Faculty of Mathematics and Computer Science

Natural Language Processing

Linguistic Inquiry and Word Count

Dinica Mircea

Department of Computer Science, Babeş-Bolyai University
1, M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania

Abstract

This research paper describes one of the tools in natural language processing (NLP). LIWC-22 (Linguistic Inquiry and Word Count) is a text analysis software that is based on years of research in computational linguistics and psychology. LIWC-22 stands out for its systematic language analysis capabilities, providing depth and precision in understanding how language reflects psychological states. Through an extensive evaluation of its processing modules, the LIWC-22 software is positioned as a accurate tool for both linguistic and psychological research, capable of interpreting the connection between language and feelings.

1. Introduction

In natural language processing, especially within the context of psychological research, the tool we choose to process and interpret the data is as important as the data itself. For this reason, our exploration of the dataset uses the latest version of a text analysis software, LIWC-22 (Linguistic Inquiry and Word Count). This tool represents the result of decades of research and development in the field of computational linguistics and psychology, designed to uncover different ways in which language reflects psychological states.

The tool is able to analyze language systematically, overcoming the complexities that early computer-based text analysis methods encountered [1].With LIWC-22, researchers have at their disposal a software tool that not only takes from previous versions but also incorporates the lat-

est advances in text analysis. Its expanded dictionary and enhanced software capabilities make it possible to analyze language samples with depth and precision. Whether one is interested in exploring the nuances of emotional expression, social connectivity, cognitive processes, or any other psychological dimension manifest in text, LIWC-22 offers platform for investigation.

In this section, we will explore the specific features of LIWC-22 that make it an invaluable tool for our research purposes, including its capabilities, reliability, and the ways in which it allows us to parse the subtle linguistic cues that signal varying psychological states.

2. The Processing Capabilities of LIWC-22

The Linguistic Inquiry and Word Count (LIWC-22) tool is a solution for processing and analyzing textual data within the domain of psychological research. This software, coupled with its comprehensive dictionary, bridges the gap between linguistic constructs and psychological theories, offering insights into the dimensions of language. Through a detailed overview of its primary and companion processing modules, we delve into how LIWC-22 serves as a tool for researchers aiming to uncover the meaning of text [1].

Upon analyzing texts, LIWC-22 evaluates the language used against its expansive dictionary, calculating the percentage of words within each text that align with specific categories. This process gives detailed metrics on the linguistic dimensions of the analyzed texts, which can be exported in various formats for further analysis.

Beyond its core functionality, LIWC-22 introduces several companion processing modules that enhance its analytical capabilities:

- Dictionary Workbench: Simplifying the creation of custom dictionaries, this module offers a user-friendly interface with built-in error checking. It facilitates the evaluation of custom dictionaries' psychometric properties, ensuring their effectiveness in research contexts [1].
- Word Frequencies and Word Clouds: These features assist in identifying the most common words within a dataset, providing visual word clouds for intuitive analysis of text samples [1].
- Topic Modeling with the Meaning Extraction Method: LIWC-22 incorporates the Meaning Extraction Method (MEM) for topic modeling, enabling researchers to uncover dominant themes and meanings within their datasets [1].
- Narrative Arc: This innovative module evaluates texts for narrative structures, offering insights into storytelling elements such as staging, plot progression, and cognitive tension [1].
- Language Style Matching (LSM): LSM analyzes the stylistic similarities between texts, offering metrics for comparing language use in various contexts, from individual communications to group dynamics [1].
- Contextualizer: Understanding the context of word use is vital. This module extracts words along with their surrounding text, allowing for a deeper examination of linguistic usage and implications [1].

- Case Studies: Tailored for in-depth analysis of individual texts, this module aggregates LIWC-22's capabilities to facilitate comprehensive study of specific documents or transcripts [1].
- Prepare Transcripts: Aiding in the preparation of conversation transcripts for analysis, this module streamlines the cleaning process, ensuring texts are optimized for LIWC analysis [1].

Together, these modules position LIWC-22 as a powerful tool for linguistic and psychological research, offering ways to explore and interpret the connection between language and feelings.

3. The Evolution of the LIWC-22 Dictionary

The LIWC-22 Dictionary is the last iteration of the software, embodying the fusion of linguistic constructs with psychosocial theories through an extensive lexicon. This core component, comprising over 12,000 words, word stems, phrases, and select emoticons, is organized into categories and subcategories designed to capture a wide array of feelings. This arrangement allows for a accurate analysis of text, offering insights into the psychological state, social relationships, and cognitive processes of individuals based on their word usage.

The LIWC-22 Dictionary has a hierarchical organization, where words are not only categorized but also interlinked across multiple dimensions. For instance, the word "cried" contributes to categories such as emotion, sadness, and past focus, illustrating the dictionary's complexity and depth. This structure enables LIWC-22 to provide a comprehensive analysis of text, reflecting various emotional and cognitive dimensions [1].

In the table 1 there are examples of the words linked to their coressponding categories and it can be seen that the words represent specifics of their category.

Social	**Culture**	**Lifestyle**	**Physical**
admiration	norwegian	free time	abs
company	nuclear	accomplish	aerobic
listener	online	real estate	ailment
locals	arabic	gaming	alcohol
refugee	political	qualify	deaf
reassure	phonecall	amusement	death
trust	person of color	god	kidney
tweets	racist	remodel	lactose
twins	bill of rights	art	salad
uncle	scanner	greed	ketogen
loyal	bots	rent	depressed
commitment	candidate	assignment	diabet
confess	opposition party	psychologist	sauna

Table 1. Examples of categories and its words in LIWC-22

The development of the LIWC-22 Dictionary represents a significant evolution from its predecessors, incorporating advances in computational linguistics and psychological research. The creation process involved multiple phases:

- Word Collection: Leveraging the foundation of the LIWC2015 dictionary, new words were generated for each category through a combination of expert input and comprehensive literature review [1].
- Judge Rating Phase: Words were qualitatively assessed by a panel of judges for their fit within each category, with disagreements resolved through in-depth analysis and consensus. Base Rate Analyses: Utilizing the Meaning Extraction Helper (MEH) tool, the frequency of dictionary words in a diverse corpus was evaluated to ensure relevance and applicability across various text samples [1].
- Candidate Word List Generation: Through statistical analysis and expert review, candidate words were identified for potential inclusion in the dictionary, ensuring a broad and relevant lexicon [1].
- Psychometric Evaluation: Each category underwent rigorous testing for internal consistency, with adjustments made to optimize the dictionary's psychometric properties [1].
- Refinement Phase: The entire process was iteratively refined to address any oversights and enhance the dictionary's accuracy and reliability [1].
- Addition of Summary Variables: New summary variables were introduced to provide additional analytical dimensions, based on cutting-edge research [1].

The LIWC-22 Dictionary has been significantly expanded to include not only traditional words but also numbers, punctuation, short phrases, and regular expressions. This expansion allows for the analysis of modern, informal communication styles found on social media and text messaging, incorporating "netspeak" and emoticons for a more comprehensive understanding of digital communication.

The dictionary's evolution reflects a balance between expert human judgment and sophisticated computational models, ensuring that LIWC-22 remains at the forefront of text analysis technology. With each iteration, LIWC has adapted to the changing landscape of language use, incorporating new categories and adjusting existing ones to better capture the psychological significance of language [1].

4. The Reliability of LIWC-22

The development of the Linguistic Inquiry and Word Count (LIWC-22) tool has consistently prioritized the establishment of a scientifically acuurate system, focusing on both reliability and validity. This commitment has guided each iteration of LIWC, with the aim of adapting to the dynamic nature of language use and leveraging the research of text-based data science. LIWC-22 represents the culmination of these efforts, integrating dictionaries with cutting-edge data analytics to offer a highly validated tool for text analysis.

The core of LIWC-22's reliability lies in the "Test Kitchen" corpus [Figure 1], a carefully chosen set of English language examples taken from many different places. This corpus serves two purposes: it is important in the selection of words for the LIWC-22 dictionary and plays a crucial role in assessing the dictionary's reliability and validity. The diversity of the Test Kitchen corpus

[Figure 1] ensure that LIWC-22's analyses are grounded in a realistic representation of language use across various contexts [1].

To capture the nature of language, the Test Kitchen corpus [Figure 1] was assembled from 15 distinct English language data sets, encompassing a wide range of communication forms, from blogs and emails to social media posts and movie dialogues. This comprehensive corpus consists of 15,000 texts, with each text sample reflecting the unique linguistic style of its author or authors. The selection process for these samples was designed to include a diverse representation of texts, ensuring a broad coverage of language use in daily life.

The construction of this corpus involved selecting 1,000 text samples from each of the 15 sources, with each text containing at least 100 words. For longer texts, a specific algorithm was employed to extract a continuous segment of 10,000 words, ensuring a manageable and consistent analysis size. In total, the Test Kitchen corpus [Figure 1] encompasses over 31 million words, providing a good foundation for the validation and refinement of the LIWC-22 dictionary [1].

Given the sensitivity and proprietary nature of some of the data sources, the Test Kitchen corpus, while invaluable for the development and testing of LIWC-22, cannot be made publicly available. This restriction shows the careful consideration given to privacy and ethical research practices in the compilation and use of the corpus. Nevertheless, the corpus's diverse and extensive dataset has been crucial in fine-tuning LIWC-22's dictionaries to reflect genuine language usage patterns .

Corpus	Description	Word Count *M* (SD)
Applications	Technical college admissions essays	1506 (501)
Blogs	Personal blogs from blogger.com	2144 (1920)
Conversations	Natural conversations	586 (510)
Enron Emails	Internal emails from Enron	316 (376)
Facebook	Facebook posts from mypersonality.com	2195 (2034)
Movies	Transcribed movie dialogue	6633 (2459)
Novels	Novels from Project Gutenberg	5703 (189)
NYT	New York Times articles	744 (494)
Reddit	Individuals' Reddit comments	1751 (1945)
Short Stories	Short stories	2977 (2211)
SOC	Stream of consciousness essays	656 (256)
Speeches	U.S. Congressional speeches	950 (1241)
TAT	Thematic Apperception Test, online website	326 (63)
Tweets	Collected tweets from individual accounts	4442 (2858)
Yelp	Restaurant reviews posted to Yelp	99 (1)
Overall mean		**2128 (2778)**

Fig. 1. The test Kitchen Corpus of 31 Million Words [1]

5. Assesing LIWC-22's Accuracy

The process of quantifying the reliability and validity of text analysis tools like LIWC-22 presents unique challenges, different from the conventional approaches used in psychological assessments. The differences between verbal behavior and structured questionnaire responses needs a nuanced approach to evaluating the properties of LIWC categories.

Unlike self-report questionnaires that measure emotions like anger, sadness or ability to cooperate, through multiple, similar questions to ensure internal consistency, natural language does not

conform to such repetitive patterns. In real-world communication, be it a social media post, an essay, or a conversation, individuals express a thought and then naturally progress to the next. This characteristic of verbal expression implies that the standards applied to language-based analyses must be adjusted to account for the unique dynamics of verbal behavior [1].

The evaluation of LIWC-22's reliability involves an innovative adaptation to the language's non-repetitive nature. Taking the LIWC-22 Anger scale as an instance, the scale encompasses 181 words and phrases associated with anger. Theoretically, the usage of one anger-related word in a text should correlate with the usage of other anger-related words within the same text. By analyzing how each of these words is employed across a selection of texts and calculating the intercorrelations among these word usages, LIWC-22's approach to determining internal consistency emerges[1].

Validating the numerous LIWC dimensions poses a significant and complex challenge. By their nature, LIWC's content categories appear to be directly relevant or face valid. Yet, the deeper question lies in determining the extent to which both personal and social psychological processes are mirrored in the use of language. For instance, the implications of using words related to "affiliation" at a high frequency raise questions about the user's social connections and needs. Are individuals using these words seeking more social interaction, or do they reflect a person's existing strong social ties? Additionally, it is important to consider whether the frequency of such language usage offers insights into or predictions about someone's social relationships and needs.

To compute these metrics, LIWC-22 employs two statistical methods: the Cronbach's alpha for continuous data, based on the percentage of total words, and the Kuder–Richardson Formula 20 (KR-20) for binary data [2], indicating the presence or absence of words. The application of Cronbach's alpha in the context of LIWC-22 encounters a problem due to the variable base rates of word usage within language categories. This variability can lead to underestimations of reliability when using traditional methods. Conversely, the Kuder–Richardson Formula 20 offers a more accurate reflection of a category's internal consistency by accommodating the binary nature of word presence, thus providing a better approximation of reliability in language analysis.

In 2021, there was a great amount of research combining text analysis and social and psychological behaviors, with more than 2,400 studies using LIWC to examine text. Findings from these studies, including those from the developers' own laboratories, reveal correlations between the affect or emotion categories detected by LIWC in texts and the authors' self-reported feelings. These correlations, although modest, show the tool's capability to capture psychological dynamics to a certain extent.

References

[1] 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 , 1–47.

[2] Kuder, G.F., Richardson, M.W., 1937. The theory of the estimation of test reliability. Psychometrika 2, 151–160.

Acknowledgement: This work is the result of my own activity, and I confirm I have neither given, nor received unauthorized assistance for this work. I declare that I used generative AI or automated tools in the creation of content or drafting of this document.