Journal of Psychological Science

Print ISSN: 1735-7462 Online ISSN: 2676-6639

Homepage: http://www.psychologicalscience.ir

The development and psychometric properties of the persian linguistic inquiry and word count (P-LIWC): Emotions and cognitive processes categories

Mohammad Ali Soltani[1], Peyman MamSharifi[2], Reza Salehi Chegeni[3], Ali Safari[4], Soudabeh Ershadi Manesh[5], Hamidreza Keshavarz[6], Majid Afshari[7], MohamadSajad Ghafouri[8], Mojtaba Mohammadi[9], Fateme Tavakoli[10]

1. M.A. in Assessment and Measurement, Allameh Tabataba'i University, Tehran, Iran. E-mail: mohalisoltani@gmail.com
2. Ph.D Candidate in Psychology, Department of Psychology, Allameh Tabataba'i University, Tehran, Iran. E-mail: peymanmamsharifi@gmail.com
3. Ph.D Candidate in Computer Engineering, 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. Assistant Professor, Department of Psychology, North Tehran Branch, Islamic Azad University, Tehran, Iran. E-mail: su_ershadi@yahoo.com
6. Ph.D Candidate in Political Sociology, Central Tehran Branch, Islamic Azad University, Tehran, Iran. E-mail: hkeshaavarz@gmail.com
7. M.A. Student of Psychology, Science and Research Branch, 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. of General Psychology, Imam Khomeini International University, Qazvin, Iran. E-mail: Tavakoli.fateme@hotmail.com

ARTICLE INFO

Article type:
Research Article

Article history:
Received 23 April 2024
Received in revised form 22 May 2024
Accepted 26 June 2024
Published Online 22 July 2024

Keywords:
persian linguistic inquiry and word count (P-LIWC),
natural languasge process,
cognitive processes,
emotions,
psychometrics,
text mining

ABSTRACT

Background: Linguistic Inquiry and Word Count (LIWC) are known as indicators of vocabulary in texts. Specifically, with the help of algorithms and computational processes, it can analyze elements such as emotions, cognitions, attitudes, vocabulary, language style and social communication in texts, and the lack of a Persian version was one of the most important reasons for conducting this research.

Aims: The purpose of this research was to develop and examine the psychometric features of the Persian Linguistic Inquiry and Word Count in categories of Affect and Cognitive Processes.

Methods: The present research method was descriptive and correlational. The method of developing categories of cognitive processes and affect consists of several stages. The development of the dictionary started with the initial translation of the words of the LIWC-22, and then, a huge corpus of Persian texts was analyzed to maximize the coverage of words of each category in the Persian language using text analysis methods. In the next step, the words were evaluated by psychologist judges, and then linguists determined the lemma of the approved words in order to distinguish different forms of the word. After completing these steps, in order to check the reliability of the dictionaries of cognitive processes and affect and their subcategories, Cronbach's alpha and Kuder-Richardson 20 were calculated, and then to check the external validity, the equivalence of the Persian Linguistic Inquiry and Word Count (P-LIWC) with the original English version of LIWC-22 was analyzed.

Results: The results showed that all categories of emotions and cognitive processes in the P-LIWC had a significant correlation with the final English version of LIWC-22 ($p<0.01$) and this tool has good validity and reliability.

Conclusion: It can be concluded that this software is used in a wide range of research fields and analysis of Persian texts and also can be used in the analysis of the texts and the treatment session of the people who refer to the psychological clinics.

Citation: Soltani, M.A., MamSharifi, P., Salehi Chegeni, R., Safari, A., Ershadi Manesh, S., Keshavarz, H., Afshari, M., Ghafouri, M.S., Mohammadi, M., & 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. 10.52547/JPS.23.137.1

Journal of Psychological Science, Vol. 23, No. 137, 2024
 DOI: 10.52547/JPS.23.137.1

✉ **Corresponding Author:** Mohammad Ali Soltani, M.A. in Assessment and Measurement, Allameh Tabataba'i University, Tehran, Iran.

E-mail: mohalisoltani@gmail.com, Tel: (+98) 9190280944

Extended Abstract

Introduction

Linguistic Inquiry and Word Count (LIWC) is a computational text mining tool used to analyze texts and linguistic content. This tool is known as a display of emotions and words in texts. Specifically, with the help of computational algorithms and processes, it can analyze elements such as emotions, attitudes, vocabulary, language style, and social communication in texts (Koutsoumpis et al., 2022). This tool is built on the basis of psychological and linguistic research and is used as a useful tool for better analysis and understanding of texts and writings in digital humanities, computational social sciences; opinion polling and other areas (Park et al., 2022). Linguistic Inquiry and Word Count operates by sorting words into distinct psychological and linguistic categories such as emotions, cognitive processes, social tendencies and even parts of conversational contexts. These categories help to understand the psychological and emotional tone of the text and provide insights about the author's mental state, personality and communication style (Pennebaker et al., 2001).

Its functions include providing statistical analyses, generating reports, and visual representations of linguistic patterns and emotional expressions in text that provide researchers with a comprehensive understanding of the emotional and cognitive aspects of communication (Lin et al., 2021). It also has dictionaries that include words related to emotions (including positive emotions, negative emotions, anxiety, anger and sadness) and cognitive processes (including insight, causal, Discrepancies, tentativeness, certainty, differentiation) and provides the possibility of detailed analysis of the content of the text. Researchers use Linguistic Inquiry and Word Count to discover linguistic patterns in social media content, analyze language in clinical settings, study differences in communication between genders, cultures, or age groups, and even predict psychological traits based on written language (Utomoa & Karyawatia, 2021). This software is constantly evolving and incorporates new dictionaries and features to increase its accuracy and use in different contexts, making it a versatile tool for text analysis and psychological research (Boyd et al., 2022).

This software has been developed in 14 international languages, including German (Meier et al., 2019), Serbian (Bjekić et al., 2014), Spanish (Del Pilar Salas-Zárate et al., 2014) and Chinese (Zhao et al., 2016). In general, Linguistic Inquiry and Word Count serve as a tool for quantifying and understanding the psychological and linguistic components of written text and provide valuable insights into the emotions, thoughts, and behavioral patterns of individuals or groups. In the present study, the categories of emotions and cognitive processes were investigated. These two classifications play a crucial role in assessing people's overall outlook towards diverse circumstances. By pinpointing these classifications, it becomes possible to enhance individuals' performance. Since the Persian Linguistic Inquiry and Word Count does not exist in Iran and its importance and application in various fields can be fruitful. As a result of this research, it sought to answer the question whether the Persian version of Linguistic Inquiry and Word Count (P-LIWC): emotions and cognitive processes categories has appropriate psychometric properties?

Method

The present research method was descriptive and correlational. The method of developing categories of cognitive processes and Affect consists of several stages. The development of the dictionary started with the initial translation of the words of the LIWC-22, and subsequently, to maximize the coverage of words of each category in the Persian language, a huge corpus of Persian texts was analyzed through text mining methods. In the subsequent phase, psychologist judges assessed the words, followed by linguists determining the lemma of the authorized words to discern various word forms. After completing these steps, in order to check the reliability of the dictionaries of cognitive processes and emotions and their subcategories, Cronbach's alpha and Kuder-Richardson 20 were calculated, and then to check the external validity, the equivalence of the Persian Linguistic Inquiry and Word Count (P-

LIWC) with the original English version of LIWC-22 was analyzed.

In the next part, the dictionary development process was done. This section provides an overview of the P-LIWC dictionary categories development process. Although this process had many steps and was somewhat recursive in nature, the overall process can be divided into 5 steps. These steps are designed based on the methods used in the localization of official non-English versions of LIWC and the method used in the construction of LIWC-22.

Step 1. Initial translation and preliminary collection of vocabulary

Step 2. Review, Scoring, Categorization and modification

Step 3. Development and expansion of the dictionary

Step 4. Evaluation of psychometric properties

Step 5. Refinement and revision

Steps 1 to 3 are related to the methodology section and steps 4 and 5 are related to the research findings section. This part was done in three stages, which will be fully explained in the following.

Step 1. Initial translation and preliminary collection of vocabulary

The purpose of this stage is to reach the desired foundation for expanding the vocabulary of the categories of cognitive processes, including the subcategories of insight, causal, discrepancies, tentativeness, certainty, and differentiation, and the categories of Affect, including the subcategories of positive emotions, negative emotions, anger, sadness, and anxiety. At this stage, first, all the words belonging to the categories of cognitive processes and emotions of the LIWC-22 dictionary and its subcategories were translated by an English translator. Based on the method used in LIWC-22 (Boyd et al., 2022). After translating the words of 13 cognitive and affect categories, each word was evaluated by 2 psychologist judges. The purpose of the initial evaluation of the words was to check the compatibility and suitability of the translated word with each of the 13 cognitive and emotional categories in terms of conceptual and cultural aspects regarding the Persian language. To do this, each of the words was given a score between 1 and 5, with score 5 indicating the highest degree of suitability and score 1 indicating the lowest degree of suitability of the word with the target category. After scoring and comparing the evaluations of the judges, the words with the highest degree of appropriateness were kept and the words whose translation had little appropriateness to the target category were changed by an English language expert to an equivalent corresponding to the relevant category.

Step 2. Review, Scoring, Categorization and modification

After the initial version of the Glossary of Cognitive Processes and Emotions was prepared, it was manually revised to identify and remove obvious conceptual inconsistencies or possible spelling errors. Then, the words were manually classified under the specified categories (insight, causal, discrepancies, tentativeness, certainty, differentiation, positive emotions, negative emotions, anxiety, anger and sadness). 4 expert psychologist judges checked all the obtained words one by one and gave one point to each word according to the appropriateness of each word with each of the intended categories. Therefore, at this stage, each of the words received 4 separate points. As the next step, the judges decided which words to eliminate. Remaining a word in one category or adding it to another category required approval by the majority of judges. In fact, each word was judged by 4 psychologists. If the majority of the judges gave a score to that word, the word remained in that dictionary, the words that did not reach the consensus of the judges were removed from the dictionary. On the other hand, for the words that the judges gave 2 marks for confirming that word and 2 marks for removing that word, group judging sessions were held and decisions were made about those words.

Step 3. Development and expansion of the dictionary

This step was done in four parts. First, the prepared dictionary was checked in terms of compatibility with the structure of the Persian language and its content and cultural characteristics. Due to the fact that the structural features of the Persian language are different from other languages, the method of doing it in the first part was determined based on the rules of terminology and syntax of the Persian language. The purpose of this section is to examine the

structural concordance of the constructed dictionary and the Persian language with dictionary-based text analysis methods. The purpose of the second part of this stage was to identify the most common words related to cognitive processes and emotions in Persian and add them to the dictionary. In order to add new words, the translated words were searched in dictionaries, encyclopedias and networks of Persian words. Synonymous words and terms were identified and added to the translated words. In the third part, the Word Similarity Analysis was done. With the aim of increasing the coverage and efficiency of the dictionary, the collection containing new words was reviewed again and the category or categories requiring further expansion were identified. At this stage, the expansion of the dictionary was completed. After finishing the judging in the fourth part, in order to recognize the different forms of the word by the computer, the lemmas of the words were determined by the linguists. The operational approach involved the linguist assuming the role of the computer and evaluating the target word based on the computer's ability to recognize it. The beginning of the desired word was considered as a lemma. The word will be registered as a lemma if the computer can recognize it and the derived words without any issues. In this case, the computer can recognize this word and the words that start with this form of the word in the Respective text. It is essential to elucidate that the quantity of words within the emotions category stood at 949 prior to any development, whereas the cognitive category contained 329 words. Subsequent to the development and expansion of the dictionary, these figures escalated to 4705 and 871 words for emotions and cognitive categories, respectively.

Results

Steps 4 and 5 are related to the findings of this research, in which the psychometric properties of this tool in Persian language are evaluated. In the following, these two steps will be fully explained.

Step 4. Evaluation of psychometric properties

The purpose of this step is to check the reliability and validity of the dictionaries of cognitive processes and emotions based on the methods used on different versions of LIWC. To calculate the reliability and validity of the dictionary, the method proposed by Pennebaker et al. (2015) and Boyd et al. (2022) was used. To measure the reliability of the dictionary, the internal consistency of each category was calculated. Due to the similarity and proximity of the words of the same category to each other, it is expected that with the increase in the repetition of a word in the text, the frequency of using other words belonging to that category will also increase in that text (Pennebaker et al., 2015; Boyd et al., 2022). In this study, a corpus of texts including more than 100 million texts of X social network (Twitter), Instagram, Telegram and official news texts including newspapers and online news agencies were used. These texts belonged to the years 2019 to 2023 and were collected by Lifeweb Company and provided to the development team. Also, this corpus included 4 billion and 800 million tokens. Therefore, with this method, an alpha coefficient was obtained for each of the dictionary categories in each of the texts. According to the type of calculations performed in this method and the nature of language categories, the alpha coefficient calculates the reliability of the categories much lower than the actual value. For this reason, in addition to Cronbach's alpha, Kuder-Richardson-20 formula was also used to check reliability and both scores were reported for each category.

In the next step, in order to check the external validity, the equivalence between the P-LIWC and the final English version of LIWC-22 was checked. In this step, two sets of texts were used for analysis. 200 texts were selected from various political, social, economic, psychology, sports, literature, biography and religion, environment, tourism, management, history, law, etc., each of which was between 700 and 1200 words. In this collection, 100 texts in Persian were translated into English, and 100 texts in English were translated into Persian. Then Pearson's correlation coefficient (r) was calculated as an index of equivalence between Persian and English versions. In fact, the more similar two versions are in terms of performance, the higher the correlation coefficients between their categories. The following table shows the correlation coefficients between the Persian and English versions. All classes of the Persian version with the final English version of LIWC-22 had a significant correlation at the level of $p < 0.01$.

Table 1. Cronbach's alpha coefficient and Kuder-Richardson 20 for all classes of cognitive processes and emotions

Category		Word count	Cronbach's alpha	Kuder-Richardson 20
cognitive processes	cognitive processes	871	0/69	0/97
	Insight	399	0/48	0/92
	Causal	312	0/43	0/89
	discrepancies	80	0/27	0/62
	tentativeness	151	0/30	0/66
	certainty	222	0/29	0/83
	differentiation	106	0/25	0/73
Emotions	Emotions	4705	0/73	0/96
	positive emotions	1642	0/62	0/94
	negative emotions	1839	0/63	0/92
	anger	707	0/42	0/79
	sadness	366	0/34	0/72
	anxiety	151	0/26	0/67

Table 2. Correlation coefficient between Persian and English version

Category		Pearson's correlation coefficient	Sig
cognitive processes	cognitive processes	0/67	p<0.01
	Insight	0/82	p<0.01
	Causal	0/62	p<0.01
	discrepancies	0/56	p<0.01
	tentativeness	0/78	p<0.01
	certainty	0/58	p<0.01
	differentiation	0/54	p<0.01
Emotions	Emotions	0/75	p<0.01
	positive emotions	0/81	p<0.01
	negative emotions	0/84	p<0.01
	anger	0/87	p<0.01
	sadness	0/68	p<0.01
	anxiety	0/92	p<0.01

As it is clear from Table 2, all categories of the P-LIWC with the final English version of LIWC-2022 had a significant correlation at the level (p<0.01).

Step 5. Refinement and revision

After completing steps 1 to 4, we partially repeated them recursively to identify any possible mistakes or oversights that may have occurred in the dictionary development process. Additionally, in an expert discussion round, decisions had to be made on how to deal with specific challenges unique to the analysis of automatic word count in Persian. This included, for example, conventions on how to treat Persian word lemmas and how to recognize them by the computer.

Conclusion

The aim of the current research was to develop and investigate the psychometrics of the categories of cognitive processes and emotions of the Persian Linguistic Inquiry and Word Count (P-LIWC). Through a series of systematic analysis, the cognitive and emotional categories of the P-LIWC dictionary, including words translated from the original English version and common words adapted from large Persian text collections, have sufficient internal consistency among different cognitive categories and also, the external validity between the Persian and English versions of LIWC-22 was confirmed by comparing the transcripts of the Persian and English texts. Therefore, the evidence shows that the Persian Linguistic Inquiry and Word Count (P-LIWC) is a powerful research tool for digital humanities and computational social science researchers to investigate psychological aspects in Persian texts. As a result, the categories of cognitive processes and emotions of the Persian Linguistic Inquiry and Word Count (P-LIWC) provide a better understanding of the psychological characteristics of Persian texts. This software is expected to serve as an important bridge between quantitative and qualitative research in Farsi, allowing researchers to gain multifaceted and deep insights into the data. We hope that this software will be used in a wide range of Persian text analysis research fields and will create more

innovations in digital humanities and computational social sciences.

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 role of each author: The first author is the executive and the responsible author of this research. The second author, manager of the psychology team; the third author, the manager of the data science team; the fourth author is the manager of the linguistics team. The fifth, seventh, eighth, and tenth authors were researchers from the psychology team, and the sixth author was a researcher from the data science team. The ninth author was the translator of the research.

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 تحقیق زبانی و شمارش کلمات (P-LIWC): طبقات هیجانات و فرآیندهای شناختی

محمدعلی سلطانی[1]✉، پیمان مام‌شریفی[2]، رضا صالحی‌چگنی[3]، علی صفری[4]، سودابه ارشادی‌منش[5]، حمیدرضا کشاورز[6]، مجید افشاری[7]، محمدسجاد غفوری[8]، مجتبی محمدی[9]، فاطمه توکلی[10]

۱. کارشناس ارشد سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران، ایران.

۲. دانشجوی دکتری روانشناسی، گروه روانشناسی، دانشگاه علامه طباطبائی، تهران، ایران.

۳. دانشجوی دکتری مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران.

۴. استادیار، گروه زبان‌شناسی، دانشگاه حضرت معصومه (س)، قم، ایران.

۵. استادیار، گروه روانشناسی، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران.

۶. دانشجوی دکتری جامعه‌شناسی سیاسی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران.

۷. دانشجوی کارشناسی ارشد روانشناسی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران.

۸. کارشناس ارشد روانشناسی بالینی خانواده، پژوهشکده خانواده، دانشگاه شهید بهشتی، تهران، ایران.

۹. کارشناس زبان و ادبیات انگلیسی، دانشگاه قم، قم، ایران.

۱۰. کارشناس ارشد روانشناسی عمومی، دانشگاه بین‌المللی امام‌خمینی (ره)، قزوین، ایران.

مشخصات مقاله

نوع مقاله:
پژوهشی

تاریخچه مقاله:
دریافت: ۱۴۰۳/۰۲/۰۴
بازنگری: ۱۴۰۳/۰۳/۰۲
پذیرش: ۱۴۰۳/۰۴/۰۶
انتشار برخط: ۱۴۰۳/۰۵/۰۱

کلیدواژه‌ها:
نسخه فارسی تحقیق زبانی و شمارش کلمات، پردازش زبان طبیعی، روان‌سنجی، فرآیندهای شناختی، هیجانات، متن‌کاوی

چکیده

زمینه: تحقیق زبانی و شمارش کلمات به‌عنوان یک نمایشگر واژگان در متون شناخته می‌شود. به‌طور خاص، به کمک الگوریتم‌ها و فرآیندهای محاسباتی، می‌تواند عناصری مانند هیجانات، شناخت‌ها، نگرش‌ها، واژگان، سبک زبانی و ارتباطات اجتماعی را در متون تحلیل کند و نبود نسخه فارسی از مهم‌ترین دلایل انجام این پژوهش بود.

هدف: هدف از انجام این پژوهش توسعه و بررسی ویژگی‌های روان‌سنجی نسخه فارسی تحقیق زبانی و شمارش کلمات: طبقات هیجانات و فرآیندهای شناختی بود.

روش: روش پژوهش حاضر نوصیفی و همبستگی بود. روش توسعه طبقات فرآیندهای شناختی و هیجانات از چند مرحله تشکیل شده است. توسعه فرهنگ لغت با ترجمه اولیه کلمات نسخه اصلی انگلیسی در سال ۱۴۰۲ شروع شد و پس از آن برای به حداکثر رساندن پوشش‌دهی کلمات هر طبقه در زبان فارسی با استفاده از روش‌های تحلیل متن، 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 آن‌ها توسط زبان‌شناسان به منظور تشخیص اشکال مختلف کلمه تعیین گردید. بعد از اتمام این مراحل به منظور بررسی پایایی فرهنگ لغت‌های فرآیندهای شناختی و هیجانات و زیر طبقات آن‌ها آلفای کرونباخ و کودر-ریچاردسون ۲۰ محاسبه گردید و سپس برای بررسی روایی بیرونی، هم ارزی نسخه فارسی تحقیق زبانی و شمارش کلمات (P-LIWC) با نسخه اصلی انگلیسی LIWC-2022 مورد تجزیه‌وتحلیل قرار گرفت.

یافته‌ها: نتایج نشان داد همه طبقات هیجان‌ها و فرآیندهای شناختی 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: با توجه به یافته‌های این پژوهش، می‌توان نتیجه گرفت که این نرم‌افزار در طیف گسترده‌ای از زمینه‌های پژوهشی و تجزیه‌وتحلیل متون فارسی کاربرد دارد و همچنین می‌توان در تحلیل متون جلسه درمان مراجعه‌کنندگان به کلینیک‌های روانشناسی نیز استفاده کرد.

استناد: سلطانی، محمدعلی؛ مام‌شریفی، پیمان؛ صالحی‌چگنی، رضا؛ صفری، علی؛ ارشادی‌منش، سودابه؛ کشاورز، حمیدرضا؛ افشاری، مجید؛ غفوری، محمدسجاد؛ محمدی، مجتبی؛ و توکلی، فاطمه (۱۴۰۳). توسعه و بررسی ویژگی‌های روان‌سنجی نسخه فارسی تحقیق زبانی و شمارش کلمات (P-LIWC): طبقات هیجانات و فرآیندهای شناختی. مجله علوم روانشناختی، دوره ۲۳، شماره ۱۳۷، ۱-۱۹.

مجله علوم روانشناختی، دوره ۲۳، شماره ۱۳۷، ۱۴۰۳. **DOI:** 10.52547/JPS.23.137.1

✉ **نویسنده مسئول:** محمدعلی سلطانی، کارشناس ارشد سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: mohalisoltani@gmail.com

تلفن: ۰۹۱۹۰۲۸۰۹۴۴

مقدمه

تحقیق زبانی و شمارش کلمات[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. 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). تحقیق زبانی و شمارش کلمات با دسته‌بندی کلمات به ابعاد روانشناختی و زبانی متعددی مانند هیجانات، فرآیندهای شناختی، تمایلات اجتماعی و حتی بخش‌هایی از گفتار عمل می‌کند. 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). همچنین دارای فرهنگ لغت‌هایی است که کلمات مربوط به هیجانات[5] (شامل هیجان مثبت[6]، هیجان منفی[7]، اضطراب[8]، خشم[9] و غم[10]) و فرایندهای شناختی[11] (شامل بینش[12]، علت[13]، مغایرت[14]، حدس و گمان[15]، قطعیت[16]، فرق‌گذاری[17]) را در برمی‌گیرد و امکان تجزیه‌وتحلیل دقیق محتوای متن را فراهم می‌کند. پژوهشگران از تحقیق زبانی و شمارش کلمات 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).

تحقیق زبانی و شمارش کلمات دارای عملکردهای مختلفی است که به تعدادی از آن‌ها اشاره می‌شود: طبقه‌بندی کلمات[18]: این نرم‌افزار کلمات را در ابعاد مختلف روانشناختی و زبانی مانند هیجانات، فرآیندهای شناختی، فرآیندهای اجتماعی و سبک‌های زبانی دسته‌بندی می‌کند. به‌عنوان مثال، می‌تواند کلمات را در گروه‌هایی مانند هیجانات مثبت یا منفی، مکانیسم های شناختی، تمایلات اجتماعی و غیره دسته‌بندی کند (باهگات و همکاران، ۲۰۲۲). بینش‌های روانشناختی[19]: با تحلیل فراوانی دسته‌بندی‌های کلمات خاص در متن، این نرم‌افزار بینش‌هایی را در مورد وضعیت روانشناختی نویسنده ارائه می‌دهد. این می‌تواند تمایلات عاطفی، الگوهای شناختی و رفتار اجتماعی را بر اساس زبان مورد استفاده آشکار کند (لیو و همکاران، ۲۰۲۳). تحلیل رفتاری[20]: این نرم‌افزار همچنین می‌تواند به پیش بینی یا درک الگوهای رفتاری بر اساس داده‌های متنی کمک کند. به‌عنوان مثال، می‌تواند به ارزیابی ویژگی‌های رفتاری مانند جهت‌گیری اجتماعی، نگرانی‌های شخصی، سبک‌های تفکر و غیره کمک کند (کیلیک و پن، ۲۰۲۲). تحقیقات و مطالعات[21]: محققان در زمینه‌های مختلف از جمله روانشناسی، زبان‌شناسی، جامعه‌شناسی، ارتباطات، علوم‌سیاسی و به‌طور کلی علوم‌انسانی دیجیتال و علوم اجتماعی محاسباتی و همچنین در پردازش زبان طبیعی[22]، از این نرم‌افزار برای تجزیه‌وتحلیل و تفسیر داده‌های متنی

1. Linguistic Inquiry and Word Count (LIWC)
2. digital humanities
3. computational social sciences (CSS)
4. polling
5. affect
6. positive emotions
7. negative emotions
8. anxiety
9. anger
10. sadness
11. cognitive processes
12. insight
13. causal
14. discrepancies
15. tentativeness
16. certainty
17. differentiation
18. classification of words
19. psychological insights
20. behavioral analysis
21. research and studies
22. natural language processing (NLP)

برای مطالعات دانشگاهی، نظرسنجی‌ها و آزمایش‌ها استفاده می‌کنند (پنه‌بیکر و همکاران، ۲۰۱۵). کاربردهای درمانی[1]: این نرم‌افزار در محیط های بالینی نیز کاربرد دارد. درمانگران و مشاوران ممکن است از آن برای تجزیه‌وتحلیل و ارزیابی محتوای عاطفی ارتباط نوشتاری یا شفاهی به‌عنوان بخشی از فرآیند درمانی استفاده کنند (دویر و همکاران، ۲۰۲۱). ۶- تجزیه‌وتحلیل محتوا[2]: این مورد را با خودکار کردن فرآیند طبقه‌بندی و کمی کردن جنبه‌های زبانی و روانشناختی متن تسهیل می‌کند که می‌تواند برای سازندگان 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، تحقیق زبانی و شمارش کلمات به‌عنوان ابزاری برای تعیین کمیت و درک مؤلفه‌های روانشناختی و زبانی متن نوشتاری عمل می‌کند و بینش‌های ارزشمندی را در مورد هیجانات، شناخت‌ها و الگوهای رفتاری افراد یا گروه‌ها ارائه می‌کند. در پژوهش حاضر طبقات هیجانات و فرآیندهای شناختی مورد بررسی قرار گرفتند. این دو طبقه از آن حیث دارای اهمیت هستند که بر اساس آن‌ها می‌توان نگرش‌های کلی افراد در خصوص وضعیت‌های مختلف را سنجید و با شناسایی آن‌ها عملکردهای افراد را بهبود بخشید. از آنجایی که نسخه فارسی تحقیق زبانی و شمارش کلمات در ایران وجود ندارد و اهمیت و کاربرد آن در حوزه‌های مختلف از جمله حوزه روان‌شناسی می‌تواند ثمربخش باشد، در نتیجه این پژوهش به دنبال پاسخگویی به این سوال بود که آیا نسخه فارسی تحقیق زبانی و شمارش کلمات: طبقات هیجانات و فرایندهای شناختی از ویژگی‌های روان‌سنجی مناسب برخوردار است؟

روش

الف) طرح پژوهش و شرکت‌کنندگان: مطالعه حاضر از نوع توصیفی و همبستگی بود. توسعه طبقات فرآیندهای شناختی و هیجانات از چند مرحله تشکیل شده است. توسعه فرهنگ لغت با ترجمه اولیه کلمات نسخه اصلی انگلیسی در سال ۱۴۰۲ شروع شد و پس از آن برای به حداکثر رساندن پوشش دهی کلمات هر طبقه در زبان فارسی با استفاده از روش‌های تحلیل متن (Text mining)، 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 آن‌ها توسط زبان‌شناسان به منظور تشخیص اشکال مختلف کلمه تعیین گردید. بعد از اتمام این مراحل به منظور بررسی پایایی فرهنگ لغت‌های فرایندهای شناختی و هیجانات و زیر طبقات آن‌ها آلفای کرونباخ و کودر-ریچاردسون ۲۰ محاسبه گردید و سپس برای بررسی روایی بیرونی، هم ارزی نسخه فارسی تحقیق زبانی و شمارش کلمات (P-LIWC) با نسخه اصلی انگلیسی LIWC-22 مورد تجزیه‌وتحلیل 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 ۱. ترجمه اولیه و جمع‌آوری مقدماتی واژگان؛ مرحله ۲. بازبینی، امتیازدهی، مقوله‌بندی و اصلاح؛ مرحله ۳. گسترش و توسعه فرهنگ لفت؛ مرحله ۴. ارزیابی ویژگی‌های روان‌سنجی و مرحله ۵. پالایش و تجدید نظر. مراحل ۱ تا ۳ مربوط به بخش روش و مراحل ۴ و ۵ مربوط به بخش یافته های پژوهش هستند. این بخش در سه مرحله انجام شد که در ادامه به صورت کامل در مورد هر کدام توضیح داده می‌شود.

مرحله ۱. ترجمه اولیه و جمع‌آوری مقدماتی واژگان

هدف از انجام این مرحله رسیدن به پایه و اساسی مطلوب برای گسترش واژگان طبقات فرایندهای شناختی شامل زیر طبقات بینش، علت، مغایرت، حدس و گمان، قطعیت و فرق‌گذاری و فرایندهای هیجانات شامل زیر طبقات هیجانات مثبت، هیجانات منفی، خشم، غم و اضطراب 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

[1]. therapeutic applications

[2]. content analysis

LIWC2001 (بژکیچ و همکاران، ۲۰۱۴؛ پیولا و همکاران، ۲۰۱۱) و روش استفاده شده در ساخت LIWC-22 (بوید و همکاران، ۲۰۲۲) طراحی شده است. در هنگام افزودن یا حذف کردن واژگان دیکشنری، علاوه بر منابع گفته شده، از روش‌های به کار رفته در ساخت واژه‌نامه‌ی هیجانات خود فرارونده (STED) (جی و رینی، ۲۰۲۰) نیز استفاده شد. STED یک واژه‌نامه برای شناسایی و تحلیل هیجانات خود فرارونده در متن است که بر اساس روش LIWC2015 (پنه‌بیکر و همکاران، ۲۰۱۵) ساخته شده است. همچنین با توجه به وجود تفاوت‌های ساختاری قابل توجه میان زبان‌های فارسی و انگلیسی، در جهت افزایش پوشش‌دهی و دقت دیکشنری، از روش‌های دیگری نیز در جمع‌آوری واژگان استفاده گردید.

در این مرحله، ابتدا تمامی واژگان متعلق به طبقات *فرایندهای شناختی و هیجانات* دیکشنری LIWC-22 و طبقات زیرمجموعه‌ی آن (شامل *بینش، علت، مغایرت، حدس و گمان، قطعیت، فرق‌گذاری، هیجان مثبت، هیجان منفی، اضطراب، خشم و غم*) توسط یک مترجم 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 بر اقدامات مشابه در ساخت دیکشنری‌های مشابه (جی و رینی، ۲۰۲۰) گرفته شده است.

پس از ترجمه کلمات ۱۳ طبقه شناختی و هیجانات، هر کلمه توسط ۲ داور روانشناس مورد ارزیابی قرار گرفت. هدف از ارزیابی اولیه کلمات، بررسی تناسب کلمه ترجمه شده با هریک از ۱۳ طبقه شناختی و احساسی 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.

جدول ۱ و ۲ تعداد کلمات ترجمه شده و مورد تأیید قرار گرفته در هر طبقه را نشان می‌دهند.

همانطور که در جدول ۱ دیده می‌شود، طبقه هیجانات ۹۴۹ کلمه و دارای دو طبقه کلی هیجانات مثبت و منفی است.

جدول ۱. تعداد کلمات طبقات هیجانات

	طبقه هیجانات					
	هیجانات	هیجان مثبت	هیجان منفی	خشم	غم	اضطراب
تعداد کلمه	۹۴۹	۴۲۴	۵۰۵	۱۶۳	۷۴	۹۰

جدول ۲. تعداد کلمات طبقات فرآیندهای شناختی

	طبقه فرایندهای شناختی						
	فرآیندهای شناختی	بینش	علت	مغایرت	حدس و گمان	قطعیت	فرق‌گذاری
تعداد کلمه	۳۲۹	۱۲۷	۵۹	۹	۶۶	۵۲	۲۱

همانطور که در جدول ۲ دیده می‌شود، طبقه فرایندهای شناختی ۳۲۹ کلمه و دارای شش طبقه بینش، علت، مغایرت، حدس و گمان، قطعیت و فرق‌گذاری است.

لازم به ذکر است، با توجه به وجود تفاوت‌های ساختاری قابل توجه میان زبان‌های فارسی و انگلیسی، در این مرحله سعی شد دقیق‌ترین واژه‌هایی که تطابق بالایی از لحاظ فرهنگی و زبانی با طبقات مذکور داشتند در دیکشنری بمانند. در مرحله بعدی به بازبینی و اصلاح طبقات پرداخته شد.

مرحله ۲. 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 فرآیندهای شناختی و هیجانات، 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 (بینش، علت، مغایرت، حدس و گمان، قطعیت، فرق‌گذاری، هیجان مثبت، هیجان منفی، اضطراب، خشم و غم) 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.

مرحله ۳. 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 می‌خواستیم مطمئن شویم که نرم‌افزار تحلیل متن می‌تواند واژه‌ها را به درستی شناسایی، طبقه‌بندی و تحلیل کند. به این منظور، کلمات

[1]. Corpora

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 فرایندهای شناختی و هیجانات 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 فرآیندهای شناختی و هیجانات 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 ۲ 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 پوشش‌دهی 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.

جدول ۳ و ۴ تعداد کلمات هر طبقه را پس از مرحله گسترش و توسعه نشان می‌دهد.

همانطور که در جدول ۳ دیده می‌شود، طبقه هیجانات پس از مرحله گسترش و توسعه دارای ۴۷۰۵ کلمه است.

جدول ۳. تعداد کلمات طبقات هیجانات پس از مرحله گسترش و توسعه

	طبقه هیجانات					
	هیجانات	هیجان مثبت	هیجان منفی	خشم	غم	اضطراب
تعداد کلمه	۴۷۰۵	۱۶۴۲	۱۸۳۹	۷۰۷	۳۶۶	۱۵۱

جدول ٤. تعداد کلمات طبقات فرآیندهای شناختی پس از مرحله گسترش و توسعه

	طبقه فرایندهای شناختی						
	فرآیندهای شناختی	بینش	علت	مغایرت	حدس و گمان	قطعیت	فرق‌گذاری
تعداد کلمه	۸۷۱	۳۹۹	۳۱۲	۸۰	۱۵۱	۲۲۲	۱۰۶

[1]. Word Similarity Analysis

[2]. Latent Semantic Analysis

همانطور که در جدول ۴ دیده می‌شود، طبقه فرآیندهای شناختی پس از مرحله گسترش و توسعه دارای ۸۷۱ کلمه است.

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 محاسباتی 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.

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 ۴ و ۵ 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 ۴. ارزیابی ویژگی‌های روان‌سنجی

هدف از این مرحله، بررسی پایایی و روایی دیکشنری‌های فرآیندهای شناختی و هیجانات 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 ۱۴۰۲ 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 ۵، ضریب آلفای کرونباخ و کودر-ریچاردسون ۲۰ برای همه طبقات فرآیندهای شناختی و هیجانات را نشان می‌دهد.

جدول ۵. ضریب آلفای کرونباخ و کودر-ریچاردسون ۲۰ برای همه طبقات فرآیندهای شناختی و هیجانات

طبقه		تعداد کلمات	آلفای کرونباخ	کودر-ریچاردسون ۲۰
فرآیندهای شناختی	فرآیندهای شناختی	۸۷۱	۰/۶۹	۰/۹۷
	بینش	۳۹۹	۰/۴۸	۰/۹۲
	علت	۳۱۲	۰/۴۳	۰/۸۹
	مغایرت	۸۰	۰/۲۷	۰/۶۲
	حدس و گمان	۱۵۱	۰/۳۰	۰/۶۶
	قطعیت	۲۲۲	۰/۲۹	۰/۸۳
	فرق‌گذاری	۱۰۶	۰/۲۵	۰/۷۳
هیجانات	هیجانات	۴۷۰۵	۰/۷۳	۰/۹۶
	هیجان مثبت	۱۶۴۲	۰/۶۲	۰/۹۴
	هیجان منفی	۱۸۳۹	۰/۶۳	۰/۹۲
	خشم	۷۰۷	۰/۴۲	۰/۷۹
	غم	۳۶۶	۰/۳۴	۰/۷۲
	اضطراب	۱۵۱	۰/۲۶	۰/۶۷

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 ۶ ضرایب همبستگی بین دو نسخه فارسی و انگلیسی را نشان می‌دهد.

جدول ٦. ضریب همبستگی بین نسخه فارسی و انگلیسی

طبقه		ضریب همبستگی پیرسون	سطح معناداری
فرآیندهای شناختی	فرآیندهای شناختی	۰/۶۷	p<۰/۰۱
	بینش	۰/۸۲	p<۰/۰۱
	علت	۰/۶۲	p<۰/۰۱
	مغایرت	۰/۵۶	p<۰/۰۱
	حدس و گمان	۰/۷۸	p<۰/۰۱
	قطعیت	۰/۵۸	p<۰/۰۱
	فرق‌گذاری	۰/۵۴	p<۰/۰۱
هیجانات	هیجانات	۰/۷۵	p<۰/۰۱
	هیجان مثبت	۰/۸۱	p<۰/۰۱
	هیجان منفی	۰/۸۴	p<۰/۰۱
	خشم	۰/۸۷	p<۰/۰۱
	غم	۰/۶۸	p<۰/۰۱
	اضطراب	۰/۹۲	p<۰/۰۱

همانطور که از جدول ۶ مشخص است، همه طبقات نسخه فارسی با نسخه نهایی انگلیسی LIWC-2022 دارای همبستگی معنادار در سطح (p< ۰/۰۱) بودند.

مرحله ۵. پالایش و تجدید نظر

پس از تکمیل مراحل ۱ تا ۴، ما تا حدی آن‌ها را به صورت بازگشتی تکرار کردیم تا هرگونه اشتباه یا نادیده گرفتن احتمالی را که ممکن است در فرآیند توسعه فرهنگ لغت رخ داده باشد، شناسایی کنیم. علاوه بر این، در یک دور بحث تخصصی، باید تصمیماتی در مورد نحوه برخورد با چالش‌های خاص منحصر به فرد برای تجزیه‌وتحلیل شمارش کلمات خودکار در زبان فارسی اتخاذ می‌شد. این شامل، برای مثال، قراردادهایی در مورد نحوه برخورد با لما کلمات فارسی و نحوه تشخیص آن‌ها توسط کامپیوتر بود. توجه داشته باشید که روان‌سنجی طبقات کلمات در طول مراحل پالایش به طرز چشمگیری تغییر کرده است. در مرحله اصلاح نهایی، یک داور که قبلاً فرهنگ لغت را ندیده بود، کل فرهنگ لغت را بررسی کرد. تغییرات پیشنهادی این داور برای انطباق در لیستی جداگانه قرار گرفت. در رتبه‌بندی نهایی کارشناسان، ۴ داور در مورد این کلمات نهایی تصمیم گرفتند و برای دستیابی به ویژگی‌های روان‌سنجی فرهنگ لغت نهایی، دوباره ارزیابی روان‌سنجی انجام گرفت.

بحث و نتیجه‌گیری

هدف پژوهش حاضر توسعه و بررسی روان‌سنجی طبقات فرایندهای شناختی و هیجانات نسخه فارسی تحقیق زبانی و شمارش کلمات P-LIWC بود. از طریق یک سری تجزیه‌وتحلیل سیستماتیک، طبقات شناختی و هیجانات نسخه فارسی فرهنگ لغت P-LIWC، شامل کلمات ترجمه شده از نسخه اصلی انگلیسی و کلمات متداول اقتباس شده از مجموعه‌های بزرگ متنی فارسی دارای سازگاری درونی کافی در بین طبقات مختلف شناختی و هیجانات است 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 تحقیق زبانی و شمارش کلمات P-LIWC یک ابزار تحقیقاتی قدرتمند برای پژوهشگران علوم انسانی دیجیتال و علوم اجتماعی محاسباتی است تا جنبه‌های روانشناختی موجود در متون فارسی را بررسی کنند.

P-LIWC پتانسیل زیادی برای تحقیقات آینده در عصر کلان داده و تجزیه‌وتحلیل 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، تحقیقات 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 #MeToo (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). شکی نیست که 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[1] (LLM) بر روی طبقات LIWC-22 در حال بررسی هستند و این شاید نقطه عطفی در کاربرد روش‌های سنتی و یادگیری ماشین به صورت همزمان باشد. به‌عنوان نتیجه، طبقات فرایندهای شناختی و هیجانات نسخه فارسی تحقیق زبانی و شمارش کلمات P-LIWC درک بهتری از ویژگی‌های روانشناختی متون فارسی ارائه می‌دهند. انتظار می‌رود این نرم‌افزار به‌عنوان پل مهمی بین تحقیقات کمی و کیفی در زبان فارسی عمل کند که به محققان اجازه می‌دهد تا بینش‌های چندوجهی و عمیقی را در مورد داده‌ها به دست آورند. امید است که این نرم‌افزار در طیف گسترده‌ای از زمینه‌های پژوهشی تجزیه‌وتحلیل متون فارسی استفاده شود و نوآوری‌های بیشتری در علوم انسانی دیجیتال و علوم اجتماعی محاسباتی ایجاد کند. از محدودیت‌های این پژوهش می‌توان به ترجمه انگلیسی به فارسی کلمات اشاره کرد که بعد از ترجمه برخی کلمات عیناً همان مفهوم کلمه خارجی را شاید نرسانند،

[1]. Larg language models

پس باید با دقت بالا و توسط افراد متخصص این کار انجام شود. همچنین تفاوت ساختاری زبان فارسی با انگلیسی موجب چالش‌هایی در تعیین لما و ریشه کلمات می‌شد که لازم بود الگویی خاص برای تعیین کلمات مختلف طراحی شود. پیشنهاد می‌شود که با توجه به نتایج پژوهش در مطالعات بعدی طبقات دیگر نرم‌افزار تحقیق زبانی و شمارش کلمات ازجمله طبقات اجتماعی و انگیزه‌ها توسعه یابند. همچنین می‌توان از این پژوهش در حوزه های مختلف از جمله علوم انسانی دیجیتال، علوم اجتماعی محاسباتی؛ افکار سنجی و حوزه‌های دیگر استفاده کرد.

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 علوم داده؛ نویسنده چهارم، مدیر تیم زبان‌شناسی هستند. نویسندگان پنجم، هفتم، هشتم و دهم از پژوهشگران تیم روانشناسی و نویسنده ششم پژوهشگر تیم علوم داده بودند. 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 می‌کنم از خداوند منان و تمامی همکاران عزیز و پرتلاشم 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

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

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. 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

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