Razi University

Research in Western Iranian Languages and Dialects

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

Linguistics Society of Iran

Automatic Recognition of Authors Identity in Persian based on Systemic Functional Grammar

Fatemeh Soltanzadeh[1✉], **Azadeh Mirzaei**[2], **Mohammad Bahrani**[3], and **Shahram Modarres Khiabani**[4]

1. Corresponding Author, Ph.D. in Linguistics, Department of Linguistics, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran. E-mail: fatemeh.slt@gmail.com
2. Associate Professor of Linguistics, Department of Linguistics, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran. E-mail: azadeh.mirzaei@atu.ac.ir
3. Assistant Professor, Department of Computer, Faculty of Statistics, Mathematics and Computer, Allameh Tabataba'i University, Tehran, Iran. E-mail: bahrani@atu.ac.ir
4. Assistant Professor, Department of English Language and Translation, Islamic Azad University, Karaj, Iran. E-mail: shmodarress@yahoo.com

Article Info

Article type:
Research Article

Article history:
Received: 19 Jul 2023
Received in revised form: 20 Sep 2023
Accepted: 23 Sep 2023
Published online: 22 Sep 2024

Keywords:
author identification,
forensic linguistics,
systemic functional grammar,function words,
conjunctive adjunct,
mood adjunct,
comment adjunct.

ABSTRACT

Automated author identification is one of the important fields in forensic linguistics. In this study, the effectiveness of systemic functional grammar (Halliday and Matthiessen, 2014) features in Persian authorship attribution was compared with that of function words. First, a corpus composed of documents written by seven contemporary Iranian authors was collected. Second, a list of function words was extracted from the corpus. Moreover, conjunction, modality and comment adjunct system networks were applied to form a lexicon using linguistics resources. Then, the relative frequency of function words in addition to systemic functional features were calculated in each document. Multilayer perceptron classifier, a type of neural network, was used for learning phase which resulted in a desirable accuracy in evaluation phase. The results of the study showed that using function words method is superior to systemic functional approach alone in Persian author identification, however, simultaneous use of the two methods increases the effectiveness in comparison to each alone.

Cite this article: Soltanzadeh, F., Mirzaei, A., Bahrani, M., & Modarres Khiabani, Sh. (2024). Automatic Recognition of Authors Identity in Persian based on Systemic Functional Grammar. *Research in Western Iranian Languages and Dialects*, 12 (3), 59-84. http//doi.org/10.22126/jlw.2023.9391.1716 (in Persian).

DOI: https://doi.org/10.22126/jlw.2023.9391.1716

Publisher: Razi University

Introduction

Recently, automated author identification has become a key focus for forensic linguistics. Author identification involves determining the writer of a text from a set of potential authors. The text in question could be a threatening letter, an email, a literary work, or a scientific article or book. The basis for author identification rests on the idea that different authors may write about the same topic using overlapping, yet distinct, lexico-grammatical units—an issue referred to as idiolect (Coulthard, 2004).

The first significant attempt to identify writing styles was Mendenhall's study of Shakespeare's plays (1887). The play *Henry VIII* is widely recognized as a collaborative work, not solely authored by William Shakespeare. Plechac (2020) investigated the use of accent or stress to identify the contributions of other authors to the play.

In Persian, several studies have been conducted to determine authorship (Farahamandpour et al., 2013; Arefi et al., 2021). These studies utilized repetitive features, such as lexical richness, frequency of syntactic groups, collocations, and the relative frequency of punctuation marks, to detect writing styles. Measuring the frequency of function words is one valid method for author identification. Function words, which have limited meanings, indicate the functional relationships between components of a sentence. Golshaie (2019) and Dabagh (2007) applied the frequency of Persian function words to identify authors. This study aims to compare the efficacy of function word frequency with systemic functional grammar methods in automatically identifying writing styles.

Theoretical Framework

Systemic Functional Grammar (SFG) is a component of the social semiotic approach to language known as systemic functional linguistics (Halliday & Matthiessen, 2014). SFG conceptualizes language as a network of systems, or interrelated sets of options for creating meaning. Since the 1960s, SFG has been applied in various contexts within computational linguistics (Matthiessen & Bateman, 1991; Teich, 1995). In SFG, the clause is considered the fundamental unit of language, and it is analyzed through three perspectives, defined as the ideational, interpersonal, and textual metafunctions. The ideational function is further divided into the experiential and logical aspects (Halliday & Matthiessen, 2014).

This study employs three system networks: conjunction, modality, and comment. These networks correspond to three types of adjuncts: conjunctive adjuncts, mood adjuncts, and comment adjuncts, respectively. In the systemic environment of conjunction, conjunctions function as conjunctive adjuncts within the clause structure. They establish relationships where one segment of text elaborates on, extends, or enhances another segment (Halliday & Matthiessen, 2014).

Modal adjuncts express the speaker's or writer's judgment or attitude toward the content of the message. There are two types of modal adjuncts: (i) mood adjuncts and (ii) comment adjuncts. Mood adjuncts and comment adjuncts are categorized within the modality and comment adjunct system networks, respectively. Modality encompasses intermediate degrees between positive and negative poles, defining the region of uncertainty between 'yes' and 'no.' The modality system allows writers to qualify events or entities in terms of their probability, typicality, obligation, or inclination (Halliday & Matthiessen, 2014). The comment adjunct system provides a means for the writer to comment on the status of a message concerning the textual and interactive context of the discourse (Argamon et al., 2007). Comments can target either the ideational content of the proposition or the interpersonal aspects of the speech function (Halliday & Matthiessen, 2014).

Method

A corpus was compiled from the works of seven contemporary Persian writers: Hoshang Golshiri, Bozorg Alavi, Ahmad Mahmoud, Mahmoud Dolatabadi, Nader Ebrahimi, Jalal Al-e Ahmad, and

Gholamhossein Saedi, totaling 2,069,243 words. From this corpus, a list of 197 function words was extracted using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. Conjunction, modality, and comment adjunct system networks were then used to create a lexicon.

An author identification system was designed using machine learning techniques. The system tokenized the texts, extracted instances of lexical units specified in the lexicon, and computed the relative frequencies of semantic attribute values for each text, resulting in an overall "feature vector" that described each text. This approach was inspired by the method introduced by Argamon et al. (2007). For the learning phase, a multilayer perceptron classifier, a type of neural network, was utilized.

Results

To evaluate the system, the collected corpus was divided into five segments, and a 5-fold cross-validation method was applied. The 5-fold cross-validation demonstrated a satisfactory accuracy when focusing exclusively on function words. The combined use of function words and SFG methods achieved an accuracy of 74.47% for Persian author identification. Subsequent feature selection identified the most effective features for the machine learning phase. The results indicated that the relative frequency of function words outperformed SFG-based attributes in terms of effectiveness.

Discussion and Conclusions

The evaluation phase revealed that the function words-based method outperformed the systemic functional grammar (SFG) approach in identifying authors. However, the simultaneous use of both methods improved effectiveness compared to using either method alone. The superior performance of the function words-based method may be attributed to the high frequency of function words and the author's unconscious control over their use.

Among the SFG-based features, the combination of top features—namely conjunctive, mood, and comment adjuncts—produced higher accuracy than any single system network alone. Additionally, the results from feature selection indicated that features derived from the modality system network were more effective than those from the conjunction and comment adjunct system networks for Persian author identification.

Overall, while the function words-based method proved to be highly effective on its own, integrating it with SFG-based methods provided a more comprehensive approach, enhancing the accuracy of author identification.

مطالعات زبان‌ها و گویش‌های غرب ایران

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

تشخیص خودکار هویت نویسندهٔ متن در زبان فارسی براساس دستور نقش‌گرای نظام‌مند

فاطمه سلطان‌زاده[۱]✉| آزاده میرزایی[۲]| محمد بحرانی[۳]| شهرام مدرس خیابانی[۴]

۱. نویسندهٔ مسئول، دکتری زبان‌شناسی، گروه زبان‌شناسی، دانشکدهٔ ادبیات فارسی و زبان‌های خارجه، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: fatemeh.slt@gmail.com

۲. دانشیار گروه زبان‌شناسی، دانشکدهٔ ادبیات فارسی و زبان‌های خارجه، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: azadeh.mirzaei@atu.ac.ir

۳. استادیار گروه رایانه، دانشکدهٔ آمار، ریاضی و رایانه، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: bahrani@atu.ac.ir

۴. استادیار گروه آموزش مترجمی زبان انگلیسی، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران. رایانامه: shmodarress@yahoo.com

اطلاعات مقاله

نوع مقاله: پژوهشی

تاریخ دریافت: ۱۴۰۲/۴/۲۸

تاریخ بازنگری: ۱۴۰۲/۶/۲۹

تاریخ پذیرش: ۱۴۰۲/۷/۱

تاریخ انتشار: ۱۴۰۳/۷/۱

کلیدواژه‌ها:
تشخیص هویت نویسنده،
زبان‌شناسی حقوقی،
دستور نقش‌گرای نظام‌مند، واژه‌های دستوری،
افزودهٔ ربطی،
افزودهٔ وجه،
افزودهٔ نگرشی.

چکیده

تشخیص خودکار هویت نویسندهٔ متن یکی از مسائل مهم زبان‌شناسی حقوقی تلقی می‌شود. در پژوهش حاضر تلاش می‌شود کارایی ویژگی‌های مبتنی‌بر مفاهیم دستور نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) با کارایی واژه‌های دستوری در تشخیص هویت نویسنده مقایسه شود. به این منظور، در ابتدا، پیکره‌ای از آثار هفت نویسندهٔ معاصر ایرانی گردآوری شد. در مرحلهٔ دوم، از واژه‌های دستوری استخراج‌شده از پیکره فهرستی تهیه شد؛ به‌علاوه، یک مجموعهٔ واژگان براساس شبکهٔ نظام حروف ربط، شبکهٔ نظام افزودهٔ وجه و شبکهٔ نظام افزودهٔ نگرشی با استفاده از منابع زبانی تهیه شد. سپس بسامد نسبی واژه‌های دستوری و ویژگی‌های مبتنی‌بر دستور نقش‌گرای نظام‌مند در هر متن محاسبه شد. طبقه‌بند پرسپترون چند لایه، نوعی شبکهٔ عصبی، برای مرحلهٔ آموزش سامانه به کار گرفته شد و به دقت مطلوبی در مرحلهٔ ارزیابی منجر شد. بررسی نتایج ارزیابی سامانه نشان داد که روش محاسبهٔ بسامد واژه‌های دستوری نسبت‌به روش مبتنی‌بر دستور نقش‌گرای نظام‌مند در تشخیص هویت نویسندهٔ متون فارسی برتری دارد؛ باوجوداین، هنگامی که ویژگی‌های دستور نقش‌گرای نظام‌مند هالیدی درکنار ویژگی بسامد واژه‌های دستوری به کار روند، کارایی سامانه نسبت‌به حالتی که تنها از ویژگی بسامد واژه‌های دستوری استفاده شود، ارتقا می‌یابد.

استناد: سلطان‌زاده، فاطمه؛ میرزایی، آزاده؛ بحرانی، محمد؛ مدرس خیابانی، شهرام (۱۴۰۳). تشخیص خودکار هویت نویسندهٔ متن در زبان فارسی براساس دستور نقش‌گرای نظام‌مند. *مطالعات زبان‌ها و گویش‌های غرب ایران*، ۱۲ (۳)، ۸۴-۵۹. https://doi.org/10.22126/jlw.2023.9391.1716

DOI: https://doi.org/10.22126/jlw.2023.9391.1716

ناشر: دانشگاه رازی

۱– مقدمه

در جهان امروز باتوجه‌به پیشرفت فنّاوری و استفادهٔ گسترده از شبکهٔ جهانی وب[1]، این امکان برای کاربران فراهم است که در فضای مجازی هویت واقعی خود را پنهان سازند، با هویت جعلی خود را معرفی کنند، رایانامه‌هایی بدون نام نویسندهٔ واقعی بفرستند و حتی به سرقت ادبی یا علمی دست بزنند. ازاین‌رو، لازم است سامانه‌هایی طراحی شوند که امنیت فضای مجازی را تأمین کنند، سرقت ادبی و علمی را کشف نمایند و به‌صورت خودکار هویت واقعی نویسندگان متون را در این فضا شناسایی کنند. به‌این‌علت، در دهه‌های اخیر تشخیص خودکار هویت نویسندهٔ متن[2] به یکی از مسائل مهم در زبان‌شناسی حقوقی[3] تبدیل شده است.

تشخیص خودکار هویت نویسندهٔ متن بر این فرض استوار است که هر فرد یک گویش فردی[4] یا سبک خاص و منحصربه‌فرد در گفتار یا نوشتار دارد. درواقع، هنگامی که مردم از زبان استفاده می‌کنند، ساخت‌های خاصی از واژگان–دستور[5] را برمی‌گزینند که دیگر افراد ممکن است از آن‌ها کمتر استفاده کنند. در مرحلهٔ بعد هر فرد این ساخت‌ها را به شکلی متفاوت از دیگر افراد ترکیب می‌کند تا پیام خود را منتقل کند و این به معنای منحصربه‌فردبودن گویش فردی و سبک گفتار یا نوشتار افراد است (کلتارد[6]، ۲۰۰۴).

در زبان فارسی دربارهٔ تشخیص هویت نویسندهٔ متن، پژوهش‌هایی مانند فرهمندپور و دیگران (۱۳۹۱) و عارفی و دیگران (۱۴۰۰) انجام شده است. در این پژوهش‌ها بیشتر جنبه‌های فنی و مهندسی کار مدنظر بوده و به جنبه‌های زبانی و نظری کمتر توجه شده است. در پژوهش‌های یادشده عمدتاً از ویژگی‌های تکراری (غنای واژگانی[7]، بسامد گروه‌های نحوی[8]، باهم‌آیی کلمات[9]، بسامد نسبی علائم نگارشی[10]) برای طراحی سامانه‌های تشخیص خودکار هویت نویسندهٔ متن استفاده شده است و به طراحی ویژگی‌های زبانی برمبنای یک نظریهٔ زبانی خاص پرداخته نشده است. در زبان فارسی همچنین دو پژوهشِ دباغ (۲۰۰۷) و گلشائی (۱۳۹۸) درخصوص تشخیص خودکار هویت نویسندهٔ متن انجام شده است که ناظر بر جنبه‌های زبانی هستند. در این پژوهش‌ها از روش محاسبهٔ بسامد واژه‌های دستوری[11] در تشخیص هویت نویسندهٔ متن استفاده شده است. محاسبهٔ بسامد واژه‌های دستوری یک روش مهم و معتبر در تشخیص هویت نویسنده است. ازآنجاکه نویسنده در به‌کارگیری واژه‌های دستوری، کنترل خودآگاه ندارد و این واژه‌ها در متن به تعداد زیاد به کار می‌روند، آن‌ها می‌توانند ویژگی‌های منحصربه‌فردی در تشخیص خودکار نویسنده تلقی شوند (سگارا[12] و دیگران، ۲۰۱۵؛ گلشائی، ۱۳۹۸).

دباغ (۲۰۰۷) براساس واژه‌های دستوری پربسامد و با روش‌های آماری به جداسازی سبک نگارش نظامی گنجوی/شهریار و عبدالحسین زرین‌کوب/سیمین دانشور پرداخته است. نتایج این پژوهش نشان می‌دهد که به‌کمک واژه‌های دستوری می‌توان سبک نویسنده‌ها در نظم و نثر فارسی را مشخص کرد. گلشائی (۱۳۹۸) باتکیه‌بر گویش فردی و با استفاده از واژه‌های دستوری زبان فارسی، تشخیص هویت نویسنده متن را برای پنج نویسنده معاصر بررسی کرده است. نتایج این پژوهش نشان می‌دهد که نویسنده‌های مختلف واژه‌های دستوری را به‌طور مشابه به کار نمی‌گیرند. درواقع، اگرچه همهٔ نویسنده‌ها، برخی واژه‌های دستوری پربسامد را به کار می‌برند، اولویت نویسندگان در به‌کارگیری آن‌ها متفاوت است.

پژوهشگران دو پژوهش نام‌برده (دباغ، ۲۰۰۷؛ گلشائی، ۱۳۹۸) اثبات می‌کنند که واژه‌های دستوری در نثر فارسی توانایی جداسازی سبک نویسندگان را دارند؛ اما آن‌ها برای واژه‌های دستوری، دسته‌بندی معنایی[13] ارائه نمی‌کنند. دباغ (۲۰۰۷) در انتخاب واژه‌های دستوری صرفاً به معیار پربسامدبودن واژه‌های دستوری بسنده کرده است. گلشائی (۱۳۹۸) نیز حداقلِ کلمات موردنیاز برای جداسازی

1. World Wide Web
2. author identification
3. forensic linguistics
4. ideolect
5. lexicogrammar
6. M. Coulthard
7. lexical richness
8. frequency of syntactic groups
9. collocation
10. relative frequency of punctuation marks
11. function words
12. S. Segarra
13. semantic classification

سبک نویسندگان براساس واژه‌های دستوری را بررسی کرده است. وی توالی‌های یک تا سه‌واژه‌ایِ (تک‌نگاشتی[1]، دونگاشتی[2]، سه نگاشتی[3]) واژه‌های دستوری را در تشخیص هویت نویسندهٔ متن ارزیابی کرده و درنهایت، واژه‌های تک‌نگاشتی یا همان واژه‌های دستوری تک‌واژه‌ای را کاراترین نوع واژه‌های دستوری در نظر گرفته است. علاوه‌براین، در پژوهش وی، گسترهٔ مفهوم واژه‌های دستوری، اندکی محدود شده و ضمایر و تمامی افعال (ربطی و غیرربطی) حذف شده است.

باتوجه‌به نکته‌های یادشده در بررسی نقش واژه‌های دستوری زبان فارسی در جداسازی سبک نویسندگان، نیاز است در انتخاب واژه‌های دستوری به مؤلفه‌های معنایی بیشتر توجه شود و برای آن‌ها دسته‌های معنایی ارائه شود. درنهایت، می‌توان انواع مختلف واژه‌های دستوری را آزمود و بهترین ترکیب واژه‌های دستوری را ازاین‌بین انتخاب کرد. در پژوهش حاضر تلاش می‌شود که این خلأ در حوزهٔ تشخیص هویت نویسندهٔ متون فارسی پُر شود و برای واژه‌های دستوری، دسته‌بندی معنایی ارائه شود. علاوه‌براین، کارایی سامانهٔ تشخیص هویت نویسنده با واژه‌های دستوری، مبنای مقایسه قرار گیرد و یک نظریهٔ معتبر معنایی در تشخیص سبک نویسندهٔ زبان فارسی آزموده شود. همچنین به نظر می‌رسد طراحی ویژگی‌های زبانی جدید که مبتنی‌بر یک نظریهٔ زبانی معتبر باشد، درکنار واژه‌های دستوری، در ارتقای کارایی سامانه‌های تشخیص هویت نویسنده مؤثر باشد. بنابراین، تشخیص خودکار نویسنده متن برای زبان فارسی، مبتنی‌بر یک نظریهٔ زبانی معتبر با استفاده از ویژگی‌های زبانی جدید، یک ضرورت برای زبان فارسی در جهان پیچیدهٔ امروز محسوب می‌شود. دراین‌زمینه، در پژوهش حاضر تلاش بر آن است تا به‌کمک ویژگی‌های زبانی مبتنی‌بر دستور نقش‌گرای نظام‌مند[4] هالیدی[5] (هالیدی و متیسن[6]، ۲۰۱۴)، هویت نویسندهٔ متن به‌صورت خودکار تشخیص داده شود و تأثیر ویژگی‌های دستور نقش‌گرا در مقایسه با ویژگی بسامد واژه‌های دستوری سنجیده شود.

پرسش‌های پژوهش حاضر عبارت‌اند از: ۱. کدام‌یک از ویژگی‌های زبانی مبتنی‌بر دستور نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) بر تشخیص سبک نویسندگان فارسی زبان تأثیرگذار است؟ ۲. اگر ویژگی‌های زبانی مبتنی‌بر دستور نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) درکنار ویژگی بسامد واژه‌های دستوری به کار روند، بر کارایی سامانه نسبت‌به حالتی که تنها از ویژگی بسامد واژه‌های دستوری استفاده شود، چه تأثیری دارد؟

برای تشخیص خودکار نویسندهٔ متن از ویژگی‌های زبانی در سطوح مختلف همچون ویژگی‌های آوایی[7]، واژگانی، نحوی، معنایی و ترکیب آن‌ها استفاده می‌شود. پس از تعریف ویژگی‌ها، اسناد موجود در پیکره که هرکدام به یک نویسندهٔ خاص در یک مجموعهٔ بسته تعلق دارند، با بهره‌گیری از ویژگی‌های زبانی تعریف‌شده به روش یادگیری ماشین دسته‌بندی و اسناد مشابه به یک نویسندهٔ واحد منسوب می‌گردد.

نخستین تلاش برای سنجیدن سبک نگارش به سدهٔ نوزدهم میلادی و بررسی نمایشنامه‌های شکسپیر[8] باز می‌گردد (مندنهال[9]، ۱۸۸۷). این پرسش که آیا شکسپیر تمام آثارش را خود به‌تنهایی نوشته است، سال‌ها برای پژوهشگران بحث‌برانگیز بوده است. برای مثال، برخی منتقدان ادبی معتقد بودند که شکسپیر نمایشنامهٔ *هنری هشتم*[10] را به‌کمک فرد دیگری نوشته است. پژوهشگران رایانه و هوش مصنوعی این ادعا را سال‌ها بررسی کرده و درنهایت، به این نتیجه رسیده‌اند که وی با همکاری فردی به نام فلچر[11] نمایشنامهٔ *هنری هشتم* را نوشته است (پلچاک[12]، ۲۰۲۱). این کار باتکیه‌بر ویژگی‌های آوایی و الگوهای غالب آهنگین[13] (توزیع هجاهای تکیه‌دار و بدون تکیه در یک سطر) انجام شده است.

از شمارش کلمات پرتکرار، میانگین طول واژه‌ها در متن و تعداد تکرار کلمات منحصربه‌فرد و غنای واژگانی نیز به‌منزلهٔ ویژگی‌های

1. unigram
2. bigram
3. trigram
4. systemic functional grammar (SFG)
5. M. A. K. Halliday
6. C. M. M. Matthiessen
7. phonetic
8. W. Shakespeare
9. T. C. Mandenhal
10. Henry VIII
11. J. Fletcher
12. P. Plecháč
13. rhythmic

واژگانی برای تشخیص خودکار نویسنده استفاده می‌شود (استاماتاتوس[1]، ۲۰۰۹). در برخی پژوهش‌ها از ترکیب چند نوع ویژگی بهره گرفته شده است. برای مثال، در پژوهش ویراسینگه[2] و دیگران (۲۰۲۱)، ویژگی بسامد واژه‌های دستوری، بسامد گروه‌های نحوی و غنای واژگانی به کار گرفته شده است. نجفی و تاوان (۲۰۲۲) که از بسامد گروه‌های نحوی، موجودیت‌های نامدار[3] و باهم‌آیی کلمات درکنار یادگیری عمیق[4] بهره می‌گیرند به کارایی مطلوبی در تعیین هویت نویسنده در زبان انگلیسی دست می‌یابند. مارتینز[5] و دیگران (۲۰۲۲) نیز ویژگی بسامد گروه‌های نحوی را با روش شبکهٔ عصبی گرافی[6] می‌آزمایند و به دقت مناسبی دست می‌یابند.

دو پژوهشِ آرگامون[7] و دیگران (۲۰۰۷) و آرگامون و کوپل[8] (۲۰۱۳) نیز برای تشخیص خودکار نویسندهٔ متن در زبان انگلیسی انجام شده است؛ در این پژوهش‌ها از مجموعه‌ای از ویژگی‌ها که براساس دستور نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) نظام یافته‌اند، بهره گرفته شده است. در پژوهش آرگامون و دیگران (۲۰۰۷)، از دستور نقش‌گرای نظام‌مند در تشخیص هویت، جنسیت[9]، ردهٔ سنی و ملیت نویسندهٔ متن استفاده شده و کارایی این دستور در مقایسه با بسامد واژه‌های دستوری سنجیده شده است. در پژوهش آرگامون و کوپل (۲۰۱۳)، از این دستور برای تشخیص جنسیت، ردهٔ سنی، ملیت و ویژگی‌های شخصیتی[10] نویسنده استفاده شده و کارایی این دستور در مقایسه با واژه‌های محتوایی[11] مقایسه شده است. به این منظور، یک واژگان خاص طراحی شده است که کلمات و عباراتی را دربرمی‌گیرد که از شبکهٔ واژگان و اصطلاح‌نامه‌های[12] برخط گردآوری شده است. هر مدخل واژگان ویژگی‌هایی دارد که برگرفته از دستور نقش‌گرای نظام‌مند است. نتایج پژوهش آرگامون و دیگران (۲۰۰۷) نشان می‌دهد که این ویژگی‌ها در ترکیب با ویژگی‌های مبتنی‌بر واژه‌های دستوری بیشترین کارایی را دارند. پژوهش آرگامون و کوپل (۲۰۱۳) نیز بیانگر آن است که در تشخیص جنسیت، ردهٔ سنی و ملیت نویسنده، ترکیب واژه‌های محتوایی و ویژگی‌های مبتنی‌بر دستور نقش‌گرای نظام‌مند بالاترین کارایی را داشته است. همچنین روش دستور نقش‌گرای نظام‌مند در قیاس با مبنای مقایسهٔ خود (واژه‌های محتوایی) کارایی پایین‌تری در تشخیص جنسیت، ردهٔ سنی و ملیت نویسندهٔ متن در زبان انگلیسی داشته است.

۲- مبانی نظری

۲-۱ دستور نقش‌گرای نظام‌مند

در پژوهش حاضر، از مفاهیم دستور نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) برای تشخیص خودکار هویت نویسندهٔ متن استفاده می‌شود. این دستور، زبان‌ها را به‌منزلهٔ نظامی از انتخاب‌ها برای بیان معنا الگوسازی و همهٔ انتخاب‌های ساختاری و واژگانی را براساس نقش‌های معنایی آن‌ها بازنمایی می‌کند. این نظریه در پردازش زبان طبیعی[13] و زبان‌شناسی رایانشی از دههٔ ۱۹۶۰ میلادی به کار گرفته شده است؛ به‌ویژه اینکه در تولید متن[14] بسیار کاربرد دارد (تیچ[15]، ۱۹۹۵؛ متیسن و بتمن[16]، ۱۹۹۱). دستور نقش‌گرای نظام‌مند هالیدی مجموعه‌ای از محدودیت‌ها را در بیان معنا تعریف می‌کند. این دستور شبکه‌ای از گزینه‌های محتمل است. این شبکه از انتخاب‌ها، شبکهٔ نظام[17] نامیده می‌شود. با داشتن این شبکه می‌توان برای توصیف سبک یک متن از ویژگی‌های آماری بهره برد؛ به این شکل که بسامد شرطی هر گره[18] را در این شبکه با داشتن گره والد[19] آن محاسبه کرد و آن را معیاری برای توصیف سبک متن به کار

1. E. Stamatatos
2. J.Weerasinghe
3. named entities
4. deep learning
5. J. A. Martinez-Galicia
6. graph neural networks
7. S. Argamon
8. M. Koppel
9. gender
10. personality
11. content words
12. thesaurus
13. natural language processing
14. text generation
15. E. Teich
16. J. A. Bateman
17. system network
18. node
19. parent node

برد. این تفسیر بر این اصل استوار است که نویسندهٔ دارای سبک خاص و منحصربه‌فرد، برای بیان معنا گزینه‌های خاصی را بر دیگر گزینه‌ها ترجیح می‌دهد (آرگامون و دیگران، ۲۰۰۷). بنابراین، در پژوهش حاضر، از دستور نقش‌گرای نظام‌مند هالیدی برای چارچوب نظری پژوهش و از روش پیشنهادی در پژوهش آرگامون و دیگران (۲۰۰۷) برای تعریف ویژگی‌ها در تشخیص هویت نویسنده استفاده می‌شود.

زبان در دستور نقش‌گرای نظام‌مند به‌صورت نظامی از معناها در نظر گرفته می‌شود که با صورت[1] همراه می‌شود. واحد مطالعه در دستور نقش‌گرای نظام‌مند بند[2] است. ازلحاظ معنایی، بند سه لایهٔ مستقل دارد که باهم در تعامل هستند. این لایه‌ها فرانقش[3] نامیده می‌شوند و عبارت‌اند از: ۱. فرانقش تجربی[4] که تجربهٔ نویسنده از هستی را بازنمایی می‌کند؛ ۲. فرانقش بینافردی[5] که به نحوهٔ تعامل میان شرکت‌کنندگان در یک رویداد گفتمانی می‌پردازد؛ ۳. فرانقش متنی[6] که انسجام و پیوستگی میان مطالب را در یک متن سامان می‌دهد (هالیدی و متیسن، ۲۰۱۴).

در پژوهش پیش‌رو، افزوده‌ها[7] در دو فرانقش بینافردی و متنی مدنظر هستند. در این پژوهش، افزودهٔ ربطی[8] که در فرانقش متنی مطرح است و افزودهٔ وجهی[9] (افزودهٔ وجه[10] و افزودهٔ نگرشی[11]) که در فرانقش بینافردی به کار می‌رود، برای تشخیص هویت نویسنده استفاده می‌شود. دراین‌زمینه، سه نوع شبکهٔ نظام معرفی می‌شود: شبکهٔ نظام حروف ربط[12]، شبکهٔ نظام وجه‌نمایی[13] و شبکهٔ نظام افزودهٔ نگرشی. در ابتدا، دربارهٔ فرانقش متنی و شبکهٔ نظام حروف ربط بحث می‌شود. سپس به فرانقش بینافردی و دو نوع شبکهٔ نظام وجه‌نمایی و شبکهٔ نظام افزودهٔ نگرشی پرداخته می‌شود.

۲-۱-۱ فرانقش متنی

در فرانقش متنی نحوهٔ چینش سازه‌ها با عنوان ساخت آغازگری-پایان‌بخشی[14] بررسی می‌شود. آغازگر به‌منزلهٔ نقطهٔ عزیمت پیام، نخستین سازهٔ بند است به‌شرط آنکه مشارک[15]، فرایند[16] یا افزوده حاشیه‌ای باشد (هالیدی و متیسن، ۲۰۱۴). بقیهٔ جمله منهای آغازگر، پایان‌بخش است.

افزوده‌های ربطی (حروف ربط) در فرانقش متنی مطرح هستند و بین دو بند یا دو جمله پیوند منطقی برقرار می‌کنند. جایگاه بی‌نشان آن‌ها در ابتدای بند یا پس از آغازگر است. این افزوده‌ها که با انسجام متن مرتبط هستند، در دستور نقش‌گرای نظام‌مند در شبکهٔ نظام حروف ربط سازمان‌بندی می‌شوند. انواع مختلف حروف ربط برای پیونددادن بندها به کار گرفته می‌شوند. حروف ربط نشان می‌دهند که چگونه یک بند براساس بافت پیشین بسط و توسعه می‌یابد. در این شبکهٔ نظام، افزودهٔ ربطی (حرف ربط) به سه دستهٔ کلی افزودهٔ تشریحی[17]، گسترشی[18] و تفصیلی[19] تقسیم می‌شود (هالیدی و متیسن، ۲۰۱۴) که در شکل (۱) مشاهده می‌شود.

در افزودهٔ تشریحی، افزوده‌ها خود به دو دستهٔ کلی تقسیم می‌شوند: بدلی[20] و واضح‌سازی[21]. در افزودهٔ بدلی، یک عنصر دوباره معرفی یا

1. form
2. clause
3. metafunction
4. experiential metafunction
5. interpersonal metafunction
6. textual metafunction
7. adjunct
8. conjunctive adjunct
9. modal adjunct
10. mood adjunct
11. comment adjunct
12. conjunction
13. modality
14. theme-rheme
15. participant
16. process
17. elaborating
18. extending
19. enhancing
20. appositive
21. clarification

بیان می‌شود که به‌صورت توضیحی[1] یا به‌صورت نمونه‌سازی[2] است. در افزودهٔ واضح‌سازی نیز آن عنصر به شکل واضح‌تری بیان می‌شود[3]. مثال‌های (۱) و (۲) به‌ترتیب نمونه‌ای از «افزودهٔ تشریحی، بدلی، توضیحی» و «افزودهٔ تشریحی، بدلی، نمونه‌سازی» هستند.

۱. آنچه به تو می‌گویم یقین و حتمی است مثل این است که خودم از دهن صاحب‌کار شنیده باشم (علوی، ۱۳۸۶).

۲. ما مسلمانیم ارباب؛ مثلاً خود من تا حالا حتی یک بار هم نشده نمازم قضا بشود (گلشیری، ۱۳۷۰).

افزودهٔ گسترشی خود به سه دستهٔ کلی تقسیم می‌شود: ۱. افزودهٔ افزایشی[4]: چیزی را به معنا می‌افزاید؛ ۲. افزودهٔ تغییری[5]: به‌نوعی معنا را تغییر می‌دهد؛ ۳. افزودهٔ تباینی[6]: برخلاف آن چیزی را بیان می‌کند. در مثال (۳) و (۴) می‌توانید به‌ترتیب نمونه‌ای از «افزودهٔ گسترشی، افزایشی، مثبت[7]» و «افزودهٔ گسترشی، تباینی» را مشاهده کنید.

۳. بچه‌ها هورا کشیدند و کف زدند (آل‌احمد، ۱۳۴۶).

۴. شب‌ها اگر از فرسنگ‌ها راه به جهت تپه‌ماهورهای شبانکاره نگاه می‌کردی، نور نارنجی پایداری می‌دیدی که بیش‌وکم می‌شد؛ اما مرگ نداشت (ابراهیمی، ۱۳۹۹).

افزودهٔ تفصیلی که پیوستگی[8] را ازطریق مقایسهٔ بین عناصر یا بندها به وجود می‌آورد خود به چهار دستهٔ کلی تقسیم می‌شود: مضمون[9]، شیوه[10]، فضایی–زمانی[11]، سببی–شرطی[12]. مثال‌های (۵) و (۶) نمونه‌هایی از «افزودهٔ تفصیلی، شیوه[13]، طریق» هستند.

۵. یکی‌دو مرتبه که مردم دِه بیچاره می‌شدند، کدخدا را پیشِ خان همسایه می‌فرستادند و از او کمک می‌گرفتند و بدین‌طریق دهکده‌ای به تصرف خانی درمی‌آمد (علوی، ۱۳۹۹).

۶. راه افتاد طرف در و پیش از اینکه خارج شود ایستاد و بعد با صدای محکمی پرسید: «برویانف را چطور کشتی؟» (ساعدی، ۱۳۹۷).

مثال (۷) نمونه‌ای از «افزودهٔ تفصیلی، فضایی–زمانی، ساده، هم‌زمانی» است که به‌لحاظ توالی زمانی، به دو رویداد هم‌زمان اشاره دارد.

۷. خودتُ عاجز نکن ننه؛ حالا کی می‌خواد بفروشه؟ (محمود، ۱۳۵۳).

افزودهٔ سببی–شرطی نیز از دو نوعِ کلی سببی و شرطی تشکیل شده است. افزودهٔ سببی خود به دو نوع عام[14] و مشخص[15] تقسیم می‌شود. در نوع «مشخص» افزوده به نتیجه[16]، علت[17] و قصد[18] اختصاص می‌یابد. در مثال (۸) و (۹) می‌توانید به‌ترتیب نمونه‌ای از «افزودهٔ سببی–شرطی، سببی، مشخص، علت» و «افزودهٔ سببی–شرطی، شرطی، منفی[19]» را مشاهده کنید.

۸. وکیل رعیت وفادارترین رعیت خویش را به هوسی به زنجیر کشید و کشان به بازار آورد و به دار مجازاتی که هیچ حقش نبود آویخت، فقط به‌خاطر آنکه خود را دربرابر اجانب کوچک‌تر از او می‌دید (ابراهیمی، ۱۳۹۹).

۹. با این حرف‌ها نمی‌خواهم ناراحتت کنم؛ اما تو باید بدانی، باید بفهمی، حتماً کریستین برایت نگفته اگرنه کار به اینجاها نمی‌کشید (گلشیری، ۱۳۵۰).

همان‌گونه که پیش‌تر اشاره شد در شکل (۱) شبکهٔ نظام حروف ربط هالیدی و متیسن (۲۰۱۴) مشاهده می‌شود که سه دستهٔ کلی

---

1. expository
2. exemplifying

۳. شایان ذکر است که افزودهٔ «واضح‌سازی» نیز خود به هفت زیربخش تقسیم می‌شود؛ به‌دلیل جزئیات زیاد، در این پژوهش این تقسیم‌بندی موردتوجه نبوده است و در شکل (۱) نمایش داده نشده است. همین مسئله دربارهٔ افزودهٔ «مرکب، فضایی–زمانی، تفصیلی» نیز صادق است.

4. additive
5. varying
6. adversative

۷. افزودهٔ افزایشی به دو صورت مثبت و منفی است.

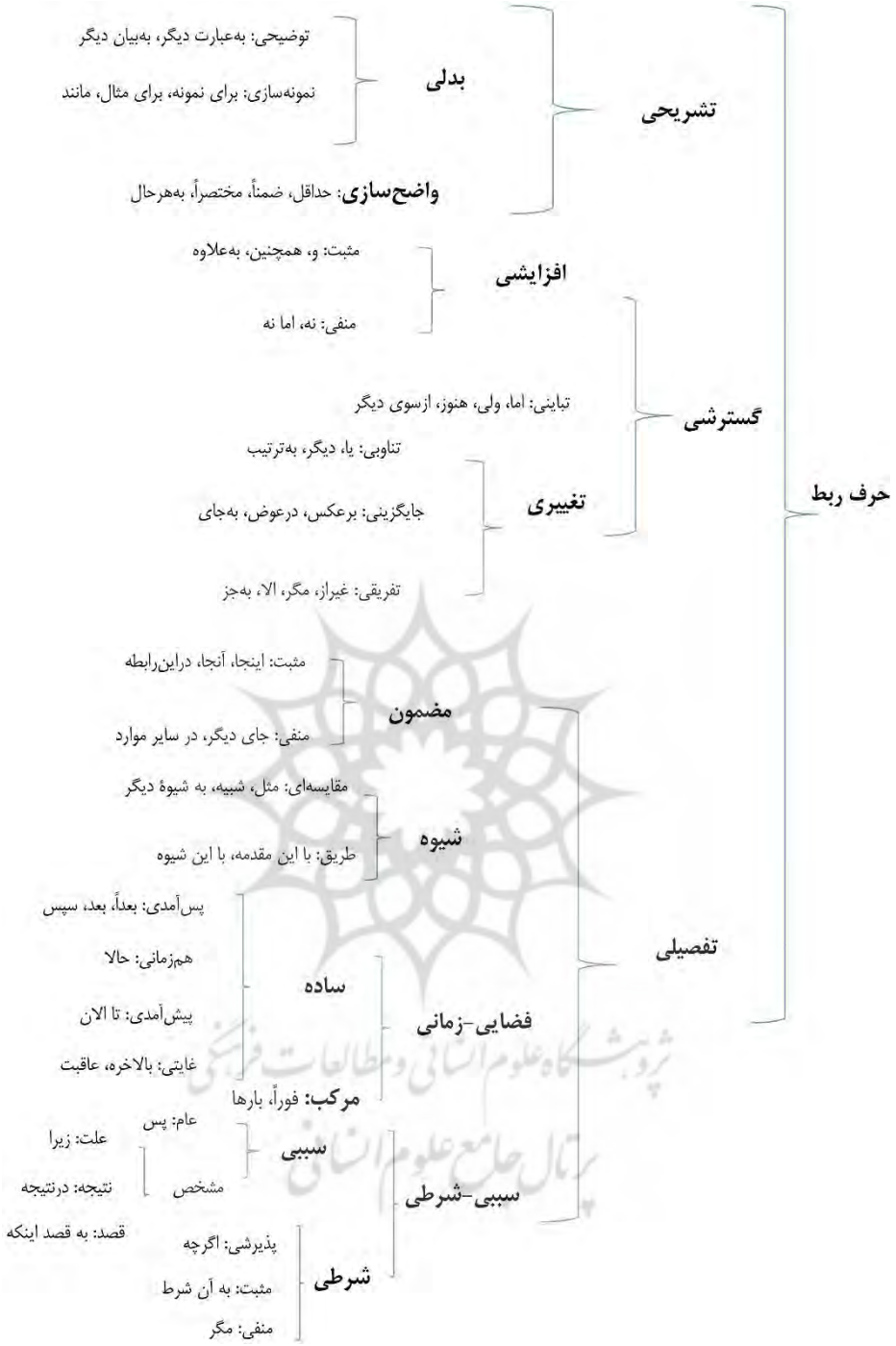
8. cohesion
9. matter
10. manner
11. spatio-temporal
12. causal-conditional

۱۳. افزودهٔ شیوه، پیوستگی را به‌وسیلهٔ مقایسه یا ارجاع بین عناصر ایجاد می‌کند و به دو گروه افزودهٔ مقایسه‌ای و طریق تقسیم می‌شوند.

14. general
15. specific
16. result
17. reason
18. purpose

۱۹. افزوده‌هایی همچون «مگر»، «اگرنه» و «در غیر این صورت» از نوع «افزودهٔ سببی–شرطی، شرطی، منفی» هستند.

افزودهٔ تشریحی، گسترشی و تفصیلی را شامل می‌شود.



شکل (۱). شبکهٔ نظام حروف ربط (هالیدی و متیسن، ۲۰۱٤: ٦١٢؛ جعفری، ۱۳۸۸: ۱۳۷-۱۳۹)

۲-۱-۲ فرانقش بینافردی

در فرانقش بینافردی هر بند نوعی تبادل است. تعامل زبانی در این فرانقش یک رابطهٔ دادوستد کالا یا خدمات یا اطلاعات است. درواقع، در این فرانقش پیام از دو نوع گزاره[1] یا پیشنهاد[2] است. در صورتِ دادن یا خواستن اطلاعات، پیام از نوع گزاره و در صورتِ عرضه یا درخواست کالا یا خدمات[3] پیام از نوع پیشنهاد است (هالیدی و متیسن، ۲۰۱۴). گوینده در فرانقش بینافردی ابزارهایی در دست دارد که

1. proposition
2. proposal
3. goods and services

با آن‌ها می‌تواند نظر خود را در پیام منعکس کند، میزان قطعیت[1] گزاره را تخمین بزند یا سبب شود که احتمال اجرایی‌شدن پیشنهاد افزایش یابد. میزان قطعیت هر گزاره با صورت‌بندی‌هایی که نمایندهٔ مفهوم احتمال[2] هستند (حتماً، احتمالاً، شاید، به‌احتمال زیاد، احتمال می‌رود و...)، همچنین با قیود تکرار (همیشه، اغلب، گاهی و...) مشخص می‌شود. میزان قطعیت یا به‌بیان بهتر، ضمانت اجرایی پیشنهاد با صورت‌بندی‌هایی منتقل می‌شود که نمایندهٔ مفهوم اجبار یا التزام[3] (اجبار، پیشنهاد و اجازه) و تمایل[4] (تصمیم، توانایی و تمایل) هستند (میرزایی، ۱۳۹۷).

در این فرانقش محتوای یک بند به دو عنصر وجه[5] و مانده[6] تقسیم می‌شود. وجه آن بخش از بند است که از دو قسمت اصلی تشکیل شده است: ۱. فاعل[7] که یک گروه اسمی است؛ ۲. عنصر زمان‌دار[8] که بخشی از گروه فعلی است و بیانگر زمان، نمود[9]، وجه، جهت[10] و قطبیت[11] است. عنصر بعدی در سازمان‌بندی نقش‌ها در معنای بینافردی، مانده است که خود از سه عنصر نقشی مختلف تشکیل شده است: الف) محمول[12]؛ ب) متمم[13]؛ ج) افزوده. محمول در تمام بندها با گروه فعلی تحقق می‌یابد. متمم آن قسمت از مانده است که فاعل نیست؛ ولی توانایی فاعل‌شدن را دارد و معمولاً به‌صورت گروه اسمی ظاهر می‌شود. افزوده، عنصری از مانده است که توانایی فاعل‌شدن را ندارد و به‌طور مشخص به‌صورت یک گروه قیدی یا حرف اضافه‌ای ظاهر می‌شود (جعفری، ۱۳۸۸: ۱۳۳).

افزودهٔ وجهی در فرانقش بینافردی مطرح است و خود به دو دسته تقسیم می‌شود: الف) افزوده‌های وجه که مفاهیمی چون احتمال، اجبار[14] (التزام)، تمایل[15]، عادت[16]، شدت[17] و زمان را بیان می‌کنند؛ ب) افزوده‌های نگرشی که تفسیر، قضاوت و اظهارنظر درمورد پیام و گزاره‌ای را که بیان می‌شود، ارائه می‌دهد. افزوده‌های وجه در شبکهٔ نظام وجه‌نمایی جای می‌گیرند. آن‌ها نویسنده را قادر می‌سازند که حوادث و موجودیت‌ها را در متن براساس احتمال، عادت، تمایل و اجبار ارزش‌گذاری کند. این شبکهٔ نظام از چهار زیرنظام نوع[18]، جهت‌گیری[19]، مقدار[20] و قطبیت تشکیل شده است که در شکل (۲) سلسله‌مراتب آن نشان داده شده است.

زیرنظام نخست، «نوع» است که از دو نوع کلی وجه‌سازی[21] (وجه احتمال–عادت) و تعدیل‌سازی[22] (وجه تمایل–التزام) تشکیل شده است. وجه‌سازی در ارتباط با گزاره موضوعیت می‌یابد. گزاره‌ها یا تأیید می‌شوند یا انکار. نقش نظام وجهیت تحلیل قلمرو عدم قطعیت مابین آری یا خیر است (حسین حمه و دیگران، ۱۴۰۰؛ هالیدی و متیسن، ۲۰۱۴). در وجه احتمال–عادت، ازیک‌سو، احتمال وقوع گزاره مطرح است (قطعاً، احتمالاً، شاید، بعیده، غیرممکنه) و ازسوی دیگر، میزان تکرار وقوع گزاره (همیشه، معمولاً، گاهی، هرگز). در موردِ اول، وجه احتمال–عادت از نوع احتمالی و دومی از نوع عادتی است. تعدیل‌سازی در ارتباط با پیشنهاد مطرح می‌شود. همان‌طور که پیش‌تر اشاره شد، میزان قطعیت پیشنهاد با صورت‌بندی‌هایی که نمایندهٔ مفهوم اجبار یا التزام (اجبار، پیشنهاد و اجازه) و تمایل (تصمیم، توانایی و تمایل) هستند، منتقل می‌شود. پس در تعدیل‌سازی نیز دو نوع وجهیت مطرح است: وجه التزامی (مانند باید، ضرورتاً، قهراً و...) و وجه تمایلی (مانند می‌خواهم، دوست داریم، تصمیم داریم و...).

همان‌طور که پیش‌تر گفته شد شبکهٔ افزودهٔ وجه از چهار زیرنظام نوع، جهت‌گیری، مقدار و قطبیت تشکیل شده است که در شکل

1. certainty
2. probability
3. obligation
4. inclination or readiness
5. mood element
6. residue
7. subject
8. finite
9. aspect
10. voice
11 . polarity
12. predicator
13. complement
14. obligation
15. desire
16. usuality
17. intensity
18. type
19. orientation
20. value
21. modalization
22. modulation

(۲) می‌توانید سلسله‌مراتب آن را مشاهده کنید.

شکل (۲). شبکهٔ نظام افزودهٔ وجه (هالیدی و متیسن، ۲۰۱٤: ۱۸۲)

دومین زیرنظام، «جهت‌گیری» است. این زیرنظام، دو نوع تقابل را نشان می‌دهد: تقابل شخصی[1] و بی‌طرفانه[2] و تقابل صریح[3] و ضمنی[4]. برای درک بهتر، این دو نوع تقابل در جدول (۱) با مثال نشان داده شده است.

جدول (۱). تحلیل نمونه‌ها براساس زیرنظام جهت‌گیری (هالیدی و متیسن، ۲۰۱٤: ۱۸۱)

جهت‌گیری	شخصی	بی‌طرفانه
ضمنی	باید	قطعاً
صریح	مطمئنم	مسلم است

در مثال‌های «باید» و «مطمئنم» که از نوع شخصی هستند، نویسنده منبع نظر را ذکر می‌کند؛ درحالی‌که در «قطعاً» و «مسلم است»، منبع آن مشخص نیست و حالت بی‌طرفانه دارند. دربارهٔ «باید»، ذکر منبع نظر به‌صورت ضمنی و در «مطمئنم» به‌صورت صریح بیان شده است (هالیدی و متیسن، ۲۰۱۴: ۱۸۱).

سومین زیرنظام، «مقدار» است که سه سطح «بالا»، «پایین» و «میانه» دارد.

چهارمین زیرنظام وجه‌نمایی، «قطبیت» است. قطبیت تقابل میان مثبت و منفی است. قطبیت را می‌توان به‌نوعی دو سر پیوستار وجهیت در نظر گرفت (میرزایی، ۱۴۰۰).

در ادامه، مثال‌هایی از افزوده‌های وجه در جملات پیکرهٔ پژوهش حاضر ذکر شده است. در جدول (۲)، افزوده‌های وجه در این مثال‌ها بنابر شبکهٔ نظام وجه‌نمایی تحلیل شده‌اند.

1. subjective
2. objective
3. explicit
4. implicit

۱۰. فولاد گفت: «یعنی ممکنه که تو تب و هذیان از خونه زده باشه بیرون و تو نخلستانا، گوشه‌ای افتاده باشه و سرما خشکش کرده باشه؟» (محمود، ۱۳۸۱).

۱۱. با این حرف‌ها نمی‌خواهم ناراحتت کنم؛ اما تو باید بدانی، باید بفهمی، حتماً کریستین برایت نگفته اگرنَه کار به اینجاها نمی‌کشید (گلشیری، ۱۳۵۰).

۱۲. عمه بزرگ گفت: «خسرو خان، از یک شازده بعید است که بادبادک پسر باغبان را بردارد (گلشیری، ۱۴۰۰).

۱۳. این بهت و درنگ گهگاهی از صبح با شیرو بود (دولت‌آبادی، ۱۴۰۱).

۱۴. من تصمیم دارم هر شنبه بیایم پیش تو، به‌شرط اینکه بازهم از این چیزها بنویسی (ساعدی، ۱۳۷۷).

جدول (۲). تحلیل نمونه‌ها براساس شبکهٔ نظام وجه‌نمایی

مثال	نوع (سطح اول)	نوع (سطح دوم)	جهت‌گیری		مقدار	قطبیت
ممکنه	وجه‌سازی	احتمالی	بی‌طرفانه	صریح	میانه	مثبت
حتماً	وجه‌سازی	احتمالی	بی‌طرفانه	ضمنی	بالا	مثبت
بعید است	وجه‌سازی	احتمالی	بی‌طرفانه	صریح	پایین	منفی
گهگاهی	وجه‌سازی	عادتی	بی‌طرفانه	ضمنی	پایین	مثبت
نمی‌خواهم	تعدیل‌سازی	تمایل	شخصی	صریح	بالا	منفی
باید	تعدیل‌سازی	التزام	شخصی	ضمنی	بالا	مثبت
تصمیم دارم	تعدیل‌سازی	تمایل	شخصی	صریح	بالا	مثبت

شبکهٔ نظام دیگر، شبکهٔ افزودهٔ نگرشی است. این افزوده‌ها نظر نویسنده یا گوینده را بیان می‌کنند و هدف آن‌ها محتوای گزاره یا بخشی از نقش گفتار است؛ ازاین‌رو، به دو گروه گزاره‌ای[1] و نقش‌گفتار[2] تقسیم می‌شوند. در افزودهٔ نگرشی از نوع گزاره‌ای، گوینده یا نویسنده نظر خود را دربارهٔ کل گزاره یا بخشی از آن (فاعل) بیان می‌کند. پس افزودهٔ نگرشی، گزاره‌ای به دو نوع بندی و فاعلی تقسیم می‌شود. افزودهٔ نگرشی از نوع نقش‌گفتار نیز به دو نوع فاقد شرایط[3] و واجد شرایط[4] تقسیم می‌شود (هالیدی و متیسن، ۲۰۱۴: ۱۹۰). در شکل (۳)، شبکهٔ نظام افزودهٔ نگرشی مشاهده می‌شود.

شکل (۳). شبکهٔ نظام افزودهٔ نگرشی (هالیدی و متیسن، ۲۰۱٤)

در شکل (۴) و (۵)، افزودهٔ بندی و فاعلی و در شکل (۶) افزودهٔ نقش‌گفتار نمایش داده شده است.

در مثال (۱۵) و (۱۶) می‌توانید به‌ترتیب نمونه‌ای از «افزودهٔ گزاره‌ای، بندی، صلاحیتی[5]، فرض، شایعه[6]» و «افزودهٔ گزاره‌ای، بندی،

1. propositional
2. speechfunctional
3. unqualified
4. qualified
5. qualificative
6. hearsay

مطلوبیت[1]، غیرمطلوب» را مشاهده کنید.

۱۵. بر فرض هم که خواهرم با این عنکبوت مشغولیتی پیدا کرده باشد، تازه به من چه؟ عنکبوت، عنکبوت است دیگر (آل‌احمد، ۱۳۵۰).

۱۶. اگر خودِ شما دختر داشتید، به همچو آدمی شوهرش می‌دادید؟ گفت: «متأسفانه من دختر ندارم» (آل‌احمد،۱۳۵۰).

شکل (۴) شبکهٔ نظام افزودهٔ گزاره‌ای بندیِ هالیدی و متیسن (۲۰۱۴) را نشان می‌دهد. در این شکل می‌توانید سلسله‌مراتب افزودهٔ گزاره‌ای بندیِ به‌کاررفته در مثال‌های (۱۵) و (۱۶) را مشاهده کنید.

شکل (٤). شبکهٔ نظام افزودهٔ گزاره‌ای بندی (هالیدی و متیسن، ۲۰۱٤: ۱۹۰)

در مثال (۱۷)، نمونه‌ای از «افزودهٔ گزاره‌ای، فاعلی، اخلاقی، منفی» ذکر شده است.

۱۷. یک کتابخانهٔ نفیس تخصصی در باب معماری داشت؛ اما حتی اشتباهاً، یک دیوان حافظ و مولوی درکنار آن مجموعهٔ نفیس، جای نگرفته بود (ابراهیمی، ۱۳۷۴).

شکل (۵) شبکهٔ نظام افزودهٔ گزاره‌ای فاعلیِ هالیدی و متیسن (۲۰۱۴) را نشان می‌دهد. در این شکل می‌توانید سلسله‌مراتب افزودهٔ گزاره‌ای فاعلیِ به‌کاررفته در مثال (۱۷) را مشاهده کنید.

شکل (۵). شبکهٔ نظام افزودهٔ گزاره‌ای فاعلی (هالیدی و متیسن، ۲۰۱٤: ۱۹۰)

1. desirability

مثال‌های (۱۸) و (۱۹) به‌ترتیب نمونه‌هایی از «افزودهٔ نقش‌گفتار، فاقد شرایط، حقیقی» و «افزودهٔ نقش‌گفتار، واجد شرایط، اعتبار[1]، عام» هستند.

۱۸. شاید حقیقتاً هم اسد نمی‌ترسید و حس ترس سامون بود که او را وامی‌داشت بپندارد برادرش هم می‌ترسد (دولت‌آبادی، ۱۳۹۵).

۱۹. به‌طور کلی، تو کم‌وکسر نداری؟ (ساعدی، ۱۳۷۷).

شکل (۶) شبکهٔ نظام افزودهٔ نقش‌گفتارِ هالیدی و متیسن (۲۰۱۴) است.

شکل (٦). شبکهٔ نظام افزودهٔ نقش‌گفتار (هالیدی و متیسن، ٢٠١٤: ١٩١؛ جعفری، ١٣٨٨: ١٤٢)

۲–۲ واژه‌های دستوری

بنابر تعریف ردفورد[2] (۲۰۰۴)، واژه‌های دستوری، اطلاعاتی دربارهٔ نقش‌های دستوری منتسب‌به برخی مفاهیم (مانند شخص، شمار، جنس و حالت) دارند و معنای قائم به ذات ندارند. با این توضیح، حرف تعریف[3]، کمی‌نما[4]، ضمیر، فعل کمکی، علامت مصدر و متمم‌نما[5] جزءِ مقوله‌های دستوری به حساب می‌آیند. در این پژوهش، در تعریف واژه‌های دستوری دو ملاک در نظر گرفته شده است. اگر واژهٔ موردنظر مسئول انتقال مفهومی دستوری باشد یا اگر خودش به‌صورت مستقل معنی نداشته باشد، آنگاه آن واژه، دستوری است. به‌این‌ترتیب، حروف اضافه که حوزهٔ معنایی‌شان در هم‌نشینی با واژه‌های دیگر مشخص می‌شود، دستوری محسوب می‌شوند (میرزایی و صفری، ۱۳۹۴: ۲۵۴–۲۵۵). درمجموع، در پژوهش حاضر، حرف اضافه، کمی‌نما، ضمیر، فعل کمکی، همکرد فعل (فعل سبک[6])، متمم‌نمای «که»، حرف تعریف «یه» و نقش‌نمای مفعولی «را» کلمات دستوری به شمار می‌آیند.

۳– پیکرهٔ پژوهش

در پژوهش حاضر، به‌منظور طراحی سامانه‌ای برای تشخیص خودکار هویت نویسندهٔ متون فارسی، پیکره‌ای از آثار هفت نویسندهٔ معاصر ایرانی به نام‌های هوشنگ گلشیری، بزرگ علوی، احمد محمود، محمود دولت‌آبادی، نادر ابراهیمی، جلال آل‌احمد و غلامحسین ساعدی گردآوری و هنجارسازی[7] شد. منبع به‌کاررفته برای گردآوری این داده، پایگاه داده‌های زبان فارسی (عاصی، ۱۹۹۷) و پیکرهٔ

1. validity
2. A. Radford
3. article
4. quantifier
5. complementizer
6. light verb
7. normalization

متنی زبان فارسی (بی‌جن‌خان و دیگران، ۲۰۱۱) بوده است. علاوه‌براین، آثار مختلف هفت نویسنده در شبکهٔ جهانی وب جست‌وجو شد و کتاب‌های آن‌ها که در قالب پی‌دی‌اف[1] بودند، با نرم افزار ایبو[2] به قالب متنی[3] درآمدند. درنهایت، کل پیکره هنجارسازی و یکدست شد. در جدول (۳)، آماره‌های این پیکره نشان داده شده است. شایان ذکر است ازآنجاکه در بیشتر این آثار از زبان «محاوره» استفاده شده است و ابزارهای پیش‌پردازشگر[4] زبان فارسی دربارهٔ زبان محاوره کارایی نامطلوبی دارند یا دسترسی به آن‌ها محدود است، پیکره به‌صورت خام و فاقد برچسب در نظر گرفته شده است. همچنین روش پیشنهادی در این پژوهش مستقل از ابزارهای پیش‌پردازشگر است.

جدول (۳). آماره‌های پیکرهٔ داستانی

نام نویسنده	تعداد آثار	مجموع کل واژه‌ها
نادر ابراهیمی	۸	۲۶۱۸۶۶
احمد محمود	۹	۳۶۱۵۶۵
جلال آل‌احمد	۹	۲۵۴۲۷۶
بزرگ علوی	۷	۳۲۶۸۰۲
هوشنگ گلشیری	۱۲	۲۴۷۷۹۱
محمود دولت آبادی	۸	۳۴۶۱۳۲
غلامحسین ساعدی	۶	۳۲۳۳۵۸
مجموع	۵۹	۲۱۲۱۷۹۰

۴- روش پژوهش

در پژوهش پیش‌رو، برای تشخیص خودکار هویت نویسندهٔ متن از رویکرد نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) و روش پیشنهادی در پژوهش آرگامون و دیگران (۲۰۰۷) استفاده شده است. به این منظور، ابتدا فهرستی شامل واژه‌های دستوری، یک مجموعهٔ واژگان برای افزوده‌های ربطی، وجه و نگرشی براساس رویکرد نقش‌گرای نظام‌مند هالیدی و پیکره‌ای از آثار هفت نویسندهٔ معاصر گردآوری شد (ارجاع به بخش ۳). سپس الگوریتمی[5] طراحی شد که کلماتی از متون پیکره را که در فهرست واژه‌های دستوری یا مجموعهٔ واژگان تعریف شده‌اند، استخراج و بسامد نسبی شاخص‌های تعریف‌شده برای هر متن را محاسبه می‌کرد. بسامد نسبی شاخص‌ها در این پژوهش شامل بسامد نسبی هریک از واژه‌های دستوری و بسامد نسبی هر گره در شبکهٔ نظام‌های مبتنی‌بر دستور نقش‌گرای هالیدی (هالیدی و متیسن، ۲۰۱۴) است. در مرحلهٔ بعد، مقدار هریک از شاخص‌ها در هر متن به‌منزلهٔ معیاری برای تمایزگذاشتن میان متون مختلف در نظر گرفته شد. به این شکل که الگوریتم یادگیری ماشینی، بُردار ویژگی‌ها[6] را که شامل هریک از شاخص‌های تعریف‌شده برای متون است، دریافت می‌کرد و باتوجه‌به تفاوت میان مقادیر این شاخص‌ها بین سبک نویسندگان تمایز قائل می‌شد. در ادامه، همهٔ این مراحل به تفصیل شرح داده می‌شود.

۴-۱ استخراج کلمات از متن

واژگان، کلمات و عباراتی را شامل می‌شود که به‌صورت نیمه‌خودکار[7] از پیکره یا شبکهٔ واژگان گردآوری شده است. هر مدخلِ واژگان شماری از ویژگی‌های معنایی براساس دستور نقش‌گرای نظام‌مند را دربردارد. برای گردآوری ویژگی‌های معنایی واژگان از سه منبع پیکرهٔ گفتمانی زبان فارسی (میرزایی و صفری، ۲۰۱۸)، فارس‌نت (شمس‌فرد و دیگران، ۲۰۱۰) و پیکرهٔ گردآوری‌شده در پژوهشی در حوزهٔ دستور نقش‌گرای نظام‌مند (جعفری، ۱۳۸۸) بهره گرفته شده است. پیکرهٔ گفتمان زبان فارسی جزءِ معدود پیکره‌های جهان است که در سطح گفتمان به‌صورت دستی برچسب‌گذاری شده است. پیکرهٔ گفتمان زبان فارسی در دو سطح درون‌جمله‌ای (بینابندی) و درون متنی (بیناجمله‌ای) تهیه شده است. کلیات شیوهٔ برچسب‌گذاری این پیکره که درمجموع حجمی بالغ بر یک میلیون کلمه دارد، براساس

1. PDF
2. https://www.eboo.ir/
3. text (txt)
4. preprocessor
5. algorithm
6. feature vector

۷. منظور از نیمه‌خودکار این است که در ابتدا فهرست کلمات به‌صورت خودکار گردآوری می‌شود؛ سپس زبان‌شناس به‌صورت دستی آن را دسته‌بندی و برچسب‌گذاری می‌کند.

استاندارد درخت‌بانک گفتمانی پن[1] است (پراساد[2] و دیگران، ۲۰۰۸). این شیوه‌نامه به‌لحاظ نظری در بحث روابط منطقی از رویکرد نقش‌گرای هالیدی و متیسن (۲۰۱۴) در فرانقش منطقی[3] پیروی می‌کند. علاوه‌براین، در پیکرهٔ گفتمان برای سه مفهوم وجه، قطبیت و قطعیت برچسب‌گذاری صورت گرفته است (شمس‌فرد و بی‌جن‌خان، ۱۴۰۱: ۷۳). فارس‌نت نیز بزرگ‌ترین شبکهٔ واژگان[4] زبان فارسی است. جعفری (۱۳۸۸)، افزوده‌ها در زبان فارسی را براساس رویکردهای نقشی و صوری بررسی و دراین‌زمینه پیکره‌ای گردآوری کرده است. در ادامه، نحوهٔ گردآوری واژگان برای انواع افزوده‌ها و فهرست واژه‌های دستوری بیان شده است.

همان‌گونه که پیش‌تر اشاره شد، افزوده‌های ربطی در فرانقش متنی مطرح هستند و بین دو بند یا دو جملهٔ پیوند منطقی برقرار می‌کنند. برای تهیهٔ فهرست افزوده‌های ربطی، کل حروف ربط موجود در پیکرهٔ گفتمان استخراج شد. در پیکرهٔ گفتمان به پیروی از درخت‌بانک گفتمانی پن برچسب‌های نمایندهٔ معنای منطقی در چهار گروه کلی شامل زمانی[5]، وابستگی[6]، مقایسه[7] و گسترش[8] دسته‌بندی می‌شوند. در این دسته‌بندی هر گره به دو زیربخش «نوع» و «زیرنوع» تقسیم‌بندی می‌شود (شمس‌فرد و بی‌جن‌خان، ۱۴۰۱: ۸۰). ازآنجاکه هریک از حروف ربط می‌تواند در پیکره بیش از یک برچسب داشته باشد، فقط آن حروفی انتخاب شد که در بیش از ۶۰ درصد موارد، یک نوع برچسب معین دارد. به‌این‌ترتیب، ۳۷۲ حرف ربط انتخاب شد. پس از استخراج این فهرست، برچسب هریک از حروف ربط به‌صورت دستی بررسی شد و براساس شبکهٔ نظام ربط نمایش‌داده‌شده در شکل (۱) بازبینی و دوباره دسته‌بندی شد.

برای تهیهٔ فهرست افزوده‌های وجه و نگرشی از پیکرهٔ گفتمان (میرزایی و صفری، ۲۰۱۸)، فارس‌نت (شمس‌فرد و دیگران، ۲۰۱۰) و پیکرهٔ پژوهش جعفری (۱۳۸۸) استفاده شد. برای تهیهٔ فهرست افزوده‌های وجه، فهرست کلمات دارای سه نوع برچسب «وجه»، «قطبیت» و «قطعیت» از پیکرهٔ گفتمان استخراج شد و براساس شبکهٔ نظام وجه‌نمایی که در شکل (۲) مشاهده می‌شود، دسته‌بندی شد. علاوه‌براین، از پیکرهٔ پژوهش جعفری (۱۳۸۸) برای تهیهٔ فهرست افزودهٔ وجه و نگرشی استفاده شد. برای افزوده‌های نگرشی مستخرج از پیکرهٔ جعفری، فهرست کلمات مترادف با آن‌ها با کمک فارس‌نت یافته و به مجموعهٔ واژگان افزوده شد. ازآنجاکه در برخی آثار نویسندگان پیکرهٔ معاصر از زبان محاوره استفاده شده است، بعضی عبارت‌ها که صورت محاوره‌ای دارند؛ مانندِ «معلومه»، «ممکنه» نیز برچسب‌گذاری و به مجموعهٔ واژگان افزوده شد. همچنین صورت‌های منفی مانند «معلوم نیست»؛ درمقابلِ «معلوم است» نیز در مجموعهٔ واژگان گنجانده شد. علاوه‌براین، صورت‌های تصریفی افعالی مانند «فکرکردن»، «حس‌کردن»، «خواستن» و «دوست‌داشتن» نیز در مجموعهٔ واژگان قرار گرفت. درمجموع، ۳۲۲ افزودهٔ وجه و ۲۰۱ افزودهٔ نگرشی برچسب‌گذاری شد.

ازآنجاکه فهرست واژه‌های دستوری و محتوایی به حوزهٔ تخصصی یا گونهٔ[9] هر متن وابسته است (میرزایی و صفری، ۱۳۹۴)، برای تهیهٔ فهرست واژه‌های دستوری، از پیکرهٔ گردآوری‌شده در پژوهش حاضر استفاده شد. به این منظور، از روش وزن‌دهی فرکانس کلمه–معکوس فرکانس سند[10] استفاده شد[11]. براساس معیار فرکانس کلمه–معکوس فرکانس سند، ۱۹۷ واژه‌ای که مقادیر کوچک‌تری داشتند، واژهٔ دستوری در نظر گرفته شدند. تا اینجا نحوهٔ گردآوری واژگان شرح داده شد. در ادامه، نحوهٔ عملکرد سامانه تشریح می‌شود.

در ابتدا، کلماتی از متن که در مجموعهٔ واژگان تعریف شده‌اند، استخراج می‌شوند؛ سپس براساس برچسب آن واژه یا عبارت در مجموعهٔ واژگان، برچسب‌گذاری می‌شوند. برای نمونه، حرف ربط «برای مثال»، در مجموعهٔ واژگان براساس نظام حروف ربط به‌ترتیب سلسله‌مراتب دارای برچسب «تشریحی»، «بدلی»، «نمونه‌سازی» است. سامانه، تمام موارد وقوع عبارتِ «برای مثال» را در متن کشف و براساس برچسب واژگان برچسب‌گذاری می‌کند. برای استخراج کلمات از متن سعی بر آن بود که اگر یک کلمه یا عبارت زیرمجموعهٔ عبارت دیگری در مجموعهٔ واژگان باشد (مانندِ «که» و «به‌محض آنکه»)، برچسب عبارت بزرگ‌تر در نظر گرفته شود.

1. Penn discourse treebank
2. R. Prasad
3. logical metafunction
4. WordNet
5. temporal
6. contingency
7. comparison
8. expansion
9. genre
10. tf-idf

۱۱. در این روش میزان تکرار یک کلمه در یک مستند درمقابلِ تعداد تکرار آن در مجموعهٔ کلیهٔ مستندات سنجیده می‌شود. در این روش به هر کلمه در یک متن عددی تخصیص می‌دهد. این معیار برای کلماتی که تقریباً در تمام اسناد پیکره یافت شوند، دارای مقادیر کمینه است. این کلمات همان واژه‌های دستوری تلقی می‌شوند (شوتز و دیگران، ۲۰۰۸).

۲-۴ محاسبهٔ بسامد نسبی شاخص‌ها

پس از استخراج کلمات از متن، بسامد نسبی مقادیر شاخص‌های معنایی برای هر متن محاسبه می‌شود. خروجی این مرحله بردار ویژگی‌های معنایی است که توصیفگر متن است. براساس ویژگی‌های مبتنی‌بر دستور نقش‌گرای هالیدی (هالیدی و متیسن، ۲۰۱۴) و واژه‌های دستوری هر متن با بردار ویژگی‌های عددی تعریف می‌شود.

بسامد نسبی هر گره در نظام‌های مبتنی‌بر دستور نقش‌گرای هالیدی[۱]، بسامد انتخاب گره O1 برحسب انتخاب O2 است که گره والد آن در شبکهٔ نظام محسوب می‌شود. درحقیقت، گره O2 دارای برچسب درشت[۲] و O1 دارای برچسب ریز[۳] است و زیرمجموعهٔ آن برچسب تلقی می‌شود. بسامد نسبی O1 با داشتن گره O2 به‌منزلهٔ گره والد به این شکل تعریف می‌شود:

$$RF_d(O_1|O_2)=\frac{N_d(O_1,O_2)}{N_d(O_2)}$$ (فرمول ۱)

(آرگامون و دیگران، ۲۰۰۷)

در این فرمول، $N_d(O1,O2)$ تعداد رخداد هم‌زمان O1 و O2 است و $N_d(O2)$ تعداد رخداد O2 در هر متن است و d به معنای متن یا همان سند است (آرگامون و دیگران، ۲۰۰۷: ۸). به‌بیان ساده‌تر، چه نسبت از تعداد رخداد کلمات در متن که دارای برچسب O2 هستند، دارای برچسب ریزتر O1 نیز هستند.

برای نمونه، در شبکهٔ نظام حروف ربط، گره‌های «تشریحی»، «گسترشی» و «تفصیلی» زیرمجموعهٔ گره حرف ربط محسوب می‌شوند. بسامد نسبی گره «تشریحی» با داشتن گره «حرف ربط» به این معناست که چه نسبت از حروف ربط به‌کارگرفته‌شده در یک متن، از نوع تشریحی هستند. دربارهٔ شبکهٔ نظام وجه‌نمایی و افزودهٔ نگرشی نیز بسامد نسبی براساس فرمول (۱) محاسبه می‌شود. برای مثال، در شبکهٔ نظام وجه‌نمایی، گره‌های «بالا»، «پایین» و «میانه» زیرمجموعهٔ گره «مقدار» هستند. بسامد نسبی گره «بالا» با داشتن گره «مقدار» به این معناست که چه نسبت از افزوده‌های وجه در یک متن، دارای مقدار «بالا» هستند.

علاوه‌بر ویژگی بسامد نسبی، صرفاً برای نظام وجه‌نمایی، ویژگی‌هایی براساس ترکیب گره‌ها تعریف می‌شود. برای هر جفت گره در زیرنظام‌های مختلف وجه‌نمایی (مانند «نوع» و «مقدار»)، بسامد نسبی واژه‌هایی که با نوع دو گره برچسب خورده‌اند، براساس ترکیب گره‌های والد آن‌ها تعریف می‌شود (آرگامون و دیگران، ۲۰۰۷). برای مثال، بسامد نسبی جفت گره «وجه احتمال-عادت» و «بالا» با داشتن جفت گره «نوع» و «مقدار» بیان می‌کند که چه نسبت از کلماتی که دارای «وجه احتمال-عادت» هستند، در زیرنظام «مقدار» دارای مقدار «بالا» نیز هستند؛ (مانند «حتماً» و «دائماً»).

برای مقایسهٔ کارایی ویژگی‌های مبتنی‌بر دستور نقش‌گرای هالیدی و ویژگی واژه‌های دستوری، ویژگی بسامد واژه‌های دستوری تعریف می‌شود. به این منظور، واژه‌های دستوری در هر متن شناخته می‌شوند و بسامد نسبی هریک از آن‌ها به‌منزلهٔ یک ویژگی محاسبه می‌گردد. برای محاسبهٔ بسامد نسبی واژه‌های دستوری از فرمول (۲) استفاده شده است:

$$\frac{count(w)}{\sum_{w'\in FW} count(w')}$$ (فرمول ۲)

(آرگامون و دیگران، ۲۰۰۷)

در این فرمول، w هریک از واژه‌های دستوری است که بسامد آن نسبت‌به بسامد کل واژه‌های دستوری در یک متن خاص محاسبه می‌شود (آرگامون و دیگران، ۲۰۰۷: ۱۳). در صورتِ کسر فرمول (۲)، تعداد رخداد یک واژهٔ دستوری خاص و در مخرج آن تعداد رخداد کل واژه‌های دستوری در متن قرار دارد.

۳-۴ یادگیری ماشینی

برای مرحلهٔ یادگیری ماشینی از روش یادگیری با ناظر[۴] استفاده می‌شود. در این روش به یک سامانه، مجموعه‌ای از جفت‌های ورودی-خروجی ارائه می‌شود و سامانه می‌کوشد تابعی از ورودی به خروجی را فراگیرد. یادگیری باناظر به تعدادی دادهٔ ورودی برای آموزش

---

۱. هر شبکهٔ نظام در هر سطح شامل چند گره است. برای مثال، شبکهٔ نظام حروف ربط در سطح اول شامل سه گره، افزودهٔ تشریحی، گسترشی و تفصیلی است. در سطح بعد، گره تشریحی شامل دو گره بدلی و واضح‌سازی است. درحقیقت، گره تشریحی والد دو گره بدلی و واضح‌سازی است. در اینجا گره تشریحی نسبت‌به دو گره بدلی و واضح‌سازی دارای برچسب درشت و دو گره بدلی و واضح‌سازی دارای برچسب ریزتر هستند.

2. coarse-grained
3. fine-grained
4. supervised learning

سامانه نیاز دارد. همان‌طور که پیش‌تر اشاره شد، خروجی مرحلهٔ دوم بردار ویژگی‌های معنایی است که توصیفگر متن است. درمورد تشخیص خودکار نویسندهٔ متن، ورودی‌ها بردار ویژگی‌های زبانی متون هستند و خروجی همان نام نویسندهٔ متن است.

در پژوهش حاضر، برای دسته‌بندی نویسندگان، از روش یادگیری عمیق[1] استفاده شده است. رویکردهای مبتنی‌بر یادگیری عمیق در پژوهش‌های نوین در حوزهٔ پردازش زبان طبیعی ازجمله تشخیص خودکار هویت نویسنده بسیار موردتوجه هستند (اوچندو[2] و دیگران، ۲۰۲۰). یکی از انواع طبقه‌بندهای روش یادگیری عمیق، طبقه‌بند پرسپترون چند لایه[3] است که از آن در این پژوهش بهره گرفته شده است. برای استفاده از این طبقه‌بند از یک ابزار[4] یادگیری ماشین به زبان پایتون[5] استفاده شده است. الگوریتم یادگیری، بردار ویژگی‌های معنایی هر متن را به‌منزلهٔ ورودی دریافت می‌کند. بردار ویژگی‌های معنایی می‌تواند انواع ویژگی‌های مبتنی‌بر دستور نقش‌گرای هالیدی (شبکهٔ نظام حروف ربط، وجه‌نمایی و افزودهٔ نگرشی) یا ترکیب انواع مختلف آن‌ها را دربرگیرد. ازسوی دیگر، این بردار می‌تواند ویژگی بسامد نسبی واژه‌های دستوری را شامل شود. در حالت دیگر، در انتخاب نوع بردار ویژگی می‌توان تمام ویژگی‌های مبتنی‌بر دستور نقش‌گرای هالیدی و ویژگی بسامد نسبی واژه‌های دستوری را باهم آمیخت. درنهایت، الگوریتم یادگیری نام نویسندهٔ متن را به‌منزلهٔ خروجی برمی‌گرداند. در ادامه، کارایی این الگوریتم در حالت‌های مختلف بردار ویژگی ارزیابی و گزارش می‌شود.

۵- یافته‌ها

برای ارزیابی کیفیت سامانهٔ طراحی‌شده در تشخیص خودکار هویت نویسندهٔ متن از روش ارزیابی متقاطع K تایی[6] استفاده شده است. اگر کل داده‌ها را به‌طور تصادفی به K زیرنمونه تقسیم کنیم، می‌توانیم در هر مرحله از فرایند، تعداد K-1 از این لایه‌ها را به‌منزلهٔ مجموعه‌دادهٔ آموزشی و یکی را به‌منزلهٔ مجموعه‌دادهٔ آزمون در نظر بگیریم. مشخص است که تعداد تکرارهای فرایند برابر با K خواهد بود (رفائیل‌زاده و دیگران، ۲۰۰۹). درواقع، یادگیری سامانه با کمک دادهٔ آموزشی صورت می‌گیرد و با استفاده از دادهٔ آزمون ارزیابی می‌شود. در این پژوهش، K= ۵ در نظر گرفته شده است. نکتهٔ درخور توجه در ارزیابی این است که نه‌تنها دادهٔ آموزش و آزمون با یکدیگر هم‌پوشانی ندارند، بلکه برای ارزیابی سامانهٔ تشخیص سبک نگارش یک نویسنده، نباید یک داستان خاص بین دادهٔ آموزش و آزمون تقسیم شود. این نکته‌ای است که در برخی پژوهش‌ها نادیده گرفته شده است؛ به‌این‌علت، در ارزیابی سامانه، کل آثار هر نویسنده به پنج قسمت تقسیم شد؛ چهار قسمت برای آموزش و یک قسمت برای آزمون. این تقسیم‌بندی به‌گونه‌ای است که داستان‌های موجود در بخش آموزش از داستان‌های موجود قسمت آزمون متمایز هستند. برای محاسبهٔ کارایی سامانه از معیار دقت[7] استفاده شد. دقت در تشخیص خودکار نویسنده عبارت است از نسبت نتایج درست به‌دست‌آمده از سامانه به کل تعداد نمونه‌های موجود در دادهٔ آزمون (درونه و شریعتی، ۲۰۱۴). ازآنجاکه فرایند ارزیابی پنج بار تکرار شد، درنهایت، از نتایج به‌دست‌آمده در پنج مرحله برای کل سامانه و همین‌طور برای هر نویسنده میانگین وزن‌دار[8] گرفته شد.[9]

الگوی طراحی‌شده در چند حالت آزموده شد: حالت نخست، بردار ویژگی شامل ویژگی‌های تعریف‌شده براساس دستور نقش‌گرای نظام‌مند است؛ حالت دوم، بردار ویژگی از نوع ویژگی‌های واژه‌های دستوری است؛ حالت سوم، ترکیب حالت نخست و دوم است. در هر حالت، دقت سامانه به روش ارزیابی متقاطع K تایی سنجیده شد. در جدول (۴) نتایج ارزیابی سامانه به‌طور خلاصه گزارش شده است. همان‌گونه که مشاهده می‌شود، بهترین نتایج حاصل از ترکیب روش مبتنی‌بر دستور نقش‌گرای نظام‌مند و بسامد واژه‌های دستوری است. همچنین دقت روش مبتنی‌بر واژه‌های دستوری از روش دستور نقش‌گرا بسیار بیشتر است. در شکل (۷)، نتایج ارزیابی سامانه به‌طور کامل و برای هر نویسنده به‌طور مجزا نمایش داده شده است. علاوه‌براین، میزان کارایی تمام ویژگی‌های تعریف‌شده در سامانه

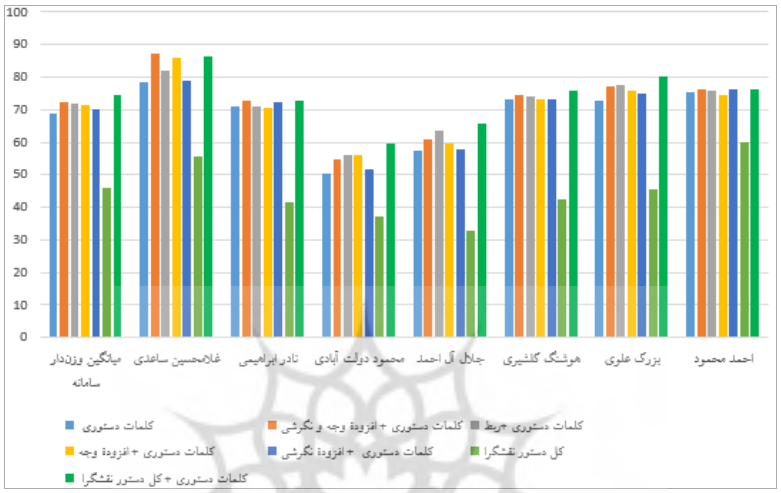
1. deep learning
2. A. Uchendu
3. multilayer perceptron
4. https://scikit-learn.org/
5. python
6. K-Fold Cross Validation
7. accuracy
8. weighted mean

۹. در محاسبهٔ میانگین وزن‌دار یک مجموعه عامل نابرابر، برای هریک از عامل‌ها، وزن یا ارزش معینی در نظر گرفته می‌شود؛ سپس آن عامل در وزن معین ضرب می‌گردد. آنگاه جمع این ارقام به‌دست‌آمده بر مجموع وزن‌ها تقسیم می‌شود.

بررسی و کاراترین ویژگی‌ها در عملکرد سامانه انتخاب شد. برای این کار از آزمون F تحلیل واریانس یک‌طرفه[1] استفاده شد.[2] در این پژوهش مقادیر بردار ویژگی (کل ویژگی‌های مبتنی‌بر دستور نقش‌گرای نظام‌مند و ویژگی واژه‌های دستوری)، براساس نتایج به‌دست‌آمده از آزمون F تحلیل واریانس یک‌طرفه مرتب و ۱۵۰ ویژگی برتر انتخاب شد. براساس این معیار، بسامد نسبی واژه‌های دستوری «آن»، «رو» و «شد» به‌ترتیب دارای بالاترین امتیاز درمیان کل ویژگی‌ها هستند.

جدول (٤). نتایج ارزیابی کلی سامانه براساس معیار دقت

نام مجموعهٔ ویژگی	کلمات دستوری	کلمات دستوری، افزودهٔ وجه و نگرشی	کلمات دستوری و افزودهٔ ربطی	کل دستور نقش‌گرا	کلمات دستوری و کل دستور نقش‌گرا
میانگین وزن‌دار	۶۸.۹۸	۷۲.۵۳	۷۲.۰۵	۴۶	۷۴.۳۱

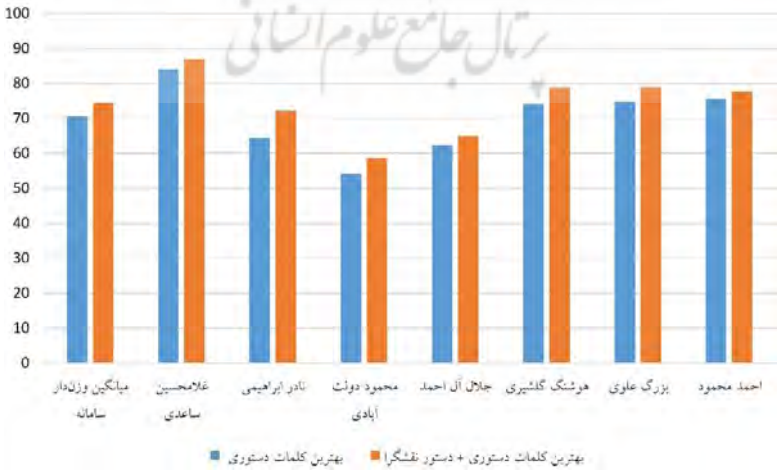


شکل (۷). نمودار مقایسه‌ای ارزیابی سامانه براساس معیار دقت

پس از انتخاب بهترین ویژگی‌ها کارایی سامانه در دو حالت سنجیده شد: ۱. کلمات دستوری؛ ۲. کلمات دستوری (بهترین ویژگی‌ها) و ویژگی‌های مبتنی‌بر دستور نقش‌گرا. این نتایج در جدول (۵) گزارش شده است. همان‌گونه که در شکل (۸) مشاهده می‌شود، بهترین نتایج را ترکیب حالت بهترین کلمات دستوری و دستور نقش‌گرا به دست می‌دهد.

جدول (۵). نتایج ارزیابی سامانه براساس معیار دقت (بهترین ویژگی‌ها)

نام مجموعهٔ ویژگی	میانگین وزن‌دار
بهترین کلمات دستوری	۷۰/۴۷
بهترین کلمات دستوری و دستور نقش‌گرا	۷۴/۴۷



شکل (۸). نمودار مقایسه‌ای ارزیابی سامانه براساس معیار دقت (بهترین ویژگی‌ها)

1. ANOVA

۲. آزمون تحلیل واریانس یک‌طرفه برای آزمون مقایسهٔ میانگین یک متغیر کمّی درمیان بیش از دو گروه مستقل استفاده می‌شود.

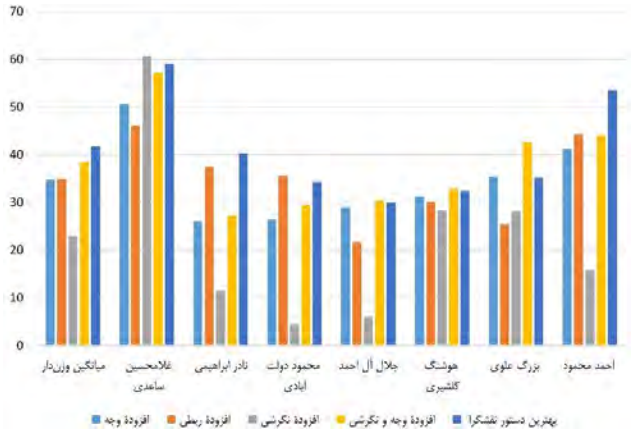
۶- بحث و نتیجه‌گیری

در پژوهش حاضر، برای تعیین هویت نویسندهٔ متن یک سامانهٔ خودکار به روش یادگیری عمیق معرفی شد. برای طراحی سامانه از ویژگی بسامد واژه‌های دستوری و مفاهیم دستور نقش‌گرای نظام‌مند استفاده شد. سپس عملکرد سامانه در سه حالت ویژگی‌های مبتنی‌بر واژه‌های دستوری، ویژگی‌های مبتنی‌بر دستور نقش‌گرای نظام‌مند و ترکیب این دو نوع ویژگی سنجیده و تحلیل شد و بهترین ویژگی‌ها انتخاب شدند. نتایج ارزیابی سامانه نشان می‌دهد روش محاسبهٔ بسامد واژه‌های دستوری نسبت‌به روش مبتنی‌بر دستور نقش‌گرای نظام‌مند (شبکهٔ نظام ربط، وجه‌نمایی وافزودهٔ نگرشی) در تشخیص هویت نویسندهٔ متن برتری دارد؛ اما درصورتی‌که ویژگی‌های زبانی مبتنی‌بر دستور نقش‌گرای نظام‌مند هالیدی درکنار ویژگی بسامد واژه‌های دستوری به کار روند، کارایی سامانه نسبت‌به حالتی که فقط از ویژگی بسامد واژه‌های دستوری استفاده شود، افزایش می‌یابد.

به نظر می‌رسد برای این نتیجه که ویژگی بسامد نسبی واژه‌های دستوری کارایی بالاتری نسبت‌به ویژگی‌های مبتنی‌بر دستور نقش‌گرا در تفکیک سبک نویسندگان داشته، دلایل متفاوتی وجود داشته است. دلیل نخست می‌تواند به کنترل‌ناپذیری خودآگاه نویسنده در استفاده از واژه‌های دستوری مربوط باشد. نویسندهٔ متن واژه‌های دستوری را به‌طور ناخودآگاه به کار می‌گیرد و برخلاف واژه‌های محتوایی کنترل خودآگاه بر به‌کاربردن یا به‌کارنبردن آن‌ها ندارد (گلشایی، ۱۳۹۸). ازآنجاکه تعیین مرز دقیق بین واژه‌های محتوایی و دستوری عملاً امکان‌پذیر نیست، می‌توان درخصوص واژه‌های دستوری یک پیوستار تعریف کرد. در این پیوستار هرچه به‌سمت دستوری‌بودن واژه پیش رویم، کنترل خودآگاه در استفاده از واژه‌ها کاهش می‌یابد. در این پژوهش، براساس آزمون F تحلیل واریانس یک‌طرفه، ضمایر «آن»، «این»، «آن‌ها»، «من» و «خود» ویژگی‌های بسیار مناسبی برای تمایزگذاشتن میان سبک نویسندگان تلقی می‌شوند. نتایج حاصل از پژوهش حاضر درمورد کاراترین ویژگی‌ها در تشخیص هویت نویسنده گواهی بر این مدعاست که ضمایر در این پیوستار به‌سمت دستوری‌بودن تمایل دارند و نویسنده کنترل خودآگاه بر به‌کارگیری آن‌ها ندارد. بدیهی است که این پیوستار خود می‌تواند موضوع یک پژوهش جدید باشد.

دومین دلیل برای اینکه در تشخیص هویت نویسنده ویژگی‌های مبتنی‌بر دستور نقش‌گرا به‌طور مستقل مؤثرتر از واژه‌های دستوری عمل نکرده‌اند، می‌تواند این باشد که استفاده از واژه‌های دستوری به‌جای یکدیگر، تفاوت سبکی بیشتری ایجاد می‌کند. برای مثال، تفاوت سبک در متونی که از واژه‌های «من» و «اینجانب» استفاده می‌کنند، درخور توجه است؛ این درحالی است که برای مثال، «مسلماً» و «مسلم است» که در زیرنظام جهت‌گیری از شبکهٔ نظام وجه‌نمایی اولی «ضمنی» و دومی «صریح» تلقی می‌شوند، تفاوت سبکی چندانی ایجاد نمی‌کنند. این مطلب دربارهٔ «باید» و «نباید» که در زیرنظام قطبیت با یکدیگر تفاوت دارند، نیز صادق است.

دلیل دیگر برای کارایی بیشتر واژه‌های دستوری نسبت‌به ویژگی‌های مبتنی‌بر دستور نقش‌گرا، می‌تواند اختلاف درصد حضور واژه‌های دستوری و کلمات مرتبط با سه نوع ویژگی دستور نقش‌گرا در پیکره باشد. به‌جز افزودهٔ ربطی «و» از شبکهٔ نظام حروف ربط که فراوانی بسیار بالایی در پیکره دارد، درصد حضور واژه‌های دستوری بیشتر از واژه‌های مرتبط با ویژگی‌های مبتنی‌بر دستور نقش‌گرا است. برای بررسی بهتر ویژگی‌های مبتنی‌بر دستور نقش‌گرای هالیدی، هریک از این سه شبکهٔ نظام حروف ربط، وجه‌نمایی و افزودهٔ نگرشی به‌صورت مجزا و در ترکیب با یکدیگر ارزیابی شدند و نتایج آن‌ها با یکدیگر و با حالت بهترین ویژگی‌های سه نوع نظام مقایسه شدند که در شکل (۹) مشاهده می‌شود.



شکل (۹). نمودار مقایسه‌ای کارایی سامانه با انواع ویژگی‌های دستور نقش‌گرا

همان‌گونه که در نمودار (۹) مشاهده می‌شود، ترکیب بهترین ویژگی‌های مرتبط با هر سه نوع افزودهٔ وجه، نگرشی و ربطی بالاترین دقت را درمیانِ دیگر حالت‌ها دارد. پس‌ازآن، افزودهٔ وجهی (افزودهٔ وجه و افزودهٔ نگرشی) قرار دارد. گواه دیگر در برتری افزودهٔ وجهی نسبت‌به افزودهٔ ربطی، درصد حضور ویژگی‌های مرتبط با افزودهٔ وجهی درمیانِ ویژگی‌های منتخب است (۳۹/۰۷ درصد). این درحالی است که درصد حضور ویژگی‌های مرتبط با افزودهٔ ربطی در این فهرست ۱۱/۲۵ درصد است.

تحلیل دقیق‌تر بهترین ویژگی‌های افزودهٔ وجهی نشان می‌دهد که ویژگی‌های مرتبط با زیرنظام‌های «مقدار» و «نوع» از شبکهٔ نظام افزودهٔ وجهی به دو صورتِ مستقل و ترکیب با سایر زیرنظام‌ها بسیار کارا هستند. زیرنظام‌های «قطبیت» و «جهت‌گیری» در ترکیب با سایر زیرنظام‌ها به بهبود عملکرد سامانه کمک شایانی می‌کنند. دربارهٔ زیرنظام وجه‌سازی (وجه احتمال–عادت) علاوه‌بر اینکه کل زیرنظام و زیرنظام «وجه احتمالی» در کارایی سامانه بسیار مؤثر هستند، زیرنظام «وجه عادتی» بیشترین تأثیر را دارد. در زیرنظام تعدیل‌سازی (وجه تمایل-التزام) نیز علاوه‌بر اثرگذاری کل زیرنظام و زیرنظام «وجه التزامی»، زیرنظام «وجه تمایلی» بالاترین کارایی را دارد. در شبکهٔ نظام افزودهٔ نگرشی، زیرنظام گزاره‌ای به‌ویژه از نوع «اخلاقی» و «فرض» بیشترین میزان اثربخشی را دارد.

در نظام شبکهٔ حروف ربط، بسامد افزوده‌های ربطی هریک از سه زیرنظام «تشریحی»، «گسترشی» و «تفصیلی» (به‌طور مجزا) به کل افزوده‌های ربطی ویژگی‌هایی ارزشمند به شمار می‌روند. برپایهٔ ویژگی‌های منتخب، زیرنظام «گسترشی» نتایجی مطلوب را به دست می‌دهد. در این زیرنظام، بسامد نسبی افزوده‌هایی از نوع «افزایشی» و «تباینی» امتیاز بالایی دارند. در زیرنظام «تفصیلی»، زیرنظام «سببی–شرطی» بسیار مؤثر است. در این زیرنظام، زیرنظام «مشخص» به‌ویژه حروف ربط از نوع «علت» و پس‌ازآن، «نتیجه» تأثیرگذار هستند. در زیرنظام «تشریحی»، افزوده‌های ربطی «بدلی» به‌ویژه از نوع «نمونه‌سازی» کارآمد هستند.

نتایج کلی این پژوهش همسو با نتایج مقالهٔ مرجع آرگامون و دیگران (۲۰۰۷) است. در پژوهش آرگومان و دیگران (۲۰۰۷)، روش بسامد نسبی واژه‌های دستوری نسبت‌به روش دستور نقش‌گرا (شبکهٔ نظام حروف ربط، شبکهٔ نظام وجه‌نمایی و شبکهٔ نظام افزودهٔ نگرشی) برتری دارد و ترکیب روش واژه‌های دستوری و دستور نقش‌گرا بالاترین کارایی را دارد. در پژوهش آرگومان و دیگران (۲۰۰۷) روش دستور نقش‌گرا به افزایش کارایی سامانه کمک کرده است که با نتایج کلی پژوهش حاضر همسو است.

یکی از تفاوت‌های پژوهش حاضر با پژوهش آرگومان و دیگران (۲۰۰۷) این است که در پژوهش آرگومان و دیگران (۲۰۰۷)، تعداد واژه‌های دستوری ۶۷۵ واژه در نظر گرفته شده است؛ درحالی که در پژوهش حاضر تعداد کل واژه‌های دستوری ۱۹۷ واژه است و پس از مرحلهٔ یافتن بهترین ویژگی‌ها تعداد ۷۴ واژه از واژه‌های دستوری انتخاب شده است. آرگومان و دیگران (۲۰۰۷) با استفاده از الگوریتم طبقه‌بندی ماشین بردار پشتیبان[1] و با ۶۷۵ واژهٔ دستوری، تنها با ویژگی بسامد نسبی واژه‌های دستوری به دقتی حدود ۸۵ درصد در تشخیص هویت نویسنده برای زبان انگلیسی دست یافته‌اند. در این پژوهش، برای مقایسهٔ نتایج کار با مقالهٔ مرجع از ۶۶۴ واژهٔ دستوری و الگوریتم ماشین بردار پشتیبان استفاده شد و دقتی معادل ۸۲.۷۵ درصد برای زبان فارسی به دست آمد که به نتایج پژوهش مرجع نزدیک است. در این مقایسه مشخص شد فرایند یادگیری دچار بیش‌برازش[2] شده است[3] و این مورد را می‌توان از نقاط ضعف پژوهش مرجع یاد کرد که البته در آن مقاله نیز چند بار به آن اشاره شده است. به‌این‌دلیل، در پژوهش حاضر برای جلوگیری از بیش‌برزاش تعداد واژه‌های دستوری بسیار کمتر از مقالهٔ مرجع و برابر با ۱۹۷ کلمه در نظر گرفته شده است. جدا از اینکه تعداد واژه‌های دستوری چقدر است، نتایج حاصل از افزودن ویژگی‌های مبتنی‌بر دستور نقش‌گرا به سامانه همسو با پژوهش آرگومان و دیگران (۲۰۰۷) است. در مقالهٔ مرجع کارایی سامانه با افزودن ویژگی‌های مبتنی‌بر دستور نقش‌گرا، حدود ۵ درصد بیشتر از حالتی است که صرفاً از واژه‌های دستوری استفاده شود. این افزایش کارایی برابر با افزایش کارایی پژوهش حاضر در حالت قرارگرفتن ویژگی‌های مبتنی‌بر دستور نقش‌گرا درکنار واژه‌های دستوری است (ارجاع به جدول ۴). تفاوت دیگر نتایج مقالهٔ مرجع با پژوهش حاضر این است که ویژگی‌های مبتنی‌بر دستور نقش‌گرای هالیدی (به‌طور مستقل از واژه‌های دستوری) برای زبان فارسی به اندازهٔ زبان انگلیسی در تشخیص هویت نویسنده کارایی نداشته‌اند. علاوه‌بر تفاوت‌های دو زبان فارسی و انگلیسی، به نظر می‌رسد تفاوت در حجم واژگان تعریف‌شده برای شبکهٔ نظام‌های دستور نقش‌گرا از عوامل این تفاوت باشد.

1. SVM
2. overfitting

۳. بیش‌برازش به این معناست که الگوریتم داده‌های آموزشی را بسیار خوب یاد گرفته است؛ ولی توانایی پیش‌بینی خوب داده‌های جدید را ندارد.

شبکهٔ نظام‌های مختلف دستور نقش‌گرا (شبکهٔ نظام حروف ربط، شبکهٔ نظام وجه‌نمایی و شبکهٔ نظام افزودهٔ نگرشی) به‌طور مستقل نتایج درخور توجهی را در تعیین هویت نویسنده به دست نمی‌دهند. درواقع، نتایج مطلوب، نتایجی است که از ترکیب ویژگی‌های منتخب از هر سه نظام و ویژگی واژه‌های دستوری به دست می‌آید. گرچه رویکرد نقش‌گرای نظام‌مند هالیدی به‌ویژه در فرانقش بینافردی دربارهٔ وجه، دسته‌بندی دقیقی ارائه می‌کند، شبکهٔ نظام‌های این دستور به‌صورت مستقل از واژه‌های دستوری عملکرد مطلوبی در تشخیص هویت نویسندهٔ متن ندارند و فقط درکنار واژه‌های دستوری به افزایش کارایی سامانه یاری می‌رسانند.

از کاستی پژوهش حاضر، می‌توان به این مسئله اشاره کرد که به‌جز شمار محدودی از افعال در فهرست افزودهٔ وجهی و افعال کمکی، به مقولهٔ فعل در طراحی سامانهٔ هویت نویسنده توجه نشد. در پژوهش‌های آینده می‌توان حجم واژگان برای افزوده‌های ربطی و افزوده‌های وجهی به‌ویژه با مقولهٔ فعل را گسترش داد و به این شکل به پژوهش‌های حوزهٔ دستور نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) کمک کرد. در پژوهش‌های آینده می‌توان الگوریتم استخراج سه نوع ویژگی دستور نقش‌گرا و واژه‌های دستوری را دربارهٔ دیگر آثار داستانی نیز به کار برد و بسامد انواع ویژگی‌ها را به دست آورد و از آن در نقد ادبی و سبک‌شناسی برای اهداف زبانی استفاده کرد.

منابع

آل‌احمد، جلال (۱۳۴۶). *نفرین زمین*. تهران: فردوس.

آل‌احمد، جلال (۱۳۵۰). *پنج داستان*. تهران: فردوس.

ابراهیمی، نادر (۱۳۷۴). *یک عاشقانهٔ آرام*. تهران: روزبهان.

ابراهیمی، نادر (۱۳۹۹). *بر جاده‌های آبی سرخ*. تهران: روزبهان.

جعفری، آزیتا (۱۳۸۸). بررسی افزوده‌ها در زبان فارسی: براساس رویکردهای نقشی و صوری. *دستور (ویژه‌نامه نامهٔ فرهنگستان)*، ۵(۱)، ۱۲۸–۱۵۵.

حسین حمه، همزه؛ علی‌اکبری، نسرین؛ کریمی، یادگار (۱۴۰۰). بررسی وجه و وجهیت در کردی سورانی: تحلیلی نقش‌گرا. *مطالعات زبان‌ها و گویش‌های غرب ایران*، ۹(۴)، ۱–۲۳.

دولت‌آبادی، محمود (۱۳۹۵). *روزگار سپری‌شدهٔ مردم سالخورده*. تهران: چشمه.

دولت‌آبادی، محمود (۱۴۰۱). *کلیدر*. چاپ ۳۷. تهران: فرهنگ معاصر.

ساعدی، غلامحسین (۱۳۷۷). *آشفته‌حالان بیداربخت*. تهران: نگاه.

ساعدی، غلامحسین (۱۳۹۷). *غریبه در شهر*. تهران: نگاه.

شمس‌فرد، مهرنوش؛ بی‌جن‌خان، محمود (۱۴۰۱). *پردازش متن و گفتار فارسی: مروری بر مبانی نظری و آخرین یافته‌های پژوهشی*. تهران: سمت.

عارفی، سمیه؛ بصیری، محمداحسان؛ روزمند، امید (۱۴۰۰). انتخاب ویژگی برای شناسایی نویسنده در متون کوتاه برخط فارسی. *فنّاوری اطلاعات و ارتباطات ایران*، ۱۳ (۴۷–۴۸)، ۳۵–۵۷.

علوی، بزرگ (۱۳۸۶). *ورق‌پاره‌های زندان*. تهران: نگاه.

علوی، بزرگ (۱۳۹۹). *گیله‌مرد*. تهران: نگاه.

فرهمندپور، زینب؛ نیک‌مهر، هومن؛ منصوری‌زاده، محرم؛ طبیب‌زاده، امید. (۱۳۹۱). یک سیستم نوین هوشمند تشخیص هویت نویسنده فارسی زبان براساس سبک نوشتاری. *محاسبات نرم*، ۱(۲)، ۲۶–۳۵.

گلشائی، رامین (۱۳۹۸). واژه‌های دستوری به‌مثابه نشانگرهای گویش فردی: رویکردی پیکره‌ای به شناسایی هویت نویسنده در زبان فارسی. *جستارهای زبانی*، ۱۰ (۳)، ۳۱۷–۲۹۳.

گلشیری، هوشنگ (۱۳۵۰). *کریستین و کید*. تهران: کتاب زمان.

گلشیری، هوشنگ (۱۳۷۰). *در ولایت هوا*. استکهلم: عصر جدید.

گلشیری، هوشنگ (۱۴۰۰). *شازده احتجاب*. چاپ ۱۸. تهران: نیلوفر.

محمود، احمد (۱۳۵۳). *همسایه‌ها*. تهران: امیرکبیر.

محمود، احمد (۱۳۸۱). *غریبه‌ها و پسرک بومی*. تهران: معین.

میرزایی، آزاده (۱۳۹۷). بازتعریف مفاهیم بندِ پایه و بندِ پیرو براساس رویکرد نقش‌گرا. *زبان و زبان‌شناسی*، ۱۳ (۲۶)، ۱۱۷–۱۳۲.

میرزایی، آزاده (۱۴۰۰). رابطهٔ قطبیت و وجهیت بندی در زبان فارسی. *مطالعات زبان‌ها و گویش‌های غرب ایران*، ۹(۱)، ۱۱۳–۱۳۵.

میرزایی، آزاده؛ صفری، پگاه (۱۳۹۴). ساختِ واژه–متن‌های تخصصی و عمومی زبان فارسی براساس بسامدگیری واژه‌های نقشی و محتوایی. مجموعه‌مقالات *نخستین همایش ملّی زبان‌شناسی پیکره‌ای* (صص. ۱۷۵–۱۹۱). تهران: نویسه پارسی.

References

Alavi, B. (2007). *Scrap papers from prison*. Tehran: Negah. (In Persian)

Alavi, B. (2020). *Gilemard*. Tehran: Negah. (In Persian)

Al-e-Ahmad, J. (1967). *The cursing of the land*. Tehran: Ferdous. (In Persian)

Al-e-Ahmad, J. (1971). *Five stories*. Tehran: Ferdous. (In Persian)

Arefi, S., Basiri, M. E., & Roozmand, O. (2021). Feature selection for author identification of Persian online short texts. *Journal of Information and Communication Technology*, *13*(47-48), 35-57. https://dorl.net/dor/20.1001.1.27170414.1400.13.47.4.0 (In Persian)

Argamon, S., & Koppel, M. (2013). A systemic functional approach to automated authorship analysis. *Journal of Low & Policy*, *21*(2), 299-315.

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.

Assi, S. M. (1997). Farsi linguistic database (FLDB). *International journal of Lexicography*, *10*(3), 5.

Bijankhan, M., Sheykhzadegan, J., Bahrani, M., & Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. *Language resources and evaluation*, *45*(2), 143-164. https://doi.org/10.1007/s10579-010-9132-x.

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, *25*(4), 431-447. https://doi.org/10. 1093/applin/25.4.431.

Dabagh, R. M. (2007). Authorship attribution and statistical text analysis. *Advances in Methodology and Statistics*, *4*(2), 149-163. https://doi.org/10.51936/uvjx7198.

Darooneh, A. H., & Shariati, A. (2014). Metrics for evaluation of the author's writing styles: Who is the best? *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *24*(3). https://doi.org/10.1063/1.4895468.

Dowlatabadi, M. (2016). The elderly people's elapsed time. Tehran: Cheshmeh. (In Persian)

Dowlatabadi, M. (2022). *Kelidar* (37th print). Tehran: Farhange Moaser. (In Persian)

Ebrahimi, N. (1995). *A quiet Romance*. Tehran: Rouzbahan. (In Persian)

Ebrahimi, N. (2020). *On the blue and red paths*. Tehran: Rouzbahan. (In Persian)

Farahmandpoor, Z., Nikmehr, H., Mansoorizade, M., & Tabibzadeh Ghamsary, O. (2013). A novel intelligent Persian authorship system based on writing style. Soft *Computing Journal*, *1*(2), 26-35. https://dorl.net/dor/20.1001.1.23223707.1391.1.2.60.9 (In Persian)

Golshaie, R. (2019). Function words as idiolect markers: A corpus-based approach to authorship attribution in Farsi. *Language Related Research*, *10*(3),293-317. (In Persian)

Golshiri, H. (1971). *Christine and kid*. Tehran: Ketab-e Zaman. (In Persian)

Golshiri, H. (1991). *Dar velayat-e Hava*. Stockholm: Asr-e Jadid. (In Persian)

Golshiri, H. (2021). *Prince Ehtejab* (18th print). Tehran: Niloufar. (In Persian)

Halliday, M. A. K., & Matthiessen, C. M. M. (2014). *Halliday's introduction to functional Grammar* (4th ed.). Oxon: Routledge.

Hussein Hama, H., Ali-akbari, N., & Karimi, Y. (2022). Mood and modality in Sorani Kurdish: A functional analysis. *Researches in Western Iranian Languages and Dialects*, *9*(4), 1-23. https://doi.org/10.22126/JLW.2021.6008.1504. (In Persian)

Jafari, A. (2008). An analysis of adjuncts in Persian: A syntacto-discoursal approach. *Dastoor*, *5*(1), 128-155. (In Persian)

Mahmoud, A. (1974). *The neighbors*. Tehran: Amirkabir. (In Persian)

Mahmoud, A. (2002). *The strangers and the little native boy*. Tehran: Moin. (In Persian)

Mandenhal, T. C. (1887). The characteristics curves of composition science. *Science*, *9*(214s), 237-246.

Martinez-Galicia, J. A., Embarcadero-Ruiz, D., Ríos-Orduña, A., & Gómez-Adorno, H. (2022). Graph-based siamese network for authorship verification. *CLEF 2022 Labs and Workshops*, *Notebook Papers*. Italy: Bologna.

Matthiessen, C. M., & Bateman, J. A. (1991). *Text generation and systemic-functional linguistics: Experiences from English and Japanese*. London and New York: Frances Pinter Publishers.

Mirzaei, A. (2017). Redefining the concepts of dependent and independent clauses according to a functional approach. *Language and Linguistics, 13*(26), 117-132. (In Persian)

Mirzaei, A. (2018). *An introduntion to corpus linguistics*. Tehran: Allameh Tabataba'i University Press. (In Persian)

Mirzaei, A. (2021). The relationship between polarity and clausal modality in Persian. *Journal of Western Iranian Languages and Dialects, 9*(1), 113-135. https://doi.org/10.22126/jlw.2020.5372.144. (In Persian)

Mirzaei, A., & Safari, P. (2014). Building specialized and general documents in Persian based on the frequency of function and content words. *In proceeding of The 1st National Conference on Corpus Linguistics* (PP. 175-192). Tehran: Neviseh Parsi. (In Persian)

Mirzaei, A., & Safari, P. (2018). Persian discourse treebank and coreference corpus. *In Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018)*. Japan: Miazaky.

Najafi, M., & Tavan, E. (2022). Text-to-text transformer in authorship verification via stylistic and semantical analysis. *CLEF 2022 Labs and Workshops, Notebook Papers*. Italy: Bologna.

Plecháč, P. (2021). Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns. *Digital Scholarship in the Humanities, 36*(2), 430-438. https://doi.org/10.1093/llc/fqaa032.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn discourse treebank 2.0. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).* Marrakech, Morocco: EUropean Language Resources Association.

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

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, 5, 532-538 . Boston: Springer. https://doi.org/10.1007/978-0- 387-39940-9_565.

Sa'edi, Gh. (1998). *Ashoftehalan-e Bidarbakht*. Tehran: Negah. (In Persian)

Sa'edi, Gh. (2018). *Stranger in the town*. Tehran: Negah. (In Persian)

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

Segarra, S., Eisen, M., & Ribeiro, A. (2015). Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing, 63*(20), 5464-5478. https://doi.org/ 10.1109/TSP.2015.2451111.

Shamsfard, M., & Bijankhan, M., (2022). *Text and speech processing for the Persian language: The state of the art and a brief review of the theoretical foundations*. Tehran: Samt. (In Persian)

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., & Assi, S. M. (2010). Semi automatic development of farsnet: The persian wordnet. *In Proceedings of 5th global WordNet conference* (Vol. 29). India: Mumbai.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology, 60*(3), 538-556 . https://doi.org/10.1002/asi.21001

Teich, E. (1995). *A proposal for dependency in systemic functional grammar–metasemiosis in computational systemic functional linguistics*. Ph.D. Doctoral dissertation, University of the Saarland and GMD/IPSI, Darmstadt, Germany.

Uchendu, A., Le, T., Shu, K., & Lee, D. (2020). Authorship attribution for neural text generation. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing)* (pp. 8384-8395). (EMNLP 2020-2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference). Association for Computational Linguistics (ACL).

Weerasinghe, J., Singh, R., & Greenstadt, R. (2021). Feature vector difference based authorship verification for open-world settings. *In Proceedings of working of the evaluation Forum* (pp.2201-2207). Romania: Bucharest.