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

Amin Hosseiny Marani
Lehigh University, CSE Department
Bethlehem, U.S
amh418@lehig.edu

Joshua Levine
University of Massachusetts Amherst
Amherst, U.S
joshualevine@umass.edu

Eric P. S. Baumer
Lehigh University, CSE Department
Bethlehem, U.S
ericpsb@lehigh.edu

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.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Topic modeling*.

KEYWORDS

topic modeling; topic labeling; performance metrics; human assessment; exploratory factor analysis.

ACM Reference Format:
Amin Hosseiny Marani, Joshua Levine, and Eric P. S. Baumer. 2022. One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3511808.3557410

1 INTRODUCTION

Machine learning (ML) and artificial intelligence (AI) algorithms are judged primarily by their ability to produce the best results, i.e., how well they perform. For labeling tasks, the "best" algorithm can be determined using numerous varied performance metrics – accuracy, precision, recall, AUC, and many others [31]. In unsupervised learning, examples range from silhouette coefficient for clustering [84] to coherence for topic modeling [83]. Across both supervised and unsupervised cases, performance metrics essentially provide an assessment of how well the model fits the data.

While valuable in some contexts, such an orientation gives rise to at least two distinct issues. First, many kinds of information about humans in which computing researchers are interested – sentiment [46, 76], social tie strength [18, 34, 50], politeness [27, 42], and others – involve a significant degree of subjective judgment. Even when leveraging techniques such as inter-rater reliability, such subjectivity calls into question the viability of establishing a definitive "correct" value as required for computing these performance metrics.

Second, machine learning metrics implicitly assume a single dimension of performance. While some metrics consider tradeoffs—precision vs. recall, sensitivity vs. specificity, optimizing multiple constraints, etc.—the machine-assigned label is still seen as either correct or incorrect, i.e., either a good fit or a bad fit. Even prior work involving human assessments of such labeling often employs a single-item scale [25, 42, 56, 59, 69, 75, 77]. However, significant work suggests that many human phenomena have multiple underlying dimensions. For constructs ranging from personality [66, 67], to privacy [16, 62, 90], to emotion [85, 99], social science research has repeatedly shown that multi-item, multi-dimensional measures can provide more robust assessments.

The same point may apply when assessing the quality of machine labeling. For example, results from classification or search tasks may be accurate from an information retrieval perspective while simultaneously embodying biases along gender, racial, political, or other lines [5, 40, 55, 80, 81, 106]. Machine learning techniques will sometimes assign labels or identify patterns that, while initially unexpected or confusing, can ultimately provide surprising insights [2, 7, 47, 74]. Thus, we hypothesize, $\mathbf{H_1}$: *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 [20, 45]. Put differently, $\mathbf{H_2}$: *These multiple dimensions can reveal differences in performance that go undetected using a single-dimensional metric.*

To test these hypotheses, this paper presents a study where human participants used a multi-item instrument to assess machine-generated labels in the specific context of topic modeling [12, 13]. Rather than assessing the quality of topics *per se*, participants were asked to rate multiple different machine-generated labels for a given

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557410

topic [21, 69]. To test $\mathbf{H_1}$, 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 [cf. 5, 40, 55, 70, 101]. We do not find evidence in support of a distinct factor for labels providing unexpected or surprising insights [cf. 2, 7, 47, 74]. These results confirm $\mathbf{H_1}$.

To test $\mathbf{H_2}$, 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 [59]. 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 $\mathbf{H_2}$.

Thus, this paper provides empirical evidence that human assessments of topic labeling quality involve multiple latent dimensions. These dimensions align with some expectations informed by prior work around bias and offensiveness [20, 40, 55, 70, 101], but not others around providing unexpected insight [2, 7, 47, 74]. Future work will be necessary to investigate if human assessment of topic quality (i.e., outputs of topic modeling techniques) is multidimensional, to test the multidimensionality of human assessment for other labeling tasks (e.g., image labeling), and to synthesize across such work to develop a validated multidimensional measure for human assessment of machine labeling.

2 RELATED WORK

2.1 Human-Centered Metrics

Various human-centered evaluation metrics have been explored in a variety of domains. For example, recent work has developed a number of techniques for computing fairness [23]. In such contexts, fairness is often operationalized as a measure of the relationships among error rates across different subpopulations [104, 105]. As a complement to such work, Woodruff et al. [101] examined how perceptions of fairness differ between algorithm designers and those who are typically discriminated against by algorithmic systems. Relatedly, Kleinberg et al. [54] argue that machine learning systems should not be evaluated in terms of performance metrics alone but rather in terms of how those systems' predictions correspond to decisions within some larger evaluative framework. In another example, traditional ML metrics, which focus on the correct classification of each individual data point, fare poorly when the result of interest is overall population proportions, e.g., in the context of voting [43].

A related line of work has also emphasized both the value and difficulty of incorporating techniques from natural language processing and related areas into interactive systems [29, 32, 103]. In such contexts, the performance of a model *per se* matters less than its utility and interpretability for human users. To wit, metrics such as F1-score treat all false positives and false negatives equally. However, some false positives may be more egregious in practice than others. As examples, a photo of an African American may be automatically assigned the tag "ape" [40], or a social media year-in-review highlight reel may prominently feature the image of a user's deceased child [70].

These examples, and others, collectively demonstrate the complexities involved in evaluating the performance of automatic labeling more broadly. This paper examines those complexities within a specific context; labeling topics in topic modeling results.

2.2 Topic Model Labeling

Topic modeling [12, 13] can be used to identify latent themes in a large text corpus. Each theme, or "topic," is represented as a probability distribution over the vocabulary of words that appear in the corpus. Each document has a topic distribution that demonstrates the proportion of the generated topics.

While a probability distribution over a vocabulary may be computationally useful, topics can be more readily interpretable when they are assigned semantically meaningful labels. Perhaps the most common labeling approach is simply to take the Top_n most probable terms (e.g. 5, 10, 20, etc.) as a label for each topic [56, 89]. While in many cases they are informative, the top-n terms can sometimes be redundant, meaningless, misleading, too general, or too specific [3, 69]. Rhody [79] describes an example where topic modeling of poetry produced a topic with high probability words including "night," "light," "moon," "stars," and others. Although this topic initially appears to be about night, Rhody suggests that the poems where this topic occurs show that it has a more metaphorical nature. The poet's "use of the tumultuous night sky [...] provides a conceit for the more significant thematic exploration of two artists' struggle with mental illness" [79, para. 8 under Interpreting Models of Figurative Language Texts]. Thus, a number of other methods have been devised to generate labels that may be more informative.

Mei et al. [69] introduced a probabilistic technique to extract the best set of labels out of generated topic terms. First, the proposed method produces candidate labels by identifying high-frequency n-grams. Second, candidate labels are ranked by maximizing mutual information between chosen candidates and topic models, as well as minimizing Kullback-Leibler (KL) divergence for chosen candidates and topic terms distribution. A final selection phase chooses an optimal label based on maximal marginal relevance and pointwise mutual information. This step ensures that each label is as specific to each topic as possible.

In their evaluation, Mei et al. [69] generated topics and labels for a corpus of SIGMOD papers and for a corpus of Associated Press (AP) articles. From their SIGMOD corpus, a topic with high probability words *clustering, clusters, video, dimensional, cluster, partitioning* was given the label "clustering algorithm." In their AP corpus, a topic with high probability words *north, case, trial, iran, documents, walsh* was assigned the label "iran contra." These examples demonstrate the ability of the Mei et al. [69] technique to identify meaningful summary n-grams whose component words may not have occurred in the Top_n words for that topic.

Work in information visualization offers other means of labeling topics. For example, Chuang et al. [21] proposed a distinctiveness

scoring technique to identify the words most specific to a given topic, rather than simply choosing the highest probability words. For a given word w, distinctiveness is computed as the KL divergence between the conditional probability $P(T|w)$ of a topic T given word w and the marginal probability $P(T)$ of a topic T:

$$distinctiveness(w) = \sum_{T} P(T|w) log \frac{P(T|w)}{P(T)} \quad (1)$$

This distinctiveness score is then used to calculate salience for each word [21]:

$$salience(w) = P(w) * distinctiveness(w) \quad (2)$$

Essentially, this technique labels a topic not with the highest probability words but with the words that most distinguish a given topic from other topics. Chuang et al. [21] provide an example from a topic model trained on abstracts of papers from the IEEE InfoVis conference between 1995 and 2010. For one topic, the Top_n label was *visualization, techniques, user, large, paper, users, approach, technique*, while their distinctiveness based label was *visualization, techniques, user, large, users, context, tasks, focus*. The second label exchanges fairly generic terms (e.g., *paper, approach, technique*) for more specific ones (e.g., *context, tasks, focus*), showing more readily how this topic pertains to focus+context visualization techniques. Although closely related to the discriminative scoring scheme from Mei et al. [69], the focus of Chuang et al. [21] on visualization means that they do not provide a direct comparison with other topic labeling techniques.

In the interest of simplicity, this paper avoids techniques that rely on external datasets or knowledge representations [e.g., 3, 11, 58]. Instead, it focuses on techniques that use only the document corpus and the trained topic model.

Most of this prior work has used human evaluation to assess the quality of topic labeling. However, human subjects in those studies were asked to rate a given label on a single, three-point ordinal question [56, 59, 69, 75]. Such coarse-grained, single-item scales may overlook aspects of a label that go beyond correct or incorrect. This paper addresses that short coming by testing a multi-item, multi-dimensional instrument for assessing topic labeling quality.

3 METHODS

This section describes the procedures for topic label generation, as well as the methods for collecting human assessments thereof.

3.1 Generating Topic Labels

To test the dimensionality of human assessments for topic label quality (H_1), three corpora of varying types and sizes were collected. These corpora were then analyzed using the R wrapper [71] for MALLET's [65] implementation of LDA [13], and labels were generated for the resulting topics. The remainder of this section describes each of these steps in detail.

3.1.1 Data: Collecting Three Varied Corpora. We collected three highly varied corpora. *1. News articles from the Associated Press (AP).* Since topic modeling often uses news articles as a test data set [13, 56, 58, 69], including this data set enables comparison with prior work. *2. Blogs posts written by parents with children on the Autism Spectrum (ASD).* This corpus provides a complement to the AP corpus in at least two ways. First, it is highly focused on a single issue: autism. Second, its tone varies widely, from scientific discussion to informal musings. *3. The diaries of 19th century writer and illustrator Thomas Butler Gunn.* These diaries were transcribed by the Missouri History Museum[1] and organized as part of the Vault at Pfaff's project[2]. Again, this corpus differs from the other two in at least two ways. First, the documents' age means their tone, vocabulary, etc. likely differ from either corpus above. Second, the corpus has only one author. These varied corpora increase the chance that the findings of this paper can generalize. Table 1 briefly summarizes each corpus.

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)

3.1.2 Extracting Topics. A separate LDA model was trained on each of the above corpora. The number of topics was chosen via two different processes. For the AP corpus, 100 topics were used to align with prior work [13]. For the other corpora, a parameter sweep was used to train separate models with $n \in [10, 15, 20, ..., 90]$ topics for each corpus. The number of topics that generated the highest coherence [56, 59] was selected. Coherence, a general quality score for topics, can be computed in a variety of ways [83]. In this paper, coherence is computed using normalized pointwise mutual information (NPMI), which has been shown to align most closely with human judgments of topic quality [59]. When multiple different numbers of topics resulted in comparable average coherence scores, these ties were resolved using guidance from domain experts and subjective judgement [7, 37].

3.1.3 Generating Labels. Using these trained models, 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. [69] technique was used. To do so, an implementation by Xiao Han[3] 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" (DistSal) uses the salience scores (Equation 2) Chuang et al. [21]. 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):

$$topic\text{-}sal(w|T) = P(w|T) \cdot distinct(w)^2 \quad (3)$$

While similar to salience (Equation 2), TopicSal includes two important differences. First, it can be calculated for each word for

[1] http://collections.mohistory.org/resource/103591
[2] https://pfaffs.web.lehigh.edu/
[3] https://github.com/xiaohan2012/chowmein

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.

We intentionally avoided more complex methods that rely on external corpora [e.g., 11] or generate natural language sentences [57], thus enabling more direct comparisons.

3.2 Human Assessment via Mechanical Turk

In line with prior work [58, 69], human subject assessments were collected for a subset of topic labels. In contrast with prior work [58], we include both high-coherence and low-coherence topics [59], 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. 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 [9, 26]. 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.

Third, each subject was asked to assess each of the four labels according to 15 different criteria (Table 2). 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) [24, 33] 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 [24, 33].

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.).

Human subjects were recruited via Amazon Mechanical Turk (MTurk). Prior work showed MTurk workers can produce quality data for NLP tasks [91] and are often more demographically diverse than a convenience sample U.S [8]. 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.

3.2.1 Rationale behind the Items Subjects Used to Rate Labels. The selection of items and their descriptions was guided by insights drawn from a variety of related prior work. First, we sought items that would assess in a fairly general sense how well a given label fit the set of documents to which it was applied. In line with prior work on scale development [62, 82, 90, 94], we included multiple synonymous and/or related items: *coherent, expected, meaningful, preferable, sensible*. Each of these items provides a way of indicating that this label fits with these documents.

We complemented this first list with a few reverse coded items: *arbitrary, confusing, unpredictable*. Doing so follows guidance from prior studies [e.g., 15, 94] wherein novel scales include both positively valenced and negatively valenced terms. This alternation of valence increases the chance that participants are attending to each item, since they need to think about whether a given item is positive or negative. Moreover, such items help determine empirically whether human assessments of a label as apt or poor lie on opposite ends of a single dimension or whether such assessments represent two distinct dimensions.

Second, some of the work surveyed above has highlighted how biases can emerge in machine labeling [4, 23, 54, 101]. Thus, we include several items related to these prior findings about human perceptions or assessments of biases: *biased, consensus, contentious, offensive, specificity*. In addition to assessing whether a label might be seen as offensive, these items also indicate whether or not different people would agree with one another about a given label. Such disagreements may go beyond whether the label is simply perceived as offensive to capture the contestational nature of some labels [cf. 41, 96].

Third, machine labeling is used not only for the completion of well-defined tasks. Particularly in social scientific or digital humanities applications [7, 36, 72, 95], machine labeling can also provide a different, novel perspective that complements human readings with labels that are: *insightful, uncanny*. These items may help determine which labels would not have been initially expected but, upon reflection, do make sense and perhaps even open up novel interpretations. Thus, these items provide another avenue by which to assess the utility of a label beyond simple task performance.

The above three aspects of labeling quality are not and should not be seen as exhaustive. Rather, they provide an initial means, informed by prior related work, to test the multidimensionality of quality in human assessment of machine labeling.

4 ANALYSIS AND RESULTS

4.1 Factor Structure

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.

4.1.1 Exploratory Factor Analysis. Exploratory factor analysis (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, as described below (Section 4.2.1). EFA computes a loading for every single item on every factor. Following Hair et al. [38] and Smith et al. [90], 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 [51]. 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 [39].

To determine the appropriate number of factors, we used four tests [19, 38]: Kaiser rule with the latent root criterion, which retains only factors with an eigenvalue greater than 1.0 [52]; parallel analysis, which compares eigenvalues of factors against eigenvalues of correlations among randomly generated variables [44]; 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 [78]; and optimal coordinates, which uses numerical methods to identify the "elbow" of a scree plot [78]. 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. Following Matsunaga [64], 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. As expected, the Varimax rotation results in orthogonal factors with near-zero pairwise correlation.

This solution also has an intuitive interpretation. 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."

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

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.

4.2 Topic Labeling Assessment

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. First, it describes the identification of a single-item measure to assess performance of the various labeling techniques, resembling previous evaluation approaches [3, 11, 69]. Second, it examines coherence score, a quantitative measure of topic quality [56, 59]. In contrast to previous

work, we find that coherence scores diverge significantly from human assessments. 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.

4.2.1 Single-Item Measure. Previous work has assessed the quality of topic labeling techniques using a single Likert scale [3, 11, 69]. 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 ($F_{3,299}$ = 34.08, p<0.001) and of corpus ($F_{4,299}$ = 4.65, p<0.009). The results also indicate a significant interaction between labeling technique and corpus ($F_{12,299}$ = 4.46, p<0.001).

A post-hoc Tukey HSD test reveals the nature of these differences (Table 3). Specifically, the Mei et al. [69] 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. First, the Top_n and TopicSal techniques perform equally well and better than all others. Second, the Mei et al. [69] technique performs worse than all others. Third, all techniques perform worse on the ASD corpus. The following subsections contrast these results against those obtained using other performance metrics.

4.2.2 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 ($F_{4,299}$ = 42.51, p<0.001) and of corpus ($F_{3,299}$ = 12.04, p<0.001), as well as a significant interaction of technique and corpus ($F_{12,299}$ = 4.82, p<0.001). A Tukey HSD test reveals that the results for technique are almost identical to those for the Preferable item (compare Table 4 and Table 3).

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 ($F_{3,299}$ = 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

Table 3: Post-hoc comparison, via Tukey's HSD, of differences between each pair of labeling techniques on the Preferable item. Each row shows, for one pair, which technique receives greater Preferable values, the absolute value of the difference in mean Preferable values, and the significance (p-value) of that difference. The Top_n technique is generally the most Preferable.

	Post-Hoc Comparison			Diff.	p-value
Technique	Top_n	>	Mei et al.	20.77	<0.001***
	Top_n	<	TopicSal	0.24	1.000
	Top_n	>	DistSal	7.52	0.009**
	TopicSal	>	Mei et al.	21.01	<0.001***
	DistSal	>	Mei et al.	13.25	<0.001***
	TopicSal	>	DistSal	7.77	0.006**
Corpus	Gunn	>	AP	1.79	0.633
	AP	>	ASD	4.86	0.059
	Gunn	>	ASD	6.65	0.008**

Note: *p<.05, **p<.01, ***p<.001.

Table 4: Post-hoc comparison, via Tukey's HSD, of differences between each pair of labeling techniques on the Suitable factor. Results are similar to the single-item measure (Table 3), except that differences between the ASD corpus and the other two corpora have larger effect sizes and become more statistically significant.

	Post-Hoc Comparison			Diff.	p-value
Technique	Top_n	>	Mei et al.	20.65	<0.001***
	Top_n	>	TopicSal	1.49	0.867
	Top_n	>	DistSal	5.30	0.031*
	TopicSal	>	Mei et al.	18.03	<0.001***
	DistSal	>	Mei et al.	14.22	<0.001***
	TopicSal	>	DistSal	3.80	0.020*
Corpus	Gunn	>	AP	0.57	0.931
	AP	>	ASD	7.97	<0.001***
	Gunn	>	ASD	7.40	<0.001***

Note: *p<.05, **p<.01, ***p<.001.

Table 5: Post-hoc comparison, via Tukey's HSD, of differences between each pair of labeling techniques on the Objectionable factor. Results differ greatly both from the Suitable factor (Table 4) and from the single-item measure (Table 3). These differences reinforce the interpretation of this factor as measuring a distinct aspect of label quality.

	Post-Hoc Comparison			Diff.	p-value
Technique	Top_n	>	Mei et al.	3.89	0.160
	Top_n	<	TopicSal	1.24	0.911
	Top_n	>	DistSal	0.53	0.993
	TopicSal	>	Mei et al.	2.66	0.488
	DistSal	>	Mei et al.	3.36	0.275
	TopicSal	>	DistSal	0.71	0.981
Corpus	Gunn	>	AP	3.78	<0.001***
	AP	>	ASD	3.94	0.049*
	Gunn	>	ASD	11.32	<0.001***

Note: *p<.05, **p<.01, ***p<.001.

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.

4.2.3 Participant Demographics and Individual Characteristics. The questionnaire items pertaining to labels being biased, offensive, etc. were informed by recent work around bias and fairness in ML and AI [6, 101]. Often, such cases go beyond quantifiable approaches to fairness[23] to incite contentious disagreements about the accuracy or correctenss of an algorithmic system [4, 35, 93]. Many such disparities occur along demographic lines, often related to race and/or gender [4, 6, 48, 60, 93, 98]. Prior work has also found that an individual's demographics can influence both their awareness and their perceptions of such situations [49, 100, 102]. Thus, 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). Models were constructed via forward stepwise selection, resulting in the most parsimonious model. A likelihood ratio test shows that there is no significant difference between the final selected model and the full model. Table 6 shows the coefficients, significance values, and R-squared for each of the three models. 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. First, greater familiarity with a given corpus predicts higher ratings on the Preferable item and on the Suitable factor. Similar to Sen et al. [88], greater familiarity with the material in the documents enabled participants to see more easily relationships between the labels and the documents.

Table 6: How each demographic item (left column) predicts changes in each performance metric (top row). Each cell shows coefficients from the selected model, i.e., the effect that a one-unit increase in each predictor has on each performance metric (or, in the case of a categorical variable, the effect of a change from the reference level). Cells with "n/a" were excluded during model selection. The last row of the table shows the R-squared of each model.

Metric	Preferable	Suitable	Objectionable
Familiarity (low)	-7.74***	-6.12***	-4.15**
Familiarity (high)	6.29**	6.37***	12.30***
Married	5.63**	5.79***	7.21***
Education	n/a	n/a	11.30***
Person of Color	n/a	n/a	9.03***
Birth Sex (Male/Man)	n/a	n/a	2.67***
R-squared	0.045***	0.056***	0.241***

Note: *p<.05, **p<.01, ***p<.001.

This pattern can also be seen in the significant effect of education on Nonsensical ratings. It is unclear, though, why being married would predict higher ratings on all metrics. Similarly, prior work provides little insight as to why male respondents would rate labels as more Objectionable [49, 100, 102].

Second, greater familiarity and higher educational attainment both predict higher Objectionable ratings. Likely, those with greater familiarity or more education are more perceptive of the potential for bias [100, 102]. Furthermore, respondents of color (those who indicated a racial category other than white) 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 [49, 100–102]. It also reinforces our interpretation of the Objectionable factor as indicating a limited or biased perspective.

4.2.4 Coherence Score. Assessing topic modeling can be done through various computational techniques; e.g., silhouette [63, 68], UMass and normalized pointwise mutual information (NPMI) coherence score [59, 83]. These computational metrics can be applied to topic labeling as well, as long as labels are set of terms [1, 28]. Normalized Pointwise Mutual Information (NPMI) Coherence score is a common topic modeling assessment approach. NPMI coherence score captures the mutual information of pairs of words. Coherence score increases if terms of a label co-occur more in a corpus.

Prior work has found measures of topic coherence that align well with human assessments of topic quality [30, 56, 59, 75]. Thus, 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). Contrast this with the correlations found by Lau et al. [59] of r=0.6 and higher. 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.

5 DISCUSSION

The results of this study provide empirical evidence that human assessments of topic labeling quality involve multiple dimensions ($\mathbf{H_1}$). 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 ($\mathbf{H_2}$).

This discussion considers in greater detail the relationship between single-dimension metrics and multiple dimensions of performance. It then offers both interpretations of these results and implications for research on human perceptions of topic modeling and of machine labeling more broadly. It concludes by noting how this study's limitations could be addressed in future work.

5.1 A Single Metric vs. Multiple Dimensions

The above results highlight several differences between a single metric (either human-assigned or computational) and multiple dimensions of human assessment. According to the single Preferable item, the difference between labels for topics from the AP corpus and labels for topics from the ASD corpus only approaches significance (diff=4.86, p=0.059). However, according to the Suitable factor, that same difference is both larger and statistically significant (diff=7.97, p<0.001). Furthermore, the single Preferable item shows no difference between labels for topics from the Gunn corpus and those from the AP corpus (diff=1.79, p=0.633), but the Objectionable factor does reveal a significant difference (diff=3.78, p<0.001).

In line with recent work [45], there was little relationship between human assessments and NPMI coherence scores [30, 56, 59, 75], despite the latter often being used as a quantitative metric for topic quality. Put differently, the identified dimensions reveal differences that were not detectable using only a single item measure or using existing performance metrics ($\mathbf{H_2}$). These previously undetectable difference help understand how topic labeling techniques perform differently across different corpora.

Finally, we identify how individual characteristics of the human raters relate to their assessments of labels. For example, participants who had less familiarity with the content in the documents rated labels as less Suitable and less Objectionable. Also, participants who identified as persons of color were more likely to perceive labels as more Objectionable than were white participants. Such results offer another example of differences that can be observed using multidimensional assessment but not using a single-item measure.

5.2 Interpretation and Implications

We now consider potential explanations for and interpretations of these findings. Some of the most notable results relate to the Objectionable factor. For example, labels for topics from the Gunn diaries were perceived as more Objectionable than labels for topics from the other corpora (Table 5). This difference may have occurred in part because Gunn's diaries often included judgmental and, at times, disparaging remarks about other people. Such remarks may have led participants to perceive all labels for a given topic as biased, offensive, etc., regardless of the computational technique used to generate the label. Also, labels for topics from the AP corpus were seen as more Objectionable than those from the ASD corpus (Table 5), perhaps because the issues covered in the AP documents (terrorism, apartheid, etc.) are highly contentious.

As one possible interpretation, human annotators (here, MTurk workers) may have rated the content itself rather than the application of a label. For instance, participants may have rated news articles in the AP corpus higher on the Objectionable factor not because of the topic label that was applied to it but because of the underlying content in those articles themselves (e.g., politics, terrorism). However, the wording of the items themselves emphasized either the label itself or the application of the label to the content (e.g., "Offensive – This label could offend someone."), decreasing the chance that participants were rating the underlying content.

As noted above, the analyses did not provide evidence to support a third dimension for labels that provided unexpected insights. Although the Kaiser rule [52] suggested a three factor solution, manual inspection of a three factor solution found that none of the factors aligned with the expected third dimension. This result may have occurred because the subjects in this study were not comparable to the social science or digital humanities researchers whose work informed the expectation of this third dimension [7, 36, 72, 95]. Including more related items could make this dimension easier to detect. Alternatively, this null result may suggest that the Suitable and Objectionable factors alone sufficiently captured the patterns of human subjects' assessments of topic labeling quality.

Analyses of the demographics of the participants shows respondents of color rated labels higher on the Objectionable factor than did white respondents. As noted above, members of minority groups are often more adept at perceiving bias and thus at noting the potentially disagreeable or offensive nature of a label [49, 100, 102]. Similarly, Woodruff et al. [101] found that members of populations usually discriminated against by algorithmic systems have different views about what constitutes fairness in those systems, and that those views may differ from developers' conceptions of fairness or bias. However, prior work on human assessments of machine labeling quality, e.g., in topic modeling [20, 59, 69, 83], has not considered how such variability among participants can influence their perceptions of quality.

Thus, this work both addresses that gap and contributes to a growing body of findings suggesting that researchers must account more fully for the complexity involved in topic labeling and perhaps topic modeling or other labeling techniques as well [14, 17, 86, 87, 92, 97]. Put differently, we must develop means of assessing performance that go beyond asking whether a given data point is assigned the correct label[4].

That said, existing metrics of topic labeling performance should not be completely abandoned. The relative speed and ease with which ML performance can be assessed helps enable rapid iterative

[4]Hopkins and King [43] make a related argument that correct labeling of each individual data point is often less important than accurate assessments of the overall proportion of each label in a data set. However, they do not focus on subjective human perceptions, neither of those overall proportions nor of individual labels.

improvement. In contrast, conducting a human assessment for every small change (e.g., adjustments to feature extraction, or hyperparameter optimization) would prove both time and cost prohibitive. Instead, developers could focus first on optimizing the ML performance metric(s) of choice first, then conduct a human assessment later on. Doing so at crucial moments in the development process, rather than after a system has been implemented and deployed, may help identify egregiously Objectionable (biased, offensive, etc.) labels before they occur in production systems [4, 5, 40, 93, 106].

As another option, focused studies could examine relationships among these easily calculable performance metrics and more labor- and time-intensive human assessments. For example, significant prior research in topic modeling has examined relationships between human perceptions of quality and various measures of coherence [56, 59, 75], However, in line with more recent work [45], this study found little evidence of correlation between coherence values and human assessments of label quality. Rather than abandon coherence as a quality metric for topic modeling, it may instead be more beneficial for future research to examine what factors influence the correlation between coherence and perceived quality. Might this relationship depend upon, for instance, the kind of corpus being analyzed (modern vs. archaic, formal vs. informal, blogs vs. Wikipedia articles, etc.), attributes of the human assessor (race, gender identity, familiarity with the content, etc.), or other factors? Moreover, as suggested by the results present here, quality may be multifaceted. A given label may seem both well suited yet simultaneously offensive or biased. Some labels or topics may seem initially confusing but, upon further inspection, reveal unanticipated insights [cf. 7]. Put differently, these results suggest that the kind of multi-dimensional, human assessments of performance presented here should be explored further in future work. Doing so will provide a deeper understanding of the relationships between efficiently calculable performance metrics and complex human perceptions of performance.

5.3 Limitations and Future Work

This work's primary limitations include the data sets used, the specific task involved (topic labeling), the dimensions of quality, and the specific annotator population.

First, we intentionally selected corpora with diversity along a variety of dimensions. That said, it would be valuable to attempt replicating these results with different corpora. While the relative performance of different labeling techniques is somewhat informative, perhaps more important is determining whether the underlying factors here generalize to other corpora.

Second, this study tested three specific dimensions of label quality and found evidence supporting two of those dimensions. Although expectations about these dimensions were informed by prior work, it is possible that a variety of other dimensions may be measurable. Put differently, while the results here demonstrate that there *are* different dimensions to human perceptions of topic labeling quality, the results do not imply that these are the *only* measurable dimensions of topic labeling quality. Future work should explore whether there are other dimensions of labeling quality.

Third, we must also keep in mind who exactly the human participants are who provided the assessments. While MTurk workers can provide a convenient and reliable source of annotations [91], such annotations can also differ from those made by experts [88]. This point is echoed by our findings regarding the influence of individual characteristics, especially familiarity, on participants' ratings. Thus, future work is required to understand how assessments of topic labeling – both in terms of relative performance among techniques and in terms of the internal factor structure of those ratings – may differ among different audiences.

Finally, it would be valuable to test whether these dimensions generalize to other kinds of topic modeling or labeling tasks. One could envision similar human subjects assessments of, e.g., sentiment analysis [61], image captioning [25, 53], machine translation [77], political partisanship [22], image manipulation detection [10, 73], or other tasks. Some work has already applied such a perspective to politeness detection [42], though using a simple three-point scale (polite, neutral, and impolite) rather than a multi-item assessment. Such work can help understand more fully the relationship between performance metrics and human assessments.

6 CONCLUSION

This paper presents empirical evidence that human perceptions about topic labeling quality have multiple dimensions ($\mathbf{H_1}$). The identified dimensions align with expectations informed by prior work on potential social aspects of machine labeling [4, 17, 86, 87, 92, 93, 101]. Data were gathered from human subjects rated automatically generated labels from a variety of techniques for labeling topics in topic modeling results [21, 69]. EFA identified two latent dimensions to human assessments: whether a label was *Suitable* and whether it was *Objectionable*.

Analysis of the different topic labeling techniques using these two dimensions reveals findings that would not have been detectable otherwise, neither using a single-item measure nor using traditional computational performance metrics. On the one hand, a single-item measure generally aligns with perceptions of how *Suitable* a label is — that the label is sensible, meaningful, or expected given the text to which it is applied. On the other hand, perceptions of suitability are distinct from perceptions of how *Objectionable* a label is — that the label may be seen as biased, offensive, or likely to spark disagreement when applied to the given text. However, we find that a common computational performance metric (NPMI coherence [59]) diverges from both dimensions of human assessment.

While this paper focuses on topic labeling techniques, future work should test the use of human assessments in topic modeling and in other labeling tasks. [e.g., 10, 22, 25, 53, 73, 77]. Doing so will help assess robustness and generalizability, both of the specific items used to gather ratings from human subjects, and of the two underlying dimensions identified here in human assessments of topic labeling performance.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant #CNS-1757787. Thanks to Lehigh University HCSC group for stimulating discussions and helpful suggestions.

REFERENCES

[1] Nikolaos Aletras and Mark Stevenson. 2014. Labelling topics using unsupervised graph-based methods. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 631–636.

[2] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. 2014. Serendip: Topic Model-Driven Visual Exploration of Text Corpora. In *Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Paris, 173–182. https://doi.org/10.1109/VAST.2014.7042493

[3] Mehdi Allahyari and Krys Kochut. 2015. Automatic topic labeling using ontology-based topic models. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 259–264.

[4] Mike Ananny. 2011. The Curious Connection between Apps for Gay Men and Sex Offenders. *The Atlantic* (April 2011).

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *Pro Publica* (May 2016).

[6] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732.

[7] Eric P. S. Baumer, Drew Siedel, Lena McDonnell, Jiayun Zhong, Patricia Sittikul, and Micki McGee. 2020. Topicalizer: Reframing Core Concepts in Machine Learning Visualization by Co-Designing for Interpretivist Scholarship. *Human–Computer Interaction* 35, 5-6 (April 2020), 452–480. https://doi.org/10.1080/07370024.2020.1734460

[8] Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. 2012. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis* 20, 3 (2012), 351–368.

[9] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. 2014. Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science* 58, 3 (2014), 739–753. https://doi.org/10.1111/ajps.12081

[10] Aparna Bharati, Richa Singh, Mayank Vatsa, and Kevin W. Bowyer. 2016. Detecting Facial Retouching Using Supervised Deep Learning. *IEEE Transactions on Information Forensics and Security* 11, 9 (Sept. 2016), 1903–1913. https://doi.org/10.1109/TIFS.2016.2561898

[11] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. Automatic Labelling of Topics with Neural Embeddings. In *Proceedings of the International Conference on Computational Linguistics (COLING)*. Osaka, Japan, 953–963.

[12] David M. Blei. 2012. Probabilistic Topic Models. *Commun. ACM* 55, 4 (April 2012), 77–84. https://doi.org/10.1145/2133806.2133826

[13] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[14] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems (NIPS)*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 4349–4357.

[15] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*, Patrick W. Jordan, Bruce Thomas, Ian L. McClelland, and Bernard A. Weerdmeester (Eds.). Taylor & Francis, London, 189–194.

[16] Tom Buchanan, Carina Paine, Adam N. Joinson, and Ulf-Dietrich Reips. 2007. Development of Measures of Online Privacy Concern and Protection for Use on the Internet. *Journal of the American Society for Information Science and Technology* 58, 2 (2007), 157–165. https://doi.org/10.1002/asi.20459

[17] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.

[18] Moira Burke and Robert E. Kraut. 2014. Growing Closer on Facebook: Changes in Tie Strength through Social Network Site Use. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM Press, Toronto, ON, 4187–4196. https://doi.org/10.1145/2556288.2557094

[19] Raymond B Cattell. 1966. The scree test for the number of factors. *Multivariate behavioral research* 1, 2 (1966), 245–276.

[20] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems (NIPS)*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates, Inc., 288–296.

[21] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI)*. ACM, Capri Island, Italy, 74–77. https://doi.org/10.1145/2254556.2254572

[22] David E Clementson. 2018. Truth bias and partisan bias in political deception detection. *Journal of Language and Social Psychology* 37, 4 (2018), 407–430.

[23] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (July 2018). https://arxiv.org/abs/1808.00023v2

[24] Mick P. Couper, Roger Tourangeau, Frederick G. Conrad, and Eleanor Singer. 2006. Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment. *Social Science Computer Review* 24, 2 (May 2006), 227–245. https://doi.org/10.1177/0894439305281503

[25] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5804–5812.

[26] Paul G. Curran. 2016. Methods for the Detection of Carelessly Invalid Responses in Survey Data. *Journal of Experimental Social Psychology* 66 (Sept. 2016), 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

[27] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, Sofia, Bulgaria, 250–259.

[28] Heidar Davoudi and Aijun An. 2015. Ontology-based topic labeling and quality prediction. In *International Symposium on Methodologies for Intelligent Systems*. Springer, 171–179.

[29] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning As a Design Material. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Denver, CO, 278–288. https://doi.org/10.1145/3025453.3025739

[30] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Topics in Tweets: A User Study of Topic Coherence Metrics for Twitter Data. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello (Eds.). Springer International Publishing, 492–504.

[31] César Ferri, José Hernández-Orallo, and R Modroiu. 2009. An experimental comparison of performance measures for classification. *Pattern Recognition Letters* 30, 1 (2009), 27–38.

[32] Rebecca Fiebrink and Marco Gillies. 2018. Introduction to the Special Issue on Human-Centered Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2 (June 2018), 7:1–7:7. https://doi.org/10.1145/3205942

[33] Frederik Funke and Ulf-Dietrich Reips. 2012. Why Semantic Differentials in Web-Based Research Should Be Made from Visual Analogue Scales and Not from 5-Point Scales. *Field Methods* 24, 3 (Aug. 2012), 310–327. https://doi.org/10.1177/1525822X12444061

[34] Eric Gilbert and Karrie Karahalios. 2009. Predicting Tie Strength with Social Media. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Boston, 211–220.

[35] Tarleton Gillespie. 2012. Can an Algorithm Be Wrong? *Limn* 1, 2 (Feb. 2012).

[36] Andrew Goldstone and Ted Underwood. 2014. The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History* 45, 3 (Nov. 2014), 359–384. https://doi.org/10.1353/nlh.2014.0025

[37] Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How Many Topics? Stability Analysis for Topic Models. In *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science)*. Springer, Berlin, Heidelberg, 498–513. https://doi.org/10.1007/978-3-662-44848-9_32

[38] J.F. Hair, W.C. Black, B.J. Babin, and R.E. Anderson. 2013. *Multivariate Data Analysis*. Pearson Education Limited. https://books.google.com/books?id=VvXZnQEACAAJ

[39] Harry H Harman and Wayne H Jones. 1966. Factor analysis by minimizing residuals (minres). *Psychometrika* 31, 3 (1966), 351–368.

[40] Alex Hern. 2015. Flickr Faces Complaints over 'offensive' Auto-Tagging for Photos. *The Guardian* (May 2015).

[41] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS) (DIS)*. ACM, Edinburgh, Scotland, UK, 95–99. https://doi.org/10.1145/3064663.3064703

[42] Erin R. Hoffman, David W. McDonald, and Mark Zachry. 2017. Evaluating a Computational Approach to Labeling Politeness: Challenges for the Application of Machine Classification to Social Computing Data. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 52:1–52:14. https://doi.org/10.1145/3134687

[43] Daniel J. Hopkins and Gary King. 2010. A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science* 54, 1 (Jan. 2010), 229–247. https://doi.org/10.1111/j.1540-5907.2009.00428.x

[44] John L. Horn. 1965. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika* 30, 2 (June 1965), 179–185. https://doi.org/10.1007/BF02289447

[45] Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. *Advances in Neural Information Processing Systems* 34 (2021).

[46] C.J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI, Ann Arbor, MI, 216–225.

[47] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration

in Qualitative Analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 23.

[48] June Park John and Martin Carnoy. 2019. The case of computer science education, employment, gender, and race/ethnicity in Silicon Valley, 1980–2015. *Journal of Education and Work* 32, 5 (2019), 421–435.

[49] Rachel L. Johnson, Somnath Saha, Jose J. Arbelaez, Mary Catherine Beach, and Lisa A. Cooper. 2004. Racial and Ethnic Differences in Patient Perceptions of Bias and Cultural Competence in Health Care. *Journal of General Internal Medicine* 19, 2 (Feb. 2004), 101–110. https://doi.org/10.1111/j.1525-1497.2004.30262.x

[50] Jason J. Jones, Jaime E. Settle, Robert M. Bond, Christopher J. Fariss, Cameron Marlow, and James H. Fowler. 2013. Inferring Tie Strength from Online Directed Behavior. *PLOS ONE* 8, 1 (Jan. 2013), e52168. https://doi.org/10.1371/journal.pone.0052168

[51] Henry F Kaiser. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 3 (1958), 187–200.

[52] Henry F. Kaiser. 1960. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement* 20, 1 (April 1960), 141–151. https://doi.org/10.1177/001316446002000116

[53] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[54] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (Feb. 2018), 237–293. https://doi.org/10.1093/qje/qjx032

[55] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland, OR, 417–432. https://doi.org/10.1145/2998181.2998321

[56] Jey Han Lau and Timothy Baldwin. 2016. The Sensitivity of Topic Coherence Evaluation to Topic Cardinality. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics, San Diego, California, 483–487.

[57] Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically Driven Neural Language Model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, BC, 355–365. https://doi.org/10.18653/v1/P17-1033

[58] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic Labelling of Topic Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Portland, OR, 1536–1545.

[59] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics, Gothenburg, Sweden, 530–539.

[60] Philipp Lenssen. 2007. Did You Mean: "He Invented"? http://blogoscoped.com/archive/2007-05-07-n56.html.

[61] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[62] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. 2004. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research* 15, 4 (2004), 336–355.

[63] Mika V Mantyla, Maelick Claes, and Umar Farooq. 2018. Measuring LDA topic stability from clusters of replicated runs. In *Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement*. 1–4.

[64] Masaki Matsunaga. 2010. How to Factor-Analyze Your Data Right: Do's, Don'ts, and How-to's. *International Journal of Psychological Research* 3, 1 (June 2010), 97–110. https://doi.org/10.21500/20112084.854

[65] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

[66] Robert R. McCrae and Paul T. Costa. 1987. Validation of the Five-Factor Model of Personality across Instruments and Observers. *Journal of Personality and Social Psychology* 52, 1 (Jan. 1987), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

[67] Robert R. McCrae and Paul T. Jr. Costa. 1997. Personality Trait Structure as a Human Universal. *American Psychologist* 52, 5 (May 1997), 509–516. https://doi.org/10.1037/0003-066X.52.5.509

[68] Vineet Mehta, Rajmonda S Caceres, and Kevin M Carter. 2014. Evaluating topic quality using model clustering. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, 178–185.

[69] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic Labeling of Multinomial Topic Models. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, San Jose, CA, 490–499.

[70] Eric Meyer. 2014, Dec 24. Inadvertent Algorithmic Cruelty.

[71] David Mimno. 2013. Mallet: A Wrapper around the Java Machine Learning Tool MALLET.

[72] John W. Mohr and Petko Bogdanov. 2013. Introduction—Topic Models: What They Are and Why They Matter. *Poetics* 41, 6 (Dec. 2013), 545–569. https://doi.org/10.1016/j.poetic.2013.10.001

[73] Daniel Moreira, Aparna Bharati, Joel Brogan, Allan Pinto, Michael Parowski, Kevin W. Bowyer, Patrick J. Flynn, Anderson Rocha, and Walter J. Scheirer. 2018. Image Provenance Analysis at Scale. *IEEE Transactions on Image Processing* 27, 12 (Dec. 2018), 6109–6123. https://doi.org/10.1109/TIP.2018.2865674

[74] Laura K. Nelson. 2017. Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research* (Nov. 2017), 0049124117729703. https://doi.org/10.1177/0049124117729703

[75] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*. ACL, Stroudsburg, PA, USA, 100–108.

[76] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135. https://doi.org/10.1561/1500000011

[77] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[78] Gilles Raîche, Theodore A. Walls, David Magis, Martin Riopel, and Jean-Guy Blais. 2013. Non-Graphical Solutions for Cattell's Scree Test. *Methodology* 9, 1 (Jan. 2013), 23–29. https://doi.org/10.1027/1614-2241/a000051

[79] Lisa Marie Rhody. 2013. Topic Modeling and Figurative Language. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

[80] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 148:1–148:22. https://doi.org/10.1145/3274417

[81] Ronald E. Robertson, David Lazer, and Christo Wilson. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the World Wide Web Conference (WWW)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 955–965. https://doi.org/10.1145/3178876.3186143

[82] Erica Robles, Abhay Sukumaran, Kathryn Rickertsen, and Cliff Nass. 2006. Being Watched or Being Special: How I Learned to Stop Worrying and Love Being Monitored, Surveilled, and Assessed. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 831–839.

[83] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the ACM Conference on Web Search and Data Mining (WSDM)*. ACM, Shanghai, China, 399–408. https://doi.org/10.1145/2684822.2685324

[84] Peter J. Rousseeuw. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* 20 (Nov. 1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[85] James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. https://doi.org/10.1037/h0077714

[86] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 144:1–144:33. https://doi.org/10.1145/3359246

[87] Carsten Schwemmer, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. 2020. Diagnosing Gender Bias in Image Recognition Systems. *Socius* 6 (Jan. 2020), 2378023120967171. https://doi.org/10.1177/2378023120967171

[88] Shilad Sen, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao (Ken) Wang, and Brent Hecht. 2015. Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*. ACM Press, Vancouver, BC, 826–838. https://doi.org/10.1145/2675133.2675285

[89] Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Niklas Elmqvist, and Leah Findlater. 2017. Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Labels. *Transactions of the Association for Computational Linguistics* 5 (Jan. 2017), 1–15.

[90] H. Jeff Smith, Sandra J. Milberg, and Sandra J. Burke. 1996. Information Privacy: Measuring Individuals' Concerns about Organizational Practices. *MIS Quarterly* 20, 2 (June 1996), 167. https://doi.org/10.2307/249477

[91] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 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)*. Association for Computational Linguistics, 254–263.

[92] Robyn Speer. 2017. How to Make a Racist AI without Really Trying. http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/.

[93] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Commun. ACM* 56, 5 (2013), 44–54. https://doi.org/10.1145/2460276.2460278

[94] Adam Tapal, Ela Oren, Reuven Dar, and Baruch Eitam. 2017. The Sense of Agency Scale: A Measure of Consciously Perceived Control over One's Mind, Body, and the Immediate Environment. *Frontiers in Psychology* 8 (Sept. 2017). https://doi.org/10.3389/fpsyg.2017.01552

[95] Ted Underwood. 2019. *Distant Horizons.* University of Chicago Press, Chicago.

[96] Kristen Vaccaro, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. Contestability in Algorithmic Systems. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. ACM, Austin, TX, 523–527. https://doi.org/10.1145/3311957.3359435

[97] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5310–5319.

[98] Yi Wang and David Redmiles. 2019. Implicit gender biases in professional software development: An empirical study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, 1–10.

[99] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology* 54, 6 (June 1988), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

[100] Ronald Weitzer and Steven A. Tuch. 2005. Racially Biased Policing: Determinants of Citizen Perceptions. *Social Forces* 83, 3 (March 2005), 1009–1030. https://doi.org/10.1353/sof.2005.0050

[101] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 656:1–656:14. https://doi.org/10.1145/3173574.3174230

[102] Scot Wortley. 1996. Justice for All? Race and Perceptions of Bias in the Ontario Criminal Justice System – A Toronto Survey. *Canadian Journal of Criminology* (Aug. 1996). https://doi.org/10.3138/cjcrim.38.4.439

[103] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)*. ACM, Hong Kong, 585–596. https://doi.org/10.1145/3196709.3196730

[104] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proceedings of the International Conference on World Wide Web (WWW)*. ACM, Perth, Australia, 1171–1180. https://doi.org/10.1145/3038912.3052660

[105] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the International Conference on Machine Learning (ICML)*. 325–333.

[106] Michael Zimmer. 2007. Google: "Did You Mean: 'He Invented'?". http://www.michaelzimmer.org/2007/05/09/google-did-you-mean-he-invented/.