Learning from Positive and Unlabeled Amazon Reviews: Towards Identifying Trustworthy Reviewers

Marios Kokkodis
mkokkodi@stern.nyu.edu
Department of Information, Operations, and Management Sciences
Leonard N. Stern School of Business, New York University
New York, New York 10012, USA

ABSTRACT

On-line marketplaces have been growing in importance over the last few years. In such environments, reviews consist the main reputation mechanism for the available products. Hence, presenting high quality reviews is crucial in achieving a high level of customer satisfaction. Towards this direction, in this work, we introduce a new dimension of review quality, the reviewer's "trustfulness". We assume that voluntary information provided by Amazon reviewers, regarding whether they are the actual buyers of the product, signals the reliability of a review. Based on this information, we characterize a reviewer as trustworthy (positive instance) or of unknown "trustfulness" (unlabeled instance). Then, we build models that exploit reviewers' profile information and on-line behavior to rank them according to the probability of being trustworthy. Our results are very promising, since they provide evidence that our predictive models separate positive from unlabeled instances with very high accuracies.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous

Keywords

Reviewer Trustfulness, Positive-only Learning

1. INTRODUCTION

Over the last few years, on-line marketplaces have been growing in importance; for example Amazon generated $48 billions in 2011[1]. In such environments, reviews consist the main reputation mechanism for the available products. As a result, identifying good quality reviews is crucial, since it can significantly enhance buyer's experience and increase his overall satisfaction. Towards this direction, most of the existing research explores factors that affect (and are able to predict) the *helpfulness* of a review. These factors include textual features of reviews as well as reviewer-based characteristics [1–4]. However *helpfulness* is not a perfect measure of review quality, since it has been proved to be biased by the posting date of the review and the assigned product rating (the same review can receive widely different *helpfulness* ratings depending on factors unrelated to the actual quality of the review) [5]. As a result, we need additional dimensions that could possibly help towards identifying high quality reviews.

In 2009, Amazon added a new feature to its product reviews, named "Amazon Verified Purchase" (AVP). This feature indicates whether or not the reviewer bought the reviewed product from Amazon. The feature does not automatically apply to all reviews, unless the reviewer explicitly opts to disclose this information. We can be reasonably certain that a review with the "verified purchase" badge is, ceteris paribus, more reliable than one without. Similarly, a reviewer with many AVP reviews is expected to be trustworthy, and therefore a reliable source of information. The key question that we try to address here is how can we use this information about such reliable reviewers (AVP-reviewers) to identify more similar reviewers that are also hopefully reliable.

In this question, the challenge is the lack of negative examples: while we do know a few AVP-reviewers, we do not have any information about the rest, since the absence of the AVP badge does not imply that the review is unreliable: the reviewer may have missed opting in to the AVP program, or bought the product from a different marketplace. We attack the problem as follows: we assume that trustworthy reviewers have the majority of their reviews AVP-labeled; we consider such reviewers as **positive** examples. In addition, we assume that the "trustfulness" of the rest of the reviewers is unknown; we consider these reviewers as **unlabeled** examples. Then we use profile and on-line behavior characteristics to build models that rank reviewers according to their chance of being trustful. We evaluate our proposed approaches on a large corpus of AVP and non-AVP Amazon Book reviews. Our results are very promising, since they provide evidence that our rankers separate positive from unlabeled instances with very high accuracies.

2. PROBLEM FORMULATION

In this work, we define "trustfulness" as the probability of a reviewer being trustworthy. We assume that "trustfulness" can be inferred by the reviewer's profile characteristics and past on-line behavior in the marketplace. For a reviewer j, these characteristics include the reviewer's Amazon rank (r_j), the reviewer's total number of posted reviews (n_j), and three user-derived features: the number of reviewer's Amazon badges[2] (b_j), the amount of personal information revealed on reviewer's profile (i_j), and the reviewer's average posting frequency (f_j). Formally, the last three features are defined as follows:

$$b_j = \sum_{k=1}^{|B|} b_{k,j}, \quad i_j = \sum_{k=1}^{|I|} c_{k,j}, \quad f_j = \frac{\sum_{k=2}^{n_j} d_j(k,k-1)}{n_j - 1},$$

[1]http://en.wikipedia.org/wiki/Amazon.com

[2]Information about badges can be found on amazon.com.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

where $b_{k,j}$ is binary and represents the existence or not of each one of the badges in the set $B = \{$"Real Name", "The", "Amazon Official", "Author", "Artist", "Manufacturer", "Vine Voice", "Holiday Team", "Community Forum" $\}$, $c_{k,j}$ is also binary, and represents the existence or not of information on reviewer's profile characteristics in the set $I = \{$Website, Email, Location, Description, Interests, Used Tags, Note, Images$\}$, and $d_j(k, k-1)$ is the number of days that elapse between the $(k-1)^{\text{th}}$ and k^{th} review of reviewer j.

Under the reasonable assumption that actual buyers are trustworthy reviewers, we can use AVP to assign positive labels. In particular, a reviewer is considered a "positive" example *iff* the majority of the reviews following his first AVP labeled review are also AVP. Otherwise, the reviewer is considered as an "unlabeled" instance.

Finally, given that $\mathbf{x}_j = (r_j, n_j, b_j, i_j, f_j)$ is the input vector of reviewer features, we can build probabilistic ranking models that learn to predict the reviewers' chance of being "positive". In other words, these models can estimate the $\Pr(\text{Positive}|\mathbf{x}_j)$. This probability represents the reviewer's estimated "trustfulness".

3. EXPERIMENTAL SETUP

Our goal is to build models that rank reviewers according to their likelihood of being trustful, following the definition in Section 2. For that purpose, we can use probabilistic classifiers such as Naive Bayes, Decision Trees and Logistic Regression. Based on the previous work by Elkan and Noto [6], such rankings have been proved to classify instances according to their actual chance of being positive, under the valid assumption that these instances are randomly chosen.

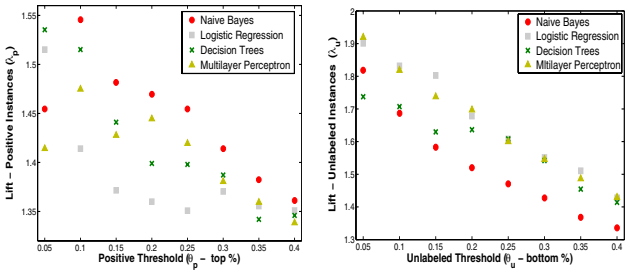
Evaluation Metrics: For our setting, accuracy might be misleading; the notion of accurately classifying an instance as "unlabeled" is unclear. A better metric to use is the **lift** (λ), because it can illustrate the relative accuracy increase in the top and bottom segments of our models' rankings (see also [7]). In particular, for our evaluation, we define two different lifts: one for the positive labeled instances (top segment, λ_p) and one for the unlabeled instances (bottom segment, λ_u). More formally:

$$\lambda_p = \frac{Pr\{\text{Positive}|\theta_p\}}{Pr\{\text{Positive}\}}, \qquad \lambda_u = \frac{Pr\{\text{Unlabeled}|\theta_u\}}{Pr\{\text{Unlabeled}\}},$$

where $Pr\{\text{Positive}|\theta\}$ is the confidence that a positive (unlabeled) instance is positioned in the top (bottom) θ percentage of a model's predicted ranking (i.e., the probability of selecting at random a positive (unlabeled) reviewer given that he is ranked in the top (bottom) θ percentage). Note that $\theta \in (0, 1)$ and that λ is proportional to the confidence of a positive (unlabeled) prediction.

Data: For our experiments, we use 321,166 real Amazon *book* reviews posted by 8,207 reviewers. The vast majority of these reviews ($\sim$ 95%) has been uploaded on Amazon between 2007 and 2011. To avoid overfitting, in all of our experiments we perform a 10-fold cross validation. Finally, we conduct our experiments on balanced data sets, where the probability of randomly drawing a positive reviewer is 0.5.

Results: We build and test four different models based on Naive Bayes, Logistic Regression, Decision trees and Multilayer Perceptron (MLP). In figures 1(a) and 1(b) we present the lift for the positive and unlabeled instances respectively, for each one of our four models and for different values of



(a) (b)

Figure 1: Lift for Positive (a) and Unlabeled (b) instances, for different θ_p and θ_u values.

$\theta_p, \theta_u \in \{0.05, 0.10, \ldots, 0.40\}$. In our balanced dataset, the lift values are constrained, $0 \leq \lambda_p, \lambda_u \leq 2$.

Our classifiers learn to predict adequately positive labeled reviewers in the top ranked instances. In particular, for specific θ_p values, the lift rises up to 1.55. Naive Bayes seems to outperform other rankers in accurately ranking the positive labeled instances. Decision Trees and MLP provide good instance positioning as well.

For the unlabeled instances, all of our rankers provide very accurate results. In particular, Multilayer Perceptron predicts an almost positive-empty set of instances in the bottom 5% of its ranking ($\lambda_u = 1.92, \theta_u = 0.05$). The rest of the classifiers also provide very high lift values.

4. DISCUSSION

Our results indicate that "trustfulness" can be a differentiating factor among reviewers. The proposed approaches in this work create a very-low density unlabeled set as well as an almost positive-free set for low values of θ_p and θ_u respectively. These results are very promising.

In the future, we intent to study the definition of "trustfulness" in more depth; we plan to include more features that we believe "trustfulness" is correlated with. In particular, dimensions that could be informative are the average opinion (positive or negative) of the reviewer, whether the reviewer reviewed complementary products and whether the review contains information exposed only to real buyers. Finally, we further intend to apply the procedure proposed by Liu et. al. [8], which can potentially provide better rankings.

5. REFERENCES

[1] Anindya Ghose and Panagiotis G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *TKDE*, 23(10), 2011.

[2] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. Exploiting social context for review quality prediction. In *WWW*, 2010.

[3] Christian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *WWW*, 2009.

[4] Soo-Min Kim, Patrick Pantel, Timothy Chklovski, and Marco Pennacchiotti. Automatically assessing review helpfulness. In *EMNLP*, 2006.

[5] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. Low-quality product review detection in opinion summarization. In *EMNLP*, 2007.

[6] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *KDD*, 2008.

[7] Foster Provost, Brian Dalessandro, Rod Hook, Xiaohan Zhang, and Alan Murray. Audience selection for on-line brand advertising: Privacy-friendly social network targeting. In *KDD*, 2009.

[8] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. Building text classifiers using positive and unlaebeled examples. In *ICDM*, 2003.