Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks

Flavio Carvalho[1], **Rafael Guimarães Rodrigues**[1], **Gabriel Santos**[1],
Pedro Cruz[1], **Lilian Ferrari**[2], **Gustavo Paiva Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

[2]UFRJ - Universidade Federal do Rio de Janeiro
Av. Brigadeiro Trompovisky s/n - Rio de Janeiro - RJ - Brasil.

flavio.carvalho@eic.cefet-rj.br, rafael.rodrigues@cefet-rj.br,
gabriel.santos@eic.cefet-rj.br, pedroparreiracruz@gmail.com,
lilianferrari@uol.com.br, gustavo.guedes@cefet-rj.br

Abstract. *LIWC is a text analysis program that categorizes words into grammatical and psychologically derived categories. The currently available LIWC lexicon for Brazilian Portuguese (LIWC_2007pt) is based on the* 2007 *version of LIWC program. As several studies indicated, LIWC_2007pt shows performance and categorization problems. In this scenario, this work highlights a new Brazilian Portuguese LIWC lexicon (LIWC_2015pt), based on LIWC* 2015 *program. This work compares the performance of LIWC_2007pt and LIWC_2015pt in classification tasks. Three experiments were conducted and the results indicate LIWC_2015pt outperforms LIWC_2007pt in all three tasks.*

1. Introduction

In the last years, data produced in social networks and other sources, such as message exchange applications, have been used to obtain useful information by identifying patterns and trends with algorithms and methods from machine learning and statistics areas [Moreira et al. 2018, Loures et al. 2017]. Before using the algorithms and methods and train machines over complex variants of mathematical models, a computational system needs to transform textual data into numerical representations, in processes such as vectorization [Liu 2012, Zhang et al. 2010]. Although these processes can be quite complex, the inherent complexity of natural language texts can be reduced utilizing lexical approaches [Grimmer and Stewart 2013].

One of the available lexical approaches for text processing adopts the Linguistic Inquiry Word Count (LIWC) program [Pennebaker et al. 2015]. It is possible to obtain different types of information from social networks users with LIWC, such as political tendencies [Caetano et al. 2017], social and economic status [Pettijohn and Sacco Jr 2009], among others. LIWC can also be used to analyze texts for health studies, e.g. where the usage of several LIWC categories shows significant differences for an Alzheimer's disease group, suggesting that the method could be used for dementia screening [Shibata et al. 2016].

An essential part of LIWC, beyond the main program, is the LIWC lexicon. It was developed to analyze emotional, social, cognitive and structural components of texts according to many categories associated to these aspects, considering the number of words that the program finds in the texts. Throughout the years, studies have been conducted in order to improve the LIWC lexicon, so that categories containing linguistic, social and psychological meaningful words could bring information to better reflect any authors' psychological processes, emotions, and social relationships [Pennebaker and Chung 2011, Ireland et al. 2011, Tausczik and Pennebaker 2010].

The most recent LIWC lexicon was released in 2015 with a new version of the program [Pennebaker et al. 2015].This lexicon - hereafter, LIWC_2015en, introduces several new categories, improving and refining the results of LIWC program for the analysis of texts in English [Pennebaker et al. 2015]. While previous versions of LIWC lexicon have been translated from English into different languages, LIWC_2015en has only been translated into German [Meier et al. 2019], Chinese [Zeng et al. 2018] and Dutch [Van Wissen and Boot 2017], to the best of our knowledge.

As for Brazilian Portuguese (BP), there is a lexicon based on the 2007 version of LIWC English lexicon [Balage Filho et al. 2013]. In this work, this lexicon is abbreviated as LIWC_2007pt. A search using Google Scholar[1] returns 44 exclusive quotes for this publication.

Observing the citation counts per year, from 2013 to 2018, we can note a growth in the number of citations to the publication introducing LIWC_2007pt lexicon, which suggests an increasing importance of this resource in academic studies in Portuguese. However, since the first published evaluations with LIWC_2007pt, some issues related to the performance of negative valence detection can be noticed [Balage Filho et al. 2013, Rodrigues and Guedes 2017]. Recent studies are also indicating several problems with this lexicon regarding spelling mistakes and words with problems related to categorization, which negatively impacts obtained results [Carvalho et al. 2018a, Carvalho et al. 2018b].

While these issues can be addressed and corrected, we also notice the fact that there is a more recent version of LIWC for English, that was developed after the release of LIWC_2007pt lexicon in 2011. To use the features of the 2015 version of LIWC program in the analysis of Portuguese texts, a Portuguese version of the lexicon with the same structure and categories as the LIWC_2015en should be available, but we are not aware of the development of one so far.

To address at the same time both problems with word spelling and categorization, as well as the introduction of several new categories, we developed a new BP lexicon resource with all the categories present in the 2015 English version of LIWC lexicon. This work evaluates our recently developed lexicon, abbreviated as LIWC_2015pt[2]. Instead of relying solely on a large number of words, the focus is to adjust words to categories that appropriately match linguistic, social and psychological characteristics to achieve better results in tasks associated with the use of LIWC for classification and Sentiment Analysis.

This work is structured so that after this introductory section, we present in section 2 works related to the evaluation of LIWC lexicons in other languages. In section 3,

[1] `https://scholar.google.com/`

[2] Access the following link to read instructions on how to cite and download the LIWC_2015pt: `https://github.com/LaCAfe/LIWC2015pt`.

we detail the materials and procedures for evaluating LIWC_2015pt lexicon. Section 4 discusses the results and section 5 presents a discussion about this work.

2. Related Work

Previous versions of the LIWC lexicon are available in different languages, such as Catalan, Spanish, French, Italian and Serbian [Bjekić et al. 2014, Massó et al. 2013, Piolat et al. 2011, Ramirez-Esparza et al. 2008, Alparone et al. 2004], among others[3]. As previously mentioned, LIWC_2007pt is a BP version of the 2007 English LIWC lexicon, which contains about $127,000$ words in 64 categories. As methodological references to evaluate the LIWC_2015pt lexicon, we searched for works that take into account the evaluation of the LIWC_2007pt lexicon and, also, works presenting the 2015 version of the LIWC lexicon in other languages.

In the evaluation of the use of the LIWC_2007pt lexicon for sentiment classification in BP texts, just the 'positive emotion' (*posemo*) and 'negative emotion' (*negemo*) categories of the lexicon were used for comparison against the Portuguese version of both the Opinion Lexicon and the SentiLex [Balage Filho et al. 2013]. The evaluations analyzed the pairwise agreement between lexicons, i.e. the number of lexicon entries with equal polarity and also measured the performance of each lexicon in the sentiment classification task using an algorithm similar to the SO-CAL [Taboada et al. 2011]. The results indicated that the LIWC_2007pt lexicon performs better in indicating positivity than negativity.

Our work differs from the publication introducing LIWC_2007pt in that it brings the comparison of classification using values from all the available categories of both the LIWC_2007pt and the LIWC_2015pt lexicons. Also, we have chosen from five different algorithms that are applicable to our task and practical to implement using off-the-shelf software tools, which contributes to the replication of this work and makes it easy to obtain results from/with any collection of texts.

Pennebaker *et al.* [2015] evaluate the 2015 version of the LIWC text analysis program using the LIWC_2015en lexicon to analyze collected text samples from a variety of studies. Then, to compare LIWC_2015en and LIWC_2007en and assist in the transition to the new version of the LIWC program, they present a table with the means, standard deviations, and correlations between the two lexicons being used to analyze the same texts samples. This is used in order to get a sense of the degree to which language varies across a variety of settings, but differing from our work, no classification results are shown in their evaluation.

In the work presenting the first LIWC translation based on the 2015 lexicon, the results of the analysis with the Dutch LIWC lexicon are compared with the results with the English lexicon, using a parallel corpus [Van Wissen and Boot 2017]. Called Dutch Parallel Corpus, this corpus is composed of Dutch texts placed alongside English texts from fields such as finance, science, culture and communication [Paulussen et al. 2013]. Results of equivalence test on translated Dutch and English lexicons are shown in tables with computed Pearson correlation coefficient or using Spearman's rank correlation coefficient, along with the values of Cohen's d for effect size [Van Wissen and Boot 2017].

[3] LIWC lexicons are available for download in `www.liwc.net/dictionaries`.

However, the 2015 Dutch lexicon is not compared against any earlier version of LIWC lexicon in the same language.

3. Materials

3.1. LIWC

LIWC is a text analysis application developed with the objective of analyzing emotional, cognitive and structural components from texts [Pennebaker et al. 2015]. It can be divided into two main parts, one is the main program and the other is a lexicon. The lexicon contains words in one or more categories that reflect linguistic, psychological, and social processes, like 'pronouns' (*pronoun*), 'positive emotions' (*posemo*), 'social processes' (*social*) and so on.

The main categories in LIWC lexicon are divided into subcategories. For example, the 'pronoun' category is divided into two subcategories: 'personal pronouns' (*ppron*) and 'impersonal pronouns' (*ipron*). The subcategory *ppron* is divided into 5 more subcategories: 'first person singular' (*i*), 'first person plural' (*we*), second person (*you*), 'third person singular' (*shehe*) and 'third person plural' (*they*).

LIWC can process any number of text files in different formats, even texts within spreadsheets from popular software applications. To process a large number of text files, it is possible to either put them in a directory or also to select multiple files within a directory. LIWC searches for each word in the text for a match with a lexicon equivalent word or word stem, and increments the percentage value of the appropriate word category if it is found [Pennebaker et al. 2015]. After processing all the selected text files, an output file is saved in a format that uses delimited text, with the variable names on the first line and, for each analyzed text file, the percentage values for each category on the subsequent lines. This output is a vectorial representation of the text file using the lexical categories as dimensional attributes, exemplified in Table 1.

Table 1. Example of an output file from LIWC after the processing of five text samples, showing some of the lexicons' categories and the percentage values of words in each text.

File	pronouns	verbs	affect	social	cogproc	bio	timeorient	informal
1	22.74	2.44	21.47	2.77	3.29	13.77	3.24	0.26
2	8.28	8.30	15.05	3.03	4.57	8.26	3.69	0.98
3	11.89	17.23	11.41	1.23	3.35	10.26	5.40	4.09
4	17.70	18.04	17.30	5.31	3.14	11.75	2.62	1.99
5	9.94	13.62	26.57	0.40	2.81	10.06	6.04	5.16

3.2. LIWC_2007pt

LIWC_2007pt lexicon is the BP version of 2007 LIWC [Balage Filho et al. 2013]. In the process of developing a lexicon with 127,149 words arranged in 64 categories, conjugations were automatically included using the NILC Unitex-PB dictionary. LIWC_2007pt lexicon was partially translated manually, but NILC's page informs that the manual work of translation was not revised, and also that lexicon can be improved[4].

LIWC_2007pt lexicon has problems according to previous studies [Carvalho et al. 2018a]. In the analysis of pronoun category (*pronoun*), the authors found that, from a total of the 128 words classified as pronouns (and included in such category of LIWC_2007pt), 40 should not be included, according to grammar resources and linguistic specialists. Likewise, the authors highlight the same situation with 8 out of 54 words classified in personal pronoun category (*ppron*). In the category of impersonal pronouns, 49 out of 88 words should not be associated with this category, such as the personal pronouns 'ele' ('he') and 'ela' ('she').

3.3. LIWC_2015pt

Since LIWC_2015en was developed after years of exploratory study of emotional, cognitive and structural components of speech samples [Pennebaker et al. 2015], we decided to take leverage of this for the development of LIWC_2015pt. As so, we used for comparison both the LIWC_2015en and also the previous version of the LIWC English lexicon, released in 2007. This approach aided in the challenging task of assigning the words into the categories according not only to their linguistic, but also their psychological and social processes.

The LIWC_2015pt has a total of 73 categories an $14,459$ words, which are related to different psychological, social and linguistic features. The LIWC_2015pt lexicon is larger than LIWC_2015en lexicon, which has $6,400$ words. The reason for this is that it is necessary to have words with semantic differentiation, including variations of gender, number, grade, etc. It is worth noting that previous studies have shown that the larger number of words in LIWC_2007pt ($127,000$) may not have a positive impact on the number of words to be identified and counted in the texts, considering the problems ascertained in other studies [Carvalho et al. 2018b].

3.4. Datasets

After developing LIWC_2015pt lexicon, we searched for datasets containing BP posts from different social networks. We collected two labeled sets, one with data extracted from MQD[5] and another from Twitter. In this work we refer to the MQD dataset as `MBAL18k` and to the dataset from Twitter as `TAS-PT`.

The `MBAL18k` dataset is a collection of randomly selected posts from MQD. After downloading the posts, we removed entries not relevant to our analysis, like sequences of just some random meaningless letters, e.g. 'xuogfjeil' and/or 'gdsgehsg', and also entries with just url links. From the remaining entries, data representing textual content in BP is divided into 3 classes: 10s, 20s and 30s, containing users of both gender with ages from 13 to 17, 23 to 27 and 33 to 42, respectively, for a clearer differentiation. The set is balanced according to the number of entries in each class, resulting in $6,000$ entries in each class (with a total of $18,000$ entries).

[4] `http://www.nilc.icmc.usp.br/portlex/index.php/en/projects/liwc`, as accessed in May 13, 2019.

[5] `http://www.meuqueridodiario.com.br` is a Brazilian social network.

TAS-PT has two files with the numerical identification (ID) from tweets in BP, one with IDs from tweets of positive sentiments and another with IDs from tweets of negative sentiment [Cavalcante and Malheiros 2017]. In these files, no textual content from the tweets is available. In this case, the Twitter API[6] is needed to obtain the textual content from the messages using the IDs information from the files. Connecting to Twitter API and downloading the textual content of TAS-PT allowed the creation of another dataset with $59,260$ files, which we named TSN-60k, containing $28,853$ files labeled as negative and $30,407$ labeled as positive.

4. Experiments

For the evaluation of LIWC_2015pt lexicon in classification tasks, typical of the area of Sentiment Analysis, we conducted sequences of experiments in which we loaded the previously mentioned datasets into the LIWC program. LIWC analyzed the textual content first using all categories from the LIWC_2015pt lexicon, and then using all categories from the LIWC_2007pt lexicon. The main objective is to compare the performance between the LIWC_2015pt and the LIWC_2007pt lexicons.

First, we produced a classification experiment to predict the age of users in text using MBAL18k dataset. Then, we conducted the experiments for classification of polarity of emotions using TSN-60k dataset. In order to generate the results, we used Weka [Hall et al. 2009] to produce the experiments with classification algorithms.

After processing the texts with LIWC_2015pt and LIWC_2007pt lexicons, the files generated by LIWC are used in the classification task, applying the following algorithms: Naive Bayes (NB), Multinomial Naive Bayes (MNB), J48, Random Forest (RF) and Logistic Model Trees (LMT). We selected NB and MNB as they are basic reference methods for classifying text [Wang and Manning 2012]. Also, RF and J48 are included since they provide good results in classifying texts [Fersini et al. 2015, Gabrilovich and Markovitch 2004]. Another algorithm chosen was LMT, because it is one of the best algorithms to classify texts using stylistic resources of Portuguese [Aires et al. 2004], even presenting good results in instances with LIWC as one of the resources used in some tasks (e.g., detection of satire, detection of sarcasm) [Ravi and Ravi 2017].

For each of the algorithms, we maintained the initial set of configurations from Weka 3.8.3 version. We obtained the mean of the F_1 Score to measure and evaluate the results of the classification algorithms. We used k-fold cross-validation technique with ten partitions [Kohavi et al. 1995] to acquire values for precision (P), recall (R) and the F_1 Score for each of the algorithms. F_1 Score ranges between 0 and 1, where 0 indicates the worst value and 1 indicates the best value, i.e. perfect values of P and R. F_1 Score is a harmonic mean of P and R, expressed by $F_1 = 2 \cdot \frac{P \cdot R}{P+R}$.

To determine whether there is statistical evidence that F_1 Score on the two sets are significantly different, we used paired T-tests to compare the F_1 Score against chance level with p-value, i.e. threshold of statistical significance, of 0.01. As a required prior condition to use paired T-tests, we used the Shapiro-Wilk test to check if sample distribution is normal, since it is a type of parametric method that can be used when the samples

[6] https://developer.twitter.com/en/docs/api-reference-index.html

satisfy the conditions of normality [Peat and Barton 2008, Öztuna et al. 2006]. In both classification experiments the results of the Shapiro-Wilk test assesses the normality of the values.

Table 2 presents the mean of the F_1 Score from inference of the age group of users of MQD with the NB, MNB, RF, J48 and LMT algorithms. It can be noted that all five algorithms perform better with the use of files processed using the LIWC_2015pt lexicon than with LIWC_2007pt lexicon. The value of F_1 Score using the algorithm LMT was the one that presented the best result (0.568).

We then used paired T-test to compare the F_1 Score from the inference of the age group of users of MQD in Table 2. Using T distribution (DF=4, two-tailed), the paired T-test shows that the difference between the average of the LIWC_2015pt minus LIWC_2007pt and μ_0 is big enough to be statistically significant, since the p-value equals 0.003. The observed standardized effect size equals to 3.0.

Table 2. Classification algorithms' F_1 Score from inference of the age group of users of MQD, using the LIWC_2007pt and the LIWC_2015pt lexicons.

	NB	MNB	RF	J48	LMT
LIWC_2007pt	0.432	0.465	0.440	0.528	0.557
LIWC_2015pt	**0.440**	**0.471**	**0.454**	**0.536**	**0.568**

The results for the dataset `TSN-60k` are described in the Table 3. The best results using LIWC_2015pt were achieved with the LMT algorithm, whereas, using LIWC_2007pt, the RF classifier achieved the best result. With this we observe that classification with RF of data from LIWC using LIWC_2015pt lexicon reaches in total an improvement of up to 37% on the value of 0.697, the best result obtained using LIWC_2007pt with the same algorithm.

Table 3. F_1 score of the algorithms used for sentiment polarity classification of text data from `TSN-60k`, using LIWC_2007pt and LIWC_2015pt lexicons.

	NB	MNB	RF	J48	LMT
LIWC_2007pt	0.615	0.649	0.697	0.644	0.683
LIWC_2015pt	**0.743**	**0.875**	**0.955**	**0.949**	**0.965**

Next, we compared the F_1 score of the algorithms used for sentiment polarity classification of text data from `TSN-60k`. In order to accomplish this task, we used all categories of LIWC_2007pt and LIWC_2015pt lexicons. Paired sample test using T distribution (DF=4, two-tailed) shows that the difference between the average of the LIWC_2015pt minus LIWC_2007pt and μ_0 is big enough to be statistically significant, since the p-value equals 0.0015, with effect size of 3.47.

In addition to the measures presented, we also note the Elapsed time for the processing of textual content of each dataset with each lexicon. Figure 1 displays the average time (95% CI) in milliseconds (ms) of ten runs of textual analysis with LIWC using either the LIWC_2007pt or LIWC_2015pt lexicons. The values were obtained from an environment with an Intel Core i3-330M processor with 2 cores of 2.13 GHz, 4.00 GB DDR3 RAM, motherboard model Calpella CRB, 5400 RPM hard disk model

WDC WD5000BEVT-00A0RT0 in ATA bus and Microsoft Windows 10 Professional 64-bit (Build 17134). In this measurement, both the CPU time and the system timeout are considered [Crowl 1994].

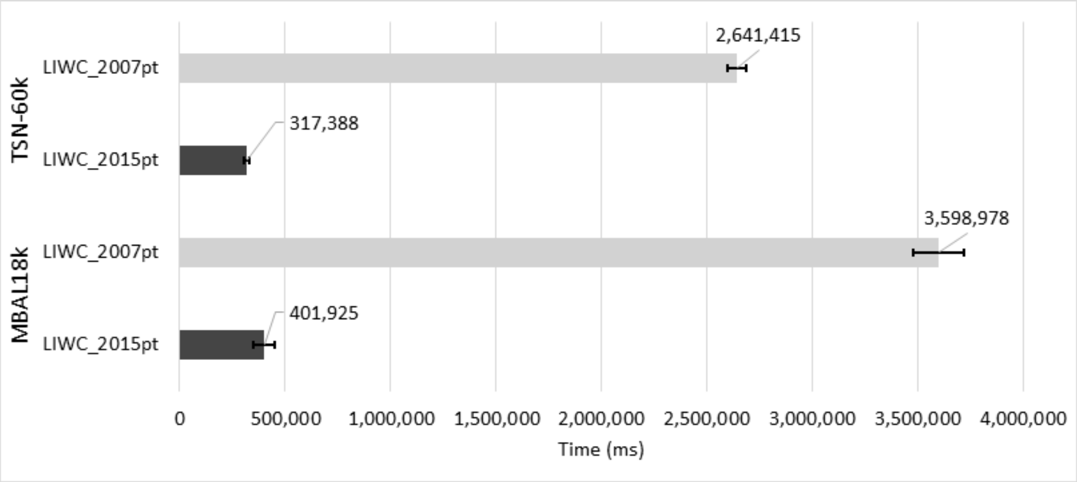

Figure 1. Avarage processing time (ms) using LIWC_2007pt and LIWC_2015pt lexicons in the LIWC program to process `MBAL18k` and `TSN-60k`, 95% CI.

It is possible to note that the time required for processing texts using LIWC_2015pt is lower than the time required for processing texts using LIWC_2007pt. Using LIWC_2015pt, the processing times of `MBAL18k` and `TSN-60k` respectively reduces up to 88.8% and 88.0% of the time required for processing using LIWC_2007pt.

5. Conclusions

The main contribution of this work is an initial evaluation of our recently developed LIWC 2015 BP lexicon (i.e., LIWC_2015pt). This lexicon enables representing written text files in dimensions of linguistic, psychological and social aspects. It is based on the 2015 English version of LIWC lexicon, which was developed after years of studies to validate categories and words in it.

Experiments executed with publicly available datasets indicate that LIWC_2015pt outperforms LIWC_2007pt in the classification task. It also indicated that the smaller size of the lexicon file of LIWC_2015pt (when compared to LIWC_2007pt) allowed for much faster textual content analysis. This is a strong indication that there is better adjustment of the words to the categories in which they are inserted.

We observed that, although better values were obtained in the experiments with `MBAL18k`, the means for F_1 Score for age inference is not very good. It is possible that the chosen set, with data from the MQD social network, is not precise with respect to age annotations. Since this information is provided by users in their profiles, we can not guarantee that they correspond to observable reality.

Still on the `MBAL18k` dataset, we could notice some entries where the posts contained texts from other authors, such as fragments of literary works or news. Sometimes, the field used by authors for publications was also used to save texts that are actually conversations with other people, that probably were copied from message exchange applications.

In this scenario, for future work, we intent to search for other sets of texts that have better control over the text entries and the information about the authors. We also plan to compare results from LIWC_2015pt with other language-specific LIWC lexicons on the same tasks, in similar datasets, and also with other lexical resources in BP. We also intend to include tasks other than classification, in order to analyze different features available in the LIWC program and to show advantages in selecting lexicons that reflects not only psychological aspects, such as emotions, but also social and linguistic features.

6. Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Aires, R., Manfrin, A., Aluísio, S., and Santos, D. (2004). Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to users' needs?

Alparone, F., Caso, S., Agosti, A., and Rellini, A. (2004). The Italian LIWC2001 Dictionary. *LIWC. net, Austin.*

Balage Filho, P. P., Pardo, T. A., and 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 (STIL)*, pages 215–219.

Bjekić, J., Lazarević, L. B., Živanović, M., and Knežević, G. (2014). Psychometric evaluation of the Serbian dictionary for automatic text analysis: LIWCser. *Psihologija*, 47(1):5–32.

Caetano, J. A., Lima, H. S., dos Santos, M. F., and Marques-Neto, H. T. (2017). Utilizando análise de sentimentos para definição da homofilia política dos usuários do Twitter durante a eleição presidencial americana de 2016. In *6th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017)*, volume 6. SBC.

Carvalho, F., Rodrigues, R. G., Ferrari, L., and Guedes, G. P. (2018a). A dictionary of pronouns for Brazilian Portuguese. In *Congresso Internacional de Informática Educativa (TISE)*, Brasília, Brasil. J. Sánchez.

Carvalho, F., Santos, G. d., and Guedes, G. P. (2018b). AffectPT-br: an affective lexicon based on LIWC 2015. In *37th International Conference of the Chilean Computer Science Society (SCCC 2018)*, Santiago, Chile. IEEE.

Cavalcante, P. E. C. and Malheiros, Y. d. A. (2017). Um dataset para análise de sentimentos na língua portuguesa. Trabalho de Conclusão de Curso, Bacharel em Sistemas de Informação, Universidade Federal da Paraíba.

Crowl, L. A. (1994). How to measure, present, and compare parallel performance. *IEEE Parallel & Distributed Technology: Systems & Technology*, 2(1):9–25.

Fersini, E., Pozzi, F. A., and Messina, E. (2015). Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In *Data Science and*

Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on, pages 1–8. IEEE.

Gabrilovich, E. and Markovitch, S. (2004). Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5. In *Proceedings of the twenty-first international conference on Machine learning*, page 41. ACM.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., and Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1145, Montreal, Canada. Morgan Kaufmann.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Loures, T. C., de Melo, P. O. V., and Veloso, A. A. (2017). É possível descrever episódios de séries de televisão a partir de comentários online? In *6th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017)*. SBC.

Massó, G., Lambert, P., Penagos, C. R., and Saurí, R. (2013). Generating New LIWC Dictionaries by Triangulation. In *Asia Information Retrieval Symposium*, pages 263–271. Springer.

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., and Horn, A. B. (2019). "LIWC auf Deutsch": The development, psychometrics, and introduction of DE-LIWC2015.

Moreira, S. F., Baklizky, M., and Digiampietri, L. A. (2018). Uso de mineração de textos para a identificação de postagens com informações de localização. In *7th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)*. SBC.

Öztuna, D., Elhan, A. H., and Tüccar, E. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*, 36(3):171–176.

Paulussen, H., Macken, L., Vandeweghe, W., and Desmet, P. (2013). Dutch parallel corpus: A balanced parallel corpus for Dutch-English and Dutch-French. In *Essential Speech and language technology for Dutch*, pages 185–199. Springer.

Peat, J. and Barton, B. (2008). *Medical statistics: A guide to data analysis and critical appraisal*. John Wiley & Sons.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Technical report, University of Texas.

Pennebaker, J. W. and Chung, C. K. (2011). Expressive writing: Connections to physical and mental health. *Oxford handbook of health psychology*, pages 417–437.

Pettijohn, T. F. and Sacco Jr, D. F. (2009). The language of lyrics: An analysis of popular billboard songs across conditions of social and economic threat. *Journal of Language and Social Psychology*, 28(3):297–311.

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., and 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.

Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., and Pennebaker, J. W. (2008). The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In *ICWSM*.

Ravi, K. and Ravi, V. (2017). A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems*, 120:15–33.

Rodrigues, R. G. and Guedes, G. P. (2017). A hybrid affective lexicon for brazilian portuguese. *CEP*, 20271:110.

Shibata, D., Wakamiya, S., Kinoshita, A., and Aramaki, E. (2016). Detecting Japanese patients with Alzheimer's disease based on word category frequencies. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 78–85.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Tausczik, Y. R. and 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.

Van Wissen, L. and Boot, P. (2017). An electronic translation of the LIWC Dictionary into Dutch. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pages 703–715. Lexical Computing.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.

Zeng, X., Yang, C., Tu, C., Liu, Z., and Sun, M. (2018). Chinese LIWC lexicon expansion via hierarchical classification of word embeddings with sememe attention. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52.