Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Suitable for All Ages: Using Reviews to Determine Appropriateness of Products

Elizabeth M. Daly, Oznur Alkan
IBM Research - Ireland
Dublin, Ireland

Michael Muller
IBM Research
Cambridge MA USA

Abstract

Product reviews provide insights in to real user experiences which can benefit others when making their purchasing decisions. Text-mining and NLP may be used to extract features and content that could influence a new user. Additionally, recommender systems and filtering interfaces rely on manufacturer reported data in order to support user preferences. In many instances this data may be absent or inaccurate. In this paper we focus on age related features mentioned in user reviews of baby and child related products in order to recommend the appropriate age range of a product. We demonstrate that manufacturer related information is frequently absent and when manufacturer specifications are available, we find they may not reflect real user experiences which could assist a buyer in their decision making process. As a result, we present a simple user interface to allow users assess the age appropriateness of the product.

The increasing market for online purchasing means users frequently have to make decisions to buy products without the opportunity to see the items in person. Users rely on images and product descriptions which are sometimes unclear or incomplete. Sometimes this information is insufficient for a user to get a sense of a product. Thankfully most retailers include a rating and reviewing facility which allows users to share their experiences.

When purchasing on-line for children, determining the age appropriateness of an item can present a problem when it is not possible to interact with the product. A children's book which a parent might flick through to get a sense of the imagery or puzzles that might be too complicated for younger children might be easy to assess in person.

There is a considerable amount of work that has been done on review analysis for different purposes. ()1 combine latent rating dimensions with latent review topics which enables them to justify ratings with text. The approach also enables to predict product ratings from review text. ()2 use reviews to answer customer queries by relevant review selection and voting mechanisms. Aspect extraction is another fundamental task of opinion mining or sentiment analysis which aims to extract opinion targets from text which is applied to reviews in order to extract important aspects of products ()hu2004mining. ()diao2014jointly presented a probabilistic model based on collaborative filtering and topic modelling in order to capture the interest distribution of users and the content distribution of movies. Sentiment analysis from the on-line dishwasher reviews are used for defect discovery in (Law, Gruss, and Abrahams 2017). ()6 presented a text-mining based solution to uncover potentially dangerous children's toys using a danger word list from injury and recall text narratives.

In contrast to existing studies we aim to use text-mining and sentiment analysis to enhance already existing recommendation capabilities by extracting specific features of a product, in this case *age appropriateness*, which can easily be extended to other dimensions. In addition, we aim to show how review based, rating based and sentiment based estimation of *age appropriateness* differ.

Analysis

We use data crawled from Amazon.com[1] which contains reviews and metadata for the products that are collected for different categories. We have used data for Baby category that includes over 160K reviews for 7050 products from 19445 distinct users. Number of reviews per product ranges between 5 and 780, 22.8 on average; whereas the number of reviews recorded per user ranges between 5 and 125 with an average of 8.27.

We extracted a subset of the reviews that are related to *age appropriateness*. To do so, we have to process the data to select the most relevant subset which is detailed in the following subsection. The resulting dataset is a subset of 160K reviews which includes 45263 reviews stored for 6154 products from 15390 distinct users.

Data Cleaning and Processing

Data cleaning and processing includes six phases:

- **Term Extraction:** Terms are annotated using the Alchemy API service[2]. Each term can be a *concept, keyword* or an *entity*. Entity terms can be associated with an entity type like organization, person, quantity, etc.
- **Review Selection:** Reviews discuss different product features such as age or weight appropriateness, long-term us-

[1] http://jmcauley.ucsd.edu/data/amazon/
[2] http://www.alchemyapi.com/

ASIN	Title	Manufacturer	Accumulated	Rating	O-Sentiment	T-Sentiment
B00012CHFI	The First Years Close And Secure Sleeper	0 Months-2 Years	0 Months-6 Months	0 Months-6 Months	0 Months-6 Months	2 Months-4 Months
B000067EH7	The First Years Deluxe Newborn To Toddler Tub	0 Months-2 Years	0 Months-1 Year	0 Months-1 Year	0 Months-1 Year	0 Months-11 Months
B0009QYTIE	Summer Infant Swaddleme Microfleece	0 Months-3 Months	0 Months-9 Months	0 Months-9 Months	1 Month-9 Months	0 Months-9 Months
B00009ZIKH	Manhattan Toy Infant Mobile for Cribs	0 Months-2 Years	1 Month-9 Months	1 Month 9 Months	1 Month 7 Months	1 Month 4 Months

Table 1: Manufacturer vs Estimated Age Ranges for Selected Products

ability, noise and price. Terms extracted from product reviews are used to map features to products. In this work, we concentrate on *age appropriateness*, therefore, we select a subset of the terms related to *age appropriateness* based on the type of the entities and a simple keyword selection mechanism. More specifically, entities that are of type *quantity* are selected. Within them, we select only the terms that have an age related substring such as *months, years, old, age*, etc.

- **Entity Value Normalization:** Terms must to be processed and normalized in order to obtain the uniform age category keywords. This step includes several preprocessing tasks such as:
  - *conversion of strings to integers*; ex: "three years" to "3 years"
  - *unification of strings*; ex: "mnts" to "month", "yrs" to "years"
  - *case normilization*
  - *separating value and unit parts*: This step results in forming (value, unit) pairs; ex: "3-years", "3years" to pair (value:3 unit:year)
  - *conversion between units* This step is needed because some terms can be expressed with different pairs like (value:1 unit:year) and (value:12 unit:month).

  This phase results in a list of *(value, unit, converted_value)* triples for each term in reviews.
- **Relevance Classification:** Phrases exist in the reviews that contain time related quantities that are not associated with the age of a person such as, "it broke 1 month later" or "I purchased this 10 years ago". As a result we train a naive Bayes classifier in order to differentiate between relevant and non-relevant phrases.
- **Outlier Detection:** Phrases related to the age of a person but not related to age range of a product exist in reviews which generally occur as outliers in the data. An example of such a review is "*My LO is 5 months and has no problem at all using this cup himself! I used it for my girls, now 12 and 10-years-old.*." for a product of age range 0-12 months. *Tukey's range test* is used to remove outliers using 10% and 90% as the lower and upper quartiles respectively.
- **Sentiment Analysis:** The sentiment is extracted in two ways, specifically; *review_sentiment* and *term_sentiment*. *Review_sentiment* reflects the sentiment of the whole review whereas *term_sentiment* represents the sentiment of the sentence where the age_related term is mentioned. The intuition is that a review can have a positive sentiment as a whole however a user may have an issue with the *age appropriateness* of a product. Alchemy API was used to annotate the text sentiment.

Estimating Age Target Group

The distribution of ages along with their sentiment are processed in order to identify a "recommended age" label. Depending on the data we base our estimations on, which refer to the four variations of the system mentioned in the previous section, we conducted four different analyses to predict age labels: *accumulation-based, rating-based, overall sentiment-based and term sentiment-based*. For the predictions, we calculated the *median* value so as to measure the central tendency of the data and we draw the appropriate age range as the lowest and highest terms. The intuition behind is that, during outlier detection, we have already removed the values far beyond the general tendency, therefore all the remaining terms give feedback around appropriate age range.

For *accumulation-based* predictions, we used all the terms that exist for that product to draw ranges. For *rating-based* predictions, we used the age terms mentioned in reviews that are positively rated. *Sentiment-based* predictions were drawn using a similar methodology, where instead of positively rated reviews' terms, we considered only terms with positive sentiment scores for both review-based and term-based analysis.

Evaluation

We evaluated four variations of our system to support the users decision making process.

- **Accumulation**: Here we aggregate reviews based on the ages mentioned in the review text. We simply show a distribution of the ages discussed in the reviews.
- **Rating**: We aggregate reviews based on the ages mentioned however we distinguish between positive and negative reviews. Reviews with 1-2 stars are considered negative. 3 stars is neutral and 4-5 stars are positive.
- **Overall Sentiment**: Reviews are aggregated based on the sentiment expressed in the review. Sentiment scores range between [-1,1] where 0 represents neutral case. We count sentiment scores greater than or equal to 0.2 as positive

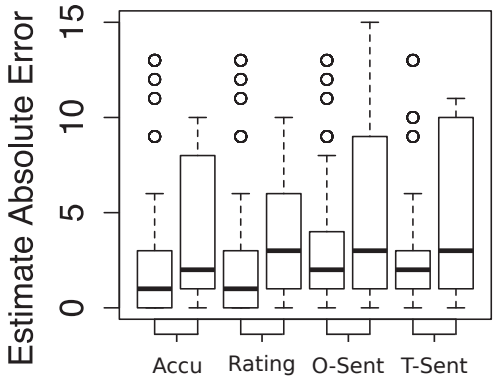

Figure 1: Absolute Error Values

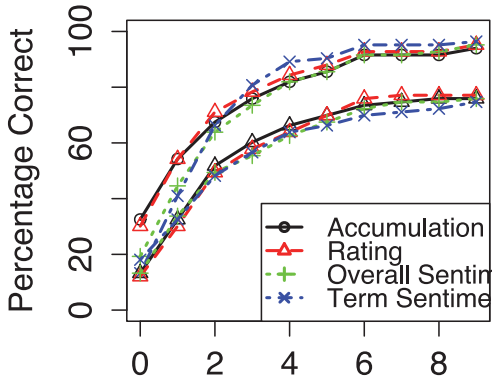

Figure 2: Percentage Accuracy

and less than or equal to -0.2 as negative. Values in range (-0.2,0.2) are counted as neutral.

- **Term Sentiment**: Reviews are aggregated based on the sentiment of the specific sentence which mentioned the age entity to cover examples such as: "I bought this for my 5 year old and she didn't seem very interested. But it is a great product so maybe when she is old." Here, in this example, the review sentiment value is positive but term sentiment value is not positive. As mentioned before, term sentiments provide us more focused feedback behind age appropriateness.

Estimating Age Target Group

In order to evaluate our age estimation method we manually annotated 200 products with the information that could be found on the Amazon product web-page. This information was sometimes included in the structured product text and sometimes within the more general product description. 60% of these reviews did not have age related ranges for the product. We use the remaining reviews to evaluate the accuracy of our proposed method to assess the recommended age for a product.

In order to measure the accuracy the ages are separated into a sequential values increasing in the order of a month before one year and year by year subsequently (e.g. 10-Months, 11-Months, 1-Year, 2-Years). The estimated max and min age range are compared with the annotated values by calculated the distance between the estimate and labelled value in the age sequence. Figure 1 shows the absolute error for the different strategies. As we can be seen the median error for all strategies whether estimating the minimum (the left) or the maximum (the right) age range values is 3 or less. The error increases somewhat when estimating the maximum. The strategies that employ sentiment both show a slight increase in error of 2 and 3 when estimating the maximum and minimum respectively. We also measure the estimation accuracy by increasing the error tolerance to understand the percentage of the product age ranges we correctly captured shown in figure 2. The upper grouping of lines represent the prediction of the maximum age, the lower grouping represents the minimum. With an error tolerance of 1 age division we get 54% accuracy for estimating the minimum and we perform slightly worse when estimating the maximum with only 33%. We again see that the sentiment strategies have a lower performance but overall the results suggest our approach generates a reasonable estimate of the appropriate age for items. In order to understand whether the estimates potentially reflect a more realistic age range than a manufacturer provided document we examine a number of scenarios where the results do not align.

In Table 1, we see four example products. The "close and secure" co-sleeper suggests it is suitable until 2 years old, however the reviews indicate people stop using it around 6-months which aligns with the recommendations from health organisations, such as the American Academy of Paediatrics, that babies should sleep in close proximity to their parents until they are 6 months (Moon 2016). The second example is a bathtub which manufacturers report to be suitable until 2 years old however one reviewer reported *"My only complaint was this was not any good for me after my baby started standing at almost 1 year."* We also find instances where products are useful for longer than estimated by the manufacturer, for example a swaddle blanket, *"It can be use for until 9 months.I like it bec it is really big lots of room to grow for little babies!"* The term sentiment method captures instances where people liked the product overall but just not for a specific age range such as a crib mobile *"My daughter enjoyed it for a few months but lost interest at about 5 months."*

User Evaluation

In order to allow users to visualise *age appropriateness* rating distribution we generated a number of examples and embedded the age-range graph into a sample Amazon product web-page. We decided to place the distribution next to the existing *ratings distribution* as a source of additional information when browsing a product. The main motivation behind this study is that, reviews are powerful resources for getting an intuition about different product features which cannot be captured in a rating distribution alone.

We conducted a brief user assessment with 25 employees. Participants ages ranged from 20 to 62. Nine women and 16 men participated. Thirteen reported parenting experience, although some participants now have teen-aged or adult chil-

Mean Assessment Statement	(1=Strongly Agree 5=Strongly Disagree)
The age-appropriateness information was easy to understand	Agree (2.24)
I would look at the age-appropriateness information.	Strongly Agree (1.4)
I would trust the age-appropriateness information.	Agree (2.48)
I would use the age-appropriateness information as part of my purchase decision.	Agree (2.12)

Table 2: Mean Responses from User Assessment

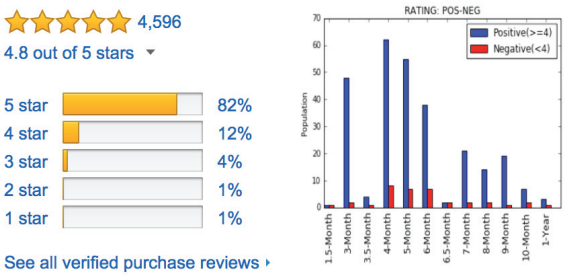

Figure 3: Review Age-Appropriateness Visualisation

dren. We asked the participants to evaluate four statements, using a Likert Scale for agreement (5 steps, from Strongly Agree to Strongly Disagree). Table 2 shows the mean responses. There were no statistically significant differences between parents and non-parents.

While the quantitative results suggest that participants liked the age-appropriateness display, their less formal comments gave us a great deal of additional information. We asked if the position of the display (adjacent to the 1-5-star summary of Customer Reviews) was the right location on the page to present this information. 18 participants (72%) agreed with the page-location; seven participants thought that it should be higher, perhaps (in summarized form) as a badge on the image of the product. We also asked for participants advice about the design of the information display. 13 participants' (52%) urged us to simplify the information. Their specific advice varied, but the common themes were to present only the positive information; to omit no-response or negative information (or to subtract negative from positive); and to provide a link from a simplified display to the full, detailed information in the current display. Four people encouraged us to provide simpler, larger labels. Finally, we asked what additional information should be summarized from the review text. Participants had quite varied suggestions. We report the interesting suggestions here, but we state clearly that none of these suggestions was proposed by a majority of the participants. Thus, as ideas to think about, participants suggested to include the materials-composition and country-of-origin of the product (concerns for chemical safety, ecological cost, and social justice); a bullet list of frequently-occurring phrases in the reviews; or the ability to filter reviews by the duration of the reviewers use of the product. Interestingly, several participants wanted to be able to formulate their own query (e.g., what are the risks for this product?) and receive an immediate summary related to their personalized interests.

Conclusion

We have presented a methodology to harness user generated reviews in order to identify the *age appropriate* range for products on amazon. We have demonstrated that while our results agree to a reasonable percentage with manufacturers recommendations, user experience may sometime differ. As a result, the estimated age range may provide useful information to purchasers and also form the foundation of a recommendation and filtering algorithm which supports age related queries. Our brief user study confirmed that this direction could be useful in the decision making process.

Acknowledgments

We are particularly grateful to Julian McAuley for making his review data available to research.

References

Diao, Q.; Qiu, M.; Wu, C.-Y.; Smola, A. J.; Jiang, J.; and Wang, C. 2014a. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). 193–202.

Diao, Q.; Qiu, M.; Wu, C.-Y.; Smola, A. J.; Jiang, J.; and Wang, C. 2014b. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 193–202. ACM.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM.

Law, D.; Gruss, R.; and Abrahams, A. S. 2017. Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications* 67:84 – 94.

Liu, B. 2012. Sentiment analysis and opinion mining.

McAuley, J., and Leskovec, J. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. 165–172.

McAuley, J., and Yang, A. 2016. Addressing complex and subjective product-related queries with customer reviews. 625–635.

Moon, R. Y. 2016. Sids and other sleep-related infant deaths: Evidence base for 2016 updated recommendations for a safe infant sleeping environment. *Pediatrics* e20162940.

Winkler, M.; Abrahams, A. S.; Gruss, R.; and Ehsani, J. P. 2016. Toy safety surveillance from online reviews. *Decis. Support Syst.* 90(C):23–32.