارائه دادند. آن‌ها با استفاده از روش «بردارسازی مفهومی»[15] که برپایهٔ انجام عملیات ریاضی بر بردارهای کلمات استوار است، کلمات را به بردارهای معنایی تبدیل کردند. این پژوهشگران با آموزش بردارهای استخراج‌شده از مجموعه‌دادهٔ اختصاصی خود به یک شبکهٔ عصبی و با استفاده از روش «اعتبارسنجی متقابل ۵–تایی»[16] به میزان صحت ۸۱/۰۹ درصد دست یافتند. همچنین حکمتیان زاده‌پور و جلالی بیدگلی (۱۳۹۸) مسئلهٔ شناسایی خودکار جنسیت براساس نظرات نوشته‌شده به زبان فارسی در سایت مایکت را بررسی کردند. آن‌ها با مقایسهٔ روش‌های کیسهٔ کلمات و ان–تایی و با استفاده از الگوریتم‌های «بیز ساده» و «شبکه های عصبی پرسپترون چندلایه»[17] دریافتند که ترکیب علائم نگارشی با رویکرد ترکیبی یک‌گرام، دوگرام و سه‌گرام و الگوریتم

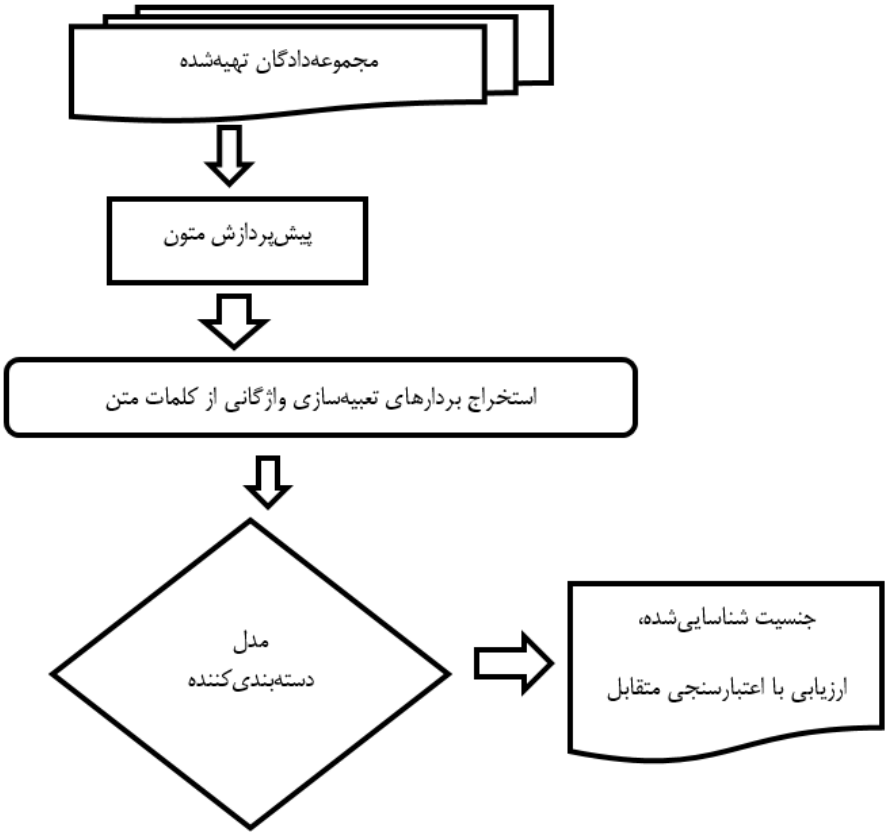
1. co-occurrence of words
2. K. Alsemearat
3. bag of words
4. F. Rangel
5. M. A. Alvarez-Carmona
6. A. Sboev
7. convolutional neutral networks
8. F. Safara
9. whale optimizer algorithm
10. Enron
11. psychologic
12. stylistic
13. n-gram
14. Bayesian random forest
15. conceptual vectorization
16. 5-fold cross validation
17. multilayer perceptron neural networks

شبکهٔ عصبی با افزایش صحت همراه است؛ به‌گونه‌ای که بالاترین میزان صحت در این مطالعه ۸۵,۷ درصد گزارش شد.

درنهایت، در این پژوهش با مرور پیشینهٔ پژوهش درزمینهٔ شناسایی جنسیت و بررسی مدل‌های نوین پردازش زبان طبیعی، سامانه‌ای برای شناسایی جنسیت برپایهٔ مدل پارس‌برت طراحی شد که به‌طور اختصاصی برای زبان فارسی توسعه یافته است. سپس برای ارزیابی عملکرد سامانهٔ طراحی‌شده، نتایج پژوهش حاضر با پژوهش‌های پیشین و سایر دسته‌بندی‌کننده‌ها مقایسه شد.

۴- روش پژوهش

هدف از پژوهش حاضر، تعیین خودکار جنسیت نگارنده در متون زبان فارسی است. ازآنجاکه نحوهٔ بیان احساسات، یکی از مؤلفه‌های متمایزکننده در رفتار زبانی زنان و مردان است، در این مطالعه، از پیکرهٔ احساسی صادقی و دیگران (۲۰۲۱) استفاده شد. این پیکره که مشتمل بر ۲۳۰۰۰ جملهٔ احساسی، شاملِ ترس، غم، شادی، خشم و حیرت، است، براساس متون روایی و به‌صورت دستی و ماشینی گردآوری شد. در مرحلهٔ بعد، برای آماده‌سازی داده‌ها، تعداد ۵۰۰۰ سند از این پیکرهٔ احساسی با توازن جنسیتی (۲۵۰۰ سند برای هر جنسیت) استخراج و برچسب‌گذاری شد. در ادامه، برای افزایش میزان دقت سامانه، متون انتخابی پیش‌پردازش شدند؛ سپس این داده‌ها برای آموزشِ مدل‌های دسته‌بندی‌کننده‌ای همچون بیز ساده، درخت تصمیم J48، ماشین بردار پشتیبان و مدل‌های یادگیری عمیق مانند حافظهٔ طولانی کوتاه‌مدت و پارس‌برت به کار گرفته شدند. درنهایت، کارایی مدل‌های ارائه‌شده با استفاده از روش اعتبارسنجی متقابل ارزیابی شد. در شکل (۱)، روند کلی طراحی سامانه نمایش داده شده است.



شکل (۱). ترسیم معماری سامانهٔ پیشنهادی

۴-۱ تهیهٔ پیکره

پیکرهٔ احساسی موردبررسی (صادقی و دیگران، ۲۰۲۱)، مشتمل بر احساس خشم، غم، حیرت، ترس و شادی است. برای هریک از این احساسات، ۱۰۰۰ سند به‌صورت متوازن انتخاب شد؛ به‌گونه‌ای که در هر دسته، ۵۰۰ سند به مردان و ۵۰۰ سند به زنان تعلق گرفت. در مرحلهٔ انتخاب جمله‌ها، با استناد به مبانی نظری بحث‌شده، تفاوت‌های زبانی میان جنسیت‌ها مانند میزان استفاده از واژه‌ها و نقش واژه‌ها به‌ویژه قیدها و صفت‌ها مدنظر قرار گرفت. مشخصات کلی این پیکره در جدول (۱) شرح داده شده است.

جدول (۱). مشخصات آماری پیکرهٔ تهیه‌شده

بسامد اسناد	نوع اسناد تهیه‌شده
۵۰۰۰	تعداد کل اسناد
۲۵۰۰	تعداد اسناد برای جنسیت مرد
۲۵۰۰	تعداد اسناد برای جنسیت زن
۲۱	میانگین تعداد واژه‌ها در هر سند برای مرد
۲۰	میانگین تعداد واژه‌ها در هر سند برای زن

نمونه‌هایی از جمله‌های موجود در پیکرهٔ صادقی و دیگران (۲۰۲۱)، که براساس احساسات و جنسیت نگارنده برچسب‌گذاری شده‌اند، در ادامه آورده شده است:

- زن:

۱- من از شوهرم بیزارم و از خودم هم بیزارم.
۲- زهرا می‌گفت: خواهر آن‌قدر افسرده می‌شدم و اندوه به سراغم می‌آمد که بی‌دلیل گریه می‌کردم.
۳- در همان شب بود که افسوس خوردم که چرا مانند دیگر هم‌جنسان خود راه‌ورَسم برخورد با مردها را نیاموخته‌ام.
۴- به حدی بیان این جمله تحکم‌آمیز و آمرانه بود که من تعجب کردم چطور جرئت می‌کند این‌طور به من امرونهی کند.
۵- این کارت واقعاً منو خوشحال کرد، خیلی ممنونم.
۶- دلم می‌خواهد بدونم نظرت درمورد این موضوع چیه.
۷- امروز یک روز واقعاً سخت بود، خیلی استرس داشتم.
۸- دوست دارم بیشتر درمورد آینده‌مون باهم صحبت کنیم.
۹- این مسئله واقعاً منو نگران کرده، نمی‌دونم چیکار کنم.

- مرد:

۱۰- از نوع رفتاری که زنش و مادرش داشتند، متنفر بود.
۱۱- امروز خیلی کار داشتم، خسته‌ام؛ ولی باید ادامه بدم.
۱۲- نگران نباش، همه‌چیز درست می‌شه.
۱۳- این مسئله رو باید سریع حل کنیم، وقت تلف نکنیم.
۱۴- تو زن کثیف و پست نه‌تنها به شوهر اولت وفادار نبودی، بلکه به من هم خیانت کردی.
۱۵- درمیان شادی پدرشدن ترس داشتم که مبادا او را از دست بدهم.
۱۶- شاهرخ وحشت‌زده به‌طرفش برگشت و برجا خشکش زد.
۱۷- مرتیکه فلان‌فلان‌شده صبر کن تا حسابت را برسم.

در مرحلهٔ پیش‌پردازش کارهایی همچون «هنجارسازی»[1]، «ریشه‌یابی»[2] و «حذف ایست‌واژه‌ها»[3] به‌منظور آماده‌سازی متن برای پردازش اصلی انجام شد؛ این فرایند به‌طور مستقیم موجب می‌شود میزان دقت و عملکرد سامانه ارتقا یابد. بااین‌حال، چالش‌های متعددی در پردازش زبان فارسی وجود دارد که از درخور توجه‌ترین آن‌ها می‌توان به مشکلات نگارشی اشاره کرد؛ برای نمونه، استفادهٔ نابجا از فاصله به‌جای نیم‌فاصله، رعایت‌نکردن علائم نگارشی فارسی و به‌کارگیری حروف غیرعربی از مواردی هستند که سلامت نگارشی متن را مختل می‌کنند. افزون‌براین، وجود واژه‌های چندامالایی و دشواری در شناسایی واژه‌های مرکب و درنظرگرفتن آن‌ها به‌منزلهٔ یک واحد واژگانی از دیگر چالش‌های موجود است. درمجموع، همان‌طور که قیومی و ممتازی (۲۰۰۹) در بررسی چالش‌های توسعهٔ پیکره برای زبان فارسی بیان کردند در پردازش الکترونیکی زبان فارسی مشکلات متعددی وجود دارد که با هنجارسازی بسیاری از آن‌ها برطرف می‌شود. تأثیر تاریخی زبان عربی بر زبان فارسی، املای واژه‌های فارسی و ترکیب صفحه‌های کد فارسی و عربی در

1. normalization
2. stemming
3. removal of stop words

سیستم‌عامل‌ها از عواملی است که در پردازش الکترونیکی متون فارسی مشکلاتی ایجاد کرده است (قیومی و ممتازی، ۲۰۰۹).

ریشه‌یابی واژه‌ها با هدفِ درنظرگرفتن یک واحد یکسان برای واژه‌های دارای ریشهٔ مشترک انجام شد. بازگردانی واژه‌ها به ریشهٔ اصلی، دقت شمارش بسامد واژه‌ها را افزایش می‌دهد و درنهایت، بر عملکرد سامانه تأثیر مثبت می‌گذارد. در این پژوهش، در فرایند ریشه‌یابی از برنامهٔ «ریشه‌یاب»[1] معنایی در زبان فارسی استفاده شد که محصول آزمایشگاه فناوری وب دانشگاه فردوسی مشهد[2] است. همچنین ایست‌واژه‌ها که واژه‌های دستوری مانند حروف ربط و حروف اضافه را دربرمی‌گیرند و بار معنایی و محتوایی مرتبط با متن ندارند، از متن حذف شدند. ویژگی این واژه‌ها، داشتن شکل یکسان در متون مختلف با موضوع‌های متفاوت است. بنابراین، در این سامانه پیش از پردازش اصلی، فهرستی از ایست‌واژه‌ها تهیه و برای پاک‌سازی متن ورودی به کار گرفته شد.

۴-۲ تعبیه‌سازی متن

پس از آماده‌سازی پیکره و اِعمال هنجارسازی برای پردازش اصلی، در گام بعد، ویژگی‌ها از متن به‌منظور آموزش مدل‌های دسته‌بندی‌کننده استخراج شدند. مفهوم اصلی تعبیه‌سازی متن، بازنمایی همهٔ واژه‌های یک زبان با مجموعه‌ای از اعداد اعشاری در قالب یک بردار است. تعبیه‌سازی‌ها بردارهایی چندبعدی هستند که معنای واژه‌ها و محتوای آن‌ها را با مقادیر عددی بازنمایی می‌کنند؛ به‌این‌ترتیب، اطلاعات زبانی به‌صورت بردارهای عددی در اختیار مدل قرار می‌گیرد.

هدف اصلی از به‌کارگیری بردارهای واژگانی، استخراج بردارهایی باکیفیت از حجم عظیمی از داده‌ها است. در این پژوهش، با استفاده از پیکرهٔ تهیه‌شده، از بردارهای واژگانی Word2Vec و مدل کیسه‌واژه‌ها به‌منزلهٔ ویژگی‌های آماری برای افزایش میزان دقت و عملکرد سامانه استفاده شد. تعداد مؤلفه‌های بردار برای هر واژه مطابق با استانداردهای رایج، ۱۰۰ در نظر گرفته شد. برای هر سند نیز برداری از میانگین طول بردارهای موجود در سند ساخته و به‌منزلهٔ ویژگی آماری سند استفاده شد. آشکارسازی روابط معنایی مخفی بین کلمات همچون «زن» و «مرد» یا «شادی» و «خوشحالی»، از مزایای این روش است که به‌دلیل کارایی بالا به‌سرعت درمیان پژوهشگران رواج یافته است. دراین‌بین، واژه‌های همسایه در یک جمله، به‌مثابهٔ بافتی که واژه در آن قرار دارد، بر میزان اهمیت و بار معنایی واژه اثر می‌گذارند.

پس از آماده‌سازی داده‌های آموزشی به‌صورت بردارهای عددی و برچسب‌گذاری آن‌ها، از مدل‌های دسته‌بندی‌کننده برای تعیین جنسیت نویسنده استفاده شد. در این پژوهش، از الگوریتم‌ها و مدل‌های متعدد برای دسته‌بندی متون بهره گرفته شد که عبارت‌اند از: بیز ساده، ماشین بردار پشتیبان، درخت تصمیم J48، حافظهٔ طولانی کوتاه‌مدت و پارس‌برت.

۴-۳ مدل پارس‌برت

در مسیر تکامل الگوریتم‌های درک بهتر زبان، استفاده از الگوریتم برت (دولین[3] و دیگران، ۲۰۱۸) یکی از تحول‌های بنیادین محسوب می‌شود. گوگل با معرفی الگوریتم برت، گامی بلند در فهم محتوا برداشته است. این مدل چارچوب یادگیری عمیقی را ارائه می‌دهد که می‌تواند به‌صورت رایگان برای عموم افراد در پردازش زبان طبیعی به کار رود. مدل‌های مبتنی‌بر برت با تحلیل عبارت‌های پیش و پس از یک واژه، بافت کامل آن واژه را در فرایند تعبیه و بازنمایی واژه‌ها در نظر می‌گیرند. دراین‌بین، مدل پارس‌برت زیرشاخه‌ای از الگوریتم برت است که به‌منظور کسب دقت بیشتر برای زبان فارسی طراحی شده است (فراهانی و دیگران، ۲۰۲۱). معماری الگوریتم برت یک «رمزگذار مبدل»[4] در دو جهت چندلایه است که براساس پیاده‌سازی اصلی توصیف‌شده (واسوانی[5] و دیگران، ۲۰۱۷) در کتابخانهٔ tensor2tensor منتشر شده است.

در این پژوهش، مدل پارس‌برت برای انجام وظیفهٔ تشخیص جنسیت نویسنده، با استفاده از پیکرهٔ تهیه‌شده، تنظیم دقیق شده است. میزان دادهٔ آموزشی درون بسته‌ها (اندازهٔ بسته)[6] برابر با ۳۲ و «میزان یادگیری»[7] نیز برابر با 10^{-4} در نظر گرفته شده است. همچنین

1. stemmer
2. http://wtlab.um.ac.ir
3. J. Devlin
4. transformer encoder
5. A. Vaswani
6. batch size
7. learning rate

ویژگی‌های پیکربندی الگوریتم برت اصلی شامل ۱۲ «لایهٔ پنهان»[1]، ۱۲ «عنوان توجه»[2] و ۷۶۸ بُعد در لایهٔ پنهان حفظ شده است. برای تنظیم دقیق مدل از «بهینه‌ساز آدام»[3] با مقادیر β_1=۰/۹ و β_2=۰/۹۸ استفاده شده است (کینگما و با[4]، ۲۰۱۴). درنهایت، برای پیش‌بینی دسته‌های خروجی (زن و مرد) در وظیفهٔ تعیین جنسیت نویسنده، یک لایهٔ «پیش‌خور»[5] با «تابع فعال‌سازی بیشینهٔ نرم»[6] به مدل پارس‌برت اضافه شده است تا عملیات دسته‌بندی تعبیهٔ متنی حاصل از پارس‌برت را انجام دهد.

۵- تحلیل داده‌ها

با اِعمال الگوریتم‌های دسته‌بندی بر داده‌های آموزشی تهیه‌شده، برای ارزیابی عملکرد مدل، از روش اعتبارسنجی متقابل ۱۰-تایی استفاده شد و کارایی سامانه براساس معیارهای دقت و بازخوانی سنجیده شد. در ادامه، برای تحلیل دقیق‌تر، نتایج پژوهش با روش‌های رایج یادگیری ماشین و مدل حافظهٔ طولانی کوتاه‌مدت مقایسه شد (جدول ۳). همچنین برای تعبیه‌سازی متن در دسته‌بندی‌های متداول «بیز ساده»، «ماشین بردار پشتیبان» و «درخت تصمیم J48» از بردارهای «فراوانی واژه-معکوس فراوانی سند»[7] و میانگین بردارهای Word2Vec استفاده شد. در مدل حافظهٔ طولانی کوتاه‌مدت نیز تعبیه‌سازی واژگان با استفاده از روش Word2Vec انجام گرفت.

مدل حافظهٔ طولانی کوتاه‌مدت گونه‌ای اصلاح‌شده از «شبکه‌های عصبی بازگشتی»[8] است. در این پژوهش، پارامترهای مختلفی برای آموزش این مدل با داده‌های پژوهش بررسی و تنظیم شد. درنهایت، براساس نتایج حاصل از ارزیابی‌ها، بهینه‌ترین مجموعه‌پارامترها برای مدل حافظهٔ طولانی کوتاه‌مدت در جدول (۲) ارائه شده است.

جدول (۲). مشخصات پارامترها

پارامترها	مقادیر
بُعد بردارهای تعبیه‌شدهٔ واژگانی	۱۰۰
اندازهٔ بسته‌ها	۳۲ تایی
مقدار حذف تصادفی	۰/۲
تعداد دور[9]	۵
تابع فعال‌سازی	بیشینهٔ نرم
تابع زیان[10]	آنتروپی متقاطع طبقه‌بندی‌شده[11]
بهینه‌ساز	آدام

بنابر نتایج حاصل از پیاده‌سازی مدل‌ها بر پیکرهٔ پژوهش (جدول ۳)، مدل پارس‌برت در تشخیص جنسیت نویسنده از سایر مدل‌ها عملکرد مطلوب‌تری دارد؛ به‌گونه‌ای که برای این مدل دقتِ برابر با ۷۷ درصد، بازخوانی برابر با ۷۶ درصد و معیار-اف[12] برابر با ۷۶/۵ درصد گزارش شده است. برتری مدل پارس‌برت را می‌توان به توانایی بالاتر این مدل در تعبیه‌سازی و نمایش متن در مقایسه با روش‌های TF-IDF و Word2Vec نسبت داد. از مزایای روش نام‌برده شناسایی ویژگی‌هایی است که به‌طور آشکار مشاهده نمی‌شوند و در سطح بالای انتزاع، در دو مرحلهٔ «آموزش و تنظیم دقیق» و «آزمون» به کار برده می‌شوند. مدل‌های مبتنی‌بر «ترنسفورمر»[13] مانند پارس‌برت به‌دلیل بهره‌گیری از سازوکار «خودتوجه»[14] قادرند روابط معنایی میان واژه‌ها و جمله‌ها را در سطح عمیق‌تری استخراج کنند. این مدل‌ها نه‌تنها بر واژه‌های آشکار متن بلکه بر الگوهای پنهان در سبک نوشتاری و ساختار جمله‌ها تمرکز دارند؛ حتی لحن

1. hidden layer
2. attention head
3. adam optimizer
4. D. P. Kingma & J. Ba
5. feedforward
6. softmax activation function
7. TF-IDF
8. recurrent neural networks (RNN)
9. epochs
10. loss function
11. categorical cross entropy
12. F-measure
13. transformer
14. self-attention

نویسنده را شناسایی می‌کنند. برای نمونه، تفاوت‌های ظریف در کاربرد ضمایر، افعال، صفات و حتی طول جمله‌ها را که میان نویسندگان زن و مرد وجود دارد، پارس‌برت می‌تواند تشخیص دهد.

جدول (۳). مقایسهٔ نتایج مدل‌های دسته‌بندی

روش‌های پیاده‌سازی	دقت	بازخوانی	معیار-اف
بیز ساده	۵۸	۶۱	۵۹/۵
ماشین بردار پشتیبان	۶۲	۶۲	۶۲
درخت تصمیم J48	۶۳	۶۴	۶۳/۵
حافظهٔ طولانی کوتاه‌مدت	۷۰	۶۹	۶۹/۵
پارس‌برت	۷۷	۷۶	۷۶/۵

در این پژوهش، میزان اثربخشی پیکرهٔ تهیه‌شده بر بهبود عملکرد سامانهٔ شناسایی جنسیت نویسنده در مقایسه با دیگر پیکره‌های موجود ارزیابی شد. به این منظور، «پیکرهٔ تشخیص خودکار جنسیت»[1] به کار گرفته شد. این پیکره را که مرادی و بحرانی (۱۳۹۴) تهیه کرده‌اند از سه بخشِ «اسناد پیکرهٔ گفتاری»، «نظرات کاربران در سایت «هلوکیش»[2] و «داستان‌های روایی» تشکیل شده است. مشخصات این پیکره در جدول (۴) ارائه شده است. در ادامه، ابتدا چند دسته‌بندی متداول شاملِ بیز ساده، درخت تصمیم J48 و ماشین بردار پشتیبانی بر دو پیکره پیاده‌سازی و ارزیابی شده‌اند. نتایج مربوط به میزان دقت به‌دست‌آمده نیز براساس ارقام موجود در جدول‌های مقالهٔ «پیکرهٔ تشخیص خودکار جنسیت» گزارش شده است.

جدول (۴). مشخصات پیکرهٔ مرادی و بحرانی (۱۳۹۴)

اسناد بر اساس جنسیت نویسنده	تعداد سند	میانگین تعداد کلمات در هر سند
اسناد برای جنسیت مرد	۱۴۴۳	۱۷۸۷۵
اسناد برای جنسیت زن	۱۴۴۳	۲۸۷۷۸

در ادامه، هر دو پیکرهٔ پژوهش حاضر و پیکرهٔ «تشخیص خودکار جنسیت» در مدل پارس‌برت و مدل‌های دسته‌بندی‌های متداول ارزیابی شدند. نتایج حاصل از این ارزیابی در جدول‌های (۵) و (۶) ارائه شده است. برپایهٔ این یافته‌ها، میزان عملکرد مدل پارس‌برت و عملکرد مدل‌های رایج براساس معیار صحت، بر پیکرهٔ احساسی جنسیتی تهیه‌شده در این پژوهش نسبت‌به پیکرهٔ تهیه‌شده در مرادی و بحرانی (۱۳۹۴) بهبود یافته است. وجه تمایز و نقطهٔ قوت پژوهش حاضر در مدیریت تعداد ویژگی‌های موردبحث نهفته است. برخلاف پژوهش مرادی و بحرانی (۱۳۹۴) که در آن بیش از ۲۰ ویژگی باتکیه‌بر دانش انسانی استخراج شده است، در پژوهش حاضر با بهره‌گیری از مدل پارس‌برت، ویژگی‌ها به‌صورت خودکار و با استفاده از روش‌های آماری از درون متن استخراج شده‌اند که به نتایج مطلوب‌تری انجامیده است.

جدول (۵). مقایسهٔ صحت مدل‌های دسته‌بندی متداول

پیکره	بیز ساده	درخت تصمیم J48	ماشین بردار پشتیبان
پیکرهٔ احساسی جنسیتی (پژوهش حاضر)	۶۱	۶۸	۶۵
پیکرهٔ تشخیص خودکار جنسیت (مرادی و بحرانی، ۱۳۹۴)	۶۰	۶۸	۶۴

جدول (۶). مقایسهٔ پیکرهٔ پژوهش حاضر با پیکرهٔ پژوهش‌های دیگر براساس مدل پارس‌برت

پیکره	دقت	بازخوانی	صحت	معیار-اف
پیکرهٔ احساسی جنسیتی (پژوهش حاضر)	۷۷	۷۶	۷۶	۷۶/۵
پیکرهٔ تشخیص خودکار جنسیت (مرادی و بحرانی، ۱۳۹۴)	۷۷	۷۵	۷۳	۷۶

همان‌طور که در جدول‌های (۳) و (۶) مشاهده می‌شود، میزان صحت دسته‌بندی با بهره‌گیری از مدل پارس‌برت ۷۶ درصد است؛ به این معنا که جنسیت نویسنده در ۲۴ درصد از متون به‌درستی تعیین نشده و تعیین جنسیت نویسنده در این موارد با خطا همراه بوده است. بروز این خطاها می‌تواند از عوامل متعددی ناشی شود که درخور توجه‌ترین آن‌ها، وجود متون دارای ابهام جنسیتی است. در برخی

1. https://www.peykaregan.ir/dataset/پیکره-تشخیص-خودکار-جنسیت
2. HelloKish

نمونه‌ها، ساختار جمله‌ها و عبارت‌ها به‌گونه‌ای است که برای هر دو جنسیت مرد و زن یکسان به کار می‌رود؛ به‌طوری که تعیین جنسیت آن‌ها حتی برای یک ارزیاب انسانی نیز دشوار است. برای نمونه، جملهٔ (۱۸) در پیکرهٔ حاضر به یک نویسندهٔ مرد تعلق دارد؛ درحالی‌که همین جمله را به همین شکل یک نویسندهٔ زن هم می‌توانست بیان کند.

۱۸- شاهرخ وحشت‌زده به طرفش برگشت و برجا خشکش زد.

بررسی خطاهای مدل پارس‌برت نشان می‌دهد که بخش عمده‌ای از خطاها در دستهٔ مذکور قرار می‌گیرند. نوع دیگر خطاها به متونی مربوط می‌شود که در آن‌ها سبک نگارش با جنسیت واقعی نویسنده همخوانی ندارد؛ به‌گونه‌ای که نویسنده از عبارت‌هایی با ویژگی‌های جنسیتی متفاوت از جنسیت خود استفاده کرده است. ازآنجاکه پیکرهٔ تهیه‌شده در این پژوهش متشکل از متون روایی است، وجود چنین مواردی در داده‌های آموزشی، موجب انحراف در فرایند تنظیم دقیق مدل پارس‌برت و درنهایت بروز خطا در تشخیص شده است. در این موارد نیز تشخیص جنسیت نویسندهٔ واقعی، حتی برای ارزیاب انسانی، به‌آسانی امکان‌پذیر نیست.

نوع سوم خطاها شامل نمونه‌هایی است که در آن‌ها مدل پارس‌برت در تشخیص جنسیت نویسنده دچار اشتباه شده است؛ درحالی‌که این تشخیص برای ارزیاب انسانی به‌وضوح امکان‌پذیر است. این خطاها به‌طور معمول از ضعف در یادگیری و تنظیم دقیق پارامترهای مدل به‌دلیل حجم ناکافی داده‌ها در پیکره ناشی می‌شود. بررسی خطاها نشان می‌دهد که این دسته، بخش کوچکی از کل خطاها را تشکیل می‌دهند و پیش‌بینی می‌شود که با افزایش حجم پیکرهٔ جنسیتی، میزان بروز آن‌ها کاهش یابد.

۶- نتیجه‌گیری

امروزه پیشرفت روزافزون سامانه‌های پردازش زبان طبیعی تعامل میان انسان و ماشین را آسان کرده است. در پژوهش حاضر با هدف شناسایی جنسیت نویسنده در متون فارسی، از روش‌های نوین تعبیه‌سازی متن و مدل پارس‌برت بهره گرفته است. به‌این‌ترتیب، از برجسته‌ترین دستاوردهای این پژوهش، تهیهٔ پیکرهٔ برچسب‌گذاری‌شدهٔ جنسیتی از متون احساسی زبان فارسی است که با ارائهٔ دسترسی آزاد به پژوهشگران، بستری برای مطالعات آینده فراهم می‌آورد. این پیکره شامل ۵۰۰۰ سند متنی (۲۵۰۰ سند برای هر جنسیت) است که با تمرکز بر احساسات مختلفِ خشم، غم، شادی، ترس و حیرت و با درنظرگرفتن تفاوت‌های زبانی میان زنان و مردان، امکان تحلیل عمیق‌تر سبک‌های نوشتاری را ایجاد می‌کند.

استفاده از مدل‌های نوین در پردازش متن فارسی از دیگر ویژگی‌های شاخصِ این پژوهش است. برخلاف مطالعات پیشین که بیشتر بر مدل‌های سنتی یادگیری ماشین مانند بیز ساده، درخت تصمیم J48 و ماشین بردار پشتیبان استوار بوده‌اند در این پژوهش با به‌کارگیری مدل پارس‌برت، گامی به‌سوی استفاده از روش‌های پیشرفته برداشته شده است.

بهره‌گیری از ویژگی‌های زبانی موجود در بافت متن برای دستیابی به نتایج دقیق‌تر از دیگر مزایای پژوهش حاضر است. به‌کارگیری اطلاعات زبانی در چهار حوزهٔ صرف، نحو، معنا و آوا در فعالیت‌های پژوهشی، با ارتقای درک ساختار، معنا و بافت متن، دقت و عمق تحلیل‌ها را افزایش می‌دهد. بی‌توجهی به این اطلاعات زبانی می‌تواند پژوهش را به سطحی‌ترین لایه‌های داده محدود کند و به نتایج نادرست یا ناقص منجر شود. برای نمونه، در تحلیل احساسات یا شناسایی موجودیت‌ها، نادیده‌گرفتن اطلاعات زبانی می‌تواند دقتِ تشخیص معنای واقعی جمله‌ها را با خطا مواجه کند. درمقابل، استفاده از این اطلاعات، کیفیت و قابلیت اطمینان پژوهش را بهبود می‌دهد و به ارائهٔ نتایج معنادارتر کمک می‌کند. بنابراین، پس از تهیه و برچسب‌گذاری داده‌های موردنیاز براساس دو جنسیت زن و مرد، مدل پارس‌برت، که یک لایهٔ پیش‌خور نیز به آن افزوده شده است، با استفاده از این داده‌ها برای انجام وظیفهٔ تشخیص خودکار جنسیت تنظیم دقیق شده است. به‌این‌ترتیب، مدل پارس‌برت یک تعبیه‌سازی متنی از متون ورودی را فراهم می‌کند و لایهٔ پیش‌خور اضافه‌شده، کار دسته‌بندی متن ورودی را به دو دستهٔ زن و مرد انجام می‌دهد.

علاوه‌بر مدل پیشنهادی، شناسایی جنسیت نویسنده با بهره‌گیری از مدل‌های مختلف دیگر مانند بیز ساده، درخت تصمیم J48، ماشین بردار پشتیبان و حافظهٔ طولانی کوتاه‌مدت انجام شده است. در این مدل‌ها از تعبیه‌سازی‌های متنی و واژگانی مبتنی‌بر روش‌های Word2Vec استفاده شده و عملکرد مدل‌ها با روش اعتبارسنجی متقابل ۱۰-تایی ارزیابی شده است. برپایهٔ نتایج، مدل پارس‌برت با کسب میزان دقت ۷۷ درصد، بازخوانی ۷۶ درصد و معیار-اف ۷۶/۵، مطلوب‌ترین عملکرد را در مقایسه با دیگر مدل‌ها ارائه می‌دهد. همچنین با مقایسهٔ پیکرهٔ جنسیتی دیگری با پیکرهٔ برچسب‌گذاری‌شدهٔ پژوهش حاضر مشخص شده است که پیکرهٔ احساسی

برچسب‌گذاری‌شده در این پژوهش در تشخیص جنسیت نویسنده با استفاده از مدل پارس‌برت، کارایی مطلوبی داشته است. درنهایت، براساس یافته‌ها می‌توان نتیجه گرفت که استفاده از مدل‌های نوین مانند پارس‌برت و توجه به اطلاعات زبانی در این پژوهش، گامی مهم به‌سوی ارتقای پردازش زبان طبیعی در زبان فارسی به‌منزلهٔ زبانی با منابع پردازشی محدود محسوب می‌شود. پیشنهاد می‌شود در پژوهش‌های آینده از پیکره‌های گسترده‌تر با تنوع موضوعی بیشتر و مدل‌های پیشرفته‌ترِ مبتنی‌بر ترنسفورمرها برای ارتقای دقت و کارایی سامانه‌ها بهره گرفته شود.

منابع

حکمتیان زاده‌پور، سیامک؛ جلالی بیدگلی، امیر (۱۳۹۸، اردیبهشت). *تشخیص خودکار جنسیت نویسنده نظرات نوشته‌شده به زبان فارسی*. سومین همایش ملی دانش و فناوری مهندسی برق، کامپیوتر و مکانیک ایران.

ساجدی، هدیه؛ تسلیمی، مهناز (۱۳۹۸). تشخیص جنسیت نویسندگان از روی متون با استفاده از جنگل تصادفی بیز. *پردازش علائم و داده‌ها*، ۱۶ (۱)، ۱۴۳-۱۵۶.

مرادی، مهدی؛ بحرانی، محمد (۱۳۹۴). تشخیص خودکار جنسیت نویسنده در متون فارسی. *پردازش علائم و داده‌ها*، ۱۲ (۴)، ۸۳-۹۴.

References

Alsmearat, K., Al-Ayyoub, M., Al-Shalabi, R., & Kanaan, G. G. (2017). Author gender identification from Arabic text. *Journal of Information Security and Applications*, *35*, 85–95.

Alvarez-Carmona, M. A., Pellegrin, L., Montes-y-Gómez, M., Sánchez-Vega, F., Escalante, H. J., López-Monroy, A. P., Villasenor-Pineda, L., & Villatoro-Tello, E. (2018). A visual approach for age and gender identification on Twitter. *Journal of Intelligent and Fuzzy Systems*, *34*(5), 3133–3145.

Brody, L. R. (1993). On understanding gender differences in the expression of emotion: Gender roles, socialization, and language. In S. L. Ablon, D. Brown, E. J. Khantzian, & J. E. Mack (Eds.), *Human feelings: Explorations in affect development and meaning* (pp. 87–121). Analytic Press.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Studies in Interactional Sociolinguistics, Vol. 4). Cambridge University Press.

Chai, C., Wu, X., Shen, D., Li, D., & Zhang, K. (2016). Gender differences in the effect of communication on college students' online decisions. *Computers in Human Behavior*, *65*, 176–188.

Cheng, N., Chandramouli, R., & Subbalakshmi, K. P. (2011). Author gender identification from text. *Digital Investigation*, *8*(1), 78–88.

Devlin, J., Chang, M. W. Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). ParseBert: Transformer-based model for Persian language understanding. *Neural Processing Letters*, *53*(6), 3831–3847.

Ghayoomi, M., Momtazi, S., & Bijankhan, M. (2009). A study of corpus development for Persian. *International Journal on Asian Language Processing*, *20*(1), 17–33.

Hekmatian Zadehpour, S., & Jalali Bidgoli, A. (2019). *Automatic gender recognition of authors of comments written in Persian*. Third National Conference on Knowledge and Technology of Electrical, Computer and Mechanical Engineering of Iran. https://civilica.com/doc/925633 (in Persian)

Holmes, J. (2013). *Women, men and politeness*. Routledge.

Hunter, D., Gambell, T., & Randhawa, B. (2005). Gender gaps in group listening and speaking: Issues in social constructivist approaches to teaching and learning. *Educational Review*, *57*(3), 329–355. https://psycnet.apa.org/doi/10.1080/00131910500149416

Kendall, S., & Tannen, D. (1997). Gender and language in the workplace. In R. Wodak (Ed.), *Gender and discourse* (pp. 81–105). Sage. https://doi.org/10.4135/9781446250204.n5

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401–412.

Lakoff, R. (1975). Language and woman's place. *Language in Society*, *2*(1), 45–79.

https://doi.org/10.1017/S0047404500000051

Moradi, M., & Bahrani, M. (2015). Automatic Gender Identification in Persian Text. *Signal and Data Processing*, *12*(4), 83–94. SID. https://sid.ir/paper/160742/fa (in Persian)

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*(3), 211–236. https://psycnet.apa.org/doi/10.1080/01638530802073712

Rangel, F., Rosso, P., Charfi, A., Zaghouani, W., Ghanem, B., & Sánchez-Junquera, J. (2019). Overview of the track on author profiling and deception detection in Arabic. *CEUR Workshop Proceedings, 2517*, 70–83.

Sadeghi, S. S., Khotanlou, H., & Rasekh Mahand, M. (2021). Automatic Persian text emotion detection using cognitive linguistic and deep learning. *Journal of AI and Data Mining*, *9*(2), 169–179. https://doi.org/10.22044/jadm.2020.9992.2136

Safara, F., Mohammed, A. S., Potrus, M., Ali, S., Quan, T. T., Souri, A., Janenia, F., & Hosseinzadeh, M. (2020). An author gender detection method using whale optimization algorithm and artificial neural network. *IEEE Access*, *8*, 48428–48437. https://doi.org/10.1109/ACCESS.2020.2973509

Sajedi, H., & Taslimi, M. (2019). Author gender identification from text using bayesian random forest. *Signal and Data Processing*, *16*(1), 143–156. https://doi.org/10.29252/jsdp.16.1.143 (in Persian)

Sboev, A., Moloshnikov, l., Gudovskikh, D., Selivanov, A., Rybka, R., & Litvinova, T. (2018). Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception. *Procedia Computer Science*, *123*, 417–423. https://doi.org/10.1016/j.procs.2018.01.064

Simaki, V., Mporas, l., & Megalooikonomou, V. (2016). Evaluation and sociolinguistic analysis of text features for gender and age identification. *American Journal of Engineering and Applied Sciences*, *9*(4), 868-876. https://doi.org/10.3844/ajeassp.2016.868.876

Tannen, D. (1999). Women and men in conversation. *The workings of language: From prescriptions to perspectives* (pp. 211–216), Bloomsbury Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Zarifi, A., & Naghavi, M. (2022). Gender identification of short text author using conceptual vectorization. *Multimedia Tools & Applications*, *82*(11), 1–17.