Human versus Machine Attention in Document Classification: A Dataset with Crowdsourced Annotations

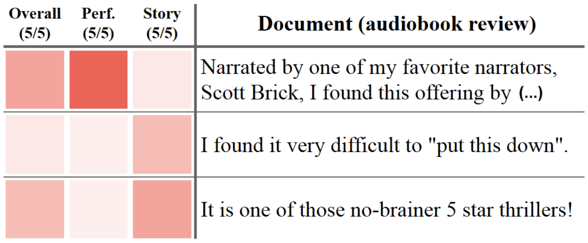
Nikolaos Pappas and **Andrei Popescu-Belis**
Idiap Research Institute
Centre du Parc, Rue Marconi 19
CH-1920 Martigny, Switzerland
{nikolaos.pappas, andrei.popescu-belis}@idiap.ch

Abstract

We present a dataset in which the contribution of each sentence of a review to the review-level rating is quantified by human judges. We define an annotation task and crowdsource it for 100 audiobook reviews with 1,662 sentences and 3 aspects: story, performance, and overall quality. The dataset is suitable for intrinsic evaluation of explicit document models with attention mechanisms, for multi-aspect sentiment analysis and summarization. We evaluated one such document attention model which uses weighted multiple-instance learning to jointly model aspect ratings and sentence-level rating contributions, and found that there is positive correlation between human and machine attention especially for sentences with high human agreement.

1 Introduction

Classifying the sentiment of documents has moved past global categories to target finer-grained ones, such as specific aspects of an item – a task known as multi-aspect sentiment analysis. An important challenge for this task is that target categories have "weak" relations to the input documents, as it is unknown which parts of the documents convey information about each category refer to. Using supervised learning to solve this task requires labeled data. Several previous studies have adopted a strongly-supervised approach using *sentence-level* labels (McAuley et al., 2012; Zhu et al., 2012), obtained with a significant human annotation effort. However, *document-level* labels are often available in social media, but learning from them requires a weakly-supervised approach. Recently, attention mechanisms for document modeling, either using hierarchical neural networks (Yang et al., 2016) or weighted multiple-instance learning (Pappas and Popescu-Belis, 2014), have proved superior in classification performance and are also able to quantify the contribution of each sentence to the document-level category.



Narrated by one of my favorite narrators, Scott Brick, I found this offering by (...)

I found it very difficult to "put this down".

It is one of those no-brainer 5 star thrillers!

Figure 1: Human attention to sentences when attributing aspect ratings (overall, performance, or story) to an audiobook review.

While explicit document models can be indirectly evaluated on aspect rating prediction or document segmentation, a more direct way to estimate their qualities is to compare the sentence-level weights or attention scores that they assign with those assigned by human judges. In this paper, we present a dataset[1] containing human estimates of the contribution of each sentence of an audiobook review to the review-level aspect rating, along three aspects: story, performance, and overall quality.

Following a pilot experiment (Sec. 2), the annotation task was fully specified and crowdsourced. Statistics about the resulting dataset are given in Sec. 3. We show how the dataset can be used to evaluate a document attention model based on multiple-instance learning (outlined in Sec. 4), by comparing

[1] Available at www.idiap.ch/paper/hatdoc/.

In this task we ask you to rate the explanatory power of sentences in a user review of an audiobook with respect to the user's opinion about the following aspects of the audiobook (recorded reading of a paper book):

***Overall**:* General rating based on all aspects, including also author attributes (writing style, imagination, etc.)
***Performance**:* Rating based on narrator attributes (acting, voice, role, etc.)
***Story**:* Rating based on the story attributes (plot, characters, setting, etc.)

We provide: the sentence under examination highlighted in the entire user review; the user's rating on a five-star scale towards an aspect of the audiobook (namely, 1: very negative, 2: negative, 3: neutral, 4: positive, 5: very positive). The question and possible answers are displayed for each required rating.

The question is: *"How much does the highlighted sentence explain the given aspect rating?"* or in other words "How much does the highlighted sentence carry the user's opinion about each aspect?" The answer is one of the following choices of how much each sentence explains the displayed aspect rating: 'not at all', 'a little', 'moderately', 'rather well', and 'very well'.

Figure 2: Main annotation instructions given to human judges in the crowdsourced task.

the sentence attention scores with those obtained by humans (Sec. 5). We find a positive correlation between human and machine attention for high confidence annotations and show that the system is more reliable than some of the qualified annotators.

2 Pilot Annotation

We defined the requirements for a pilot experiment to reflect our interest in capturing sentence-level justifications of the aspect ratings indicated in a review. The focus is on the sentiment of a sentence, and not merely its topic. For example, in an audiobook review, a sentence that lists the main characters of the book is about the story, but it is factual and does not explain the reviewer's sentiment with respect to the story, i.e whether they liked it or not.

Definition. We recruited three annotators with good command of English among our colleagues. They were given ten audiobook reviews in self-contained files, along with the aspect rating scores (1–5 stars for 3 aspects) assigned by the authors of the reviews. The aspects, namely 'overall', 'performance' and 'story' were briefly defined, e.g. as "about plot, characters or setting" for the latter. The annotators had to answer on a 5-point scale the following question for each sentence and aspect: *"How much does the sentence explain why the user rated the aspect as they did?"* We instructed the annotators to assign explanatory scores only when they met opinionated sentences (expressing sentiment) and to ignore factual sentences about the aspects, as well as subtle or indirect expressions of opinions.

Results. We obtained 684 sentence-level scores for 3 aspects in 10 reviews. The agreement between each pair of annotators was computed using Pearson's correlation coefficient r (Pearson, 1895) and Cohen's kappa coefficient κ (Cohen, 1960). For the latter, since we do not want to treat two different labels as a complete disagreement, we incorporated a distance measure, namely the absolute differences of normalized values between annotators.

The pairwise scores between annotators a, b and c are listed in Table 1. When computed over all rating dimensions, the average r coefficient is 0.72 (strong positive linear relationship) and the average κ is 0.79 (substantial agreement). Both values show that the obtained sentence labels are to a great extent reliable. When considering each aspect separately, the largest agreement was achieved on 'performance', followed by 'story', and then 'overall'. This is most likely due to our definition of the latter aspect to include all other aspects as well as author attributes.

	$a \leftrightarrow b$		$b \leftrightarrow c$		$c \leftrightarrow a$	
	r	κ	r	κ	r	κ
Ov.	0.80	0.81	0.44	0.60	0.48	0.64
Pr.	0.96	0.97	0.87	0.92	0.89	0.92
St.	0.73	0.79	0.63	0.72	0.72	0.78
All	0.84	0.86	0.64	0.75	0.70	0.78

Table 1: Pearson's correlation (r) and Kohen's kappa (κ) scores computed for each aspect (Ov: overall, Pr: performance, St: story) and each pair of annotators (a, b and c) in the pilot study.

3 Crowdsourced Task

Definition. For the definitive task, we wrote detailed instructions to annotators, providing a precise definition of the explanatory value of each sentence with respect to the aspect rating of the review. The main instructions are shown in Fig. 2, and they were complemented with additional tips and observations, as well as two fully-annotated sample reviews. The annotation interface showed for each task the question and possible answers (listed at the bottom of Fig. 2), along with the target sentence, highlighted within

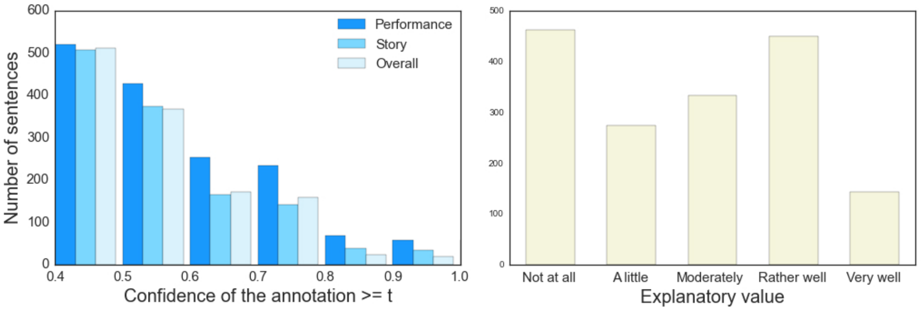


Figure 3: Number of sentences for different confidence values (left) and annotation labels (right).

the review. Each of the three aspects was annotated separately, to avoid confusion.

Results. We collected 100 reviews of audiobooks from Audible (`www.audible.com`) with 1,662 sentences. There are 20 reviews for each rating value of the 'overall' aspect (1–5 stars), to balance the distribution of positive vs. negative reviews. We obtained human judgments over the set of 100 reviews by crowdsourcing the task via Crowdflower (`www.crowdflower.com`).

The reliability of the judges was controlled by randomly inserting test questions with known answers ("gold" questions). Using these questions, Crowdflower computed a confidence score for each judge and then used it to compute the confidence for each annotated example. We only kept the answers of judges who achieved at least 70% success rate on the gold questions. For each non-gold question, we collected answers from at least four reliable annotators, and the majority answer was considered as the gold truth.

We obtained 7,121 judgments of the 1,662 sentences, on the entire spectrum of the rating distributions, as shown in Fig. 3, right side. The confidence of the annotations was computed by Crowdflower as 57% for the 'overall' and 'story' aspects, and 63% for 'performance'. The percentages of sentences with a confidence ≥ 0.8 were quite low, at respectively 4%, 7% and 12% for each aspect. Still, a substantial proportion of sentences have a confidence above 0.5, as shown in Fig. 3, left side. These numbers suggest that the task was the most difficult for the 'overall' aspect, followed by the 'story' and 'performance' aspects.

For evaluating an automatic system, high-confidence annotations (e.g. above 0.6) can be directly compared with labels assigned by a system. An alternative evaluation approach keeps all annotations, but replaces some of the human ratings with system ones, and examines the variation of inter-annotator agreement.

4 System: A Model of Document Attention

We use the data to evaluate a document attention model (Pappas and Popescu-Belis, 2014) which uses multiple-instance regression (MIR, Dietterich et al., 1997) to deal with coarse-grained input labels. The input is a set of bags (here, reviews), each of which contains a variable number of instances (here, sentences). The labels used for training (here, the aspect ratings) can be at the bag level (weak supervision), and not at the instance level. Our system learns to assign importance scores to individual instances, and to predict the labels of unseen bags.

In past models, the influence of instance labels on bag labels has been modeled with simplifying assumptions (e.g. averaging), whereas our system learns to aggregate instances of a bag according to their importance, like attention-based neural networks (Luong et al., 2015). To jointly learn instance weights and target labels, the system minimizes a regularized least squares loss. While in our 2014 paper this was done using alternating projections (as in Wagstaff and Lane, 2007), we use here stochastic gradient descent (Bottou, 1998) with the efficient ADAGRAD implementation (Duchi et al., 2011). In particular, the attention is modeled by a normalized exponential function, namely a softmax and a linear activation between a contextual vector and the document matrix (sentence vectors). Essentially, this formulation enables learning with stochastic gradient descent while preserving the initial instance relevance assumption in the MIR framework and the constraints in our 2014 paper.

The system is trained on a uniform sample of 50,000 audiobook reviews from Audible, with

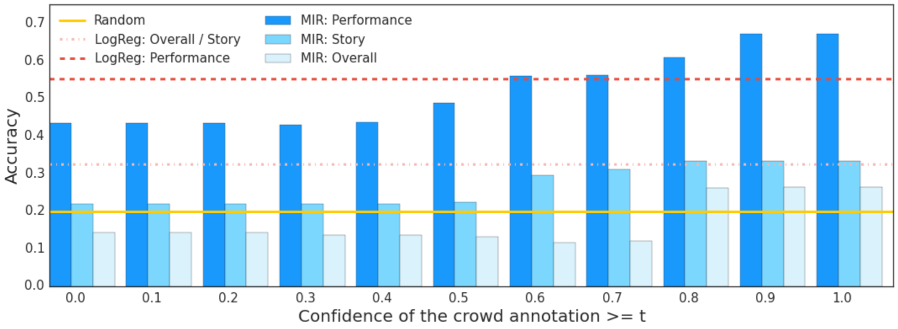


Figure 4: Accuracy of the evaluated system (MIR) on predicting the explanatory value of sentences with respect to review-level ratings of the three aspects, for subsets of increasing crowd confidence values. The accuracy of a supervised system, Logistic Regression, trained on the attention labels with 10-fold cross-validation, is noted LogReg. Random accuracy is 1 out of 5 (20%).

10,000 reviews for each value of the 'overall' aspect (1–5 stars). The training set does not include the 100 annotated reviews, used for testing only.

5 Comparison of System to Humans

Attention prediction. To evaluate the system's estimates of the contribution of each sentence to the review rating, a first and simple metric is the number of sentences for which system and human labels are identical, i.e. *accuracy*. Identity of labels is however hard to achieve, given that even humans do not have perfect agreement. Fig. 4 displays the accuracy of the system, for each aspect, for test subsets of increasing crowd confidence, from the entire test set to only the most reliable labels. Our MIR system appears to achieve the highest accuracy on the 'performance' aspect, exceeding 60% for labels assigned with at least 0.8 confidence by humans. The accuracy for 'story' is 33%, while for 'overall' it is the lowest, at 26%. The system outperforms the random baseline at 20% for 'performance' and 'story'. When compared with the expected accuracy of a supervised system (10-fold cross-validation over the ground-truth labels), namely Logistic Regression, our system achieves similar accuracy on sentences with confidence greater or equal to 0.6.

When relaxing the constraints of exact label matching, i.e. accepting as matches neighboring labels as well (distance 1), the accuracies at the 0.8 confidence level increase to 71%, 43% and 52% respectively for each aspect. Interestingly, the 'overall' aspect benefits the most from this relaxation, showing that many predictions were actually close to the gold label. The MIR performance is greater for higher crowd confidence values, which shows that both the system and the humans find similar difficulties in assigning importance scores to sentences wrt. document-level aspects.

While accuracy gives an indication of a system's quality, it is not entirely informative in the absence of a direct comparison term, such as a better baseline than random guesses. A second evaluation metric enabled by our dataset compares the system's quality with that of human annotators.

Reliability analysis. This more nuanced evaluation places the system on the same scale of qualification, from the most reliable judges (those who most agree with the average) to the least reliable ones. We consider the average standard deviation (STD) among humans, which decreases when the answers of the least reliable judges are removed, and ask: what happens if certain judges are replaced by our system? Fig. 5 displays the difference obtained from the STD of all judges for three replacement strategies:

Random: Select a random label per sentence and replace it with a random value.

Human: Replace the least reliable human judge for each sentence (i.e. largest distance to the average) with the average label of each sentence.

Model: Replace at random an annotator label per sentence with a system one.

As shown in Fig. 5, 'Model' consistently outperforms 'Random' for all aspects and confidence levels, as it leads to a larger decrease (or a smaller increase) in STD. The system performs better than the least agreeing judges on the 'story' and 'overall' aspects, as it leads to a smaller STD than the 'Human' configuration, sometimes even smaller than the initial STD of all judges. Given the qualification

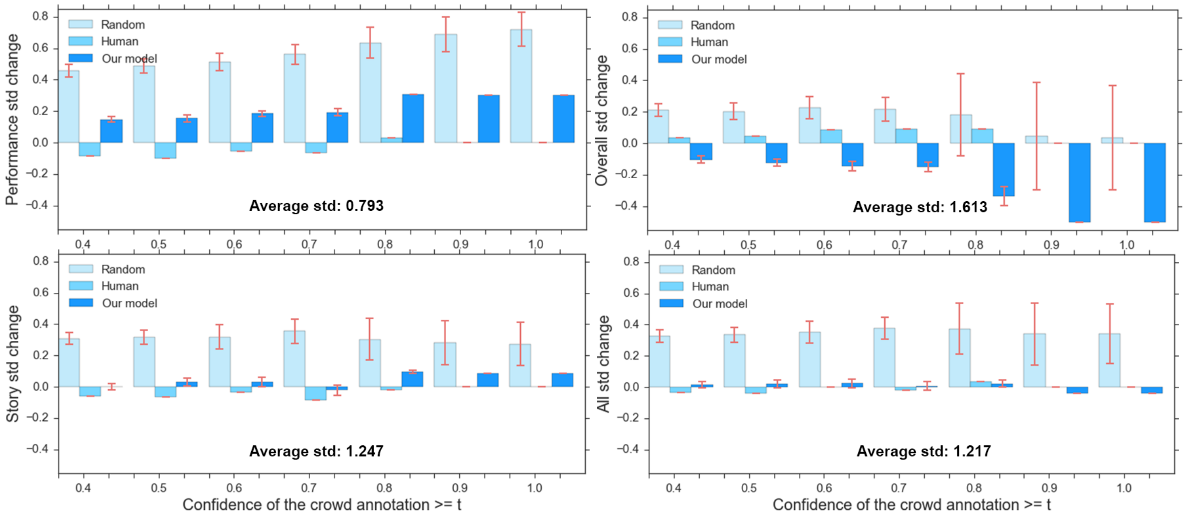


Figure 5: Changes in average STD of the explanatory sentence labels in three replacement experiments (color coded), for each of the three aspects separately and then jointly for all of them.

controls enforced by the Crowdflower, we conclude that the labels assigned by the system are comparable to those of qualified human judges for 'story' and 'overall'. For 'performance', however, the high agreement of judges cannot be matched by the system, according to this metric. Still, these results provide evidence that the weights found by the system capture the explanatory value of sentences in a way that is similar to humans.

6 Related Work

Multi-aspect sentiment analysis. This task usually requires aspect segmentation, followed by prediction or summarization (Hu and Liu, 2004; Zhuang et al., 2006). Most related studies have engineered various feature sets, augmenting words with topic or content models (Mei et al., 2007; Titov and McDonald, 2008; Sauper et al., 2010; Lu et al., 2011), or with linguistic features (Pang and Lee, 2005; Qu et al., 2010; Zhu et al., 2012). McAuley et al. (2012) proposed an interpretable probabilistic model for modeling aspect reviews. Kim et al. (2013) proposed an hierarchical model to discover the review structure from unlabeled corpora. Previous systems for rating prediction were trained on segmented texts (Zhu et al., 2012; McAuley et al., 2012), while our system (Pappas and Popescu-Belis, 2014) used weak supervision on unsegmented text. Here, we introduced a new evaluation of such models on sentiment summarization considering human attention.

Document classification. Recent studies have shown that attention mechanisms are beneficial to machine translation (Bahdanau et al., 2014), question answering (Sukhbaatar et al., 2015), text summarization (Rush et al., 2015), and document classification (Pappas and Popescu-Belis, 2014). Most recently, Yang et al. (2016) introduced hierarchical attention networks for document classification. Despite the improvements, it is yet unclear what exactly this attention mechanism captures for the task at hand. Our dataset enables the direct comparison of such mechanism and human attention scores for document classification, thus contributing to a better understanding of the document attention models.

7 Conclusion

We presented a new dataset with human attention to sentences triggered when attributing aspect ratings to reviews. The dataset enables the evaluation of attention-based models for document classification and the explicit evaluation of sentiment summarization. Our crowdsourcing task is sound and can be used for larger-scale annotations. In the future, statistical properties of the data (e.g. numeric scale), should be exploited even further to provide more accurate evaluations, for instance by relaxing the exact match rule to tolerate marginal mismatches.

Acknowledgments

We are grateful to the European Union for support in the Horizon 2020 SUMMA project n. 688139.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Léon Bottou. 1998. On-line learning and stochastic approximations. In David Saad, editor, *On-line Learning in Neural Networks*, Cambridge University Press, New York, pages 9–42.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* 89(12):31 – 71.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12:2121–2159.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge discovery and data mining*. Seattle, WA, KDD '04, pages 168–177.

Suin Kim, Jianwen Zhang, Zheng Chen, Alice Oh, and Shixia Liu. 2013. A hierarchical aspect-sentiment model for online reviews. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*. Bellevue, WA, AAAI'13, pages 526–533.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Proc. of the 11th IEEE Int. Conf. on Data Mining Workshops*. Washington, DC, ICDMW '11, pages 81–88.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, EMNLP '15, pages 1412–1421.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of the 12th IEEE International Conference on Data Mining*. Brussels, ICDM '12, pages 1020–1025.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on the World Wide Web*. Banff, Canada, WWW '07, pages 171–180.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Ann Arbor, MI, ACL '05, pages 115–124.

Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, EMNLP '14, pages 455–466.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58:240–242.

Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China, COLING '10, pages 913–921.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *CoRR* abs/1509.00685.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2010. Incorporating content structure into text analysis applications. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA, EMNLP '10, pages 377–387.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. Weakly supervised memory networks. *CoRR* abs/1503.08895.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference* *on World Wide Web*. Beijing, China, WWW '08, pages 111–120.

Kiri L. Wagstaff and Terran Lane. 2007. Salience assignment for multiple-instance regression. In *Proceedings of the ICML 2007 Workshop on Constrained Optimization and Structured Output Spaces*. Corvallis, OR.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, NAACL'16.

Jingbo Zhu, Chunliang Zhang, and Matthew Y. Ma. 2012. Multi-aspect rating inference with aspect-based segmentation. *IEEE Transactions on Affective Computing* 3(4):469–481.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. Arlington, VA, CIKM '06, pages 43–50.