Journal of Psychological Science
Print ISSN: 1735-7462 Online ISSN: 2676-6639
Homepage: http://www.psychologicalscience.ir

Development and Psychometric Evaluation of the Persian Linguistic Inquiry and Word Count System (P-LIWC)

Mohammad Ali Soltani[1]✉, Peyman MamSharifi[2], Reza Salehi Chegeni[3], Ali Safari[4], Hamidreza Keshavarz[5], Daniel Rezaei Rouzbahani[6], Majid Afshari[7], Mohamad Sajad Ghafouri[8], Mojtaba Mohammadi[9], Fateme Tavakoli[10]

1. M.A. in Assessment and Measurement, Department of Measurement, Allameh Tabataba'i University, Tehran, Iran. E-mail: mohalisoltani@gmail.com
2. PhD of Psychology, Department of Psychology, Faculty of Psychology and Educational Sciences, Allameh Tabataba'i University, Tehran, Iran. E-mail: peymanmamsharifi@gmail.com
3. PhD student in Computer Engineering, Department of Computer Engineering, Faculty of Mathematics and Computer Science Amirkabir University of Technology, Tehran, Iran. E-mail: reza.salehi@aut.ac.ir
4. Assistant Professor, Department of Linguistics, Hazrat-e Masoumeh University, Qom, Iran. E-mail: alisfari228@gmail.com
5. PhD student of Political Sociology, Department of Political Sociology, CT.C., Islamic Azad University, Tehran, Iran. E-mail: hkeshaavarz@gmail.com
6. Master of Science in IT, Department Science in IT, Faculty of Computer Engineering, Yazd University, Yazd, Iran. E-mail: daniyal.rezaii@gmail.com
7. M.A. student of Psychology, Department of Psychology, SR.C., Islamic Azad University, Tehran, Iran. E-mail: majidafshari.sch@gmail.com
8. M.A. of Family Clinical Psychology - Family Research Institute, Shahid Beheshti University, Tehran, Iran. E-mail: msq.1373@gmail.com
9. Bachelor of English Language and Literature, University of Qom, Qom, Iran. E-mail: Mmohammadi161@gmail.com
10. M.A., Department of General Psychology, Imam Khomeini International University, Qazvin, Iran. E-mail: Tavakoli.fateme@hotmail.com

ARTICLE INFO

Article type:
Research Article

Article history:
Received 11 October 2025
Received in revised form 05 January 2026
Accepted 09 February 2026
Published Online 23 July 2026

Keywords:
Linguistic Inquiry and Word Count System,
natural language processing,
psychometrics,
text mining.

ABSTRACT

Background: The Linguistic Inquiry and Word Count (LIWC) is recognized as one of the most prominent tools for text analysis. By utilizing computational algorithms and linguistic processing techniques, it can identify and examine various elements within textual data, such as emotions, cognitive processes, attitudes, vocabulary, language styles, and patterns of social interaction. The absence of a localized Persian version of this tool served as one of the primary motivations for conducting the present study.

Aims: The aim of this study was to develop and examine the psychometric properties of the Persian Linguistic Inquiry and Word Count system (P-LIWC).

Methods: The present study employed a descriptive-correlational design. The process of developing the LIWC categories consisted of several stages. The development of the dictionary began with the initial translation of words from the original English version. Then, to maximize word coverage in each category in Persian, a large corpus of Persian texts was analyzed using text analysis methods. In the next step, the translated words were evaluated by psychologist judges, and for the approved words, their lemmas were determined by linguists to identify different word forms. After completing these stages, the reliability of the category and subcategory dictionaries was examined using Cronbach's alpha and the Kuder–Richardson 20 coefficients. To assess external validity, the equivalence between the Persian version (P-LIWC) and the original English LIWC-2022 was analyzed.

Results: The results showed that all categories of the Persian Linguistic Inquiry and Word Count system were significantly correlated with the final English version (LIWC-2022) ($p<0.01$), indicating that the instrument possesses good validity and reliability.

Conclusion: Based on the findings of this study, it can be concluded that the software has the potential to be applied across a wide range of research domains and in the analysis of Persian textual data. Furthermore, it can be effectively utilized for examining the content of therapy session transcripts from clients in psychological clinics.

Citation: Soltani, M.A., MamSharifi, P., Salehi Chegeni, R., Safari, A., Keshavarz, H., Rezaei Rouzbahani, D., Afshari, M., Ghafouri, M.S., Mohammadi, M., & Tavakoli, F. (2026). Development and psychometric evaluation of the Persian Linguistic Inquiry and Word Count system (P-LIWC). *Journal of Psychological Science*, 25(161), 1-26. 10.66224/jps.25.161.5

Journal of Psychological Science, Vol. 25, No. 161, 2026
 DOI: 10.66224/jps.25.161.5

✉ **Corresponding Author:** Mohammad Ali Soltani, M.A. in Assessment and Measurement, Allameh Tabataba'i University, Tehran, Iran. E-mail:. E-mail: mohalisoltani@gmail.com, Tel: (+98) 9190280944

Extended Abstract

Introduction

The Linguistic Inquiry and Word Count (LIWC) system, developed by Pennebaker et al. (2001), is a computational text-mining tool designed to analyze linguistic and psychological content in texts. It categorizes words into various linguistic and psychological dimensions, enabling researchers to infer cognitive, emotional, and social patterns from language use (Boyd et al., 2022; Koutsoumpis et al., 2022). LIWC has been widely utilized across multiple disciplines, including digital humanities, computational social sciences, psychology, and linguistic research. Despite its global recognition and adaptation into more than 14 languages such as German (Meier et al., 2019), Serbian (Bjekić et al., 2014), Spanish (Del Pilar Salas-Zárate et al., 2014), and Chinese (Zhao et al., 2016), no validated Persian version previously existed. The absence of a culturally adapted and psychometrically validated Persian LIWC prompted the current study to develop and evaluate the Persian Linguistic Inquiry and Word Count (P-LIWC) system. This study aimed to develop and psychometrically evaluate the Persian version of the Linguistic Inquiry and Word Count system (P-LIWC), focusing on all linguistic, cognitive, affective, and social categories. The objective was to ensure that P-LIWC achieves sufficient internal consistency, structural validity, and external equivalence with the original LIWC-22 English version. This study seeks to answer the research question: What are the psychometric properties of the Persian version of the Linguistic Inquiry and Word Count (LIWC)?

Method

Research design

The research followed a descriptive-correlational design. The P-LIWC dictionary development process consisted of five major stages based on the methodologies used in previous LIWC localization efforts (Pennebaker et al., 2015; Boyd et al., 2022). Step 1 involved initial translation and preliminary vocabulary collection from the LIWC-22 dictionary. Step 2 focused on review, scoring, categorization, and revision by expert psychologist judges. Step 3 covered dictionary expansion and refinement using Persian linguistic corpora and synonym analysis to maximize category coverage. Steps 4 and 5 involved psychometric evaluation and refinement of the dictionary based on reliability and validity analyses. Reliability was examined using Cronbach's alpha and Kuder–Richardson 20 (KR-20) coefficients, following the procedures outlined by Pennebaker et al. (2015). External validity was tested through Pearson correlation coefficients (r) between the Persian and English versions across multiple categories. A corpus of more than 100 million Persian-language texts (approximately 4.8 billion tokens) from social media platforms (Twitter/X, Instagram, Telegram) and formal news outlets (2019–2023) was used.

Results

The reliability analysis showed acceptable to high internal consistency across categories. Table 1 presents Cronbach's alpha and KR-20 coefficients for major linguistic, cognitive, and emotional categories.

Table1. Cronbach's alpha coefficient and Kuder-Richardson 20 for all classes of P-LIWC

Category	Words/ Entries in category	Internal Consistency: Cronbach's α	Internal Consistency: KR-20
Linguistic Dimensions			
Total function words	6040	0.97	0.97
Total pronouns	1869	0.95	0.97
Personal pronouns	407	0.78	0.83
1st person singular	309	0.65	0.70
1st person plural	80	0.57	0.86
2nd person	25	0.52	0.83
3rd person singular	89	0.43	0.78
3rd person plural	87	0.54	0.82
Impersonal pronouns	28	0.47	0.80
Determiners	99	0.69	0.73
Articles	392	0.65	0.76
Numbers	121	0.69	0.75
Prepositions	112	0.81	0.93
Auxiliary verbs	284	0.88	0.91
Adverbs	575	0.93	0.93
Conjunctions	49	0.68	0.83
Negations	231	0.83	0.95
Common verbs	1708	0.95	0.96
Common adjectives	2598	0.95	0.97
Quantities	340	0.78	0.82
Psychological Processes			
Drives			
Drives	1768	0.88	0.91
Affiliation	589	0.73	0.80
Achievement	302	0.66	0.73
Power	942	0.78	0.85
Cognition			
Cognition	2269	0.96	0.96
All-or-none	84	0.65	0.85
Cognitive processes	2160	0.95	0.96
Insight	766	0.85	0.88
Causation	427	0.85	0.88
Discrepancy	207	0.61	0.68
Tentative	371	0.78	0.81
Certitude	203	0.69	0.70
Differentiation	454	0.86	0.88
Memory	39	0.34	0.74

Category	Words/ Entries in category	Internal Consistency: Cronbach's α	Internal Consistency: KR-20
Affect			
Affect	5798	0.92	0.95
Positive tone	2207	0.88	0.92
Negative tone	3190	0.83	0.89
Emotion	1632	0.71	0.79
Positive emotion	469	0.64	0.81
Negative emotion	1117	0.63	0.79
Anxiety	216	0.44	0.75
Anger	334	0.65	0.89
Sadness	239	0.67	0.91
Swear words	442	0.63	0.82
Social processes			
Social processes	3467	0.90	0.93
Social behavior	2920	0.89	0.92
Prosocial behavior	531	0.68	0.74
Politeness	261	0.59	0.78
Interpersonal conflict	700	0.68	0.75
Moralization	1116	0.70	0.78
Communication	681	0.80	0.82
Social referents	621	0.67	0.76
Family	244	0.56	0.68
Friends	137	0.43	0.77
Female references	221	0.47	0.79
Male references	155	0.64	0.83
Culture			
Culture	982	0.83	0.87
Politics	528	0.83	0.86
Ethnicity	245	0.51	0.57
Technology	211	0.53	0.85
Lifestyle			
Lifestyle	2419	0.86	0.91
Leisure	728	0.60	0.86
Home	241	0.43	0.77
Work	607	0.82	0.86
Money	351	0.74	0.82
Religion	543	0.77	0.83
Physical			
Physical	1623	0.72	0.84
Health	555	0.71	0.81
Illness	292	0.58	0.78
Wellness	131	0.43	0.72
Mental health	135	0.41	0.67
Substances	181	0.40	0.65

Category	Words/ Entries in category	Internal Consistency: Cronbach's α	Internal Consistency: KR-20
Sexual	380	0.51	0.74
Food	380	0.67	0.79
Death	189	0.49	0.81
States			
Need	129	0.50	0.79
Want	151	0.38	0.73
Acquire	179	0.43	0.81
Lack	238	0.57	0.65
Fulfilled	137	0.67	0.69
Fatigue	184	0.42	0.75
Motives			
Reward	82	0.45	0.76
Risk	189	0.42	0.69
Curiosity	163	0.39	0.73
Allure	209	0.77	0.82
Perception			
Perception	2536	0.83	0.94
Attention	165	0.41	0.78
Motion	769	0.72	0.93
Space	831	0.72	0.90
Visual	357	0.65	0.76
Auditory	316	0.47	0.71
Feeling	169	0.52	0.82
Time orientation			
Time	759	0.87	0.89
Past focus	705	0.81	0.80
Present focus	367	0.58	0.84
Future focus	186	0.58	0.79
Conversational			
Conversational	200	0.69	0.89
Netspeak	123	0.65	0.82
Assent	19	0.48	0.75
Nonfluencies	15	0.37	0.66
Fillers	50	0.25	0.62
Emoji / Emoticon	48	0.12	0.66
Punctuation marks	92	0.66	0.70

The internal consistency of categories ranged from 0.25 to 0.73 for Cronbach's alpha and from 0.62 to 0.97 for KR-20, indicating that the dictionary categories were stable across contexts. External validity analysis showed that all P-LIWC categories significantly correlated ($p < 0.01$) with their English LIWC-22 counterparts (see Table 2).

Table2. Correlation coefficient between Persian and English version

Category	Pearson's (r)	Significance
Linguistic Dimensions		
Total function words	0.73	p < 0.01
Total pronouns	0.71	p < 0.01
Personal pronouns	0.71	p < 0.01
1st person singular	0.70	p < 0.01
1st person plural	0.83	p < 0.01
2nd person	0.68	p < 0.01
3rd person singular	0.67	p < 0.01
3rd person plural	0.86	p < 0.01
Impersonal pronouns	0.62	p < 0.01
Determiners	0.55	p < 0.01
Articles	0.12	p < 0.01
Numbers	0.74	p < 0.01
Prepositions	0.61	p < 0.01
Auxiliary verbs	0.66	p < 0.01
Adverbs	0.59	p < 0.01
Conjunctions	0.50	p < 0.01
Negations	0.68	p < 0.01
Common verbs	0.74	p < 0.01
Common adjectives	0.67	p < 0.01
Quantities	0.59	p < 0.01
Psychological Processes		
Drives		
Drives	0.81	p < 0.01
Affiliation	0.70	p < 0.01
Achievement	0.86	p < 0.01
Power	0.85	p < 0.01
Cognition		p < 0.01
Cognition	0.84	p < 0.01
All-or-none	0.66	p < 0.01
Cognitive processes	0.85	p < 0.01
Insight	0.86	p < 0.01
Causation	0.67	p < 0.01
Discrepancy	0.54	p < 0.01
Tentative	0.87	p < 0.01
Certitude	0.61	p < 0.01
Differentiation	0.75	p < 0.01
Memory	0.63	p < 0.01
Affect		p < 0.01
Affect	0.76	p < 0.01
Positive tone	0.80	p < 0.01
Negative tone	0.82	p < 0.01

Category	Pearson's (r)	Significance
Emotion	0.74	p < 0.01
Positive emotion	0.82	p < 0.01
Negative emotion	0.86	p < 0.01
Anxiety	0.96	p < 0.01
Anger	0.82	p < 0.01
Sadness	0.78	p < 0.01
Swear words	0.51	p < 0.01
Social processes		
Social processes	0.67	p < 0.01
Social behavior	0.62	p < 0.01
Prosocial behavior	0.52	p < 0.01
Politeness	0.54	p < 0.01
Interpersonal conflict	0.74	p < 0.01
Moralization	0.61	p < 0.01
Communication	0.72	p < 0.01
Social referents	0.75	p < 0.01
Family	0.78	p < 0.01
Friends	0.71	p < 0.01
Female references	0.84	p < 0.01
Male references	0.81	p < 0.01
Culture		
Culture	0.83	p < 0.01
Politics	0.82	p < 0.01
Ethnicity	0.81	p < 0.01
Technology	0.83	p < 0.01
Lifestyle		
Lifestyle	0.90	p < 0.01
Leisure	0.92	p < 0.01
Home	0.61	p < 0.01
Work	0.94	p < 0.01
Money	0.88	p < 0.01
Religion	0.95	p < 0.01
Physical		
Physical	0.73	p < 0.01
Health	0.84	p < 0.01
Illness	0.83	p < 0.01
Wellness	0.63	p < 0.01
Mental health	0.82	p < 0.01
Substances	0.63	p < 0.01
Sexual	0.78	p < 0.01
Food	0.83	p < 0.01
Death	0.91	p < 0.01
States		

Category	Pearson's (r)	Significance
Need	0.81	p < 0.01
Want	0.52	p < 0.01
Acquire	0.61	p < 0.01
Lack	0.64	p < 0.01
Fulfilled	0.58	p < 0.01
Fatigue	0.54	p < 0.01
Motives		
Reward	0.75	p < 0.01
Risk	0.81	p < 0.01
Curiosity	0.53	p < 0.01
Allure	0.61	p < 0.01
Perception		
Perception	0.60	p < 0.01
Attention	0.70	p < 0.01
Motion	0.64	p < 0.01
Space	0.74	p < 0.01
Visual	0.53	p < 0.01
Auditory	0.52	p < 0.01
Feeling	0.85	p < 0.01
Time orientation		
Time	0.79	p < 0.01
Past focus	0.77	p < 0.01
Present focus	0.66	p < 0.01
Future focus	0.57	p < 0.01
Conversational		
Conversational	0.35	p < 0.01
Netspeak	0.25	p < 0.01
Assent	0.37	p < 0.01
Nonfluencies	0.16	p < 0.01
Fillers	0.18	p < 0.01
Emoji / Emoticon	0.73	p < 0.01
Punctuation marks	0.70	p < 0.01

All correlations were statistically significant, confirming strong equivalence between the Persian and English LIWC-22 categories. These findings support the cross-linguistic validity of the P-LIWC.

Conclusion

The findings demonstrated that the Persian LIWC (P-LIWC) exhibits strong psychometric properties. Reliability indices indicate that the internal consistency of categories such as emotions and cognitive processes are comparable to other validated versions (e.g., Meier et al., 2019; Zhao et al., 2016). High correlations with the English version confirm that P-LIWC effectively captures linguistic and

psychological dimensions across Persian-language texts.

This study provides a validated computational tool for Persian-speaking researchers to analyze linguistic data in psychology, social sciences, and digital humanities. The system can be used to examine therapy session transcripts, social media communication, and written narratives, offering insights into emotional tone, cognitive processes, and social behavior.

In conclusion, the P-LIWC bridges a significant methodological gap, allowing Persian-language research to join global linguistic and psychological text analyses with a culturally and linguistically adapted tool.

Ethical Considerations

Compliance with ethical guidelines: Following the principles of research ethics: This article is taken from an independent research in the field of psychology. Since there are no participants in this research, the only ethical consideration has been related to accuracy in judging words without bias.

Funding: This research is in the form of an independent research that has no financial sponsor.

Authors' contribution: The first author is the principal investigator and corresponding author of this study. The second author is the head of the Psychology Team; the third author is the head of the Artificial Intelligence Team; and the fourth author is the head of the Linguistics Team. The fifth author was the Sociology Researcher, and the sixth author served as the Chief Technology Officer (CTO). The seventh, eighth, and tenth authors were researchers on the Psychology Team, and the ninth author was the research translator.

Conflict of interest: The authors declare no conflict of interest for this study.

Acknowledgments: My gratitude goes out to God Almighty and my dedicated colleagues for their invaluable assistance in successfully fulfilling this research in alignment with the national scientific principles and objectives.

مجله علوم روانشناختی

شاپا چاپی: ۷۴۶۲-۱۷۳۵ شاپا الکترونیکی: ۶۶۳۹-۲۶۷۶

Homepage: http://www.psychologicalscience.ir

توسعه و بررسی ویژگی‌های روان‌سنجی نسخه فارسی سامانه واکاوی زبانی و شمارش واژگان (P-LIWC)

محمد علی سلطانی[۱]✉، پیمان مام شریفی[۲]، رضا صالحی چگنی[۳]، علی صفری[۴]، حمیدرضا کشاورز[۵]، دانیال رضایی روزبهانی[۶]، مجید افشاری[۷]، محمدسجاد غفوری[۸]، مجتبی محمدی[۹]، فاطمه توکلی[۱۰]

۱. کارشناس ارشد، گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: mohalisoltani@gmail.com

۲. دکتری روانشناسی، گروه روانشناسی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه:peymanmamsharifi@gmail.com

۳. دکتری مهندسی کامپیوتر، گروه علوم کامپیوتر، دانشکده علوم ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران. رایانامه:reza.salehi@aut.ac.ir

۴. استادیار، گروه زبان‌شناسی، دانشگاه حضرت معصومه (س)، قم، ایران. رایانامه:alisfari228@gmail.com

۵. دانشجوی دکتری، گروه جامعه شناسی سیاسی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران. رایانامه:hkeshaavarz@gmail.com

۶. کارشناسی ارشد، گروه فناوری اطلاعات، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران. رایانامه:daniyal.rezaii@gmail.com

۷. دانشجوی کارشناسی ارشد،گروه روانشناسی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران. رایانامه:majidafshari.sch@gmail.com

۸. کارشناسی ارشد روانشناسی بالینی خانواده، پژوهشکده خانواده، دانشگاه شهید بهشتی، تهران، ایران. رایانامه:msq.1373@gmail.com

۹. کارشناسی زبان و ادبیات انگلیسی، دانشگاه قم، قم، ایران. رایانامه:Mmohammadi161@gmail.com

۱۰. کارشناس ارشد، گروه روانشناسی عمومی، دانشگاه بین‌المللی امام خمینی (ره)، قزوین، ایران. رایانامه:Tavakoli.fateme@hotmail.com

مشخصات مقاله

نوع مقاله:
پژوهشی

تاریخچه مقاله:
دریافت: ۱۴۰۴/۰۷/۱۹
بازنگری: ۱۴۰۴/۱۰/۱۵
پذیرش: ۱۴۰۴/۱۱/۲۰
انتشار برخط: ۱۴۰۵/۰۵/۰۱

کلیدواژه‌ها:
سامانه واکاوی زبانی و شمارش واژگان، پردازش زبان طبیعی، روان¬سنجی، متن‌کاوی.

چکیده

زمینه: ابزار واکاوی زبانی و شمارش واژگان یکی از شناخته‌شده‌ترین روش‌های تحلیل متون به شمار می‌رود. این ابزار با بهره‌گیری از الگوریتم‌ها و فرآیندهای رایانشی، قادر است جنبه‌هایی مانند احساسات، فرآیندهای شناختی، نگرش‌ها، واژگان، سبک‌های زبانی و الگوهای تعامل اجتماعی را در داده‌های متنی شناسایی و بررسی کند. فقدان نسخه‌ای بومی‌سازی‌شده به زبان فارسی از این ابزار، یکی از انگیزه‌های اصلی اجرای پژوهش حاضر بود.

هدف: هدف از انجام این پژوهش توسعه و بررسی ویژگی‌های روان‌سنجی نسخه فارسی سامانه واکاوی زبانی و شمارش واژگان بود.

روش: روش پژوهش حاضر توصیفی و همبستگی بود. روش توسعه طبقات سامانه واکاوی زبانی و شمارش واژگان از چند مرحله تشکیل شده است. توسعه فرهنگ لغت با ترجمه اولیه کلمات نسخه اصلی انگلیسی شروع شد و پس از آن برای به حداکثر رساندن پوشش دهی کلمات هر طبقه در زبان فارسی با استفاده از روش‌های تحلیل متن، پیکره عظیمی از متون فارسی مورد تجزیه و تحلیل قرار گرفت. در گام بعدی کلمات توسط داوران روانشناس مورد ارزیابی قرار گرفت و سپس برای کلمات تأیید شده لما آن‌ها توسط زبان‌شناسان به منظور تشخیص اشکال مختلف کلمه تعیین شد. بعد از اتمام این مراحل به منظور بررسی پایایی فرهنگ لغت‌های طبقات و زیر طبقات آن‌ها آلفای کرونباخ و کودر-ریچاردسون ۲۰ محاسبه شد و سپس برای بررسی روایی بیرونی، هم ارزی نسخه فارسی واکاوی زبانی و شمارش واژگان (P-LIWC) با نسخه اصلی انگلیسی LIWC-2022 مورد تجزیه و تحلیل قرار گرفت.

یافته‌ها: نتایج نشان داد همه طبقات سامانه واکاوی زبانی و شمارش واژگان در نسخه فارسی با نسخه نهایی انگلیسی LIWC-2022 دارای همبستگی معنادار بودند ($p<۰/۰۱$) و این ابزار از روایی و پایایی مناسبی برخوردار است.

نتیجه‌گیری: بر اساس نتایج به‌دست‌آمده از این پژوهش، می‌توان اظهار کرد که ابزار واکاوی زبانی و شمارش واژگان، قابلیت استفاده در حوزه‌های متنوع پژوهشی و تحلیل متون فارسی را داراست. افزون بر این، این ابزار می‌تواند در تحلیل محتوای گفت‌وگوها و متون جلسات درمانی مراجعان کلینیک‌های روان‌شناسی نیز مورد بهره‌برداری قرار گیرد.

استناد: سلطانی، محمد علی؛ مام شریفی، پیمان؛ صالحی چگنی، رضا؛ صفری، علی؛ کشاورز، حمیدرضا؛ رضایی روزبهانی، دانیال؛ افشاری، مجید؛ غفوری، محمدسجاد؛ محمدی، مجتبی؛ و توکلی، فاطمه (۱۴۰۵). توسعه و بررسی ویژگی‌های روان‌سنجی نسخه فارسی سامانه واکاوی زبانی و شمارش واژگان (P-LIWC). *مجله علوم روانشناختی*، دوره ۲۵، شماره ۱۶۱، ۱ -۲۶.

مجله علوم روانشناختی، دوره ۲۵، شماره ۱۶۱، ۱۴۰۵. DOI: 10.66224/jps.25.161.5

✉ **نویسنده مسئول:** محمد علی سلطانی، کارشناس ارشد سنجش و اندازه‌گیری، دانشکده روان¬شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: mohalisoltani@gmail.com

مقدمه

سامانه واکاوی زبانی و شمارش واژگان[1] یک ابزار متن کاوی محاسباتی است که برای تحلیل متون و محتوای زبانی استفاده می‌شود. این ابزار اصطلاحاً به‌عنوان یک نمایشگر هیجانات و واژگان در متون شناخته می‌شود (سلطانی و همکاران، ۱۴۰۳). به‌طور خاص، به کمک الگوریتم‌ها و فرآیندهای محاسباتی، می‌تواند عناصری مانند هیجانات، نگرش‌ها، واژگان، سبک زبانی و ارتباطات اجتماعی را در متون تحلیل کند (کوتسومپیس و همکاران، ۲۰۲۲). این ابزار بر اساس پژوهش‌های روان‌شناختی و زبان‌شناسی ساخته شده و به‌عنوان یک ابزار مفید برای تحلیل و فهم بهتر متون و نوشتارها در علوم انسانی دیجیتال[2]، علوم اجتماعی محاسباتی[3]؛ افکار سنجی[4] و حوزه‌های دیگر به کار می‌رود (پارک و همکاران، ۲۰۲۲). واکاوی زبانی و شمارش واژگان با دسته‌بندی کلمات به ابعاد روان‌شناختی و زبانی متعددی مانند هیجانات، فرآیندهای شناختی، تمایلات اجتماعی و حتی بخش‌هایی از گفتار عمل می‌کند (مام‌شریفی و همکاران، ۱۴۰۵). این ابعاد به درک لحن روان‌شناختی و عاطفی متن کمک می‌کند و بینش‌هایی را در مورد وضعیت ذهنی، شخصیت و سبک ارتباطی نویسنده ارائه می‌دهد (پنه‌بیکر و همکاران، ۲۰۰۱). کارکردهای آن شامل ارائه تجزیه‌وتحلیل‌های آماری، تولید گزارش‌ها و بازنمایی‌های بصری الگوهای زبانی و عبارات احساسی در متن است که به محققان درک جامعی از جنبه‌های احساسی و شناختی ارتباط را ارائه می‌دهد (لین و همکاران، ۲۰۲۱).

ابزار واکاوی زبانی و شمارش واژگان با دسته‌بندی واژگان در ابعاد مختلف، بینش‌های روان‌شناختی و رفتاری ارزشمندی ارائه می‌دهد. این ابعاد شامل ابعاد زبان‌شناختی[5] (کل واژگان دستوری، کل ضمایر، حروف تعریف، اعداد، حروف اضافه، افعال کمکی، قیود، حروف ربط، واژگان نفی، افعال رایج، صفات رایج و واژگان کمیت)، سائق‌ها/کشش‌های درونی[6] (پیوندجویی/تعلق اجتماعی[7]، پیشرفت[8] و قدرت[9])، فرایندهای شناختی[10] (تفکر مطلق‌نگر/ همه یا هیچ، بینش/ درک[11]، علیت/استدلال[12]، ناهماهنگی/ناسازگاری[13]، تردید/ابهام‌گویی[14]، قطعیت[15]، تمایزگذاری[16] و حافظه)، هیجانات[17] (لحن مثبت و منفی[18]، هیجان مثبت[19]، اضطراب[20]، خشم[21]، غم[22] و دشنام[23])، فرایندهای اجتماعی[24] (رفتار اجتماعی، رفتار جامعه‌پسند/یاریگرانه[25]، ادب/مؤدبانه‌گویی[26]، تعارض بین‌فردی[27]، اخلاق‌گرایی[28]، ارتباط[29]، ارجاعات اجتماعی[30]، خانواده، دوستان، ارجاع به زن، ارجاع به مرد)، فرهنگ[31] (فرهنگ، سیاست[32]، قومیت[33]، فناوری)، سبک زندگی[34] (فراغت[35]، خانه، کار، پول، دین/مذهب)، بدنی/جسمانی[36] (سلامت، بیماری، تندرستی، سلامت روان، مواد مخدر، جنسی[37]، غذا، مرگ) حالات[38] (نیاز[39]، خواست[40]، به‌دست‌آوردن[41]، فقدان[42]، برآورده‌سازی[43]، خستگی[44])، انگیزش‌ها[45] (پاداش، ریسک/خطرپذیری،

[1] Linguistic Inquiry and Word Count (LIWC)
[2] digital humanities
[3] computational social sciences (CSS)
[4] polling
[5] Linguistic Dimensions
[6] Drives
[7] Affiliation
[8] Achievement
[9] Power
[10] Cognitive processes
[11] insight
[12] causal
[13] discrepancies
[14] tentativeness
[15] certainty
[16] differentiation
[17] affect
[18] Positive & Negative tone
[19] positive emotions
[20] anxiety
[21] anger
[22] sadness
[23] Swear words
[24] Social processes
[25] Prosocial behavior
[26] Politeness
[27] Interpersonal conflict
[28] Moralization
[29] Communication
[30] Social referents
[31] Culture
[32] Politics
[33] Ethnicity
[34] Lifestyle
[35] Leisure
[36] Physical
[37] Sexual
[38] States
[39] Need
[40] Want
[41] Acquire
[42] Lack
[43] Fulfilled
[44] Fatigue
[45] Motives

کنجکاوی[1]، جاذبه/کشش[2]، ادراک[3] (ادراک، توجه[4]، حرکت[5]، فضا[6]، دیداری، شنیداری، حسی/لامسه‌ای)، جهت‌گیری زمان[7] (زمان، تمرکز بر گذشته، تمرکز بر حال، تمرکز بر آینده)، محاوره‌ای[8] (محاوره‌ای، زبان اینترنتی[9]، تأیید/موافقت[10]، ناپیوستگی‌های گفتاری[11]، کلمات پرکننده[12]، ایموجی/شکلک، علائم نگارشی[13]) را در برمی‌گیرد و امکان تجزیه‌وتحلیل دقیق محتوای متن را فراهم می‌کند. پژوهشگران از واکاوی زبانی و شمارش واژگان برای کشف الگوهای زبانی در محتوای رسانه‌های اجتماعی، تجزیه‌وتحلیل زبان در محیط‌های بالینی، مطالعه تفاوت‌های جنسیتی، فرهنگ‌ها یا گروه‌های سنی و حتی پیش‌بینی ویژگی‌های روان‌شناختی بر اساس زبان نوشتاری استفاده می‌کنند (اوتوموا و کاریاواتیا، ۲۰۲۱). این نرم‌افزار به‌طور مداوم در حال پیشرفت است و فرهنگ لغت‌ها و ویژگی‌های جدیدی را برای افزایش دقت و کاربرد آن در زمینه‌های مختلف ترکیب می‌کند و آن را به ابزاری همه‌کاره برای تجزیه‌وتحلیل متن و تحقیقات روان‌شناختی تبدیل می‌کند (بوید و همکاران، ۲۰۲۲).

واکاوی زبانی و شمارش واژگان دارای عملکردهای مختلفی است که به تعدادی از آن‌ها اشاره می‌شود: طبقه‌بندی کلمات[14]: این نرم‌افزار قادر است کلمات را در ابعاد مختلف روان‌شناختی و زبانی مانند هیجانات، فرآیندهای شناختی، فرایندهای اجتماعی و سبک‌های زبانی دسته‌بندی نماید. به‌عنوان مثال، می‌تواند کلمات را در گروه‌هایی مانند هیجانات مثبت یا منفی، مکانیسم‌های شناختی، تمایلات اجتماعی و سایر گروه‌ها طبقه‌بندی کند (باهگات و همکاران، ۲۰۲۲). بینش‌های روان‌شناختی[15]: با تحلیل فراوانی دسته‌بندی‌های مشخص کلمات در متن، این ابزار توانایی ارائه بینش‌هایی درباره وضعیت روان‌شناختی نویسنده را دارد. این امر می‌تواند تمایلات عاطفی، الگوهای شناختی و رفتارهای اجتماعی فرد را بر اساس زبان مورد استفاده آشکار سازد (لیو و همکاران، ۲۰۲۳). تحلیل رفتاری[16]: این نرم‌افزار مچنین قابلیت پیش‌بینی و درک الگوهای رفتاری را بر اساس داده‌های متنی فراهم می‌کند. برای نمونه، می‌تواند ویژگی‌های رفتاری مانند جهت‌گیری اجتماعی، نگرانی‌های شخصی، سبک‌های تفکر و موارد مشابه را ارزیابی نماید (کیلیک و پن، ۲۰۲۲). تحقیقات و مطالعات[17]: محققان در زمینه‌های مختلف از جمله روان‌شناسی، زبان‌شناسی، جامعه‌شناسی، ارتباطات، علوم سیاسی و به‌طور کلی علوم انسانی دیجیتال و علوم اجتماعی محاسباتی و همچنین در پردازش زبان طبیعی[18]، از این نرم‌افزار برای تجزیه‌وتحلیل و تفسیر داده‌های متنی برای مطالعات دانشگاهی، نظرسنجی‌ها و آزمایش‌ها استفاده می‌کنند (پنه‌بیکر و همکاران، ۲۰۱۵). کاربردهای درمانی[19]: این نرم‌افزار در محیط‌های بالینی نیز کاربرد دارد. درمانگران و مشاوران می‌توانند از آن برای تحلیل و ارزیابی محتوای عاطفی ارتباط نوشتاری یا شفاهی به‌عنوان بخشی از فرآیند درمانی استفاده کنند (دویر و همکاران، ۲۰۲۱). تجزیه‌وتحلیل محتوا[20]: این مورد را با خودکار کردن فرآیند طبقه‌بندی و کمی کردن جنبه‌های زبانی و روان‌شناختی متن تسهیل می‌کند که می‌تواند برای تولیدکنندگان محتوا، بازاریابان و دانشمندان علوم اجتماعی مفید باشد (اندی و اندی، ۲۰۲۱).

این نرم‌افزار در ۱۴ زبان بین‌المللی از جمله آلمانی (میر و همکاران، ۲۰۱۹)، صربستان (بژیکیچ و همکاران، ۲۰۱۴)، اسپانیایی (دل پیلار سالاز زاراته و همکاران، ۲۰۱۴) و چینی (ژائو و همکاران، ۲۰۱۶) توسعه پیدا کرده است. به‌طور کلی، واکاوی زبانی و شمارش واژگان به‌عنوان ابزاری برای تعیین کمیت و تحلیل مؤلفه‌های روان‌شناختی و زبانی متن نوشتاری عمل می‌کند و بینش‌های ارزشمندی درباره طیف عظیمی از الگوهای زبانی و روانی افراد یا گروه‌ها ارائه می‌دهد. در پژوهش حاضر، کلیه طبقات تحلیلی این ابزار مورد بررسی قرار گرفتند. از آنجا که نسخه فارسی واکاوی زبانی و شمارش واژگان در ایران موجود نیست و کاربرد آن در حوزه‌های مختلف، از جمله روان‌شناسی، می‌تواند مفید و ثمربخش باشد، این پژوهش به بررسی این

[1] Curiosity
[2] Allure
[3] Perception
[4] Attention
[5] Motion
[6] Space
[7] Time orientation
[8] Conversational
[9] Netspeak
[10] Assent
[11] Nonfluencies
[12] Fillers
[13] Punctuation Marks
[14] classification of words
[15] psychological insights
[16] behavioral analysis
[17] research and studies
[18] natural language processing (NLP)
[19] therapeutic applications
[20] content analysis

سوال پرداخت که نسخه فارسی واکاوی زبانی و شمارش واژگان، دارای چه ویژگی‌های روان‌سنجی است؟

روش

طرح پژوهش

مطالعه حاضر از نوع توصیفی و همبستگی بود. توسعه تمامی طبقات نسخه فارسی واکاوی زبانی و شمارش واژگان از چند مرحله تشکیل شده است. توسعه فرهنگ لغت با ترجمه اولیه کلمات نسخه اصلی انگلیسی شروع شد و پس از آن برای به حداکثر رساندن پوشش دهی کلمات هر طبقه در زبان فارسی با استفاده از روش‌های متن‌کاوی[1]، پیکره عظیمی از متون فارسی مورد تجزیه‌وتحلیل قرار گرفت. در گام بعدی کلمات توسط داوران روانشناس مورد ارزیابی قرار گرفت و سپس برای کلمات تأیید شده لما[2] آن‌ها توسط زبان‌شناسان به منظور تشخیص اشکال مختلف کلمه تعیین شد. بعد از اتمام این مراحل به منظور بررسی پایایی فرهنگ تمامی طبقات نسخه فارسی واکاوی زبانی و شمارش واژگان و زیر طبقات آن‌ها آلفای کرونباخ و کودر-ریچاردسون ۲۰ محاسبه گردید و سپس برای بررسی روایی بیرونی، هم ارزی نسخه فارسی واکاوی زبانی و شمارش واژگان (P-LIWC) با نسخه اصلی انگلیسی LIWC-22 مورد تجزیه وتحلیل قرار گرفت.

در بخش بعدی فرایند توسعه فرهنگ لغت انجام شد. این قسمت یک نمای کلی از روند توسعه طبقات فرهنگ لغت P-LIWC ارائه داده می‌شود. با اینکه این فرایند دارای مراحل بسیاری بود و تا حدی ماهیت بازگشتی داشت، می‌توان فرایند کلی را به ۵ مرحله تقسیم کرد. این مراحل بر اساس روش‌های بکار گرفته شده در بومی‌سازی نسخه‌های رسمی غیر انگلیسی زبان LIWC و روش استفاده شده در ساخت LIWC-22 طراحی شده است. مرحله اول: ترجمه اولیه و جمع‌آوری مقدماتی واژگان؛ مرحله دوم: بازبینی، امتیازدهی، مقوله‌بندی و اصلاح؛ مرحله سوم: گسترش و توسعه فرهنگ لفت؛ مرحله چهارم: ارزیابی ویژگی‌های روان‌سنجی و مرحله پنجم: پالایش و تجدید نظر

مراحل اول تا سوم مربوط به بخش روش و مراحل چهار و پنج مربوط به بخش یافته‌های پژوهش هستند. این بخش در سه مرحله انجام شد که در ادامه به صورت کامل در مورد هر کدام توضیح داده می‌شود.

مرحله اول: ترجمه اولیه و جمع‌آوری مقدماتی واژگان

هدف از انجام این مرحله رسیدن به پایه و اساسی مطلوب برای گسترش واژگان تمامی طبقات نسخه فارسی واکاوی زبانی و شمارش واژگان بوده است. روش انجام در این مرحله بر اساس روش‌های استفاده شده در ترجمه و بومی‌سازی نسخه‌های رسمی غیرانگیسی زبان دیکشنری LIWC-22، LIWC2015 (مایر و همکاران، ۲۰۱۸) و نسخه‌ی قبلی آن یعنی LIWC2001 (بژکیچ و همکاران، ۲۰۱۴؛ پیولا و همکاران، ۲۰۱۱) و روش استفاده شده در ساخت LIWC-22 (بوید و همکاران، ۲۰۲۲) طراحی شده است. همچنین با توجه به وجود تفاوت‌های ساختاری قابل توجه میان زبان‌های فارسی و انگلیسی، در جهت افزایش پوشش دهی و دقت دیکشنری، از روش‌های دیگری نیز در جمع‌آوری واژگان استفاده گردید. در این مرحله، ابتدا تمامی واژگان متعلق به طبقات و زیر طبقات نسخه فارسی واکاوی زبانی و شمارش واژگان از دیکشنری LIWC-22 توسط مترجمان زبان انگلیسی ترجمه شد. با استناد به روش استفاده شده در LIWC-22 (بوید و همکاران، ۲۰۲۲)، برای افزودن واژه‌ها به دیکشنری هیجانات و با توجه به ماهیت دیکشنری مورد نظر که شامل واژگان با بار هیجانی خواهد بود، تعداد قابل توجهی از مقیاس‌های مرسوم و معتبر سنجش متغیرهای روانی نیز مورد بررسی قرار گرفتند و واژگان استخراج شده از آن‌ها به واژگان به دست آمده از واژه‌نامه‌های فارسی و LIWC-22 افزوده شدند. این آزمون‌ها که توسط پژوهشگران متخصص و با توجه به ادبیات پژوهشی مرتبط انتخاب شدند، مجموعه‌ای از مقیاس‌های سنجش هیجانات، عواطف، نشانه‌های اضطراب، اختلالات اضطرابی و اختلالات وسواس، افسردگی، استرس، نشانه‌های تروما و اختلال پس از ضربه، قلدری، پرخاشگری و خشم، شادکامی و رضایت، نشانه‌های مانیک و اختلال مانیا را دربر می‌گیرند. سپس از محتوای آزمون‌ها فهرستی از واژگانی که از نظر مفهومی با دیکشنری نهایی مرتبط هستند تهیه شد که به واژگان قبلی اضافه گردیدند. این تصمیم به منظور حفظ روشمندی و وفاداری به روش‌شناسی LIWC-22 (بوید و همکاران، ۲۰۲۲) و با تکیه

[1] text mining

[2] lemma

بر اقدامات مشابه در ساخت دیکشنری‌های مشابه (جی و رینی، ۲۰۲۰) گرفته شده است.

پس از ترجمه کلمات تمامی طبقات نسخه واکاوی زبانی و شمارش واژگان، هر کلمه توسط ۲ داور روانشناس مورد ارزیابی قرار گرفت. هدف از ارزیابی اولیه کلمات، بررسی تناسب کلمه ترجمه شده با هریک از طبقات از لحاظ مفهومی و فرهنگی با زبان فارسی بود. برای انجام این کار به هریک از واژه‌ها نمره‌ای بین ۱ تا ۵ داده شد که نمره ۵ نشان دهنده بیشترین میزان تناسب و نمره ۱ نشان دهنده کمترین میزان تناسب کلمه با طبقه مورد نظر بود. پس از اتمام امتیازدهی و مقایسه ارزیابی داوران، کلمات دارای بیشترین میزان تناسب حفظ شدند و کلماتی که ترجمه آن‌ها دارای تناسب کمی با طبقه مورد نظر بود، ترجمه آن‌ها توسط کارشناس زبان انگلیسی به معادلی متناسب با طبقه مذکور تغییر یافت. بعد از آن، مجدداً معادل‌های جدید کلمات مورد ارزیابی و داوری کارشناسان روانشناسی و زبان‌شناسی قرار گرفت. این مراحل تا تأیید تناسب همه کلمات ترجمه شده با طبقات مورد نظر تکرار شد.

مرحله دوم: بازبینی، امتیازدهی، مقوله‌بندی و اصلاح

در این مرحله نیز مانند مرحله‌ی قبل از روش‌های استفاده شده در تألیف، ترجمه و بومی‌سازی LIWC استفاده شد. در مواردی که میان جزییات روش‌های طبقه‌بندی در منابع موجود ناهمخوانی دیده شد، از روش استفاده شده در جدیدترین نسخه اصلی (انگلیسی زبان) یعنی LIWC-22 (بوید و همکاران، ۲۰۲۲) استفاده گردید. همچنین روش‌های برگرفته از این منابع با روش‌های طبقه‌بندی واژگان مرتبط با هیجان در زبان فارسی نیز مطابقت داده شدند (کاویانی و همکاران، ۱۳۸۴ و ۱۳۸۶).

پس از آماده شدن نسخه اولیه واژه‌نامه طبقات نسخه فارسی واکاوی زبانی و شمارش واژگان، این نسخه به صورت دستی بازبینی شد تا ناهم‌خوانی‌های مفهومی آشکار یا اشتباهات املایی احتمالی شناسایی و حذف شوند. پس از آن، واژه‌ها به‌طور دستی تحت طبقات مشخص شده طبقه‌بندی شدند. ۴ داور متخصص روانشناس دارای حداقل مدرک کارشناسی ارشد برای داوری انتخاب شدند و به آن‌ها نحوه داوری کلمات آموزش داده شد. سپس همه‌ی واژه‌های به دست آمده را یکی یکی بررسی کردند و به میزان تناسب (همخوانی) هر واژه با هر یک از طبقه‌های مورد نظر، به آن واژه یک امتیاز دادند؛ بنابراین در این مرحله هر یک از واژه‌ها ۴ امتیاز جداگانه دریافت کردند. هنگام بررسی تناسب واژه با طبقه مورد نظر، معانی تحت‌اللفظی و استعاری واژه هر دو مورد توجه قرار گرفتند. علاوه بر این، در جهت به حداقل رساندن خطا، واژه‌های هم‌نگاره حذف شدند (بژکیچ و همکاران، ۲۰۱۴؛ پیولا و همکاران، ۲۰۱۱). در صورتی که یکی از واژه‌های هم‌نگاره به‌طور قابل‌توجهی متداول‌تر از سایرین بود، آن واژه نگه داشته شد و واژه‌های دیگر حذف گردیدند.

در این مرحله داوران تصمیم گرفتند که کدام واژه‌ها در دیکشنری حفظ شوند. باقی ماندن یک کلمه در یک طبقه یا افزوده شدن آن به طبقه دیگر، نیازمند تأیید توسط اکثریت داوران بود. در واقع هر کلمه توسط ۴ روانشناس داوری شدند. اگر اکثریت داوران به آن کلمه نمره می‌دادند، کلمه در آن دیکشنری باقی می‌ماند، کلماتی که به اجماع داوران نمی‌رسیدند از دیکشنری حذف می‌شدند. از سوی دیگر برای کلماتی که داوران ۲ نمره برای تأیید آن کلمه و ۲ نمره به حذف آن کلمه داده بودند، جلسات داوری گروهی برگزار شد و در مورد آن کلمات تصمیم‌گیری شد. با استناد به روش پنه‌بیکر و همکاران (۲۰۱۵) و بوید و همکاران (۲۰۲۲) در صورت عدم توافق میان داوران، با تحلیل پیکره‌های متون[1] (مجموعه‌ای گسترده از متن‌هایی در موضوعات و قالب‌های متنوع از جمله شبکه‌های اجتماعی، نشریات و ادبیات داستانی و ...)، معیارهای دیگر مانند بیشترین موارد استفاده از واژه و معانی متعدد آن مشخص شد و بر اساس آن‌ها تصمیم‌گیری صورت گرفت.

سپس بازبینی کامپیوتری انجام شد. در این مرحله واژگان و اصطلاحات واژه‌نامه‌ی ساخته شده در پیکره‌های متون جست‌وجو شدند تا مشخص شود که در متن چگونه به کار می‌روند. هدف از انجام این تحلیل این بود که مطمئن شویم کلمات واژه‌نامه در زبان روزمره متداول هستند. کلمات و اصطلاحاتی که حداقل یک بار در یکی از منابع بررسی شده به کار نرفته باشند از واژه‌نامه حذف شدند. علاوه بر این، بر اساس روش جی و رینی (۲۰۲۰) واژگانی که کمتر از ۱۰ بار در این منابع تکرار شده بودند نیز دوباره مورد بررسی و داوری قرار گرفتند.

[1] Corpora

مرحله سوم: گسترش و توسعه فرهنگ لغت

این مرحله در چهار بخش انجام شد. ابتدا واژه‌نامه‌ی تهیه شده از نظر تناسب با ساختار زبان فارسی و ویژگی‌های فرهنگی آن بررسی شد. با توجه به اینکه ویژگی‌های ساختاری زبان فارسی با زبان‌های دیگر متفاوت است، روش انجام در بخش اول بر اساس قواعد واژه‌شناسی و صرف و نحو زبان فارسی تعیین شد. هدف از این بخش بررسی هم‌خوانی ساختاری دیکشنری ساخته شده و زبان فارسی با روش‌های تحلیل متن دیکشنری محور است. در واقع این مرحله برای اطمینان از این مورد انجام شد که نرم‌افزار تحلیل متن می‌تواند واژه‌ها را به درستی شناسایی، طبقه‌بندی و تحلیل کند. به این منظور، کلمات طبقه‌های دیکشنری ساخته شده بررسی گردیدند و با توجه به نتایج به دست آمده از این بررسی، تغییرات لازم در ساختار طبقات اعمال شد. بررسی تناسب با زبان و فرهنگ از نظر محتوایی و مفهومی نیز اهمیت دارد؛ بنابراین، محتوای هر طبقه با هدف افزودن واژگان مرتبط با فرهنگ ایرانی بررسی گردید تا دیکشنری نهایی تا جای ممکن مجموعه واژگان مرتبط موجود در زبان فارسی را پوشش دهد. پس از این که تغییرات لازم اعمال و کاستی‌های موجود شناسایی شدند، با توجه به ماهیت و وسعت این کاستی‌ها، با استفاده از روش‌های متناسب برآمده از ادبیات پژوهشی مرتبط و پس از مطابقت دادن این روش‌ها با روش‌شناسی استفاده شده در ساخت نسخه‌های مختلف LIWC، بار دیگر مجموعه‌ی جدیدی از واژه‌ها شناسایی و به دیکشنری افزوده گردیدند.

هدف بخش دوم از این مرحله شناسایی‌های متداول‌ترین واژه‌های مرتبط با طبقات نسخه فارسی واکاوی زبانی و شمارش واژگان در زبان فارسی و افزودن آن‌ها به دیکشنری بود. به منظور افزودن واژگان جدید، کلمات ترجمه شده در واژه‌نامه‌ها، دایره‌المعارف‌ها و شبکه‌های واژگان فارسی جست‌وجو گردیدند، واژگان و اصطلاحات مترادف آن‌ها مشخص شد و به واژه‌های ترجمه شده افزوده گردیدند. همچنین، بار دیگر پیکره‌های متنی فارسی بررسی شدند تا پرتکرارترین واژه‌های زبان فارسی شناسایی شوند. سپس واژه‌های به دست آمده با واژه‌های دیکشنری طبقات نسخه فارسی واکاوی زبانی و شمارش واژگان مطابقت داده شد و واژه‌های تکراری حذف گردیدند. همچنین همبستگی همه واژه‌های غیرتکراری با هر کدام از طبقه‌های دیکشنری سنجیده شدند و از کلماتی که حداقل با یکی از طبقه‌های دیکشنری همبستگی مثبت داشتند یک فهرست تهیه شد. به همین ترتیب سپس، داوران واژگان این فهرست را یکی یکی و با روش به کار رفته در مرحله‌ی دوم از نظر تناسب با طبقات واژه‌نامه بررسی کردند، به آن‌ها امتیاز دادند و آن‌ها را در مقوله‌ها طبقه‌بندی و در نهایت کلمات نامتناسب را حذف کردند.

در بخش سوم به انجام تحلیل همانندی واژه‌ها[1] پرداخته شد. با هدف افزایش پوشش دهی و کارآمدی دیکشنری، مجموعه‌ی شامل کلمات جدید بار دیگر بازبینی شدند و طبقه یا طبقاتی که به گسترش بیشتر نیاز داشتند، مشخص گردیدند. در این صورت، با استناد به روش استفاده شده در نسخه‌ی آلمانی D-LIWC (مایر و همکاران، ۲۰۱۸) تحلیل معنایی نهان[2] انجام شد. این روش فرض می‌کند که گروه‌هایی از کلمات که در یک قسمت از متن آمده‌اند با یکدیگر ارتباط دارند؛ بنابراین با انجام این تحلیل بر روی کلمات طبقات انتخاب شده از دیکشنری در پیکره‌های متون، فهرست جدیدی از واژگان تهیه شد. این فهرست بار دیگر به روش اشاره شده در مراحل قبل توسط داوران ارزیابی گردید و کلمات به طبقات متناسب اضافه شدند. در این مرحله گسترش دیکشنری به پایان رسید.

پس از اتمام داوری در بخش چهارم، به منظور تشخیص اشکال مختلف کلمه توسط کامپیوتر، لما کلمات توسط زبان‌شناسان تعیین شدند. روش کار به این صورت بود که زبان‌شناس خود را به‌جای کامپیوتر قرار می‌داد و کلمه مورد نظر از نظر تشخیص کامپیوتر و الگورتیم محاساباتی بررسی و ابتدای کلمه مورد نظر به‌عنوان لما در نظر گرفته می‌شد. در صورتی که الگوریتم در تشخیص کلمه و کلمات مشتق از آن دچار مشکل نمی‌شد کلمه مورد نظر به‌عنوان لما ثبت می‌گردید. در این صورت الگوریتم می‌تواند کلمه مورد نظر و کلماتی که با این صورت کلمه شروع می‌شوند را در متن مربوطه تشخیص دهد. مثلاً کلمه وهم به‌عنوان لما در نظر گرفته شد و از این طریق کلماتی مثل وهم‌ها و یا وهم‌آلود و وهمناک نیز که کد و طبقه یکسان دارند به درستی تشخیص داده می‌شوند. کلمه هوش نیز می‌تواند به‌عنوان لما در نظر گرفته شود زیرا سیستم می‌تواند علاوه بر این کلمه، کلماتی مثل هوشمند، هوشیار، هوشم و دیگر کلمات مرتبط با کد و طبقه یکسان را تشخیص دهد و کلمه دیگری که کد متفاوت داشته باشد تشخیص داده نمی‌شود.

[1] Word Similarity Analysis

[2] Latent Semantic Analysis

اما در مورد یک سری کلمات در صورتی که کلمه مورد نظر به‌عنوان لما در نظر گرفته می‌شد سیستم دچار خطا می‌گردید. در این صورت کلمه مورد نظر به‌عنوان لما در نظر گرفته نمی‌شد و در عوض شکل‌های مختلف کلمه که از آن مشتق شده و دارای کد و طبقه یکسان هستند توسط کامپیوتر تولید می‌شدند و به دیکشنری اضافه شدند؛ مثلاً کلمه ول نمی‌تواند به‌عنوان لما در نظر گرفته شود زیرا در این صورت سیستم کلماتی مثل ولوله یا ولد را نیز تشخیص می‌دهد که دارای کد و طبقه متفاوتی هستند؛ بنابراین در مورد این کلمه، طبق فرمول داده شده به سیستم، کلماتی مثل ول شده، ولیم و ولش توسط سیستم تولید شده و در دیکشنری قرار داده شدند. در مورد کلمه وام نیز به همین ترتیب عمل شد. این کلمه نمی‌تواند به‌عنوان لما در نظر گرفته شود زیرا در این صورت الگوریتم کلماتی مانند وامانده را نیز تشخیص می‌دهد که از نظر معنایی و کد روانشناسی با کلمه وام متفاوت است؛ بنابراین کلمه وام به‌عنوان لما در نظر گرفته نشد و در عوض شکل‌های مختلف آن مانند وام‌ها، وامم و وامت توسط سیستم تولید و به دیکشنری اضافه گردید. همچنین در برخی موارد مانند کلمات وازده و وازدگی، یک کلمه کامل به‌عنوان لما در نظر گرفته نمی‌شد. در مورد این کلمات، صورت وازد به‌عنوان لما در نظر گرفته شد که در این صورت الگوریتم هم کلمه وازدگی و هم کلمه وازده را تشخیص می‌دهد.

در مورد فعل‌ها، ابتدا فعل‌ها به سه دسته فعل‌های ساده، فعل‌های پیشوندی و فعل‌های مرکب تقسیم شدند و پس از آن شکل‌های مختلف فعل در زمان‌های دستوری مختلف از مصدر فعل تولید شده و به دیکشنری اضافه گردیدند. البته در مورد برخی فعل‌های مرکب مثل هدر رفتن، جزء غیر فعلی به‌عنوان لما در نظر گرفته می‌شد تا در این صورت شکل‌های مختلف کلمه و فعل مرکب تشخیص داده شوند. در مورد اسم‌ها و صفاتی که برای آن‌ها لما در نظر گرفته نشد و همچنین فعل‌ها، شکل‌های مختلف کلمه تولید و به دیکشنری اضافه گردیدند. توضیح اینکه اسم‌ها و صفات با توجه به رفتار صرفی و پسوندهایی که می‌توانند بگیرند به چند دسته تقسیم شدند تا از تولید صورت‌های نادرست جلوگیری شود.

یافته‌ها

مراحل چهار و پنج مربوط به بخش یافته‌های این پژوهش هستند که ویژگی‌های روان‌سنجی این ابزار در زبان فارسی مورد ارزیابی قرار می‌گیرند. در ادامه این دو مرحله به صورت کامل توضیح داده خواهند شد.

مرحله چهارم: ارزیابی ویژگی‌های روان‌سنجی

هدف از این مرحله، بررسی پایایی و روایی دیکشنری‌ها بر اساس روش‌های استفاده شده بر روی نسخه‌های متفاوت LIWC است. برای محاسبه پایایی و روایی دیکشنری از روش پیشنهادی پنه‌بیکر و همکاران (۲۰۱۵) و بوید و همکاران (۲۰۲۲) استفاده شد. از دیدگاه روان‌سنجی، یک معیار روان‌شناختی پایا و معتبر باید به‌طور مداوم ویژگی‌های روان‌شناختی هدف را نشان دهد. به گفته پنه‌بیکر و همکاران. (۲۰۱۵)، اگر معیار مبتنی بر فرهنگ لغت مانند LIWC به درستی جنبه‌های روان‌شناختی فردی را که متنی را تولید می‌کند به تصویر بکشد، متن حاوی چندین کلمه مرتبط با یک طبقه روان‌شناختی در فرهنگ لغت خواهد بود. به‌عنوان مثال، اگر شخصی در مورد تجربه مثبت خود به زبان فارسی خاطراتی بنویسد، کلمات متعددی در طبقه هیجان مثبت LIWC، مانند «زیبا»، «سرگرم‌کننده» و «شگفت‌انگیز» در یک متن وجود دارد؛ بنابراین الگوی بیانی تا حدی که هیجانات مثبت یک فرد در متن بالا باشد، از نظر درونی سازگار است. همچنین پنه‌بیکر و بوید (۲۰۱۵، ۲۰۲۲) هشدار می‌دهند که محاسبه پایایی و روایی دیکشنری‌ها ساده نیست و نتایج به دست آمده از روش‌های پیشنهاد شده باید با احتیاط تفسیر شوند. با وجود این که روش‌های مورد استفاده بر روی واژه‌نامه از نظر اجرا مشابه با روش‌های استفاده شده در سنجش پایایی و روایی پرسشنامه‌ها هستند، اما با توجه به ماهیت متفاوت زبان طبیعی، آستانه‌ی قابل قبول برای ضرایب پایایی در زبان‌های طبیعی پایین‌تر است و تصمیم‌گیری درباره‌ی روایی واژه‌نامه نیز باید توسط روش‌های پیچیده‌تری انجام شود.

برای سنجش پایایی دیکشنری، همسانی درونی هر طبقه محاسبه شد. با توجه به همانندی و نزدیکی واژه‌های یک طبقه با یکدیگر انتظار می‌رود با افزایش تکرار یک واژه در متن، دفعات استفاده از واژه‌های دیگر متعلق به آن طبقه نیز در آن متن افزایش یابد (پنه‌بیکر و همکاران، ۲۰۱۵؛ بوید و همکاران، ۲۰۲۲). در این بررسی از پیکره‌ای از متون که شامل بیش از ۱۰۰ میلیون متن شبکه‌های اجتماعی ایکس (توییتر)، اینستاگرام، تلگرام و متون رسمی خبری شامل روزنامه‌ها و خبرگزاری‌های آنلاین استفاده شد. این متن‌ها متعلق به سال‌های ۱۳۹۸ تا ۱۴۰۴ بودند و توسط شرکت لایف‌وب

جمع آوری شده و در اختیار تیم توسعه دهنده قرار گرفته است. همچنین این پیکره شامل ۴ میلیارد و ۸۰۰ میلیون توکن بود. برای بررسی این همبستگی متقابل، در ابتدا واژه‌های هر طبقه به صورت جداگانه در نظر گرفته شدند. سپس تعداد دفعاتی که هر واژه در مجموعه‌ی پیکره‌های متون به کار رفته است شمرده و این عدد به تعداد کل کلمات موجود در پیکره‌های متون تقسیم گردید تا نمره‌ی هر واژه به‌عنوان درصدی از تعداد کل کلمات به دست آید. سپس هر کدام از این نمره‌ها به‌عنوان یک آیتم در نظر گرفته شدند و ضریب همبستگی آلفای کرونباخ و به روش معمول محاسبه گردید؛ بنابراین با این روش برای هر یک از طبقات دیکشنری در هر کدام از پیکره‌های متون یک ضریب آلفای به دست آمد. با توجه به نوع محاسبات انجام شده در این روش و ماهیت طبقات زبانی، ضریب آلفا پایایی طبقات را بسیار پایین‌تر از مقدار واقعی محاسبه می‌کند. به همین دلیل علاوه بر آلفای کرونباخ از فرمول کودر-ریچاردسون-۲۰ نیز برای بررسی پایایی نیز استفاده شد و هر دو نمره برای هر طبقه گزارش گردید. جدول ۱، ضریب آلفای کرونباخ و کودر-ریچاردسون ۲۰ برای همه طبقات را نشان می‌دهد.

جدول ۱. ضریب آلفای کرونباخ و کودر-ریچاردسون ۲۰ برای همه طبقات

طبقه		تعداد کلمات	آلفای کرونباخ	کودر-ریچاردسون ۲۰
ابعاد زبان‌شناختی	ابعاد زبان‌شناختی	۶۰۴۰	۰/۹۷	۰/۹۷
	کل واژگان دستوری	۱۸۶۹	۰/۹۵	۰/۹۷
	کل ضمایر	۴۰۷	۰/۷۸	۰/۸۳
	ضمایر شخصی	۳۰۹	۰/۶۵	۰/۷۰
	ضمیر اول شخص مفرد (من)	۸۰	۰/۵۷	۰/۸۶
	ضمیر اول شخص جمع (ما)	۲۵	۰/۵۲	۰/۸۳
	ضمیر دوم شخص (تو/شما)	۸۹	۰/۴۳	۰/۷۸
	ضمیر سوم شخص مفرد (او)	۸۷	۰/۵۴	۰/۸۲
	ضمیر سوم شخص جمع (آنها)	۲۸	۰/۴۷	۰/۸۰
	ضمایر غیرشخصی / اشاره‌ای	۹۹	۰/۶۹	۰/۷۳
	حروف تعریف (معرفه/نکره)	۳۹۲	۰/۶۵	۰/۷۶
	اعداد	۱۲۱	۰/۶۹	۰/۷۵
	حروف اضافه	۱۱۲	۰/۸۱	۰/۹۳
	افعال کمکی	۲۸۴	۰/۸۸	۰/۹۱
	قیود	۵۷۵	۰/۹۳	۰/۹۳
	حروف ربط	۴۹	۰/۶۸	۰/۸۳
	واژگان نفی	۲۳۱	۰/۸۳	۰/۹۵
سائق‌ها/کشش‌های درونی	افعال رایج	۱۷۰۸	۰/۹۵	۰/۹۶
	صفات رایج	۲۵۹۸	۰/۹۵	۰/۹۷
	واژگان کمّیت	۳۴۰	۰/۷۸	۰/۸۲
	سائق‌ها/کشش‌های درونی	۱۷۶۸	۰/۸۸	۰/۹۱
	پیوندجویی / تعلق اجتماعی	۵۸۹	۰/۷۳	۰/۸۰
	پیشرفت	۳۰۲	۰/۶۶	۰/۷۳
	قدرت	۹۴۲	۰/۷۸	۰/۸۵
شناخت	شناخت	۲۲۶۹	۰/۹۶	۰/۹۶
	تفکر مطلق‌نگر / همه یا هیچ	۸۴	۰/۶۵	۰/۸۵
	فرایندهای شناختی	۲۱۶۰	۰/۹۵	۰/۹۶
	بینش / درک	۷۶۶	۰/۸۵	۰/۸۸
	علیت / استدلال	۴۲۷	۰/۸۵	۰/۸۸
	ناهماهنگی / ناسازگاری	۲۰۷	۰/۶۱	۰/۶۸

طبقه		تعداد کلمات	آلفای کرونباخ	کودر-ریچاردسون ۲۰
هیجانات	تردید / ابهام‌گویی	۳۷۱	۰/۷۸	۰/۸۱
	قطعیت / یقین	۲۰۳	۰/۶۹	۰/۷۰
	تمایزگذاری	۴۵۴	۰/۸۶	۰/۸۸
	حافظه	۳۹	۰/۳۴	۰/۷۴
	هیجانات	۵۷۹۸	۰/۹۲	۰/۹۵
	لحن مثبت	۲۲۰۷	۰/۸۸	۰/۹۲
	لحن منفی	۳۱۹۰	۰/۸۳	۰/۸۹
	هیجان	۱۶۳۲	۰/۷۱	۰/۷۹
	هیجان مثبت	۴۶۹	۰/۶۴	۰/۸۱
	هیجان منفی	۱۱۱۷	۰/۶۳	۰/۷۹
	اضطراب	۲۱۶	۰/۴۴	۰/۷۵
	خشم	۳۳۴	۰/۶۵	۰/۸۹
	غم	۲۳۹	۰/۶۷	۰/۹۱
	دشنام	۴۴۲	۰/۶۳	۰/۸۲
فرایندهای اجتماعی	فرایندهای اجتماعی	۳۴۶۷	۰/۹۰	۰/۹۳
	رفتار اجتماعی	۲۹۲۰	۰/۸۹	۰/۹۲
	رفتارجامعه‌پسند/یاریگرانه	۵۳۱	۰/۶۸	۰/۷۴
	ادب / مؤدبانه‌گویی	۲۶۱	۰/۵۹	۰/۷۸
	تعارض بین‌فردی	۷۰۰	۰/۶۸	۰/۷۵
	اخلاق‌گرایی	۱۱۱۶	۰/۷۰	۰/۷۸
	ارتباط	۶۸۱	۰/۸۰	۰/۸۲
	ارجاعات اجتماعی	۶۲۱	۰/۶۷	۰/۷۶
	خانواده	۲۴۴	۰/۵۶	۰/۶۸
	دوستان	۱۳۷	۰/۴۳	۰/۷۷
	ارجاع به زن	۲۲۱	۰/۵۷	۰/۷۹
	ارجاع به مرد	۱۵۵	۰/۶۴	۰/۸۳
فرهنگ	فرهنگ	۹۸۲	۰/۸۳	۰/۸۷
	سیاست	۵۲۸	۰/۸۳	۰/۸۶
	قومیت	۲۴۵	۰/۵۱	۰/۵۷
	فناوری	۲۱۱	۰/۵۳	۰/۸۵
سبک زندگی	سبک زندگی	۲۴۱۹	۰/۸۶	۰/۹۱
	فراغت	۷۲۸	۰/۶۰	۰/۸۶
	خانه	۲۴۱	۰/۴۳	۰/۷۷
	کار	۶۰۷	۰/۸۲	۰/۸۶
	پول	۳۵۱	۰/۷۴	۰/۸۲
	دین/مذهب	۵۴۳	۰/۷۷	۰/۸۳
بدنی / جسمانی	بدنی / جسمانی	۱۶۲۳	۰/۷۲	۰/۸۴
	سلامت	۵۵۵	۰/۷۱	۰/۸۱
	بیماری	۲۹۲	۰/۵۸	۰/۷۸
	تندرستی	۱۳۱	۰/۴۳	۰/۷۲
	سلامت روان	۱۳۵	۰/۴۱	۰/۶۷

طبقه		تعداد کلمات	آلفای کرونباخ	کودر-ریچاردسون ۲۰
حالات	مواد مخدر	۱۸۱	۰/۴۰	۰/۶۵
	جنسی	۳۸۰	۰/۵۱	۰/۷۴
	غذا	۳۸۰	۰/۶۷	۰/۷۹
	مرگ	۱۸۹	۰/۴۹	۰/۸۱
	نیاز	۱۲۹	۰/۵۰	۰/۷۹
	خواست	۱۵۱	۰/۳۸	۰/۷۳
	به‌دست‌آوردن	۱۷۹	۰/۴۳	۰/۸۱
	فقدان	۲۳۸	۰/۵۷	۰/۶۵
	برآورده‌سازی	۱۳۷	۰/۶۷	۰/۶۹
	خستگی	۱۸۴	۰/۴۲	۰/۷۵
انگیزش‌ها	پاداش	۸۲	۰/۴۵	۰/۷۶
	ریسک / خطرپذیری	۱۸۹	۰/۴۲	۰/۶۹
	کنجکاوی	۱۶۳	۰/۳۹	۰/۷۳
	جاذبه / کشش	۲۰۹	۰/۷۷	۰/۸۲
ادراک	ادراک	۲۵۳۶	۰/۸۳	۰/۹۴
	توجه	۱۶۵	۰/۴۱	۰/۷۸
	حرکت	۷۶۹	۰/۷۲	۰/۹۳
	فضا	۸۳۱	۰/۷۲	۰/۹۰
	دیداری	۳۵۷	۰/۶۵	۰/۷۶
	شنیداری	۳۱۶	۰/۴۷	۰/۷۱
	حسی / لامسه‌ای	۱۶۹	۰/۵۲	۰/۸۲
جهت‌گیری زمانی	زمان	۷۵۹	۰/۸۷	۰/۸۹
	تمرکز بر گذشته	۷۰۵	۰/۸۱	۰/۸۰
	تمرکز بر حال	۳۶۷	۰/۵۸	۰/۸۴
	تمرکز بر آینده	۱۸۶	۰/۵۸	۰/۷۹
محاوره‌ای	محاوره‌ای	۲۰۰	۰/۶۹	۰/۸۹
	زبان اینترنتی	۱۲۳	۰/۶۵	۰/۸۲
	تأیید / موافقت	۱۹	۰/۴۸	۰/۷۵
	ناپیوستگی‌های گفتاری	۱۵	۰/۳۷	۰/۶۶
	کلمات پرکننده	۵۰	۰/۲۵	۰/۶۲
	ایموجی / شکلک	۴۸	۰/۱۲	۰/۶۶
	علائم نگارشی	۹۲	۰/۶۶	۰/۷۰

در مرحله بعد، به منظور بررسی روایی بیرونی، هم ارزی بین نسخه فارسی و نسخه نهایی انگلیسی LIWC-22 بررسی شد. در این مرحله دو مجموعه متن برای تجزیه‌وتحلیل مورد استفاده قرار گرفت. ۲۰۰ متن از موضوعات مختلف سیاسی، اجتماعی، اقتصادی، روانشناسی، ورزشی، ادبیات، زندگینامه و مذهبی، محیط زیست، گردشگری، مدیریت، تاریخ، حقوق و... انتخاب شد که هریک بین ۷۰۰ تا ۱۲۰۰ کلمه بودند. در این مجموعه ۱۰۰ متنی که به زبان فارسی وجود داشت به زبان انگلیسی برگردانده شد و ۱۰۰ متن به زبان انگلیسی وجود داشت که به زبان فارسی ترجمه شد. سپس ضریب همبستگی پیرسون (r) به‌عنوان شاخص هم ارزی میان نسخه فارسی و انگلیسی محاسبه شد. در واقع هر چه دو نسخه از لحاظ عملکرد مشابه باشند ضرایب همبستگی بین طبقات آن‌ها بالاتر است. جدول ۲ ضرایب همبستگی بین دو نسخه فارسی و انگلیسی را نشان می‌دهد.

جدول۲. ضریب همبستگی بین نسخه فارسی و انگلیسی

طبقه		ضریب همبستگی پیرسون	سطح معناداری
ابعاد زبان‌شناختی	ابعاد زبان‌شناختی	۰/۷۳	p<۰/۰۱
	کل واژگان دستوری	۰/۷۱	p<۰/۰۱
	کل ضمایر	۰/۷۱	p<۰/۰۱
	ضمایر شخصی	۰/۷۰	p<۰/۰۱
	ضمیر اول شخص مفرد (من)	۰/۸۳	p<۰/۰۱
	ضمیر اول شخص جمع (ما)	۰/۶۸	p<۰/۰۱
	ضمیر دوم شخص (تو/شما)	۰/۶۷	p<۰/۰۱
	ضمیر سوم شخص مفرد (او)	۰/۸۶	p<۰/۰۱
	ضمیر سوم شخص جمع (آنها)	۰/۶۲	p<۰/۰۱
	ضمایر غیرشخصی / اشاره‌ای	۰/۵۵	p<۰/۰۱
	حروف تعریف (معرفه/نکره)	۰/۱۲	p<۰/۰۱
	اعداد	۰/۷۴	p<۰/۰۱
	حروف اضافه	۰/۶۱	p<۰/۰۱
	افعال کمکی	۰/۶۶	p<۰/۰۱
	قیود	۰/۵۹	p<۰/۰۱
	حروف ربط	۰/۵۰	p<۰/۰۱
	واژگان نفی	۰/۶۸	p<۰/۰۱
	افعال رایج	۰/۷۴	p<۰/۰۱
	صفات رایج	۰/۶۷	p<۰/۰۱
	واژگان کمّیت	۰/۵۹	p<۰/۰۱
سائق‌ها/کشش‌های درونی	سائق‌ها/کشش‌های درونی	۰/۸۱	p<۰/۰۱
	پیوندجویی / تعلق اجتماعی	۰/۷۰	p<۰/۰۱
	پیشرفت	۰/۸۶	p<۰/۰۱
	قدرت	۰/۸۵	p<۰/۰۱
شناخت	شناخت	۰/۸۴	p<۰/۰۱
	تفکر مطلق‌نگر / همه یا هیچ	۰/۶۶	p<۰/۰۱
	فرایندهای شناختی	۰/۸۵	p<۰/۰۱
	بینش / درک	۰/۸۶	p<۰/۰۱
	علیت / استدلال	۰/۶۷	p<۰/۰۱
	ناهماهنگی / ناسازگاری	۰/۵۴	p<۰/۰۱
	تردید / ابهام‌گویی	۰/۸۷	p<۰/۰۱
	قطعیت / یقین	۰/۶۱	p<۰/۰۱
	تمایزگذاری	۰/۷۵	p<۰/۰۱
	حافظه	۰/۶۳	p<۰/۰۱
هیجانات	هیجانات	۰/۷۶	p<۰/۰۱
	لحن مثبت	۰/۸۰	p<۰/۰۱
	لحن منفی	۰/۸۲	p<۰/۰۱
	هیجان	۰/۷۴	p<۰/۰۱
	هیجان مثبت	۰/۸۲	p<۰/۰۱
	هیجان منفی	۰/۸۶	p<۰/۰۱

طبقه		ضریب همبستگی پیرسون	سطح معناداری
فرایندهای اجتماعی	اضطراب	۰/۹۶	p<۰/۰۱
	خشم	۰/۸۲	p<۰/۰۱
	غم	۰/۷۸	p<۰/۰۱
	دشنام	۰/۵۱	p<۰/۰۱
	فرایندهای اجتماعی	۰/۶۷	p<۰/۰۱
	رفتار اجتماعی	۰/۶۲	p<۰/۰۱
	رفتارجامعه‌پسند/یاریگرانه	۰/۵۲	p<۰/۰۱
	ادب / مؤدبانه‌گویی	۰/۵۴	p<۰/۰۱
	تعارض بین‌فردی	۰/۷۴	p<۰/۰۱
	اخلاق‌گرایی	۰/۶۱	p<۰/۰۱
	ارتباط	۰/۷۲	p<۰/۰۱
	ارجاعات اجتماعی	۰/۷۵	p<۰/۰۱
	خانواده	۰/۷۸	p<۰/۰۱
	دوستان	۰/۷۱	p<۰/۰۱
	ارجاع به زن	۰/۸۴	p<۰/۰۱
	ارجاع به مرد	۰/۸۱	p<۰/۰۱
فرهنگ	فرهنگ	۰/۸۳	p<۰/۰۱
	سیاست	۰/۸۲	p<۰/۰۱
	قومیت	۰/۸۱	p<۰/۰۱
	فناوری	۰/۸۳	p<۰/۰۱
سبک زندگی	سبک زندگی	۰/۹۰	p<۰/۰۱
	فراغت	۰/۹۲	p<۰/۰۱
	خانه	۰/۶۱	p<۰/۰۱
	کار	۰/۹۴	p<۰/۰۱
	پول	۰/۸۸	p<۰/۰۱
بدنی / جسمانی	دین/مذهب	۰/۹۵	p<۰/۰۱
	بدنی / جسمانی	۰/۷۳	p<۰/۰۱
	سلامت	۰/۸۴	p<۰/۰۱
	بیماری	۰/۸۳	p<۰/۰۱
	تندرستی	۰/۶۳	p<۰/۰۱
	سلامت روان	۰/۸۲	p<۰/۰۱
	مواد مخدر	۰/۶۳	p<۰/۰۱
حالات	جنسی	۰/۷۸	p<۰/۰۱
	غذا	۰/۸۳	p<۰/۰۱
	مرگ	۰/۹۱	p<۰/۰۱
	نیاز	۰/۸۱	p<۰/۰۱
	خواست	۰/۵۲	p<۰/۰۱
	به‌دست‌آوردن	۰/۶۱	p<۰/۰۱
	فقدان	۰/۶۴	p<۰/۰۱
	برآورده‌سازی	۰/۵۸	p<۰/۰۱
انگیزش‌ها	خستگی	۰/۵۴	p<۰/۰۱
	پاداش	۰/۷۵	p<۰/۰۱

طبقه		ضریب همبستگی پیرسون	سطح معناداری
ادراک	ریسک / خطرپذیری	۰/۸۱	p<۰/۰۱
	کنجکاوی	۰/۵۳	p<۰/۰۱
	جاذبه / کشش	۰/۶۱	p<۰/۰۱
	ادراک	۰/۶۰	p<۰/۰۱
	توجه	۰/۷۰	p<۰/۰۱
	حرکت	۰/۶۴	p<۰/۰۱
	فضا	۰/۷۴	p<۰/۰۱
	دیداری	۰/۵۳	p<۰/۰۱
	شنیداری	۰/۵۲	p<۰/۰۱
	حسی / لامسه‌ای	۰/۸۵	p<۰/۰۱
جهت‌گیری زمانی	زمان	۰/۷۹	p<۰/۰۱
	تمرکز بر گذشته	۰/۷۷	p<۰/۰۱
	تمرکز بر حال	۰/۶۶	p<۰/۰۱
	تمرکز بر آینده	۰/۵۷	p<۰/۰۱
محاوره‌ای	محاوره‌ای	۰/۳۵	p<۰/۰۱
	زبان اینترنتی	۰/۲۵	p<۰/۰۱
	تأیید / موافقت	۰/۳۷	p<۰/۰۱
	ناپیوستگی‌های گفتاری	۰/۱۶	p<۰/۰۱
	کلمات پرکننده	۰/۱۸	p<۰/۰۱
	ایموجی / شکلک	۰/۷۳	p<۰/۰۱
	علائم نگارشی	۰/۷۰	p<۰/۰۱

همانطور که از جدول ۶ مشخص است، همه طبقات نسخه فارسی با نسخه نهایی انگلیسی LIWC-22 دارای همبستگی معنادار در سطح (p<۰/۰۱) بودند.

بحث و نتیجه‌گیری

هدف این پژوهش توسعه و بررسی ویژگی‌های روان‌سنجی تمامی طبقات نسخه فارسی سامانه واکاوی زبانی و شمارش واژگان (P-LIWC) بود. با انجام یک سری تجزیه و تحلیل‌های سیستماتیک، مشخص شد که تمامی طبقات این نسخه فارسی، شامل واژگان ترجمه‌شده از نسخه اصلی انگلیسی و واژگان متداول اقتباس‌شده از مجموعه‌های بزرگ متنی فارسی، از همبستگی درونی کافی و سازگاری قابل‌قبول بین طبقات مختلف برخوردار هستند. همچنین روایی بیرونی بین نسخه فارسی و انگلیسی LIWC-22 با مقایسه رونوشت‌های متون فارسی و انگلیسی تأیید شد؛ بنابراین شواهد نشان می‌دهد که نسخه فارسی سامانه واکاوی زبانی و شمارش واژگان P-LIW یک ابزار تحقیقاتی قدرتمند برای پژوهشگران علوم انسانی دیجیتال و علوم اجتماعی محاسباتی است تا جنبه‌های روان‌شناختی موجود در متون فارسی را بررسی کنند.

بخش تحلیل یافته‌ها نشان داد که تمامی طبقات نسخه فارسی واکاوی زبانی و شمارش واژگان (P-LIWC)، شامل هیجانات، فرآیندهای شناختی، ابعاد زبان‌شناختی، سائق‌ها و انگیزه‌ها، فرایندهای اجتماعی، فرهنگ، سبک زندگی، بدنی/جسمانی، حالات، ادراک، جهت‌گیری زمانی و ویژگی‌های محاوره‌ای، از همبستگی درونی کافی و اعتبار روان‌سنجی مناسبی برخوردار هستند. این یافته‌ها تأکید می‌کند که P-LIWC قادر است تحلیل‌های روان‌شناختی دقیق و گسترده‌ای از متون فارسی ارائه دهد و امکان مقایسه و همسان‌سازی بین متون فارسی و انگلیسی را فراهم سازد که این امر نقش مهمی در تحقیقات بین‌فرهنگی و بین‌زبانی ایفا می‌کند. علاوه بر این، نتایج نشان می‌دهد که استفاده از این ابزار می‌تواند دیدگاه‌های چندوجهی و جامع‌تری نسبت به حالات روانی، الگوهای شناختی، انگیزه‌ها

و رفتارهای اجتماعی افراد ارائه دهد و به پژوهشگران امکان دهد تا تحلیل‌های کمی و کیفی متنی را به‌صورت همزمان انجام دهند.

نسخه فارسی واکاوی زبانی و شمارش واژگان ظرفیت بالایی برای پژوهش‌های آینده در عصر کلان داده و تجزیه و تحلیل داده‌های اجتماعی دارد. محققان می‌توانند از فرهنگ لغت برای تجزیه‌وتحلیل حالات روان‌شناختی فرد تا فرآیندهای روان‌شناختی پیچیده گروه‌ها که در چندین شکل از متون کلان فارسی منعکس شده است استفاده کنند، مانند پست‌های شبکه‌های اجتماعی، متون رسمی و غیر رسمی، رونوشت‌های صوتی مصاحبه‌ها و فیلم‌ها. صورت‌جلسه و غیره. به‌عنوان مثال، ساساهارا و همکاران (۲۰۲۱) پست‌های توییتر را با نسخه ژاپنی J-LIWC-2015 تجزیه‌وتحلیل کردند تا الگوهای تغییرات واکنش مصرف‌کنندگان به فروش مجدد کالاهای ضروری (مانند ماسک‌های صورت و ضدعفونی‌کننده‌های دست) در ژاپن را در مرحله اولیه COVID-19 بررسی کنند. در همه‌گیری کرونا، پژوهش‌های طولی نشان داد که اوج استفاده از واژه‌های طبقات خشم و انگیزه در توییتر با انتشار اخبار مربوط به فروش مجدد ماسک، تحریم‌های قانونی علیه فروش مجدد ماسک برای کسب سود و انتقاد از کسانی که بیش از حد از آن‌ها سود می‌برند، مطابقت دارد. این یافته‌ها منعکس‌کننده کاربرد ابزارهای معتبر تحلیلی مبتنی بر متن برای سنجش روندهای روانی در افکار عمومی در اینترنت است. همچنین، معادل‌سازی بین دیکشنری‌های نسخه فارسی (P-LIWC) و نسخه انگلیسی (LIWC-22) تجزیه‌وتحلیل محتوای متون نوشته شده در موضوع و زمینه یکسان توسط گویشوران زبان‌های مختلف را تسهیل می‌کند. به‌عنوان مثال، یک مطالعه بین فرهنگی (لوپز و همکاران، ۲۰۱۹) از LIWC-2007 برای تجزیه‌وتحلیل محتوای پست‌های توییتر سال ۲۰۱۷ با هشتگ می‌تو[1] (هشتگ رسانه‌های اجتماعی که برای اعتراف و به اشتراک‌گذاری تجربیات آزار جنسی در محل کار و اشکال دیگر استفاده می‌شد)، به زبان فرانسوی و انگلیسی استفاده کرد. یافته‌ها نشان داد که توییت‌های فرانسوی شامل عبارات تهاجمی‌تر از توییت‌های انگلیسی است. همین طرح پژوهشی اکنون می‌تواند برای تحلیل تطبیقی متون فارسی و انگلیسی به کار رود. در همین حال، تحقیقات اخیر (ویندزور و همکاران، ۲۰۱۹) ادعا می‌کنند که LIWC-2015 برای تجزیه‌وتحلیل متون اسناد سازمان ملل متحد که به صورت ماشینی از چندین زبان مبدأ غیر انگلیسی به انگلیسی ترجمه شده‌اند مناسب است. با این وجود، این رویکرد ممکن است کلمات با فراوانی بالا را که فقط در زبان‌های مبدأ مشاهده می‌شوند، از دست بدهد. حداقل در حال حاضر، پوشش این رویکرد محدود به متونی است که شامل کلمات غیررسمی یا عبارات خاص فرهنگ نمی‌شود. نرم‌افزار P-LIWC از رویکرد شمارش کلمات بر اساس یک فرهنگ لغت ثابت بدون در نظر گرفتن زمینه استفاده از کلمه استفاده می‌کند. در همین حال، تحقیقات اخیر مزیت بالقوه الگوریتم‌های یادگیری ماشینی پیشرفته، مانند بازنمایی‌ها و تبدیل‌های رمزگذار دوطرفه (BERT؛ دولین و همکاران، ۲۰۱۸) و ترانسفورماتور پیش‌آموزشی ۳ (GPT-3؛ براون و همکاران، ۲۰۲۰) را نسبت به رویکرد سنتی شمارش کلمات در پردازش زبان طبیعی گزارش کرده است (لیک و مورفی، ۲۰۲۳). استفاده از فناوری یادگیری ماشین مبتنی بر داده در این زمینه آینده امیدوار کننده‌ای دارد. با این حال، یکی از بزرگ‌ترین اشکالات رویکرد یادگیری ماشینی، فقدان داده‌های آموزشی موجود است. مدل‌های زبان عصبی مانند BERT به داده‌های «برچسب‌دار» در مقیاس بزرگ برای آموزش و تنظیم دقیق نیاز دارند، اما چنین داده‌هایی اغلب به راحتی در تحقیقات علوم اجتماعی در دسترس نیستند. علاوه بر این، ما نباید فقط مراقب عملکرد الگوریتم‌ها باشیم، بلکه باید در مورد ویژگی‌های شفافیت در روش و قابلیت اعتماد در پردازش داده‌ها نیز مراقب باشیم (فلزمان، ۲۰۱۹).

با این حال، در حال حاضر استفاده از مدل‌های زبانی بزرگ[2] (LLM) بر روی طبقات LIWC-22 در حال بررسی است و این امر می‌تواند نقطه عطفی در هم‌افزایی روش‌های سنتی تحلیل متنی و یادگیری ماشین محسوب شود. در نتیجه، طبقات فرآیندهای شناختی و هیجانات نسخه فارسی سامانه واکاوی زبانی و شمارش واژگان (P-LIWC) توانایی ارائه درک عمیق‌تر و جامع‌تری از ویژگی‌های روان‌شناختی متون فارسی را فراهم می‌آورند. انتظار می‌رود این ابزار به‌عنوان یک پل مهم بین پژوهش‌های کمی و کیفی در زبان فارسی عمل کند و به پژوهشگران امکان دهد تا بینش‌های چندوجهی و جامع در مورد داده‌های متنی به دست آورند. پیش‌بینی می‌شود که P-LIWC در طیف گسترده‌ای از زمینه‌های پژوهشی مرتبط با تحلیل متون فارسی کاربرد داشته باشد و بتواند نوآوری‌های

[1] #MeToo

[2] Larg language models

بیشتری را در حوزه علوم انسانی دیجیتال و علوم اجتماعی محاسباتی ایجاد کند. با وجود هنجار شدن تمامی طبقات نسخه فارسی سامانه واکاوی زبانی و شمارش واژگان (P-LIWC)، برخی محدودیت‌ها و چالش‌ها همچنان وجود دارد. تفاوت‌های ساختاری زبان فارسی با انگلیسی چالش‌هایی را در تعیین ریشه، صیغه و شکل واژگان ایجاد می‌کند که با طراحی الگویی خاص برای شناسایی و طبقه‌بندی واژگان در تمامی طبقات، شامل ابعاد زبان‌شناختی، هیجانات، شناخت، سائق‌ها و انگیزه‌ها، فرایندهای اجتماعی، فرهنگ، سبک زندگی، بدنی/جسمانی، حالات، ادراک، جهت‌گیری زمانی و ویژگی‌های محاوره‌ای تا حد زیادی مدیریت شد؛ اما بهبود این الگوریتم‌ها در پژوهش‌های آینده می‌تواند دقت بیشتری ایجاد کند. پیشنهاد می‌شود که در مطالعات بعدی، کاربرد تمامی طبقات P-LIWC در تحلیل متون طولی، مقایسه‌ای و بین‌فرهنگی و همچنین در شبکه‌های اجتماعی و متون رسمی و غیررسمی مورد آزمایش قرار گیرد و توسعه مدل‌های پیشرفته مبتنی بر یادگیری ماشین و مدل‌های زبانی بزرگ (LLM) برای تحلیل همزمان تمامی طبقات بررسی شود. از نظر کاربردی، نسخه هنجار شده P-LIWC ابزار توانمندی برای تحلیل حالات روان‌شناختی فردی، فرآیندهای شناختی، انگیزه‌ها، رفتارهای اجتماعی، سبک زندگی، تعاملات فرهنگی و سایر ویژگی‌های روانی و رفتاری در متون فارسی است و همسان‌سازی آن با نسخه انگلیسی LIWC امکان تحلیل بین‌زبانی و بین‌فرهنگی را فراهم می‌کند و زمینه‌ساز پژوهش‌های بین‌رشته‌ای و نوآوری در علوم انسانی دیجیتال و علوم اجتماعی محاسباتی خواهد بود.

ملاحظات اخلاقی

پیروی از اصول اخلاق پژوهش: این مقاله برگرفته از پژوهشی مستقل در رشته روانشناسی است. از آنجا که هیچ شرکت‌کننده‌ای در این پژوهش وجود ندارد، بنابراین تنها ملاحظه اخلاقی مربوط به دقت در انجام داوری کلمات بدون سوگیری بوده است.

حامی مالی: این پژوهش در قالب یک پژوهش مستقل است که هیچ حامی مالی ندارد.

نقش هر یک از نویسندگان: نویسنده اول مجری و نویسنده مسئول این پژوهش است. نویسنده دوم، مدیر تیم روان‌شناسی؛ نویسنده سوم، مدیر تیم هوش مصنوعی؛ نویسنده چهارم، مدیر تیم زبان‌شناسی هستند. نویسنده پنجم پژوهشگر جامعه‌شناسی و نویسنده ششم مدیرارشد فنی تیم بودند. نویسندگان، هفتم، هشتم و دهم از پژوهشگران تیم روان‌شناسی و. نویسنده نهم نیز مترجم پژوهش بودند.

تضاد منافع: نویسندگان، هیچ تضاد منافعی را در رابطه با این پژوهش اعلام نمی‌نمایند.

تشکر و قدردانی: در نهایت سپاس و قدردانی از خداوند منان و تمامی همکاران عزیز و پرتلاش که با یاری رساندن در این امر مهم، کمک کردند که این پژوهش با توجه به آرمان‌ها و اهداف علمی کشور به اتمام رسد.

منابع

سلطانی، محمد علی؛ مام شریفی، پیمان؛ صالحی چگنی، رضا؛ صفری، علی؛ ارشادی منش، سودابه؛ کشاورز، حمیدرضا؛ و همکاران. (۱۴۰۳). توسعه و بررسی ویژگی های روان سنجی نسخه فارسی تحقیق زبانی و شمارش کلمات (P-LIWC) طبقات هیجانات و فرآیندهای شناختی. *مجله علوم روانشناختی*، ۲۳ (۱۳۷)، ۱۹-۱. http://dx.doi.org/10.52547/JPS.23.137.997

کاویانی، حسین؛ پور ناصح، مهرانگیز؛ و گلفام، ارسلان. (۱۳۸۴). تقابل واژه‌های هیجان و شناخت در لغت‌نامه‌های زبان فارسی. *تازه‌های علوم شناختی*، ۷ (۲)، ۲۹-۳۷. http://icssjournal.ir/article-1-134-fa.html

مام شریفی، پیمان؛ سهرابی، فرامرز؛ برجعلی، احمد؛ زمان پور، عنایت اله؛ و حسین ثابت، فریده. (۱۴۰۵). واکاوی روانی-زبانی ابعاد شناختی، هیجانی و جنسی هرزه‌نگاری در شبکه‌های اجتماعی اینستاگرام، تلگرام و ایکس. *مجله علوم روانشناختی*، ۲۵ (۱۶۰)، ۱۸-۱. http://dx.doi.org/10.66224/jps.25.160.1

کاویانی، حسین؛ موسوی، اشرف‌سادات؛ و گلفام، ارسلان. (۱۳۸۶). کاوش در واژه‌های روان‌شناختی (هیجان، شناخت، آسیب‌شناسی، درمان و شخصیت)، در لغت‌نامه (فرهنگ) های زبان فارسی. *زبان و زبان‌شناسی*، ۳ (۵)، ۸۹-۱۰۲. https://lsi-linguistics.ihcs.ac.ir/article_1606.html

References

Andy, A., & Andy, U. (2021). Understanding Communication in an Online Cancer Forum: Content Analysis Study. *JMIR cancer*, *7*(3), e29555. https://doi.org/10.2196/29555

Bahgat, M., Wilson, S., & Magdy, W. (2022, June). LIWC-UD: classifying online slang terms into LIWC categories. In *Proceedings of the 14th ACM Web Science Conference 2022* (pp. 422-432). https://doi.org/10.1145/3501247.3531572

Bjekić, J., Lazarević, L. B., Živanović, M., & Knežević, G. (2014). Psychometric evaluation of the Serbian dictionary for automatic text analysis-LIWCser. *Psihologija*, *47*(1), 5-32. https://doi.org/10.2298/PSI1401005B

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin. https://www.liwc.app/help/psychometrics-manuals

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Del Pilar Salas-Zárate, M., López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., & Alor-Hernández, G. (2014). A study on LIWC categories for opinion mining in Spanish reviews. *Journal of Information Science*, *40*(6), 749-760. http://dx.doi.org/10.1177/0165551514547842

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

Dwyer, A., de Almeida Neto, A., Estival, D., Li, W., Lam-Cassettari, C., & Antoniou, M. (2021). Suitability of Text-Based Communications for the Delivery of Psychological Therapeutic Services to Rural and Remote Communities: Scoping Review. *JMIR mental health*, *8*(2), e19478. https://doi.org/10.2196/19478

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 2053951719860542. https://doi.org/10.1177/2053951719860542

Ji, Q., & Raney, A. A. (2020). Developing and validating the self-transcendent emotion dictionary for text analysis. *PLoS ONE*, *15*(9), Article e0239050. https://doi.org/10.1371/journal.pone.0239050

Kaviani, H., Pournaseh, M., & Golfam, A. (2005). Emotion versus Cognition Words in Persian Dictionaries. *Advances in Cognitive Sciences*, 7 (2), 29-37. (In Persian) http://icssjournal.ir/article-1-134-fa.html

Kaviani, H., Mousavi, A., & Golfam, A. (2007). A Study on Psychology-Related Words in Persian Dictionaries. *Language and Linguistics*, 3(5), 89-102. (In Persian) https://lsi-linguistics.ihcs.ac.ir/article_1606.html

Kilic, I. Y., & Pan, S. (2022, June). Incorporating LIWC in neural networks to improve human trait and behavior analysis in low resource scenarios. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4532-4539). https://aclanthology.org/2022.lrec-1.482

Koutsoumpis, A., Oostrom, J. K., Holtrop, D., van Breda, W., Ghassemi, S., & de Vries, R. E. (2022). The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the Big Five and the Linguistic Inquiry and Word Count (LIWC).Psychological Bulletin, 148(11-12), 843–868. https://doi.org/10.1037/bul0000381

Lake, B. M., & Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological Review, 130*(2), 401–431. https://doi.org/10.1037/rev0000297

Lin, Y., Yu, R., & Dowell, N. (2020). LIWCs the Same, Not the Same: Gendered Linguistic Signals

of Performance and Experience in Online STEM Courses. *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I, 12163*, 333–345. https://doi.org/10.1007/978-3-030-52237-7_27

Lopez, I., Quillivic, R., Evans, H., & Arriaga, R. I. (2019). Denouncing sexual violence: A cross-language and cross-cultural analysis of# MeToo and# BalanceTonPorc. In *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part II 17* (pp. 733-743). Springer International Publishing. https://dx.doi.org/10.1007/978-3-030-29384-0_44

Lyu, S., Ren, X., Du, Y., & Zhao, N. (2023). Detecting depression of Chinese microblog users *via* text analysis: Combining Linguistic Inquiry Word Count (LIWC) with culture and suicide related lexicons. *Frontiers in psychiatry*, *14*, 1121583. https://doi.org/10.3389/fpsyt.2023.1121583

Soltani, M. A., MamSharifi, P., Salehi Chegeni, R., Safari, A., Ershadi Manesh, S., Keshavarz, H., ... & Tavakoli, F. (2024). The development and psychometric properties of the persian linguistic inquiry and word count (P-LIWC): Emotions and cognitive processes categories. *Journal of Psychological Science*, *23*(137), 1-19. (In Persian). http://dx.doi.org/10.52547/JPS.23.137.997

MamSharifi, P., Sohrabi, F., Borjali, A., Zamanpour, E., HosseinSabet, F. (2026). The Psycho–Linguistic Analysis of Cognitive, Emotional, and Sexual Dimensions of Pornography in Social Networks: Instagram, Telegram, and X. *Journal of Psychological Science*. *25*(160), 1-18. (In Persian). http://dx.doi.org/10.66224/jps.25.160.1 \

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2019). "LIWC auf Deutsch": The development, psychometrics, and introduction of DE-LIWC2015. *PsyArXiv*, (a). https://doi.org/10.17605/OSF.IO/TFQZC

Park, C., Shim, M., Eo, S., Lee, S., Seo, J., Moon, H., & Lim, H. (2022). Empirical analysis of parallel corpora and in-depth analysis using LIWC. *Applied Sciences*, *12*(11), 5545. https://doi.org/10.3390/app12115545

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. http://hdl.handle.net/2152/31333

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J.W. (2011). La version française du dictionnaire pour le liwc: Modalités de construction et exemples d'utilisation [The French dictionary for LIWC: Modalities of construction and examples of use]. *Psychologie Française*, *56*(3), 145–159. https://doi.org/10.1016/j.psfr.2011.07.002

Sasahara, K., Okuda, S., & Igarashi, T. (2021). Text mining approach to quantify consumer psychology and behavior in COVID-19 pandemic. In *Proceedings of the Annual Conference of JSAI, JSAI2021*. https://confit.atlas.jp/guide/event/jsai2021/subject/1D3-OS-3b-04/detail

Utomoa, P. A., & Karyawatia, A. E. (2021). Sentiment analysis of tribal, religion, and race with LIWC. *Jurnal Elektronik Ilmu Komputer Udayana p-ISSN*, *2301*, 5373. https://jurnal.harianregional.com/jlk/full-64361

Windsor, L. C., Cupit, J. G., & Windsor, A. J. (2019). Automated content analysis across six languages. *PloS one*, *14*(11), e0224425. https://doi.org/10.1371/journal.pone.0224425

Zhao, N., Jiao, D., Bai, S., & Zhu, T. (2016). Evaluating the Validity of Simplified Chinese Version of LIWC in Detecting Psychological Expressions in Short Texts on Social Network Services. *PloS one*, *11*(6), e0157947. https://doi.org/10.1371/journal.pone.0157947