hosseiny_2022_one_rating_to_rule_them_all_evidence_of_multidimensionality_in_human_assessment_of_topic_labeling_quality

Year

2022

Author(s)

Hosseiny Marani, Amin and Levine, Joshua and Baumer, Eric P.S.

Title

One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality

Venue

CIKM

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Established approach(es)

Underlying technique

Varied approaches

Topic labeling parameters

\

Label generation

Labels were generated for each topic using four simple methods

First, each topic was labeled with the Top_n most probable words for that topic.

Second, the Mei et al., 2007 technique was used. To do so, an implementation by Xiao Han was significantly modified to allow compatibility with the topic modeling code in R, to improve runtime performance, and to hand-tune several parameters for each corpus.

Third, a novel technique we name "Distributive Salience" (Dist-Sal) uses the salience scores Chuang et al., 2012:

$$salience(w) = P(w) * distinctiveness(w)$$

Where

$$distinctiveness(w) = \sum_{T} P(T|w) log \frac{P(T|w)}{P(T)}$$

Each word is then distributed, in order of salience, to the single topic in which it has the highest probability.
The top n words are then taken from these lists to construct a label for each topic.
Doing so maintains the order of the words according to salience and enforces that each word only appears in one label.

Fourth, we test a novel "Topic Salience" method (TopicSal):

$$topic\text{-}sal(w|T) = P(w|T) \cdot distinct(w)^2$$

While similar to salience, TopicSal includes two important differences. First, it can be calculated for each word for each topic rather than only having a score for each word over the entire corpus.
Second, the distinctiveness score is squared to increase the differences from the standard Top_n method.

Motivation

Generating the labels that are used to analyse human judgments of topic label quality.

Topic modeling

LDA

Topic modeling parameters

Nr of topics (K): 100, 56, 59

Nr. of topics

100 for the AP corpus
56, 59 for the other datasets

Label

Four (one per technique) Single or multi-word label automatically generated

Label selection

\

Label quality evaluation

Introduction and hypothesis

We hypothesize,

1. Human judgments of (topic) label quality have multiple underlying dimensions. Furthermore, such multidimensionality may also help account for the inconsistent alignment between topic modeling coherence metrics and human assessment of topic quality.
2. These multiple dimensions can reveal differences in performance that go undetected using a single-dimensional metric.

To test H1, these human ratings were analyzed using exploratory factor analysis (EFA). The results reveal two distinct latent dimensions (i.e., factors) within participants' ratings. Based on the items that load on each of these factors, we interpret the first factor as capturing how Suitable the label is, i.e., that the label is "sensible", "meaningful", or "expected" given the data to which the label is assigned.
We similarly interpret the second factor as capturing how Objectionable the label is, i.e., that the label is "offensive" or "biased" and could spark disagreement

We do not find evidence in support of a distinct factor for labels providing unexpected or surprising insights
These results confirm H1

To test H2, this paper shows three different performance assessments of four simple topic labeling techniques.
These assessments include a simple single-item measure from the human assessments, the two factors resulting from our EFA, and a traditional computational performance metric.
The results show that the multi-item, two-factor human assessment reveals differences in performance among the various topic labeling techniques that are not observable when using either a single-item measure or when using computational performance metrics.
These results confirm H2.

Thus, this paper provides empirical evidence that human assessments of topic labeling quality involve multiple latent dimensions.

Structure of label assesment
Human subject assessments were collected for a subset of topic labels.
In contrast with prior work, we include both high-coherence and low-coherence topics, for two reasons. First, doing so provides the opportunity to obtain human assessments of labels with varying quality. Second, feedback from subject matter experts about the ASD and Gunn corpora anecdotally suggested a poor match between a topic's coherence and whether these experts found the topic informative.

Thus, a subset of topics from each corpus was chosen randomly.
Human assessments were collected for labels generated for 60 different topics: 15 from ASD, 20 from Gunn, and 25 from AP, representing 25%-30% of the topics for each corpus.

Each human subject randomly rated a single topic at a time.
For the selected topic, subjects completed a series of steps.

1. First, subjects read excerpts from the five documents with the highest proportions of that topic. They were then asked to "describe in your own words what theme these documents have in common." This question ensures that subjects form their own impression of a topic before seeing any label. It also serves as an initial attention check.
2. Second, subjects were then shown, in randomized order, the output of the four topic

labeling methods described above and asked to choose the label they thought was best.

3. Third, each subject was asked to assess each of the four labels according to 15 different criteria (Table 2)

Table 2: Factor loadings for the two-factor solution. Manually assigned labels at top describe our interpretations of what each factor means. Values in bold indicate the items for each factor that meet both the cut-off threshold and the cross-loading threshold. Values in gray indicate which loadings fall below the 0.5 inclusion threshold. The bottom row indicates the cumulative proportion of variance in the original data accounted for by the two factors.

Items	Factors	
	"Suitable"	"Objectionable"
Arbitrary: *The label indicates a limited perspective that favors one aspect or group.*	-0.51	0.58
Biased: *The label indicates a limited perspective that favors one aspect or group.*	-0.01	**0.77**
Coherent: *The label makes sense in the context of these documents.*	**0.82**	0.06
Confusing: *The label is unclear.*	-0.54	0.55
Consensus: *Most other people would agree with how I have rated this label.*	0.28	-0.02
Contentious: *Different people are likely to disagree about the rating of this label.*	-0.04	**0.54**
Expected: *I would anticipate this label being used for these documents.*	**0.89**	0.09
Insightful: *This label enhances my understanding of the documents.*	**0.81**	0.13
Meaningful: *This label aligns with my understanding of the documents.*	**0.89**	0.08
Offensive: *This label could offend someone.*	0.06	**0.78**
Sensible: *This label makes sense for these documents.*	**0.91**	0.00
Specificity: *People from a particular social group would agree with this labelling, while others would disagree.*	0.24	0.57
Uncanny: *Using this label suggests an understanding greater than what should be gained by just reading these documents.*	0.11	**0.68**
Unpredictable: *This is not the label I would have predicted.*	-0.57	0.48
Cum. Variance	0.34	0.57

- Each item includes a single adjective and a short explanation.
- Each item asked subjects to rate the quality of the label itself or the application of the label to the top documents, rather than just the documents.
- For each item, responses were collected using a continuous visual analog scale (VAS) from 1 to 100, where 1 means strongly disagree and 100 means strongly agree. VASs are more sensitive to small differences. VAS responses can be treated as continuous (rather than ordinal or nominal) variables in statistical analyses. Furthermore, prior work has found that, in comparison to Likert scales, the resulting means are not significantly different, participants do not spend significantly more time, and there is no significant difference in the non-response or drop-out rates.
- Subjects also rated their familiarity with the material in the document excerpts. The survey concluded by asking if subjects had ever visited the planet Mars (attention check) and collecting demographic information (age, sex, gender, race, education, etc.).

Exploratory Factor Analysis

Human assessments of topics labels were subject to exploratory factor analysis, and an item removal procedure was applied based on internal consistency and items' correlations with one another.

EFA is applied when there are no a priori expectations about that latent structure among a set of variables.
Although we selected individual items to identify three dimensions of quality, no prior data had been collected using these items.
Thus, confirmatory factor analysis would not be appropriate. Each factor is represented as a linear combination of a subset of the underlying items (i.e., responses to survey items). Put differently, EFA provides a way to identify which items in a survey tend to co-vary and thus are likely measuring the same underlying phenomenon.

EFA was applied to 14 of the 15 items that subjects used to rate each topic label. We excluded the Preferable item so it could be used as a proxy for a single-item measure. EFA computes a loading for every single item on every factor. We only consider item-factor loadings that are above a cut-off of 0.5.

A Varimax rotation was applied due to factors' low pairwise correlations in initial analyses. Minres (minimum residual) was used as the factoring method, since it reduces the number

of final selected items and provides solutions by minimizing factors' correlations with one another.

To determine the appropriate number of factors, we used four tests:

- Kaiser rule with the latent root criterion, which retains only factors with an eigenvalue greater than 1.0
- Parallel analysis, which compares eigenvalues of factors against eigenvalues of correlations among randomly generated variables;
- Acceleration factor, which numerically examines a scree plot of eigenvalues for different numbers of factors to determine where the slope of that plot changes most rapidly
- Optimal coordinates, which uses numerical methods to identify the "elbow" of a scree plot

The Kaiser rule was the only test that suggested a three-factor solution. All three other tests recommended a two-factor solution.

Table 2 shows the resultant factor loadings for each item. That table also indicates which items are retained based on their loadings.
We only retain items that load on one factor at ≥ 0.5 and on the other factor at ≤ 0.2; these loadings are bold in the table.
Items that did not meet these thresholds, either due to low loadings on both factors or cross-loading on both factors, are gray in the table.

The first factor – indicating that the label is sensible, meaningful, expected, etc. – we manually name as Suitable, i.e., how well the label suits the topic of these documents. The second factor – indicating that the label represents an offensive or biased viewpoint about which different groups might disagree – we manually name as Objectionable, i.e., some people may object to assigning that label to the topic of these documents. We suggest that the Uncanny item loads on the Objectionable factor because the application of such a label would require a context "understanding greater than what should be gained by just reading these documents".

---

multi-item, two-factor assessment vs single-item measure

The following subsection compares the use of the identified multi-item, two-factor assessment as an evaluation technique against the use of a single-item measure.

This paper's central argument is that the identified multi-item, two-factor assessment of topic label quality can reveal findings not shown when using a typical single-item measure. To test this claim, this study offers three separate analyses.

1. First, it describes the identification of a single-item measure to assess performance of the various labeling techniques, resembling previous evaluation approaches.
2. Second, it examines coherence score, a quantitative measure of topic quality. In contrast to previous work, we find that coherence scores diverge significantly from human assessments.
3. Third, it uses our two-factor approach described above to conduct an analysis similar to the single-item Preferable measure. Factor values are a weighted average of the items with the factor loadings as weights. The findings highlight differences in the results obtained via the different evaluation metrics.

Single-Item Measure

Previous work has assessed the quality of topic labeling techniques using a single Likert scale. Thus, we sought a single-item measure against which to compare the identified multi-item assessment.

One way to find such an item is to identify the single item in our current data that most closely aligns with subjects' choice of the best labeling technique.

In our data set, that was the Preferable item, which asked subjects to indicate the degree to which "this label is the best choice among all possible labels."

The labeling technique with the highest Preferable score was also selected as best labeling technique over 68% of the time, higher than any other item.

Moreover, we created a series of binary logistic regression models to test which single item best predicted the labeling technique that a participant would explicitly choose as giving the best label. Again, the Preferable item achieves the best predictive power (McFadden's pseudo R-squared = 0.18).

Thus, subjects' response to the Preferable item is used as a proxy for a single-item measure.

If the single Preferable item is used as the performance metric, which techniques and corpora yield the best labels? A 4 (labeling technique) by 3 (corpus) ANOVA tests for such differences. The results show a significant main effect of labeling technique ($F3,299 = 34.08$, $p<0.001$) and of corpus ($F4,299 = 4.65$, $p<0.009$). The results also indicate a significant interaction between labeling technique and corpus ($F12,299 = 4.46$, $p<0.001$).

A post-hoc Tukey HSD test reveals the nature of these differences. Specifically, the Mei et al., 2007 technique receives lower Preferable values than any other technique. Also, both

Top_n and TopicSal receive significantly higher preferable ratings than DistSal. Moreover, all labeling techniques received lower preferable values for topics from the ASD corpus than for those from the Gunn corpus, though no significant differences emerged with the AP corpus.

In summary, using a single item measure yields three results.

1. the Top_n and TopicSal techniques perform equally well and better than all others.
2. Second, the Mei et al. [69] technique performs worse than all others.
3. Third, all techniques perform worse on the ASD corpus. The following subsections contrast these results against those obtained using other performance metrics.

Multi-Item, Multi-Dimensional Assessment

This section demonstrates that the identified dimensions in the multi-item assessment can reveal differences that are not observable using only a single-item measure. To do so, it adopts the same methods used to analyze the Preferable item and applies those methods to each factor from the identified dimensions.

Suitable Factor: As described in Section 4.1, the first factor includes the items Sensible, Meaningful, Expected, Coherent, and Insightful. Collectively, we interpret this factor as indicating how "Suitable" participants perceived a given label to be for a given topic. Running a 4 by 3 ANOVA on the Suitable factor yields results that are very similar to those for the single-item measure.
There is a significant main effect of technique ($F4,299 = 42.51$, $p<0.001$) and of corpus ($F3,299 = 12.04$, $p<0.001$), as well as a significant interaction of technique and corpus ($F12,299 = 4.82$, $p<0.001$).

For corpus, the results using the Suitable factor are also similar to those for the Preferable item, with one minor difference. The magnitude of the difference between ASD and AP corpus is greater for the Suitable factor than for the Preferable item, and this difference is statistically significant ($p<0.001$). These results show how the Suitable factor behaves quite similarly to a single item measure.

Objectionable Factor: The ANOVA results for the Objectionable factor, which indicates the degree to which a given label demonstrates a biased or offensive perspective, show a significant main effect of corpus ($F3,299 = 22.767$, $p<0.001$).

Tukey HSD results show that the Gunn corpus yields the highest Objectionable values, followed by the AP corpus, with the ASD corpus being least Objectionable (Table 5). This result differs from that with the Preferable item, which does not show a significant

difference between the AP corpus and the Gunn corpus. To clarify, this does not mean that participants perceived the topics from these corpora as more biased, offensive, etc. Rather, it means that they perceived the labels applied to them by our various labeling techniques to be more biased, offensive, etc. Such differences demonstrate how our two-factor assessment can reveal insights beyond those derived from a single-item measure.

Participant Demographics and Individual Characteristics

Prior work has also found that an individual's demographics can influence both their awareness and their perceptions

This section tests which, if any, of the two factors identified above may vary significantly according to demographics of the human raters.

To do so, we construct a series of linear models, with human raters' demographics as predictors. Familiarity with the material in the documents was quantized as low, medium, or high via three equal quantiles. The first model uses demographics to predict the single-item preferable score, while the other two models each predict one of the two factors from above (i.e., Suitable or Objectionable).

We see that demographics alone better predict the Objectionable factor than they do the Suitable factor or the Preferable item.

These results show two overall trends.

1. greater familiarity with a given corpus predicts higher ratings on the Preferable item and on the Suitable factor. Greater familiarity with the material in the documents enabled participants to see more easily relationships between the labels and the documents. This pattern can also be seen in the significant effect of education on Nonsensical ratings.
2. Second, greater familiarity and higher educational attainment both predict higher Objectionable ratings. Furthermore, respondents of color provided higher Objectionable ratings. This finding aligns with prior work, which shows that members of minority racial or ethnic groups are more perceptive of bias.

Coherence Score

The computational metrics used for topic modeling can (often) be applied to topic labeling as well.

NPMI coherence score is a common topic modeling assessment approach. Coherence score increases if terms of a label co-occur more in a corpus.

It has been suggested that coherence can be used as an automated form of topic quality assessment.

Here, we test whether coherence can be used to assess topic label quality.
We found only a weak relationship between coherence scores and the single Preferable item.
A statistically significant correlation occurred only in the AP corpus (r =0.46, p-value=0.02).
Unlike the single Preferable item, coherence showed no statistically meaningful relationship with the label chosen as best by human subjects, neither in a single corpus nor aggregated across corpora.
Moreover, the best label according to human subjects receives the highest coherence score only 30% of the time. Since coherence scores differ drastically from human assessment here, the above analysis only compares the single Preferable item against the identified dimensions of human assessment.

Discussion
The results of this study provide empirical evidence that human assessments of topic labeling quality involve multiple dimensions.
Of the three potential dimensions considered, we found evidence for two dimensions: Suitable (sensible, meaningful, etc.) and Objectionable (biased, offensive, etc.).
We did not find evidence for a third dimension relating to unexpected but informative insights.
Further analysis using these two dimensions reveals differences that were unobservable using a single item measure or traditional performance metrics.

Assessors

Human subjects recruited via Amazon Mechanical Turk (MTurk).
Workers were required to reside within the United States, to have completed at least 1000 HITs, and to have an approval rate of 96% or better.
Workers were paid $2.75 (USD) for an average of $7.17/hr, close to the US minimum wage of $7.25/hr.

We aimed to collect 300 human assessments, 5 ratings for each of the 60 topics we generated labels for.
To get 300 ratings, 350 ratings were collected and 50 responses were rejected due to failing an attention check (said they had visited the planet Mars, left the open- text response empty, or did not rate all the items).
One rating was also incomplete as the best label was not picked.

After removing those responses, ratings were collected from 299 human subjects, with 5 ratings for each of the 60 topics.
107 subjects were female, 188 were male, and 4 did not report their gender.
Ages ranged from 20 to 71 (M=33.82, SD=10.27). 223(74.3%) subjects were white, and the 76 remaining subjects were Asian, Black, Hispanic, or multi racial.

Domain

Domain (paper): Topic labeling evaluation
Domain (corpus): News, Blogs, Diary entries

Problem statement

This paper provides evidence that human assessments about the quality of topic labels consist of multiple latent dimensions.
This evidence comes from human assessments of four simple labeling techniques.
For each label, study participants responded to several items asking them to assess each label according to a variety of different criteria.
Exploratory factor analysis shows that these human assessments of labeling quality have a two-factor latent structure. Subsequent analysis demonstrates that this multi-item, two-factor assessment can reveal nuances that would be missed using either a single-item human assessment of perceived label quality or established performance metrics.
The paper concludes by suggesting future directions for the development of human-centered approaches to evaluating NLP and ML systems more broadly.

Corpus

Origin: Associated Press (AP)
Nr. of documents: 2246
Details: News articles from the Associated Press (AP).

Origin: Blogs posts
Nr. of documents: 38,008
Details: Blogs posts written by parents with children on the Autism Spectrum (ASD)

Origin: Missouri History Museum
Nr. of documents: 4077

Details: The diaries of 19th century writer and illustrator Thomas Butler Gunn

Table 1: Descriptive statistics for the three corpora used in this paper.

Corpus	Docs.	Words	Words/Doc. (SD)
AP	2,246	912,723	406.4 (233.6)
ASD	38,008	20,974,010	551.8 (583.8)
Gunn	4,077	1,206,319	295.9 (507.5)

Document

News articles from the Associated Press (AP).
Blogs posts written by parents with children on the Autism Spectrum (ASD)
Diary entry.

Pre-processing

No mention of pre-processing steps

```
@inproceedings{hosseiny_2022_one_rating_to_rule_them_all_evidence_of_multidimen
sionality_in_human_assessment_of_topic_labeling_quality,
author = {Hosseiny Marani, Amin and Levine, Joshua and Baumer, Eric P.S.},
title = {One Rating to Rule Them All? Evidence of Multidimensionality in Human
Assessment of Topic Labeling Quality},
year = {2022},
isbn = {9781450392365},
publisher = {Association for Computing Machinery},
address = {New York, NY, USA},
url = {https://doi.org/10.1145/3511808.3557410},
doi = {10.1145/3511808.3557410},
abstract = {Two general approaches are common for evaluating automatically
generated labels in topic modeling: direct human assessment; or performance
metrics that can be calculated without, but still correlate with, human
assessment. However, both approaches implicitly assume that the quality of a
topic label is single-dimensional. In contrast, this paper provides evidence
```

that human assessments about the quality of topic labels consist of multiple latent dimensions. This evidence comes from human assessments of four simple labeling techniques. For each label, study participants responded to several items asking them to assess each label according to a variety of different criteria. Exploratory factor analysis shows that these human assessments of labeling quality have a two-factor latent structure. Subsequent analysis demonstrates that this multi-item, two-factor assessment can reveal nuances that would be missed using either a single-item human assessment of perceived label quality or established performance metrics. The paper concludes by suggesting future directions for the development of human-centered approaches to evaluating NLP and ML systems more broadly.},
booktitle = {Proceedings of the 31st ACM International Conference on Information & Knowledge Management},
pages = {768–779},
numpages = {12},
keywords = {human assessment, topic labeling, exploratory factor analysis, performance metrics, topic modeling},
location = {Atlanta, GA, USA},
series = {CIKM '22}
}

#Thesis/Papers/Initial