



Provided by the author(s) and University College Dublin Library in accordance with publisher policies. Please cite the published version when available.

Title	An assessment of machine learning techniques for review recommendation
Authors(s)	O'Mahony, Michael P.; Cunningham, Pádraig; Smyth, Barry
Publication date	2009-08
Publication information	L. Coyle, J. Freyne (ed.s). Artificial Intelligence and Cognitive Science : 20th Irish Conference, AICS 2009 Dublin, Ireland, August 19-21, 2009 : Revised Selected Papers, LNAI 6206
Conference details	Presentation at the 20th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 09), Dublin, 19th - 21st August 2009
Publisher	Springer
Link to online version	http://dx.doi.org/10.1007/978-3-642-17080-5_26
Item record/more information	http://hdl.handle.net/10197/1656
Publisher's version (DOI)	10.1007/978-3-642-17080-5_26

Downloaded 2020-07-06T16:58:47Z

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)



Some rights reserved. For more information, please see the item record link above.



An Assessment of Machine Learning Techniques for Review Recommendation^{*}

Michael P. O'Mahony[†], Pádraig Cunningham[‡], and Barry Smyth[†]

[†]CLARITY: Centre for Sensor Web Technologies,

[‡]UCD Complex and Adaptive Systems Laboratory,
School of Computer Science and Informatics, University College Dublin
{michael.p.omahony, padraig.cunningham, barry.smyth}@ucd.ie

Abstract. In this paper, we consider a classification-based approach to the recommendation of user-generated product reviews. In particular, we develop review ranking techniques that allow the most *helpful* reviews for a particular product to be recommended, thereby facilitating users to readily assess the quality of the product in question. We apply a supervised machine learning approach to this task and compare the performance achieved by several classification algorithms using a large-scale study based on TripAdvisor hotel reviews. Our findings indicate that our approach is successful in recommending helpful reviews compared to benchmark ranking schemes, and further we highlight an interesting performance asymmetry that is biased in favour of reviews expressing negative sentiment.

1 Introduction

Recommendations are now a familiar feature of many online services. They help us to navigate through complex information spaces from music (iTunes) and movies (NetFlix) to books (Amazon) and consumer electronics (BestBuy). In these scenarios, recommender systems have largely been used to suggest catalog items (songs, movies, books etc.) to users based on their learned interests, which are often derived from their past purchasing behaviour. Lately, as online services embrace the world of the *social web*, *user-generated content* is playing an increasingly important role when it comes to supporting user buying decisions. For example, many online stores now include comprehensive consumer reviews to complement product descriptions, and it is not uncommon for popular products to attract hundreds of reviews from users who are only too happy to share their thoughts and opinions. Indeed many of us use sites like Amazon, Yelp and TripAdvisor primarily for their review information, even when we are planning to make our purchases elsewhere. In the world of recommender systems, these reviews can serve as a form of *recommendation explanation* [1–3] and can play a key role in helping the user to better evaluate the goodness of the product suggestion.

^{*} This work is supported by Science Foundation Ireland under Grant Nos. 07/CE/I1147 and 08/SRC/I1407

The growing volume of these reviews motivates a new recommendation challenge: namely, the recommendation of balanced and helpful reviews. Although reviews are becoming increasingly commonplace, they can vary greatly in their quality and helpfulness. For example, reviews are often biased and sometimes are contributed by self-interested parties, while others can be very balanced and insightful. For this reason, the ability to recommend helpful reviews will add considerable value to the user’s online experience. While some services are addressing this by allowing users to rate the helpfulness of each review, this type of feedback can be sparse and varied, with many reviews, particularly recent ones, failing to attract any feedback.

In this paper, we focus on this review recommendation task and describe a classification-based approach to suggesting helpful reviews. Briefly, we develop a classifier capable of classifying a review as either *helpful* or *unhelpful* with some confidence and go on to show how this information can be used as the basis for ranking reviews by the predicted helpfulness during recommendation. In particular, we extend our recent work on this topic [4] by evaluating a variety of different machine learning approaches, using feature selection, on a large-scale dataset from the TripAdvisor site. We highlight some interesting properties of this task and dataset from a machine learning perspective and demonstrate the considerable benefits of a *random forest* learning technique, over more conventional Bayesian and decision tree techniques, in this particular domain. We go on to show how the resulting recommender system is capable of suggesting superior reviews, compared to a number of alternative recommendation benchmarks, and highlight an interesting recommendation performance asymmetry that is biased in favour of reviews expressing negative sentiment.

2 Review Recommendation as Classification

Reviews can vary in two dimensions – they can be either *helpful* or *unhelpful* from a reader’s perspective and they can express either *positive* or *negative* sentiment in respect of the product or service in question. Importantly, TripAdvisor facilitates feedback on reviews where users can indicate whether or not they found them to be helpful. In order to distinguish the most unambiguously helpful reviews from the rest, we consider a review to be helpful if and only if 75% or more of the raters of that review have found it helpful. Each TripAdvisor review also contains a score (on a 5-point scale) which indicates the user’s sentiment toward the hotel. Following the example of Amazon.com (see Figure 3), we define a positive review as any review in which a score of ≥ 4 -stars has been assigned to the hotel, and a negative review as any review in which a score of < 4 -stars has been assigned. Overall, there are significantly more helpful positive reviews in our evaluation dataset than helpful negative ones, which indicates that users are much less inclined to perceive negative reviews as helpful. We discuss this matter in detail, and its impact on our evaluation, in Section 3.2.

The main objective of this work is to rank a collection of reviews in order to present a user with a small set of positive and negative reviews that will be

considered helpful. In the evaluation presented here, we assess the use of *ranking* classifiers to achieve this objective. Ranking classifiers are classifiers that assign a score rather than a class label so that an unlabeled set of examples can be ranked by degree of belonging to a class.

Thus reviews can be recommended by returning the top-ranking positive and negative reviews to the user. An obvious criterion for assessing the effectiveness of the process is to score the *enrichment* of the reviews presented, i.e. if 30% of reviews selected at random are helpful and 60% of reviews selected by the classifier are helpful, then a 2-fold enrichment is achieved.

2.1 The TripAdvisor Chicago Dataset

The dataset used for the evaluation comprises reviews of Chicago hotels gathered from the TripAdvisor service¹. This dataset is described in detail in [4] and so only summary details are provided here. The dataset contains 17,038 reviews of 7,646 hotels by 7,399 users. All reviews in the dataset have received feedback on review helpfulness on a minimum of 5 occasions. We use this feedback as the ground truth in respect of classification and recommendation performance and reviews are labeled as either helpful or unhelpful as described above.

In gathering the data each review has been represented by 23 different features. These features are divided into four categories [4]:

- *Social network features* – the degree distribution statistics in the bipartite user-hotel graph,
- *Sentiment features* – the overall rating score and (optional) sub-scores for particular hotel features assigned by users to hotels,
- *Content features* – review length, readability and completeness, and
- *User reputation features* – the helpfulness of reviews from users in the past.

It is important to distinguish the reputation features in particular, given that these features will not be available in situations where feedback on review helpfulness is not facilitated or is limited in quantity.

2.2 Classifier Selection and Configuration

In setting up a classification-based review selection system along the lines we propose, there are a number of issues to be considered:

- The amount of training data required for good performance should be determined.
- An optimal subset of the 23 features has to be selected.
- A classifier with good ranking performance on the data has to be identified.

¹ <http://www.tripadvisor.com/>

Training set size: In supervised machine learning terms a training set of over 17,000 examples is very large. In most application domains, less than a thousand examples will provide good coverage of the problem domain and the addition of new examples will not improve generalisation accuracy (i.e. accuracy on unseen examples). For this reason, we performed a cross-validation analysis to see what size of training set would offer adequate coverage. The results of this evaluation are shown in Section 3.1.

Feature Selection: While the 23 features have been chosen to be predictive of review helpfulness, there is no guarantee that some of the features are not irrelevant or redundant in the presence of some of the other features. It may be that a subset of the features are sufficient to provide good performance. Indeed for some classification techniques the presence of irrelevant features may damage performance. We have tested three feature selection techniques for identifying a good feature subset. These are feature ranking based on information gain, ranking based on variable importance derived using random forest [5] and wrapper search [6]. Somewhat surprisingly, feature selection based on information gain proved to be as effective as the others and that was the method we used.

The information gain for a feature is the reduction in entropy achieved by partitioning the data based on that feature. A simple and effective subset selection strategy is to rank the features using information gain and then, starting with the highest scoring feature, evaluate using cross-validation the performance of a classifier built with that feature. Then add the next highest ranking feature and reevaluate; repeat until no further improvements are achieved [6]. The results of this feature selection strategy on the Chicago data are shown in Section 3.1.

Classifier Selection The two criteria for selecting a good classifier for this recommendation task are that it should have good accuracy on the task and that it should be able to rank reviews in terms of expected helpfulness. In practice, probabilities or confidence scores can be produced for most classifier types and Weka², the toolkit used in our evaluation, can assign a ‘probability’ to the predictions of most classifiers it supports. The classifiers we have chosen for detailed analysis are naïve Bayes, the J48 implementation of the C4.5 decision tree algorithm, the JRip implementation of the Ripper rule learner and random forests.

It is straightforward to assign a confidence to a prediction produced by naïve Bayes because the classifier directly calculates a posterior probability on which the classification is based. JRip and J48 produce confidence scores based on the distribution of training instances classified by the rule or leaf node (in the J48 case). For example, if 10 training samples are classified by a single Ripper rule and the distribution of those samples is 7 positive and 3 negative, then that rule will assign a confidence of 0.7 to all its predictions. J48 assigns confidence scores in the same way. Clearly, if there are few rules in the Ripper model or if a decision tree is very simple, then a ranking based on these confidence scores will be very coarse. Happily this is not the case with this recommendation problem –

² <http://www.cs.waikato.ac.nz/ml/weka/>

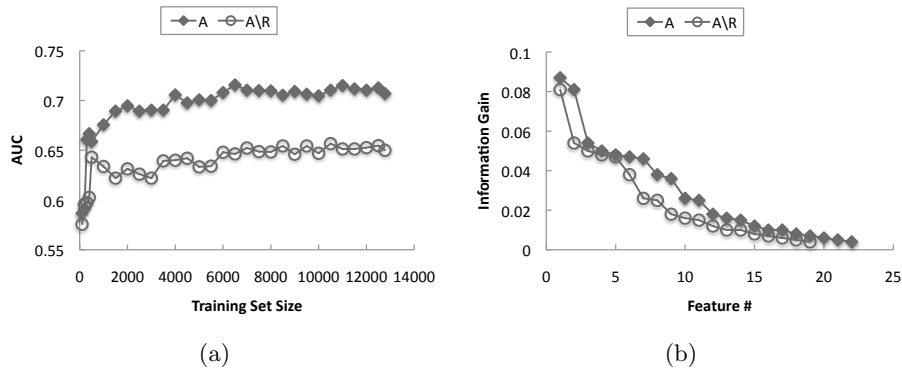


Fig. 1. Classification performance versus training set size (a) and information gain for individual features (b) when reputation features are included (labeled ‘A’) and excluded (labeled ‘A\R’)

the rulesets are large and the decision trees are bushy so the resulting ranking is sufficiently fine for our purposes. Given that the random forest is an ensemble of decision trees, it can also assign a confidence to all the predictions it produces.

3 Evaluation

In the previous section, we have described how a classification approach can be used for recommendation. The success of this approach will depend on the accuracy achieved by the classifiers and how this translates into the recommendation of helpful reviews. In this section, we examine these issues in the context of a large-scale study using the TripAdvisor dataset described above.

In the following sections, feature selection by information gain and the classification performance versus feature subset size experiments were performed using a randomly selected 25% of dataset instances. The training set size and recommendation experiments were performed using the remaining data. Classification performance is evaluated using area under ROC curve (AUC) [7]. AUC produces a value between 0 and 1 and is equal to the probability that a classifier will rank a randomly chosen positive instance (in our case, a helpful review) higher than a randomly chosen negative one (unhelpful review).

3.1 Classification Results

The classification performance achieved by J48 versus training set size is shown in Figure 1(a). It is surprising to note that AUC continues to improve with the addition of new data, even beyond 10,000 examples. This indicates that the relationship between review helpfulness and the predictive features is complex and that very large training sets will be required to build good models.

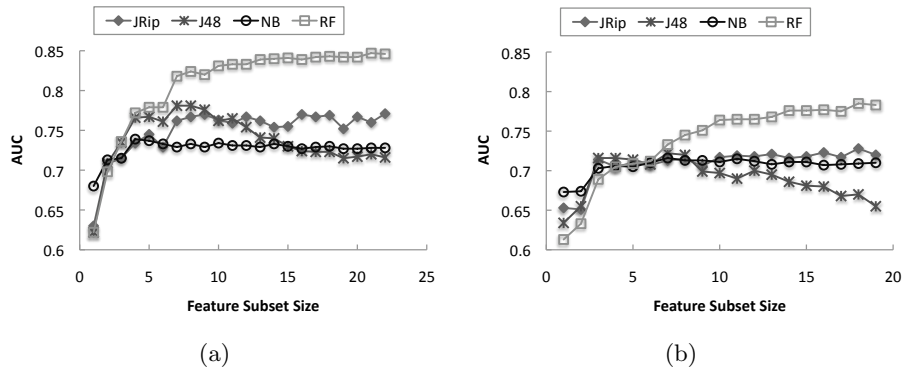


Fig. 2. Performance of JRip, J48, naïve Bayes (NB) and random forest (RF) versus feature subset size when reputation features are included (a) and excluded (b)

Figures 2(a) and 2(b) show classification performance for increasing feature subset sizes when reputation features are included and excluded, respectively. While classifiers trained using reputation features performed better in terms of AUC, similar trends were observed for both conditions. (We discuss reputation features in more detail in Section 3.2.).

These results indicate that J48 and to a lesser extent, naïve Bayes, were much more sensitive to a small number of optimal features than JRip or random forest. As can be seen from Figure 1(b), where information gain for individual features ranked in descending order of performance is shown, information gain declined rapidly with low values observed for the lower-ranked features. These latter features appear to be very noisy and lead to overfitting in the case of J48 and, to a lesser extent, naïve Bayes. Thus the removal of lower-ranked features improved the performance achieved by these classifiers. Interestingly, this was not the case for random forest and JRip, which proved to be robust in the presence of noisy features. Indeed, random forest can get some useful information from these noisy features with performance improving until all features are included. This is probably due to the ability of the random forest to reduce the variance component of error because it is an ensemble technique.

3.2 Recommendation Results

Ultimately, the use of classification techniques are a means to enable the recommendation of reviews to a user. The above findings indicate that reasonable classification performance has been obtained, and thus we can be optimistic that the approach can provide a basis for high quality recommendations. In this section, we evaluate the quality of these recommendations.

We adopt a form of recommendation similar to that implemented by Amazon.com, where the most helpful contrasting reviews (i.e. one positive and one

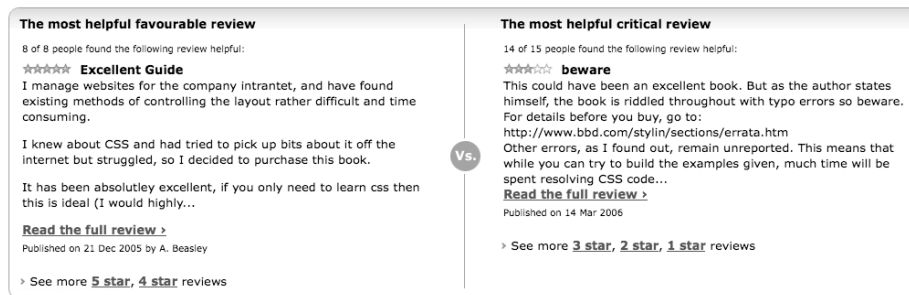


Fig. 3. An example of the Amazon.com approach to review ranking, where the most helpful positive and negative product reviews are listed side-by-side

negative review³) are presented to users side-by-side (see Figure 3). Of course this approach is limited to cases where sufficient feedback on review helpfulness has been amassed; for example, in our dataset, only 31% of all reviews were rated 5 times or more. In addition, we consider two alternatives to our classification-based recommendation technique by ranking reviews by *date* (recommending the most recent positive and negative reviews) and ranking reviews at *random* (recommending a randomly selected positive and negative review).

We construct positive and negative recommendation test sets using only those hotels which have a minimum of 5 positive or negative reviews, respectively. There are 178 and 96 such hotels in the dataset, respectively. During recommendation we adopt a leave-one-out approach such that, for each test set hotel, we recommend its most helpful positive and negative reviews using classifiers which are trained on the reviews of all other hotels in the dataset.

Recommendation performance is evaluated in terms of how frequently our recommenders manage to select a review that is unambiguously helpful according to the definition given in Section 2.1; that is, a review where 75% or more of the raters of that review have found it helpful. In particular, we consider the enrichment in the percentage of helpful reviews recommended across test set hotels with respect to the random ranking scheme as described in Section 2.

We compare the enrichment provided by the two best performing classifiers from Section 3.1 – the ensemble random forest approach and the J48 decision tree. Given that random forest and J48 gave best performance using feature sets consisting of all available features and the top 8 features, respectively, we present results in Figure 4 for each classifier using both of these feature sets. Results obtained using the top 8 features are indicated by an asterisk (*).

Positive versus Negative Reviews To begin, consider the recommendation performance that was achieved when reputation features were included. It is clear that both classifiers provided significant enrichment benefits relative to

³ Recall from Section 2 that a positive (resp. negative) review is defined as any review in which a score of ≥ 4 -stars (resp. < 4 -stars) has been assigned to the hotel.

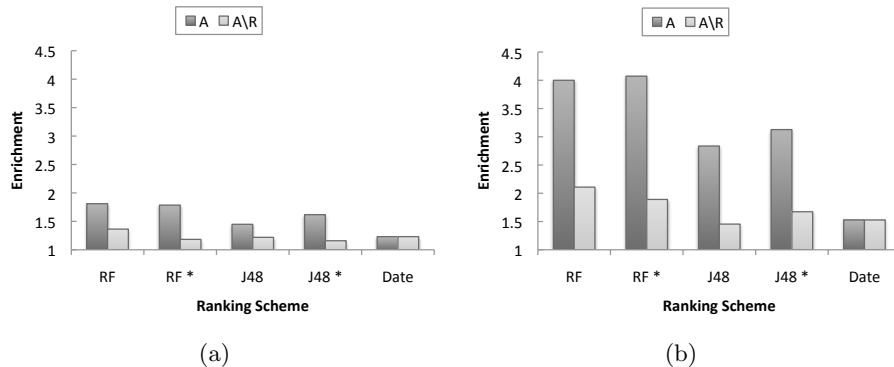


Fig. 4. Enrichment in the percentage of helpful reviews recommended with respect to the random ranking scheme for positive reviews (a) and negative reviews (b) when reputation features are included (labeled ‘A’) and excluded (labeled ‘A\R’)

random and date-based ranking. Random forest performed the best, achieving an enrichment of 1.8 and 4 for positive and negative reviews, respectively. Similar performance was seen using the two feature subset sizes considered. As expected, J48 achieved greater enrichment using a feature subset consisting of the top 8 features, where an enrichment of 1.6 and 3.1 was achieved for positive and negative reviews, respectively. The effect of feature subset size on the classifiers is consistent with the trends observed in Figure 2(a).

Interestingly, the classification-based approaches provided the greatest benefits in relation to the recommendation of negative reviews. As can be seen in Figure 5, the mean percentage of helpful reviews in positive and negative review test sets was 47% and 14%, respectively. Clearly, users are less inclined to find reviews expressing negative sentiment helpful, and consequently ranking such reviews by date and at random was unable to achieve good performance. Hence the need for high-performing review ranking schemes is greater for reviews expressing negative sentiment, and our results indicate that the classification-based recommenders indeed achieved particularly good enrichment for these reviews.

Reputation Features Reputation features are designed to capture the helpfulness of reviews that individual users have authored in the past. We can therefore expect that such features are likely to exert considerable influence on recommendation performance. These features are not, however, available for all reviews (because many TripAdvisor reviews receive little or no feedback). In addition, not all online services facilitate the provision of feedback on reviews, and it is therefore important to consider performance when reputation features are excluded from training instances.

When reputation features were excluded, much more modest enrichment benefits were observed from a recommendation perspective. For example, in the case of positive reviews, random forest and J48 achieved an enrichment of only 1.4 and

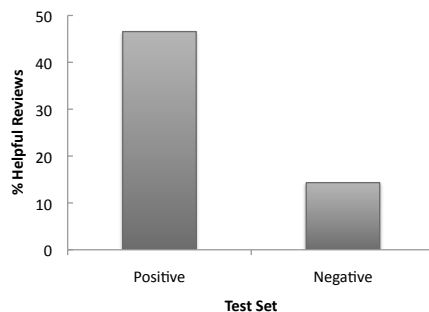


Fig. 5. Mean percentage of helpful reviews in positive and negative review test sets

1.2, respectively, compared to 1.2 for ranking by date. Once again, classification performance for negative reviews was considerably better, where enrichments of 2.1, 1.7 and 1.5 were achieved for random forest, J48 and ranking by date, respectively. Thus, while the classification-based approaches have provided enrichment benefits in the absence of reputation features (and particularly so with respect to negative reviews), these results nevertheless highlight the importance of the reputation features in terms of recommendation performance.

4 Conclusions

In this paper, we have presented a classification-based approach to the recommendation of helpful product reviews. We have considered various classifiers and examined their performance in terms of training set size, feature selection, classification accuracy and the enrichment of recommended reviews with respect to benchmark ranking schemes. The learning task proved to be complex with AUC performance continuing to improve even for very large-sized training sets. JRip and random forest were robust to the presence of noisy features, while the performance of J48 in particular, and naïve Bayes to a lesser extent, improved when learning was based on feature subsets consisting of the top features as ranked by information gain. The ensemble random forest classifier provided best overall performance, and the importance of the reputation features was also discussed. The results also indicated a bias in respect of review recommendation, where the classification-based approach delivered greater enrichment for negative reviews. This finding is significant, given that such reviews are perceived as being less helpful by users, and hence the need for ranking schemes that can accurately recommend helpful reviews which express negative sentiment.

The proliferation of user-generated content continues in the world of the social web and the following related work is of interest. Recently, the effect of credibility indicators in topical blog post retrieval has been investigated [8]. Several indicators (features) were considered; for example topical consistency, regularity of posts and various measures such as spelling quality, length of post

and the appropriate use of capitalisation and emoticons etc. in the text. The use of such indicators were found to significantly improve retrieval performance in [8]. Machine learning techniques were employed in [9] to classify text-based reviews from a sentiment perspective (i.e. positive versus negative reviews) using content-based feature sets. A study based on TripAdvisor reviews demonstrated the effectiveness of the approach. A classification approach was also adopted in [10] to distinguish between conversational and informational questions in social Q&A sites (e.g. Yahoo! Answers, Answerbag). In this work, features such as question category, text categorization and social network metrics were selected as the basis for classification and good performance was achieved.

In future work, we will include additional features from the above and other related work in our classification and recommendation model. We are particularly interested in developing a sufficiently rich feature set where performance does not rely as strongly on user reputation features, given that these features are not always available for particular reviews and that not all services support such feedback on reviews. In addition, we plan on applying our approach to other product review domains, e.g. Amazon.com, Hostelworld.com and blippr.com. The latter domain is of interest given that review texts are constrained to 160 characters in length, which poses additional challenges from a classification perspective.

References

1. Bilgic, M., Mooney, R.J.: Explaining recommendations: Satisfaction vs. promotion. In: Beyond Personalization Workshop, held in conjunction with the 2005 International Conference on Intelligent User Interfaces, San Diego, CA, USA (2005)
2. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceeding on the ACM 2000 Conference on Computer Supported Cooperative Work, Philadelphia, PA, USA (2000) 241–250
3. Gretzel, U., Fesenmaier, D.R.: Persuasion in recommender systems. *International Journal of Electronic Commerce* **11**(2) (2006) 81–100
4. O’Mahony, M.P., Smyth, B.: Learning to recommend helpful hotel reviews. In: 3rd ACM Conference on Recommender Systems (RecSys 2009). (2009)
5. Breiman, L.: Random forests. *Machine Learning* **45**(1) (2001) 5–32
6. Cunningham, P.: Dimension Reduction. In Cord, M., Cunningham, P., eds.: *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Springer (2008) 91–112
7. Fawcett, T.: Roc graphs: Notes and practical considerations for researchers. In: Technical Report HPL-2003-4, HP Laboratories, CA, USA. (2004)
8. Weerkamp, W., de Rijke, M.: Credibility improves topical blog post retrieval. In: Proceedings of the Association for Computational Linguistics with the Human Language Technology Conference (ACL-08: HLT). (June 16 –18 2008) 923 – 931
9. Baccianella, S., Esuli, A., Sebastiani, F.: Multi-facet rating of product reviews. In: *Advances in Information Retrieval*, 31th European Conference on Information Retrieval Research (ECIR 2009), Springer (April 6 – 9 2009) 461 – 472
10. Harper, F.M., Moy, D., Konstan, J.A.: Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI’09), Boston, MA, USA (April 2009) 759 – 768