Revista de Sistemas de Informação da FSMA
n. 18 (2017) pp. 2-8

SISTEMAS DE INFORMAÇÃO

http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL

The influence of readability aspects on the user's perception of helpfulness of online reviews

Vinicius Woloszyn, *PhD Student in Computer Science, PPGC-UFRGS*,
Henrique Dias Pereira dos Santos, *Master in Computer Science, PPGC-UFRGS* e
Leandro Krug Wives, *Associate Professor, PPGC-UFRGS*

***Abstract*—Getting information about other people's opinions is an important part of the purchase decision process. Therefore, consumers publish and obtain information about products in collaborative reviews sites. However, such sites have a huge amount of data that can hinder the user to get a useful product review. This article analyzes textual aspects of reviews, in terms of readability, to identify factors that can be used as indicators of their helpfulness. Considering the set of indicators used in the literature, our experiments show 16 other relevant aspects that can be applied in the automatic assessment of online review helpfulness.**

***Index Terms*—readability, reviews, helpfulness**

I. Introduction

USER-generated online reviews have become essential components of B2C. Empirical studies have shown a significant influence of reviews on the user's purchase decision [1]. Thus, in order to aid consumers to perform their purchase choices, most of the collaborative reviewing sites apply ranking schemes to sort reviews based on their helpfulness. Typically, the helpfulness score of a review is based on votes that are manually given by (other) users, but this approach fails on newly written reviews and those with few votes, configuring the "cold start problem" [2].

Several studies have been addressed to analyze review helpfulness in terms of the interplay between sentiments (favorable, unfavorable and mixed) and product types (search and experience) [3]–[7]. However, the ways in which readability (concerning the easiness in which text reviews could be understood) relates to review helpfulness, remain largely unknown. Hence, in this paper, we aim to answer the following questions: How does readability relate to the helpfulness of user-generated reviews across sentiments and product types? Which features are more correlated to helpfulness in each sentiment?

We conduct our experiments using reviews extracted from a well-known e-commerce website in Brazil called "Buscape"[1] [8]. To analyze the relation of readability with helpfulness, we employed Multiple Linear Regression, using full grid search over Linear Regression, SVM Linear Regression and SVM RBF Regression, as used by Chua in his work [9].

[1]http://www.buscape.com.br/

The rest of this paper is structured as follows. Section II presents some works found in the literature that are related to the context of our research and which comprises the state of the art in terms of Online Review Helpfulness. We also highlight some important aspects of this field. In Section III, we describe the set of readability metrics employed in this work. Section IV describes the methodology used. Section V presents and discuss the results obtained. Finally, in Section VI, we expose our conclusions.

II. Background and Related Work

USUALLY, collaborative reviewing sites require consumers to quantify the sentiment of a given product in the form of a five stars rating. The higher the score (i.e., the number of stars), the more favorable is the opinion of the consumer about the product being reviewed. Several studies have been addressed to the task of analyzing the sentiment influence in the user's perception of the helpfulness, yet, this relation showed not to be precise [9]. Some of these studies are presented bellow.

A. Sentiment and Product Type

There are three competing findings reporting the relation between sentiment and helpfulness. Based on Uncertainty Reduction Theory, according to [10], both favorable and unfavorable reviews can be useful to reduce uncertainty by confirming or eliminating purchasing options.

Secondly, according to Prospect Theory [11], which says that individuals take decisions based on potential losses or gains, consumers might find mixed reviews helpful since they may contain both pros and cons of the product under evaluation. Based on this, mixed reviews are perceived as more valuable than only favorable or unfavorable ones.

Finally, the third one, based on Interpersonal Evaluation Theory [12], states that negative evaluations are perceived as being more clever than lenient ones, and then unfavorable reviews are perceived as entries articulated by an intelligent individual, thereby enhancing their helpfulness [13]. Thus, unfavorable reviews are seen as more valuable than favorable or mixed reviews.

Additionally, studies have shown that there is a variation of the perception of helpfulness across the types of products reviewed. According to [14], goods (products, in our case) can be classified on 'search' and 'experience'. 'Search' products like digital cameras and cell phones are a class of commodities whose quality is easier to assess because it is based on quantitative specifications [15]. On the other hand, 'experience' products such as music albums and services are those whose quality is harder to assess since it is based on qualitative aspects and other users' experiences [1]. Thus, consumers tend to explore the opinions about 'experience' products more than they would do for 'search' products.

III. Readability

THERE are many traditional methods for assessing comprehension difficulty of a text. Previous work on readability [16] guided the development of the metrics used in our research and they can be classified into the following categories: 'Descriptive', 'Word information', 'Diversity', 'Connectives', 'Readability', and 'Ratios'. Those types and their corresponding metrics are detailed next.

A. Descriptive

Descriptive metrics provide basic information about the text, such as the number of sentences and words. They help to check the output to make sure that the information makes sense, as well as being important values for computing more sophisticated metrics. In fact, mean word length in syllables combined with mean sentence length is a standard readability metric that makes the core principles of the Flesch-Kincaid and Lexile readability scores, remaining some of the most important metrics for measuring textual readability. The descriptive metrics used in this work are the following:

- **Word, sentence and syllable count.** The number of words, sentences and syllables in a text.
- **Words per sentence.** Some statistical data regarding the size of sentences in words. For instance, in this work, we calculate the mean, median, percentiles (25, 50, 75 and 90) and also the percentage of sentences above 30 words long.
  The average (mean) sentence length is a classical feature in readability. For example, the work of [17] explored the use of the 90th percentile sentence and inspired the computing of the other percentiles and the median.
- **Syllables per word.** Statistical data of the length of words in syllables. We also compute the mean, median and four percentiles (25, 50, 75 and 90). It is important to state that Flesch [18] found that the average number of syllables per word had a correlation of .66 with the comprehension difficulty.

B. Word information

Metrics in this category are based on the concept that to each word is assigned a syntactic part-of-speech category. These categories are separated in content words (adjectives, nouns, verbs, and adverbs) and function words (determiners, adpositions[2], pronouns and conjunctions). These metrics are a reflex of the elements of a text that are likely to support a reader's construction of a coherent situation model [19]. A single part-of-speech category is attributed to each word of the text based on its syntactic context. The parser of our choice, *NLPNET* [20] achieves about 97% of accuracy when compared to other state-of-the-art taggers for the Portuguese language [21].

- **Incidences of part-of-speech elements.** We calculate the rate of word categories (adjectives, nouns, verbs, adverbs, pronouns) per 1000 words in the text, and also the incidence of content and function words per 1000 words.

C. Diversity

Diversity metrics provide information on the variety of words in a text. For instance, a high word diversity means many unique words need to be decoded and integrated with the discourse context, which should make comprehension more difficult. In contrast, if some words are being more often repeated, it tends to increase cohesion, and, thus, make for a more readable text. Otherwise, it may appear as a lack of vocabulary. Word diversity is a common metric on the measuring of readability, with works as early as [22], which found a correlation between difficulties of passages and the mean frequency of the words in such passages.

- **Lexical diversity.** This measures how varied is the vocabulary of a text. It is defined as the ratio of unique words in the text in comparison to the total number of words in the same text.
- **Content diversity.** This metric is similar to Lexical diversity, but it only takes in consideration content words (e.g., adjectives, nouns, verbs and adverbs). It is defined as the ratio of unique content words that appear in the text in comparison to the total number of content words in the text.

D. Connectives

The traditional unit for analyzing grammatical complexity has been the sentence, i.e., the text enclosed with a starting capital letter and a punctuation mark. However, [23] found evidence that independent clauses might be a more valid unit of analysis. For instance, sentences with connectives such as "O telefone é leve e rápido" (The phone is light and fast) may be treated as if there were two separate syntactic units since the conjunction "e" (and) serves roughly the same function as a punctuation mark. Hence, although the presence of connectives creates longer sentences, their use might decrease the difficulty of understanding of a text by creating cohesive links between ideas and providing clues about text organization [19], [24]–[27]

- **Incidence of connectives.** Connectives are divided into five general classes [28], [29], additive, logic, temporal, causal and negative. We calculate the incidence

[2] adposition is a cover term for prepositions and postpositions

of the total number of connectives, as well as the extent of each distinct category. For this, we created a dictionary of connectives for the Portuguese language.

- **Logic operators.** Within the logical connectives, there are some specific logical particles like “e” (and), “ou” (or), “se” (if), and “não”, “nem”, “nenhum”, etc (not, neither, none, etc). Measuring such logical particles and how they relate with cohesion and readability in a text was explored by Coh-Metrix-Port [30].

E. Readability

There are many traditional methods for assessing text difficulty. Klare stated that more than 40 readability formulas had been developed over the years [31]. The most common, however, are the Flesch-Kincaid formulas. Since the Flesch reading ease score has a validated Portuguese adaptation [32].

- **Flesch reading ease, Portuguese version.** The Flesch reading ease adaptation for Brazilian Portuguese developed by [32], and it basically consists on a shift of 42 points on the result of the original Flesch reading ease formula to compensate for the fact that words in Portuguese typically have a bigger number of syllables.

$$RE = 206.835 - (1.015 * ASL) - (84.6 * ASW) \quad (1)$$

Where RE is the Readability Ease, ASL is the Average Sentence Length (i.e., the number of words divided by the number of sentences) and ASW is Average number of syllables per word (i.e., the number of syllables divided by the number of words).

F. Ratios

In contrast to the incidence metrics, which compute the number of a specific word type in a span of 1000 words, ratios are a more relative measure, comparing the incidence of a certain class of units to the incidence of another class of units. For example, [33] demonstrated that the ratios of some part-of-speech elements of the text could be a good predictor of the text’s complexity.

- **Ratio of pronouns on prepositions.** Inspired by the model of [17], we implemented a metric of pronouns on prepositions as a measure of the syntactic complexity of sentences. This specific ratio was found to be related to the readability of texts in the French language and inspired us to check if it could be an interesting metric to compute for the Portuguese.
- **Ratio of verbs on nouns.** Considering the Portuguese language, the ratio of verbs on noun could indicate a good use of transitive verbs in the text.

IV. Methodology

In this section, we describe how our experiments were performed. Firstly, we present the dataset we have used. In addition, we define the helpfulness formula we have used and present several features that we have investigated for the evaluation of reviews’ helpfulness in terms of sentiments expressed.

A. The Dataset and its preprocessing task

We conduct our analysis using Brazilian Portuguese reviews, motivated mostly by the lack of Natural Language Techniques resources to this language. The corpus used for this study comes from a larger Brazilian e-commerce site called “Buscape”[3], which is a well-known e-commerce Website in Brazil [8].

The original corpus has 85,910 reviews containing *search* and *experience* products. For each review, the following attributes were used: rating (rate of a product as given by the reviewer), text (i.e., the review content in natural language), the number of positive and negative votes, and the (total) number of votes. We follow the methodology described in [34] by considering only reviews that have received more than five votes, either positive or negative. Additionally, to eliminate uninformative comments, we have considered only reviews with more than 10 words. Finally, we opted to examine only reviews related to search products, since, in Buscape corpus, were left a low quantity of experience products after the previous filtering (in future works, we intend to get additional reviews from other sites such as http://www.amazon.com/). Table I describes the categories of products of the remaining dataset containing 7,262 reviews.

TABLE I
Top 10 categories of ’search’ products reviews extracted from Buscape.

Category	**Reviews**
Cell Phones	1,985
Televisions	1,617
Digital Cameras	694
Washers	466
Refrigerators	397
Air Conditioners	349
Tablets	340
Laptops	268
Video-games Consoles	227
Printers	197
Others	722

To better understand the user behavior on this domain, Table II summarizes the distribution of votes, rating, word count and helpfulness across sentiments (favorable, unfavorable and mixed). It shows that users gave more positives votes (i.e. people who considerate this review helpful) than negative in all sentiments (favorable, unfavorable, mixed). It also indicates that sentiments have no influence on the user’s perception of utility.

B. Helpfulness definition

Previous works on Online Review Helpfulness assessment have defined review helpfulness as a proportion of users who gave a positive vote [1], [34]. Thus, we specified our helpfulness function as follows:

[3]http://www.buscape.com.br/

TABLE II
DESCRIPTIVE STATISTICS (MEAN ± SD) OF REVIEWS AS A FUNCTION OF THEIR SENTIMENT.

	Favorable (1,666)	**Unfavorable (490)**	**Mixed (5,106)**
Total votes	15.16 ± 17.01	21.57 ± 35.16	17.88 ± 22.99
Positive votes	9.98 ± 12.62	14.78 ± 26.47	13.61 ± 19.12
Negative votes	5.18 ± 7.51	6.79 ± 12.28	4.26 ± 6.72
Word Count	62.56 ± 59.99	95.20 ± 71.19	88.06 ± 73.43
Helpfulness	0.65 ± 0.28	0.66 ± 0.29	0.75 ± 0.23

$$h(r \epsilon R) = \frac{vote_{+}(r)}{vote_{+}(r) + vote_{-}(r)} \tag{2}$$

where $votes_{+}(r)$ is the number of people who find a review useful and $votes_{-}(r)$ is the number of users that do not consider a review useful.

C. Data Analysis

Previous works [9], [34] also present an investigation of the relation of readability aspects with review's helpfulness. Similarly, we use the readability features as independent variables and the helpfulness as the dependent variable.

We begin evaluating each type of linguistic feature in isolation to investigate its predictive power of user review helpfulness; we then examine them together in various combinations to find the most useful feature for modeling user review helpfulness. Although predicted helpfulness ranking could be directly used to compare the helpfulness of a given set of reviews, predicting helpfulness rating is desirable in practice to examine the helpfulness between existing reviews and newly written ones without re-ranking all previously ranked reviews.

We adopted the Spearman correlation coefficient to perform significance tests since it is the most commonly used measure of correlation between two sets of ranked data points. Thus, we created three models using different supervised machine learning algorithms in our experiments: a) Linear Regression (Linear); b) Support Vector Machine (SVM-L) based on linear kernel; and c) SVM based on radial basis function (SVM-R). To assess the performance of each model we used 10-fold cross-validation, and found that Linear Regression was the model that showed better accuracy (details in Figure 1). For the remaining test fold, each set of product reviews was ranked according to the regression model prediction for the helpfulness value.

V. RESULTS

WE conducted a series of experiments to evaluate the utility of the proposed helpfulness linguistic features. The main focus of this paper is to analyze the readability features as a function of helpfulness and sentiment. Thus, we have followed the methodology employed on previous works on this field [6], [34] to reveal relevant textual characteristics that can be used as predictors to a helpfulness assessment.

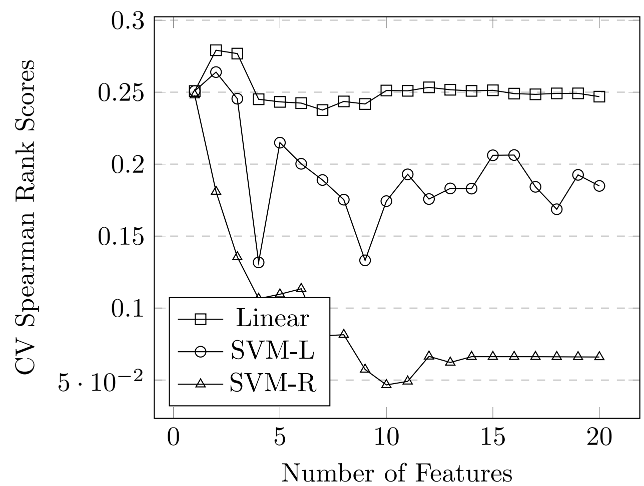

Fig. 1. Regression Model Evaluation for Buscape's Reviews.

As indicated earlier, considering the dataset of 7,262 reviews, the distribution of review sentiment was as follows: favorable (1,666), unfavorable (490), and mixed (5,106). The dominant proportion of favorable reviews was expected. Each review was quantified using a set of 33 metrics, as previously described in Section III. Afterward, each feature was independently employed to train a Linear Regression model. Table III shows 16 (out of the 32) features that achieved better correlation ($\beta > 0.10$) with the dependent variable (helpfulness).

Among all reviews sentiments, helpfulness was found to be very related to syllable and sentence count ($\beta > 0.20, p < 0.001$). It is an important feature to measure review helpfulness, and it had a similar behavior in other works [6], [34]. The helpfulness of favorable product reviews was positively related to the lexical diversity ($\beta = 0.16, p < 0.001$), and the 25 percentile of word length ($\beta = 0.20, p < 0.001$). Reviews with rich lexical diversity and with higher word length were likely to be voted as helpful.

The helpfulness ratio of unfavorable product reviews was positively related to various incidences of features like logical connectives (negation and operators with $\beta = 0.20, p < 0.001$), functional incidence and pronouns incidence ($\beta = 0.22, p < 0.001$). Unfavorable reviews that are rich in lexical features were likely to be voted as helpful.

TABLE III
MULTIPLE REGRESSION RESULTS FOR FEATURES OF THE REVIEWS, WITH $p < 0.001$.

Features	**Favorable (1666)**	**Unfavorable (490)**	**Mixed (5106)**
Syllable Count	0.28	**0.30**	**0.26**
Sentence Count	**0.28**	0.21	0.23
Percentile 25 Word Length	**0.20**	0.04	0.05
Lexical Diversity	0.16	0.11	0.11
Content Diversity	0.11	0.05	0.06
Connective Temporal Incidence	0.10	0.13	0.06
Connective Casual Incidence	0.09	0.12	0.05
Pronouns Incidence	0.09	**0.22**	0.16
Logic Negation Incidence	0.09	**0.20**	0.09
Verb Incidence	0.09	0.13	0.07
Logic Operators Incidence	0.08	**0.20**	0.09
Adverb Incidence	0.07	0.14	0.05
Mean Sentence Length	0.05	0.10	0.01
Adjective Incidence	0.05	0.15	0.03
Average Word Per Sentence	0.05	0.10	0.00
Content Incidence	0.05	0.11	0.02
Functional Incidence	0.03	**0.22**	0.07
Pronouns on Prepositions Ratio	0.02	0.13	0.10
Verbs on Nouns Ratio	0.06	0.13	0.09

VI. CONCLUSION

ALTHOUGH our experiments were performed using the Portuguese language, the metrics used in this paper can be applied to tune the accuracy of the models in other languages. Despite the difference between peer reviews and other types of reviews, our work demonstrates that many generic linguistic features are also effective in predicting user-review helpfulness. Therefore, we analyzed the information quality dimension of reviews in order to reveal textual characteristics that can be used as indicators of the helpfulness of reviews. Thus, in addition to the set of predictors employed in previous works, the main contribution of our work is a collection of another 16 statically relevant predictors that can be successfully applied to assess the helpfulness of reviews.

The gold standard stated [3], [34] showed that the best single feature to predict helpfulness was 'sentence count', but here we show that, at least in our experiments, 'syllable count' has a better correlation. Near to syllable count, pronouns and functional incidence are good features to measure helpfulness for unfavorable reviews.

The behavior of the users in this domain is slightly different compared with similar works that consider the number of words [9], [34]. The word count in the favorable sentiment category has a lower mean (62.56) than in unfavorable reviews (95.20). It may indicate that users who express unfavorable sentiments tend to explain their discontent using more words than the users that have expressed a favorable opinion about a product.

Another interesting characteristic on this domain is that unfavorable reviews tend to receive more votes than favorable or mixed ones. It seems that consumers are more interested in unfavorable reviews than other ones.

In this study, we also provide valuable insights about the user behavior on the analyzed domain, such as that consumers give more positives votes than negative in all sentiments confirming the presumption that the sentiments have no influence on the user's perception of utility.

The primary focus of this paper was to provide an analysis of readability features to reveal relevant textual characteristics that can be used as predictors to a helpfulness assessment. Thus, as future work, we highlight the importance of combining features from other dimensions to create a model capable of assessing the helpfulness of Online Reviews. We are also considering the application of these features in the domain of Recommender Systems, aiming to assess the similarity of users based on the textual characteristics of their reviews and comments, and also to estimate the similarity of items based on their reviews. This may be relevant to recommend more appropriated products and services to users.

REFERENCES

[1] S. M. Mudambi and D. Schuff, "What makes a helpful review? a study of customer reviews on amazon. com," *MIS quarterly*, vol. 34, no. 1, pp. 185–200, 2010.

[2] X. N. Lam, T. Vu, T. D. Le, and A. D. Duong, "Addressing cold-start problem in recommendation systems," in *Proceedings of the 2Nd International Conference on Ubiquitous Information Management and Communication*, ser. ICUIMC '08, New York, NY, USA: ACM, 2008, pp. 208–211, ISBN: 978-1-59593-993-7. DOI: 10.1145/1352793.1352837. [On-

line]. Available: http://doi.acm.org/10.1145/1352793.1352837.

[3] W. Xiong and D. Litman, "Automatically predicting peer-review helpfulness," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 502–507. [Online]. Available: http://www.aclweb.org/anthology/P11-2088.

[4] A. Mukherjee and B. Liu, "Modeling review comments," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jeju Island, Korea: Association for Computational Linguistics, Jul. 2012, pp. 320–329. [Online]. Available: http://www.aclweb.org/anthology/P12-1034.

[5] Y.-C. Zeng and S.-H. Wu, "Modeling the helpful opinion mining of online consumer reviews as a classification problem," in *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, Nagoya, Japan: Asian Federation of Natural Language Processing, Oct. 2013, pp. 29–35. [Online]. Available: http://www.aclweb.org/anthology/W13-4205.

[6] A. Y. Chua and S. Banerjee, "Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth," *Journal of the Association for Information Science and Technology*, vol. 66, no. 2, pp. 354–362, 2015, ISSN: 2330-1643. DOI: 10.1002/asi.23180. [Online]. Available: http://dx.doi.org/10.1002/asi.23180.

[7] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1014–1023. [Online]. Available: http://www.aclweb.org/anthology/P15-1098.

[8] N. Hartmann, L. Avanço, P. P. Balage Filho, M. S. Duran, M. d.G. V. Nunes, T. A. S. Pardo, S. M. Aluísio, *et al.*, "A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words.," in *LREC*, 2014, pp. 3865–3871.

[9] A. Y. Chua and S. Banerjee, "Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality," *Computers in Human Behavior*, vol. 54, pp. 547 –554, 2016, ISSN: 0747-5632. DOI: http://dx.doi.org/10.1016/j.chb.2015.08.057. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S074756321530131X.

[10] C. Forman, A. Ghose, and B. Wiesenfeld, "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets," *Information Systems Research*, vol. 19, no. 3, pp. 291–313, 2008, cited By 356. DOI: 10.1287/isre.1080.0193. [Online]. Available: https://www.scopus.com/inward/record.url?eid=2-s2.0-48449101733&partnerID=40&md5=f17d25f2f9bb185a0eae7a4fdb7ee752.

[11] A. Kühberger, "The influence of framing on risky decisions: A meta-analysis," *Organizational behavior and human decision processes*, vol. 75, no. 1, pp. 23–55, 1998.

[12] T. M. Amabile and A. H. Glazebrook, "A negativity bias in interpersonal evaluation," *Journal of Experimental Social Psychology*, vol. 18, no. 1, pp. 1–22, 1982.

[13] P. F. Wu, H. Van Der Heijden, and N. Korfiatis, "The influences of negativity and review quality on the helpfulness of online reviews," in *International conference on information systems*, 2011.

[14] P. Nelson, "Information and consumer behavior," *Journal of Political Economy*, vol. 78, no. 2, pp. 311–329, 1970, ISSN: 00223808, 1537534X. [Online]. Available: http://www.jstor.org/stable/1830691.

[15] L. M. Willemsen, P. C. Neijens, F. Bronner, and J. A. de Ridder, "“highly recommended!” the content characteristics and perceived usefulness of online consumer reviews," *Journal of Computer-Mediated Communication*, vol. 17, no. 1, pp. 19–38, 2011, ISSN: 1083-6101. DOI: 10.1111/j.1083-6101.2011.01551.x. [Online]. Available: http://dx.doi.org/10.1111/j.1083-6101.2011.01551.x.

[16] M. J. B. Finatto, C. E. Scarton, A. Rocha, and S. Aluísio, "Características do jornalismo popular: Avaliação da inteligibilidade e auxílio à descrição do gênero," in *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, 2011.

[17] T. François and C. Fairon, "An ai readability formula for french as a foreign language," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, 2012, pp. 466–477.

[18] R. Flesch, "A new readability yardstick.," *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[19] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behavior research methods, instruments, & computers*, vol. 36, no. 2, pp. 193–202, 2004.

[20] E. R. Fonseca and S. M. Aluísio, "A deep architecture for non-projective dependency parsing," in *Proceedings of NAACL-HLT*, 2015, pp. 56–61.

[21] E. R. Fonseca and J. a.L. G. Rosa, "Mac-morpho revisited: Towards robust part-of-speech tagging," in *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013, pp. 98–107.

[22] I. Lorge, "Predicting readability," *The Teachers College Record*, vol. 45, no. 6, pp. 404–419, 1944.

[23] E. B. Coleman, "Improving comprehensibility by shortening sentences," *Journal of Applied Psychology*, vol. 46, no. 2, pp. 131–134, 1962. DOI: 10.1037/h0039740.

[24] K. Cain and H. M. Nash, ""the influence of connectives on young readers' processing and comprehension of text," *Journal of Educational Psychology*, vol. 103, no. 2, pp. 429–441, 2011. DOI: 10.1037/a0022824.

[25] A. Crismore, R. Markkanen, and M. S. Steffensen, "Metadiscourse in persuasive writing a study of texts written by american and finnish university students," vol. 10, no. 1, pp. 39–71, 1993.

[26] T. J. Sanders and L. G. Noordman, "The role of coherence relations and their lingustic markers in text processing," *Discourse processes*, vol. 29, no. 1, pp. 37–60, 2000.

[27] W. J. V. Kopple, "Some exploratory discourse on metadiscourse," *College composition and communication*, vol. 36, no. 1, pp. 82–93, 1985. [Online]. Available: http://www.jstor.org/stable/357609.

[28] M. Halliday and R. Hasan, *Cohesion in english*. London: Longman, 1976.

[29] M. Louwerse, "An analytic and cognitive parametrization of coherence ralations," *Cognitive Linguistics*, vol. 12, no. 3, pp. 291–316, 2001.

[30] C. E. Scarton and S. M. Alu'isio, "Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: Adaptando as métricas do cohmetrix para o português," *Linguamática*, vol. 2, no. 1, pp. 45–61, 2010.

[31] G. R. Klare, "A second look at the validity of readability formulas," *Journal of literacy Research*, vol. 8, no. 2, pp. 129–152, 1976.

[32] T. B. Martins, C. M. Ghiraldelo, M. d.G. V. Nunes, and O. N. de Oliveira Junior, *Readability formulas applied to textbooks in brazilian portuguese*. Icmsc-Usp, 1996.

[33] J. R. Bormuth, "Readability: A new approach," *Reading research quarterly*, vol. 1, no. 3, pp. 79–132, 1966. [Online]. Available: http://www.jstor.org/stable/747021.

[34] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '06, Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 423–430, ISBN: 1-932432-73-6. [Online]. Available: http://dl.acm.org/citation.cfm?id=1610075.1610135.