Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining

Marco Passon[†], **Marco Lippi**[‡], **Giuseppe Serra**[†], **Carlo Tasso**[†]

[†]Università degli Studi di Udine
[‡]Università degli Studi di Modena e Reggio Emilia
marco.passon@spes.uniud.it
marco.lippi@unimore.it
{giuseppe.serra, carlo.tasso}@uniud.it

Abstract

Internet users generate content at unprecedented rates. Building intelligent systems capable of discriminating useful content within this ocean of information is thus becoming a urgent need. In this paper, we aim to predict the usefulness of Amazon reviews, and to do this we exploit features coming from an off-the-shelf argumentation mining system. We argue that the usefulness of a review, in fact, is strictly related to its argumentative content, whereas the use of an already trained system avoids the costly need of relabeling a novel dataset. Results obtained on a large publicly available corpus support this hypothesis.

1 Introduction

In our digital era, reviews affect our everyday decisions. More and more people resort to digital reviews before buying a good or deciding where to eat or stay. In fact, helpful reviews allow users to grasp more clearly the features of a product they are about to buy, and thus to understand whether it fits their needs. The same can be said for users who want to book hotels or restaurants.

Companies have started to exploit the importance of reviews. For example, when browsing for a specific product, we are usually presented reviews that have been judged helpful by other users. Moreover, we are often given the possibility to sort reviews according to the number of people who judged them as helpful. That said, a review can also be helpful for companies who want to monitor what people think about their brand. Being able to identify helpful reviews has thus many important applications, both for users and for companies, and in multiple domains.

The automatic identification of helpful reviews is not as easy as it may seem, because the review content has to be semantically analyzed. Therefore, this process is traditionally done by asking users for a judgment.

To overcome this issue, some approaches have been proposed. One of the earliest studies (Kim et al., 2006) aims to rank Amazon reviews by their usefulness by training a regressor with a combination of different features extracted from text and metadata of the reviews, as well as features of the product. Similar approaches employ different sets of features (Ngo-Ye and Sinha, 2012), for example including the reputation of reviewers too (Baek et al., 2012). Another significant work (Mudambi and Schuff, 2010) builds a customer model that describes which features of an Amazon review affect its perceived usefulness, and then it uses such features to build a regression model to predict the usefulness, expressed as the percentage of the number of people who judged a review to be useful. A hybrid regression model (Ngo-Ye and Sinha, 2014) combines text and additional features describing users (recency, frequency, monetary value) to predict the number of people who judged as useful reviews taken from Amazon and Yelp. A more complete work considers both regression and classification (Ghose and Ipeirotis, 2011). It proves different hypotheses, starting with expressing the usefulness of an Amazon review as a function of readability and subjectivity cues, and then converting the usefulness, expressed with a continuous value, into a binary usefulness, that is predicting if a review is useful or not useful.

Another recent work (Liu et al., 2017) presents an approach that explores an similar assumption to ours: helpful reviews are typically *argumentative*. In fact, what we hope to read in a review is something that goes beyond plain opinions or sentiment, being rather a collection of reasons and evidence that motivate and support the overall judgment of the product or service that is reviewed. These characteristics are usually cap-

tured by an argumentation analysis, and could be automatically detected by an argumentation mining system (Lippi and Torroni, 2016a). The work in (Liu et al., 2017) considers a set of 110 hotel reviews, it presents a complete and manual labeling of the arguments in such reviews, and it exploits such information as additional features for a machine learning classifier that predicts usefulness. In this paper, instead, we investigate the possibility to predict the usefulness of Amazon reviews by using features coming from an automatic *argumentation mining system*, thus not directly using human-annotated arguments. A preliminary experimental study conducted on a large publicly dataset (117,000 Amazon reviews) confirms that this could be really doable and a very fruitful research direction.

2 Background

Argumentation is the discipline that studies the way in which humans debate and articulate their opinions and beliefs (Walton, 2009). Argumentation mining (Lippi and Torroni, 2016a) is a rapidly expanding area, at the cross-road of many research fields, such as computational linguistics, machine learning, artificial intelligence. The main goal of argumentation mining is to automatically extract arguments and their relations from plain textual documents.

Among the many approaches developed in recent years for argumentation mining, based on advanced machine learning and natural language processing techniques, the vast majority is in fact genre-dependent, or domain-dependent, as they exploit information that is highly specific of the application scenario. Due to the complexity of these tasks, building general systems capable of processing unstructured documents of any genre, and of automatically reconstructing the relations between the arguments contained in them, still remains an open challenge.

In this work, we consider a simple definition of argument, inspired by the work by Douglas Walton (2009), that is the so-called claim/premise model. A *claim* can be defined as an assertion regarding a certain topic, and it is typically considered as the conclusion of an argument. A *premise* is a piece of evidence that supports the claim, by bringing a contribution in favor of the thesis that is contained within the claim itself.

3 Methodology

Our goal is to develop a machine learning system capable of predicting the usefulness of a review, by exploiting information related to its argumentative content. In particular, we consider to enrich the features of a standard text classification algorithm with features coming from an argumentation mining system. To this aim, we use MARGOT (Lippi and Torroni, 2016b), a publicly available argumentation mining system[1] that employs the claim/premise model (to our knowledge, there are no other off-the-shelf systems that perform argumentation mining). Two distinct classifiers, based on Tree Kernels (Moschitti, 2006) are trained to detect claims and premises (also called evidence), respectively. When processing a document, MARGOT returns two scores for each sentence, one computed by each kernel machine, that are used to predict the presence of a claim or a premise within that sentence (by default, MARGOT uses a decision threshold equal to zero).

Consider for example the following excerpt of a review, where the proposition in italics is identified by MARGOT as a claim:

> The only jam band I ever listen to now is Cream, simply because they were geniuses. They were geniuses because the spontaneity, melodicism, and *fearlessness in their improvisation has never been equaled in rock*, and rarely so in jazz.

Clearly, such a review is very informative, since it comments on very specific aspects of the product, bringing motivations that can greatly help users in taking their decisions. Similarly, the following excerpt of another review brings very convincing arguments in favor of an overall positive judgment of the product. In this case, both sentences are classified by MARGOT as argumentative.

> *The music indeed seems to transcend so many moods that most pianists have a very hard time balancing this act and there is an immense discography of these concertos of disjoint and loosely-knit performances. Pletnev pushes a straightforward bravura approach with lyrical interludes – and his performance pays off brilliantly.*

[1] http://margot.disi.unibo.it

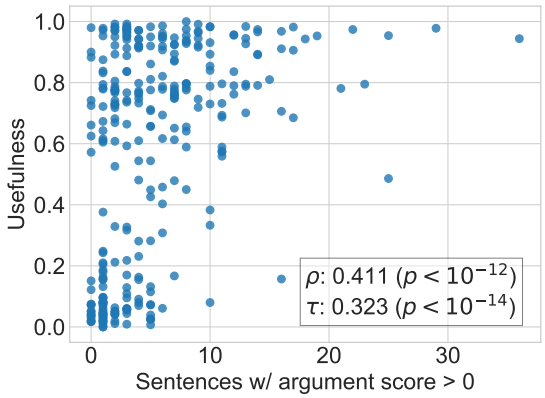

Figure 1: Relation between usefulness and number of sentences whose claim or evidence score is above zero for category "CDs and Vinyl".

Within this work, we compute simple statistics from the output of MARGOT: the average claim/evidence/argument score, the maximum claim/evidence/argument score, the number and the percentage of sentences whose claim/evidence/argument score is above 0 (that is, the number and the percentage of sentences that contain a claim, an evidence or simply one of those). From a preliminary analysis, in fact, we observed how the presence of arguments within a review is highly informative of its usefulness. Figure 1, for example, shows the correlation of the number of sentences whose claim or evidence score, according to MARGOT, is above 0, with the usefulness for a subset of 200 reviews in the Amazon category "CDs and Vinyl". While it is true that a low number of sentences that contain a claim or an evidence does not necessarily mean that the review is useless, yet the figure shows that a review with a high number of sentences containing a claim or an evidence is most likely a useful review, which confirms our intuition that useful reviews are in fact argumentative. We use these simple statistics as an additional set of features to be used within a standard text classification algorithm, in order to assess whether the presence of argumentative content can help in predicting how useful is a review.

We hereby remark that using MARGOT within this framework is not optimal, because MARGOT was trained on a completely different genre of documents, that is Wikipedia articles. Therefore, we are dealing with a *transfer learning* task, where the argumentation mining system is tested on a different domain with respect to the one it was originality trained on. Using such a classifier adds a challenge to our approach, but it has the advantage of not needing a labeled corpus of argumentative reviews to train a new argumentation mining system from scratch. Indeed, more sophisticated systems that take into account argumentation could be developed: here, we just want to exploit a straightforward combination of features in order to test our hypothesis.

4 Experimental Results

To evaluate the proposed approach we use the public Amazon Reviews dataset (McAuley and Leskovec, 2013), in particular, we worked with the so called "5-core" subset, that is, a subset of the data in which all users and items have at least five reviews. Each element of this dataset contains a product review and metadata related to it.

Since we aim to predict usefulness, for each review we compute the ratio between the number of people who voted and judged that review as useful, and the total number of people who expressed a judgment about that review. Then, we define useful reviews as the ones whose percentage of usefulness is equal or above 0.7 (that means that at least 70% of the people who judged a review, judged it as useful), while the remaining are considered not to be useful, and thus they represent our negative class.

The Amazon Review dataset is split into product categories. For our experiments we picked three of them, chosen among those with the highest number of reviews. Our choice has fallen upon the "CDs and Vinyl" "Electronics" and "Movies and TV" categories. We further selected only the reviews having at least 75 rates, in order to assess usefulness on a reasonably large set of samples. Finally, we randomly selected 39,000 reviews for each category, ending up with an almost balanced number of helpful and unhelpful reviews.

Our goal in executing the experiments is to predict whether a review is considered useful, by taking into account either its textual content only, or, additionally, also the argumentation mining data coming from MARGOT. In other words, we are working in a binary classification scenario.

In these experiments we use a stochastic gradient descent classifier[2] with a hinge loss, which is a classic solution in binary classification tasks. We

[2]We used `SGDClassifier` in *scikit-learn*.

Table 1: Performance on three Amazon categories using different sets of features: Margot features (M), Bag-of-Words (BoW), Bag-of-Words weighted by TF-IDF (TF-IDF), and combinations thereof.

Category	Data	A	P	R	F_1
CDs and Vinyl	M	.600	.544	.772	.638
	BoW	.756	.716	.769	.742
	BoW + M	.784	.744	**.799**	.771
	TF-IDF	.769	.736	.767	.752
	TF-IDF + M	**.787**	**.751**	.797	**.773**
Electronics	M	.583	.529	**.744**	.618
	BoW	.676	.639	.656	.648
	BoW + M	**.689**	.640	.714	**.675**
	TF-IDF	.672	**.651**	.612	.631
	TF-IDF + M	**.689**	.649	.684	.666
Movies and TV	M	.564	.517	.792	.625
	BoW	.745	.705	.748	.726
	BoW + M	.773	**.741**	.767	.754
	TF-IDF	.757	.719	.761	.740
	TF-IDF + M	**.777**	.739	**.784**	**.761**

performed the tuning of the α and ϵ parameters with a 5-fold cross validation over the training set, and we then used the best model to predict over the test set. From the original set of 39,000 reviews, 50% of them is used as training set, and the other half as the test set. Each category is treated singularly.

We run experiments both employing a plain Bag-of-Words model, and with TF-IDF features. Both preprocessing variants perform tokenization and stemming[3] and exclude stopwords and words that do not appear more than five times in the whole training set. To regularize the different magnitude of the features, both textual features and argumentation mining features are normalized using the L2 normalization in all our experiments. Textual and argumentative features are simply concatenated into a single vector. The performance is measured in terms of accuracy (A), precision (P), recall (R), and F_1, as in standard text classification applications.

Table 1 shows that, even using only the features obtained from MARGOT, thus completely ignoring the textual content of the review, the accuracy of the classifier is far above a random baseline. Moreover, results clearly highlights how the improvement obtained by using argumentative features is consistent across all product categories, both using plain BoW and TF-IDF weighting. For the "CDs and Vinyl"and "Electronics"categories the difference between the classifier exploiting TF-IDF with MARGOT and the one using TF-IDF only is statistically significant according to a McNemar's test, with p-value < 0.01. The same holds for the BOW classifier, for the "Electronics"and "Movies and TV"categories.

It is interesting to notice that, while the "CDs and Vinyl" and the "Movies and TV" categories have similar performance, even when using textual data only, the category "Electronics" results to be the most difficult to predict. One plausible explanation for this is the heterogeneity of such category, that includes many different types of electronic devices. The other two categories, instead, include more homogeneous products. It would be very interesting to further investigate whether certain product categories result to be more suitable for argumentation studies.

5 Conclusions

When reading online reviews of products, restaurants, and hotels, we typically appreciate those that bring motivations and reasons rather than plain opinions. In other words, we often look for *argumentative* reviews. In this paper, we proposed a first experimental study that aims to show how features coming from an off-the-shelf argumentation mining system can help in predicting whether a given review is useful.

We remark that this is just a preliminary study, which yet opens the doors to several research directions that we aim to investigate in future works. First, we certainly plan to use more advanced machine learning systems, such as deep learning architectures, that have achieved significant results in many applications related to natural language processing. In addition, we aim to address different learning problems, for example moving to multi-class classification, or directly to regression.

The combination of textual and argumentative features exploited in this work was effective in confirming our intuition, but it can certainly be improved. While building a dedicated argumentation mining system for product reviews could require an effort in terms of corpus annotation, we believe that transfer learning here could play a crucial role. Beyond using statistics obtained from the output of an argumentation mining system as an additional input for a second-stage classifier, a unified model combining the two steps could result to be a smart compromise for this kind of application.

[3] We used `Snowball` from python `nltk` library.

References

Hyunmi Baek, JoongHo Ahn, and Youngseok Choi. 2012. Helpfulness of online consumer reviews: Readers' objectives and review cues. *International Journal of Electronic Commerce*, 17(2):99–126.

A. Ghose and P. G. Ipeirotis. 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.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 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.

Marco Lippi and Paolo Torroni. 2016a. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.

Marco Lippi and Paolo Torroni. 2016b. MARGOT: A web server for argumentation mining. *Expert Syst. Appl.*, 65:292–303.

Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using argument-based features to predict and analyse review helpfulness. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1358–1363. Association for Computational Linguistics.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In Johannes Frnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *Machine Learning: ECML 2006*, volume 4212 of *LNCS*, pages 318–329. Springer Berlin Heidelberg.

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

Thomas L Ngo-Ye and Atish P Sinha. 2012. Analyzing online review helpfulness using a regressional relieff-enhanced text mining method. *ACM Transactions on Management Information Systems (TMIS)*, 3(2):10.

Thomas L Ngo-Ye and Atish P Sinha. 2014. The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems*, 61:47–58.

Douglas Walton. 2009. Argumentation theory: A very short introduction. In Guillermo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 1–22. Springer US.