Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Prediction of Helpful Reviews Using Emotions Extraction

Lionel Martin and **Pearl Pu**
Human Computer Interactions Group
School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPFL)
CH–1015, Lausanne, Switzerland
{lionel.martin, pearl.pu}@epfl.ch

Abstract

Reviews keep playing an increasingly important role in the decision process of buying products and booking hotels. However, the large amount of available information can be confusing to users. A more succinct interface, gathering only the most helpful reviews, can reduce information processing time and save effort. To create such an interface in real time, we need reliable prediction algorithms to classify and predict new reviews which have not been voted but are potentially helpful. So far such helpfulness prediction algorithms have benefited from structural aspects, such as the length and readability score. Since emotional words are at the heart of our written communication and are powerful to trigger listeners' attention, we believe that emotional words can serve as important parameters for predicting helpfulness of review text.

Using GALC, a general lexicon of emotional words associated with a model representing 20 different categories, we extracted the emotionality from the review text and applied supervised classification method to derive the emotion-based helpful review prediction. As the second contribution, we propose an evaluation framework comparing three different real-world datasets extracted from the most well-known product review websites. This framework shows that emotion-based methods are outperforming the structure-based approach, by up to 9%.

Introduction

Product reviews tend to have a consequent impact on markets (Duan, Gu, and Whinston 2008). Products, businesses, services; every purchase you plan or decision you have to take can be alleviated by the advices of those that tried in the past the given product (interchangeably denote any item that can be reviewed, including businesses and services). Nowadays, more and more reviews are available on products and often the latest ones are not the most useful to take the decision to purchase this product. For this reason, it would be practical to filter the relevant reviews so as to speed up and improve the decision process.

Moreover, as time passes, review website users increasingly rely on the quality of the reviews and less on their quantity (Park and Lee 2009). Thus, users would still benefit from the information that a product is popular to make his first opinion (e.g., displaying the total number of reviews for this product) but then it gradually gets more relevant to this person to read others' experiences to decide. However, studies showed that even if people tend to read more often the content of the reviews compared to the previous years, they also read fewer reviews in average. For this reason, it is really important for retailers to provide them the "best" reviews about the product to avoid information overload (Malhotra 1984). Best in this context doesn't mean the most eulogistic ones; it rather concerns the most helpful ones present in the social media.

Review comments have been largely studied during the last ten years because of the different sociological aspects related to them and the ease of access. For example, Amazon[1] counted millions of unique visitors and thousands of reviews posted in 2013. Review websites form interesting corpora for information retrieval and have been studied as such for sentiment analysis – the detection of positive or negative comments. In this sense, predictions of positive and negative reviews, objective and subjective texts have been widely proposed to help people make their choice.

People studied the process that make someone's vote helpful or unhelpful for some time and tried to recognize the interesting comments. Indeed many reviews contain personal critics or a simple transcript of the product's specifications without any helpful information. Those reviews don't require to be read and a large amount of time could be saved by focusing only on the reviews that present an interesting point of view, describing an experience (Otterbacher 2009; Siersdorfer et al. 2010). However, we considered that previous work labeled helpful a too large portion of the reviews in their training sets. Their definition would keep the consumers in a situation of information overload, which is not desired. Thus, we voluntarily selected a much smaller quantity of reviews as helpful in this work. Consequently, we also had to investigate existing rarity mining techniques to compensate this effect to avoid biased a priori classification due to unbalanced testing sets.

Several work previously explored the importance of the content of the reviews to characterize helpfulness such as

[1] www.amazon.com

(Cao, Duan, and Gan 2011; Mudambi and Schuff 2010; Willemsen et al. 2011). However, the impact of the content of the reviews can be studied under different perspectives. Unlike most of the work done previously that focused on statistics computed from the texts, we decided to perform an emotion analysis that we found to be more intuitive to explain helpfulness. Indeed, emotions are powerful tools for communication as they are most likely to evoke the feelings of others and engage their responses. They also drive people's action and regulate their decision process. On top of that, recent research presented the study of emotions as an interesting way to classify documents (Mohammad 2011; Martin, Sintsova, and Pu 2014). In the same way that subjectivity has been addressed to predict helpfulness (Ghose and Ipeirotis 2011), we extract the emotionality contained in the reviews in order to determine whether a review is helpful. The extraction of emotional features and the development of machine learning algorithms are fields that we don't want to contribute to with this paper. However, we will use some of the techniques that belong to the respective states-of-the-art.

In this paper, we compare our results with three different datasets obtained from various review websites, which face different helpfulness measurements and attendances. Then, two media will probably identify helpful reviews differently. A simple example to illustrate this notion lies in the design difference between those social media. On Amazon, people can say that a review is either helpful or not, while they can only acknowledge the helpfulness on Yelp[2] or Trip Advisor.[3] Thus, we can wonder if ordering the comments in decreasing order of helpfulness and cutting after a certain ratio would allow compatibility of the method and how we should select this threshold. Additionally, different definitions of helpfulness on different datasets might impact the comparison of the prediction techniques.

Ultimately, our goal is to answer the following questions with our study:

1. Can the emotionality of reviews be an accurate predictor of helpfulness in product reviews?
2. Are our predictive models resistant to the differences contained in various review websites?

Related Work

The classification of emotions has been an active area of research for more than 30 years. Those emotions have been separated into different families since the beginning. Throughout the time and researchers, different models of classification (i.e., the emotion models) have been introduced to serve different theories.

(Plutchik 1980) associated emotions by pairs: Joy and Sadness, Trust and Disgust, Fear and Anger, Surprise and Anticipation. He also introduced two additional ideas with these families. First, they could be expressed at different intensity levels defining three layers: soft, medium, and strong (e.g., Annoyance, Anger and Rage) and moreover they could be mixed together (e.g, Joy and Trust combine to become Love). Nonetheless, retrieving only these 8 families of emotions is not precise enough to attempt to characterize entirely the emotions in our datasets of product reviews. Moreover, even though Plutchik pairs his emotions with opposite feelings, some of the pairs are composed with two negative emotions breaking the leverage (e.g., Fear and Anger).

(Ekman 1992) on his side defined 6 and then 7 emotions (i.e., Happiness, Sadness, Surprise, Fear, Disgust, Anger, and later Contempt) as the basis to any feeling. More complex emotions can then be expressed as a combination picked from this basis. However, a single emotion represent positive feelings, which was not enough for our analysis.

More recently, (Scherer 2005) proposed a solution that patches the issues discussed so far with 20 different categories, 10 of which are positive emotions and the other 10 are negative ones. These emotions are arranged on a wheel separated in 4 quadrants describing the valence (positive or negative emotions on the horizontal axis) and control (low-control or high-control on the vertical axis) for each of them. This classification globally best fits our objectives of characterizing the influence via the recognition of emotions.

Extraction of writers' impressions raised a lot of attention (Pang, Lee, and Vaithyanathan 2002; Turney 2002) first to decide if a review is whether positive or negative and later to understand the emotional state of the writer. To the best of our knowledge, this understanding of emotionality of the writer has never been used to predict the helpfulness of the text he wrote. For this very reason, we believe that our contribution might improve forward the detection of helpful reviews.

The study of reviewers' *helpfulness scores* has been addressed in several papers aimed at improving the visualization of comments (Danescu-Niculescu-Mizil, Lee, and Kleinberg 2009; Ghose and Ipeirotis 2011; Lu et al. 2010; O'Mahony and Smyth 2009). These works characterized the helpful reviews using different sets of features and predicted the future score of newly created comments to allow a classification that is not dependent of the time elapsed since the review was written.

(Kim et al. 2006) identified the helpfulness to be affected mainly by three factors: the length of the review, the product rating and the TF-IDF score of words used. Later (O'Mahony and Smyth 2010) presented the gain introduced by the computation of readability measures on classification. These two works show the impact of the content in the characterization of influence even though the authors focused on statistics computed on the content and not the actual meaning of the words.

Few works combined the study of structure and content to understand the influence of the users. (Liu et al. 2008) described and predicted the helpfulness of the reviews based on three attributes. These attributes have the particularity to belong to three different classes. The first one regards content and compute statistics on the review to determine the writing style of the reviewer. The second one is structural and considers the time elapsed since the review was written. Finally, the third one is a combination of the two first kinds: it represents the expertise of a user for a given class of items and links the structure and content. The idea is to classify the

[2] www.yelp.com

[3] www.tripadvisor.com

	Amazon	TA	Yelp
Reviews	303,937	68,049	229,906
Products	7,507	216	11,537
Users	233,710	59,067	45,981
Helpfulness Ratings	2,399,158	94,860	318,823
Avg. # Reviews/Product	40.49	315.04	19.93
Avg. Product Rating	3.92	3.83	3.77
Avg. Helpfulness Rating	54.48%	1.39	1.39
Avg. Length (# words)	143.83	206.55	133.00

Table 1: Statistics of the three datasets

newly reviewed item and to associate it with a class of items. Then, it checks if the reviewers is known to be influential for this class of items or not.

Helpfulness Prediction

Datasets

We based our research on three datasets extracted from different product review websites. A first dataset of 68,049 reviews has been crawled from Trip Advisor, which is one of the largest social media platforms for hotel reviews. Those reviews are anonymized versions of comments posted on the website between 2008 and 2011 in 215 different hotels in Las Vegas. A second one contains reviews from Yelp, a social platform of restaurant reviews mostly. The dataset is formed of 229,908 reviews extracted from the website for the 2013 Recsys Challenge.[4] Finally our last dataset contains reviews from Amazon, the world's largest online retailer. 303,937 reviews about 7,507 products are compared in order to see if what can be applied to business reviews also stands for product reviews.

In all the datasets, a review contains at least: (a) an overall rating for the item giving information about the positivity of the comment; (b) a review text (plus a review title for 2 of them) from which we extracted the emotionality of the review; and (c) the helpful score providing an estimate of the influence of the comment. Amazon has the particularity among the three dataset to allow visitors to vote on a review either positively (saying that the review is helpful) or negatively (saying that the review is unhelpful). We use thus the proportion of helpfulness votes among the total number of votes $\frac{r_+}{r_+ + r_-}$ for this dataset. Table 1 presents the main statistics about our three real-world datasets.

Helpful vs. Unhelpful

While predicting the helpfulness score of a review or its rank can use all available information, deciding if it is helpful or not brings a few challenges in our setup. Indeed, as quickly presented in the introduction, we need to define what helpful reviews are, before training on our supervised datasets. (Ghose and Ipeirotis 2011) compared experts' opinions about review helpfulness to a dataset where proportions of helpful votes were used as the metric (like our Amazon dataset). They concluded that reviews with at least 60% of helpful votes had to be classified as helpful reviews to best match experts' analysis. This definition separates the reviews into two relatively balanced sets. Hence, the number of reviews to consider remains extremely high and with it the overload for the users that try to decide. Moreover, depending of the dataset, the affluence is quite different and the average of helpfulness votes too. Thus, a percentage of top-comments might be more appropriate in this case than a constant threshold to split on since for example the number of reviews with 4 votes on Yelp is more frequent than the same score on Trip Advisor. During the evaluation, we will compare Ghose's choice with the selection of the top 1% reviews that we decided to use to extract the best reviews.[5]

Then, a second challenge comes from the selection of the samples to train on once the threshold is set: the relative rarity issue. Given our attempt to extract the most helpful comments from the rest, we create necessarily a dataset with a large class (of unhelpful reviews) and a small one (of helpful reviews). Thus, classification is biased by this a priori distribution of the samples and we need to compensate this effect. This problem is surveyed in (Weiss 2004) where the author proposed a handful number of solutions depending on the specificity of the rarity. Our first approach was to undersample the largest class to obtain a balanced dataset but we later compared this with the SMOTE algorithm (Chawla et al. 2002) that generates new samples in the neighborhood of the original samples of the minority class. In the evaluation, we will see the impact of the two chosen methods on classification and the correlation of this choice with that of the selected machine learning algorithm.

Prediction Models

In this paper, we try to estimate the number of helpfulness votes based on contextual informations about the review texts to determine the impact of emotional content on their influence. We want to test our emotion features extraction on three similar problems. We check first if the presence of emotions in text is enough to decide if a given review is part of the best ones or not, but also in a second step if it allows accurate estimation of the number of users that found this particular review helpful. Finally we also want to estimate the efficiency in ranking reviews based on this score.

The applications associated with the three approaches are diverse. Obviously, as explained in the introduction, the main application is to help readers selecting products by improved visualization of the reviews. In this case, we need an accurate selection of the best reviews, or a good ranking procedure. Regression could, on another side, be interesting as a measure of the impact of a review text. This knowledge would be helpful to authors of reviews if they could see in real-time the estimated impact of their draft comment.

To achieve these goals, we compare the benefit of emotions with three different well-known models: Support Vector Machine, Random Forest and Naive Bayes. The latter one is only used for the classification problem here. The selection of these algorithms has been motivated mostly by the

[4] www.kaggle.com/c/yelp-recsys-2013

[5] This choice was directly motivated by our willingness to filter a quantity of reviews that one can read with real-world applications.

frequency at which they appear to outperform their competitors in the state-of-the-art. Even though Naive Bayes is usually a bit less accurate, its speed of execution makes it sometimes a good candidate when the difference of efficiency is negligible or if the number of features becomes very large.

First, we tune the parameters of the algorithms such as the kernel function (radial basis functions are performing the best), the penalty parameter C, and the kernel coefficient γ for SVM following (Hsu, Chang, and Lin 2003)'s recommendation for grid-searching over 110 combinations.

Then, we evaluate the methods with different metrics depending on the problem: regression is evaluated with mean squared error, ranking with the Spearman's rank correlation (Spearman 1904), and classification is measured with three metrics that often appear in the literature: accuracy, F1-score, and AUC. In all cases the evaluation is performed over a 10-fold stratified cross-validation: the dataset is separated into 10 chunks of equal size and class distribution, and 10 rounds are performed where each time a tenth of the dataset constitute the testing set while the rest is the training set; the outputted results are the averages of the ten runs.

Features

We compared our method against different state-of-the-art techniques of helpfulness prediction as well as common statistics about the review texts. For this reason we extracted different features from each of our three datasets that we will describe now.

Emotionality of reviews. To test our hypothesis that emotions are a good predictor of the helpfulness votes in review websites, we extracted the words that were associated with particular emotions. We selected the Scherer's emotion model (called Geneva Emotion Wheel) with its 20 emotion categories for our work and we used the GALC lexicon (Scherer 2005) to determine which words were conveying which emotion. For each category, we counted the number of words present in the review that belong to the lexicon. We improved this extraction with negation treatment to take into account the sentences flipped by negation and avoid mislabeling of the words in such neighborhood. We defined a first feature vector with the number of occurrences of each emotion and called it **GEW20**. To study the impact of the negation treatment, we also constructed a feature set called **GEW_NO_NEG** that is not applying the additional computations on negated sentences. Moreover, we also constructed a set with 21 features containing additionally a counter for non-emotional words, **GEW21**, to accurately fit GEW model.

Sentiment analysis. Highly studied aspect of the reviews posted online, the overall sentiment expressed in the texts can be mined thanks to lexicons that label words either positively or negatively. These lexicons map words to confidence scores depending whether the word was mostly present in one type of reviews in the training set or almost equally among the two (positive and negative reviews are the two types). We selected a lexicon that was constructed on hotel reviews and created a vector of two features counting the positive and the negative occurrences separately, **SENTI**. Regarding sentiment, we also retrieved the global rating that users gave to the product and called in **RATE**.

Part of Speech tagging. Several papers used the syntactic characterization of words and their proportion in prediction models (Lu et al. 2010; Kim et al. 2006; Liu et al. 2008). They usually counted the number of nouns, verbs, and adjectives contained in the text and constructed a feature set with the results. We extended this feature set to all the tags we encountered and called it **POS**.

Text statistics. In order to compare our method with the state-of-the-art, we extracted different features presented by two papers. On the one hand, (Kim et al. 2006) established that taking into account **LEN1**, **RATE** and **TF-IDF** scores was the best combination for the regression of helpfulness scores. On the other hand, (O'Mahony and Smyth 2010) focused on readability measures, including the Flesch Reading Ease (**FLES**), which is a combination of the number of words (**LEN1**) and the number of sentences (**LEN2**) . Moreover, **FLES** would give an intuitive explanation to the helpfulness: reviews that are easy to read tend to be more influential.

Evaluation

In this section we compare the effect of different features presented above on the prediction of top review comments. Tables 2 and 3 summarizes the result of the 10-fold cross validation for all datasets and features. By lack of space, the plots are presenting subsets of our evaluation and discussions focus around these sub-problems. However, all results can be generalized to the other datasets and to both sampling techniques unless specified differently.

First, we see that our emotion features perform quite well among the different feature sets we tested. Negation treatment on 20 emotions slightly improves the results in some setups while it gives similar results in the others. Taking into consideration the 21^{st} component counting non-emotional words contained in the review significantly improves the learning especially when combined with SVM (up to 16.5% with under-sampling, see figure 1).

Following the results of the different features on Trip Advisor presented in figure 1, we noticed an important impact of Part of Speech tagging for helpfulness classification. Indeed the precision-recall curve covers more than 95% of the cases with a Random Forest algorithm trained on POS. This class of features moreover outclasses all the features we tested on Yelp and Amazon (see tab. 2) but it does not outperform our 21 emotion features with SVM on Trip Advisor.

In comparison with (Kim et al. 2006)'s work, GEW21 improves by 12% the measure of AUC on SVM (the algorithm that the authors used in their experiment) but on the two other methods their features achieve an equivalent AUC score than GEW21.

On another hand, rating is not really an appropriate predictor of helpfulness. Indeed, compared to random guessing, the improvement of AUC in our Trip Advisor dataset is only of 10% in the best setup (taking SVM with SMOTE

	Amazon			Trip Advisor			Yelp		
	A	F1	AUC	A	F1	AUC	A	F1	AUC
RATE	0.54	0.57	0.54	0.57	0.62	0.61	0.50	0.60	0.52
LEN1	0.59	0.62	0.61	0.73	0.71	0.79	0.73	0.75	0.80
LEN2	0.58	0.60	0.60	0.71	0.70	0.78	0.72	0.73	0.80
POS	**0.68**	**0.68**	**0.80**	0.72	0.73	0.79	**0.84**	**0.87**	**0.85**
SENTI	0.56	0.54	0.57	0.70	0.69	0.76	0.63	0.63	0.67
FLES	0.52	0.61	0.53	0.53	0.66	0.54	0.53	0.61	0.53
GEWN	0.57	0.56	0.58	0.69	0.70	0.73	0.65	0.68	0.70
GEW20	0.56	0.58	0.57	0.71	0.68	0.70	0.65	0.68	0.69
GEW21	0.63	0.63	0.62	**0.74**	**0.73**	**0.82**	0.73	0.75	0.80

Table 2: Evaluation of the SVM algorithm on the features extracted over three datasets sampled with SMOTE using Accuracy (A), F1-score (F1) and Area Under the Curve (AUC) as measures. GEWN refers to GEW_NO_NEG.

	Amazon			Trip Advisor			Yelp		
	A	F1	AUC	A	F1	AUC	A	F1	AUC
RATE	0.54	0.64	0.55	0.57	0.59	0.61	0.53	0.60	0.53
LEN1	0.62	0.57	0.66	0.80	0.78	0.85	0.75	0.75	0.83
LEN2	0.58	0.60	0.61	0.71	0.70	0.78	0.72	0.73	0.79
POS	**0.88**	**0.86**	**0.93**	**0.89**	**0.89**	**0.96**	0.78	**0.87**	0.85
SENTI	0.67	0.65	0.73	0.74	0.75	0.82	0.72	0.72	0.78
FLES	0.65	0.64	0.70	0.62	0.62	0.67	0.65	0.64	0.70
GEWN	0.60	0.52	0.65	0.76	0.72	0.83	0.71	0.67	0.77
GEW20	0.61	0.51	0.65	0.77	0.74	0.84	0.71	0.67	0.77
GEW21	0.71	0.70	0.78	0.79	0.79	0.87	**0.81**	0.81	**0.89**

Table 3: Evaluation of the RF algorithm on the features extracted over three datasets sampled with SMOTE using Accuracy (A), F1-score (F1) and Area Under the Curve (AUC) as measures. GEWN refers to GEW_NO_NEG.

sampling). This result was predictable since both positive and negative reviews can be helpful. Whereas we expected sentiment extraction to follow the same logic since they are closely related in polarity prediction, it appears to be a relatively good characteristic to consider due to its lexicon by exceeding 80%.

Conclusively, GEW21 outperforms all the statistics that we compared that are based on the texts but do not take the meaning of the words into account (up to 9% against LEN1, the best structural feature), such as the number of words, the readability measure or the rating. However, we discovered that the number of words belonging to the classes of Part of Speech completely overpassed our expectations with an AUC score of 95% using Random Forest and SMOTE.

Discussion

Emotionality of review texts performs well for the classification task on the datasets that we presented. We have seen that our attempt to correct the emotion symbolized by negative sentences was an interesting idea that could probably be improved further since all prediction tasks did not benefit uniformly of this processing step and some datasets even gained nothing from it. However, this result is encouraging to interpret the characteristics of influence intuitively. Indeed, as it was the case with (O'Mahony and Smyth 2010)'s work on prediction using readability, the prediction using emotions has the advantage to convey a message: reviewers expressing their thoughts about a product emotionally are more helpful than those that are less involved. This intuitive characterization can also benefit to people writing a review before they publish it. Indeed, we can imagine a tool that would be able to prevent the review authors that they could improve their comments making precise changes.

Furthermore, our evaluation gave us the idea to later construct a lexicon to discover directly helpfulness. Indeed, as one constructed sentiment lexicons or emotion lexicons, the same thing could be done with helpfulness directly to improve prediction. Nonetheless, this will not help us interpret the main components of the construction of a helpful comment intuitively as discussed previously.

We use the emotion features to present the results on the different datasets (fig. 2). First, considering under-sampling, we always observe that SVM classifies the samples the best, followed by Random Forest and finally Naive Bayes. However, we remark here again that Random Forest benefits strongly from the SMOTE technique for sampling. We also used this opportunity to compare our threshold for classification with that of (Ghose and Ipeirotis 2011), used recently by several authors. Even though measured scores are better in this setup in comparison with ours on Amazon, the differences between the techniques remain identical to our previous study. Thus, depending on the needs an alternate splitting strategy on the comments is acceptable and our analysis holds for the selection of features and method. The results discussed here and showed on the figure also hold for F1-score and accuracy measures.

Regression and ranking on their side benefit from a similar analysis. Once again, the datasets can all be predicted similarly. The main difference in this case lies in the difference of attendances and thus the scale of the helpfulness scores of the different reviews. Indeed since helpfulness in Trip Advisor reviews range between 0 and 37 while Yelp reviews can reach 120, the mean squared error scales differently. Except from this aspect, features performing well in classification usually also perform better in regression than the rest. For example, our evaluation shows that Part of Speech and emotion extraction using GALC are the features that minimize best MSE.

We further studied the regression of helpfulness scores with ranking. After that regression has been estimated, we ranked these results in decreasing order and compared to the effective ranking of the review in the test set. Spearman correlation was computed and Part of Speech performed best. Since regression does not require balanced datasets, we used all available data. The low correlation values compared to previous work that used this measure are probably explained by the difference in the size of the datasets. In this setup as well, we noticed a beneficial effect of negation treatment on emotion extraction and also of the 21^{st} emotion category. The results of the test using SVM are depicted in figure 3.

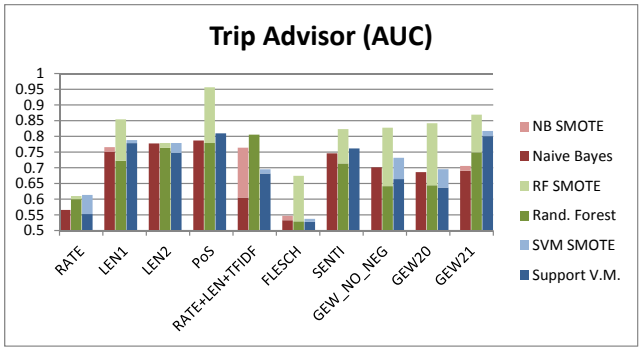

Figure 1: Effect of sampling on classification with our Trip Advisor dataset.

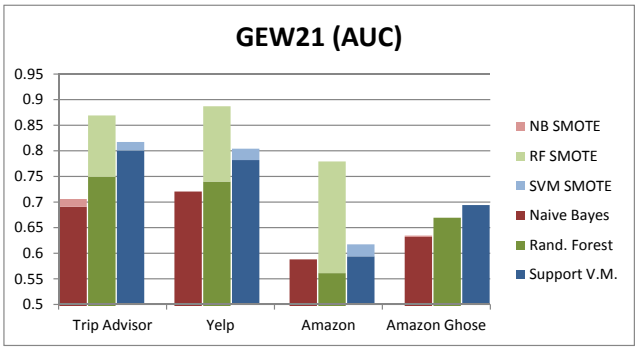

Figure 2: Evaluation of different classification algorithms for helpfulness prediction with emotion features.

Conclusion

We develop in this paper a method to predict reviews' helpfulness using primarily emotionality. Since emotions are good to trigger reactions, we were suspecting it to help review readers putting themselves in the writer's place. We thus extracted from the review texts the words that convey emotions to the readers by the mean of the GALC lexicon. In our benchmark, comparing this with state-of-the-art and structural features, we notice an improvement up to 9%. The combination of emotion detection with the other features is left as a future work as well as the robustness in transfer learning (training on one dataset and testing on a different one). We also define a stricter criterion for review's helpfulness than the one used in recent works about helpfulness classification. Indeed, we focused on the massive overload due to the display of too many reviews to the users and wanted to keep only few texts that can suffice to users willing to take their decision. We quickly present results about regression and ranking still using the same feature sets. We acknowledge the variance that different measures and datasets bring into the evaluation and we emphasize on the fact that Part of Speech tagging also performed really well in some situations.

We discuss on the results of different datasets and point out the similar behavior of all features presented in the evaluation including ours. Sampling techniques, machine learning algorithms, and the differences between the evaluation measures overall follow a similar pattern between our product and service review datasets.

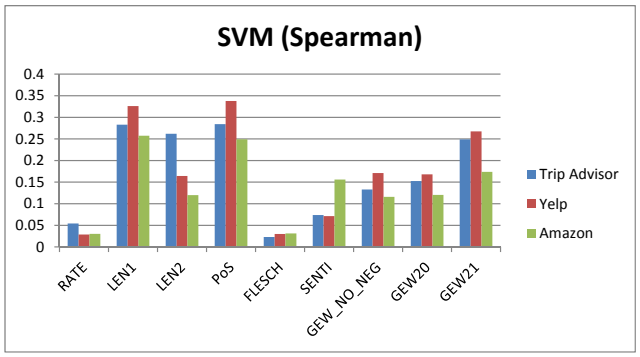

Figure 3: Evaluation of ranking over all datasets using Spearman's ranking correlation.

Additionally, we use our evaluation to compare the effect of different techniques both for the classification and the selection of the samples to train on. We discover an important improvement using SMOTE algorithm on Random Forests for the majority of the feature sets on all our datasets compared with under-sampling and a smaller significant improvement on Support Vector Machine. As requested, this technique leverages training sets in a way that improve the efficiency of classification. Applying it on an almost balanced set of Amazon product review using Ghose's threshold for classification (with 60% of positive votes) did not produce any enhancement compared with under-sampling though, as expected since no rarity occurs.

In future work, we will try to improve further the treatment of negation. We have seen that our first approach was already giving positive results in some classification problems but this is probably not the best that one can do. There is also a need to compare different lexicons to ensure if we could enhance prediction with emotions further. However, this requires a deep evaluation of GALC and its alternatives. Another approach would be to construct a lexicon of words triggering the sentiment of helpfulness as shortly discussed in the evaluation. Such lexicon could be used in real-time to improve users' comment by giving them their helpfulness score.

Acknowledgments

We would like to thank the anonymous reviewers of this paper for their comments. The authors are also grateful to the Swiss National Science Foundation for their support.

References

Cao, Q.; Duan, W.; and Gan, Q. 2011. Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach. *Decision Support Systems* 50(2):511–521.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling tech-

nique. *Journal of Artificial Intelligence Research* 16(1):321–357.

Danescu-Niculescu-Mizil, C.; Lee, L.; and Kleinberg, J. 2009. How Opinions are Received by Online Communities : A Case Study on Amazon.com Helpfulness Votes. In *Proc. International conference on World wide web WWW'09*, 141–150. ACM.

Duan, W.; Gu, B.; and Whinston, A. B. 2008. The dynamics of online word-of-mouth and product salesan empirical investigation of the movie industry. *Journal of Retailing* 84(2):233–242.

Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion* 6(3-4):169–200.

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.

Hsu, C.-W.; Chang, C.-C.; and Lin, C.-J. 2003. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Technology, National Taiwan University.

Kim, S.-M.; Pantel, P.; Chklovski, T.; and Pennacchiotti, M. 2006. Automatically assessing review helpfulness. In *Proc. Conference on Empirical Methods in Natural Language Processing*, 423–430. ACL.

Liu, Y.; Huang, X.; An, A.; and Yu, X. 2008. Modeling and predicting the helpfulness of online reviews. In *IEEE International Conference on Data Mining*, 443–452. IEEE.

Lu, Y.; Tsaparas, P.; Ntoulas, A.; and Polanyi, L. 2010. Exploiting social context for review quality prediction. In *Proc. International conference on World Wide Web*, 691–700. ACM.

Malhotra, N. K. 1984. Reflections on the information overload paradigm in consumer decision making. *The Journal of Consumer Research* 10(4):436–440.

Martin, L.; Sintsova, V.; and Pu, P. 2014. Are influential writers more objective?: An analysis of emotionality in review comments. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, 799–804.

Mohammad, S. M. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proc. Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 105–114. ACL.

Mudambi, S. M., and Schuff, D. 2010. What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly* 34(1):185–200.

O'Mahony, M. P., and Smyth, B. 2009. Learning to recommend helpful hotel reviews. In *Proc. conference on Recommender systems, RecSys'09*, 305–308. ACM.

O'Mahony, M. P., and Smyth, B. 2010. Using readability tests to predict helpful product reviews. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, 164–167. Centre De Hautes Etudes Internationales D'informatique Documentaire (CID).

Otterbacher, J. 2009. 'helpfulness' in online communities: a measure of message quality. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 955–964. ACM.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proc. conference on Empirical methods in natural language processing*, 79–86. ACL.

Park, D.-H., and Lee, J. 2009. ewom overload and its effect on consumer behavioral intention depending on consumer involvement. *Electronic Commerce Research and Applications* 7(4):386–398.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience* 1(3):3–33.

Scherer, K. R. 2005. What are emotions? and how can they be measured? *Social science information* 44(4):695–729.

Siersdorfer, S.; Chelaru, S.; Nejdl, W.; and San Pedro, J. 2010. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proc. international conference on World wide web*, 891–900. ACM.

Spearman, C. 1904. The proof and measurement of association between two things. *The American journal of psychology* 15(1):72–101.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. Annual meeting on association for computational linguistics*, 417–424. ACL.

Weiss, G. M. 2004. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter* 6(1):7–19.

Willemsen, L. M.; Neijens, P. C.; Bronner, F.; and de Ridder, J. A. 2011. highly recommended! the content characteristics and perceived usefulness of online consumer reviews. *Journal of Computer-Mediated Communication* 17(1):19–38.