Article

Not peer-reviewed version

Health Literacy: Structural Topic Modeling and Citation Trends

Benjamin M. Tabak *, Matheus Britto Froner, Rubiane Daniele C. de Almeida, Luzia Cláudia D. Couto, Thiago Christiano Silva, Sayonara de Fatima Faria Barbosa

Posted Date: 5 August 2024

doi: 10.20944/preprints202408.0278.v1

Keywords: Health Literacy; Mental Health Literacy; Digital Health Literacy; HLQ; HLS-EU; MHLS; REALM; TOFHLA; eHeals; NVS

Article

Health Literacy: Structural Topic Modeling and Citation Trends

Benjamin M. Tabak [1,*,†,‡], Matheus B. Froner [1,‡], Rubiane Daniele C. de Almeida [1,‡], Luzia Cláudia D. Couto [2,‡], Thiago C. Silva[3,‡] and Sayonara de Fatima Faria Barbosa [4,‡]

1 School of Public Policy and Government, Getulio Vargas Foundation; SGAN 602, mod. A-C, Asa Norte, Brasilia, DF, Brazil
2 Ministério da Saúde. Esplanada dos Ministérios, Bloco G, Térreo. CEP: 70058-900. Brasília/DF
3 Universidade Católica de Brasília, Brasília/DF
4 University of Cincinnati College of Nursing. 2600 Clifton Ave, Cincinnati, OH 45221
* Correspondence: benjaminm.tabak@gmail.com
† Current address: School of Public Policy and Government, Getulio Vargas Foundation, SGAN 602 Módulos A,B,C, Asa Norte, Brasília, DF 70830-020, Brazil.
‡ These authors contributed equally to this work.

Abstract: Health literacy is a growing research area with specific aspects and different instruments to measure health literacy. This article uses natural language processing model to analyze the academic corpora regarding seven health literacy instruments - Health Literacy Questionnaire, Mental Health Literacy Scale, Rapid Estimate of Adult Literacy in Medicine, Test of Functional Health Literacy in Adults, Health Literacy Survey, The Newest Vital Sign and eHealth Literacy Scale. We apply Structural Topic Modeling to all the abstracts of the selected academic corpora, label the topics based on their focus, and use the topic distribution with metadata to train a Random Forest algorithm for predicting yearly citations. We estimate Regression models based on the ranking of the most relevant features, which serves as a robustness check and to infer their impact on citation dynamics. We have found that Digital Health Literacy is positively associated with yearly citations, while other topics such as Functional Health Literacy and Women's Health reduce citation likelihood. Other article characteristics also have shown a significant role in the citation likelihood, such as the publication year, amount of articles and certain keywords. These findings portray the current landscape of health literacy research, highlighting literature gaps and popular topics.

Keywords: health literacy; mental health literacy; digital health literacy; HLQ; HLS-EU; MHLS; REALM; TOFHLA; eHeals; NVS

1. Introduction

The recent COVID-19 pandemic has demonstrated the importance of adopting public health policies that promote knowledge about health decisions and existing healthcare options. In this context, health literacy gained special importance [1]. In times of crisis, when people need to use information quickly to make decisions related to health, which can involve high stakes and, possibly, even life or death scenarios, greater health literacy becomes essential.

However, a low health literacy of the population can result in negative effects both to individuals and society as a whole. Patients with low health literacy do not adhere to the medications prescribed by health professionals [2–4]. Those with low health literacy may also need help distinguishing fake information to the detriment of which can lead to poor decision making when dealing with health issues. Trust in public health authorities can be reduced since people with low health literacy may need help understanding the language, which is often technical, and looking for alternatives when facing health issues [5–8].

Many of these problems have been clearly observed during the Covid-19 pandemic. Those with lower e-health literacy have been shown to be more susceptible to fake information, which results in an inadequate response to the risk of the pandemic [5–7]. Lower health literacy has also been shown to

 doi:10.20944/preprints202408.0278.v1

be related to hesitancy towards Covid-19 vaccination as well as a lack of trust in government health institutions [8].

Therefore, it is essential to develop instruments to measure health literacy properly. It is possible to analyze the different dimensions of literacy to develop evidence-based public policies. There are currently several instruments available. Some are more subjective in that the health system users' perceptions are asked. Other tools are more objective and allow for the elaboration of questions with right and wrong answers, making it possible to determine whether the patient can comprehend what the healthcare providers are saying.

People navigating the healthcare system must have a broad understanding of their problems, the possibilities of treatments, their risks, and their benefits. It is essential to empower everyone who seeks the health system so that they can make well-informed decisions, increasing people's well-being.

Understanding which dimensions are most difficult for health system users to understand is essential so that they can discuss an intelligence system for health literacy. It is essential to find ways to bring knowledge to the citizen in an accessible language. Health institutions and all those participating in the process must be equipped to provide those seeking health assistance with quality information that allows them to act consciously and properly evaluate their situation.

To contribute to this debate, in this article, we use Structural Topic Modeling (STM), a natural language processing model applied to the abstract of articles discussing seven different health literacy. This approach. Which enables us to categorize and understand the latent themes present in these articles. Based on the output of this model, along with other metadata, we implement a machine learning algorithm to determine the most important variables to predict citation and use them in a linear regression. This not only quantifies the relationship between these variables and citations but also allows for inference about their relative importance and impact within the larger corpus.

We discuss in this paper the following questionnaires and their respective corpus:

1. Health Literacy Questionnaire - HLQ
2. Mental Health Literacy Scale - MHLS
3. Rapid Estimate of Adult Literacy in Medicine - REALM
4. Test of Functional Health Literacy in Adults - TOFHLA
5. Health Literacy Survey - HLS
6. The Newest Vital Sign - NVS
7. eHealth Literacy Scale - eHeals

2. Materials and Methods

We have selected seven widely used instruments for measuring health literacy. We have searched for the most cited scientific references to each one of the instruments in the Scopus database. Using the most cited articles, along with the seminal article for each instrument we have summarized each instrument.

We have opted to use Structural Topic Modeling since this natural language processing approach allows for the interpretation of a high volume of textual data. The resulting information serves to train the Random Forest algorithm and to estimate regression models. This allows the use of more detailed information present in the corpus to examine the topics being studied, rather than depending solely on the keywords provided by the authors.

2.1. Structural Topic Modeling

We use the Structural Topic Modeling approach applied to the abstract of all the articles in the seven corpora. However, to obtain the best results from the STM model, we have cleaned the data, by tokenizing, stemming, and removing any stopwords and non-alphabetic characters from the abstracts of the corpus.

STM is based on the Latent Dirichlet Allocation (LDA) method. LDA assumes that a document inside its corpus arises from a probabilistic distribution of latent topics and it also assumes that all

 doi:10.20944/preprints202408.0278.v1

documents share a common Dirichlet prior. These latent topics are represented as a probabilistic distribution over words, with those also sharing a common Dirichlet prior [9–11].

However, due to the STM model not being based on a Dirichlet Distribution, it allows for the incorporation of other metadata present in the corpus. This facilitates the integration of relevant aspects to the topic modeling, discarding LDA's assumption of constant that topic prevalence and content are constant across documents [10–13]. We have opted to use the publication year of each article as well as seven dummy variables indicating in which corpora each article was present, as covariates with topic prevalence. We have made this choice given that different surveys have different focuses and that topics might receive more or less attention with time.

The STM model, similarly to the LDA approach, assumes that each document is made out of K topics, with each document having different topic proportions (θ). However, unlike in LDA models, STM allows for the θ to be correlated, and the prevalence of those topics can be modified by the metadata X set as covariates of a logistic normal distribution regression model [12,13].

First, the document-level attention for each topic is derived from a logistic-normal generalized linear model based on a vector of document covariates X_d:

$$\theta_d | X_d, \gamma, \sum \sim LogisticNormal(\mu = X_d\gamma, \sum) \tag{1}$$

where θ is the different topic proportions, d is the document index, X_d is a $1 - by - p$ vector, γ is a $p - by - (K - 1)$ matrix of coefficients and $\sum$ $(K - 1)$-by-$(K - 1)$ covariance matrix.

The STM model assigns a thematic composition to a document, symbolized by θ, which is inferred from the document's words (w). The association of each word (w) with a topic is governed by a probability distribution unique to that document. Hence, the chance of a word aligning with a specific topic depends on the document's overall thematic framework, as represented by θ. This document-specific probability distribution is parameterized by β, which is the deviations from the word frequencies (m) in natural logarithm [12,13]. Being formally represented below:

$$\beta_d k \propto \exp m + \kappa^(t)_k \tag{2}$$

where m signifies the baseline distribution of words across all topics, indicating commonality in word usage. $\kappa^(t)_k$ represents the deviation attributable to the specific topic k, highlighting unique word preferences within that topic.

Then for each in word in each document ($n \in 1, ..., N_d$ the word's topic assignment is drawn based on the document-specific distribution over topics.

$$z_d, n | \vec{\theta}_d \sim Multinomial(\vec{\theta}_d) \tag{3}$$

Based on the chosen topic, draw a word observed from that topic.

$$w_d, n | z_d, n, \beta_{d,k=z_d,n} \sim Multinomial(\beta_{d,k=z_d,n}) \tag{4}$$

In our analysis, we have opted to not include any content covariates, which would be used for variations in word meanings that arise due to specific metadata associated with the documents. By omitting these covariates, our focus remains solely on the inherent textual content present in the corpora.

One important methodological aspect of the STM is that it needs the number of K topics to be specified before running the model. In order to determine the optimal K value, we have implemented the searchK function from the "stm" package [13]. This method evaluates the output from several K specifications.

Our choice of K was done aiming to balance the mean topic exclusivity and semantic coherence values estimated by the searchK function. Based upon these values exhibited in Figure 1, we have opted to select a K value of 14.

 doi:10.20944/preprints202408.0278.v1

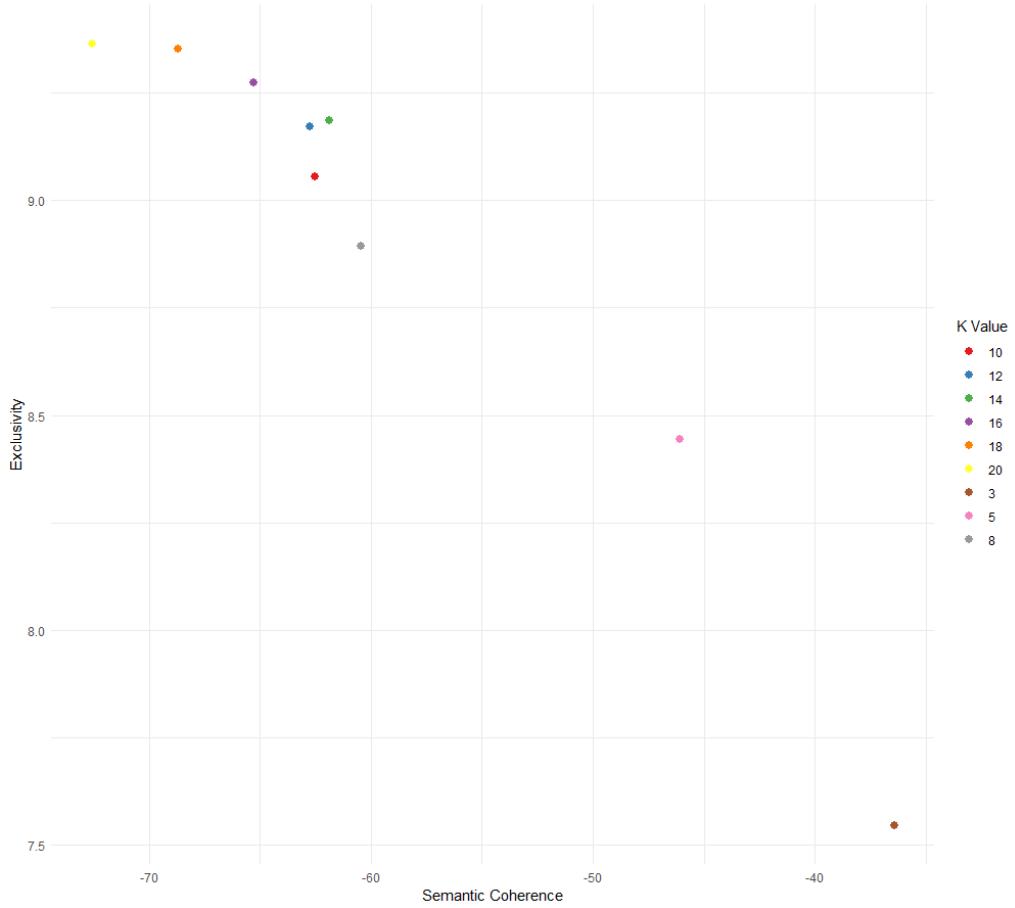

Figure 1. Semantic Coherence and Exclusivity Values.

2.2. Citation Dynamics

With the distribution of topics for each document, as well as other metadata of the corpus, such as the age of the article, which keywords were present, and the number of authors we have implemented a Random Forest algorithm to predict citation. The Random Forest algorithm works as an ensemble of decision or regression trees, each tree based upon a bootstrap sample of the original data. At each node of the trees, a random subset of variables is used, which increases even more the variability among the trees.

The target variable of our analysis is the log of the yearly citations + 1, which is predicted by regression trees as a part of the Random Forest algorithm, given its continuous nature. We have divided our data into the testing set comprising of 20% of the total papers and the training set consisting of the remaining 80%. After this division, we filtered variables and removed those that remained with the same value in over 98% of the observations to reduce complexity. In order to prevent multicollinearity we have also removed keywords that presented a Pearson correlation coefficient higher than 0.99 with another predictor.

We have selected the hyperparameters by minimizing the Root Mean Squared Error (RMSE) due to its sensitivity to large errors and its interpretations being relatable to the original scale of the data [14,15]. We defined the mtry (number of attributes that each tree in the forest uses during training) through repeated k-fold cross-validation. In this tuning process, we have used 5 k-folds, 5 separate runs, and had a tune length of 30 with the number of trees being 500 a minimum node size of 5, and no maximum depth value defined. With this, we selected a mtry of 32.

This algorithm helps us predict citations and establish the most relevant variables impacting citations. However one of the problems for the Random Forest algorithm is interpretability, to overcome

 doi:10.20944/preprints202408.0278.v1

this issue we follow the methodology of using the most relevant variables in an OLS regression [1,16]. This allows for the interpretation of coefficients to evaluate the relationship between the variables and citations. The regression can be seen formally in Equation (5).

$$y_i = \alpha + \beta_1 \text{Age} + \beta_2 \text{SingleAuthored}_i + \beta_3 \text{QtyAuthors}_i + \beta_4 \text{Keywords}_i + \sum_{\substack{j=1 \\ j \neq 13}}^{K} \beta_j \text{Topic}_{ji} + \epsilon_i \quad (5)$$

where 'y_i' is the yearly citations of the paper 'i', 'α' is the intercept, the value of the dependent variable when all independent variables are zero, 'β_1' refers to the years since the publication of the paper,'β_2' is a dummy variable representing whether paper 'i' was written by a single author, 'β_3' is the number of authors in the paper 'i', 'β_4' are dummy variables that represent whether each of the top keywords for predicting citations, as estimated by the Random Forest algorithm, are present in paper 'i'. β_j represents the coefficient for the distribution of topic j on article i of each of the K topics except for Topic 13, due to it being the most prevalent topic, which was removed to avoid multicollinearity, given that the sum of all topic proportion on a given topic is always 1. The 'ϵ' represents the residuals of the regression. We opted to implement fixed effects for the age of the paper and for the country of the corresponding author. In this way, we can make our model robust for unobserved heterogeneity.

3. Results

3.1. Instruments

It is important to note that each instrument has their specific peculiarities and intended uses. The REALM [17], while being more focused on literacy and being one of the first health literacy instruments, was developed to be used as way to pre-screen patients in order to properly communicate health information with them; The HLQ manages to create a nuanced group of scores that is able to provide guidance for public policy [18–20]; The HLS-EU-Q is capable of measuring different aspects of health literacy and summing it in one single score [21,22]; The NVS is able to focus the instrument on the evaluation of a food label, making it a quick instrument that measures the ability of an individual to interpret and use health related information [23]; The TOFHLA evaluates health literacy conjoined with numeracy and reading comprehension [24]; The eHeals and the MHLS are both instruments focused on specific aspects of health literacy, the former on the use of electronic sources for health information [25] and the latter on mental health literacy and mental health stigma [26].

The instruments mentioned, each with their unique focus and methodology, collectively contribute to a comprehensive understanding of health literacy. The REALM, HLQ, HLS-EU-Q, NVS, TOFHLA, eHeals, and MHLS, while diverse in their approaches, together encapsulate the multifaceted nature of health literacy. Their diverse applications range from clinical settings, where quick screening is crucial, to policy-making arenas, where nuanced insights into population health literacy are needed. This diversity underscores the importance of choosing the right tool for the right context, as the implications of these assessments can significantly influence patient outcomes and health policy decisions.

3.2. Structural Topic Modeling Results

The resulting 14 topics from the STM model and their main lemmatized words are presented in Table 1. We have labeled each top according to those words, as a way to facilitate the understanding and interpretation of the underlying themes within our dataset. By assigning descriptive labels based on the predominant words associated with each topic, we aim to encapsulate the essence of the discussions and trends captured by the model. This approach allows for a more intuitive grasp of the contents of each topic.

In Table 1, the Highest Prob refers to words that have the highest probability of occurring within a specific topic; FREX is a metric that combines word frequency and exclusivity to identify words that

 doi:10.20944/preprints202408.0278.v1

are not only frequent in a topic but also exclusive to it, providing a balance between common and distinctive words when characterizing topics [27,28]; Lift measures how much more likely a word is to appear in a topic compared to its overall frequency across all documents. A higher lift value indicates that a word is more unique to a topic, helping to distinguish between topics [13,29]. These categories encompass both most probable words, more exclusive words and those that are most probable and most exclusive simultaneously.

Table 1. Topic Labels.

Topic Number	Topic Label	Main Words
Topic 1	NVS	Highest Prob: patient, nvs, sign, vital, newest, visit, use FREX: nvs, hrc, sign, heart, vital, visit, physician
Topic 2	Functional Health Literacy	Highest Prob: literaci, health, adult, function, test, measur, tofhla FREX: tofhla, realm, read, function, numeraci, mmse, stofhla
Topic 3	Self-care	Highest Prob: health, behavior, diabet, control, literaci, intervent, knowledg FREX: diabet, selfcar, behavior, glycem, control, spss, diet
Topic 4	Mental	Highest Prob: mental, health, depress, mhl, scale, asthma, ill FREX: mental, mhl, disord, mhls, wellb, asthma, helpseek
Topic 5	General	Highest Prob: health, literaci, score, studi, level, correl, signific FREX: dental, oral, univers, pearson, adolesc, correl, reserv
Topic 6	Instrument Validation	Highest Prob: valid, item, reliabl, measur, factor, instrument, scale FREX: psychometr, cronbach, properti, alpha, confirmatori, converg, cfa
Topic 7	Treatment	Highest Prob: patient, medic, adher, diseas, hospit, associ, literaci FREX: adher, nonadher, transplant, hiv, kidney, hemodialysi, ckd
Topic 8	HLS EU	Highest Prob: health, literaci, studi, use, research, popul, need FREX: european, will, project, migrant, compet, hls, review
Topic 9	Women Health	Highest Prob: women, health, literaci, cancer, qualiti, inform, life FREX: cancer, breast, women, decisionmak, life, pregnant, pregnanc
Topic 10	Healthcare	Highest Prob: low, group, particip, literaci, caregiv, care, intervent FREX: caregiv, vaccin, franci, taylor, low, llc, aor
Topic 11	Children	Highest Prob: parent, screen, use, children, comprehens, question, particip FREX: parent, label, dose, children, screen, child, instruct
Topic 12	eHeals	Highest Prob: ehealth, inform, use, literaci, eheal, internet, student FREX: ehealth, internet, digit, ehl, mhealth, eheal, onlin
Topic 13	General Method	Highest Prob: health, literaci, associ, level, age, factor, educ FREX: resid, status, regress, incom, logist, sociodemograph, age
Topic 14	HLQ	Highest Prob: health, literaci, inform, use, healthcar, activ, hlq FREX: hlq, domain, healthcar, engag, profil, navig, rehabilit

Figure 2 shows the topic prevalence of each of the topics, according to the labels shown in Table 1. Here we can see that the General Method topic is the most prevalent one, which makes sense, given that it is not a specific theme. It is followed by the Functional Health Literacy and the HLS EU topics, the former being composed by the REALM and TOFHLA instruments and the latter by the European Health Literacy Survey, all three being widely used instruments. The least prevalent topic is the mental one, which can be explained since it is a specific topic inside of health care as a whole, being the focus of only one of the instruments being studied, the MHLS.

 doi:10.20944/preprints202408.0278.v1

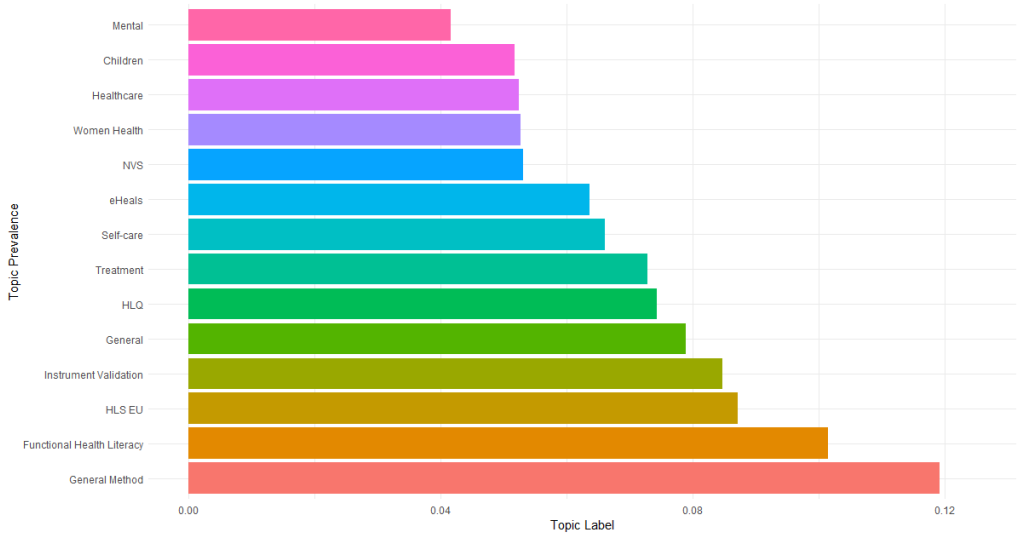

Figure 2. Topic Prevalence.

In Figure 3 we present the topic network based on the topic co-occurrence on the same documents. We can see that the topic with the highest number of connections is Healthcare, which indicates that these different approaches are being applied in the healthcare context. Functional Health Literacy and NVS topics are both connected to the Children and Treatment topics, which is interesting to note given that the capacity to understand and properly administer prescriptions is relevant both in the case of treatment as a whole and also for guardians tasked with the care of a child. This relationship underscores the vital role that health literacy plays in ensuring appropriate health outcomes. Three topics can be seen separated, are the eHeals, Mental, and Instrument Validations topics, which, in the case of the former two, are topics focused on specialized domains of health literacy, and the latter stands apart as a methodological topic dedicated to ensuring the reliability and accuracy of tools used to measure health literacy.

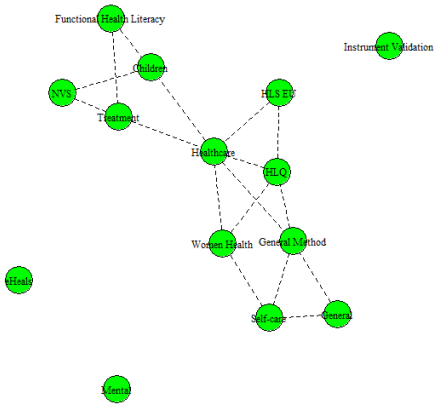

Figure 3. Topic Correlation.

One interesting aspect is that the year of publication covariate can also help us understand what topics have been receiving more attention as of late and what topics are on the decline. To do so we

 doi:10.20944/preprints202408.0278.v1

present the coefficients of the publication year predicting each of the topic' proportions through a linear model, exhibited on Figure 4. The following topics have presented negative coefficients: NVS, Functional Health Literacy, Treatment, and Children. It's interesting to note that all four are correlated according to Figure 3, which indicates that those topics used to receive more attention, but with time the literature has focused on different aspects of health literacy. The topics with positive coefficients, Self-care, Mental, General, Instrument Validation, HLS EU, eHeals, General Method and HLQ, on the other hand, indicate areas of growing interest within the field of health literacy research. This shift in focus reflects changes in societal health concerns, advancements in technology and healthcare practices, and the development of the body of work regarding health literacy as a whole. For instance, the rising interest in mental health literacy parallels a broader societal acknowledgment of mental health's importance. Regarding the eHeals topic, the access to electronic devices expanded a significantly since the 1990s. With electronic devices now commonplace, they serve as a primary source of health information for many individuals, which reflects on the growth of academic interest in digital health literacy. The growth of instruments like the HLQ and the HLS EU make sense in this context given that they have been developed to try to capture the multi-dimensional side of health literacy, expanding on previous instruments that were focused on numeracy and literacy. Both the Women Health and the Healthcare have a large confidence interval which indicates a high standard error and low significance in the estimation.

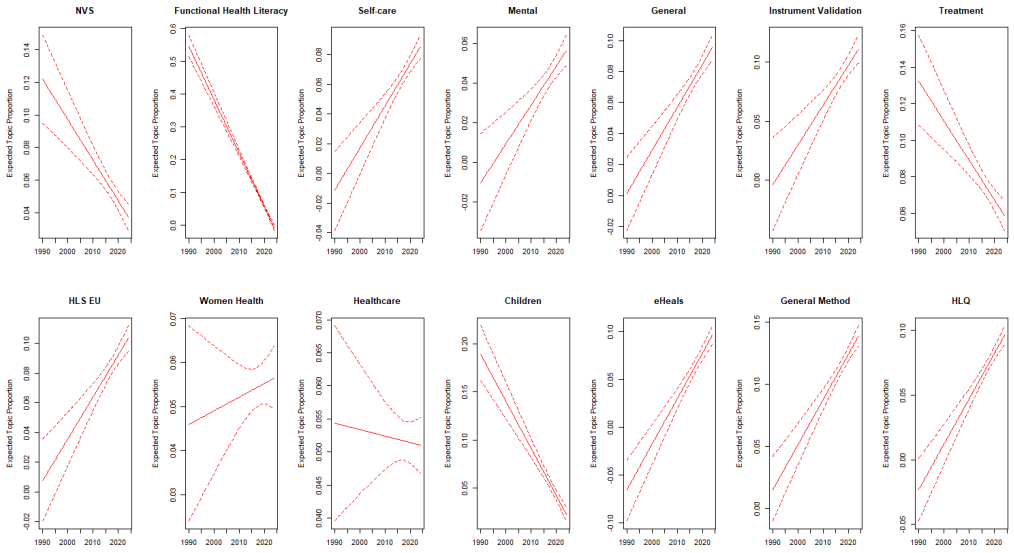

Figure 4. Year of publication.

3.3. Citation Dynamics

Implementing the Random Forest algorithm we have used the RMSE parameter to estimate the optimal mtry value. In Figure 5 we exhibit the resulting RMSE for each mtry, which designated the mtry value of 32 as the optimal setting for our data.

 doi:10.20944/preprints202408.0278.v1

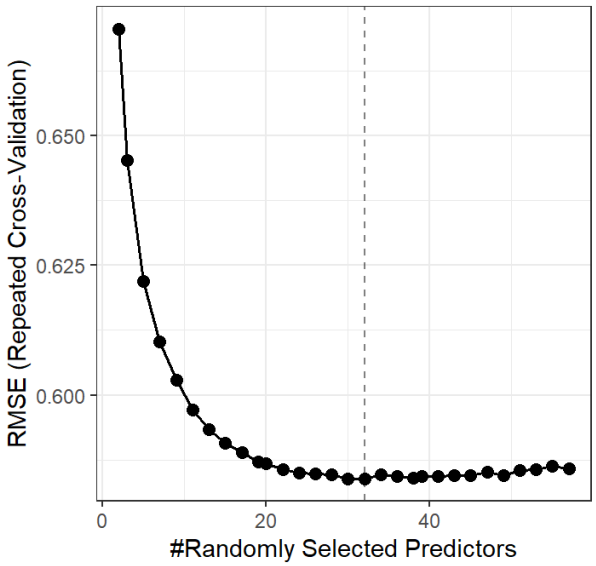

Figure 5. Estimation of Optimal mtry value.

Having trained our algorithm we were able to identify the most relevant variables to predict the yearly citations. We present the 25 most important variables to predict yearly citations in Figure 6. It is noteworthy that, other than the article's age, the most important variables are the topic distributions of each article. This denotes that the result of the STM model captures relevant differences amidst the different topics that are reflected in the different citation dynamics in each topic.

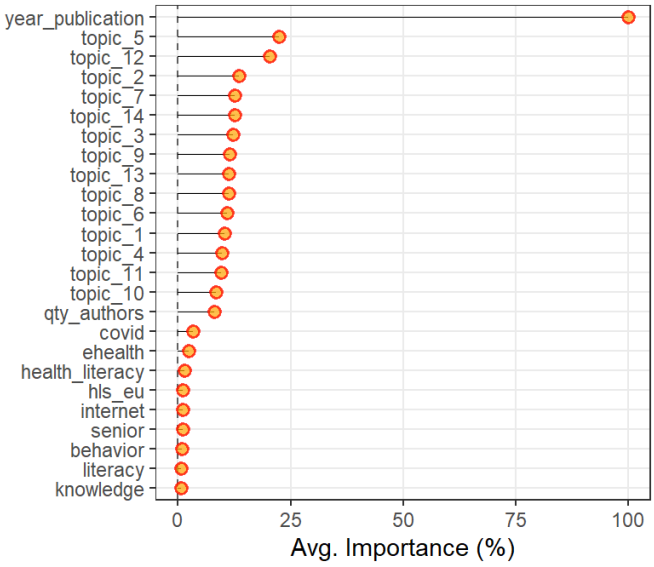

Figure 6. Most Important Variables to Predict Yearly Citations

In Table 2 we present the regressions shown in Equation (5) both with and without using the top keywords in Figure 6. We have also opted to present the results with and without fixed effects for the articles' age since it is the most important variable according to the Random Forest algorithm, which allows us to estimate the effects of the other variables while isolating the impact of the year of publication of the articles. On columns 1 and 2 we can see the coefficients for the models without keywords and on columns 3 and 4 the coefficients for the models with the keywords. It's important to note that the most prevalent topic, Topic 13, has been removed from the model in order to prevent multicollinearity.

 doi:10.20944/preprints202408.0278.v1

Table 2. Citation Dynamics Regression

	Dependent variable:			
	Citation per year			
	(1)	(2)	(3)	(4)
Paper's Age	0.581***		0.590***	
	(0.032)		(0.033)	
Topic 1 (NVS)	−6.255***	−6.338***	−4.837***	−4.825***
	(1.396)	(1.376)	(1.438)	(1.415)
Topic 2 (Functional Health Literacy)	−6.726***	−4.788***	−6.489***	−4.513***
	(1.573)	(1.564)	(1.597)	(1.582)
Topic 3 (Self-care)	−1.542	−1.546	−1.236	−1.376
	(1.571)	(1.535)	(1.598)	(1.558)
Topic 4 (Mental)	−2.602*	−2.358	−0.623	−0.445
	(1.470)	(1.435)	(1.550)	(1.511)
Topic 5 (General)	−7.133***	−6.947***	−6.574***	−6.358***
	(1.661)	(1.619)	(1.682)	(1.636)
Topic 6 (Instrumental Validation)	−0.629	−0.424	−0.113	0.118
	(1.193)	(1.169)	(1.214)	(1.189)
Topic 7 (Treatment)	−3.735**	−3.533**	−2.731*	−2.475
	(1.542)	(1.513)	(1.568)	(1.536)
Topic 8 (HLS EU)	−1.952	−2.052	−1.787	−1.768
	(1.456)	(1.424)	(1.459)	(1.425)
Topic 9 (Women Health)	−7.150***	−6.370***	−5.834***	−4.929***
	(1.713)	(1.679)	(1.732)	(1.695)
Topic 10 (Healthcare)	−3.694	−3.608	−3.970	−3.900
	(2.727)	(2.665)	(2.741)	(2.673)
Topic 11 (Children)	−4.233***	−3.781**	−3.154**	−2.552
	(1.578)	(1.560)	(1.600)	(1.579)
Topic 12 (eHeals)	2.987**	3.397**	3.173**	3.559**
	(1.357)	(1.328)	(1.603)	(1.565)
Topic 14 (HLQ)	0.559	0.758	1.458	1.728
	(1.653)	(1.613)	(1.674)	(1.629)
Amount of Authors	0.156***	0.161***	0.146***	0.151***
	(0.037)	(0.036)	(0.037)	(0.036)
Single Author	−0.805	−0.916	−0.835	−0.941
	(0.738)	(0.723)	(0.734)	(0.717)
Keyword Covid			1.861***	2.075***
			(0.616)	(0.605)
Keyword eHealth			0.452	0.348
			(0.482)	(0.469)
Keyword Health Literacy			0.966***	0.919***
			(0.285)	(0.281)
Keyword HLS EU			1.333**	1.310**
			(0.597)	(0.582)
Keyword Internet			0.483	0.520
			(0.651)	(0.634)
Keyword Senior			0.414	0.506
			(0.548)	(0.536)
Keyword Behavior			0.120	0.253
			(0.478)	(0.465)
Keyword Literacy			1.207**	1.799***
			(0.525)	(0.526)
Keyword Knowledge			−0.684	0.119
			(0.783)	(0.780)
Constant	1.448		−0.074	
	(1.048)		(1.116)	
Fixed Effects	None	Paper's Age	None	Paper's Age
Observations	1,857	1,857	1,857	1,857
R^2	0.211	0.267	0.224	0.282
Adjusted R^2	0.204	0.248	0.213	0.260
Residual Std. Error	4.886 (df = 1840)	4.748 (df = 1811)	4.858 (df = 1831)	4.710 (df = 1802)
Note:				*p<0.1; **p<0.05; ***p<0.01

 doi:10.20944/preprints202408.0278.v1

Here we can see that in the models without keywords 7 out of the 13 topics present a p-value lower than 0.05. In the models with keywords, the model with no fixed effects presents 6 topics with that level of significance, and in the model with age-fixed effects that number goes down to 5. This indicates that the effects of the topics are partially represented by the keywords, which makes sense given that both are related to the themes being explored in the articles. It's relevant to note that the increase in the adjusted R^2 due to the addition of keywords in the model is incremental, going from 0.204 to 0.213 in the models without fixed effects and from 0.248 to 0.260 in the models with fixed effects.

The topics with a p-value higher than 0.05 in the models without keywords are the topics labeled as NVS, Functional Health Literacy, General, Treatment, Women Health, Children, and the eHeals. All of them have presented a negative coefficient save from the eHeals topic. By adding keywords to the model the Treatment topic has a higher p-value, still maintaining a p-value lower than 0.1, and by adding keywords and age-fixed effects the Children topic loses significance.

The significant keywords are Covid, Health Literacy, HLS EU and Literacy. All of them show positive coefficients and Covid has the highest coefficient.

4. Discussion

In this article, we have provided a brief review of seven instruments used to measure health literacy. We have also provided a discussion of the main articles of each one of those instruments. This can serve as an overview of these instruments, which can guide the choice of instruments for future researchers in the area.

The use of accurate and contextually adequate health literacy measuring instruments is essential in order to properly assess and propose adequate solutions to communicate properly with low health literacy patients and populations. It is also important to develop public policies to improve health literacy in the medium and long term. This challenge and its impact in public health have made themselves especially clear in the pandemic and its consequences.

Our Structural Topic Modeling allowed us to thematically divide the literature through the latent topics identified. We have labeled the topics according to their most representative words and shown their correlation as well as what topics have gained interest throughout the years.

By coupling these results with the Random Forest algorithm and OLS regression models, our analysis has explored the role that topics and keywords play in citation dynamics. This expands on the methodology used to evaluate citation likelihood in other areas of knowledge, by incorporating the thematic information contained in the abstracts [1,16]. We've been able to isolate and identify the impact of specific topics and keywords on yearly citations. Notably, our findings reveal that certain topics, as well as keywords such as Covid, Health Literacy, HLS EU, and Literacy, significantly affect the yearly citations of the articles, however, the inclusion of keywords only slightly enhanced the explanatory power of our models as evidenced by incremental increases in adjusted R^2 values, which indicates that the topics resulting from the STM model play a larger role on prediction yearly citations.

4.1. Conclusion

In the course of research, we look into seven instruments designed to measure health literacy, while some seem to offer insightful results, others are limited to specific aspects of health literacy. This specific focus of some instruments is important, given that health literacy is a concept with multiple dimensions. We these instruments, outlining their key aspects, advantages, and disadvantages. Given the immense cultural, social, and geographical aspects that influence health literacy, it is challenging to obtain a satisfactory measure for health literacy in its multiple facets, and therefore the choice of instrument is extremely relevant.

With our use of Structured Topic Modeling, we have categorized the literature into distinct thematic areas, identifying latent topics based on their most representative words and analyzing their correlations and trends over time. By integrating these insights with findings from Random

 doi:10.20944/preprints202408.0278.v1

Forest and OLS regression models, we investigated how topics and specific keywords, notably Covid, Health Literacy, HLS EU, and Literacy, influence the yearly citations of articles. Our analysis revealed that while these factors do significantly impact citation counts. This suggests that the latent topics identified through STM are more influential in predicting yearly citations, underscoring the importance of thematic content in driving scholarly impact and engagement within the field.

In conclusion, this article offers an overview of the seven instruments being studied that highlights the strengths and limitations of each tool, this serves as a guide for researchers looking to select the most appropriate instrument for their specific research goals. Notably, our analysis using Structural Topic Modeling is useful for understanding the thematic divisions within the literature, which, when incorporated into Random Forest and regression analyses, offers predictive insights on scholarly impact. This insight emphasizes the importance of thematic relevance in the visibility and impact of health literacy research.

Author Contributions: All authors have contributed equally to this work.

Funding: Financial support for this research was provided by Fundação de Apoio à Pesquisa do Distrito Federal (FAP-DF). All the authors acknowledge FAP-DF for their financial support through the Project "Um diagnóstico da Educação em Saúde no Distrito Federal" (Process No. 33435.154.29827.20102022). Matheus B. Froner and Rubiane Daniele Cardoso de Almeida gratefully acknowledge financial support from Fundação de Apoio à Pesquisa do Distrito Federal (Project no. 00193.00002349/2022-43). Thiago C. Silva (Grant no. 302703/2022-5) and Benjamin M. Tabak. (Grant no. 305485/2022-9) gratefully acknowledge financial support from the CNPq foundation.

Data Availability Statement: Data presented in this study are publicly available on Scopus and Web of Science and can be made available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DOAJ	Directory of open access journals
eHeals	eHealth Literacy Scale
HLS-EU-Q	The European Health Literacy Survey Questionnaire
MHLS	Mental Health Literacy Scale
TOFHLA	Test of Functional Health Literacy in Adults
REALM	Rapid Estimate of Adult Literacy in Medicine
NVS	Newest Vital Signt
HLQ	Health Literacy Questionnaire

References

1. Tabak, B.M.; Froner, M.B.; Corrêa, R.S.; Silva, T.C. The Intersection of Health Literacy and Public Health: A Machine Learning-Enhanced Bibliometric Investigation. *International Journal of Environmental Research and Public Health* **2023**, *20*. doi:10.3390/ijerph20206951.
2. Huang, C.L.; Chiang, C.H.; Yang, S.C.; Wu, F.Z. The Associations among Gender, Age, eHealth Literacy, Beliefs about Medicines and Medication Adherence among Elementary and Secondary School Teachers. *International Journal of Environmental Research and Public Health* **2022**, *19*, 6926.
3. Fabricius, P.K.; Aharaz, A.; Stefánsdóttir, N.T.; Houlind, M.B.; Steffensen, K.D.; Andersen, O.; Kirk, J.W. Shared Decision Making with Acutely Hospitalized, Older Poly-Medicated Patients: A Mixed-Methods Study in an Emergency Department. *International Journal of Environmental Research and Public Health* **2022**, *19*, 6429.
4. Kim, M.; Suh, D.; Barone, J.A.; Jung, S.Y.; Wu, W.; Suh, D.C. Health literacy level and comprehension of prescription and nonprescription drug information. *International Journal of Environmental Research and Public Health* **2022**, *19*, 6665.

5. Mohammed, F.; Al-Kumaim, N.H.; Alzahrani, A.I.; Fazea, Y. The impact of social media shared health content on protective behavior against COVID-19. *International Journal of Environmental Research and Public Health* **2023**, *20*, 1775.
6. Vrdelja, M.; Vrbovšek, S.; Klopčič, V.; Dadaczynski, K.; Okan, O. Facing the growing COVID-19 infodemic: digital health literacy and information-seeking behaviour of university students in Slovenia. *International journal of environmental research and public health* **2021**, *18*, 8507.
7. Okan, O.; Bollweg, T.M.; Berens, E.M.; Hurrelmann, K.; Bauer, U.; Schaeffer, D. Coronavirus-related health literacy: a cross-sectional study in adults during the COVID-19 infodemic in Germany. *International Journal of Environmental Research and Public Health* **2020**, *17*, 5503.
8. Pickles, K.; Copp, T.; Meyerowitz-Katz, G.; Dodd, R.H.; Bonner, C.; Nickel, B.; Steffens, M.S.; Seale, H.; Cvejic, E.; Taba, M.; others. COVID-19 vaccine misperceptions in a community sample of adults aged 18–49 years in Australia. *International Journal of Environmental Research and Public Health* **2022**, *19*, 6883.
9. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* **2019**, *78*, 15169–15211.
10. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *Journal of machine Learning research* **2003**, *3*, 993–1022.
11. Blei, D.M.; Lafferty, J.D. Topic models. In *Text mining*; Chapman and Hall/CRC, 2009; pp. 101–124.
12. Roberts, M.E.; Stewart, B.M.; Tingley, D.; Lucas, C.; Leder-Luis, J.; Gadarian, S.K.; Albertson, B.; Rand, D.G. Structural topic models for open-ended survey responses. *American journal of political science* **2014**, *58*, 1064–1082.
13. Roberts, M.E.; Stewart, B.M.; Tingley, D. Stm: An R package for structural topic models. *Journal of Statistical Software* **2019**, *91*, 1–40.
14. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific model development discussions* **2014**, *7*, 1525–1534.
15. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *International journal of forecasting* **2006**, *22*, 679–688.
16. Tabak, B.M.; Silva, T.C.; Fiche, M.E.; Braz, T. Citation likelihood analysis of the interbank financial networks literature: A machine learning and bibliometric approach. *Physica A: Statistical Mechanics and its Applications* **2021**, *562*, 125363. doi:https://doi.org/10.1016/j.physa.2020.125363.
17. Murphy, P.W.; Davis, T.C.; Long, S.W.; Jackson, R.H.; Decker, B.C. Rapid estimate of adult literacy in medicine (REALM): a quick reading test for patients. *Journal of reading* **1993**, *37*, 124–130.
18. Batterham, R.W.; Buchbinder, R.; Beauchamp, A.; Dodson, S.; Elsworth, G.R.; Osborne, R.H. The OPtimising HEalth LIterAcy (Ophelia) process: Study protocol for using health literacy profiling and community engagement to create and implement health reform. *BMC Public Health* **2014**, *14*. Cited by: 116; All Open Access, Gold Open Access, Green Open Access, doi:10.1186/1471-2458-14-694.
19. Beauchamp, A.; Batterham, R.W.; Dodson, S.; Astbury, B.; Elsworth, G.R.; McPhee, C.; Jacobson, J.; Buchbinder, R.; Osborne, R.H. Systematic development and implementation of interventions to OPtimise Health Literacy and Access (Ophelia). *BMC Public Health* **2017**, *17*. Cited by: 67; All Open Access, Gold Open Access, Green Open Access, doi:10.1186/s12889-017-4147-5.
20. Osborne, R.H.; Batterham, R.W.; Elsworth, G.R.; Hawkins, M.; Buchbinder, R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). *BMC public health* **2013**, *13*, 1–17.
21. Sørensen, K.; Van den Broucke, S.; Pelikan, J.M.; Fullam, J.; Doyle, G.; Slonska, Z.; Kondilis, B.; Stoffels, V.; Osborne, R.H.; Brand, H. Measuring health literacy in populations: illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q). *BMC public health* **2013**, *13*, 1–10.
22. Sørensen, K.; Pelikan, J.M.; Röthlin, F.; Ganahl, K.; Slonska, Z.; Doyle, G.; Fullam, J.; Kondilis, B.; Agrafiotis, D.; Uiters, E.; others. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *The European journal of public health* **2015**, *25*, 1053–1058.
23. Pfizer Inc. Newest Vital Sign: A Health Literacy Assessment Tool, 2011. Accessed: 2023-07-28.
24. Parker, R.M.; Baker, D.W.; Williams, M.V.; Nurss, J.R. The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills. *J Gen Intern Med* **1995**, *10*, 537–541.
25. Norman, C.D.; Skinner, H.A. eHEALS: the eHealth literacy scale. *Journal of medical Internet research* **2006**, *8*, e507.

 doi:10.20944/preprints202408.0278.v1

26. O'Connor, M.; Casey, L. The Mental Health Literacy Scale (MHLS): A new scale-based measure of mental health literacy. *Psychiatry Research* **2015**, *229*, 511 – 516. Cited by: 187; All Open Access, Green Open Access, doi:10.1016/j.psychres.2015.05.064.
27. Airoldi, E.M.; Bischof, J.M. Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association* **2016**, *111*, 1381–1403.
28. Bischof, J.; Airoldi, E.M. Summarizing topical content with word frequency and exclusivity. Proceedings of the 29th international conference on machine learning (icml-12), 2012, pp. 201–208.
29. Taddy, M. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association* **2013**, *108*, 755–770.