Current Psychology (2022) 41:3597–3614
https://doi.org/10.1007/s12144-020-00872-4

The development and validation of the Romanian version of Linguistic Inquiry and Word Count 2015 (Ro-LIWC2015)

Diana Paula Dudău[1] · Florin Alin Sava[1]

Published online: 24 June 2020

Abstract

Today, performing automatic language analysis to extract meaning from natural language is one of the top-notch directions in social science research, but it can be challenging. Linguistic Inquiry and Word Count 2015 (LIWC2015; Pennebaker et al. 2015) is one of the most versatile, yet easy to master instruments to transform any text into data, meeting the needs of psychologists who are not usually proficient in data science. Moreover, LIWC2015 is already available in multiple languages, which opens the door to exciting intercultural quests. The current article introduces the first Romanian version of LIWC2015, Ro-LIWC2015, and thus, contributes to the line of research concerning multilingual analysis. Throughout the paper, we describe the challenges of creating the Romanian dictionary and discuss other linguistics aspects, which could be useful for new adaptations of LIWC2015. Also, we present the results of two studies for assessing the criterion validity of Ro-LIWC2015. The first study focuses on the consistency between the Romanian and the English dictionaries in analyzing a corpus of books. The second study tests whether Ro-LIWC2015 can acquire linguistic differences in contrasting corpora. For this purpose, we analyzed posts from help-seeking forums for anxiety, depression, and health issues, and leveraged supervised learning to address several classification problems. The selected algorithm allows feature ranking, which facilitates more thorough interpretations. The linguistic markers extracted with Ro-LIWC2015 mirrored a number of disorder-specific features of depression and anxiety. Given the obtained results, this research encourages the use of Ro-LIWC2015 for hypothesis testing.

Keywords LIWC2015 · Text analysis · Text mining · Content analysis · Machine learning · Mental health

Introduction

The Rationale of the Current Research

Language-based communication is one of the most valuable gifts that evolution has offered to humans. Paradoxically, language not only defines us as a species (Harari 2014) but also differentiates us as individuals (e.g., Pennebaker and King 1999). We use it to give shape to our inner processing and build bridges between ourselves and the outside world. These bridges are paved not only with explicit contents (i.e., the open meaning of the words) but also with more subtle, hard-to-control features such as the grammatical structure of the message. Both types of linguistic components can mirror parts of who we are – thoughts, feelings, attitudes, motivations, etc. (e.g., Settanni et al. 2018). The question is: How could social scientists acquire such valuable insights and use them cross-culturally, considering that language is so unstructured and different around the world and that the social science curriculum usually does not include courses in programming and advanced data processing?

The history of using language as a vehicle towards a deeper understanding of humans started way before 1961, the year when the Webster's Dictionary of English Language coined the term *content analysis*, given that some empirical quests into the meaning of linguistic contents have been noticed as far as 400 years ago, in theology (Krippendorff 2004). Traditionally, content analysis has been performed manually by raters who must follow a set of coding rules, depending on the scope of their inquiry (e.g., Drisko and Maschi 2016).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12144-020-00872-4) contains supplementary material, which is available to authorized users.

✉ Florin Alin Sava
florin.sava@e-uvt.ro

[1] Department of Psychology, West University of Timisoara, 4 Vasile Pârvan Blvd., 300223 Timişoara, Romania

Nowadays, human coding remains an indispensable research method to extract meaning from natural language but is far from being regarded as an optimal solution, considering some of its limitations, such as the difficulty of achieving inter-rater agreement when the analyzed content is very broad or personal, or the fact that it can be extremely time-consuming and expensive (Tausczik and Pennebaker 2010). Furthermore, the technological advances of the last three decades gave birth to a fast-growing repository of written natural communication, which created the urge to take traditional content analysis to a new level (e.g., Piryani et al. 2017).

Thus, researchers have started to combine cutting-edge computer science tools and techniques, with knowledge from psychology or other social science, to gain insights into human thinking, emotions/affect, and behavior, from natural language (Mäntylä et al. 2018). There are several approaches for automatic language analysis, ranging on a spectrum from hand-driven closed-vocabulary methods, which include manual and crowd-sourced dictionaries, to more data-driven and open-vocabulary methods like derived dictionaries, topics, or words and phrases (for a review and critical analysis, see Schwartz and Ungar 2015). Both closed- and open-vocabulary approaches bring advantages, and disadvantages and, ideally, researchers should be able to apply both depending on their research questions and resources. However, the open-vocabulary approach, as opposed to the closed-vocabulary approach, does not suit datasets of any size and is unreachable for psychologists who do not work in interdisciplinary teams or who are not self-taught programmers and data scientists (Kern et al. 2016).

The goal of the current paper is to present Linguistic Inquiry and Word Count 2015 (LIWC2015; Pennebaker et al. 2015), one of the most popular and easy-to-use computer-based language analysis tools for social science research, and provide the first version for the Romanian language. LIWC2015 is a closed-vocabulary approach tool. With its intuitive software and variety of grammatical and psychological components selected and refined rigorously, LIWC2015 meets the psychologists' needs for a practical and objective solution to manage even large amounts of unstructured linguistic data, irrespective of the research topic. In this regard, LIWC2015 and its previous versions have received growing attention and have been used in 593 papers indexed in Web of Science by the middle of April 2020, according to our search. Almost half of them, namely 288 of the total, were published in the last three years and covered different subdisciplines of psychology, among other domains such as communication, computer science, linguistics, or psychiatry, to name a few. The top psychology subfields in which LIWC2015 or its previous versions were recently used were social psychology (32 papers), multidisciplinary psychology (24 papers), experimental psychology (21 papers), and clinical psychology (18 papers).

Furthermore, LIWC2015 could enable researchers to pursue new and exciting intercultural quests, since it has already been translated into Dutch (van Wissen and Boot 2017), Ukrainian (Zasiekin et al. 2018), German (Meier et al. 2018), Brazilian Portuguese (Carvalho et al. 2019), and Chinese (e.g., Huang, Lin, Seih, Lin, & Lee, n. d.). With the expansion of the digital universe and the realization that most tools for computerized text analysis were developed in English, the problem of how to perform multilingual analysis has gained increased interest (e.g., Balahur and Perea-Ortega 2015). Likewise, given the popularity and versatility of LIWC2015, new language versions will probably emerge soon. Our paper contributes to this line of research regarding multilingual analysis by presenting two validation studies for the first Romanian version of LIWC2015 (Ro-LIWC2015).

Whilst the first study follows the common procedure for testing the equivalence between the English dictionary and a new language version, the second study leverages supervised learning to address several classification problems increasingly difficult. These problems culminate with distinguishing the language of depression from that of anxiety, given the cognitive and linguistic profiles that tend to characterize these disorders (e.g., Hendriks et al. 2014; Thorstad and Wolff 2019). Computing the classification accuracy might be more informative for our hypotheses testing than a classical comparison-type problem, especially because depression and anxiety are two highly comorbid conditions (e.g., Gorman 1996; Kessler et al. 2015). Moreover non-traditional statistics might be more suitable in a validation context of this sort, given that the LIWC2015 dictionary contains tens of components. The algorithm that we employed allows variables to simultaneously enter into the model, which reduces the accumulation of type-one error specific to repeatedly using the *t*-test or other classical procedure (e.g., Field 2018). Also, it creates a hierarchy of features, which facilitates a better understanding of the data and more thorough interpretations.

In the remainder of the introductory section, we will cover an overview of LIWC2015 and provide background information on the process of obtaining the Romanian equivalent of LIWC2015. Then, we will discuss the validation strategies that we applied.

LIWC2015 as a Valuable Research Tool

Linguistic Inquiry and Word Count (LIWC) is a lexicon and a software solution developed to enable researchers to automatically extract various psychological and style characteristics of any piece of text. The first version of LIWC was released in the early 1990s as part of Pennebaker, Francis and their collaborators' quest for understanding why writing or talking about negative life experiences can lead to improvements in physical and mental health (Pennebaker and Graybeal 2001; Pennebaker et al. 2015). From the outset, the LIWC program

comprised a dictionary and a module for text processing. Since then, LIWC has been modified three times – LIWC2001 (Pennebaker et al. 2001) and Francis 1996), LIWC2007 (Pennebaker et al. 2007), and LIWC2015 (Pennebaker et al. 2015). Each new version has brought improvements to both the dictionary and software. However, the latest version, LIWC2015, which is also the focus of the current paper, is significantly different from the previous versions since both components have been rebuilt, rather than updated (Pennebaker et al. 2015).

The creation of the default LIWC2015 dictionary was a laborious process of several years, which led to a list of 6549 words, word stems, and emoticons assigned to approximately 90 higher- and lower- level categories, based on psychometric standards (for a thorough presentation of each development stage, see Pennebaker et al. 2015). The validity and reliability methods that determined the composition of LIWC2015 are one of the main reasons why LIWC2015 is a powerful resource (Boyd 2017). Furthermore, whilst other dictionary-based tools are more specialized, LIWC2015 covers a variety of features, including four structural linguistic dimensions, 21 parts of speech and other function words, 41 categories with psychological connotation, six types of personal concerns, five forms of informal language, and four summary variables (analytical thinking, clout, authenticity, and emotional tone). The summary variables are not available for translation; they remain unique features of the English version. The comprehensive list of LIWC2015 categories displayed hierarchically, with examples, is provided by Pennebaker et al. (2015).

In contrast, other well-known tools, such as Affective Norms for English Words (ANEW; Bradley and Lang 1999), SentiStrength (Thelwall et al. 2010), SentiWordNet (Baccianella et al. 2010), OpinionFinder (Wilson et al. 2005), or General Inquirer (Stone et al. 1966) are more limited. For instance, ANEW measures only three emotional dimensions: pleasure, arousal, and dominance (Bradley and Lang 1999). Likewise, SentiStrength is a dictionary of words related to emotions, designed to extract positive and negative sentiment strength (Thelwall et al. 2010). SentiWordNet also focuses mostly on the polarity of words; it extracts three features: positivity, negativity, and neutrality/objectivity (Baccianella et al. 2010). OpinionFinder is slightly different because it analyses the subjectivity of textual data on four components, three of which do not target sentiment (Wilson et al. 2005). General Inquirer is more similar to LWC2015 because it covers a wider range of linguistic features, including two valence categories, Osgood semantic dimensions, words referring to pleasure, pain, virtue, and vice, language associated with particular institutions, references to places and objects, motivation-related words, cognitive orientation, and others.[1] However, General Inquirer is much less preferred than LIWC2015 – a search we conducted in Web of Science yielded only 19 papers referring to this tool and indexed in the last three years.

LIWC2015 is regarded as a closed-vocabulary approach to language analysis, which is a more feasible alternative to the open-vocabulary approach (Schwartz and Ungar 2015). As opposed to the open-vocabulary methods, LIWC2015 is accessible even for people with no background in computer science or data science. Also, like any other closed-vocabulary approach tool, it can be implemented even on samples with regular sizes of tens to hundreds of participants (Schwartz and Ungar 2015). These advantages are possible because the closed-vocabulary approach typically consists of using a piece of software to compare the linguistic inputs with a predefined list of items.

The LIWC2015 software supports various machine-readable formats of the input text and demonstrates flexibility in the options that the user can choose to investigate the linguistic contents of interest. By operating a user-friendly menu, the researcher can instantly compare each target word of each uploaded text file, with the dictionary words. A target word is part of the text introduced in the software for analysis, whereas a dictionary word belongs to the LIWC2015 dictionary. Every time the software finds a match, an item for the category or categories attached to the dictionary word is counted. Moreover, as the target file is crossed, other structural elements of the text such as punctuation or the total number of words are also recorded (Pennebaker et al. 2015). Hence, the LIWC2015 processor acts like a tokenizer and word counter; it can calculate frequencies adjusted by the total number of words and display them as percentages. Although the word counting approach can occasionally lead to incorrect classifications since a system like LIWC2015 could not detect sarcasm or semantic nuances, it is generally efficient because people naturally tend to express themselves using words grouped into meaningful clusters (Boyd 2017). Thus, usually, if a target word is misclassified, other related words would compensate for the same dictionary category.

Since its development, LIWC has been used as a research tool in various contexts, leading to exciting results regarding language use and individual differences, mental health, or social processes (for a representative review, see Boyd 2017). To give a flavor of its diverse applications, we would mention several examples coming from different areas. One of them is the study of Kleim et al. (2018) who managed to predict post-trauma adjustment based on the linguistic features of victims' narratives. Also, Bond et al. (2017) analyzed the language used in the 2016 US presidential debates and

[1] The General Inquirer categories can be seen at http://www.wjh.harvard.edu/~inquirer/homecat.htm

identified the differences between truthful and untruthful statements. Likewise, Scheuerlein et al. (2018) showed how language reflected the changes in the transformational leadership qualities of CEOs during the financial crisis. In the public health domain, for instance, Faasse et al. (2016) investigated the language of pro- and anti-vaccination comments and provided an example of how LIWC2015 could be useful in detecting health perceptions.

The Romanian Translation of the LIWC2015 Dictionary

Over time, LIWC dictionaries have been translated into multiple languages, including Spanish (Ramírez-Esparza et al. 2007), French (Piolat et al. 2011), German (Wolf et al. 2008; Meier et al. 2018), Dutch (Boot et al. 2017; van Wissen and Boot 2017), Brazilian-Portuguese (Balage Filho et al. 2013; Carvalho et al. 2019), Chinese (Huang et al. 2012, n.d.), Serbian (Bjekić et al. 2012), Italian (Agosti and Rellini 2007), and Russian (Kailer and Chung 2011). There was also an attempt to translate Ro-LIWC2001 made available by Fofiu (2012), but it was never validated, nor updated to meet the particularities of LIWC2015. Our project is the first attempt to build a Romanian version of LIWC2015 and to test its validity.

The process of translating tools like LIWC is not straightforward since every language has specific grammar rules and semantics (e.g., Levshina 2016; Patard 2014) that need to be accounted for in order for the software to reveal accurate results. The biggest challenge is to decide what translations and word variations to include in the dictionary and what categories to attach to specific words when there are language inconsistencies such as changes in meanings due to translation, or different ways to form verb tenses, distinguishing between masculine and feminine words, articulating words, or dealing with diacritics. Other authors discussed such adaptation issues in the context of developing, for example, the Spanish LIWC2001 (Ramírez-Esparza et al. 2007), the French LIWC2007 (Piolat et al. 2011), or the Dutch LIWC2007 (Boot et al. 2017). If we analyze the existent LIWC versions, we notice that a different dictionary has emerged with every translation. In this regard, based on the files downloaded from the *dictionaries.liwc.net* webpage, the Spanish LIWC2001 (Ramírez-Esparza et al. 2007) contains 12,656 words, the French LIWC2007 (Piolat et al. 2011) contains 39,164 words, the Italian LIWC2007 (Agosti and Rellini 2007) contains 5153 words, and the Dutch LIWC2007 (Boot et al. 2017) contains 11,091 words. However, even though LIWC versions differ in length, they tend to generate results consistent with those obtained with the English version, and also to show good validity, as shown in the papers dedicated to presenting them, which we have already cited. In other words, translation challenges tend not to be a significant obstacle.

The process of developing the Romanian LIWC2015 took one year and a half and involved three main steps. First, the 6539 words of the English dictionary were equally assigned to six translators, and a first draft of the Romanian dictionary was obtained. This first draft contained up to five synonyms for every English word, without any adjustments to the categories. The translators held periodic meetings to discuss the problems they encountered throughout the process, how to solve them, and whether the translation procedure should be refined. Each word was translated from English to Romanian using several dictionaries. In the second phase of the development of Ro-LIWC2015, the first author revised all the translations, following the same procedure as in the first step. At this point, every word was assigned the appropriate categories according to the Romanian grammar and semantics while keeping the duplicates. Finally, all files containing the second Romanian draft were copied in a single file. Then, the duplicates were marked automatically using a function in Microsoft Excel and removed manually. If the duplicates had different categories, those categories that stood out were assessed in terms of whether they should be kept or not, retaking the same steps as before. Specifically, we rechecked the definitions of the LIWC2015 words and their translations using several dictionaries and relying on our grammar and semantics knowledge as native Romanian speakers. In general, the categories were merged, given that this step was more of a chance to detect any mistakes that might have slipped. The translation protocol had to include several specific rules derived from the language differences between English and Romanian. More details about the translation procedure are available in Supplementary Material 1. The final Ro-LIWC2015 contains 47,825 entries altogether, but not all of these entries represent unique words because some words were spelled in two different forms, with or without diacritics.

One of the advantages of using a dictionary with a higher number of entries is that the researchers could detect the meaning of more words from the input text according to the predefined labels of the dictionary. For instance, as Meier et al. (2018) showed, the German version of the dictionary, DE-LIWC2015, which comprises 18,000 words and 77 categories, captured 87.84% of the total words in the analyzed text. In contrast, the German version of LIWC2001 (Wolf et al. 2008), which contains only 7598 words and 68 categories, detected only 70% of the same input text (Meier et al. 2018).

Overview of the Current Research

This paper proposes two main approaches to assess the criterion validity of Ro-LIWC2015. Typically, criterion validity emerges from evidence based on the relationships between the test of interest – Ro-LIWC2015, in this case – and other

variables. More precisely, determining the criterion validity involves testing whether the scores established with our instrument are related to other variables to which we would expect them to relate, and vice versa, showing that they are not related to other variables to which we would not expect them to relate (Miller and Lovler 2016). The measurements can be taken at the same time (concurrent validity) or with a delay (predictive validity). Each approach employed in the current paper leverages a different type of criterion for concurrent validity.

First, we test whether the Romanian dictionary and the English version developed by Pennebaker et al. (2015) provide similar outcomes on two homologous corpora that differ only by language. If the two dictionaries are alike, we expect a strong Pearson correlation (i.e., $r \geq .50$) between the sets of features extracted with the two tools. We expect large effect sizes especially for the psychological categories, and not necessarily for the grammatical ones, given that Romanian and English have different origins. Other authors such as Meier et al. (2018) also considered high coefficients as an appropriate metric of equivalence between two counterpart dictionaries. Nevertheless, we argue that smaller effect sizes or not statistically significant correlations could signal not only translation peculiarities or errors but also structural differences between the languages themselves, particularly for the grammar categories such as verbs, articles, etc. Both unfavorable scenarios would raise concerns primarily on using Ro-LIWC2015 together with the English LIWC2015 to fulfill the research need of including language as an independent variable in a direct comparison-type scenario.

In our second approach, we address the criterion validity of Ro-LIWC2015 solely within the Romanian language. For this purpose, we will investigate how efficient Ro-LIWC2015 is in detecting between-group differences when such differences should occur. This strategy aims to assess not the equivalence of Ro-LIWC2015 with another translation but its ability to correctly identify the content focus within a text. In this regard, previous research has shown that the language of individuals with mental health disorders tends to stand out on multiple contents (e.g, Gkotsis et al. 2016), which is an insight we aim to leverage in our study. Thus, for instance, texts about mental health issues should contain more references to affect categories, especially negative emotions (e.g., sadness, anxiety, or anger words) or biological processes (e.g., body, health, or ingest words) than texts on economic topics. However, such a comparison (health vs. economics) might be perceived as not strict enough to support the criterion validity of Ro-LIWC2015. Therefore, in the second study, we seek to extract the linguistic markers of depression and anxiety from posts on Romanian help-seeking forums, which is a more challenging task due to the comorbidity of these conditions. Towards this end, we will employ a supervised learning procedure to address several binary classification problems increasingly difficult: discriminating depression and anxiety posts from orthopedics posts, then from endocrinology posts, and, finally, from one another. We assume that distinguishing mental health posts from endocrinology posts is slightly harder than the analogous scenario with the orthopedics posts because many endocrinology issues cause emotional imbalance. Also, messages about medical issues should contain more words referring to biological and health matters than the mental health corpora, although both types of posts should contain such references.

Throughout both approaches, we focus only on the lower-level features of the LIWC2015 dictionary, given that the hierarchically superior ones represent the cumulative percentage of the constituent categories. For example, *affect* is a higher-level category comprising *positive emotions* and *negative emotions,* which means that it yields values equal to the sum of the word percentages for the two valence features. Furthermore, *sadness*, *anxiety*, and *anger* categories are subordinated to *negative emotions*. Therefore, for instance, for the supervised learning approach, we included in the model only the *positive emotions* category, which does not have subcomponents, along with *sadness*, *anxiety*, and *anger*, while excluding the *affect* and *negative emotions* categories. In Table 1, the higher-level features are aligned to the left and the lower-level ones are indented.

Study 1

The goal of this study is to estimate the equivalence between the Romanian dictionary and the English version, as a measure of criterion validity. For this purpose, we applied a method similar to the one employed by other authors to validate LIWC in other languages (e.g., Spanish – Ramírez-Esparza et al. 2007; German – Wolf et al. 2008, Meier et al. 2018; Dutch – van Wissen and Boot 2017). The method consists of analyzing a set of texts available in both English and the language under test. Thus, a number of benchmark linguistic characteristics of the input text were extracted with the English dictionary and used to assess the results obtained with our new instrument. We relied on this approach to test the following general trend hypothesis:

Hypothesis 1 For most style and content features of the input corpus, as the word percentages established with the English LIWC2015 increase, so do those obtained with Ro-LIWC2015. In statistical terms, we would expect a positive correlation between the percentages computed with the English and Romanian LIWC2015, especially for the psychological variables.

This hypothesis is based on previous validation studies that reported strong correlations between the English LIWC2015 and other translated versions, like the German LIWC2015 (Meier et al. 2018) and the Dutch LIWC2015 (van Wissen and Boot 2017), for most categories. To the best of our

Table 1 The Romanian versus English LIWC2015 – Pearson's correlation coefficients and paired sample t-test results for the lower-level features

	Differences					Equivalence
	Ro-LIWC2015 *M (SD)*	English LIWC2015 *M (SD)*	*t-values*	*p-values*	*Cohen's d*	*Pearson's coefficients, r*
Pronouns						
I	1.46 (0.67)	3.01 (1.33)	−12.37	0.00	2.09	0.93**
We	0.32 (0.18)	0.53 (0.39)	−5.19	0.00	0.88	0.92**
You	1.13 (0.56)	1.85 (0.64)	−5.69	0.00	0.96	0.22
She and he	4.10 (0.98)	6.86 (2.32)	−8.89	0.00	1.50	0.66**
They	1.08 (0.19)	0.78 (0.32)	5.50	0.00	0.93	0.26
Impersonal	4.19 (0.38)	4.80 (0.72)	−5.18	0.00	0.87	0.33
Other function words						
Articles	3.93 (0.40)	7.18 (1.06)	−16.92	0.00	2.86	0.01
Prepositions	11.71 (0.94)	13.51 (0.40)	−10.93	0.00	1.85	0.12
Auxiliary verbs	3.37 (0.53)	8.85 (0.48)	−50.17	0.00	8.48	0.19
Adverbs	4.99 (1.08)	4.56 (0.41)	2.39	0.00	0.40	0.19
Conjunctions	2.40 (0.95)	5.70 (0.96)	−18.23	0.00	3.08	0.37*
Negations	2.81 (0.36)	1.95 (0.18)	16.84	0.00	2.85	0.53**
Other grammar						
Verbs	16.31 (1.25)	17.42 (0.99)	−4.18	0.00	0.71	0.04
Adjectives	6.72 (0.91)	4.07 (0.25)	17.51	0.00	2.96	0.21
Comparisons	1.82 (0.37)	2.06 (0.19)	−3.57	0.00	0.60	0.15
Interrogatives	2.80 (0.42)	1.56 (0.32)	21.11	0.00	3.57	0.58**
Numbers	3.84 (0.48)	1.05 (0.32)	33.53	0.00	5.67	0.29
Quantifiers	1.21 (0.18)	1.74 (0.25)	−16.77	0.00	2.83	0.66**
Affect						
Positive	3.53 (0.51)	2.99 (0.42)	8.87	0.00	1.50	0.72**
Negative	3.30 (0.56)	2.18 (0.36)	15.43	0.00	2.61	0.65**
Anxiety	0.63 (0.13)	0.48 (0.09)	7.78	0.00	1.31	0.56**
Anger	0.99 (0.24)	0.60 (0.17)	13.05	0.00	2.21	0.69**
Sadness	0.82 (0.14)	0.51 (0.11)	12.10	0.00	2.04	0.34*
Social						
Family	0.58 (0.20)	0.56 (0.19)	0.77	0.44	0.13	0.75**
Friend	0.23 (0.06)	0.19 (0.05)	3.62	0.00	0.61	0.52**
Female	0.78 (0.17)	4.07 (1.58)	−12.37	0.00	2.09	0.09
Male	2.16 (0.62)	3.86 (1.25)	−13.07	0.00	2.21	0.87**
Cognitive processes						
Insight	2.41 (0.45)	2.53 (0.44)	−1.38	0.18	0.23	0.28
Causation	2.30 (0.23)	1.26 (0.23)	24.62	0.00	4.16	0.42*
Discrepancy	2.58 (0.32)	1.87 (0.25)	10.89	0.00	1.84	0.09
Tentative	3.31 (0.40)	2.24 (0.34)	18.31	0.00	3.09	0.57**
Certainty	2.13 (0.31)	1.58 (0.34)	11.16	0.00	1.89	0.58**
Difference	3.68 (0.47)	2.98 (0.41)	9.03	0.00	1.53	0.45**
Perceptual processes						
See	1.47 (0.36)	1.46 (0.24)	0.16	0.87	0.03	0.75**
Hear	1.34 (0.28)	1.13 (0.29)	5.34	0.00	0.90	0.66**
Feel	0.80 (0.28)	0.97 (0.32)	−4.32	0.00	0.73	0.70**
Biological processes						
Body	1.52 (0.57)	1.67 (0.76)	−2.17	0.04	0.37	0.84**
Health	0.57 (0.27)	0.55 (0.24)	0.85	0.40	0.14	0.83**
Sexual	0.13 (0.13)	0.16 (0.10)	−2.01	0.05	0.34	0.60**

Table 1 (continued)

	Differences					Equivalence
	Ro-LIWC2015 *M (SD)*	English LIWC2015 *M (SD)*	*t-values*	*p-values*	*Cohen's d*	*Pearson's coefficients, r*
Ingest	0.50 (0.17)	0.48 (0.20)	1.02	0.32	0.17	0.79**
Drives						
Affiliation	1.20 (0.24)	1.64 (0.54)	−5.83	0.00	0.99	0.58**
Achievement	2.02 (0.20)	0.97 (0.20)	26.22	0.00	4.43	0.31
Power	3.02 (0.37)	2.25 (0.27)	12.51	0.00	2.11	0.40*
Reward	1.06 (0.15)	1.04 (0.17)	0.99	0.33	0.17	0.53**
Risk	1.00 (0.21)	0.56 (0.12)	18.81	0.00	3.18	0.75**
Time orientation						
Past	9.73 (1.23)	7.31 (1.45)	12.87	0.00	2.18	0.67**
Present	6.88 (1.38)	7.41 (1.76)	−3.11	0.00	0.53	0.82**
Future	0.93 (.20)	1.25 (.24)	−8.32	0.00	1.41	0.48**
Relativity						
Motion	2.81 (0.30)	2.25 (0.27)	10.24	0.00	1.73	0.38*
Space	8.86 (0.75)	7.32 (0.71)	11.34	0.00	1.92	0.40*
Time	5.74 (0.65)	4.79 (0.58)	8.27	0.00	1.40	0.38*
Personal concerns						
Work	1.28 (0.39)	1.24 (0.39)	1.08	0.29	0.18	0.83**
Leisure	0.76 (0.22)	0.78 (0.23)	−0.74	0.47	0.12	0.79**
Home	0.55 (0.20)	0.64 (0.22)	−7.05	0.00	1.19	0.94**
Money	0.40 (0.17)	0.54 (0.26)	−4.67	0.00	0.79	0.74**
Religion	0.36 (0.31)	0.41 (0.34)	−1.90	0.07	0.32	0.88**
Death	0.21 (0.11)	0.23 (0.16)	−1.06	0.30	0.18	0.69**
Informal language						
Swear	0.14 (0.07)	0.13 (0.10)	1.31	0.20	0.22	0.75**
Net speak	0.19 (0.60)	0.06 (0.06)	1.29	0.21	0.22	0.25
Agreement	0.42 (0.13)	0.19 (0.08)	13.09	0.00	2.21	0.59**
Non-fluencies	0.05 (0.04)	0.22 (0.10)	−10.85	0.00	1.83	0.47**
Filler words	0.001 (0.003)	0.013 (0.01)	−7.32	0.00	1.24	0.16

M = mean word percentage; *SD* = standard deviation of the word percentages; *N* = 35 books; * $p < 0.05$; ** $p < 0.01$

knowledge, the Brazilian Portuguese LIWC2015 (Carvalho et al. 2019) was validated only against the LIWC2007 version of the same language (Balage Filho et al. 2013). As far as the Ukrainian LIWC2015 (Zasiekin et al. 2018) and the Chinese LIWC2015 (Huang et al. n.d.) are concerned, we did not have access to any validation study presented in English. Other LIWC2015 translations have not been developed yet. Nevertheless, to strengthen our hypothesis we could also mention the validation studies of the German LIWC2001 (Wolf et al. 2008) and the Serbian LIWC2007 (Bjekić et al. 2014) in which the full list of the correlations with the homologous English dictionary was reported.

In light of the challenges that we encountered in the translation process of the LIWC2015 dictionary from English to Romanian, which we presented in the introductory section of the current article and in the Supplementary Material 1, we did not exclude the possibility that the grammar categories would show lower correlations. In the same line of thought, Romanian is a Romance language, while English is a Germanic language, and it is well known that languages of such different roots differ one from another on a number of features (e.g., Levshina 2016; Patard 2014). Since we could not find another similar study addressing the equivalence between the LIWC2015 dictionary in another Romance language and the English version, our hypothesis, especially the part regarding the grammar categories, is rather exploratory.

Method

A sample of 35 contemporary literature books written by popular authors such as Nora Roberts, Sandra Brown, or Amanda

Quick was collected. These books were accessible in English and Romanian, in machine-readable formats compatible with the LIWC2015 software (see the list of books in Table S1 from Supplementary Material 2). The rationale for selecting these books was to acquire representative text materials, containing samples of language that resemble real-life communication. Also, we chose the whole book as the unit of analysis – not chapters or other fractions of the book – because we sought to cover as many words as possible per item. Thus, the intention was to reduce the risk of inferring about the whole dictionary if the input encompassed a limited, random number of dictionary words, which would lead to biased conclusions.

The procedure to transform the words found in the selected books into data was straightforward. The English version of the books was processed with the English version of the LIWC2015 dictionary, whereas the Romanian version of the same books was processed with Ro-LIWC2015. Data is available on our Open Science Framework account (Sava and Dudău 2020).

Results

Preliminary Descriptive Analysis According to the LIWC2015 tokenizer, the English corpus contained 101,384.54 words per book ($SD = 66{,}602.93$), whereas the Romanian corpus 94,647.86 words per book ($SD = 65{,}696.20$) on average. The mean percentage of words in the Romanian corpus covered by Ro-LIWC2015 was 66.90% ($SD = 4.39\%$), which is less than the number of words labeled by the English dictionary ($M = 87.04\%$; $SD = 2.47\%$) but not worrying considering the performance of other translations. For example, the Serbian LIWC2007 was able to analyze 64.28% of the input text, as opposed to the English LIWC2007 that included, on average, 80.32% of the total words (Bjekić et al. 2014). DE-LIWC2015 retrieved about 85% of words in the processed text that also belonged to the dictionary (Meier et al. 2018). However, German has similar origins with English – both are Germanic languages –, whereas Romanian is a Romance language and Serbian a Slavic language. Therefore, such differences in coverage could be explained by the linguistic particularities of each language.

One method that could be used to increase the coverage of new instruments might be inspecting different corpora in the target language for most common words and checking whether they are already part of the dictionary obtained by the mere translation of the English LIWC2015. If they have not already been included in the dictionary, they should be assigned to the appropriate LIWC2015 categories. Thus, the new LIWC2015 tool would be extended with words that best define how native speakers express themselves through language, and the content analysis should improve. However, such a method would require both programming and linguistics skills and would increase the time necessary for obtaining the new dictionary. Therefore, a cost-benefit analysis after testing the quality of the new tool comprising the translated words from English to the language of interest should be considered before deciding to proceed with a strategy to detect and add to the dictionary the high-frequency words specific to that language.

Main Analysis The equivalence analysis mainly relied on the correlation coefficient between the variables measured with the English and Romanian LIWC2015. We considered that proofs of good validity were effects that accounted for at least 25% of the variation found in our data (i.e., $r \geq .50$), which is in line with Meier et al. (2018). This analysis was complemented with a direct comparative approach, testing whether statistically significant differences occurred between the English and the Romanian results. However, previous findings revealed that other LIWC translations significantly departed from the English version in terms of direct comparisons (e.g., Meier et al. 2018; Ramírez-Esparza et al. 2007). Therefore, we did not expect that such differences would be absent in our case, mainly because of the language specificities. Table 1 contains the results of both analyses for the lower-level categories in the hierarchy of LIWC2015 features. Table S2 from the Supplementary Material 2 presents the results for the higher-level features. All punctuation variables were excluded from the analysis since they are not part of the dictionary per se but the software, and we did not have any intervention upon them when we created Ro-LIWC2015. Therefore, Table 1 contains only 62 categories.

The equivalence test based on correlations suggested that the Romanian version of LIWC2015 tended to resemble the original dictionary developed by Pennebaker et al. (2015) on multiple linguistic domains, with regard to how words usage covaried between the two paired samples of books. Most correlation coefficients – 72.58%, namely 45 out of 62 – were statistically significant and 56.45% of the correlation coefficients – 35 out 62 – were higher than 0.50. It is worth mentioning that most categories that did not obtain statistically significant correlation coefficients were function words and grammar categories, which could be accounted for by language specificities. The LIWC2015 adaptations for German (Meier et al. 2018) and Dutch (van Wissen and Boot 2017) showed Pearson's coefficients higher than 0.50 for most correlations with the English measures, even in the case of grammar features. Nevertheless, in this line of thought, we reiterate that German and Dutch share their origins with English, which is not the case of Romanian. Romanian resembles languages like Spanish and Brazilian Portuguese, but we could not find records of the correlations between the LIWC2015 versions for these languages and the English dictionary. Moreover, in our study, the *p*-values for personal pronouns "you" and "they" were higher than 0.05, whereas the use of the other personal pronouns demonstrated statistically significant correlations. These results might support the hypothesis of

language particularities, given that Romanian significantly departs from English in terms of second-person pronouns and third-person plural pronouns. Specifically, in Romanian, there is a clear-cut distinction between the singular and the plural forms of the second-person pronouns. Also, in Romanian, the feminine forms of the third-person plural pronoun differ from the masculine ones. All in all, investigating the details regarding the linguistic particularities of languages and the equivalence of different versions of LIWC2015 remains an open research topic.

To conclude, our results revealed that for most content categories, there was a high positive association between the word percentages generated with the two dictionaries, which is in line with our first general hypothesis. In other words, overall, with few exceptions, the two instruments tend to similarly detect the changes that occurred from one book to another, in terms of meaningful content. Thus, for most categories, as the word percentages for the English books increased, the word percentages for the Romanian books also increased.

Secondary Analysis The results of the *t* statistics presented in Table 1 revealed significant differences between the word percentages obtained with Ro-LIWC2015 and those acquired with the English LIWC2015. The Romanian dictionary seemed to detect less meaningful words for some categories (e.g., *first-person pronouns*, orientation towards *present* and *future*, *affiliation*, or personal concerns regarding *home* and *money*, etc.), and more for other categories (e.g., all *emotion* categories, orientation towards *past*, *achievement*, *power*, *risk*, or all *cognitive processes* except *insight*, etc.). The effect sizes were mostly large (Cohen's $d > 0.80$). Overall, the Romanian LIWC2015 tended to capture more meaningful content from the analyzed items than to underestimate it, compared to the English dictionary – there were 24 out of 62 categories with significantly fewer word percentages, and 38 out of 62 categories with significantly more word percentages measured with Ro-LIWC2015 than with the English version of LIWC2015.

Table 2 The composition of the help-seeking forums corpora according to Ro-LIWC2015 – Means (M) and standard deviations (SD) of the word counts and dictionary words variables

			Word counts		Dictionary words	
Condition	Source	N_{posts}	*M*	*SD*	*M*	*SD*
Depression	SM	796	211.89	247.21	85.77%	8.61%
	Tro	160	339.32	383.17	88.12%	3.97%
Anxiety	SM	679	156.92	150.84	85.36%	8.26%
	Tro	80	201.85	198.30	88.12%	4.28%
Control	O-RoM	1712	133.03	100.34	78.70%	7.48%
	E-RoM	1322	113.29	123.04	75.76%	12.00%

Word counts = the raw number of words; Dictionary words = the percentage of words in the analyzed text covered by the dictionary

Discussion

This first validation study addressed the equivalence between the Romanian and the English version of the LIWC2015 dictionary. The association statistics suggested that, overall, the Romanian and English LIWC2015 similarly measured trends in psychological meanings of words across units of analysis. The fact that the correlation coefficients were not statistically significant mainly for the function words and grammatical categories is understandable, given the distinctive rules that define the Romanian language. For example, in Romanian, the definite article is part of nouns endings, whereas in English it is established by the word "the" that precedes the nouns. The tenses of the verbs have different forming rules. However, apart from the function words and grammar categories, there were very few categories for which Pearson's correlation coefficient was not statistically significant or large.

The differences between the Romanian and English LIWC2015 revealed by the *t* statistics might indicate more clearly the particularities of one language against the other. Such statistically significant results have been found in other languages, too, and have been explained as a possible effect of the uniqueness of each language (e.g., Meier et al. 2018; Ramírez-Esparza et al. 2007). Given the specific features of the Romanian language, which sometimes led to changes in category assignment, as well as the fact that we expanded the original dictionary with synonyms, the recorded differences against the English dictionary were rather expectable. In the same line of thought, for most categories, the Romanian LIWC2015 tended to capture more – not less – meaningful content than the English dictionary. Such differences cast doubt mainly on the extent to which LIWC2015 can be used to directly compare samples of different languages, not on the quality of the Romanian dictionary.

The bottom line is that the results of the two statistical approaches do not necessarily contradict one another. Instead, they catch different types of equivalence: (1) the capacity of the two dictionaries to capture the same trends within similar datasets; (2) the extent to which the results differ quantitatively by language, with Ro-LIWC typically covering more psychological content in the recognized text for most categories.

Study 2

The second study focused on the criterion validity of Ro-LIWC2015 considering only the Romanian language. The aim was to extract the linguistic markers of depression and anxiety from posts on Romanian help-seeking forums, using contrast groups and supervised learning. Specifically, to assess the criterion validity of Ro-LIWC2015, we build on the research indicating that people with depression and anxiety show disorder-specific cognitive and linguistic profiles (e.g.,

Hendriks et al. 2014; Thorstad and Wolff 2019). Also, we gathered corpora for two control conditions – orthopedics and endocrinology – besides the depression and anxiety corpora, and checked the following hypotheses:

Hypothesis 2a Both mental health corpora would substantially depart from each of the two control corpora. Thus, the supervised learning algorithm would attain good performance in classifying the orthopedics and endocrinology posts against the depression and anxiety posts. This expectation is in line with the previous research showing that depressed and anxious individuals tend to express themselves differently than others (e.g., Dao et al. 2014; De Choudhury et al. 2013; Thorstad and Wolff 2019).

Hypothesis 2b The linguistic features obtained with Ro-LIWC2015 can be used to accurately discriminate between depression and anxiety posts. As indicated by the literature – both theoretical and research papers – depression and anxiety disorders are defined not only by overlaps but also by distinct cognitive vulnerabilities (for a review, see, for example, Hendriks et al. 2014). Thus, we state that there is evidence for us to assume that depression and anxiety would leave a mark on the language that people use to describe their problems. Footprints like specific emotional load or worries would be found in the linguistic profiles captured from our corpora, although depression and anxiety disorders are highly comorbid (e.g., Gorman 1996).

This second hypothesis is, however, very ambitious, given the high comorbidity between the two conditions, with co-existing symptoms in up to 90% of the patients (Gorman 1996). Likewise, Lamers et al. (2011) found on a large Dutch cohort that the percentage of patients with a current anxiety disorder who had a lifetime history of a depression disorder was 75%, whereas the percentage of patients with current depression who had a lifetime history of anxiety disorder was 81%. In a similar vein, Hirschfeld (2001) showed that a patient with an anxiety disorder had a very high likelihood of developing an additional diagnosis of major depression within the following year. Also, the epidemiological study of Kessler et al. (2015) revealed that anxiety disorders tend to precede the onset of depression and to predict its persistence. For all these reasons, we expect a lower, but still acceptable accuracy in differentiating depression corpus from anxiety corpus using the outcome of Ro-LIWC2015 as an input for the classification approach.

Method

Posts from three Romanian help-seeking forums were collected. The forums were: (1) *sfatulmedicului.ro* (SM), a popular forum where patients with different health issues seek advice directly from professionals or patients with similar problems; (2) *romedic.ro* (RoM), also a popular medical advice forum; (3) *terapeuti.ro* (TRo), a website specialized in psychotherapy, where people who experience mental health or personal issues can seek the help of licensed psychologists.

From SM and TRo, all posts available in the sections dedicated to depression and anxiety disorders were saved. Next, posts common between the two sections were eliminated. We used RoM as a source for the control corpora, saving all posts from the "Orthopedics" (O-RoM) and "Endocrinology" (E-RoM) sections. Before analysis, the SM corpus was cleaned from characters that were introduced automatically by the website developers to censor the cursing or sexually explicit content. The TRo and RoM corpora did not have this problem. The number of posts included in each corpus, as well as the Ro-LIWC2015 coverage of our linguistic inputs, is depicted in Table 2, according to the three conditions, i.e., depression, anxiety, and control. To obtain the depression and anxiety datasets, respectively, we concatenated the posts from SM and TRo. The final depression corpus contains 956 posts, while the anxiety corpus comprises 759 posts. Notably, Ro-LIWC2015 captured more words from the posts collected from the help-seeking forums than from the books we used for analysis in our first study (see Table 2).

To test our hypotheses, we employed the linear discriminant analysis (LDA), which is a machine learning algorithm for classification. LDA assumes that the covariance matrix is equal between classes and uses Bayes theorem, Gaussian densities, and logit transformation to set linear decision boundaries – for a thorough explanation regarding the derivation of the linear discriminant function, see Hastie et al. (2017). To assess the performance of the classification model, we computed seven parameters: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, accuracy, and the area under the receiver operating characteristic curve (AUC).

Results

Preliminary Analysis The mean percentage of words covered by Ro-LIWC2015 ranged from 75.76% in the case of endocrinology corpus to 88.12% in the case of depression and anxiety corpora (see Table 2). This performance is better than the one obtained in the first study, showing that the Ro-LIWC2015 coverage is higher for the forum-type text than for language of popular books.

Before the implementation of the LDA algorithm, we checked for collinearity problems across all classification scenarios. For this purpose the Variance Inflation Factor (VIF) was computed for each feature. VIF indicates whether an input variable has a strong linear relationship with other input variables (Field 2018). We used five as the VIF cut-off for signaling the violation of the independence of attributes, which is an important assumption for LDA (Bickel and Levina 2004). Typically, the multicollinearity is considered

Table 3 LDA statistics on the training sets

			Class means (scaled data)		
C_1 vs. C_2	Hierarchy of features		C_1	C_2	LDA coefficients
Depr vs. Ortho	1.	Space	−0.59	0.33	0.381
	2.	Body	−0.59	0.33	0.326
	3.	Health	−0.48	0.27	0.292
	4.	Anxiety	0.29	−0.16	−0.254
	5.	Negations	0.53	−0.29	−0.253
	6.	Anger	0.52	−0.29	−0.228
	7.	Focus on future	−0.05	0.03	0.228
Depr vs. Endo	1.	Sadness	0.44	−0.32	−0.392
	2.	Health	−0.43	0.31	0.312
	3.	Numbers	−0.40	0.29	0.309
	4.	Anxiety	0.20	−0.15	−0.264
	5.	Negations	0.46	−0.34	−0.257
	6.	Anger	0.43	−0.31	−0.226
	7.	Work	−0.29	0.21	0.215
Anx vs. Ortho	1.	Anxiety	0.75	−0.33	−0.499
	2.	Space	−0.62	0.27	0.400
	3.	Death	0.50	−0.22	−0.303
	4.	Focus on future	−0.07	0.03	0.291
	5.	Ingest	0.35	−0.16	−0.258
	6.	Discrepancy	0.08	−0.03	0.256
	7.	Question mark	0.002	−0.001	0.213
Anx vs. Endo	1.	Anxiety	0.60	−0.35	−0.581
	2.	Numbers	−0.41	0.24	0.346
	3.	Sexual	−0.21	0.12	0.230
	4.	Death	0.31	−0.18	−0.211
	5.	Health	−0.23	0.13	0.200
	6.	Auxiliary verbs	−0.06	0.04	0.197
	7.	Work	−0.28	0.16	0.186
Depr vs. Anx	1.	Sadness	0.18	−0.23	−0.434
	2.	Anxiety	−0.21	0.26	0.346
	3.	Body	−0.23	0.29	0.333
	4.	Discrepancy	0.11	−0.14	−0.253
	5.	Male	0.18	−0.23	−0.207
	6.	Negations	0.20	−0.25	−0.201
	7.	Word counts	0.14	−0.17	−0.199

C_1 = class 1; C_2 = class 2; Depr = depression; Anx = anxiety; Ortho = orthopedics; Endo = endocrinology; $N_{depression} = 717$; $N_{anxiety} = 569$; $N_{orthopedics} = 1284$; $N_{endocrinology} = 992$; LDA coefficients = the loadings of each variables on the discriminant function, also called "slopes" or "weights"; The features are listed in the descending order of their influence on classification, according to the absolute values of the LDA coefficients; Only the top seven features are mentioned for each pair of classes

severe if the VIF value exceeds ten, and moderately severe if the VIF is greater than five (Bowerman et al. 2015).

Only the lower-level features in the hierarchy of LIWC2015 categories were considered. Additionally, the percentages of question and exclamation marks were included in the analysis since they are punctuation with a clear function and could also carry emotional meanings. Thus, initially, we used 64 linguistic features for analysis. For all binary classification problems, due to collinearity, we removed the use of verbs, which was the only variable for which the VIF value was greater than five across all scenarios. Also, for the same reason, we eliminated the *focus on the future* variable but only from the analysis concerning the distinction between anxiety and endocrinology posts. Table S3 in the Supplementary

Table 4 The performance of the LDA algorithm in classifying posts in each pair of corpora on the test sets

Pair	Class	Sensitivity	Specificity	PPV	NPV	F1-score	Accuracy	AUC
Pair 1								
	Depression	0.85	0.97	0.94	0.92	0.90	93%	0.91
	Orthopedics	0.97	0.85	0.92	0.94	0.95		
Pair 2								
	Depression	0.81	0.93	0.90	0.87	0.85	88%	0.87
	Endocrinology	0.93	0.81	0.87	0.90	0.90		
Pair 3								
	Anxiety	0.78	0.97	0.93	0.91	0.85	92%	0.88
	Orthopedics	0.97	0.78	0.91	0.93	0.94		
Pair 4								
	Anxiety	0.78	0.93	0.87	0.88	0.82	88%	0.86
	Endocrinology	0.93	0.78	0.88	0.87	0.91		
Pair 5								
	Depression	0.77	0.66	0.74	0.70	0.76	72%	0.72
	Anxiety	0.66	0.77	0.70	0.74	0.68		

$N_{depression} = 239$; $N_{anxiety} = 190$; $N_{orthopedics} = 428$; $N_{endocrinology} = 330$

Material 2 contains the average number of verbs and words indicating the focus on the future, along with their standard deviations and several examples extracted from our corpora and identified in the English dictionary based on the translation that we made as part of the Ro-LIWC2015 development. Typical examples belonging to the *verbs* and *focus on the future* categories of the original LIWC2015 dictionary can also be found in Pennebaker et al. (2015). Nevertheless, it is important to emphasize that those words per se were not eliminated from the analysis if they also belonged to other categories besides *verbs* and *focus on the future.*

Main Analysis To address the risk of overfitting, for each LDA, the dataset was randomly divided into two subsamples: 75% of posts were assigned to the training set, while the remaining 25% to the test set. The training subset served as a data source for supervised learning, whereas the test subset was used to assess the accuracy of the classifier.

Overall, the LDA statistics provide evidence for the fact that Ro-LIWC2015 demonstrates good criterion validity on our data. Table 3 depicts the top seven linguistic markers based on the absolute values on the LDA classifier in each classification scenario. These characteristics were the most influential for classification on the training set. Tables S4-S8 in the Supplementary Material 2 present the entire hierarchy of features, as established according to their impact on the classification decision. The means of the classes mirror the differences between the components of each pair of contrast samples on the training set. Supplementary Material 2 also includes the normalized confusion matrices (Table S9), which show the percentage of the correct and incorrect classifications of the whole dataset for each pair of corpora.

Depression vs. Control Conditions – Markers of Depression from Text Mining As expected, the linguistic profile of depression posts differed significantly from the contents and style of both control corpora. The parameters presented in Table 4 suggest that the LDA classifier performed well and very well in distinguishing between depression and each of the two health corpora on the test set. When the orthopedics corpus was the contrast group (see Table 3), the most influential linguistic features were words related to space, body, and health, which were less present in the depression corpus, as one would expect. On the other hand, the depression corpus contained more negative emotional content (anxiety, anger), more negations, and less focus on the future. Among the top features for distinguishing depression from endocrinology (see Table 3) were sadness, health, numbers, anxiety, negations, and anger. Sadness, anxiety, negations, and anger were markers of depression, whereas more words referring to health and numbers defined the endocrinology corpus. Likewise, in both contrast groups scenarios, depression posts contained more words suggesting anxiety and anger, and more negations, which is explainable given the symptoms of depression – these three types of linguistic content were also among the most influential features for classification on the training set.

Anxiety vs. Control Conditions – Markers of Anxiety from Text Mining Also, according to the accuracy parameters in Table 4, the LDA model for identifying the anxiety posts against orthopedics and endocrinology posts achieved good to excellent performance, which was consistent with our hypothesis. The use of more words related to anxiety in the anxiety corpus was the most impactful feature that affected the decision in both classification

problems (anxiety vs. orthopedics and anxiety vs. endocrinology), as shown in Table 3. Another noteworthy marker of anxiety revealed in both scenarios was the use of more words related to death. The higher number of words regarding space was once again a distinctive and important feature of the orthopedics corpus. The endocrinology posts contained more words related to numbers, sex, and health than the anxiety corpus, which influenced a lot the classification decision.

Depression vs. Anxiety – Linguistic Markers that Distinguish between the Two Mental Health Conditions In line with the *Hypothesis 2b*, the LDA model managed to discriminate between depression and anxiety posts, in a fair manner, demonstrating a 72% accuracy, as Table 4 shows. The probability of a post being recognized by the model as belonging to the depression corpus when it did was 0.74. The analogous probability of a post from the anxiety corpus was 0.70. Although it is possible to improve the classification accuracy, the high comorbidity of these two conditions might account for the lower classification rate compared to the previous two scenarios.

The linguistic markers with the highest absolute values on the LDA classifier were words referring to sadness, anxiety, body, and discrepancy (see Table 3). The differences between the two corpora on these most influential features were consistent with what one would expect given the disorder-specific cognitive profiles of depression and anxiety. In the texts belonging to the depression corpus, *sadness* and *discrepancy* (e.g., *should*, *would*) are more present than in the anxiety corpus, while in the texts belonging to the anxiety corpus, *anxiety* and *body* categories are more present than in the depression corpus.

Discussion

In the second study, we assessed the criterion validity of Ro-LIWC2015 using posts found in help-seeking forums under the sections dedicated to depression, anxiety, orthopedics, and endocrinology issues. Overall, our tool demonstrated good proprieties since the linguistic features that we measured were successful predictors in our binary classification problems. Also, the linguistic markers revealed in each scenario were consistent with what one would expect, given the results of previous research and the characteristics of each disorder.

Depression and anxiety corpora not only differed from the control samples in a way that assured good classification accuracy but also the dissimilarities seemed to capture several disorder-specific features. The users who posted in the depression sections demonstrated higher self-focus (i.e., they used more first-person pronouns) than those who discussed orthopedics and endocrinology problems. This result is in line with the previous studies revealing that higher use of first-person pronouns is a common marker of depression (e.g., Edwards and Holtzman 2017). Other distinctive features of depression compared to control samples, including more words suggesting emotions and certainty, more swear words, and more conjunctions, were also consistent with previous research investigating the language of depression on social media (e.g., Dao et al. 2014; De Choudhury et al. 2013). Although such features were not in the top ten influential markers for classification, they suggest promising results in support of good criterion validity. Their lower position in the hierarchy of features could be explained by the fact that the control samples also had some strong particularities.

Thus, for example, one impactful feature in distinguishing between orthopedics and mental health corpora (both depression and anxiety) was the use of more words related to space. Such a result could be expectable since the orthopedics injuries have a specific location and coverage (e.g., "left elbow", where "left" is a *space* word) and impair patients' relationship with the environment (e.g., "I can't walk the stairs", where "stairs" is *a space* word). The common sense tells that describing such injuries requires more words related to space. Also, the use of words related to health was a consistent marker of endocrinology posts compared to both depression and anxiety corpora. This result could be in line with the fact that the endocrinological problems have a wide range of consequences on the body. Also, typically, the endocrinological deficiencies are controlled with medication taken according to a thorough plan, which could explain the importance of words related to numbers and ingestion in the classification scenarios involving endocrinology posts.

The LDA statistics for the distinction between depression and anxiety posts also provided evidence for the validity of Ro-LIWC2015 since the algorithm attained a fair accuracy despite the high comorbidity between these conditions. The top-four impactful categories in defining depression against anxiety corpora were sadness, anxiety, body, and discrepancy, which is consistent with previous research.

Sonnenschein et al. (2018) showed that depressed patients used more words suggesting sadness and a similar amount of first-person singular pronouns than patients with anxiety disorders, during cognitive-behavior therapy. In our study, words related to sadness were more frequent in depression posts than in anxiety posts, while the use of first-person singular pronouns had a small influence on the classification decision. Higher self-focus emerged not only from the depression corpus but also from the anxiety posts when they were compared to both control samples.

Also, our findings might reflect, at least partly, the disorder-specific cognitive profiles resulted from previous research and well-known theories regarding the distinctiveness/similarity of depression and anxiety, as depicted, for example, in the paper of Hendriks et al. (2014). In line with this literature, our model might suggest that rumination was higher in depression posts than in anxiety posts, as indicated by the fact that, on average, the former contained more word counts than the latter and that discrepancy

was among the top markers that differentiated between the two conditions. In the same line of thought, having negative evaluations of the self, the world, and the future typically defines depression, not anxiety (Hendriks et al. 2014). The linguistic markers extracted with Ro-LIWC2015 could have mirrored such disorder-specific features, too, in the fact that depression corpus contained more negations compared to anxiety corpus.

Our results were consistent also with the cognitive profile of anxiety, which is characterized by higher worry and physical concerns (Hendriks et al. 2014). We found that the anxiety corpus included more words related to anxiety (the worry component) and biological and perceptual processes – as represented by the top categories *body* and *health*, and the less impactful categories for classification *ingest* and *feel* – which could signal sensitivity to physical issues. Other findings concerning the language of anxiety on social media converged with this picture. For example, Thorstad and Wolff (2019) applied cluster analysis on posts from various clinical subreddits. They revealed that the anxiety corpus was characterized by words referring to panic, fear, worry, drugs, and obsessive thoughts, among others. In the *anxiety* and *body processes* categories, Ro-LIWC2015 covers many of the words that formed these anxiety clusters.

General Discussion

This paper we focused on LIWC2015 (Pennebaker et al. 2015), one of the highly used and most powerful computer-based language analysis tools worldwide, and developed and tested its Romanian version – Ro-LIWC2015. To assess the criterion validity of our tool, we proposed two studies. In the first study, we used as input a Romanian corpus of 35 books and its English counterpart. In the second study, we processed texts about anxiety, depression, orthopaedics, and endocrinology problems posted in help-seeking forums and created five binary classification scenarios. Both studies were consistent with our hypotheses, revealing promising results to support the fact that Ro-LIWC2015 is a valid tool.

In the first study, the correlation analysis was used to test the equivalence between the original LIWC2015 and the Romanian LIWC2015. Overall, the results sustained the equivalence between the Romanian and the English version of LIWC2015, which was in line with previous research (e.g., Meier et al. 2018; Ramírez-Esparza et al. 2007). However, given the particularities of each language, we argue that direct between-group comparisons might be problematic in a multilingual setting. One easy solution to this problem would be to standardize the scores within-group (e.g., *z*-scores) or to center the scores around the mean for each language when researchers seek to correlate the LIWC2015 scores with other variables of interests in a multilingual setting. This solution is in line with other views (see Meier et al. 2018).

The second study also provided evidence for the criterion validity of the Romanian LIWC2015. Our hypothesis that depression and anxiety corpora would depart substantially from the orthopedics and endocrinology corpora was supported by the obtained results. As expected, considering previous research (e.g., De Choudhury et al. 2013; Edwards and Holtzman 2017), depression posts contained more first-person-singular pronouns, conjunctions, and certainty words, to name a few linguistic markers, than the control samples. Also, the anxiety corpus was more abundant in words related to anxiety than the control corpora, which remarkably influenced the classification decision. A number of potential disorder-specific particularities also emerged from the orthopedics and endocrinology corpora.

Our second study also provided evidence consistent with the hypothesis that the linguistic features obtained with Ro-LIWC2015 can be used to discriminate between depression and anxiety posts fairly accurately. The linguistic profiles of each condition, as identified by the LDA algorithm, were consistent with previous research and well-known theories regarding the distinct features of depression and anxiety, which constitute another clue for good criterion validity. The language describing depression problems contained more words referring to sadness, discrepancy, and negations and contained more word counts. These linguistic characteristics could reflect higher rumination and more negative views on own inner and outer experiences, as it typically happens more in depression than in anxiety (e.g., Hendriks et al. 2014). In contrast, the anxiety posts carried more words related to anxiety, body parts, corporal sensations, ingestion and health matters. The worrying and sensitivity to physical issues are top components of the cognitive and linguistic profile of anxiety, as shown by previous research (e.g., Hendriks et al. 2014; Thorstad and Wolff 2019). Also, our model indicated that the use of the first-person pronouns was a weak criterion in distinguishing between depression and anxiety posts, which was also in line with other findings (e.g., Sonnenschein et al. 2018).

Overall, in light of the obtained results, the current paper brings to the forefront the first valid Romanian version of the LIWC2015 dictionary, which could already be used in research on various topics. Introducing this new tool has two major practical implications. First, the automatic content analysis instruments like LIWC2015 can help psychologists and other social scientists leverage data that are less affected by the problems commonly associated with the self-report and implicit methods. Both types of assessment are extensively applied in social science, despite their shortcomings. Usually, the self-report method is affected by self-presentation or memory biases (e.g., Gosling et al. 1998; Tourangeau 2000), whereas the implicit methods involve unknown mechanisms (Goodall 2011). The possibility of extracting meanings from the natural language with Ro-LIWC2015 could enhance the Romanian research, leading to powerful results. Second, our paper contributes to the line of research regarding multilingual analysis,

which is an important topic today given the technological developments that allowed the accumulation of vast amounts of linguistic data from all around the world. In this regard, the current research adds the Romanian language to the repertoire of languages amenable to LIWC2015 analysis, which, so far, comprises German, Dutch, Brazilian Portuguese, Ukrainian, and Chinese, besides English.

Limitations

Although our findings broadly converged with the existent literature and verified our hypotheses, they should be considered within the boundaries of several methodological shortcomings. One of the major concerns is that both studies focused only on the corpora of informal and semi-formal speech since we opted for the language from contemporary books and help-seeking forums dedicated to patients and professionals. As Meier et al. (2018) stated, the language analysis can be affected not only by how a dictionary was built but also by the context embedded in the analyzed texts. Testing the equivalence between the German and the English version of LIWC2015, they obtained different between-language correlations in corpora of formal versus semi-formal language. Therefore, future research should address the validity of Ro-LIWC2015 on additional linguistic samples.

In the same line of thought, Ro-LIWC2015 showed better word coverage on the help-seeking forums dataset, especially on depression and anxiety posts, than on the corpus of books. This difference in the percentage of recognized words for analysis could suggest that our dictionary might be more suitable for processing texts about mental and physical health issues than other matters. Other variations in coverage could also occur on other contents. In this regard, an important limitation in the process of creating Ro-LIWC2015 is that we did not consider which words are most frequent in real-world communication in Romanian. Our translation entirely relied on the words found in several dictionaries. Thus, although we extended the dictionary by up to five synonyms for every English word, the coverage of Ro-LIWC2015 could be improved.

Another noteworthy limit of our research is the fact that the groups in our second study were self-formed. The labeling was determined solely by the users' judgment when they decided in which section of the forum to post their message. The absence of an objective criterion in establishing the samples could have introduced bias in our statistical models. In this line of thought, we recommend further research with improved methodology. One suggestion would be to screen participants for depression and anxiety based on specific criteria such as clinical interview, questionnaire scores, or language analysis before using their digital traces or other linguistic data to assess the validity of Ro-LIWC2015.

In the same vein, we used only two criteria to test the validity of Ro-LIWC2015 – the results obtained with the English LIWC2015, and the type of problem that characterized the corpora collected from Romanian help-seeking forums (orthopedics, endocrinology, depression, and anxiety). To strengthen the criterion-related validity evidence, future research should investigate the relationship between the linguistic features extracted with Ro-LIWC2015 and other variables. Such variables could be different psychological constructs such as personality traits (e.g., comparing introverts and extraverts), or linguistic features acquired with other versions of LIWC2015 and other computer-based tools. Also, methodologies that enable the assessment of predictive validity, which is also a measure of criterion validity, should be implemented. In both studies, we tested the concurrent validity. In this regard, for instance, a sample of depressed individuals could be asked to write meaningful essays at two time-points. The language of those who would follow a cognitive-behavioral therapy program should be different at the second measurement than the language of the control subsample, according to the Ro-LIWC2015 analysis, given that at that time, they would also display lower depressive symptoms. Likewise, we would recommend testing the internal consistency of Ro-LIWC2015. Assessing other types of validity than criterion validity might be more problematic. For instance, content validity could be established with human assessors who should be different than the persons who built the Romanian version of LIWC2015. They could rate how well each item in the Romanian dictionary was assigned to each category. However, considering that Ro-LIWC2015 contains a large number of entries, it is very likely that the process would be time consuming or would require many trained raters. Moreover, although the instrument as a whole does not measure a specific construct, a number of categories do refer to well-established psychological variables. Thus, the construct validity – both convergent and discriminant types – could be addressed for some components of Ro-LIWC2015. For example, the percentage of words that indicate negative emotions should show strong negative correlations with measures of depression and anxiety (convergent validity), and negative correlations with happiness (discriminant validity).

Conclusion

From the very beginning, traditional content analysis was the key to extract inferences from natural language in a systematic, rigorous manner. Although it remains a valuable approach to specific research problems, it has shortcomings that make it be outdated for many current quests. The technological advances of the last three decades opened up promising avenues of social science research by providing an enormous and ever-increasing repository of written language. However, to automatically convert text for statistical analysis can be a challenging task, especially for those who do not have skills in data science. LIWC2015 is one of the most versatile and popular tools for language analysis worldwide and comes with a user-friendly software solution that

anyone can manipulate instantly. This paper introduced the first Romanian version of LIWC2015. Our studies revealed that Ro-LIWC2015 shows good criterion validity. Although further research is needed to cover additional validity-check scenarios, we already encourage the use of Ro-LIWC2015 for hypothesis testing.

Acknowledgements This work has received funding from the BID grant (PN-III-P1-PFE-28) funded by the Romanian Ministry of Research and Innovation.

Compliance with Ethical Standards

Ethical approval All procedures performed in studies were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. For this type of study formal consent is not required.

Conflict of Interest No conflict of interest.

References

Agosti, A., & Rellini, A. (2007). *The Italian LIWC dictionary*. Austin, TX: LIWC.net.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (pp. 2200–2204).

Balage Filho, P. P., Pardo, T. A. S., & Aluísio, S. M. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology* (pp. 215–219). Sociedade Brasileira de Computação.

Balahur, A., & Perea-Ortega, J. M. (2015). Sentiment analysis system adaptation for multilingual processing: The case of tweets. *Information Processing & Management, 51*(4), 547–556. https://doi.org/10.1016/j.ipm.2014.10.004.

Bickel, P. J., & Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli, 10*(6), 989–1010.

Bjekić, J., Lazarević, L., Erić, M., Stojimirović, E., & Đokić, T. (2012). Razvoj srpske verzije rečnika za automatsku analizu teksta (LIWCser). *Psihološka Istraživanja, 15*(1), 85–110.

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.

Bond, G. D., Holman, R. D., Eggert, J. A. L., Speller, L. F., Garcia, O. N., Mejia, S. C., Mcinnes, K. W., Ceniceros, E. C., & Rustige, R. (2017). 'Lyin'Ted', 'Crooked Hillary', and 'Deceptive Donald': Language of lies in the 2016 US Presidential Debates. *Applied Cognitive Psychology, 31*(6), 668–677. https://doi.org/10.1002/acp.3376.

Boot, P., Zijlstra, H., & Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics, 6*(1), 65–76. https://doi.org/10.1075/dujal.6.1.04boo.

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S. (2015). Regression analysis. Unified concepts, practical applications, and computer implementation. Business Expert Press.

Boyd, R. L. (2017). Psychological text analysis in the digital humanities. In S. Hai-Jew (Ed.), *Data analytics in digital humanities* (pp. 161–189). Springer International Publishing.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical report C-1. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Carvalho, F., Rodrigues, R. G., Santos, G., Cruz, P., Ferrari, L., & Guedes, G. P. (2019). Evaluating the Brazilian Portuguese version of the 2015 LIWC lexicon with sentiment analysis in social networks. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining* (pp. 24–34). SBC.

Dao, B., Nguyen, T., Phung, D., & Venkatesh, S. (2014). Effect of mood, social connectivity and age in online depression community via topic and linguistic analysis. In B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali, & Y. Zhang (Eds.), *Web Information Systems Engineering – WISE 2014. WISE 2014. Lecture Notes in Computer Science* (vol. 8786, pp. 398–407). Cham: Springer. https://doi.org/10.1007/978-3-319-11749-2_30.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations. *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 47–56). https://doi.org/10.1145/2464464.2464480.

Drisko, J. W., & Maschi, T. (2016). *Content analysis. Pocket guides to social work research methods*. New York: Oxford University Press.

Edwards, T., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality, 68*, 63–68. https://doi.org/10.1016/j.jrp.2017.02.005.

Faasse, K., Chatman, C. J., & Martin, L. R. (2016). A comparison of language use in pro-and anti-vaccination comments in response to a high profile Facebook post. *Vaccine, 34*(47), 5808–5814. https://doi.org/10.1016/j.vaccine.2016.09.029.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. California: SAGE Publications Ltd..

Fofiu, A. (2012). The Romanian version of the LIWC2001 dictionary and its application for text analysis with Yoshikoder. *Studia Universitatis Babes-Bolyai-Sociologia, 57*(2), 139–151.

Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., & Dutta, R. (2016). The language of mental health problems in social media. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 63–73). https://doi.org/10.18653/v1/W16-0307.

Goodall, C. E. (2011). An overview of implicit measures of attitudes: methods, mechanisms, strengths, and limitations. *Communication Methods and Measures, 5*(3), 203–222. https://doi.org/10.1080/19312458.2011.596992.

Gorman, J. M. (1996). Comorbid depression and anxiety spectrum disorders. *Depression and Anxiety, 4*(4), 160–168.

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and*

Social Psychology, 74(5), 1337–1349. https://doi.org/10.1037/0022-3514.74.5.1337.

Harari, Y. N. (2014). *Sapiens: A brief history of humankind*. London: Vintage Books.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning. Data mining, inference, and prediction* (2nd ed.). Springer Science + Business Media. https://doi.org/10.1007/b94608.

Hendriks, S. M., Licht, C. M., Spijker, J., Beekman, A. T., Hardeveld, F., de Graaf, R., & Penninx, B. W. (2014). Disorder-specific cognitive profiles in major depressive disorder and generalized anxiety disorder. *BMC Psychiatry, 14*(96). https://doi.org/10.1186/1471-244X-14-96.

Hirschfeld, R. M. (2001). The comorbidity of major depression and anxiety disorders: Recognition and management in primary care. *Primary Care Companion to the Journal of Clinical Psychiatry, 3*(6), 244–254. https://doi.org/10.4088/pcc.v03n0609.

Huang, C.-L., Chung, C. K., Hui, N., Lin, Y.-C., Seih, Y.-T., Lam, B. C. P., Chen, W.-C., Bond, M. H., & Pennebaker, J. W. (2012). The development of the Chinese Linguistic Inquiry and Word Count dictionary. *Chinese Journal of Psychology, 54*(2), 185–201.

Huang, C.-L., Lin, W.-F., Seih, Y.-T., Lin, Y.-C., & Lee, C.-L. (n.d.). *Traditional Chinese LIWC2015 Dictionary*. Austin, TX: LIWC.net.

Kailer, A., & Chung, C. K. (2011). *The Russian LIWC2007 dictionary*. Austin, TX: LIWC.net.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods, 21*(4), 507–525. https://doi.org/10.1037/met0000091.

Kessler, R., Sampson, N., Berglund, P., Gruber, M., Al-Hamzawi, A., Andrade, L., et al. (2015). Anxious and non-anxious major depressive disorder in the World Health Organization world mental health surveys. *Epidemiology and Psychiatric Sciences, 24*(3), 210–226. https://doi.org/10.1017/S2045796015000189.

Kleim, B., Horn, A. B., Kraehenmann, R., Mehl, M. R., & Ehlers, A. (2018). Early linguistic markers of trauma-specific processing indicate vulnerability for later chronic posttraumatic stress disorder. *Frontiers in Psychiatry, 9*, 645. https://doi.org/10.3389/fpsyt.2018.00645.

Krippendorff, K. (2004). *Content analysis. An introduction to its methodology* (2nd ed.). Thousand Oakes, California: Sage.

Lamers, F., van Oppen, P., Comijs, H. C., Smit, J. H., Spinhoven, P., van Balkom, A. J. L. M., et al. (2011). Comorbidity patterns of anxiety and depressive disorders in a large cohort study: The Netherlands study of depression and anxiety (NESDA). *Journal of Clinical Psychiatry, 72*(3), 341–348. https://doi.org/10.4088/JCP.10m06176blu.

Levshina, N. (2016). Verbs of letting in Germanic and romance languages: A quantitative investigation based on a parallel corpus of film subtitles. *Languages in Contrast, 16*(1), 84–117. https://doi.org/10.1075/lic.16.1.04lev.

Mäntylä, Graziotin, & Kuutila. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review, 27*, 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002.

Meier, T., Boyd, R.L., Pennebaker, J.W., Mehl, M.R., Martin, M., Wolf, M., & Horn, A.B. (2018). *"LIWC auf Deutsch": The development, psychometrics, and introduction of DE-LIWC2015.* Retrieved from https://osf.io/tfqzc/.

Miller, L. A., & Lovler, R. L. (2016). Foundations of psychological testing. *A practical approach* (5th ed.). SAGE Publications, Inc.

Patard, A. (2014). When tense and aspect convey modality. Reflections on the modal uses of past tenses in Romance and Germanic languages. *Journal of Pragmatics, 71*, 69–97. https://doi.org/10.1016/j.pragma.2014.06.009.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC 2001.* Mahwah: Erlbaum.

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science, 10*(3), 90–93.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count (LIWC): LIWC2007*. Austin, TX: LIWC.net.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

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. *Psychologie Française, 56*(3), 145–159. https://doi.org/10.1016/j.psfr.2011.07.002.

Piryani, R., Madhavi, D., & Singh, V. K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management, 53*(1), 122–150. https://doi.org/10.1016/j.ipm.2016.07.001.

Ramírez-Esparza, N., Pennebaker, J. W., Garcia, A. F., & Suriá, R. (2007). La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista mexicana de psicología, 24*(1), 85–99.

Sava, F. A., & Dudău, D. P. (2020). RoLIWC2015 and mental health. Retrieved from osf.io/6tn9k

Scheuerlein, J., Chládková, H., & Bauer, K. (2018). Transformational leadership qualities during the financial crisis-a content analysis of CEOs letter to shareholders. *International Journal for Quality Research, 12*(3), 551–572. https://doi.org/10.18421/IJQR12.03-01.

Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The Annals of the American Academy of Political and Social Science, 659*(1), 78–94. https://doi.org/10.1177/0002716215569197.

Settanni, M., Azucar, D., & Marengo, D. (2018). Predicting individual characteristics from digital traces on social media: A meta-analysis. *Cyberpsychology, Behavior and Social Networking, 21*(4), 217–228. https://doi.org/10.1089/cyber.2017.0384.

Sonnenschein, A. R., Hofmann, S. G., Ziegelmayer, T., & Lutz, W. (2018). Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cognitive Behaviour Therapy, 47*(4), 315–327. https://doi.org/10.1080/16506073.2017.1419505.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The general inquirer: A computer approach to content analysis*. M.I.T. Press.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544–2558. https://doi.org/10.1002/asi.21416.

Thorstad, R., & Wolff, P. (2019). Predicting future mental illness from social media: A big-data approach. *Behavior Research Methods., 51*, 1586–1600. https://doi.org/10.3758/s13428-019-01235-z.

Tourangeau, R. (2000). Remembering what happened: Memory errors and survey reports. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 29–47). Mahwah: Lawrence Erlbaum Associates Publishers.

van Wissen, L., & Boot, P. (2017). An electronic translation of the LIWC dictionary into Dutch. In *Electronic lexicography in the 21st century:*

Proceedings of eLex 2017 conference (pp. 703–715). Lexical Computing.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., ... & Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations* (pp. 34–35).

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative textanalyse: äquivalenz und robustheit der deutschen version des linguistic inquiry and word count. *Diagnostica, 54*(2), 85–98.

Zasiekin, S., Bezuglova, N., Hapon, A., Matiushenko, V., Podolska, O., & Zubchuk, D. (2018). Psycholinguistic aspects of translating LIWC dictionary. *East European Journal of Psycholinguistics, 5*(1), 111–118. https://doi.org/10.5281/zenodo.1436335.