The Challenges of Modeling and Predicting Online Review Helpfulness

Rogério Figueredo de Sousa, Thiago Alexandre Salgueiro Pardo

[1]Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
Av. Trabalhador São Carlense, 400 – 13.566-590 – São Carlos – SP – Brazil

rogerfig@usp.br, taspardo@icmc.usp.br

Abstract. *Predicting review helpfulness is an important task in Natural Language Processing. It is useful for dealing with the huge amount of online reviews on varied domains and languages, helping and guiding users on what to read and consider in their daily decisions. However, there are limited initiatives to investigate the nature of this task and how hard it is. This paper aims to fulfill this gap, providing a better understanding of it. Two complementary experiments are performed in order to uncover patterns of usefulness evaluation as performed by humans and relevant features for machine prediction. To assure our results, we run the experiments for two different domains: movies and apps. We show that humans agree on the process of assigning helpfulness to reviews, despite the difficulty of the task. More than this, people perform this process systematically and consistently. Finally, we empirically identify the most relevant content features for machine learning prediction of review helpfulness.*

1. Introduction

Web popularized access to large sets of information. Frequent actions as buying products and purchasing services may be done more consciously, as there are millions of reviews about products, movies, apps, and so forth. Unfortunately, such amount of information is a double-edged sword. On one hand, it provides valuable material to the users, but, on the other hand, it contains more information than a person can handle. This is a problem that is the subject of several areas. One of them is Natural Language Processing (NLP). In this paper, we are particularly interested in the subtask of Modeling and Predicting Online Review Helpfulness.

Among the large amount of data on the Web, User-Generated Content (UGC) is a major source, and product and service comments form a great portion of that content. However, not every comment (or opinion or review) is considered useful or relevant by other users. Indeed, some of this content may be considered unwanted, such as poorly written texts, vague opinions, texts with questionable content, etc [Kim et al. 2006]. This shows that user-generated content varies a lot in quality and such texts do not necessarily help readers' decision-making. A helpful review, according to [Mudambi and Schuff 2010] is a "peer-generated product evaluation that facilitates the consumer's purchase decision process". In such situation, modeling and predicting review helpfulness comprise the definition of models for characterizing good quality content and the proposition of methods for classifying opinions regarding their helpfulness degree.

Despite the importance of such research line, few studies have focused on the nature of this task and on determining how systematic and difficult it may be. The purpose of this paper is to bring some understanding on what influences people perception on review helpfulness and which features are more relevant for machines to automatically deal with online reviews. We run two complementary experiments on two different domains (movies and apps). We show that humans agree on the process of assigning helpfulness to reviews, despite the difficulty of the task. Moreover, we show that people perform this process systematically and consistently. Finally, we also identify the most relevant content features for machine learning prediction of review helpfulness.

The paper is organized as follows. Section 2 presents the main definitions about the task and also describes the main related work. Section 3 details the corpus that is used in this work. Section 4 describes the adopted methodology. In Section 5, we report the achieved results. Finally, Section 6 concludes the paper, indicating future research.

2. Related Work

Modeling and prediction of online review helpfulness are part of a task that studies the factors that determine review helpfulness and attempts to accurately predict it [Diaz and Ng 2018].

Helpfulness is relevant for ranking and displaying content to users who search comments on products or services on e-commerce websites. These websites usually present the most helpful ones first and delegate to the users the task of evaluating whether they are helpful or not. Questions like "Was this review helpful to you?" are presented to the users, and the feedback allows the system to eventually re-rank the set of reviews. However, some reviews can take a long time to accumulate a good number of user feedback. Recent reviews and the product with low user traffic are more affected by this fact. Therefore, automating the task is very beneficial. The automatic helpfulness prediction can benefit the websites that do not have ranking systems as well as can improve the manual rankings. In addition, the prediction of helpfulness can be used to filter off low-quality reviews, which can improve other tasks, such as review summarization [Anchiêta et al. 2017].

The main works in helpfulness prediction attempt to perform one of these three tasks: score regression, binary review classification, or review ranking methods. They depend on the helpfulness score that is usually calculated for each review by Equation 1. Score regression aims to predict the helpfulness score $h \, \epsilon \, [0, 1]$. Binary review classification seeks to decide whether comments are helpful or not based on a specific threshold (e.g., $h > 0.5$). Review ranking needs to order the reviews by their helpfulness according to a reference ranking.

$$h = \frac{helpful\ votes}{helpful\ votes + unhelpful\ votes} \tag{1}$$

Several features have been used to characterize helpfulness in the literature. They are usually split in two categories: content and context features [Diaz and Ng 2018]. The content features are related to the information that can be extracted directly from the review, such as the text and the "stars" given by the author. Context features are those extracted from outside the review, such as reviewer information. For the more interested reader, we recommend the survey of [Diaz and Ng 2018].

Most of the works try to generate a model using a set of those features. For instance, [Kim et al. 2006] used structural features (length of the review, number of sentences, etc.), lexical features (Term Frequency - Inverse Document Frequency (*TF-IDF*) statistic of words) or even syntactic and meta-data features (number of stars) in order to predict the helpfulness of reviews. They generated a regression model using a dataset extracted from *Amazon.com* and achieved their best results with the combination of all the features, obtaining 0.656 and 0.604 on Spearman correlation coefficient. More recently, [Baowaly et al. 2019] achieved the state of the art results for helpfulness classification. They used a dataset collected from the *Steam* game database and generated a model with the Gradient Boosting Machine algorithm. Some of their features were categorized in metadata features (e.g., recommendation, posting date of a review, etc.), reviewer features (e.g., number of reviews, number of acquired games, etc.), semantic features, TF-IDF and Word2Vec features, among others. Their model achieved more than 0.99 of f-measure in several categories of games, such as action, survival and RPG.

Some works attempted to understand the impact of the features in the task. [Mudambi and Schuff 2010] investigated what makes reviews helpful to a consumer. They evaluated three features: review extremity, review depth, and product type. Using a dataset from *Amazon.com*, they found out that the product type ("experience" or "search") influences the effect of the review extremity and the review depth over users. For experience goods, the extreme reviews are less helpful than moderate ones. The review depth has a positive influence on both product types, but has a bigger influence on search goods than for experience goods. [Tsur and Rappoport 2009], generated an algorithm to classify the reviews and, in addition, attempted to understand the nature of book review evaluation. Three human annotators evaluated 360 reviews and the authors concluded that review evaluation is subjective, but people still get a high agreement, achieving a Fleiss' kappa value of 73.3%.

Table 1. UTLCorpus numbers.

	Movies	**Apps**
# texts	$1,833,691$	$898,847$
# objects	$4,283$	243
# types	$1,828,647$	$419,713$
# tokens	$60,177,264$	$11,919,636$
Avg. of Tokens p/ doc	32.7994	12.9384
Helpfulness Label	*helpful:* $381,083$ *(20%)*	*helpful:* $50,166$ *(5%)*

In this paper, inspired by the previous initiatives, we present a deeper investigation of human behavior on evaluating helpfulness and of useful features for machine learning-based helpfulness prediction. We start by briefly describing in the next section the corpus that we use for our experiments.

3. The UTLCorpus

In this paper, we use the UTLCorpus [Sousa et al. 2019] as our dataset. This corpus is composed by reviews written in Portuguese for two domains: movies and apps. An amount of $2,732,538$ reviews ($1,833,691$ for movies and $898,847$ for apps) were collected using two web crawlers.

The authors of UTLCorpus anonymized the dataset and made it publicly available. They preserved important metadata fields from the original reviews, such as star rating, publication date, and, specifically in the movie domains, information on whether a reviewer saw a movie or whether the movie is a favorite.

Table 1 synthesizes the basic statistics of the corpus and shows some interesting information. One may see that the average size of movie reviews is much higher than that of apps. The information of helpfulness label shows that the corpus is highly unbalanced, mainly for the apps domain, which can be a problem in some cases. It is worth mentioning that this unbalancing problem does not interfere with the results presented here. The correlation experiments were performed on the balanced (with undersampling) and on the original (unbalanced) datasets, and the results were similar.

4. Research Methodology

Trying to understand the textual and non-textual features that characterize the helpfulness of online reviews, this work proposes a study of review helpfulness modeling and prediction. In this section, we present the proposed configuration of our study.

We investigated two complementary questions to guide our study, each one trying to understand a specific property of the helpfulness of reviews on apps and movies. In summary, the questions are as follows:

1. How difficult is the task for humans?
2. Which features are relevant for the task of helpfulness prediction?

Answering such questions may drive research in the area and foster the development of better systems in the future. In the following subsections, we explore each of the questions.

4.1. Helpfulness Evaluation is Difficult for Humans?

To answer this question, we need to discover if humans agree with each other while evaluating the helpfulness of reviews. For this purpose, we conduct a manual annotation process, counting with some annotators to accomplish this task.

The annotation process was to read and evaluate the helpfulness of 24 reviews extracted from the UTLCorpus, 12 from each domain, equally distributed in helpful and not helpful categories. These reviews were selected from only a movie and an app, randomly. The respondents needed only to choose among three options: *The review is helpful*, or *the review is unhelpful*, or *I don't know*.

To approximate the annotation process to that found in the ordinary process of evaluating the helpfulness of reviews, we decided to add an "information need" for annotators. Looking at the ordinary process of voting on the helpfulness of reviews on websites, we have found that users do not arbitrarily decide on the helpfulness of reviews. If they are reading reviews about a product, they are concerned with getting some relevant information about it. And because of their interest in the product, they can be more critical when evaluating reviews. This "information need" was specified to the annotators through the following sentences: "*You are deciding whether to download the app [app name] (to watch the movie [movie name]), and you have come across these reviews. You must answer*

the following question for each review: "Is this opinion helpful to you?" Evaluate whether the review helps you to decide to download or not the app (to watch or not the movie)." Note the underlined excerpts, they vary for each domain as highlighted in brackets. Figure 1 shows an example of a review in the form with an "information need" text.

We distributed a form to fourteen annotators, and they had a few days to accomplish the task. By the end of the deadline, only ten annotators completed the process.

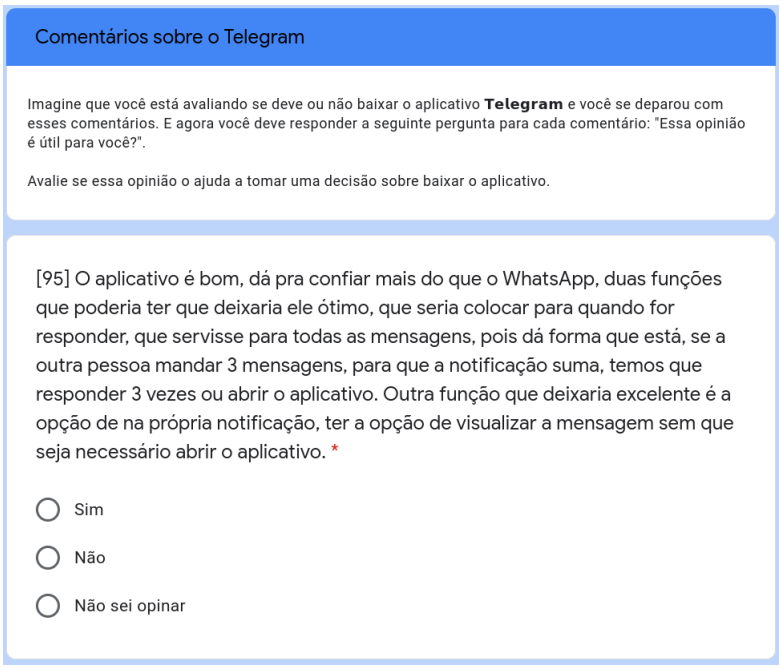

Figure 1. An example review (in Portuguese) on the form distributed to the annotators. It also shows the "information need" provided to annotators.

Although the main objective of the annotation process is to evaluate the agreement of annotators, we aggregate some other side objectives that could help us to understand the evaluation process of helpfulness by humans. We randomly selected the 24 reviews, but in sets with specific conditions. The first condition is the domain, 12 of each domain, as commented before. The second condition is the helpfulness category, being six of each class (helpful or not helpful), and, finally, the last condition is the length of review: three reviews are long and three are short. The short ones have 30 words at most, while the long ones have more than 60 words. Figure 2 helps to illustrate the subset we ended up for human evaluation. The decision to select 24 reviews for manual annotation was due to the nature of the comments. [Liu et al. 2007] and [Tsur and Rappoport 2009] show that the domains where characteristics are not so well-defined generate more open reviews, making evaluation difficult and expensive. Another reason is that this approach brings the process closer to the real voting conditions, where customers typically rate few comments.

We expect with this configuration to get some additional information about what influences people perceived helpfulness, more specifically, if the length of review influences the evaluation of review helpfulness.

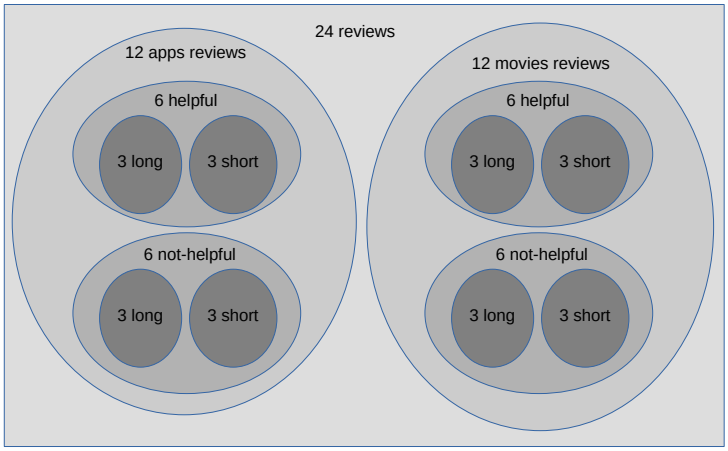

Figure 2. A graphical representation of the subsets of reviews for the evaluation.

4.2. Which features are relevant for the task?

The researchers in helpfulness modeling and prediction, along the years, developed many types of features trying to characterize the helpfulness of reviews in different domains and languages.

To answer the question "*Which features are relevant for the task?*", we followed a tiny pipeline:

1. To select the relevant candidate features from the literature in the area.
2. To select the necessary resources to implement and adapt the features to the language of the corpus (which is in Portuguese).
3. To implement the selected features.
4. To calculate the contribution of the features for the target task.

The first step of our pipeline revealed many features in many works. Considering that the features can be classified into different categories, we decided to limit the selection to content features only. The content features extract information directly from the reviews, such as review text and star rating. Most of these features are simple and easy to understand and to replicate, therefore, we were able to adapt and evaluate more features. And it is worth to mention that we selected and adapted the most common features of helpfulness prediction literature.

The second step shows us the necessary resources and tools to adapt the features to our language. Despite the differences in accuracy of many tools between languages, we choose the equivalent resources for each selected feature.

In the third step, we try to adapt the features as accurately as possible[2], considering the particularities of the language

The last step is the most important in our pipeline. In this step, we calculate the impact of the features, individually comparing to the helpfulness of the reviews. We decided to compute the correlation between feature values and the helpfulness class (not

[2]Our entire adaptation code of features is available at `https://github.com/RogerFig/features_experiments`

Table 2. List of Features.

Feature	Description
Average Sentence Length (Avg-SL)	Ratio between the number of words and the number of sentences in the review [Liu et al. 2007, Lu et al. 2010]
Number of Sentences (Num-S)	Total of sentences in the review [Liu et al. 2007, Lu et al. 2010]
Number of Words (Num-W)	Total of words in the review [Kim et al. 2006, Mudambi and Schuff 2010]
Star Rating (Star-R)	The review-assigned product star rating [Huang et al. 2015]
Readability Features (READ)	Measures how easy a text is to read and contains the following features: Automated Readability Index (ARI), Coleman-Liau Index (CLI), Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), Gunning fog index (GFI) and SMOG [DuBay 2004, Ghose and Ipeirotis 2011]
Spelling errors (SPELL)	Total words not found in a lexicon composed of words from the Wiktionary[1] and the Unitex-PB lexicon [Ghose and Ipeirotis 2011, Muniz 2004]
Dominant Terms (Dom-Terms)	Presence of important terms in reviews, considering their specificity for the domain and movie/app [Tsur and Rappoport 2009]
Product features (Prod-Feat)	Presence of product features in the reviews [Kim et al. 2006, Hong et al. 2012, Liu et al. 2007]
Sentiment Words (SENT)	Word count that may reflect opinions, analyses, emotions etc. [Kim et al. 2006]. We use some categories of LIWC dictionary [Balage Filho et al. 2013, Pennebaker et al. 2001] to calculate these features. The categories are: Negate, Swear, Affect, Posemo, Negemo, Anxiety, Anger and Sad.
Sentiment divergence (Sent-Div)	Difference between the general sentiment about the movie/app and the sentiment expressed by the author of a review [Hong et al. 2012]. We used the Sentilex sentiment lexicon [Silva et al. 2012] to calculate this feature.
Subjectivity (SUB)	The probability of a review and its sentences being subjective [Ghose and Ipeirotis 2011]
Syntactic tokens (SYN)	Number of tokens with the following Part-of-Speech tags: Noun (N), Verb (V), Adverb (ADV) and Adjective (ADJ). It also includes counting for open class words (Open) [Kim et al. 2006]
Star Deviation (Star-Dev)	Difference between the amount of stars in a review and the average star rating for the movie/app [Hong et al. 2012]

helpful: 0 and helpful: 1) of reviews using the correlation coefficients of Pearson and Spearman. All features have been normalized and Section 5 presents the correlation results.

With this process, we expect to find clues about the impact of features in helpfulness definition, determining which features are more or less relevant to the task. Table 2 presents and describe all features used in this work, including citations to some of the main previous works that used them.

5. Results and Discussion

Considering the methodology described in Section 4, we present in this section the results achieved in the annotation process and the correlation study between features and helpfulness.

5.1. The Annotation Process and Evaluation of the Lexical Similarity

In order to evaluate the annotation process, we used a well-known inter-annotators agreement metric: Krippendorff Alpha [Krippendorff 1970]. For the sake of better visualization, the results are divided into some groups.

Figure 3 shows the results of the annotation process. It is worth to remember that we impose some conditions to select the reviews. We split the reviews on these three groups: length (short, long), domain (movies, apps) and helpfulness (helpful, not helpful). Hence, the figure presents the inter-annotator agreement considering the combination of groups. The first part of the figure shows the agreement for bigger groups. The second part of the figure presents the agreement values for composition of two groups: *length* X *domain*. The third part of the figure presents the agreement results considering all three groups: *domain* X *class* X *length*.

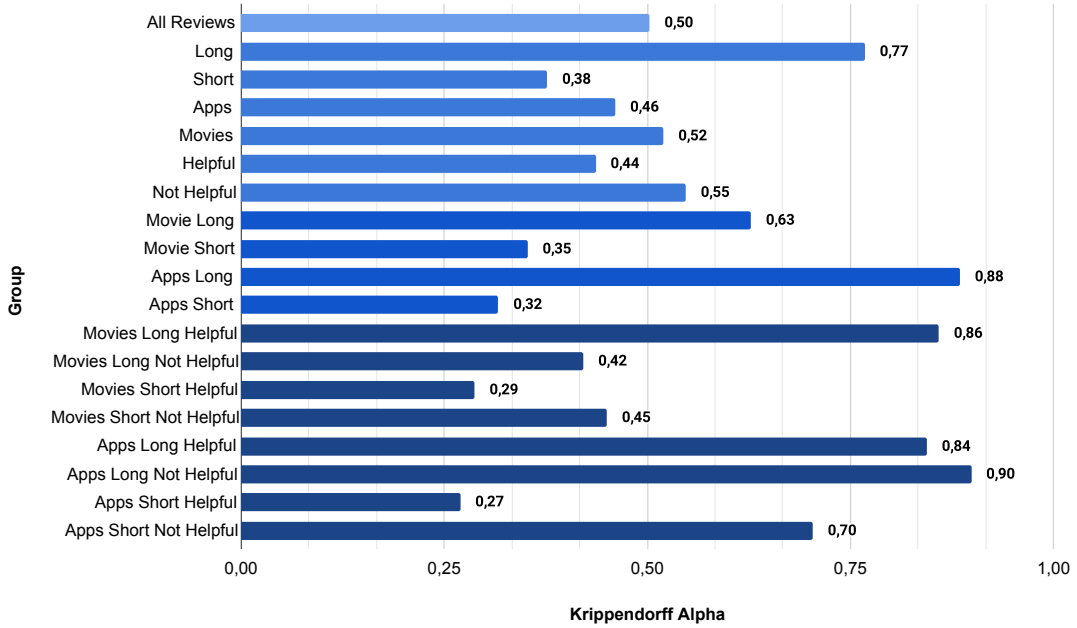

Figure 3. General results for agreement.

There are some information that stands out in the table. In the first place, we can observe an agreement pattern. The long reviews produce a better agreement than the short ones, probably because the longer ones tend to include more information to support the user decision (considering the information need). Apps also produce better agreement values, which may be possibly explained by the less subjective reviews (as they frequently comment on technical aspects of the apps). The best agreement results were achieved by apps' long reviews for the helpful category. It is also interesting how short reviews (for both domains) do not produce good agreement results for the helpful category. Overall, the high agreement results achieved for some cases show that the task is clear enough for humans under certain circumstances, as enough amount of available information (as provided by the longer reviews).

We proposed an additional experiment, which consists of evaluating the lexical similarity of reviews and comparing their categories. If humans are consistent in their annotation, we expect to see higher helpfulness agreement as the lexical similarity increases.

For this experiment, we use the training part of the UTLCorpus, which contains 80% of reviews (1,466,952 movie reviews and 719,077 app reviews). The process was conducted as follows:

1. For each domain, the reviews were split in long and short ones;

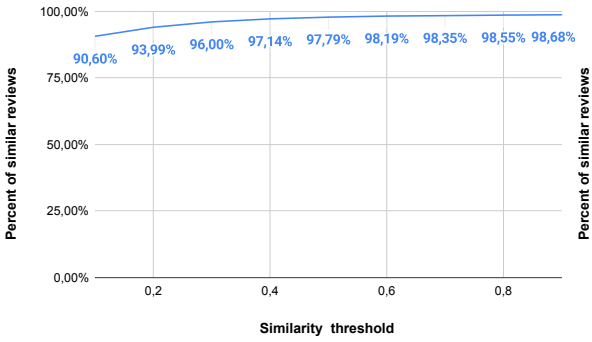

Figure 4. Short apps reviews.

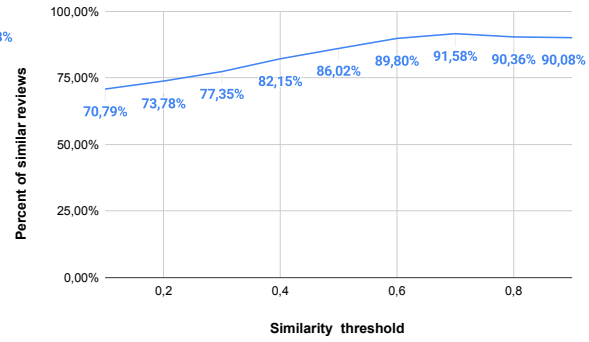

Figure 5. Short movies reviews.

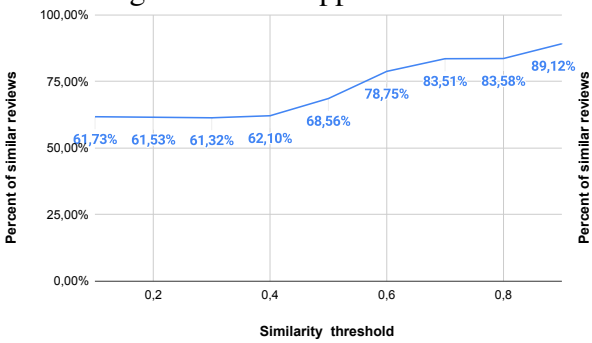

Figure 6. Long apps reviews.

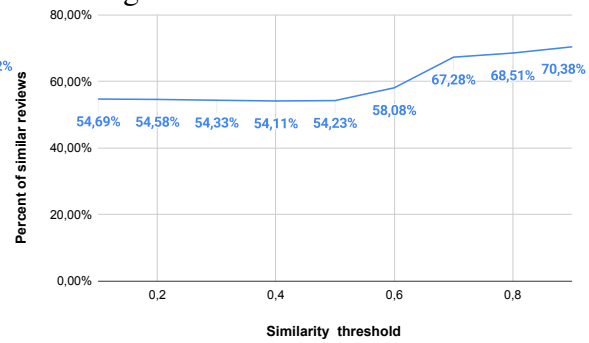

Figure 7. Long movies reviews.

Figure 8. Similarity Experiment.

2. Each review was represented by a Bag-of-Words Vector;
3. The Cosine similarity was calculated among all vectors (all vs. all);
4. We calculate the percentage of the reviews that have a cosine similarity above a threshold and have the same helpfulness category.

Several similarity thresholds have been considered, ranging from 0.1 to 0.9 and the results are presented in Figure 8. The X axis shows the similarity thresholds and the Y axis shows the percentage of reviews with the same helpfulness category. As expected, we may see that the proportion of reviews with the same category grows with the increase of the lexical similarity. The short reviews have a higher proportion of similarity than the long ones. One possible explanation is that users have a tendency to use less diverse vocabulary to write shorter comments. On the other hand, the authors need to use a diversified vocabulary to write the long ones.

Taken together, these results suggest that there is strong evidence that people agree with each other on the process of assigning helpfulness to reviews in domains of movies and apps, and they perform this process systematically and consistently. Moreover, the lexical similarity curves support the evidence that human judgment is not aleatory.

5.2. Correlation of features with helpfulness

For the purpose of finding relevant features for determining the helpfulness of reviews, we calculate the correlation coefficients of Pearson and Spearman for all features in Table 2 in relation to the helpfulness class (not helpful: 0 and helpful: 1). For this experiment,

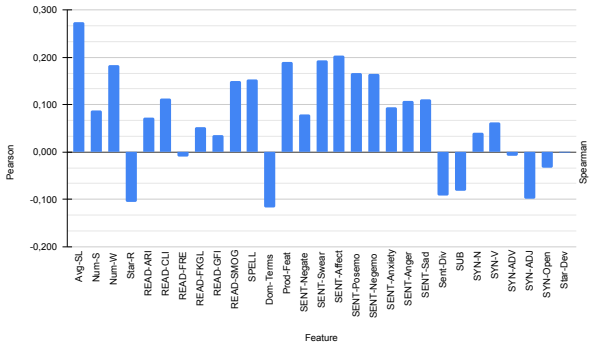

Figure 9. Pearson for Apps Domain.

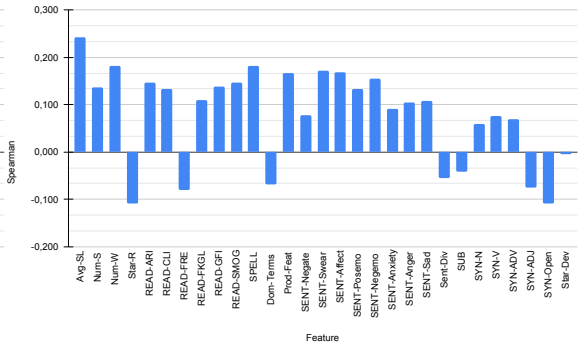

Figure 10. Spearman for Apps Domain.

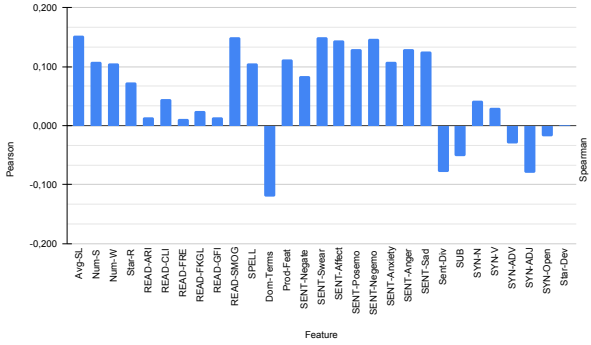

Figure 11. Pearson for Movies Domain.

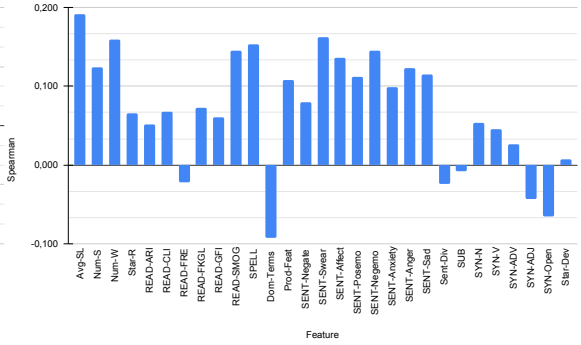

Figure 12. Spearman for Movies Domain.

Figure 13. Feature correlation results.

we used the training subset of UTLCorpus. As explained in the previous subsection, the training subset contains 80% of all the reviews of UTLCorpus.

Figure 13 summarize the results. Figure 11 and Figure 12 provides the results for the movies domain, and Figure 9 and Figure 10 presents the results for the apps domain.

An inspection of the figures shows that in both domains the simple features are among the most positively correlated features, for example, *Average Sentence Length, Number of Sentences, Number of Words, and Spelling Errors*. Some readability scores and the LIWC [Silva et al. 2012] features also showed a noticeable positive correlation. Each of the features in the sentiment words category refers to the category of the same name in the LIWC ("negate", "swear", "affect", "posemo", "negemo", "anxiety", "anger" and "sad"). In the opposite direction, we can highlight some features with inverted correlation, for example, *dominant terms* in both domains and *star rating* for apps domain. Most of the remaining features have not achieved important values of correlation, with intermediate results.

Being more specific, among the content features presented in this subsection, the most correlated ones with movie review helpfulness are (according to the two used correlation measures): Average Sentence Length, Readability-SMOG, and some Sentiment Features. Exclusively for Apps, we have: Average Sentence Length, Number of Words, Readability-SMOG, Spelling Errors, Product Features, and some Sentiment Features. It is interesting to notice that some of the features are relevant for both domains, indicating that they might be useful for building general domain classifiers.

The presence of common relevant features in the two domains is specially important for the area of sentiment analysis, as it is widely known that the domain usually makes a lot of difference in the performance of systems. More experiments must be carried out for obtaining irrefutable conclusions, but our domains (movies and apps) are different enough to allow us to infer that such features might be also relevant for other domains. Some evidence of the domain differences come from some researches that have shown that reviews on topics like movies and books tend to be more "passionate", while reviews on electronic devices and apps tend to be more "technical" (see, e.g., [Vargas and Pardo 2018] for some interesting discussion on this).

6. Final Remarks

In this paper, we presented a study of review helpfulness, trying to answer how hard the task is and which features appear to be more useful for prediction. We show that people agree with each other in the task of evaluating the helpfulness of reviews for movie and app domains (specially for longer texts). Moreover, through lexical similarity, we show that people are consistent in the task. We also evidence that some features are clearly correlated to task of helpfulness prediction, independently of the domain, which might help producing better general domain helpfulness classifiers. To the best of our knowledge, the work reported here is the most comprehensive one on such topics. The interested reader may find more information at the web portal of the POeTiSA project[3].

Future work includes generating machine learning classification models with the best features and testing context features, as these new features may bring more understanding about the task.

Acknowledgments

The authors are grateful to the USP/IBM/FAPESP Center for Artificial Intelligence (C4AI, grant #2019/07665-4) and *Instituto Federal do Piauí* (IFPI).

References

Anchiêta, R., Sousa, R. F., Moura, R., and Pardo, T. (2017). Improving opinion summarization by assessing sentence importance in on-line reviews. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 32–36.

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (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*.

Baowaly, M. K., Tu, Y.-P., and Chen, K.-T. (2019). Predicting the helpfulness of game reviews: A case study on the steam store. *Journal of Intelligent & Fuzzy Systems*, 36(5):4731–4742.

Diaz, G. O. and Ng, V. (2018). Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 698–708.

DuBay, W. H. (2004). The principles of readability. *Online Submission*.

[3] https://sites.google.com/icmc.usp.br/poetisa

Ghose, A. and Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

Hong, Y., Lu, J., Yao, J., Zhu, Q., and Zhou, G. (2012). What reviews are satisfactory: Novel features for automatic helpfulness voting. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 495–504, New York, NY, USA. ACM.

Huang, A. H., Chen, K., Yen, D. C., and Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, 48:17–27.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430. Association for Computational Linguistics.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., and Zhou, M. (2007). Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L. (2010). Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, pages 691–700. ACM.

Mudambi, S. M. and Schuff, D. (2010). Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, pages 185–200.

Muniz, M. C. M. (2004). *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto Unitex-PB*. PhD thesis, Universidade de São Paulo.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. In *International Conference on Computational Processing of the Portuguese Language*, pages 218–228. Springer.

Sousa, R. F., Brum, H. B., and Nunes, M. d. G. V. (2019). A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Proceedings of the 12th Brazilian Symposium in Information and Human Language Technology*, pages 209–218. Sociedade Brasileira de Computação.

Tsur, O. and Rappoport, A. (2009). Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *ICWSM*.

Vargas, F. and Pardo, T. (2018). Hierarchical clustering of aspects for opinion mining: a corpus study. In Finatto, M., Rebechi, R., Sarmento, S., and Bocorny, A., editors, *Linguística de Corpus: Perspectivas*, pages 69–91. Porto Alegre: Instituto de Letras da UFRGS.