Challenges to the Development of Psycholinguistic Dictionary in the Hindi Language Intended for Personality Recognition

[1]Jayshri Patil, [2]Dr. Jikitsha Sheth

[1]Assistant Professor, [2]Associate Professor
[1]BMCCA, Bhagwan Mahavir University,
[2]SRIMCA, Uka Tarsadia University,

***Abstract*: In this digital era user expresses their feelings and emotions through social media or other online platforms. The user-generated content on such platforms is a key resource for inferring users' personalities. For the evaluation of human perception and personality, there is a need for sentiment analysis tools to infer users' personalities. In the field of personality recognition, growing interest in lexical re-sources provides distinct psychologically meaningful word categories. One such resource is Linguistic Inquiry and Word Count (LIWC) which was manually prepared in English and provides word counts in psychologically meaningful categories. In India many peoples survive in rural areas hence they have the problem of understanding and speaking the English language and increasing the Indian language content on social media. For the evaluation of hu-man perception and personality, it needs to get an insight into user-generated content in the Indian language using a psycholinguistic dictionary. In this paper, we present a literature review on the development of the psycholinguistic dictionary LIWC in various languages and the challenges in developing a psycholinguistic dictionary in the Hindi language.**

***Index Terms*: Personality, Psycholinguistic Dictionary, LIWC, Personality Recognition.**

I. INTRODUCTION

The natural language used by humans in daily life reveals their personality characteristics and social relationships. The words used by the users in natural language are the way by which psychologists try to understand human beings and their personalities. To make the analysis of such natural language there is a need for a tool that relates text with psychologically relevant categories. One such computerized text analysis tool is Linguistic Inquiry Word Count (LIWC) to link everyday language used by a human with a behavioral measure of personality [1]. The automatic text analysis is very important in the field of personality to know the user sentiments. This tool consists of dictionary and a software intended for tokenization and word counting. Each word in the dictionary relates to one or more psycholinguistic categories. LIWC2015 dictionary consists of approximately 90 psycholinguistic categories (e.g., the total number of words, the number of words per sentence), language constructs (e.g., auxiliary verbs, ad-verbs), psychological constructs (e.g., emotions, drives), relativity (e.g., motion, space, time), personal concerns (work, achievement, home), informal language (e.g., social media-specific words, swear words), and punctuation [1,2,4]. Currently translated version of the LIWC dictionary is available in Brazilian Portuguese, Chinese (Simplified), Chinese (Traditional), Dutch, French, German, Italian, Japanese, Norwegian, Romanian, Russian, Serbian, Spanish, Turkish, and Ukrainian.

The history of LIWC started in the early 1990s. The revised three versions have been launched LIWC2001, LIWC2007, and LIWC2015. The latest altered version of both LIWC2007 and LIWC2015 is LIWC2022 is recently launched [10]. To perform multilingual analysis in the digital era, many re-searchers have adopted the LIWC2015 dictionary and translated it from the English version to new languages [4]. The translation of LIWC into other language raise various difficulties and challenges because of cross-cultural languages. Every language is different from other languages because of morphological, syntax, and semantics structure. For overcoming these challenges there is a need to use se-mantic information coded in lexical resources of the target language. The following section describes the existing translation of LIWC English version into other languages.

II. RELATED WORK

This section gives a comprehensive review of an existing translation of LIWC into other languages. In the paper [3] authors have translated LIWC207 dictionary into Dutch language. They make an automated translation of the LIWC 2007 version and then compare it to the manually translated version of the dictionary. Authors have developed a translation pipeline to translate an English LIWC dictionary into Dutch, the pipeline procedure includes the steps: Initialisation, Wild-card expansion, translation, filtering, tagging, adding lemmas, Adding other word forms, Handling function words, Remove function words from content categories, Extending hierarchy, Wrap-up, and Manual correction. Authors have evaluated the automatic approach by com-paring results with the English lexicon, using a parallel corpus.

In the work [5], the authors have developed LIWCser Serbian dictionary from LIWC2007 English dictionary. It consists of 12103 words and word stems organized into 65 categories. Serbian language-specific characteristics and culture have been incorporated into the dictionary. It is larger in the number of words than English (4500), Dutch (6568), and Spanish (7515), but smaller than French (39230) [5]. The development of LIWCser comprises several phases. First, translation of all English dictionary words, and inclusion of synonyms, antonyms, and jargon words. For defining Linguistic categories words authors have utilized word-lists grammatical categories given in Serbian grammar book. Then applied appropriate inflections to all the words. In next phase classified all words into categories defined by LIWC2007 dictionary. In the final phase, two different judges re-viewed all word with related categories and further enhanced the dictionary with some culturally specific words.

The researchers [7] have developed Ro-LIWC2015 Romanian Translation of the LIWC2015 Dictionary. The construction of Ro-LIWC2015 comprises various steps. First, the English dictionary 6539 words were equally allocated to six translators and obtained a first draft of the translated Romanian dictionary. This draft contained a maximum of five synonyms for every English word, without any adjustments to the categories. With the periodic discussion of translators, these issues were resolved. In the further phase of the development of Ro-LIWC2015, the first author improved all the translations, following the same procedure as in the first step. Then every word was assigned the related categories according to the Romanian grammar and semantics while keeping the duplicates. The final version of Ro-LIWC2015 contains 47,825 entries. For evaluating Ro-LIWC2015 linguistic differences in different corpora, authors have analyzed posts from help-seeking forums for anxiety, depression, and health issues.

The work in [8] has explored a triangulation-based semi-automatic approach for constructing Catalan dictionary from the LIWC English dictionary. Authors have utilized Romance languages dictionaries as well as original English LIWC2001 dictionary and performed multi-lingual translations. Then automatically align different source words of each Catalan word. Evaluation of translations from different dictionaries allows to identify common words in different languages and the issue of assigning correct categories to the words has been re-solved.

III. Challenges to the development of Psycholinguistic Dictionary in the Hindi Language

LIWC is a text analysis tool developed by social psychologist James Pennebaker and his team at the University of Texas [1]. It calculates the degree to which any text uses psycholinguistic features. LIWC gives an efficient technique for analyzing emotional, cognitive, and structural components present in user-generated content [1,2]. In India many peoples live in rural areas hence they have the problem of understanding and speaking the English language. They express their emotions and thoughts in Indian languages. For the evaluation of hu-man perception and personality, it needs to get an insight into user-generated content in the Indian language specific to the Hindi language.

The Hindi language is resource-scarce and morphologically rich so a lot of information expresses in words [9]. The process of translating LIWC is not straightforward since every language has precise gram-mar rules and semantics. The biggest challenge is to decide what translations need to implement and word variations in different languages and which psychological categories relate to each word after translation. The major challenges involved in language inconsistencies such as translation error, changes in meanings due to translation, variation in verb tenses and part of speech tag, distinguishing between masculine and feminine words, and articulating words. There is a need to manually assign the correct POS tag to the translated dictionary words. The English language word can correspond to more than one word in the Hindi language. Word mapping is a tough task and re-quires manual efforts because of resources scarce in the Hindi language [11].

Another challenge is verb tenses and the meaning of words is dependent on their phrase in the Hindi language, in that case, there is a need to process word phrases instead of a single word.   Hindi language verb tense depends on its auxiliary verb and postposition, so we need to count words in Past Focus, Future Focus, and Present Focus categories at the time of processing text. For example In the translation process, while translating the LIWC dictionary word ‘depend’, and ‘de-pended’ both words translated into the same word ‘निर्भर’(Nirbhar) in Hindi, and the word ‘depends’ and ‘depending’ translated into the same phrase ‘निर्भर करता है’(Nirbhar Karta Hai). It becomes difficult to distinguish the tenses and gender forms of the words. In the translation process, LIWC dictionary English words were translated into the Hindi phrase because of Morphological variations like ‘सिर घूमना’(Sir Ghumana), ‘चक्कर आना’(Chakkar Aana), ‘दिमाग चाटना’(Dimmag Chatna). There is a high possibility of losing context information which leads to a translation error. Also, the translation of English words returns duplicate words for the different translations. For the same root word in the English language, there can be many words in the Hindi language with varying information for the tense, gender, and person. The Hindi language is morphologically rich where auxiliary words add with the verb for extra information concerning tense, gender, and person.

Conclusion

LIWC is an important text analysis tool for the multilingual analysis task to link the words daily used by human to their behavior and personality. For analyzing the user-generated content in different languages researchers have translated LIWC into a different language. In this paper, we presented a literature review on different translations of LIWC and also discuss the challenges of developing a psycholinguistic dictionary using LIWC in the Hindi language. The Hindi language is morphologically rich so there is a need for manual work to add Hindi language-specific words to the dictionary and need to process word phrases instead of a single word to retrieve correct verb tenses and gender information.

References

[1] Yla R. Tausczik1 and James W. Pennebaker, “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods”

[2] Pennebaker, James W., et al., The development and psychometric properties of LIWC2015. 2015

[3] Van Wissen, Leon, and Peter Boot, An electronic translation of the LIWC Dictionary into Dutch." Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference. Lexical Computing, 2017.

[4] Dudău, Diana Paula, and Florin Alin Sava. "Performing multilingual analysis with Linguistic Inquiry and Word Count 2015 (LIWC2015). An equivalence study of four languages." Frontiers in Psychology 12 (2021): 2860.

[5] Bjekić, Jovana, et al. "Psychometric evaluation of the Serbian dictionary for automatic text analysis: LIWCser." Psihologija 47.1 (2014): 5-32.

[6] Piolat, Annie, et al. "The French dictionary for LIWC: Modalities of construction and examples of use." Psychologie Francaise 56.3 (2011): 145-159.

[7] Dudău, Diana Paula, and Florin Alin Sava. "The development and validation of the Romanian version of Linguistic Inquiry and Word Count 2015 (Ro-LIWC2015)." Current Psychology (2020): 1-18.

[8] Massó, Guillem, et al. "Generating new LIWC dictionaries by triangulation." Asia Information Retrieval Symposium. Springer, Berlin, Heidelberg, 2013.

[9] Kumar, Yaman, et al. "BHAAV-A Text Corpus for Emotion Analysis from Hindi Stories."

[10] Boyd, Ryan L., et al. "The Development and Psychometric Properties of LIWC-22." (2022).

[11] Arora, Piyush. "Sentiment analysis for hindi language." MS by Research in Computer Science (2013).