What Makes a Review Helpful?
Predicting the Helpfulness of Italian TripAdvisor Reviews

Giulia Chiriatti•◇, Dominique Brunato◇, Felice Dell'Orletta◇, Giulia Venturi◇
• University of Pisa
chiriattigiulia@gmail.com
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
chiriattigiulia@gmail.com
{dominique.brunato,felice.dellorletta,giulia.venturi}@ilc.cnr.it

Abstract

In this paper we introduce a classification system devoted to predict the helpfulness of Italian online reviews. It is based on a wide set of features reflecting the different factors involved and tested on different categories of TripAdvisor reviews. For this purpose, we collected the first Italian corpus of online reviews enriched with metadata related to their helpfulness and we carried out an in-depth analysis of the most predictive features.[1]

1 Introduction

Predicting and modeling relevant factors that determine the helpfulness of online reviews have been attracting a growing attention in the Natural Language Processing (NLP) community. Both practical applications and the interest to study human variables underlying the assignment of helpful/unhelpful votes are mainly involved. The identification of product reviews which are useful to customers can be important for several e-business purposes (e.g. the development of product recommendation systems) as well as to investigate persuasive elements that make a review helpful for a review reader (Hong et al., 2012; Park, 2018). Several approaches have been devised, differing at the level of predicting methods (mainly regression or classification algorithms) and of typologies of factors considered, including content elements found within the review and contextual ones referring to user profiles. Although various strategies have already been followed, according to the recent survey by Diaz and Ng (2018), a number of issues are still open and deserve to be explored. Among others, they include *i)* the need for "more sophisticated textual features" that can be useful to model a writing style typical of helpful reviews, and *ii)* the lack of studies focused on languages other than English.

In this paper, we address these open issues and we present a study devoted to predict Italian review helpfulness with a specific focus on the role played by linguistic features in modelling the style of helpful reviews. Similarly to previous studies, we tackled the task as a text classification problem but with two main novelties. Firstly, we relied on different sets of predictors, considering both lexical (content) and structural features (i.e. morphosyntactic and syntactic) aimed at reconstructing the style of a text (the linguistic "form"). Secondly, we investigated which typology of features are the most effective to predict the helpfulness of online reviews and whether they remain the same across different review categories.

Our contribution. *i*) We collected a corpus of Italian online reviews enriched with metadata related to their helpfulness[2]. *ii*) We developed the first classification system devoted to predict the helpfulness of Italian online reviews, based on features modelling both lexical and linguistic factors involved, and tested it in two experimental scenarios, i.e. in- and out-domain with respect to the training category of reviews. *iii*) We identified and ranked the most predictive features, showing the key role played by linguistic features, especially to predict the helpfulness of reviews belonging to a category very different from the training one.

2 Corpus

We collected a sample of almost 1 million user-generated reviews from the Italian section of TripAdvisor, focusing on two travel-related categories, restaurants and attractions (e.g. parks, historical sites), and two geographical areas, Rome

[1]

[2] The corpus is available for research purposes at http://www.italianlp.it/resources/

and Milan. We also gathered two types of metadata associated with each review: review rating and number of helpful votes. Firstly, we filtered our data according to language (Italian) and length ($>$ 7 tokens), discarding 52.29% of the total reviews. Then we empirically[3] set a threshold at a minimum of 3 votes in order to distinguish helpful reviews (3+ votes) from unhelpful ones (0 votes). Some examples of reviews that belong to the two classes are reported in Table 2. In line with studies carried out for the English language (Park, 2018), also in our case review votes tend to be sparse across all categories: in particular reviews with 3+ votes constitute only 5.10% of the unfiltered dataset. For this reason we balanced the data by selecting a comparable number of helpful and unhelpful reviews per restaurant or attraction. As shown in Table 1, our final corpus consists of 42,107 reviews from 1,218 restaurants and 383 attractions for a total of 4,133,312 tokens.

Category	#Helpful	#Unhelpful	#Reviews
Rome rest.	12,635	12,404	25,039
Milan rest.	6,105	5,991	12,096
Attractions	2,564	2,408	4,972
TOTAL	21,304	20,803	42,107

Table 1: Corpus of helpful and unhelpful TripAdvisor reviews.

3 Helpfulness Predictors

According to our research purposes, we considered various categories of features aimed at modeling both the content and the linguistic "form" of online reviews. They can be grouped into three main classes: *lexical*, *linguistic* and *metadata* features. The first typology has already been tested in the literature (Diaz and Ng, 2018) in order to predict review helpfulness on the basis of meaningful words. On the contrary, the use of linguistic features extracted from sentence structure is introduced for the first time in this paper. Differently from previous studies (Kim et al., 2006; Hong et al., 2012) where the distribution of some Parts-Of-Speech was exploited as helpfulness predictor, we rely here on a wide set of linguistic features automatically extracted from the corpus of reviews linguistically annotated. Since they have been shown to have a high discriminative power in different tasks, e.g. assessment of text readability (Collins, 2014), identification of textual genre of a document (Cimino et al., 2017), we investigated in this study whether they are able to model the linguistic "form" (the style) of helpful reviews. In addition, we explored the contribution of a kind of metadata feature (i.e. the star rating given by the reviewer) that has also been widely tested in studies on helpfulness prediction, as reported in Diaz and Ng (2018).

In order to extract lexical and linguistic predictors of helpfulness, the corpus was linguistically annotated at different levels of analysis. In particular, it was tagged by the PoS tagger described in Dell'Orletta (2009) and dependency-parsed by the DeSR parser (Attardi et al., 2009).

Lexical features. They include two types of features: (*i*) the distribution of unigrams and bigrams of characters, words and lemmas (hereafter *NGR*); (*ii*) word embedding combinations (*WE*) obtained by separately computing the average of the vector representations of nouns, verbs and adjectives in the review. The word embeddings were trained on the ItWaC corpus (Baroni et al., 2009) and a collection of Italian tweets[4] using the *word2vec* toolkit (Mikolov et al., 2013).

Linguistic features. They refer to four main types, modelling diverse aspects of writing style:
raw text features, i.e. review, sentence and word length, calculated in terms of sentences, tokens and characters, respectively;
features related to lexical richness, which is captured considering *i)* the internal composition of the vocabulary of review with respect to the *Basic Italian Vocabulary* and its usage repertories (De Mauro, 2000), and *ii)* Type/Token Ratio;
morpho-syntactic features, i.e. the distribution of unigrams of Parts-of-Speech, and verb moods, tenses and persons;
syntactic features, which refer to diverse characteristics of sentence structure: *i)* the depth of the whole parse tree (calculated in terms of the longest path from the root of the dependency tree to some leaf); *ii)* the length of dependency links (i.e. the tokens occurring between the head and the dependent); *iii)* the distribution of dependency types, *iv)* the average depth and the distribution of embed-

[3] In order to choose the threshold value, we considered the mean and the standard deviation of the number of votes in the initial dataset (2.21 ± 0.59).

[4] http://www.italianlp.it/resources/italian-word-embeddings/

Label	Category	Example (*Italian*)	Example (*English*)
Helpful	Rome restaurants	La prima regola di un buon ristorante che fa pizza no stop è: Scegliere la pizza che preferisco. Qui non solo non si può scegliere la pizza ma capita spesso che escano le stesse pizze più volte così uno è costretto a mangiare sempre la stessa!! Per non parlare dell'ambiente poi, un vero casino, capisco che l'area bambini è la principale attrazione del ristorante, rivolto soprattutto alle famiglie, ma il casino che si crea non è cmq giustificabile. La pizza è di una qualità davvero scadente, praticamente era cruda!!! La pizza con la Lonza....una semplice focaccia con un pezzo di prosciutto preso molto probabilmente al discount! Ragazzi, carina l'idea di prendersi cura dei pargoli, ma non prendiamoci in giro però.	The first rule of a good restaurant that makes pizza no stop is: Choose the pizza I prefer. Here you can not only choose the pizza but it often happens that the same pizzas come out more times so one is forced to always eat the same one!!! Not to mention the environment then, a real mess, I understand that the children's area is the main attraction of the restaurant, aimed above all at families, but the mess that is created is not justifiable anyway. The pizza is of a really poor quality, practically it was raw!!! Pizza with Lonza....a simple focaccia with a piece of ham most probably taken at the discount store! Guys, nice idea to take care of the little ones, but let's not fool around.
Unhelpful	Milan restaurants	Devo dire che trovandomi per caso in quella zona con i miei amici abbiamo provato il posto è devo dire ché è molto accogliente e che la zona per mangiare nel cortile è proprio intima e carina...Per quanto riguarda il mangiare posso dire di essere soddisfatto perché le portate erano nelle mie corde ed avendo preso il pesce ero soddisfatto di quanto cucinato dal cuoco. Bravi mica male.	I must say that finding myself by chance in that area with my friends we tried the place and I must say that it is very welcoming and that the area to eat in the courtyard is really intimate and pretty... As for eating I can say I'm satisfied because the courses were on my ropes and having caught the fish I was satisfied with what the cook had cooked.

Table 2: Examples of helpful vs unhelpful reviews.

ded prepositional chains modifying a noun; *v)* a set of features aimed at modeling the behaviour of verbal predicates, i.e. the number of verbal roots, the average verbal arity and the distribution of verbs by arity, the distribution of verbal predicates with elliptical subject; *vi)* the usage of subordination, calculated considering the ratio between principal and subordinate clauses, and the average depth and the distribution of embedded chains of subordinate clauses; *vii)* a last set of features related to the canonical construction of a sentence in Italian, i.e. the relative ordering of subordinates with respect to the main clause and of subject and object with respect to their verbal head.

The effectiveness of these features to predict helpful online reviews is confirmed by the fact that according to the Wilcoxon rank sum test, 75% of the considered features (i.e. 160 out of 212) turned out to vary in a statistically significant way between helpful and unhelpful reviews. As shown in Table 3, helpful reviews are on average 1-sentence longer than unhelpful ones and they also contain much longer sentences. The correlation between length and helpfulness is not surprising since longer sentences are likely to be more informative, thus offering more contents that might influence the voting process outcome. The higher sentence length also has an expected effect on some syntactic features correlated to complexity. Sentences occurring in helpful texts have deeper syntactic trees (*Avg. max depth*) and contain more subordinate clauses and embedded prepositional chains. However, they appear as simpler with respect to other features related for instance to canonicity effects. They show a more standard syntactic structure, with a higher distribution of objects in post verbal position and subjects preceding the main verb. Interestingly, helpfulness is also positively correlated with a reader-focused style, as shown by the greater use of pronouns and verbs in the first and second person.

Metadata feature. *Review star rating (STR)* is the rating score assigned by the reviewer, ranging from 1 to 5. Previous research reported in Diaz and Ng (2018) has shown that a connection exists between the rating of the review and its helpfulness. In our dataset rating scores are unequally distributed across the different review categories. Restaurant reviews are more likely to have an extreme rating, either low or high, rather than a neutral one, and helpful reviews follow the same pattern: e.g., in the Rome restaurant cate-

Feature	*Help*	*UnHelp*	*Diff.*
N. sent	4,61	3,46	1,15
Avg. sent length	36.79	26.22	10.57
Avg. clause length	10	11.65	-1.65
% Nouns	23.5	24.5	-1
% Verbs	14.28	12.79	1.49
% Adj	8.32	10.37	-2.41
% Negative adv	1.33	0.97	0.36
% Pronouns	4.99	4.14	0.85
% 1st sing p.	9.23	8.15	1.08
% 2nd pl p.	1.34	1.08	0.26
Avg. prep chains length	11,4	6,3	5,1
Avg. max depth	7,64	6,28	1,36
% Subord clause	62,09	43,89	18,2
% Post obj	78,84	68,66	10,18
% Pre subj	73,13	65,03	8.01

Table 3: A subset of linguistic features whose values vary in a statistically significant way between helpful and unhelpful reviews.

gory 37.05% of the helpful reviews have a rating of 1 and 25.76% a rating of 5. On the contrary, attractions reviews tend to have higher ratings, with 56.12% of the helpful ones belonging to the highest-rated class. Only the attractions category seems to confirm the presence of the *positivity bias* that is discussed in Diaz and Ng (2018), according to which reviews with positive ratings are seen as more helpful.

4 Experiments and Results

We addressed the helpfulness prediction task as a binary classification problem. In order to assess the contribution of each set of features illustrated in Section 3, we defined two experimental scenarios differing at the level of review categories chosen as test data and set-up (in terms of feature configurations). We built a classifier based on the LIBLINEAR implementation of Support Vector Machines with a linear kernel (Fan et al., 2008) and trained on a set of 12,516 reviews written for 411 Rome restaurants. All the features were previously scaled in the same range $[0, 1]$. We evaluated our system by computing the accuracy score for each feature configuration. As baseline for each review category we implemented the score of a classifier which always outputs the most probable class according to the class distribution of the dataset (in this case the *helpful* class).

In the first experimental scenario we tested the feature models generated by the SVM classifier on a test set of 12,523 reviews that belong to the same domain of the training data (i.e. the Rome restaurants category) but were written for restaurants different from the ones in the training set. As shown in Table 4, we obtained a general improvement over the baseline with all feature configurations apart from the one that exploits only the metadata feature (*STR*, the star rating of the reviews). Nevertheless, this feature does improve the accuracy score of all models by at least one point, thus confirming its usefulness for helpfulness prediction (Diaz and Ng, 2018). The results also highlight the prominent role of lexical information (*NGR+WE*) in assessing helpfulness, although this is primarily explained by the in-domain scenario. Even if the accuracy of the linguistic model (*LING*) is lower with respect to the one obtained by the other feature models, we found out that linguistic information plays a main role in the helpfulness prediction. It allows achieving an accuracy score of 66% and of 70.81% by also adding review ratings, a value that is in line with that of the lexical model.

Model	**Accuracy**
STR	49.6%
NGR	69.9%
NGR+STR	71.13%
WE	68.54%
WE+STR	69.96%
NGR+WE	70.17%
NGR+WE+STR	**71.14%**
LING	66%
LING+STR	70.81%
ALL	70.04%
ALL+STR	71.05%
Baseline	50.46%

Table 4: In-domain classification of helpful vs. unhelpful reviews using different feature models.

In the out-domain scenario we tested the considered feature models on reviews that belong to the other two categories (Milan restaurants and attractions). As reported in Table 5, we observed that the performances of the classifier tested on the reviews of Milan restaurants, even if slightly worse, are very similar to the ones obtained on the test set of Rome restaurants. This result suggests that the system may perform consistently across different geographical areas, although further experiments should be carried out. For example, we might test our models on a greater number of cities or other types of geographical areas. As we expected, the accuracy decreases mainly in the domain more distant from the training one (i.e. the attractions category). This is especially the case of the lexical classification model, that has a drop of 10.5 points.

The star rating feature is also shown to worsen the accuracy scores, probably because of the way the ratings are distributed in the attractions category with respect to the restaurant ones. It is interesting to note that the best performing model resulted to be the one exploiting the linguistic features (with a lower drop of 5.24%), thus showing the predictive power of sentence structure information in predicting review helpfulness.

Model	Milan	Attractions
NGR+WE	69.38%	59.67%
NGR+WE+STR	**70.92%**	58.02%
LING	65.82%	**60.76%**
LING+STR	**70.92%**	60.28%
ALL	69.2%	59.9%
ALL+STR	70.78%	58.49%
Baseline	50.47%	51.56%

Table 5: Out-domain classification of helpful vs. unhelpful reviews in terms of accuracy using different feature models.

5 Discussion

As discussed in the previous section, we found out that linguistic features allow achieving an accuracy almost in line with the one obtained using only lexical information. Interestingly enough, they are the most predictive ones in the out-of-domain scenario. In order to gain insight into which of these features are the most effective in the task of automatic classification, we ranked them according to the absolute value of their weight in the linear SVM model generated with the linguistic feature configuration. Among the 50 top-ranked ones, besides the raw text features (whose role in predicting helpfulness has already been proven in the literature), we found morpho-syntactic and syntactic features. They are typically related to a rich and articulated writing style. This is the case for example of features concerning nominal modification, in particular the number of prepositional chains (holding the 1st position in the ranking) and their average length but also the distribution of adjectives and determiners. Others involve verbal structures, e.g. the number of dependents instantiated by the verbal heads and the frequency of adverbs (especially negation ones). Features related to the usage of subordination, such as the number of subordinate structures and the average depth of parse trees, also appear among the top-ranked. Finally, another group of high-ranked features concerns a subjective writing style, as shown by the distribution of verbs in the first and second person. These types of features resulted to be discriminant in the comparison between helpful and unhelpful reviews (Section 3). This shows that the writing style of helpful reviews, informative but also personal and reader-focused, has an high predictive power.

The importance of the linguistic features is further confirmed by a second inspection in which the same ranking method was applied to the all-feature model. Also in this case, we found out that 59.6% of the whole set of 212 linguistic features we considered is in the 90th percentile of the ranking of the total 741,339 features.

6 Conclusion

In this paper, we have presented the first approach to the task of review helpfulness prediction for the Italian language. Two experimental scenarios have been tested in a corpus of TripAdvisor reviews belonging to different categories (restaurants and attractions). In line with previous findings obtained for the English language, we confirmed that lexical information plays a significant role in classifying helpful reviews. In addition, we proved for the first time the highly predictive power of linguistic features modeling the writing style independently from the content. This is particularly true in the two out-domain experiments: in the first case (same category, different geographical area), the classifier based on the linguistic features achieves the same accuracy of the model using lexical features and it even outperforms all the other configuration models when tested on the most distant review category (restaurants vs attractions).

Among the possible future issues that we would like to investigate, an interesting one concerns the role played by metadata features. In the reported results, we showed that star ratings are not relevant when considered alone, but they give a plus when combined with both lexical and linguistic features. Beyond this metadata, we would like to extend the analysis to further user information possibly related to review helpfulness.

Acknowledgments

This work was partially supported by the 2-year project PERFORMA (Personalizzazione di pERcorsi FORMativi Avanzati), co-funded by Regione Toscana (POR FSE 2014-2020).

References

G. Attardi, F. Dell'Orletta, M. Simi and J. Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian* , Reggio Emilia, December.

M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta. 2009 The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3), pp. 209-226.

A. Cimino, M. Wieling, F. Dell'Orletta, S. Montemagni S. and G. Venturi. 2017 Identifying Predictive Features for Textual Genre Classification: the Key Role of Syntax. *Proceedings of 4th Italian Conference on Computational Linguistics (CLiC-it)*, 11-13 December, 2017, Rome.

K. Collins-Thompson. 2014. Computational assessment of text readability: a survey of current and future research. *Recent Advances in Automatic Readability Assessment and Text Simplification*, Special issue of International Journal of Applied Linguistics, (165-2), 97-135.

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian* , Reggio Emilia, December.

T. De Mauro. 2000. Grande dizionario italiano dell'uso (GRADIT). Torino, UTET.

G. O. Diaz and V. Ng. 2018. Modeling and Prediction of Online Product Review Helpfulness: A Survey. ACL, 2018.

R. E. Fan, K.-W. Chang, C.-J. Hsieh, X. Wang and C.-J. Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Y. Hong, J. Lu, J. Yao, Q. Zhu and G. Zhou. 2012. What Reviews are Satisfactory: Novel Features for Automatic Helpfulness Voting. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 495-504.

S. M. Kim, P. Pantel, T. Chklovski and M. Pennacchiotti. 2006. Automatically assessing review helpfulness. *Proceedings of the the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 423-430.

T. Mikolov, K. Chen, G. Corrado and J. Dean. 2013 Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781

Y-J. Park. 2018. Predicting the helpfulness of online customer reviews across different product types. *Sustainability*, 10(1735), pp. 1-20.