A bunch of helpfulness and sentiment corpora in Brazilian Portuguese

Rogério Figueredo de Sousa, Henrico Bertini Brum, Maria das Graças Volpe Nunes

[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,henrico.brum}@usp.br, gracan@icmc.usp.br

Abstract. *This paper presents the UTLcorpus, a novel corpus in Brazilian Portuguese for helpfulness classification of online reviews. There is a lack of corpora in Brazilian Portuguese annotated with helpfulness information, therefore there are also few works on modeling and predicting helpfulness of online reviews in this language. Moreover, there is no reference corpus to ground those results. This work tries to partially solve this problem by presenting UTLcorpus, a huge amount of annotated online reviews regarding helpfulness. Since the source data also contain star score labels, this paper also explores polarity labels in the data set. Some experiments show that both tasks of predicting helpfulness and polarity are benefited by the use of this corpus.*

1. Introduction

Navigating websites for buying clothes, picking travel locations or choosing a movie can be a hard task considering the number of choices we can find. It may be even harder if we are looking for user reviews to support our final decision. The high amount of comments in such websites can be a hold back for user looking for opinions that may help them to choose products/services, or even alert them for flaws already known to previous acquirers. Helpfulness Prediction (HP) is the task that aims to correctly predict whether a review or opinion is helpful for a user to read before acquiring a product or hiring a service.

Prediction models are usually based on supervised learning, therefore, demanding for linguistic resources (corpora) of labeled reviews. One of the challenges for helpfulness research in Brazilian Portuguese is the few number of available data sets. We present UTLcorpus, a data set composed of two automatic annotated corpora for helpfulness in that language. The data were extracted from two different domains: movie reviews from a Brazilian social network for movies[1] and app reviews from Google App Store[2].

The data was anonymized and preprocessed. Evaluations were carried out to bootstrap the corpora for the HP task and the results were compared to other literature corpora in Brazilian Portuguese. Since UTLcorpus also contains labels for binary polarity classification (star score indicative of positive and negative reviews) we also used literature methods for evaluating the data for this task.

[1]www.filmow.com. Accessed in May 19th, 2019.
[2]play.google.com. Accessed in May 19th, 2019.

The main contribution of this work is the creation of resources mainly for helpfulness prediction, but also for the polarity classification task. The corpus created in this work should be useful to increase the research in these areas and help to find the particularities of the helpfulness modeling task, thus enabling the understanding of this phenomenon in Brazilian Portuguese.

The paper is organized as follows. Section 2 presents an overview of Helpfulness Modeling and Prediction Task. The UTLcorpus is presented in Section 3. Experiments with the corpus in the tasks of helpfulness prediction and polarity classification are presented in Section 4. In Section 5 some important literature works on Polarity Classification and Helpfulness Prediction are discussed. Finally in Section 6 some conclusions and future works are presented.

2. Helpfulness Prediction Task

Modeling and prediction online reviews helpfulness (quality, usefulness or utility [Liu 2012]) are relevant for ranking and displaying comments to users who search comments on products or services. Most e-commerce websites present the most useful 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 re-rank eventually the set of reviews.

The drawback of this functionality is that the reviews can take a long time to accumulate a good number of user feedback. This is especially noticeable in new reviews, which can even be useful, but because of their low posting time, they can not get sufficient votes to achieve the top of the ranking. This fact demonstrates one of the advantages of automating the task. Websites that do not have ranking systems can benefit as well as the rankings themselves can be improved by the use of helpfulness prediction. In addition, the prediction of helpfulness can be used to filter off low-quality reviews, which can improve other tasks, such as the reviews summarization [Anchiêta et al. 2017].

Helpfulness prediction tasks mainly include score regression, binary review classification and review ranking. These three methods depend on the helpfulness score which is usually calculated for each review by the Equation 1. The score regression aims to predict the helpfulness score $h \, \epsilon \, [0, 1]$. The binary review classification seeks to decide whether comments are helpful or not based on a specific threshold (e.g. $h > 0.5$). And the 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. Usually they are 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. And the context features are those extracted from outside the review, such as reviewer information. In the survey of [Diaz and Ng 2018] one can find the most important features of the literature.

Contrary to what occurs for Portuguese, some large English corpora with utility annotations are available:

- Multi-Domain Sentiment Dataset (MDSD) [He and McAuley 2016] [3]: Collected from Amazon.com. Contains 25 product categories and 1.422.530 reviews.
- Amazon Review Dataset (ARD) [Blitzer et al. 2007, McAuley et al. 2015] [4]: Also collected from Amazon.com, contains 24 product categories and 142.8 million reviews and includes more metadata information than MDSD.
- Ciao Dataset [Tang et al. 2013] [5]: Was collected from an extinct e-commerce website and contains 302.232 reviews. The main difference from the previous ones is that it contains a social network between their users.

For the best of our knowledge, there is only one available corpus in Brazilian Portuguese, the *Buscapé* [Hartmann et al. 2014], containing 28.774 product reviews annotated with information that can be used for calculating helpfulness. Therefore, this work presents a new *corpus* containing information of helpfulness to promote researches in this task.

3. The UTLcorpus

The data set is a collection of reviews extracted from two domains: movies and apps. 2.881.589 reviews (1.839.851 of movies and 1.041.738 of apps) were collected using two web crawlers. The domains were chosen for the popularity, the high amount of data and the presence of a public "like" counter in each review, which makes possible to infer a helpfulness label. Besides the "like" counter, the data also contains scores given by users to the movie/app they are evaluating. We used the later for inferring positive and negative labels.

The methodology for labeling the polarity was proposed in [Avanço 2015]. Each review has a 5-star score according to the author's evaluation of the related movie/app. Reviews with 0 and 5 stars are ignored to avoid those cases in which the users stars are not coherent with the review text. Also the 3 star reviews are discarded because they usually contain positive and negative sentiment about the entity.

In order to label the data we looked for the utility labels in the data set. Both domains provide the number of "likes" a review received (indicating it was helpful for other users) and the main issue we faced was the lack of a counterpart indicating the number of "dislikes" were attributed to the review. The majority of works in the literature [Kim et al. 2006, Malik and Hussain 2017] divide the number of positive likes by the sum of likes and dislikes to obtain a value and determine a threshold of helpfulness for a data set.

Since we can not count on dislikes, we define helpfulness in UTLcorpus as following. First, we group the data by category (movie titles and app names). This is performed because more popular apps/movies aggregate more likes by review than the less popular ones. Then we sort the reviews by the number of likes each of them has received (ignoring the ones with zero likes, since we can not identify if they have anything helpful in

[3]https://www.cs.jhu.edu/ mdredze/datasets/sentiment/

[4]http://jmcauley.ucsd.edu/data/amazon/

[5]https://www.cse.msu.edu/ tangjili/trust.html

	Movie review subset	App review subset
# documents	1.834.702	921.257
# types	1.828.647	419.713
# tokens	60.177.264	11.919.636
Avg. token per doc.	32.7994	12.9384
Helpfulness Labeled	1.833.691	898.847
	helpful: 381.083 *(20%)*	*helpful:* 50.166 *(5%)*
Sentiment Labeled	862.768	320.255
	positive: 702.720 *(81%)*	*positive:* 113.351 *(35%)*

Table 1. UTLcorpus information

Subset	**Movie Review**		**App Review**	
	Positive	**Negative**	**Positive**	**Negative**
Helpful	155.672	37.699	6.533	18.473
Unhelpful	546.357	122.029	98.374	174.465

Table 2. Intersection between classes

it). Next, we remove replications and then consider helpful any comment with more likes then the first percentile of the distribution. To determine the negative samples we consider any review with fewer likes than the threshold previously defined and crawled at least five days after the review was published; the observation of the timespan is important since recent reviews take longer to achieve higher like counts thus becoming a false negative noise in the data set.

Online reviews are classified as User Generated Content (UGC) [Krishnamoorthy 2015], a type of text which carries many noisy linguistic phenomenons such as typos and Internet slangs. In order to reduce the noise in UTLcorpus the data was normalized using Enelvo [Bertaglia 2017], a tool for normalizing UGC in Brazilian Portuguese[6].

The data set is composed of two subsets representing two different domains:

Movie review corpus: The movie reviews corpus was obtained by crawling Filmow, a popular film social network containing reviews, scores, evaluations and general movie information. We crawled reviews from the 4.283 most popular movies in the platform (we stopped for storage reasons) and the reviews generally represent opinions about films, actors/actresses and all kind of experience the users can have in watching a movie. Reviews can be directed to recent blockbusters as well as classics and we made available the movie titles which the reviews are about. The class distribution is skewed and the majority class for helpfulness is of unhelpful (80% of the documents) and positive (around 80% of the documents). Even though the helpfulness and polarity tags come from the

[6]Available in `https://github.com/tfcbertaglia/enelvo`. Visited in March 19th, 2019.

same data set, some reviews do not have enough information to be part of both of the subsets (972.945 documents do not the overlap).

The most frequent terms in the corpus (ignoring stop-words such as demonstratives pronouns, conjunctions and punctuation) are: *movie*, *well*, *good*, *story* and *best*.

App review corpus: The App Review corpus was obtained by crawling Google Play Store[7]. The corpus contains app reviews and one important feature of this data set is the absence of reviews with zero stars, since it is mandatory to evaluate with a star score any review. We gathered reviews from 243 apps (the most popular ones) and the whole corpus contains 921.257 reviews. The data is also skewed in both labels, being unhelpful the majority class for helpfulness (95%), and negative the majority class for polarity (65%). 371.651 reviews have only one label and do not overlap.

The most frequent terms in the corpus (also ignoring stop-words) are: *app*, *best*, *great*, *can* and *cool*. It is interesting no notice that several words are more frequent in both corpora even though they are skewed for different polarity classes.

Table 1 contains detailed data set information. The skewing of the data is a challenge for machine learning classification methods and we address this issue in section 4. Table 2 presents the intersection between the classes (Helpfulness and Polarity) for both datasets. It is possible to see that, in the movie reviews subcorpus, 80% of the helpful comments have positive polarity, and so are 81% of the unhelpful comments. Moreover, in the app reviews subcorpus, most of the helpful reviews (73%) as well as of the unhelpful ones are negative (63%).

4. Corpus Evaluation

In order to observe and evaluate the characteristics of the corpora on classification tasks we performed experiments in both subsets of UTLcorpus, Movie Review corpus (MR) and App Review corpus (AR), using a baseline and machine learning classifiers for comparison purposes. Helpfulness Prediction can be seen as a task very similar to polarity classification thus we defined a baseline and also used three machine learning classifiers (Support Vector Machines[8], Multi-layer Perceptron[9] and Random Forest[10]) following the work of [Brum and Nunes 2018], originally proposed for polarity classification of sentences. The work of [Brum and Nunes 2018] used a grid search technique to set the hyper parameters.

For baseline purposes we represented each sentence using a 2-dimensional vector with the number of positive and negative terms using a Brazilian Portuguese sentiment lexicon – Sentilex [Silva et al. 2012], which contains Portuguese terms (eg. *bom*, *ruim*, *péssimo*) and their respective polarity label. We trained a SVM model using this feature representation and evaluated the data sets in a 10-fold cross validation scheme. Furthermore, we used three classifiers trained and evaluated on a 10-fold cross validation. To avoid the skewing of the majority class, the data were balanced randomly (by the minority class), thus reducing the data sets considerably. The final sizes of the corpora are

[7]`https://play.google.com/store`. Visited in March 19th, 2019.

[8]Hyper parameters – C: 1; *alpha*: 0.1; linear kernel.

[9]Hyper parameters – Activation: $tanh(x)$; learning rate: 0.001; alpha: 0.0001; neurons: 200; layers: 2.

[10]Hyper parameters – number of estimators: 200.

762.078 documents in the MR corpus and 100.322 in the AR corpus, nevertheless, the final corpus is still larger than the Brazilian Portuguese corpus *Buscapé*.

To the other three classifiers, differently from the baseline method, the data was represented using pre-trained 600-dimensional word2vec embeddings trained in more than 1 billion Portuguese Brazilian user-generated content (tweets and forums). The representation is described in [Corrêa et al. 2017] and has also been used for polarity classification in [Brum and Nunes 2018].

Classifier	Movie Review			App Review			Buscapé		
	F1-Help	F1-No-Help	F1-Measure	F1-Help	F1-No-Help	F1-Measure	F1-Help	F1-No-Help	F1-Measure
Baseline	0.4499	0.6493	0.5496	0.5896	0.6617	0.6256	0.4967	0.6343	0.5655
Linear SVM	0.6341	0.6039	0.6189	0.7115	0.6436	0.6775	0.6072	0.6142	0.6107
MLP	0.6387	0.6118	**0.6252**	0.7082	0.6516	0.6799	0.6114	0.5983	0.6048
Random Forest	0.6220	0.5920	0.6072	0.7267	0.7182	**0.7224**	0.6361	0.6436	**0.6398**

Table 3. Helpfulness detection results

The results obtained in the classification are shown in Table 3. The F1 values presented in the table are acquired with 10-fold cross-validation technique (Mean of 10 executions). Before classifying the data the class distribution was balanced by using *under-sampling*, in other words, we removed samples of the majority class before performing the cross-validation. Experiments with the unbalanced corpora resulted in F1 far below the ones presented in Table 3, even the best results had the minority class F1 below 0.1.

The baseline worked pretty well in relation to F1-Measure, but one can see that it does not handle well the positive class. The best results for helpfulness detection in both corpora were obtained using Random Forest classifier which predicts the class based on several estimators (Decision Trees). It is still uncertain if our method for defining the helpful class is reliable enough, but with this methodology it is still possible to predict the correct label in 60% of the time for movie reviews and 70% of the time for app reviews ($std.dev. = 0.004$). We believe that one possible explanation for the results is that the prediction of the utility does not depend on text only. We understand that helpfulness may be affected by the context (domain, category, website, etc.) in which the comment is inserted, as well as by the intention with which the reader is reading the comment.

Since UTLcorpus also has polarity labels we were able to perform experiments using them. The main difference was the baseline used: for polarity classification we represented the data similarly (positive and negative term frequency) but predicted as positive any sentence with more positive terms than negative ones. In Table 4 we can see the results for polarity classification in the corpora.

Classifier	Movie Review			App Review			Buscapé		
	F1-Pos	F1-Neg	F1-Measure	F1-Pos	F1-Neg	F1-Measure	F1-Pos	F1-Neg	F1-Measure
Baseline	0.6167	0.1675	0.3920	0.6378	0.1541	0.3959	0.5467	0.2031	0.3748
Linear SVM	0.6878	0.6517	0.6697	0.7687	0.7710	0.7698	0.8106	0.8243	0.8174
MLP	0.6602	0.6843	**0.6722**	0.7818	0.7814	**0.7815**	0.8146	0.8121	0.8133
Random Forest	0.6528	0.6644	0.6586	0.7588	0.7786	0.7686	0.8310	0.8128	**0.8218**

Table 4. Polarity classification results

Finally, we merged the corpora and perform experiments using the two domains. The choices of candidates reviews are the same for each domain and all selected by the criteria above mentioned are used this time. The results are presented on Table 5.

Classifier	Helpfulness Prediction			Polarity Classification		
	F1-Help	F1-No-Help	F1-Measure	F1-Pos	F1-Neg	F1-Measure
Baseline	0.6708	0.4583	0.5645	0.1570	0.6179	0.3874
Linear SVM	0.6299	0.6759	0.6529	0.7011	0.7578	0.7294
MLP	0.6383	0.6863	**0.6623**	0.7304	0.7646	0.7475
Random Forest	0.6398	0.6834	0.6616	–	–	–

Table 5. Helpfulness prediction and Polarity classification results using the whole UTLcorpus

For polarity classification we are able to better compare the results than for helpfulness since the literature contains more works relating that task. The results of Table 4 reached almost 0.8 F1 and one of the reasons may be that polarities are easier to separate from each other – usually people use different expressions and different words when evaluating positively or negatively a movie or app, the same does not always apply for helpfulness. Even though the best result obtained in *Buscapé* corpus was 0.8174 in F1, other authors achieved 0.8935% in the same corpus [Avanço et al. 2016].

One of the reasons for the low results is that the representation used (pre-trained word embeddings) usually works well with neural models since they basically rearrange the data using weights. It may explain why the best results for the whole corpus were obtained using MLP (Table 5), which follows the same principle with less layers. Another limitation was the size of the dataset. Linguistic approaches (which use n-grams for example) demand more resources for storage and processing. The $t\text{-}value$ between results was measured in order to calculate the significance of the differences and all of them were significant at $p < 0.05$.

5. Related Work

This paper relates to several other works both in helpfulness detection and sentiment analysis due to its similarities between fields.

For helpfulness prediction as a classification task, [Krishnamoorthy 2015] examines the impact of some specific linguistic features based on a model named Linguistic Category Model (LCM) [Semin 2011], on helpfulness prediction task. The author builds three machine learning methods for helpfulness binary classification, using a threshold $h = 0.60$.

Using a corpus extracted from Amazon.com (MDSD), the Random Forest method achieved the best result reaching an average of 84% of F-measure using all features. Individually the LCM features obtained the best results.

[Zeng et al. 2014] addressed the helpfulness prediction problem as a three-class classification problem. The classes are (1) Helpful positive reviews (star rating ϵ [4,5] and helpfulness score $h > threshold$); (2) Helpful negative reviews (star rating ϵ [1,2]

and helpfulness score $h > threshold$), and (3) Unhelpful reviews (helpfulness score $h < threshold$). They collected 8.690 reviews from Amazon.com. The experimentation included an empirical test to decide the helpfulness score threshold. The best value obtained 72.82% of accuracy on ten-fold cross-validation. Specifically, regarding each class, the helpful positives reached 69% in macro-f1; the helpful negatives, 79,5% in macro-f1 and the unhelpful ones, 80% in macro-f1.

[Malik and Hussain 2017] used an emotion score of reviews (confidence, surprise, anger, etc.) as a feature to predict helpfulness. The authors modeled and evaluated a set of learning methods on Amazon.com corpus (MDSD), and they achieved 89% of f-measure, using emotion features as input for a deep neural network method.

In [Hartmann et al. 2014] the authors introduce *Buscapé*, a corpus for user-generated content research constructed using product reviews from an e-commerce website in Brazilian Portuguese. The authors extracted 85.910 documents and annotated typos and Internet slangs for normalisation task. The corpus also contains a 5-star-based score and "like" votes that we used in this paper (section 4) for comparison with our own results in UTLcorpus.

This data set has been used several times in literature [Avanço et al. 2016, Brum and Nunes 2018, Bertaglia 2017]. We emphasize the work of [Avanço et al. 2016] because the authors classified the data using machine learning classifiers (SVM and Naive Bayes), lexical-based classifiers and ensemble of classifiers (both machine learning-based and lexical-based) and achieved the state-of-the-art for the corpus, 0.8935 in f1. The main difference of this paper with ours is that we only used machine learning classifiers and we used embeddings for data representation, whilst those authors used a combination of *bag-of-words* and linguistic features such as number of sentiment words and PoS tags.

We can also compare our work with [Corrêa et al. 2017] since they also annotated a large corpus for semantic purposes (polarity classification). In this paper the authors crawled Twitter for Brazilian Portuguese posts and used Distant Supervision, automatically labeling documents based on semantic clues, to form a large corpora for sentiment analysis. Pelesent is composed of 980.067 tweets that contained emojis and/or emoticons indicating negative or positive polarity (eg. ":)" for positive and ":(" for negative).

6. Discussion and future work

In this paper we presented UTLcorpus, a review corpus of two domains (movies and apps) with automatic labels for helpfulness detection and polarity classification. We proposed an automatic label methodology for helpfulness using "like" votes and used a literature inspired method for attaching a polarity (positive or negative) to $2.755.959$ forming two corpora – one of Movie Reviews ($1.834.702$ documents) and one of App Reviews (921.257). The methodology was replicated in a similar corpus (Buscapé) in order for comparing sizes and results obtained in classification experiments.

UTLcorpus is one of the first data sets for helpfulness detection in Brazilian Portuguese, but it can also be used for sentiment analysis (polarity classification, aspect extraction or else) and for others NLP tasks such as language modeling, normalization, discourse analysis or semantic parsing, for example.

We evaluated the corpus using machine learning methods from the literature and obtained results up for replication and comparison with other models. The dataset is available in the github `github.com/RogerFig/UTLCorpus` already pre-processed, normalized with Enelvo and anonymised in order to be used for research purposes.

One of the future works to be conduct is the exploration of state-of-the-art models for classification such as convolutional neural networks [Kim 2014] and Long-short Term Memory architectures as well as the investigation of different representation models – morphological-based embeddings [Bojanowski et al. 2017] or context embeddings such as Elmo [Gardner et al. 2017]. Another future work is to expand the usefulness prediction for handling comment rating so that a list of best comments can be presented to users. Finally, manual evaluation of helpfulness can be performed on reviews, although we believe that a reader who is not interested in a product will handle a review differently from an interested one.

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.

Avanço, L. V. (2015). Sobre normalização e classificação de polaridade de textos opinativos na web.

Avanço, L. V., Brum, H. B., and Nunes, M. d. G. V. (2016). Improving opinion classifiers by combining different methods and resources. *XIII Encontro Nacional de Inteligência Artificial e Computacional*.

Bertaglia, T. F. C. (2017). Normalização textual de conteúdo gerado por usuário.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brum, H. B. and Nunes, M. d. G. V. (2018). Building a sentiment corpus of tweets in brazilian portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resources Association.

Corrêa, E. A., Marinho, V. Q., dos Santos, L. B., Bertaglia, T. F. C., Treviso, M. V., and Brum, H. B. (2017). Pelesent: Cross-domain polarity classification using distant supervision. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 49–54. IEEE.

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.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. S. (2017). Allennlp: A deep semantic natural language processing platform.

Hartmann, N., Avanço, L., Balage Filho, P. P., Duran, M. S., Nunes, M. D. G. V., Pardo, T. A. S., Aluísio, S. M., et al. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *LREC*, pages 3865–3871.

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

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.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Malik, M. and Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, 73:290–302.

McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52, New York, NY, USA. ACM.

Semin, G. R. (2011). The linguistic category model. *Handbook of theories of social psychology*, 1:309–326.

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

Tang, J., Gao, H., Hu, X., and Liu, H. (2013). Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 1–8, New York, NY, USA. ACM.

Zeng, Y.-C., Ku, T., Wu, S.-H., Chen, L.-P., and Chen, G.-D. (2014). Modeling the helpful opinion mining of online consumer reviews as a classification problem. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 2, June 2014*, 19(2).