AAAI Technical Report SS-12-06
Wisdom of the Crowd

Crowdsourcing Evaluations of Classifier Interpretability

Amanda Hutton, Alexander Liu, Cheryl Martin

Applied Research Laboratories, The University of Texas at Austin, Austin, TX, USA
{ahutton, aliu, cmartin}@arlut.utexas.edu

Abstract

This paper presents work using crowdsourcing to assess explanations for supervised text classification. In this paper, an *explanation* is defined to be a set of words from the input text that a classifier or human believes to be most useful for making a classification decision. We compared two types of explanations for classification decisions: human-generated and computer-generated. The comparison is based on whether the type of the explanation was identifiable and on which type of explanation was preferred.

Crowdsourcing was used to collect two types of data for these experiments. First, human-generated explanations were collected by having users select an appropriate category for a piece of text and highlight words that best support this category. Second, users were asked to compare human- and computer-generated explanations and indicate which they preferred and why. The crowdsourced data used for this paper was collected primarily via Amazon's Mechanical Turk, using several quality control methods.

We found that in one test corpus, the two explanation types were virtually indistinguishable, and that participants did not have a significant preference for one type over another. For another corpus, the explanations were slightly more distinguishable, and participants preferred the computer-generated explanations at a small, but statistically significant, level. We conclude that computer-generated explanations for text classification can be comparable in quality to human-generated explanations.

Introduction

Many recent studies have demonstrated that crowdsourcing is an effective method for collecting labels for supervised machine learning. This paper demonstrates that crowdsourcing is also a natural method for evaluating the results of machine learning algorithms, particularly for evaluations that are inherently subjective. In particular, crowdsourcing is applied to assess the interpretability of supervised text classification.

Text classification is a specific type of machine learning problem where the goal is to automatically categorize a document as belonging to a predefined class such as a topic or overall sentiment. Human effort is often required to label a large number of documents for use by the text classification algorithm for training (i.e., for building a model useful for classifying future documents).

The accuracy of a classification system plays an important role in the acceptance of the system for use in practice. This acceptance is increased when a useful and understandable justification is provided alongside the classification decision (Symeonidis et al. 2008). Providing justifications with the classification decision may also help users quickly determine the validity of a decision. Tintarev and Masthoff (2007) discuss seven possible goals against which justifications provided by classifier systems might be judged. The importance of each of these goals may vary depending on the application of the system.

Unfortunately, justifications are not typically produced by machine learning systems. The output of machine learning classifiers is often limited, frequently consisting only of the classification decision itself. Moreover, justifications provided by machine learning algorithms can be difficult to interpret. Few methods of presenting explanations to non-machine learning experts have been widely accepted, and the most user-friendly method of presenting machine learning justifications is still an open question.

This paper offers insight into the interaction of human users with a text classification system. In particular, we address effective methods of presenting output of text classification algorithms to users through human- and machine-provided justification. We rely on crowdsourcing methods to gather the human-provided data for this paper. We also depend on crowdsourced responses to evaluate the output of a machine learning algorithm that is used to provide an explanation of the classification decision.

Related Work

For many applications, the use of crowdsourcing is a practical and cost-effective technique. There are many advantages to using a crowdsourcing service such as Amazon's Mechanical Turk.

Despite potential advantages, there are concerns with some users providing spam responses. Previous research has developed methods to identify and control spammers as well as to improve usability within the task assignment. Some question types seem to be more vulnerable to spam responses than others. Eickhoff and de Vries (2011) specify several common spamming and cheating techniques that are used and provide suggestions for counteracting these nuisance behaviors. We address many of these issues in our quality control methods.

Most crowdsourcing quality control techniques include using questions for which ground truth exists to check for response validity. Often, these questions are used to estimate the true worker error rate (Snow et al. 2008; Ipeirotis et al. 2010). Grady and Lease (2010) provide additional insight into improving user responses. They find that there is a correlation between worker effort and tasks that provide bonus opportunities.

Applications similar to ours, which use crowdsourcing as a method for evaluation, can be found in Alonso et al. (2008) and Paiement et al. (2010), among others. Alonso et al. (2008) use crowdsourcing to evaluate relevance of search query results. Crowdsourcing techniques were also used in Paiement et al. (2010) to evaluate relevance of results from local business searches.

The application-specific component of this paper concerns the application of machine learning to text classification. Although machine learning in this context is useful and cost effective, it is difficult to explain classification decisions to users. Most text classification algorithms can only present a predicted class label for the document, which does not offer insight into why or how the label was chosen. Some classifiers can present probabilistic output, or be modified to produce probabilities, such as in Platt (1999). Rule-based or tree-based classifiers can produce some explanation for a predicted label, since the created model in these cases consists of fairly intuitive, human-understandable rules (e.g., Cohen and Singer 1999). However, these explanations can be difficult to interpret for laypersons.

Traceback for Linear Classifiers

We define an *explanation* to be a set of words that a classifier or human believes to be most useful for making a decision. We use a straightforward method of extracting a list of words during text mining that are most important in making a classifier decision. This paper assesses whether the results of this algorithm are useful as explanations.

Consider a classifier that learns a decision rule to classify a document $\boldsymbol{x}$ as positive if $\sum_{i=1}^{n_v} \theta_i x_i > \tau$ (where θ_i is the weight learned for the ith feature, x_i is the value of the ith feature, n_v is the number of features, and τ is some learned threshold). If a document is classified as positive, then the largest positive term $\theta_i x_i$ for some i is the term that contributes the most to the classifier's decision. Conversely, if $\boldsymbol{x}$ is classified as negative, then the largest negative term $\theta_i x_i$ is the term that contributes most to the classifier decision. Thus, one can provide explanations for a linear text classifier by extracting the n_e most indicative words.

The approach is similar but not identical to the keyword-based approach described in Stumpf (2009), and can be used with any classifier that builds a linear model, such as SVM with linear kernels and multinomial naïve Bayes. (Note that multinomial naïve Bayes is usually not presented as a linear model, but for binary class-problems, one can show that the model learned by multinomial naïve Bayes is of this form.) We use the multinomial naïve Bayes algorithm to generate the list of words used as computer-generated explanations in our experiments. We selected naïve Bayes because it tended to perform best (in terms of accuracy) in preliminary experiments on our test corpora.

Using Crowdsourcing to Evaluate Explanations for Machine Learning Decisions

We use the reviews dataset[1] and a subset of the Enron e-mail dataset[2] as test corpora. The reviews dataset is distributed as a fully labeled dataset, and the classes indicate whether a movie review is positive or negative. The Enron dataset is not pre-labeled. Our class distinctions for the Enron corpus indicate whether an e-mail is work-related or not work-related. We had our lab members assign these class labels. Each e-mail was labeled twice and when the labels were inconsistent, a third labeler was used as a tie-breaker.

The first aspect of this study consisted of collecting the human explanations for both corpora. We ran several crowdsourced rounds of data collection where we had users label the most important words in determining the class label of a particular document. For this collection of human explanations we utilized both Mechanical Turk and members of our lab. For the reviews dataset, Mechanical Turk participants were asked to read a given review and determine whether it was positive or negative and to select a minimum of five explanatory words. For the Enron dataset, lab members were asked to read a given email, determine whether it was work-related or not work-related and select words that supported this classification. The participants creating these explanations did not have access to the pre-determined (ground-truth) class labels, which were used for quality control.

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/
[2] http://www.cs.cmu.edu/~enron/

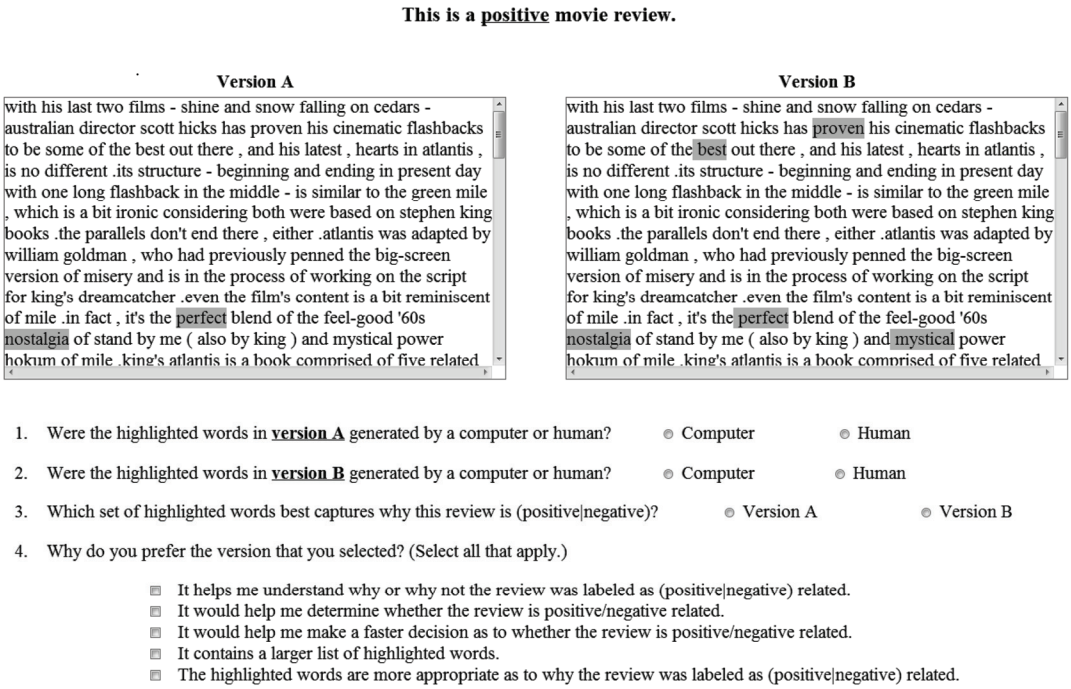

Figure 1: Screen shot of the Mechanical Turk HIT to evaluate the quality of particular explanations. Version A presents an explanation generated by a human and Version B presents a computer-generated explanation.

We next evaluated explanations and compared human- versus computer-generated explanations. This study can be framed similarly to work by product marketing researchers who want to determine what types of products are preferred by consumers and/or other research asking humans to subjectively rate algorithmic results (e.g., results of speech synthesis, search ranking). Thus, one possibility is to present a single explanation and have each user rate each explanation using some scale. However, there are known problems with this approach. Instead, the approach we chose to use is an "A-B comparison" between objects. In this case, two objects are presented at a time, and the user evaluating the two objects can simply make a binary choice of whether they prefer choice A or choice B. Figure 1 provides a screen shot corresponding to this approach using a Mechanical Turk Human Intelligence Task (HIT).

This assessment determines whether there is a clear preference for a human-generated versus a computer-generated explanation. We were also interested in determining whether a user could differentiate between a human-generated explanation and a computer-generated explanation, as well as collecting reasons for why a particular explanation was better than another. For this comparison between explanation types, we used the word lists generated from the naïve Bayes algorithm as the computer-generated explanations, and we used the word lists collected from the previous step as the human-generated explanations.

There was, not surprisingly, a notable difference between the time it took to gather the human-generated explanations and the time it took to differentiate between the explanations. Collecting the human explanations took approximately 20 times longer.

Quality Control

The probability of users providing spam feedback in Mechanical Turk is high; therefore, we applied several checks for spamming. Eickhoff and de Vries (2011) note that closed-class questions (i.e., questions with radio buttons or check boxes) are subject to random responses, and it is best to have some sort of ground truth comparison to filter out spam responses. The paper also provides warning for using open-ended questions without a method of checking the quality of responses. Sometimes individuals will copy and paste random bits of text from the HIT and submit that as the response.

Along these lines, we provided participants with a list of behaviors that would result in HIT rejection. These included routine incorrect labeling of the reviews, providing fewer than 5 words, and providing words that did not exist in the review or were misspelled. We also allowed users to provide more than 5 words, which many users did. Additionally, we provided a bonus under the terms that for every word a user provided that matched a word given from a different user for the same review they would receive $0.01 bonus.

To reduce the chance of attracting an abundance of low-quality workers, only those with a 90% acceptance rate were allowed to participate. We also used one form of ground truth as a check against spamming and/or cheating. This ground truth was the pre-labeled review category (positive/negative). Participants were asked to classify a review as positive or negative and we then filtered out incorrectly labeled HITs on the following basis: (1) we assumed that workers who habitually labeled the reviews incorrectly were not providing valid or high quality results, and (2) if a worker thought that a negative review was positive and provided justification at the positive level, this result was thrown out because the justification would be incorrect for the second phase of the experiment. Lastly, we rejected HITs from users who had labeled more than 10 HITs with an accuracy of less than 90%.

We also checked to verify that the words that workers provided were in fact found in the original review, that each word was unique, and that no words were duplicated. Once we had the final count of valid words for each HIT, we verified that there were at least 5 words present. If there were not, we rejected the HIT per warnings in the HIT instructions. All rejected HITs were republished. We reviewed the batch of HITs for quality control three times a week until the batch was fully completed and until all HITs had been approved. We observed that each time the batch of HITs was reviewed, approximately 30% of HITs were filtered out or rejected.

We found that 67% of workers who completed more than one HIT and who were not removed in the quality control phase provided more than the minimum 5 words. Figure 2 shows the distribution of the HITs that contained more than 5 words provided by the worker. Workers may have been more inclined to provide more than 5 words because of the bonus that we offered. This gave potential for workers to significantly increase their pay. We believe that this not only helped motivate workers to provide more than the minimum five words, but it likely helped motivate workers to provide useful and clear explanations.

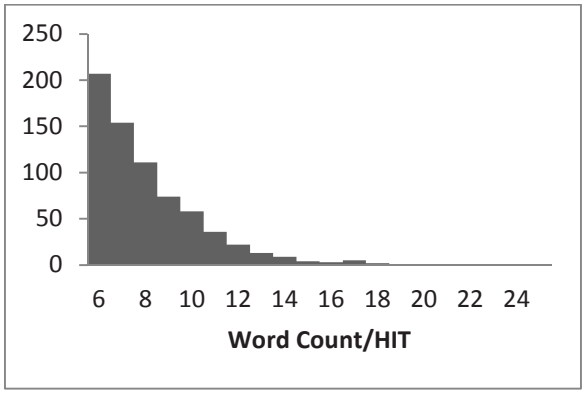

Figure 2: Number of HITs containing more than 5 words provided by a worker.

In the second phase of this experiment, we asked workers to compare two different explanations of the classification for a review or email. Ensuring quality for this phase proved to be more difficult because we did not have an obvious ground truth to check against. We could not use the same technique from the previous phase that asked workers to classify the text because we were fundamentally testing the quality of classification explanations *given* a specific class. For ensuring quality, we again only used workers with a 90% HIT acceptance rate. Additionally, we did not use responses that were completed faster than two standard deviations from the mean.

Results

We assessed whether participants could correctly identify the type of explanation (human- or computer-generated) and which type of explanation participants preferred. For each HIT, participants were presented with two explanations (Version A and Version B, as shown in Figure 1). Participants were instructed that both explanations could be the same type (e.g., both human-generated). In this assessment, participants had to correctly identify the source of both the A and B explanations, each of which is a binary decision. This means that a correct identification rate of 0.25 for each HIT is no better than random.

For the Enron corpus, there were 765 valid HITs. Out of those, 177 had both explanations correctly identified. This gives a proportion of 0.231, which is slightly worse than 0.25. These results indicate that for the Enron corpus, participants are unable to differentiate well between the two types of explanations. Figure 3 presents examples of emails with both correctly identified explanations and incorrectly identified explanations. Since these are examples of personal emails, proper names, phone numbers, and web addresses have been removed in the paper examples for privacy, but were not removed in the actual HITs.

Results for the reviews corpus consisted of 1000 valid HITs. Of these, the types for both A and B explanations were correctly identified for 285 responses, which gives a proportion of 0.285. Therefore, for the reviews corpus, the distinction between the explanations types was better than random and more obvious to participants than in the Enron corpus.

We also assessed user preference towards one explanation type over another. The results were compared to 0.5, which is the random baseline proportion that one would expect if there was no preference between human- and computer-generated explanations.

Examples Emails for which Participants Identified Both Sources Correctly		
Email Text	Dear Ken, I have been told that you will not have time to meet with me, prior to Mr. Fox dinner in New York. I have to be in N.Y., anyhow, so let me know if you feel like having a night cap, after your dinner. Here is some relevant information regarding the participants at the dinner and their interests in Mexico. Best regards, jaime - Alatorre.doc	Michelle, I took advantage of the long weekend and wanted to share this good news with you first because you have been supportive of my digital endeavors. Color me excited! My digital waterfall electronic storefront is open for business. You may have seen some of the images in the Oregon 2000 vacation report on my personal web site. I've placed the best of the lot for viewing and purchase on the new storefront site. The address is: http://[WEB_ADDRESS] Come linger awhile. I think you will enjoy a tour of the gallery. The images are magnificent! Michelle, don't worry about purchasing; just enjoy the views. I wanted to show off my web design skills and share these delightful waterfall images with you. Click on a thumbnail to see an enlargement of the image. Paddrick (the digiographer formerly known as Pat)
Naïve Bayes	meet, me, prior, cap, relevant, interests, regards, doc	weekend, my, digital, color, lot, awhile, image
Human	meet, information, participants	long weekend, waterfall, images, tour, gallery, thumbnail
Examples Emails for which Participants Incorrectly Identified Both Sources		
Email Text	Rick, Unfortunately, contrary to the performance management system's reply (see below), performance reviews are not confidential in the DC office as the office manager has access to the passwords of some of the Vice Presidents. I would like to comply with your request to give a performance review for Cynthia [LAST_NAME], but unless I can send it to you in a confidential envelope I have no guarantee it will not be seen at some point by my colleagues. (You are listed as her supervisor according to the PEP system.) This poses a problem for me. Your advice is welcome. Lora [LAST_NAME] [PHONE_NUMBER]	As we move forward with our global LNG strategy, it is obvious that the best way to capture value for Enron is to bring the same merchant activity to this business unit. This merchant philosophy allows aggressive book and position management from a worldwide perspective. Our ability to gather data and to show only one Enron face in the market is crucial to our success. As important is one pricing desk. Many of you know, Eric [LAST_NAME] in Houston is running our pricing desk, Larry [LAST_NAME] is working on all finance type business for us, and Jen [LAST_NAME] is running our fundamentals group. To bring full merchant capabilities to our worldwide LNG business, Mike and I wanted to clearly outline a point that seems to need attention. ALL LNG contacts, negotiations, and contracts, (physical, financial, buys, sales, etc.), with external counterparties must and will be managed through our commercial LNG group headed by Eric [LAST_NAME], and Rick [LAST_NAME]. It is crucial to manage market perspective, appearance, and information. We all know how important each project (Metgas, and other initiatives in India) could be, but a coordinated front will help us better grow all of our businesses. Regards Jeff
Naïve Bayes	performance, management, confidential, review, cynthia, confidential, pep, system, sullivan, perez	enron, merchant, pricing, desk, fundamentals, external, counterparties, bergsieker, market
Human	performance, management, reviews, confidential, supervisor, enron, process	forward, global, capture, value, enron, merchant, business, unit, data, market, pricing, finance, contacts, negotiations, contracts, buys, sales, coordinated, grow

Figure 3: Example emails with human- and computer-generated explanations.

For the Enron corpus, 370 of the 765 responses preferred the explanations that were computer generated. This indicates that there is a slight preference towards the human-generated explanations (0.516), but the preference is not significant.

For the reviews corpus, only 426 HITs contained one explanation that was human generated and one explanation that was computer generated. Of these, a preference for the computer-generated explanation was indicated in 235 HITs. This is a proportion of 0.552, which is compared against the value of no preference, 0.5. Results from a z-test found $p<0.005$, which supports the hypothesis that the computer-generated explanations were preferred over the human-generated. Recall that for this dataset, the computer-generated explanations were also more identifiable. Future work will explore whether this correlation can be generalized.

Finally, we also asked participants to explain why they preferred one explanation over another. Specifically, participants were given the following five options, which are based on points raised in Tintarev (2007) for explaining classifier decisions:

1. It helps me understand why or why not the document was labeled as (positive|negative) related. (TRANSPARENCY)
2. It would help me make a faster decision as to whether the document is positive/negative related. (EFFICIENCY)
3. It would help me determine whether the document is positive/negative related. (EFFECTIVENESS)

4. The highlighted words are more appropriate as to why the document was labeled as (positive|negative). (APPROPRIATENESS)
5. It contains a larger list of highlighted words. (LENGTH)

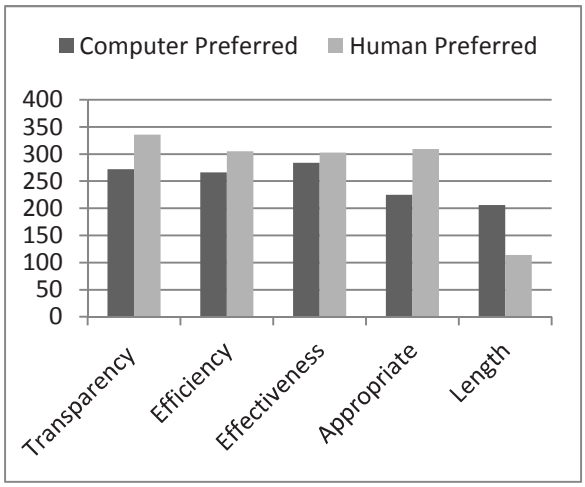

Figure 4: Frequencies of each goal trait for which explanation type was preferred.

As shown in Figure 4, based on the five questions asked, the main advantage for computer-generated explanations is length. For the remaining four metrics, the human-generated explanations tend to score better. Despite this, there was a statistically insignificant preference for human-generated explanations in the Enron corpus, and a slight but statistically significant preference for computer-generated explanations in the reviews corpus.

Conclusion

We have presented an effective method for using computer-generated explanations to justify classification decisions. We have also provided a validation technique for these explanations using crowdsourcing.

In order to collect the data used in the explanation validation phase of the paper, we relied on crowdsourced responses. We implemented several quality-control methods that removed poor-quality and spam results from our data. We also found success in providing bonuses to workers when they provided explanations that matched other workers' responses.

We compared human-generated and computer-generated explanations based on their identifiableness and user preference. We further explored the quality of the computer-generated explanations by comparing preference between the two types.

We found that in the Enron email corpus, the two explanation types were virtually indistinguishable, and that participants did not have a statistically significant preference for one type over another. For the movie reviews corpus, we found that the explanations were slightly more distinguishable. Participants preferred the computer-generated explanations at a statistically significant level. Thus, for the datasets used, computer-generated and human-generated explanations are not always distinguishable, but when they are, computer-generated explanations are more often preferred. Future work is warranted for determining whether there is a general correlation between the preference for computer-generated explanations and the ability to identify them.

This study has shown that computer-generated explanations can be at least as good as human-generated explanations. Users in this study either (1) could not distinguish between the two types of explanations or (2) demonstrated a slight preference for computer-generated explanations. This justifies using our straightforward method of providing classifier traceback to provide explanations for supervised text classification decisions.

References

Alonso, O.; Rose, D. E.; and Stewart, B. Crowdsourcing for relevance evaluation. SIGIR Forum, 42:9–15, December 2008.

Cohen, W.W., and Singer, Y. 1999. Context-sensitive learning methods for text categorization. In *ACM Transactions on Information Systems* 17(2):141-173.

Eickhoff, C., and de Vries, A. P. 2011. How Crowdsourcable is your task? In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM),* Hong Kong, China.

Grady, C., and Lease, M. 2010. Crowdsourcing document relevance assessment with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (CSLDAMT '10). Association for Computational Linguistics, Stroudsburg, PA, pp.172-179.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (HCOMP '10). ACM, New York, NY, pp. 64-67.

Paiement, J.-F.; Shanahan, J. G.; and Zajac, R. 2010. Crowdsourcing local search relevance. In *Proceedings of the CrowdConf 2010*, CrowdConf 2010.

Platt, J. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, ed. P. J. Bartlett, B. Schölkopf, D. Schuurmans and A. J. Smola, pp. 61-74. Cambridge, MA: MIT Press.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, pp. 254-263.

Stumpf, S., et al. 2009. Interacting meaningfully with machine learning systems: Three experiments. In *International Journal of Human-Computer Studies* 67(8):639-662.

Symeonidis, P.; Nanopoulos, A.; and Manolopoulos, Y. 2008. Providing justifications in recommender systems. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 38(6):1262-1272.

Tintarev, N., and Masthoff, J. 2007. A survey of explanations in recommender systems. In *Proceedings of the IEEE 23rd International Conference on Data Engineering Workshop*, pp. 801-810.