Web Content Classification Using Distributions of Subjective Quality Evaluations

MARIA RAFALAK and DOMINIK DEJA, Polish-Japanese Academy of Information Technology
ADAM WIERZBICKI, Wrocław University of Technology
RADOSŁAW NIELEK, Polish-Japanese Academy of Information Technology
MICHAŁ KĄKOL, Polish-Japanese Academy of Information Technology

Machine learning algorithms and recommender systems trained on human ratings are widely in use today. However, human ratings may be associated with a high level of uncertainty and are subjective, influenced by demographic or psychological factors. We propose a new approach to the design of object classes from human ratings: the use of entire distributions to construct classes. By avoiding aggregation for class definition, our approach loses no information and can deal with highly volatile or conflicting ratings. The approach is based on the concept of the Earth Mover's Distance (EMD), a measure of distance for distributions. We evaluate the proposed approach based on four datasets obtained from diverse Web content or movie quality evaluation services or experiments. We show that clusters discovered in these datasets using the EMD measure are characterized by a consistent and simple interpretation. Quality classes defined using entire rating distributions can be fitted to clusters of distributions in the four datasets using two parameters, resulting in a good overall fit. We also consider the impact of the composition of small samples on the distributions that are the basis of our classification approach. We show that using distributions based on small samples of 10 evaluations is still robust to several demographic and psychological variables. This observation suggests that the proposed approach can be used in practice for quality evaluation, even for highly uncertain and subjective ratings.

CCS Concepts: • **Information systems** → **Crowdsourcing**; • **Human-centered computing** → **Collaborative filtering**; **Social tagging**; **Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Cluster analysis*; *Uncertainty quantification*

Additional Key Words and Phrases: Web content quality, rating distribution, earth mover's distance, classification design, sample composition, robustness

ACM Reference Format:

Maria Rafalak, Dominik Deja, Adam Wierzbicki, Radosław Nielek, and Michał Kąkol. 2016. Web content classification using distributions of subjective quality evaluations. *ACM Trans. Web* 10, 4, Article 21 (November 2016), 30 pages.
DOI: <http://dx.doi.org/10.1145/2994132>

1. INTRODUCTION

Many Web-based systems use quality evaluation that is usually carried out by human evaluators. Examples of systems that have used such an approach include the Article

---

This research is supported by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097[ENGINE].
Authors' addresses: A. Wierzbicki (corresponding author); email: adamw@pjwstk.edu.pl.

This work is licensed under a Creative Commons Attribution-ShareAlike International 4.0 License.

2016 Copyright is held by the owner/author(s).
ACM 1559-1131/2016/11-ART21 \$15.00
DOI: <http://dx.doi.org/10.1145/2994132>

Feedback Tool on Wikipedia, the MovieLens¹ system for evaluating films, and the WOT² system or Reconcile³ system for evaluating credibility. Quality evaluation can also be supported by machine classification approaches that attempt to learn quality ratings and predict the ratings of new content. However, human ratings may be uncertain (i.e., may be characterized by a low level of inter-rater agreement). Ratings may also be subjective, random, biased, or generally unreliable. A method is needed to construct quality classes based on human ratings to recommend content to users and to teach and evaluate machine learning algorithms.

One possible solution to this problem is an attempt to obtain expert ratings that can be used as ground truth. However, this solution has several drawbacks. The first is scalability. While it is possible to imagine using experts in a topical domain, in a case where we are trying to build a system for predicting the credibility (or quality) of general Web content, such an approach must fail. It is extremely rare to construct a community with expert knowledge in all domains around a system (with the possible exception of Wikipedia).

Another problem is that it may be impossible to find experts for some domains. This may be because a domain is too specialized and experts are unavailable, or because the domain is too broad and requires expertise in many areas (and such combined expertise is unavailable). Experts may also be expensive and hard to motivate to participate in crowdsourcing or social recommender systems. Last but not least, experts may disagree with each other or with non-expert users, which in extreme cases means that we cannot use their ratings as a standard.

In this article, we propose a new solution to the problem of constructing quality classes for Web content based on human evaluations. The idea is to use the entire distribution of ratings to construct classes.

When we considered this idea, we were faced with the first and most practical objection: the difficulty of obtaining enough ratings to create distributions. This objection is answered by simply considering the four datasets used in this article. All of these datasets were obtained using various crowdsourcing or collaborative filtering approaches, and all of them included at least 10 ratings for a significant set of Web content, enabling the computation of empirical rating distributions from the data.

When a data scientist today wishes to define classes of Web content quality – usually for the purpose of designing a machine classification approach – she/he needs to address the issue of class design. Usually, this issue is solved by selecting thresholds for averages or other statistics (median values, modes, etc.) obtained from a sample of evaluations, and using these thresholds as boundaries for classes. The thresholds will typically be based on examination of the data, but the data scientist often needs to make an ad hoc decision in order to set threshold values. Clustering or other unsupervised approaches based on averages or medians will not work well in cases where raters often disagree, leading to dispersed distributions of ratings. We propose a new approach that uses clustering of data based on entire distributions of ratings, rather than aggregated values.

Our approach has the advantage of naturally dealing with distributions with high dispersions and other issues of distribution shape such as bias (or skewness). We show that this approach yields good results on several datasets. Furthermore, clusters should be the basis for creating operational definitions of quality classes that can be used to classify new Web content. We show how to define simple and intuitive classes using entire distributions of ratings, and demonstrate that the proposed

¹<https://movielens.org/>.

²<http://www.mywot.com/>.

³<http://www.reconcile.pl/>.

Fig. 1. Steps for designing classes based on rating distributions.

classes fit discovered clusters better than classes defined using distribution means. The fit is good not just for individual datasets: we are able to fit our classes to four datasets using two parameters. We also compare our proposed method for defining classes for highly dispersed distributions with a method that uses known dispersion measures.

The second concern when using entire distributions to assign objects to classes is that distribution shapes are more sensitive to sample variations than aggregated values. We address this concern using a sensitivity analysis. We demonstrate that distributions of Web content credibility and quality ratings are robust to extreme variations in sample composition with respect to several social, demographic, and psychological characteristics of users.

An overview of our approach is presented in Figure 1. The first step is to create an empirical distribution of ratings for an object. In this work, we focus on ratings on the Likert scale. We cluster rating distributions from four datasets using the EMD distance measure. Next, based on a qualitative investigation of the clusters (Section 5), we propose a simple, yet generic method of designing classes (Section 6). The proposed classes are simple, but sufficiently general to be able to fit to clusters discovered in all datasets (using the same values of two class parameters to evaluate the generality of our approach). Selecting these classes based on a rating distribution is simply equivalent to checking two threshold values. We also address a second objection to the proposed approach: whether distributions obtained from small samples (10 ratings) are sufficiently robust to be used for class selection. In Section 7, we have investigated the robustness of rating distributions to various demographic, social, and psychological factors. We compare distributions based on random samples with distributions based on extremely biased samples composed of ratings from respondents with the same values of a factor (for example, only male respondents). We are able to show that the EMD distance between random and extremely biased samples is usually small. What follows is that it is unlikely that the sample composition would affect distribution shape, and therefore it cannot affect the class that would be chosen for the distribution.

Our approach is not alternative but complementary to existing iterative aggregation methods that are also capable of dealing with adversarial ratings [Mukherjee et al. 2012]. These approaches usually compute a measure of rater credibility or reputation that is used to weigh the rater's evaluations. After several iterations, these weights converge and the final object rating is calculated as a weighted average of ratings. We propose that at this stage, the weighted distribution of ratings can be used to assign the evaluated object to a class. The set of classes can be created by a method

Fig. 2. Context and applications of our approach.

similar to ours based on weighted distributions of object ratings computed using the iterative algorithm. The combination of the use of an iterative aggregation algorithm with our proposed approach is described on Figure 2. The figure also shows potential applications of our approach: our classes can be used for machine classification, or in recommender systems – see related work in the next section.

We use four datasets in this work, containing diverse quality evaluations (two datasets of credibility evaluations of Web pages, one of trustworthiness evaluations of Wikipedia articles, and one of quality evaluations of movies). The datasets also differ in terms of user motivation and selection (some were obtained by recruiting and controlling users, while others were created by self-motivated users). The studied datasets contain diverse attributes of users and rated objects.

The two parts of the study, class design and evaluation of distribution robustness, contribute to answering the question of whether it is possible and practical to base a machine classification system on classes defined using rating distributions.

The **contributions of this article** are as follows:

- The proposal of an approach to clustering distributions using the Earth Mover’s Distance to study patterns of quality evaluations in four datasets of actual Web content quality evaluations.
- The proposal of a simple but general and flexible method of defining quality classes based on rating distributions, taking disagreement into account. The proposed classes are defined independently of the data and therefore can be used for training machine learning approaches with no danger of information leaks. Moreover, the proposed classes are fitted to clusters of rating distributions in the four datasets using two parameters, resulting in a satisfying overall accuracy.
- Evaluation of the degree of robustness of rating distributions to sample construction with respect to social, demographic and psychological factors of respondents as well as content features using three diverse datasets.

Organization of this Article. The article is organized as follows: In Section 2, we review related work. Section 3 introduces mathematical formalisms used in the article, especially the Earth Mover’s Distance. Section 4 describes the datasets used in this study. The main part of our work starts in Section 5, which describes the results of clustering rating distributions from our datasets. We attempt to interpret discovered clusters as highly positive, highly negative, or controversial. Section 6 introduces a definition of quality classes based on rating distributions. We define classes for highly

positive, highly negative, and controversial (disagreeing) ratings. Labeled clusters are then compared with our class definitions to evaluate the descriptive ability and generality of the class definitions proposed in this article. Section 7 describes the study of the robustness of rating distributions with respect to social, demographic, and psychological factors. We also investigate the impact of the topic of evaluated content on the rating distributions and find that it has the most significant impact on rating distribution shape. Section 8 concludes the article.

2. RELATED WORK

Our work is related to three main areas: aggregation of Crowdsourced ratings, Web content classification based on user ratings, and recommender systems. In all of these areas, subjective human quality evaluations are used to classify or compare objects. In our work, the considered objects are Web pages, Wikipedia articles, or movies, but our results could be generalized to other objects that are evaluated subjectively by humans.

2.1. Aggregation of Crowdsourced Ratings

Crowdsourcing is often used to obtain ratings of various objects, which can be anything from sushi to books. Increasingly often, crowdsourcing is used to evaluate results of other crowdsourcing jobs [Yukino and Kashima 2013] – this kind of crowdsourced quality evaluation can be treated similarly to ordinary crowdsourced ratings. A common problem is the aggregation of crowdsourced quality ratings into one result.

The most basic approach for aggregating crowdsourced ratings is to use majority voting. As Nowak and Rüger [2010] state, using central tendency measures may yield positive effects. A theoretical perspective on the bounds of majority voting is given by Kuncheva et al. [2003]. One can ask labelers to appraise objects more than once in order to obtain better results [Ipeirotis et al. 2014]. More sophisticated algorithms use the expectation-maximization method (EM). Dawid and Skene [1979] made the first attempt at using EM to estimate labelers' expertise and the true values of labels. Snow et al. [2008] use it to combat with a voting bias. Whitehill et al. [2009] propose an algorithm which learns both labelers' expertise and the difficulties associated with certain objects. A method which incorporates all of these features was proposed by Welinder et al. [2010]. Their algorithm estimates labelers' expertise in various areas, labelers' bias, and the difficulties of objects.

From our point of view, the relevant results of aggregation algorithms are the weights that can be assigned to each Crowdsourced worker – and hence, to each of his ratings. Using these weights, it is possible to produce a weighted distribution of ratings (or a simple weighted average, which is the output of many aggregation methods). The weighted distribution can be an input to our method for classifying evaluated objects. In our approach, this weighted distribution can be used to determine a quality class, even if the original quality ratings are uncertain (see Figure 2). From this point of view, our method is a new proposal of dealing with uncertainty of crowdsourced ratings.

All aggregation algorithms for crowdsourced ratings must deal with noise and uncertainty in crowdsourced data that may be caused by various factors. Frénay and Verleysen [2013] enumerate noise which is caused by: insufficient information, insufficient expertise, subjectivity, and communication problems. From our point of view, it is worth adding to this list noise which results from maliciously motivated or lazy labelers.

Cheaters, spammers, and malicious labelers are among the most irritating sources of noise in crowdsourcing, especially when users are being paid (with the possible exception of Wikipedia). Sorokin and Forsyth [2008] showed that the curve-linking quality of evaluations with salaries per rated object is concave, and that in order to optimize a process one needs to find its maximum. To deter cheaters, one might follow Kittur

et al. [2008], who propose implementation of features which make cheating more time-consuming than normal voting. Many additional measures to discourage misbehavior are suggested by Eickhoff and de Vries [2013]. Specific methods for spam detection are proposed in Hovy et al. [2013] and Raykar and Yu [2012]. Another interesting method to limit fake or lazy labels in crowdsourcing systems is to increase users' motivation by creating an efficient recommender system that suggests content for evaluation [Liu et al. 2015]. All such methods, similar to the use of iterative algorithms, are complementary to our approach, as they can be thought of as producing weights (possibly zero) for each rating.

Noise which comes from subjectivity is often unavoidable due to the subject in question or the high costs of acquiring gold labels. Smyth et al. [1995] describe the process of labeling images of volcanoes on Venus. Malossini et al. [2006] describe how to deal with subjective results in the medical informatics field. In the social informatics field, Kkol et al. [2013] investigate ways to cope with web content evaluations.

From the point of view of our work, other sources of noise (such as insufficient information, insufficient expertise level, and, to some extent, communication problems) can be treated in the same manner as the issues mentioned above.

To the best of our knowledge, this is the first proposal involving the direct use (without aggregation) of distributions of ratings to construct classes from uncertain ratings. One reason for the lack of such research is the high cost of gathering many ratings. However, contemporary crowdsourcing systems enable gathering several ratings (without aggregation) to construct empirical rating distributions for large corpora of objects.

2.2. Studies of Bias in Crowd Opinions

Another issue with crowdsourced data is its susceptibility to potential distortions. The rapid evolution of Internet technologies has afforded new means of gathering data and information retrieval. Currently, many diverse online services rely heavily on the wisdom-of-crowds approach using opinions or evaluations uploaded by users to gain knowledge. However, recent findings from market and scientific areas call attention to the question of the reliability of crowdsourced data. It has been observed that, in many cases, evaluations uploaded by service users are overly positive. Such a negatively skewed distribution of evaluations or opinions in an online environment is often addressed as positive bias and has been observed in different online contexts.

The primary and probably most common example of positive bias is online auction platforms where users are encouraged to leave comments about their transactions. The popular auction service eBay noted that about 99% of uploaded opinions can be classified as positive. This observation suggests that it might be justifiable to treat neutral comments as negatives [Morzy and Wierzbicki 2006; Kittur et al. 2008]. However, positive bias in such a context can be easily explained by reciprocity and the strategic reasoning adopted by users [Hargittai and Hsieh 2012].

The second context in which positive bias has been observed is online review systems. It has been noted that services such as BookCrossing and IMDB (a popular movie database) are also burdened with a negatively skewed distribution of uploaded opinions [Kostakos 2009]. It might be suspected that motivations to actively participate in the system play a key role in this phenomenon. Users who read books or watch movies tend to select items that they perceive as interesting. Therefore, they may be inclined to provide more positive reviews.

However, positive bias in online evaluations has also been observed in the context of scientific experiments where the abovementioned conditions do not apply. Kkol et al. [2013] asked a general population of netizens (over two thousand users) to evaluate the credibility of websites on a 1-to-5-point Likert scale. The website corpus was balanced in terms of website credibility and topic. Web pages were assigned at random

to experiment participants, who did not see the evaluations given by others. Despite a great effort made to eliminate factors related to positive bias, the obtained distribution of credibility evaluations also showed characteristic negative skewness.

2.3. Web Content Quality Classification Based on User Ratings

The quality of Web content can be rated in several proposed applications, such as the Article Feedback Tool on Wikipedia (the source of one of our datasets), or crowdsourced systems for Web content credibility evaluation such as myWOT.com or Reconcile.pl. Research on Wikipedia article quality has been using AFT ratings [Jankowski-Lorek et al. 2014; Flekova et al. 2014].

Research on Web content credibility has aimed to construct machine classifiers that could classify credibility based on sets of Web page features (related to Web page content, linguistic features, or related to a Web page's pagerank and social media mentions) [Wawer et al. 2014; Olteanu et al. 2013]. Our approach could be applied in both domains.

2.4. Hybrid Collaborative Filtering Recommender Systems

Recommender systems usually attempt to produce personalized recommendations, in contrast to our approach that produces a class design based on samples of ratings from many users. However, recommender systems also face cold-start problems due to extremely large numbers of objects for recommendation or extremely large sets of users. These problems can be partially solved using clustering of ratings. So-called hybrid collaborative recommender systems [Bobadilla et al. 2013; Shinde and Kulkarni 2012] aim to solve the cold-start problem by clustering users and objects. Our approach could be used in a hybrid collaborative recommender system as follows: first, objects need to be grouped into topical categories. Next, within each category, our approach can be applied to design quality classes. Finally, a user can be mapped to topical categories based on her interests or previous ratings. Objects from the highest quality class in these categories can be recommended to a user.

3. THE EARTH MOVER'S DISTANCE, STATISTICAL TESTS AND CONSENSUS MEASURES

In this section, we introduce the basic tools used in our research, the Earth Mover's Distance, which is a measure of distance between distributions, non-parametric tests of distribution similarity, and measures of distribution dispersion.

3.1. The Earth Mover's Distance (EMD)

Many measures designed to compare probability distributions are extensively described in the literature. These methods can be grouped into two categories: bin-by-bin and cross-bin dissimilarity measures. The former are based on comparisons between corresponding histogram bins of analyzed distributions, whereas the latter also include non-corresponding bin comparisons. Rubner et al. [2000] lists the Minkowski difference, histogram intersection, Kullback-Leibler divergence, and c_2 statistics as illustrations of bin-by-bin dissimilarity measures. The cross-bin category includes quadratic form distance, Kolmogorov-Smirnov distance, and the Earth Mover's Distance.

The Earth Mover's Distance (EMD), also known as the Wasserstein metric, is a measure based on the solution of a transportation problem raised in linear programming. It reflects the minimal cost necessary to transform one distribution into another. When comparing two probability distributions, one distribution can be treated as a supplier and the second one as a consumer. All of a distribution's mass must be moved from the first distribution to the second. The costs of transportation of one unit of goods to every consumer by different suppliers are known: they can be set to the distance between elements of the first and of the second distribution. The goal is to find the cheapest

flow of goods that will satisfy the consumers' needs. The solution to this problem is the minimal amount of work required to transform one distribution into another [Rubner et al. 2000]. We will now give a formal definition of EMD for the general case of two distributions with different values. However, in this article, the general definition will be applied to distributions of five Likert scale values.

U is a distribution with m discrete values. V is a distribution with n discrete values,

$$\begin{aligned} U &= \{(u_l, p_{ul}), \dots, (u_m, p_{um})\} \\ V &= \{(V_l, p_{vl}), \dots, (u_m, p_{um})\} \end{aligned}$$

where

p_{ui}, p_{vj} are probabilities of values. $D = [d_{ij}]$ is the ground distance matrix of distances between u_i and v_j

To calculate EMD, it is necessary to find the optimal flow $F = [f_{ij}]$ between u_i and v_j that would minimize the general cost:

$$WORK(U, V, F) = \sum_{i=1}^m \sum_{j=1}^n d_{ij} f_{ij}$$

where the following constraints are imposed:

$$\begin{aligned} f_{ij} &\geq 0; \quad 1 \leq i \leq m; \quad 1 \leq j \leq n \\ \sum_{j=1}^n f_{ij} &\leq p_{ui}; \quad 1 \leq i \leq m \\ \sum_{i=1}^m f_{ij} &\leq p_{vj}; \quad 1 \leq j \leq n \\ \sum_{i=1}^m \sum_{j=1}^n f_{ij} &= \min \left(\sum_{i=1}^m p_{ui}, \sum_{j=1}^n p_{vj} \right) = 1 \end{aligned}$$

When the optimal flow is found (i.e., the transportation problem is solved), EMD is defined by the work normalized by the total flow:

$$EMD(U, V) = \frac{\sum_{i=1}^m \sum_{j=1}^n d_{ij} f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}}$$

EMD, due to its properties, is most often used in image retrieval and texture analysis. It is a nonparametric and symmetric measure, meeting the requirement of the triangle inequality. EMD allows partial matches and is suitable for multivariate comparisons. Its major disadvantage is its relative computational complexity, which renders it unsuitable for some types of data analysis.

In our data sets, we are dealing with discrete distributions of evaluations given on a 1-to-5-point Likert scale. In other words, referring to the adopted labels, $n = m = 5$. To establish distances d_{ij} between the values, we shall simply assume that the scale is ordinal and calculate the interval difference.

In order to facilitate better understanding of how the EMD measure works for a Likert scale, some sample comparisons between distributions have been prepared. As illustrated in Figure 3 (Example A), transportation of a distribution mass equal to 0.1 increases EMD values proportionally to the distance this mass is transported. Figure 3 (Example B) illustrates sample EMDs for distributions concentrated on opposite ends of the scale. The EMD for the generated example equals 1.9 (see Figure 3, Example B).

Fig. 3. Sample ratings distributions and EMDs from reference distribution (A); Example of extreme ratings distribution (B).

3.2. Distribution Similarity Tests

In order to test the sensitivity of distributions to various features of respondents or evaluated Web pages (see Section 7), we used statistical tests. The Kruskal-Wallis test is a non-parametric statistical method used to compare multiple samples in terms of their original distribution. This method is suitable for data measured on an ordinal scale, as it bases its calculations on ranks assigned to observations. The test does not assume a normal distribution of the analyzed data.

The null hypothesis in the Kruskal-Wallis test states that mean ranks in analyzed groups are equal. In other words, data from compared samples come from the same distribution.

The test statistic H is expressed by the following formula:

$$H = (N - 1) \frac{\sum_{i=1}^g n_i (\bar{r}_i - \bar{r})^2}{\sum_{i=1}^g \sum_{j=1}^{n_i} (r_{ij} - \bar{r})^2},$$

where n_i is the number of observations in group i , r_{ij} is the rank of observation j from group i (from all observations), and N is the total number of observations in the analyzed sample.

$$\bar{r}_i = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$$

$\bar{r}$ is an average of all r_{ij} ; $\bar{r} = 1/2 (N + 1)$

3.3. Leik's Measure of Consensus

An important issue for the aggregation of crowdsourcing results is whether or not users agree on an evaluation (class, or value on a numerical scale). Inter-rater agreement measures, or dispersion measures of results, are usually used to assess agreement. In this study, we focus on a Likert-type 5-point ordinal scale, which makes the use of classic dispersion measures difficult (since the use of standard deviation is unjustified [Jamieson et al. 2004]). There is a need to address this issue with a proper measure tailored to ordinal scales. Historically, the first measure of this type was Leik's ordinal consensus, introduced in 1966 [Leik 1966], based on cumulative frequency distributions. This measure is used throughout our study; however, other ordinal consensus measures exist [Tastle and Wierman 2007; Van der Eijk 2001]. Ordinal consensus measures are capable of identifying polarization or agreement in ordinal scaled values; they are typically normalized from 0 (representing polarization) to 1 (representing perfect agreement). However, for a uniform distribution of variable levels, the frequencies of

Leik's consensus value depend on the rating scale levels (0.4 in the case of a 5-point scale).

4. OVERVIEW OF THE DATASETS USED

In this article, we report results obtained from four distinct datasets:

- A dataset of ratings obtained from a diversified sample of human users (Diversified);
- The Web Content Credibility Corpus (C3) containing ratings obtained from Amazon Mechanical Turk users;
- A dataset obtained from the Wikipedia Article Feedback Tool (AFT);
- MovieLens.

4.1. Dataset 1 (Diversified)

DATASET PROVENANCE

The first dataset was created in a study conducted in September 2012; by researchers from the Polish-Japanese Academy of Information Technology (PJATK) in cooperation with one of the largest Polish research institutes specializing in online social opinion polling. The study was conducted between 9 and 25 September 2012; The respondent group was randomly selected from the population of Polish Internet users and was controlled in terms of demographic variables.

The main task of the experiment's participants was to evaluate the credibility of presented websites. Websites were assigned to respondents at random. Participants did not interact and did not see any previous evaluations.

The website corpus used in the experiment was constructed manually and consisted of Polish web content dedicated to 16 different medical topics.

DATASET CHARACTERIZATION

The dataset contains 4,354 credibility evaluations submitted by 1,503 experiment participants. The website corpus used in the study consisted of 154 websites representing different topic categories, or 28 evaluations per page on average.

Controlled independent variables included: age, gender, education level (primary school/secondary school/higher education), Internet experience (*users' characteristics*), and topic category (*websites' characteristics*).

Internet experience was measured with the *Web-Use Skill Index* [Hargittai and Hsieh 2012], a simple scale consisting of 10 items referring to Internet-related terms, rated on a 1-to-5-point Likert scale. The possible scores that can be obtained on this scale range from 10 to 50 points.

RATINGS CRITERIA

Experiment participants were asked to evaluate the credibility of the presented websites on a 1-to-5-point Likert scale, where 1 stands for '*completely not credible*' and 5 stands for '*completely credible*'.

4.2. Dataset 2 (Content Credibility Corpus, C3)⁴

DATASET PROVENANCE

The second dataset used in this study was created in an experiment conducted by our research team as a part of the Reconcile⁵ project. The study concentrated on the tendency presented by the online reviewers to give overly positive judgments about the credibility of presented websites (*positive bias*). The study was carried out between February and November 2013 on the Amazon Mechanical Turk platform. Participants

⁴This dataset is available for public use: <http://sourceforge.net/projects/reconcile2011/files/data/>.

⁵<http://www.reconcile.pl/>.

were recruited from the U.S. population. Their task was to evaluate the credibility of the presented websites. Each participant could evaluate any number of websites (but no more than 50) and was given a proportional monetary reward. Additional characteristics of users were controlled during the study.

Literature studies dedicated to *positive bias* have led to the selection of the following psychological dimensions in the study: *trust* (expectancy that others can be relied upon), *need for cognition* (deriving pleasure from cognitive activities), *risk-taking* (tendency to involve in risky behaviors), *conformity* (understood as a sub-dimension of cooperativeness), and *intellect* (estimations of one's own intellectual abilities). In our study, we hypothesized that users showing a high level of trust and risk-taking and low level of need for cognition, conformity, and intellect will more often overestimate the credibility of the evaluated website [Rafalak et al. 2014a].

Psychological traits were measured using selected scales from the International Personality Item Pool⁶: NEO:A1 to measure trust, JPI:Cpr to measure conformity, JPI:Rkt to measure risk taking, AB5C:V+/V+ vs. V-/V- to measure intellect, and CHS to measure need for cognition. The results were analyzed using a graded response model from Item Response Theory (IRT). Internet experience was measured with the *Web-Use Skill Index* scale [Hargittai and Hsieh 2012].

The website corpus used in this experiment was constructed using three methods: customized Google queries, manual selection, and the use of RSS feed subscriptions. The selected websites represented five major topical categories.

DATASET CHARACTERIZATION

The dataset contained 15,750 evaluations of 5,543 pages from 2,041 participants. The following independent variables were collected, describing users and the websites used.

Characteristics referring to users:

- Age (groups: 10–20, 20–30, 30–40, 40–50, 50–60, 60–70, 70+)
- Education level (groups: no schooling completed, to grade 12 but no diploma or lower, high school graduate, some college credit but no degree, associate's degree, bachelor's degree, master's degree or higher)
- Internet experience (groups: high, average, low)
- Political views (groups: 10-step slider, from 'left wing' to 'right wing')
- Gender (groups: male, female)
- Income (less than \$10,000, \$10,000 to \$25,000, \$25,000 to \$50,000, \$50,000 to \$75,000, \$75,000 to \$100,000, more than \$100,000)
- Psychological traits (trust, risk taking, conformity, intellect, need for cognition; analyzed groups: low or high in every analyzed trait)

Characteristics referring to websites:

- Websites included in the corpus represented 17 distinctive detailed subject categories, grouped into five large topic categories.

RATINGS CRITERIA

Respondents were assigned to one subject category from which the pages to be assessed were randomly chosen. Their task was to evaluate presented websites in terms of credibility on a 1-to-5-point Likert scale. In order to eliminate uploading of careless evaluations, two protection mechanisms were implemented. Each evaluation had to be justified by providing a short comment or set of three similar URLs. Additionally, we excluded from further analysis evaluations that took users less than 2 minutes, treating them as random.

⁶<http://ipip.ori.org/>.

Rate this page
[What's this?](#) [View page ratings](#)

☐ Trustworthy
 ☐ Objective
 ☐ Complete
 ☐ Well-written

★★★★★
 ★★★★★
 ★★★★★
 ★★★★★

☒ I am highly knowledgeable about this topic (optional)

☐ I have a relevant college/university degree
☒ It is part of my profession
☒ It is a deep personal passion
☐ The source of my knowledge is not listed here
☒ I would like to help improve Wikipedia, send me an e-mail (optional)

example@example.com

We will send you a confirmation e-mail. We will not share your e-mail address with outside parties as per our feedback privacy statement.

Your ratings have not been submitted yet
 [Submit ratings](#)

Fig. 4. AFT GUI.

4.3. Dataset 3 (AFT)

DATASET PROVENANCE

The Article Feedback Tool (AFT) is a Wikimedia survey for article feedback, aimed at engaging readers in assessment of article quality. AFT was used on Wikipedia in five different versions. Currently discontinued, it was in use for over a year, enabling the compilation of a large dataset of Wikipedia article evaluations made by volunteer Wikipedia users. AFT GUI is illustrated in Figure 4.

DATASET CHARACTERIZATION

The original AFTv4 dataset contains over 11M ratings of articles based on 5.6M different revisions of over 1.5M distinct articles, collected between July 2011 and July 2012.

RATINGS CRITERIA

In our research, we used the Article Feedback Tool v4 (AFT) which enabled readers of English Wikipedia to rate a given article using four different criteria: trustworthiness, objectivity, completeness, and quality of writing style. Each user was presented with a survey at the end of every article and could submit his/her ratings for each dimension by choosing a rank ranging from 1 to 5 (depicted by stars).

4.4. Dataset 4 (MovieLens)

DATASET PROVENANCE

MovieLens is a project conducted by the Social Computing Research group at the University of Minnesota, which aims to produce reliable recommender systems designed specifically for recommending movies. The fourth dataset used in our work was derived from the MovieLens system.

DATASET CHARACTERIZATION

The dataset contains over a million anonymized ratings of approximately 4,000 movies by more than 6,000 users who joined the MovieLens platform in 2000. Basic demographic information (gender, age, occupation) was added to the dataset. Every user rated at least 20 movies. On average, every user rated 166 (with a median of 96). The most active user rated 2,314. Over 70% of evaluators were males (a 72:28 ratio); 73% of the evaluators were in the 18–44 age range. The three most frequent occupations were managers, students, and various unspecified.

Fig. 5. Reference distributions for analyzed datasets.

The number of ratings per movie was more diversified. The minimum number of evaluations was 1, while the most frequently evaluated movie (*American Beauty*) gathered 3,428 evaluations. The average number of ratings for a movie was 270 evaluations (a median of 123.5).

RATINGS CRITERIA

Every appraisal was made on an ordinal scale, which ranged from 1 (the lowest score) to 5 (the highest). Anyone could join the project and, after evaluating 15 movies, could start accessing his or her own recommendations from the system.

4.5. Reference Distributions

In this section, aggregated distributions of ratings from the analyzed datasets are briefly described. These cumulative distributions are used as a benchmark (we compare them to other distributions derived from different subsamples). Therefore, hereafter, they will be called *reference distributions*. As illustrated in Figure 5, all reference distributions calculated for analyzed datasets show negative skewness. This means that positive evaluations (rating options 4 and 5) are more common than negative ones (rating options 1 and 2). Therefore, the datasets used for the analysis are not free from the positive bias described in Section 2.2.

5. CLUSTERING RATING DISTRIBUTIONS

In this section, we turn to the first application of EMD: the analysis of our four datasets, aimed at discovering clusters of distributions. Clustering, one of the unsupervised machine learning techniques, helps to find a hidden structure in unlabeled data [Friedman et al. 2001]. From our perspective, clustering can help us to better understand gathered data, find typical distributions, and show any similarities between different datasets. This section represents the step called *Clustering Rating Distributions* depicted in Figure 1 that illustrates the general logic behind the proposed method.

As we use EMD to measure distances between distributions, we use the k -medoids algorithm to find cluster centers. This algorithm has the advantage that every center of a cluster must be an actual observation. This enables us to interpret the results in an intuitive way. However, one of the drawbacks of this method is that it is computationally expensive and requires constructing an $N \times N$ distance matrix between observations.

Table I. Average Silhouette Values for Various Numbers of Clusters

Number of clusters:	2	3	4	5
AFT	0.47	0.37	0.30	0.28
C3	0.55	0.44	0.39	0.33
Diversified	0.49	0.34	0.34	0.25
MovieLens	0.53	0.46	0.43	0.39

Table II. Definitions of Extreme Distributions of Values on the Likert Scale

$P_{Neg} = \{p_1 = 1\}$	Extremely negative rating distribution
$P_{Pos} = \{p_5 = 1\}$	Extremely positive rating distribution
$P_{Con} = \{p_1 = 0.5, p_5 = 0.5\}$	Extremely controversial rating distribution

For an AFT dataset, this is virtually impossible, as it would require the construction of a matrix with over 10^{10} entries. Therefore, a sample consisting of 1,000 observations randomly chosen from the dataset has been used for clustering.

Depending on the number of clusters we choose to use, the results yield different information. Creating $K = 1$ clusters enables us to find an average distribution. The distance between this and every other distribution is minimal; thus the medoid distribution may be treated as typical. If we take $K > 1$, then we can interpret the obtained clusters as distributions belonging to the same quality class.

In selecting the number of clusters, it is possible to evaluate the quality of the obtained clustering for various values of K , using the average silhouette measure (see Table I).

The largest average silhouette is obtained for $K = 2$ for all datasets. However, because of the positive bias, using just two classes of Web content quality results in great diversity within one of the two classes. For example, warning users about highly non-credible Web content would be impossible with a binary classification that matched the clusters discovered in the C3 dataset, as the cluster of non-positive evaluations contains distributions focused on the middle of the Likert scale together with distributions focused on the negative ratings. A binary classification may also be insufficient for applications that require a zero false positive (or false negative) rate, for example, the recommendation of highly credible medical Web content. For this reason, we consider $K = 3$ or $K = 4$ classes for various datasets. Using $K = 3$, we hope to discover a cluster of positive quality evaluations of Web content called **Pos**, a cluster of negative quality evaluations called **Neg**, and a cluster of neutral quality evaluations called **Neu**. Adding a fourth class should result in the discovery of a cluster of conflicting evaluations of Web content, which we refer to as controversial, called **Con**. To decide whether it is possible to identify such clusters in the data, we examine the shape of medoids for $K = 4$.

In order to better evaluate the discovered medoids, we can calculate their EMD distances from extreme distributions focused on the positive or negative evaluations on the Likert scale. Let us introduce some notation:

$P = \{p_1, p_2, p_3, p_4, p_5\}$: distribution of ratings on a 1–5 scale.

$$p_1 + p_2 + p_3 + p_4 + p_5 = 1$$

$\mathcal{P}$: family of all distributions P

$EMD(P_1, P_2)$: the Earth Mover's Distance between two distributions

In this notation, p_i denotes the probability that an evaluation has a value of i on the Likert scale, while a notation of $\{p_1 = x_1, p_2 = x_2, \dots, p_5 = x_5\}$ denotes the shape of a distribution. For the sake of concision, we skip $p_i = 0$.

Table II contains the definitions of extreme distributions. The first two distributions focus all mass on one end of the Likert scale, while the last focuses all mass on both ends and represents extreme disagreement of ratings.

Fig. 6. Four distribution clusters discovered in the Diversified dataset.

Fig. 7. Four distribution clusters discovered in the C3 dataset.

Fig. 8. Four distribution clusters discovered in the AFT dataset.

The shape of medoids for the four clusters differs visibly in every dataset (see Figures 6–9). Nevertheless, for every dataset, a medoid that can be referred to as the Pos class representation can be identified. This is the first medoid shown in Figures 6–9. These medoids are, to various extents, focused on the right end of the Likert scale. The closest is the medoid from the C3 dataset, whose EMD from P_{Pos} is equal to 0.29. The furthest is the Pos medoid from the MovieLens dataset; its EMD from P_{Pos} is 1.10.

A similar approach can be applied in the search for the Neg cluster. Again, a medoid that is shifted towards the negative end of the Likert scale exists in every dataset, as

Fig. 9. Four distribution clusters discovered in the MovieLens dataset.

illustrated by the second medoid in Figures 6–9. However, due to the positive bias of ratings in all datasets, this medoid is focused less distinctively on the extreme negative evaluations. The tendency of ratings to be J-shaped [Hu et al. 2009] makes most distributions left-skewed. This results in greater EMDs of discovered Neg medoids from the ideal P_{Neg} distribution (Diversified: 1.96, C3: 1.29, AFT: 1.25, MovieLens: 1.16).

Since we have successfully identified a Pos and Neg cluster for each dataset, all of the remaining observations could be grouped into a Neu cluster, resulting in $K = 3$. However, we still need to take into consideration potential disagreement of ratings for a significant set of Web content. Knowledge of the domains suggests that this is a real possibility; for example, Wikipedia (which is the source of content evaluated in the AFT dataset) includes procedures for identifying controversial articles and dealing with controversy.

It is not possible to identify medoids that belong to the Con class for every dataset. It seems that for most of the datasets, controversy is not a part of their natural structure. Only in the AFT dataset (from Wikipedia) are controversial observations frequent enough to constitute a separate cluster (third medoid in Figure 8). This medoid is characterized by a clearly U-shaped distribution quite close to the extremely conflicting P_{Con} distribution (an EMD distance of less than 0.3). The fourth cluster's medoid for AFT is characterized by a J-shaped distribution that cannot be interpreted as representing clear disagreement.

Medoids that have not been selected as representatives of Pos or Neg can be thought of as representatives of the Neu class. If no cluster matches the Con class, the two leftover clusters can be merged into one. The shape of the Neu class may vary, as it is simply the class that contains all remaining observations.

For every dataset, two distinct clusters exist that can be treated as the Pos and Neg class, respectively. For the AFT dataset, the Con class can also be introduced, whereas for other datasets this class does not occur. This results in our choice of $K = 3$ for the MovieLens, C3, and Diversified datasets, and $K = 4$ for the AFT dataset.

Following this qualitative analysis, we turn now to a quantitative evaluation of the fit of our proposed classes to discovered clusters.

6. DEFINING QUALITY CLASSES BASED ON RATING DISTRIBUTIONS

In this section, we propose a simple and generic method for defining classes of objects based on the rating distributions of the objects. The proposed classes are used to interpret the clusters discovered in our datasets. We show that the fit between the clusters and proposed classes is good even for very simple class parameterization. The

fit can be further improved using dispersion measures. This section represents the step called *Class Design* and *Class Selection* depicted in Figure 1 that illustrates the general logic behind the proposed method.

6.1. Definitions of Pos, Neg, Con and Neu Classes

This section introduces definitions of rating distribution classes that can be used to interpret clusters discovered in the data. These definitions are dedicated to ratings on a 1-to-5-point Likert scale but can be easily generalized to other scoring options. The definitions do not require rating aggregation, since they use entire distributions, so the question of whether the scale is ordinal or cardinal is irrelevant. Moreover, the class definitions can be made independent of the data, so they can be applied to training machine classification algorithms without the danger of information leaks.

We define our classes of distributions using the extreme distributions P_{pos} , P_{Neg} , and P_{Con} introduced in the previous section. The classes use the two parameters R and α :

$$— Pos_{\alpha}^R = \{P : EMD(P_{pos}, P) < R_1; p_1 + p_2 + p_3 < \alpha_1\}$$

(Class of rating distributions close to Extremely Positive Rating Distribution)

$$Neg_{\alpha}^R = \{P : EMD(P_{Neg}, P) < R_2; p_4 + p_5 < \alpha_2\}$$

(Class of rating distributions close to Extremely Negative Rating Distribution)

$$Con^R = \{P : EMD(P_{Con}, P) < R_3\}$$

(Class of rating distributions close to Extremely Disagreeing Rating Distribution, referred to as Controversial)

$$Neu = \mathcal{P} \setminus Pos_{\alpha_1}^{R_1} \setminus Neg_{\alpha_2}^{R_2} \setminus Con^{R_3}$$

(Class of all remaining rating distributions)

We shall use the same values for $\alpha_1 = \alpha_2 = \alpha$. The parameter α is used to avoid moving too much of the distribution mass to the other end of the Likert scale. This may happen especially with distributions that are close to the threshold of R for the EMD distance. Note that for $\alpha = 1$, the class definitions only use the radius parameters.

6.2. Simple Fitting of Proposed Classes to Discovered Clusters

Having clustered the data, we decide to check whether it is possible to derive a simple set of distance thresholds from the reference distributions. This would enable us to label any distribution for new Web content. We also want to validate the generality of the proposed classes by fitting them to all considered datasets using identical parameters. If this approach was to result in low accuracy, even assuming the fit could be improved for individual datasets using fitted parameter values, we would consider our approach to be insufficiently generic.

By iterating over a matrix of possible values for Pos, Neg, and Con radii (we consider radii from 0 to 4 with a step of 0.01) and changing the value of α (we consider the same values for both Pos and Neg classes, $\alpha = 0.1, 0.2, 0.3$, or 1) we find an optimal point, which maximizes the geometric mean of the percentage of correctly assigned labels for all datasets. We choose the geometric mean because it takes a value of zero if the percentage of correctly assigned labels drops to zero for any of the analyzed categories. Moreover, it is influenced to a greater extent by smaller values than the arithmetic mean.

The optimal point is found for the following parameters R of the Pos, Neg, and Con classes. The Con class is considered only for the AFT dataset, since this is the only

Table III. Accuracy of Optimal Class Fit to Data Clusters for Various Values of α

α	0.1	0.2	0.3	1
AFT	64.1%	78.0%	85.3%	85.8%
C3	78.4%	85.3%	81.8%	75.2%
Diversified	65.2%	75.6%	84.5%	92.3%
MovieLens	60.0%	67.9%	62.7%	48.6%
Geometric Mean of Accuracy	66.6%	76.5%	78.0%	73.3%

Table IV. Comparison of Results Obtained Via Clustering and Results Obtained Using Optimal Parameters of Classes. $R_{Pos} = 1$; $R_{Neg} = 2$; $R_{Con} = 0.7$; $\alpha = 0.3$

AFT					C3				
	Pos	Con	Neg	Neu		Pos	Con	Neg	Neu
Cluster 1 – Pos	262	0	0	6	Cluster 1 – Pos	204	0	0	0
Cluster 3 – Con	0	211	3	42	Cluster 2 – Neg	0	0	39	2
Cluster 2 – Neg	0	18	37	5	Cluster 3 and 4 – Neu	80	3	12	192
Cluster 4 – Neu	19	53	1	344	MovieLens				
Diversified						Pos	Con	Neg	Neu
	Pos	Con	Neg	Neu	Cluster 1 – Pos	348	1	0	738
Cluster 1 – Pos	35	0	0	0	Cluster 2 – Neg	0	3	378	3
Cluster 2 – Neg	0	0	9	15	Cluster 3 and 4 – Neu	0	9	494	1376
Cluster 3 and 4 – Neu	9	0	0	87					

dataset with four clusters.

$$Pos_{0.3}^1 = \{P : EMD(P_{Pos}, P) < 1; p1 + p2 + p3 < 0.3\}$$

$$Neg_{0.3}^2 = \{P : EMD(P_{Neg}, P) < 2; p4 + p5 < 0.3\}$$

$$Con_{0.7}^0 = \{P : EMD(P_{Con}^0, P) < 0.7\}$$

The optimal radius R for these three classes has the following values: $R_{Pos} = 1$; $R_{Neg} = 2$; and $R_{Con} = 0.7$. The optimal value of $\alpha = 0.3$.

In order to evaluate the effect of the alpha parameter on accuracy, we vary this parameter and find the optimal radius for proposed classes for all datasets. This enables us to compare obtained fit accuracies for various values of alpha. The results are shown in Table III.

The results show that increasing α from 0.1 to 1 causes in an increase of the accuracy for some datasets, while for other datasets the optimal value of α is 0.2 or 0.3. Overall, the optimal value of $\alpha = 0.3$. Note that we observe a severe deterioration in the accuracy for the MovieLens dataset for $\alpha = 1$.

Knowing the optimal parameters for the Pos, Neg, and Con classes, we can determine the Neu class, which, by definition, contains all observations that do not belong to the other classes. Now that we have selected parameters for all classes, we are ready to compare them with the clusters found in analyzed datasets.

Table IV contains numbers of observations belonging to discovered clusters (rows) and theoretical classes (columns). The diagonal contains the number of correctly classified observations. For the Con class, apart from the AFT dataset, we show only

misclassifications: instances of distributions that are in extreme disagreement, but have not been assigned into a Con cluster (because there are too few of them).

While overall agreement of the fitted classes with discovered clusters is high (the geometric mean of accuracy for all datasets is 78%), some interesting features can be observed. The highest agreement rate is observed for the Pos, Con, and Neg classes. Class Neu, containing all unclassified ratings, shows the highest misclassification rate.

6.3. Fitting the Controversy Class to Discovered Clusters

The above analysis uses only two parameters (the radius R and α) to fit the proposed classes to clusters discovered in the empirical data. The obtained result shows quite a good fit. The only exception is the Con class, which is too small in the analyzed datasets. To improve the fit of the Con class, we attempt to introduce a more sophisticated definition of this class. This enhanced definition uses classic dispersion measures that have also been frequently used to measure inter-rater agreement.

The definition of the extremely conflicting distribution (Con) can be extended by a parameter β that enables a shift of the mass from one end to the other: $P_{Con}^\beta = \{p_1 = 0.5 - \beta, p_5 = 0.5 + \beta\}$. The value of β must be limited so that P_{Con}^β does not change into P_{Neg} or P_{Pos} , respectively, for $\beta = -0.5$ and $\beta = 0.5$.

The controversy defined using the P_{Con} distribution serves as a perfect illustration of opinion polarization, which can be identified using classic dispersion measures. For the sake of this article we decide to use one of the earliest proposed ordinal dispersion measures, Leik's ordinal consensus (see Section 3.3). We test the performance of Leik's ordinal consensus against EMD in the detection of controversy by using the following variants of definitions for the Con class:

$$\begin{aligned} ConL^T &= \{P : Leik(P) < T\} \\ ConD_\beta^{R,T} &= \{P : EMD(P_{Con}^\beta, P) < R \vee Leik(P) < T\} \\ ConC_\beta^{R,T} &= \{P : EMD(P_{Con}^\beta, P) < R \wedge Leik(P) < T\} \end{aligned}$$

We measure the goodness of fit of these definitions by comparing the classes Con_β^R , $ConD_\beta^{R,T}$, and $ConC_\beta^{R,T}$ with the clusters from the AFT and C3 datasets containing clusters with distributions that could be considered evidence of controversy. We assume a binary classification case: the controversial cluster vs. all other distributions in the dataset.

The space of parameters α , β , R , and T is searched for the most promising parameter values yielding the best precision and recall for both datasets. Although no promising outcome is produced for a search executed on the C3 dataset, the results for AFT show that the best performance is achieved for the last $ConC_\beta^{R,T}$ class (using a conjunction of Leik and EMD). The optimal parameter values of $T = 0.35$, $\beta = 0.1$, and $R = 0.8$ show the following performances: precision = 78%, recall = 82%, $f_1 = 80\%$. In comparison, predicting the Con cluster in the AFT dataset using only the EMD radius and the simpler definition of the Con class results in precision of 75% and recall of 82%.

Note that the EMD of the uniform distribution for the 5-point Likert scale from the P_{Con} distribution is 0.8. This value is the same as the optimal EMD radius for the Con class found using our fitting approach for the AFT dataset.

The poor results obtained for the C3 dataset can be explained by the insufficient number of controversial examples in the dataset, which do not enable acquisition of a meaningful cluster.

6.4. Advantages of Proposed Method over Arithmetic Mean

In this section, we compare the use of the proposed classes with that of classes that use a simple arithmetic mean.

The first advantage of using the classes defined in Section 6.1 is evident when dealing with conflicting ratings. If a significant amount of Web content generates disagreement regarding its quality evaluation, as in the AFT dataset, using means to define classes will introduce classification errors. Our approach deals with this issue. However, it might be argued that the mean can be used together with a dispersion measure instead.

A comparison of the use of the mean with our approach can be done by considering the extreme case of $\alpha = 1$. In such a case, in the proposed class definitions, the condition of α is not used. Only the radius parameter determines the class. It can be shown that, in this case, the proposed classes are equivalent to using the mean. This is because $EMD(P_{NEG}, P) = \bar{P} - 1$ and $EMD(P_{POS}, P) = 5 - \bar{P}$, where $\bar{P}$ is the mean of distribution P . The proof of this can be found in Appendix A. This implies that classes constructed using the mean with threshold values $R_{Neg} + 1$ and $5 - R_{Pos}$ for the Neg and Pos class, respectively, would be identical to our classes with $\alpha = 1$. A comparison of the accuracy obtained using such classes with other values of α has already been shown in Table III. The optimal fit is obtained for $\alpha = 0.3$ with a geometric average accuracy of 78%, while the accuracy for classes constructed using the mean would be 73.3%. However, for the MovieLens dataset, the loss of accuracy from using means is quite severe (48.6% versus 62.7% obtained for $\alpha = 0.3$). This comparison shows the clear advantage of using a definition of classes that utilizes the α parameter, which is only possible if we use the entire distribution to define the Pos and Neg classes.

The classes proposed in this article are merely simple examples of how entire distributions can be used to define Web content classifications. The advantage of the proposed approach is its simplicity and generality. However, it is possible to use entire distributions to define classes that are much better fitted for specific applications. For example, our class definitions can be tailored so that the classes take into account possible bias of ratings (similarly as in the previous section for the Con class).

The exclusive use of arithmetic means for clustering or interpretation of the obtained clusters is likely to disregard important patterns in the data. This especially concerns the Con class, as its distribution concentrates mainly on the extreme points of the scale and therefore cannot be adequately reflected by the arithmetic mean.

7. SENSITIVITY OF RATING DISTRIBUTIONS

In the previous sections, we demonstrated how distributions of Web content quality or credibility can be clustered using EMD, and how these clusters can be interpreted as classes of distributions. Our goal is to classify Web content into distinct categories reflecting its perceived credibility or quality. In order to do this, we need to obtain a sample of evaluations for the Web content under consideration, construct a distribution, and assign it to a class. In this section, we consider the question of whether this assignment would be strongly affected by sample composition. This section represents the step called *Robustness Analysis to Sample Composition* depicted in Figure 1 that illustrates the general logic behind the proposed method.

In an ideal situation, we could ask a representative sample of Internet users to evaluate each website and, based on the collected data, draw reliable conclusions. However, such a scenario is not realistic in practical applications. Considering the unlimited resources of the Internet, assuming a high response rate for every website or movie included in the rating system is unrealistic. Rather, we should expect small samples of evaluations for a subset of all objects.

Table V. Summary Statistics of EMD Between Reference Distribution and Random Subsamples for Analyzed Datasets

Dataset	Min	2nd quintile	Median	Mean	3rd quintile	Max
Diversified N = 10	0.11	0.32	0.36	0.40	0.41	1.92
Diversified N = 50	0.04	0.14	0.16	0.17	0.18	0.35
Diversified N = 100	0.04	0.11	0.12	0.13	0.14	0.32
C3 N = 10	0.10	0.30	0.34	0.35	0.36	0.77
C3 N = 50	0.02	0.12	0.13	0.15	0.15	0.47
C3 N = 100	0.01	0.1	0.10	0.12	0.13	0.23
Movies N = 10	0.13	0.31	0.36	0.39	0.38	1.82
Movies N = 50	0.05	0.14	0.15	0.17	0.17	0.43
Movies N = 100	0.03	0.11	0.11	0.12	0.13	0.29
Aft N = 10	0.03	0.09	0.11	0.14	0.14	0.59
Aft N = 50	0.01	0.04	0.06	0.07	0.07	0.58
Aft N = 100	0.01	0.04	0.05	0.06	0.06	0.31

Therefore in this section, we examine how the proposed classification method reacts to the sample structure. We study how certain changes in the values of independent variables (the characteristics of Internet users or of websites) affect the sample distribution of evaluations and how they may thus potentially change the final content classification. This can also be thought of as an examination of the robustness of the proposed classification method to the sample structure.

The sensitivity analysis follows a plan that aims to investigate various factors that can have an impact on the sample composition. We consider demographic factors such as age, gender, education, and Internet experience. The Diversified dataset was especially prepared to investigate the impact of such factors on Web content credibility evaluations. We also consider the impact of psychological traits. The C3 dataset was especially prepared to take these factors into account (crowdsourcing workers needed to fill in a psychological survey). The topic of content is also in the Diversified, C3, and MovieLens datasets. Some datasets like MovieLens contained additional information about the raters (such as occupation) that we did not consider originally, but thought interesting. Therefore, we decided not to exclude it from our analysis. On the other hand, the MovieLens dataset does not contain information that we originally considered, such as education or Internet experience. The AFT dataset contains no information about raters, who were all anonymous.

7.1. Setting a Benchmark for Interpreting Distribution Comparisons

The reference ratings distributions described in Section 4.5 were calculated using all evaluations from the analyzed datasets. However, we need to examine cases in which we have much smaller samples of evaluations. This line of reasoning requires setting a benchmark for further comparisons. Random samples of user evaluations (size 10, 50, and 100) are drawn from analyzed datasets and compared to respective reference distributions using EMD. This experiment is then repeated 1,000 times. Summary statistics of the obtained results are presented in Table V.

The results, not surprisingly, show that the larger subsets differ to a lesser degree from the respective reference distribution. However, in real-life situations, smaller samples of evaluations for every object are to be expected. Therefore, summary statistics calculated for subsamples of size 10 will be treated in this article as a benchmark for further comparisons.

Table VI. Maximum Medians of EMD for Samples of Size 10 Selected Using Demographic Variables and Respective Benchmarks (in Gray)

	Age	Gender	Education	Internet experience	Politics	Income	Occupation	2nd quintile(benchmark)	Median(benchmark)	3rd quintile(benchmark)
C3	0.35	0.33	0.34	0.34	0.34	0.34	NA	0.30	0.34	0.36
MovieLens	0.38	0.38	NA	NA	NA	NA	0.42*	0.31	0.36	0.38
Diversified	0.39	0.39	0.38	0.39	NA	NA	NA	0.32	0.36	0.41

NA – factors not included in the dataset.

7.2. Method for Evaluating the Sensitivity of the Sample Structure to an Independent Variable

One of the major objectives of this article is to determine how the specific background of users (demographic, sociological, or psychological profile, as well as topic of content) may affect final classifications based on evaluations uploaded by random and relatively small groups of system participants. As described by Rafalak et al. [2014b], Wawer et al. [2014], and Nielek et al. [2013], psychosocial determinants play an important role in predicting evaluations in crowdsourcing systems dedicated to website trustworthiness. In order to explore how extreme cases affect evaluations uploaded to a quality evaluation system, for every value of an independent variable included in the analyzed datasets, random samples of size 10 are drawn (for example, for the gender variable, we draw samples of size 10 of evaluations submitted only by men and samples submitted only by women). We compare distributions obtained from these samples using EMD with respective reference distributions. The experiment is repeated 1,000 times. We then calculate the median EMD for each value of an independent variable. The maximum median of EMD for all values of a variable is used to evaluate the impact of that variable on sample composition. We compare this value to the benchmark EMD obtained from perfectly random samples of size 10.

7.3. Demographic and Social Factors

The analyzed datasets contain several demographic characteristics of users. The datasets related to Web content credibility evaluation (C3, Diversified) contain richer information about users, the remaining datasets (AFT, MovieLens) only basic demographic data. Additional information about the relation between demographic factors and credibility evaluation can be found in Rafalak et al. [2014b]. In subsequent analysis, we sometimes omit a certain feature if it is not present in a dataset. Common variables contained in different datasets are unified in terms of measurement scales. Maximum medians of EMD distributions for subsamples selected using demographic variables are summarized in Table VI.

The obtained results show that maximum medians for the analyzed homogeneous samples representing demographic variables fit almost perfectly into the range between the second and third quintiles of respective benchmark distribution. Only for the “occupation” variable in the MovieLens dataset are those limits slightly exceeded (by 0.04). The range between the second and third quintiles contains the middle 20% of analyzed distributions. Therefore, even in the most extreme cases, where all users

Table VII. Kruskal-Wallis H Statistics Referring to Demographic Variables Calculated for Different Datasets

Dataset		C3				Diversified		MovieLens		
Demographic characteristics	df	H	<i>p</i>	df	H	<i>p</i>	df	H	<i>p</i>	
Age	6	84.3	<0.001*	4	10.2	<0.05*	4	3123.1	<0.001*	
Gender	1	21.3	<0.001*	1	1.1	>0.05	1	422.9	<0.001*	
Education	7	42.6	<0.001*	2	0.5	>0.05	NA	NA	NA	
Internet experience	2	55.7	<0.001*	2	4.0	>0.05	NA	NA	NA	
Politics	10	43.7	<0.001	NA	NA	NA	NA	NA	NA	
Income	5	25.3	<0.001*	NA	NA	NA	NA	NA	NA	
Occupation	NA	NA	NA	NA	NA	NA	20	2341.3	<0.001*	

NA – factors not included in the datasets.

represent the same level of a demographic factor (i.e., all are men), the observed EMD from respective reference distribution is very similar to the benchmark. Therefore, we postulate that the proposed clustering method is invariant to demographic characteristics of users who upload ratings to the system.

Statistical comparisons of differences between distributions generated for all levels of available demographic features are summarized in Table VII.

The analyses show that when subsample distributions for various values of demographic features are compared to one another, almost all demographic features are significant. However, this effect is not observed in all datasets (e.g., gender, education and Internet experience). This observation leads us to conclude that while demographic features affect the shape of distributions, as demonstrated by statistical tests, this effect is weak enough to be disregarded when classifying distributions using our proposed method.

7.4. Psychological Traits

The C3 dataset contains information about Internet users' psychological profiles (for more details, see Section 4.2). More details about the influence of psychological traits on the online credibility evaluations can be found in Rafalak et al. [2014a]. The results obtained from psychological questionnaires are analyzed using the graded response model from Item Response Theory (IRT). The final result is typically expressed on a standardized scale, where $M = 0$ and $SD = 1$. Based on the distribution of results from questionnaires, for every scale we can classify users into three distinct categories: those who received low, high, or average results on a particular psychological scale. Results lower than the first quartile of given distribution are categorized as low, higher than the third quartile as high, and all other results as average. The maximum median of EMD distributions for the subsamples selected with respect to the psychological factors included in the C3 dataset are summarized in Table VIII.

The obtained results show that the maximum median for the analyzed homogeneous samples representing the psychological factors in the C3 dataset are all equal to the second quintile of C3 benchmark distribution. Therefore, we might conclude that even a population of users showing extreme psychological traits is not likely to differ much from the proposed benchmark distribution in terms of uploaded ratings.

Statistical comparisons for differences distinguished by intensity of monitored psychological traits are summarized in Table IX. According to the results of the Kruskal-Wallis test, only groups distinguished by levels of trust and conformity presented differences in distributions of credibility evaluations.

Table VIII. Maximum Medians of EMD for Samples of size 10 Selected Using Psychological Factors and C3 Benchmarks (in gray)

Psychological traits in the C3 dataset (subsamples $N = 10$)		C3 benchmark	
Psychological traits:	Max median	Min	0.10
Trust	0.30	2nd quintile	0.30
Conformity	0.30	Median	0.34
Need for cognition	0.30	Mean	0.35
Risk taking	0.30	3rd quintile	0.36
Intellect	0.30	Max	0.77

Table IX. Kruskal-Wallis H Statistics Referring to Psychological Traits

Dataset		C3	
Psychological traits:	df	H	p
Risk taking	2	1.886	0.389
Trust	2	8.564	<0.05*
Conformity	2	6.139	<0.05*
Need for cognition	2	3.986	0.136
Intellect	2	1.667	0.434

Table X. Maximum Medians of EMD for Samples of Size 10 Selected Using Website Topic Category Compared to Respective Benchmarks (in gray)

	Topic	2nd quintile (benchmark)	Median (benchmark)	3rd quintile (benchmark)
C3	1.68	0.30	0.34	0.36
MovieLens	0.41	0.31	0.36	0.38
Diversified	0.42	0.32	0.36	0.41

Only in groups distinguished by level of trust and conformity are statistically significant differences in users' ratings observed. However, this effect is not strong enough to cause an observable difference when compared to the C3 benchmark.

7.5. Topic of Rated Content

In three datasets, subject categories of evaluated websites can be distinguished. These categories are different for every dataset, and thus cannot be unified. Maximum medians from EMD comparisons between subsamples distinguished by website topic categories and respective reference distributions are summarized in Table X.

In all three datasets, the maximum median of EMD distributions is greater than the third quintile of the respective benchmark. In the MovieLens and Diversified datasets, this difference is rather negligible, but in C3 it is significant. It is worth mentioning that C3 is more diversified than the MovieLens and Diversified datasets in terms of website topic category. MovieLens contains users' ratings on movies, Diversified mostly ratings of websites dedicated to medicine, C3 ratings of websites representing five distinct and diversified topic categories. This is probably why the obtained results show the strongest impact for the website topic in the C3 dataset.

Kruskal-Wallis comparisons of credibility evaluation distributions calculated for websites grouped by topic category are presented in Table XI. In all analyzed datasets, these comparisons reached the level of statistical significance. This result suggests that the credibility ratings given by users are dependent on the subject category of the evaluated website.

Table XI. Kruskal-Wallis H Statistics for Credibility Evaluations Distributions in Different Datasets for Website Topic Categories

Dataset	C3			Diversified			MovieLens		
	df	H	<i>p</i>	df	H	<i>p</i>	df	H	<i>p</i>
Topic	56	108.75	<0.005*	8	72.51	<0.001*	300	78079.55	<0.001*

Table XII. Comparisons of Maximum Medians of EMD for Samples of Size 10 Selected Using Various Variables in the the C3, Diversified and MovieLens Datasets

	Dataset	C3	Diversified	MovieLens	MAX
Demographics	Age	0.35	0.39	0.38	0.39
	Gender	0.33	0.39	0.38	0.39
	Education	0.34	0.38		0.38
	Internet experience	0.34	0.39		0.39
	Politics	0.34			0.34
	Income	0.34			0.34
	Occupation			0.42	0.42
Psychological traits	Risk taking	0.30			0.30
	Trust	0.30			0.30
	Conformity	0.30			0.30
	Intellect	0.30			0.30
	Need for cognition	0.30			0.30
Websites' features	Topic	1.68	0.42	0.41	1.68

7.6. Comparison of the Impacts of Various Factors

Summarized results illustrating how certain factors affect distributions of ratings are shown in Table XII. When compared to the respective benchmarks, only the category of website topic seems to significantly modify the ratings distribution. The maximum medians from EMD distributions generated for subsamples representing homogeneous topic categories in the C3 dataset greatly exceed the third quintile of benchmark distribution. A median slightly higher than the third quintile of the MovieLens benchmark distribution was also observed for the occupation factor.

7.7. Interaction of Independent Factors

The results described in the previous subsection suggest that the only factor remarkably affecting distribution of quality evaluation ratings is the topic category of a website. However, bearing in mind the great complexity of social and psychological phenomena, it seems appropriate to consider interactions between the topic category of an evaluated website and the characteristics of system users. Given that aspects of evaluated content can be easily controlled by the system administrator, focusing on characteristics of system users in relation to evaluated stimuli seems even more reasonable.

However, fragmenting data into more detailed categories often leads to low frequencies of observations in analyzed data bins and consequently disables reliable intergroup comparisons. Therefore, we decide that each bin for every category level should contain at least 40 observations in order to be included in the analysis. Due to data collection design, the Diversified dataset seems most appropriate for the purpose of examining interactions.

Analogously to previous analyses, subsamples of size 10 are selected 1,000 times with respect to the combination of different factors included in the Diversified dataset. The obtained distributions are compared using EMD with the reference distribution of the Diversified dataset. The maximum median EMD for every combination is illustrated in Table XIII.

Table XIII. Maximum Median of EMD Distributions for 1,000 Subsamples of Size 10 Selected by a Combination of Factors in the Diversified Dataset Compared with the Benchmark (in gray)

	Topic category			Benchmark	
	Hormonal contraception	Oral chemotherapy	Dukan diet	Min	0.11
Gender	0.39	0.37	0.41	2nd quintile	0.32
Age	0.38	0.43	0.47	3rd quintile	0.41
Education	0.38	0.37	0.42	Max	1.92

Table XIV. Kruskal-Wallis H Statistics for Credibility Evaluations Distributions for Selected Interactions of Independent Variables (the Diversified Dataset)

	df	Hormonal contraception		Oral chemotherapy		Dukan diet	
		H	<i>p</i>	H	<i>p</i>	H	<i>p</i>
Gender	1	1.580	>0.05	0.144	>0.05	2.083	>0.05
Age group	4	1.630	>0.05	5.870	>0.05	8.709	>0.05
Education	2	0.905	>0.05	1.879	>0.05	0.653	>0.05

The results show that the maximum medians from combinations of different factors are in the range of, or slightly above, the second and third quintiles of the benchmark distribution. Therefore, we might conclude that analyzed combinations of independent factors do not strongly modify ratings distributions and are not likely to change classifications obtained with the proposed methodology.

In order to check whether the ratings' distributions differ between categories of factor combinations, the Kruskal-Wallis statistical test is executed. The results are summarized in Table XIV. None of the analyzed inter-group comparisons are statistically significant.

8. CONCLUSIONS

In this article, we have considered the research question: is it possible and practical to base a machine classification system for Web content quality on classes defined using rating distributions? To answer it, the first step was to propose an approach to define classes using rating distributions.

We proposed a set of four Web content quality classes, Pos, Neg, Con, and Neu, that could be intentionally defined using rating distributions and EMD. We also attempted to match these classes to clusters of distributions obtained from four diverse empirical datasets. The results of this analysis revealed a relatively good match between the proposed classes and the discovered clusters. Moreover, this match was obtained using just two parameters of proposed classes – the radius of EMD and a parameter limiting the mass that could be shifted towards the other end of the Likert scale – and was good for all four studied datasets. We used the same parameters for all datasets to prove the generality of the proposed approach. The Con class was discovered in only one dataset, but we were also able to match it well in this set using the proposed definition.

We compared the proposed classes against classes constructed using distribution means (which are a special case of our classes) and using the Leik distribution dispersion measure for ordinal scales (for the Con class). The comparison proves that the use of means in the considered datasets results in a worse fit to clusters discovered using EMD than the use of the proposed class definitions. Also, the use of only the Leik measure results in a worse fit than the use of our proposed definition for the Con class.

Among the concerns regarding the practicality of using distributions, the acquisition of sufficiently good samples is important. For this reason, we have studied the robustness of distributions in our datasets to changes in the characteristics of raters (demographic and social) as well as to the topic of the rated content. The method we used is relatively insensitive to the sample size, as it takes into account only the shape of the distribution. Obtained results proved to be stable for distributions consisting of 10 ratings per object and more. For this reason, we complemented it with non-parametric statistical tests. The results indicate that the social and psychological characteristics of raters have a small impact on the shape of rating distributions, leading us to the conclusion that in practice, rating samples obtained using crowdsourcing would suffice to assign rated content to proposed classes. On the other hand, the topic of content was clearly a variable with a much stronger and universal (observed in three datasets) impact on distribution shape.

The approach proposed in this work is very simple and quite generic. The number of classes, as well as their shape, can be adapted to the desired application – for example, using unsupervised machine learning. The definitions of classes can also be more sophisticated, as shown by the example of fitting the Con class to the AFT dataset using the Leik dispersion measure and EMD. Our contribution is only a first step; however, we believe that using rating distributions for machine learning is a promising approach. Classes derived from the shape of distributions have the following advantages:

- (1) Using any form of aggregation that is the function of a rating distribution (mean or majority rating, for example) results in a loss of information that may be significant for uncertain ratings.
- (2) Dealing with rater disagreement by directly introducing a class of controversial content enables learning of the characteristic features of such content, which can be useful in many applications. Jankowski-Lorek et al. [2014] proves that early detection of controversy can boost effectiveness of peer production systems.
- (3) Using distributions based on samples of moderate size results in reduced uncertainty about an object's class.
- (4) The proposed definitions of classes are independent of data, since they were fitted to four datasets at the same time. This enables the use of the proposed definitions for machine learning while avoiding information leaks.

Of course, using distributions also has drawbacks, the most important of which is the necessity of obtaining a sufficient number (at least 10) of ratings per object. This, though it may seem costly, is realistic when using current crowdsourcing services such as Amazon Mechanical Turk (as demonstrated by our research leading to the creation of the C3 and Diversified datasets).

Among the next steps for our research is the creation of iterative aggregation algorithms for crowdsourced ratings aimed at the creation of rating distributions rather than the choice of a single value on the rating scale. These algorithms should also be able to evaluate the credibility or quality of a rating user, which requires a new approach if rating distributions are used instead of single values.

APPENDIX A. RELATIONSHIP BETWEEN EMD FROM EXTREME DISTRIBUTIONS AND DISTRIBUTION MEAN

In this appendix, we prove two simple facts concerning the relationship of EMD from the extreme distributions with the arithmetic mean. We prove these facts for the distributions using the Likert scale; however, the reasoning would be the same for all rating scales.

Let us begin by repeating the notation used in the article:

$P = \{p_1, p_2, p_3, p_4, p_5\}$: distribution of ratings on a 1–5 scale

$$p_1 + p_2 + p_3 + p_4 + p_5 = 1$$

$\bar{P}$ denotes the mean of P .

$EMD(P_1, P_2)$: the Earth Mover's Distance between two distributions

In the notation used, p_i denotes the probability that an evaluation has a value of i on the Likert scale, and a notation of $\{p_1 = x_1, p_2 = x_2, \dots, p_5 = x_5\}$ denotes the shape of a distribution. (For the sake of concision, we skip $p_i = 0$.)

Using this notation, we defined the extreme distributions $P_{Neg} = \{p_1 = 1\}$ and $P_{Pos} = \{p_5 = 1\}$. We shall now consider the EMD of an arbitrary distribution $P = \{p_1 = x_1, p_2 = x_2, p_3 = x_3, p_4 = x_4, p_5 = x_5\}$ to the extreme distributions.

FACT 1. $EMD(P_{Neg}, P) = \bar{P} - 1$.

PROOF. The proof follows from the observation that the shortest and least costly path of moving the distribution mass from the extremely negative distribution P_{Neg} to distribution P consists of moving the mass to the right. We can choose an order; for example, we can first move x_2 of the distribution mass from rating 1 to rating 2, then x_3 of the distribution mass from rating 1 to rating 3, and so on. There is no shorter or cheaper way to move this mass.

Thus, $EMD(P_{Neg}, P) = 1 \cdot x_2 + 2 \cdot x_3 + 3 \cdot x_4 + 4 \cdot x_5$.

However, $\bar{P} = 1 \cdot x_1 + 2 \cdot x_2 + 3 \cdot x_3 + 4 \cdot x_4 + 5 \cdot x_5$.

Therefore,

$$\begin{aligned} \bar{P} - 1 &= x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 - (x_1 + x_2 + x_3 + x_4 + x_5) \\ &= x_2 + 2x_3 + 3x_4 + 4x_5 = EMD(P_{Neg}, P). \end{aligned}$$

This completes the proof. $\square$

FACT 2. $EMD(P_{Pos}, P) = 5 - \bar{P}$

PROOF. The reasoning is similar to that behind Fact 1.

$$EMD(P_{Pos}, P) = 1 \cdot x_4 + 2 \cdot x_3 + 3 \cdot x_2 + 4 \cdot x_1$$

$$\begin{aligned} 5 - \bar{P} &= 5(x_1 + x_2 + x_3 + x_4 + x_5) - (x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5) \\ &= 4x_1 + 3x_2 + 2x_3 + x_4 = EMD(P_{Pos}, P) \end{aligned}$$

This completes the proof. $\square$

Note that the proven relationships may be useful for reducing the computational complexity of calculating the EMD radius from the extreme distributions for the proposed Pos and Neg classes.

REFERENCES

- Jesus Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. 2013. Recommender systems survey. *Knowledge-Based Systems* 46, (July 2013), 109–132.
- Alexander P. Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*. (1979), 20–28.
- Carsten Eickhoff and Arjen P. de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval* 16, 2 (2013), 121–137.
- Lucie Flekova, Oliver Ferschke, and Iryna Gurevych. 2014. What makes a good biography?: Multidimensional quality analysis based on Wikipedia article feedback data. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*. ACM, New York, 855–866.
- Benoît Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (2014), 845–869.

- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Springer, Berlin.
- Eszter Hargittai and Yuli Patrick Hsieh. 2012. Succinct survey measures of web-use skills. *Social Science Computer Review* 30, 1, 95–107.
- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *HLT-NAACL*, 1120–1130.
- Nan Hu, Jie Zhang, and Paul A. Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Communications of the ACM* 52, 10, 144–147.
- Panagiotis G. Ipeirotis, Foster Provost, Victor S. Sheng, and Jing Wang. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28, 2, 402–441.
- Susan Jamieson. 2004. Likert scales: How to (ab)use them. *Medical Education* 38, 12, 1217–1218.
- Michał Jankowski-Lorek, Radosław Nielek, Adam Wierzbicki, Kazimierz Zieliński, 2014. Predicting controversy of Wikipedia articles using the article feedback tool. In *Proceedings of the 2014 International Conference on Social Computing*. ACM, 22.
- Michał Kałol, Michał Jankowski-Lorek, Katarzyna Abramczuk, Adam Wierzbicki, and Michelle Catasta. 2013. On the subjectivity and bias of web content credibility evaluations. In *Proceedings of the 22nd International Conference on World Wide Web Companion*. 1131–1136.
- Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 453–456.
- Vassilis Kostakos. 2009. Is the crowd's wisdom biased? A quantitative analysis of three online communities. In *Proceedings of the 2009 International Conference on Computational Science and Engineering (CSE'09)*. IEEE, 251–255.
- Ludmila Kuncheva, Christopher J. Whitaker, Catherine A. Shipp, and Robert P. W. Duin. 2003. Limits on the majority evaluation accuracy in classifier fusion. *Pattern Analysis & Applications* 6, 1, 22–31.
- Robert Leik. 1966. A measure of ordinal consensus. *Pacific Sociological Review* 9, 2, 85–90.
- Xiu Liu, Radosław Nielek, Paulina Adamska, Adam Wierzbicki, and Karl Aberer. 2015. Towards a highly effective and robust Web credibility evaluation system. *Decision Support Systems* 79 (2015), 99–108.
- Andrea Malossini, Enrico Blanzieri, and Raymond T. Ng. 2006. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics* 22, 17 (2006), 2114–2121.
- Mikołaj Morzy and Adam Wierzbicki. 2006. The sound of silence: Mining implicit feedbacks to compute reputation. In *International Workshop on Internet and Network Economics*. Springer, 365–376.
- Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, New York, 191–200. DOI: 10.1145/2187836.2187863.
- Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*. ACM, 557–566.
- Radosław Nielek, Aleksander Wawer, Michał Jankowski-Lorek, and Adam Wierzbicki. 2013. Temporal, cultural and thematic aspects of web credibility. In *Social Informatics*. Springer International Publishing, 419–428.
- Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. 2013. Web credibility: Features exploration and credibility prediction. *Advances in Information Retrieval*. Springer, Berlin, 557–568.
- Maria Rafalak, Katarzyna Abramczuk, and Adam Wierzbicki. 2014a. Incredible: Is (almost) all web content trustworthy? Analysis of psychological factors related to website credibility evaluation. *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1117–1122.
- Maria Rafalak, Piotr Bilski, and Adam Wierzbicki. 2014b. Analysis of demographical factors' influence on websites' credibility evaluation. *Human-Computer Interaction. Applications and Services*. Springer International Publishing, 57–68.
- Vikas C. Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research* 13, 491–518.
- Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*. 40, 2 (2000), 99–121.
- Subhash Shinde and Uday Kulkarni. 2012. Hybrid personalized recommender system using centering–bunching-based clustering algorithm. *Expert Systems with Applications* 39, 1, 1381–1387.
- Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1995. Inferring ground truth from subjective labelling of Venus images. In *Advances in Neural Information Processing Systems*. 1085–1092.

- 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*. Association for Computational Linguistic, 254–263.
- Alexander Sorokin and David Forsyth. 2008. Utility data annotation with Amazon Mechanical Turk. *Urbana* 51, 61(2008), 820.
- William J. Tastle and Mark J. Wierman. 2007. Consensus and dissent: A measure of ordinal dispersion. *International Journal of Approximate Reasoning* 45, 3, 531–545.
- Cees Van der Eijk. 2001. Measuring agreement in ordered rating scales. *Quality and Quantity* 35, 3, 325–341.
- Aleksander Wawer, Radosław Nielek, and Adam Wierzbicki. 2014. Predicting webpage credibility using linguistic features. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1135–1140.
- Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. 2010. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*. 2424–2432.
- Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose evaluation should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*. 2035–2043.
- Baba Yukino and Hisashi Kashima. 2013. Statistical quality estimation for general Crowdsourcing tasks. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 554–562.

Received April 2015; revised August 2016; accepted August 2016