A Machine Learning Approach for Classifying Textual Data in Crowdsourcing

Marcel Rhyn[1], Ivo Blohm[1]

[1] University of St. Gallen (HSG), Institute of Information Management, St. Gallen, Switzerland
{marcel.rhyn,ivo.blohm}@unisg.ch

Abstract. Crowdsourcing represents an innovative approach that allows companies to engage a diverse network of people over the internet and use their collective creativity, expertise, or workforce for completing tasks that have previously been performed by dedicated employees or contractors. However, the process of reviewing and filtering the large amount of solutions, ideas, or feedback submitted by a crowd is a latent challenge. Identifying valuable inputs and separating them from low quality contributions that cannot be used by the companies is time-consuming and cost-intensive. In this study, we build upon the principles of text mining and machine learning to partially automatize this process. Our results show that it is possible to explain and predict the quality of crowdsourced contributions based on a set of textual features. We use these textual features to train and evaluate a classification algorithm capable of automatically filtering textual contributions in crowdsourcing.

Keywords: Crowdsourcing, Machine Learning, Text Mining, Automatization

1 Introduction

In recent years, crowdsourcing has increasingly gained attention as an innovative approach to harness the collective resources of a broad and diverse network of people over the internet. The fundamental idea of crowdsourcing is that an organization proposes the voluntary undertaking of a task to an independent group of contributors in an open call [1, 2]. It seeks to mobilize the creativity, knowledge, or distributed workforce of a large panel of people who perform value creation activities that have previously been carried out by designated agents, such as employees or third-party contractors. The approach grants scalable access to remote resources and allows tasks to be completed in a parallelized fashion regardless of time and location. In this vein, crowdsourcing has been found to greatly improve the efficiency and effectiveness of problem-solving in organizations [3, 4].

However, the potential that arises from the decentralized contributions provided by a crowd comes with a critical challenge. The quantity and complexity of information that needs to be processed and evaluated in crowdsourcing is high – especially when the contributions are submitted in a raw, textual format. In 2006, for example, more than 140’000 international participants joined the IBM Innovation Jam and submitted

13th International Conference on Wirtschaftsinformatik,
February 12-15, 2017, St. Gallen, Switzerland

Rhyn, M.; Blohm, I. (2017): A Machine Learning Approach for Classifying Textual Data in Crowdsourcing, in Leimeister, J.M.; Brenner, W. (Hrsg.): Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017), St. Gallen, S. 1171-1185

over 46’000 ideas in a single crowdsourcing contest [5]. Similarly, the devastating earthquake in Haiti during January 2010 generated over 13’500 crowdsourced messages on online maps that were used to locate emergencies and distribute relief supplies [6]. As these contributions are submitted by a diverse network of people with different backgrounds and degrees of expertise, textual data in crowdsourcing usually entail a high amount of noise and ambiguity. Thus, the process of manually evaluating the data and filtering out low quality contributions is arduous and lengthy [2]. It generally accounts for one of the most time-consuming and cost-intensive steps in crowdsourcing [7]. For example, it took Google almost three years and 3’000 employees to condense the 150’000 proposals submitted to its *Project 10 to the 100* [2].

Text mining and machine learning algorithms represent promising solutions to cope with the vast amount of contributions in crowdsourcing [8]. They provide the means to discover patterns and extract useful information from textual data in a fast, scalable, and repeatable way [9]. In this vein, they offer the potential to automatically evaluate and filter contributions in crowdsourcing. Although multiple studies have asked for such automated approaches, research on crowdsourcing is still lacking feasible models for this task [7, 10]. Our study aims to close this gap by addressing the following research question: “What textual characteristics can be used to assess and automatically predict the quality of contributions in crowdsourcing?” To answer this question, we choose a two-pronged approach that has already been used similarly in related studies [11]. First, we apply an explanatory regression analysis to examine textual characteristics that are associated with contribution quality in crowdsourcing. Then, we use these textual characteristics for predictive modeling with machine learning algorithms. That is, we build a classifier capable of predicting the quality of the contributions based on their textual characteristics.

Hence, the contribution of our study is twofold. For researchers, we provide a set of variables and models to explain and predict contribution quality in crowdsourcing. These models and variables can be used to assess textual contributions with machine learning algorithms and, thus, contribute to a partial automatization of the evaluation process. For practitioners, we build a classifier based on the Random Forest algorithm that incorporates these variables. It is capable of automatically filtering high quality and low quality contributions submitted by a crowd and makes the process of reviewing large volumes of textual feedback more efficient.

The remainder of this paper is structured as follows. In Section 2, we discuss the characteristics of textual contributions in crowdsourcing and review existing evaluation methods for this type of data. In Section 3, we derive hypotheses regarding the relationship between textual characteristics and contribution quality in crowdsourcing. In Section 4, we describe the methodology for testing these hypotheses with a regression analysis and outline our approach for predictive modeling with machine learning algorithms. Finally, in Section 5 and 6, we analyze the results and illustrate their implications for both researchers and practitioners.

2 Related Work

2.1 Textual Data in Crowdsourcing

The fundamental principle of crowdsourcing is the use of an open call to engage a wide network of potential contributors who submit their solutions to a set of tasks broadcasted by a company [1]. In this vein, crowdsourcing facilitates the collection of information and the distribution of problem-solving to a mass of users that are coerced into productive labor [6]. On the flipside, opening up the participation to a decentralized crowd of individuals makes it more difficult to control the content and format of the data [12]. This is especially challenging for the broad range of crowdsourcing settings that are based on contributions submitted in an free text format, such as ideas on open innovation platforms [5] or user feedback in crowdsourced software testing [7]. These textual contributions represent an unstructured data format and come with several problematic characteristics regarding both their contextual and their representational quality [13]. First, there is no ground truth to contributions such as ideas, feedback, or reviews. Hence, for these types of textual contributions, it is inherently complex to assess and compare contextual characteristics such as the relevancy or the completeness of the information [14]. Members of a crowd may have different perceptions of what is relevant or interesting for such a task and will typically cover a broad range of topics in their contributions [12]. Some contributions may lack focus and specificity; others may even include contradictory or false information [2]. Second, the representation of information in textual contributions is generally of high variance and diversity [6]. Depending on their background and their degree of expertise, members of a crowd may express themselves in very distinct ways, using different expressions for similar issues or similar expressions for different issues [2]. Hence, not only is there a wide range of potential topics but also a wide range of potential descriptions for these topics. This is aggravated by the fact that textual data generated by a crowd typically entail a high amount of noise due to spelling mistakes, grammatical errors, excessive punctuation, or informal writing styles [6].

2.2 Evaluation Methods for Textual Data

Given the previously described characteristics of textual data in crowdsourcing, it is difficult to use traditional approaches to quality control [15]. For example, it is not possible to employ gold standard data as there is typically no ground truth to which the contributions can be compared. Hence, companies rely on a manual assessment of the contributions. That is, someone has to read the contributions, evaluate the quality of the content, compare it to the requirements of the task, and either accept or dismiss the input for further consideration by the company [7]. Expert panels that review and select relevant inputs represent one of the most reliable yet impractical means for this step [14]. The volume of textual data and the rate at which they are created in crowdsourcing often exceed their information processing capacities [2]. Other approaches rely on the crowd itself for the evaluation of the contributions. However, multiple studies have shown that the design of ratings scales is highly challenging and often fails to produce reliable results [16]. For example, rating scales have been found to frequently face the problems of bimodal distributions or self-selection bias [17].

In consequence, a number of studies have experimented with text mining and machine learning algorithms to support the evaluation of textual data in crowdsourcing. Walter and Back [18] use text mining algorithms to cluster ideas submitted to innovation jams in an attempt to provide decision support for expert panels reviewing the contributions. Similarly in the domain of crowdsourced software testing, existing research has used text mining approaches to automatically cluster bug reports and prioritize them for the developers [19]. In the humanitarian aid sector, Rogstadius et al. [20] and Barbier et al. [6] outline the use of text mining algorithms for clustering crowdsourced incident reports and extracting named entities (e.g., locations or family names) in order to make the coordination of appropriate responses more efficient.

Hence, existing studies have already examined how the large number of textual contributions can be clustered and organized for companies trying to analyze the multitude of diverse topics and content submitted by the crowd. We extend this body of literature by analyzing textual characteristics that can be used to explain and predict the quality of the contributions. This allows companies not only to organize the variety of contributions, but also to automatically identify relevant inputs with machine learning algorithms and filter out those that are likely not to bear any value.

3 Hypotheses Development

For developing our model, we draw upon well-established textual features discussed in related literature [21–25] to operationalize the previously described contextual and representational characteristics of crowdsourced data and examine how these features are associated with contribution quality in crowdsourcing. Contextual characteristics account for the amount and the relevancy of the information provided in textual data. Representational characteristics account for the extent to which the text is presented in a clear and intelligible manner [21].

First, the amount of information in a textual contribution has frequently been discussed as one of its most important features by related literature [22, 24, 25]. Longer contributions contain more information that could potentially be relevant for the company than shorter ones [21]. It is also easier for companies to act on feedback that is well elaborated [2], as it allows them to build a more comprehensive and coherent representation of the information in the text [14]. For example, Riedl et al. [16] note that “more accurate, understandable, and comprehensive information enables decision makers to perform better” (p. 12). On the other hand, they emphasize that contributions that are short and less elaborated tend to deliver less information that could be required for an accurate understanding of the contributions and appropriate decision making [16]. Second, related literature also emphasizes the need to consider the relevancy of the information in a contribution [21, 25]. Otterbacher [21] quantifies the extent to which a product review contains terms that are statistically important across other reviews. Similarly, Weimer and Gurevych [25] use similarity features to measure the relatedness of a post to a forum topic. For crowdsourcing in particular, relevant contributions are typically characterized as containing clear and specific information for the companies to act on [2], while vague and blurry descriptions have been found to be detrimental to contribution quality [16]. Hence, we hypothesize as follows:

Hypothesis 1. The length of a textual contribution is positively associated with the quality of the contribution.

Hypothesis 2. The specificity of the terms used in a textual contribution is positively associated with the quality of the contribution.

Besides contextual characteristics accounting for the amount and the relevancy of the information, a second layer of analysis is concerned with the representational characteristics of a contribution [21, 26]. On the one hand, representational characteristics can be used as means to measure the sophistication of a contribution [21]. For example, the readability [27] is frequently used to analyze the syntactic and semantic complexity of a text [26]. In crowdsourcing, a higher readability of a contribution should enable companies to better understand the submitted content and extract relevant cues or information more easily [14]. On the other hand, representational characteristics can be broken down to purely superficial aspects, such as the extent to which a contribution respects common writing standards or reveals irregularities [11, 25]. Poorly written contributions containing spelling errors and grammatical mistakes increase the noise and ambiguity in the data [26]. Such irregularities impose a higher cognitive load on the recipient in the company and make the contributions prone to misinterpretation [14]. Hence, they are likely to be detrimental to the interpretability or clarity of crowdsourced contributions and may render the acquisition of the embedded information more difficult for companies. Thus, we define the second set of our hypotheses as follows:

Hypothesis 3. The readability of a textual contribution is positively associated with the quality of the contribution.

Hypothesis 4. The number of spelling mistakes in a textual contribution is negatively associated with the quality of the contribution.

4 Methods and Data

In order to answer our research question, we combine two independent data sources: textual contributions from a crowdsourcing project for which we apply text mining algorithms to make them eligible for statistical analysis and an expert-based baseline measure of contribution quality. This allows detailed insights into the automated classification of the contributions with machine learning algorithms.

4.1 Data Collection

For our study, we retrieved textual data from a crowdsourcing project in the field of software testing. We conducted a crowdsourced software test in cooperation with a

German-based intermediary that ranks amongst Europe's leading platforms in this domain and manages a crowd of more than 100'000 international software testers. The test was designed as a user acceptance test for a website and has been carried out in August 2015 over the course of 5 days. It consisted of open tasks that asked the testers about their opinion on positive and negative aspects of the website as well as suggestions for further improvement. This setting was chosen for several reasons. First, user acceptance tests for websites represent one of the most frequently performed types of software tests by crowdtesting platforms, as they allow companies to gather feedback from real end users of the software [7]. Second, user acceptance tests typically lead to a large amount of textual data which are especially time-consuming to evaluate by experts or developers. Third, the feedback retrieved during user acceptance tests resemble contributions in other domains, such as ideas in innovation management or reviews in product development. This allows the results of our study to be transferred to other crowdsourcing contexts and ensures their generalizability.

We received 309 contributions in a raw textual format from 104 testers who represent the target demographic of the website and who were randomly assigned to the software test by the intermediary. On average, the contributions contained 41 words with a standard deviation of 38 words. All contributions were written in English.

4.2 Expert Evaluation of Contribution Quality

As discussed previously, there is no ground truth to contributions such as ideas, feedback, or reviews. In the absence of objective measures, it is necessary to employ an expert-based baseline measure for contribution quality [14]. Therefore, we adapted the Consensual Assessment Technique for our study [28]. We asked two software experts to manually review the feedback. Both experts are involved in the development of the website for which the user acceptance test has been conducted. Thus, they are qualified to evaluate the contributions of the crowd. They independently reviewed all test reports by using the same evaluation scheme. The evaluation scheme is based on the framework proposed by Blohm et al. [14] for crowdsourcing and includes four criteria: relevance, elaboration, feasibility, and novelty. To cover these criteria, we used questions developed by Nørgaard and Hornbæk [29] who applied them analogously for assessing usability feedback in software testing. Hence, they are suitable for our study which is concerned with similar feedback to user acceptance tests. Each criterion was rated on a 5-point Likert scale. To validate the ratings of the experts, we calculated the weighted Cohen's Kappa for each criterion [30].

Table 1. Cohen's Kappa Statistics

Relevance	**Elaboration**	**Feasibility**	**Novelty**
0.78**	0.76**	0.77**	0.73**

Note: **substantial agreement, see Landis and Koch [31]

The strength of agreement as listed in **Table 1** is substantial [31] for all criteria, indicating that we have reliable quality measures. We used the mean to aggregate the expert ratings. Since we analyze contribution quality as a multidimensional construct [14], we followed past research [32–34] and calculated a composite score for contribution quality by averaging the ratings.

4.3 Variables and Measurements

We draw upon related literature [21–25] and use the textual features derived in Section 3 as variables to explain and predict the quality of the crowdsourced contributions. We use two variables (i.e., length and specificity) to account for their contextual characteristics and two variables (i.e., readability and spelling) to account for their representational characteristics.

Length. We measure the length of a contribution by counting the total number of words per contribution.

Specificity. We measure the specificity by building the sum of all TF.IDF-indices for a contribution. The TF.IDF-index represents a term weighting scheme that accounts for the importance of a particular term in the data set based on the term frequency and the inverse document frequency [35]. Generally speaking, broad and frequently used terms by the crowd (e.g., “bad” or “design”) will receive lower values than more specific terms (e.g., “unintuitive” or “navigation”). For calculating these TF.IDF-indices, we follow the commonly used bag-of-words approach with a vector space model and apply standard preprocessing steps [36]. More specifically, we tokenize the contributions by breaking them up into individual terms. We apply standard transformations to the single terms, including normalization (i.e., transforming all characters to lower-case), stop word filtering (i.e., removing terms such as articles or prepositions that bear no value for the analysis) and stemming (i.e., reducing terms to their root form to avoid duplications) with the Porter stemmer [37].

Readability. We follow Ghose and Ipeirotis [11] as well as Blohm et al. [14] and measure the readability of the text by calculating the Coleman-Liau index [27] for each contribution. This index captures the complexity of the contributions by analyzing part-of-speech tags and measuring the average length of their terms and sentences. A higher index indicates a better readability for the text.

Spelling. Finally, we measure irregularities and non-conformance to writing standards by counting the number of spelling errors per contribution. In order to ensure that the spelling errors were accurately captured, we manually reviewed all 309 contributions.

5 Models and Results

5.1 Explanatory Regression Analysis

In this section, we use regression modeling to analyze whether the textual features of the contributions are associated with their quality. The length of a contribution, the specificity of the terms, the readability of the text, and the number of spelling errors represent the independent variables. The contribution quality as rated by the experts represents the dependent variable. The results are depicted in **Table 2**.

Table 2. Regression Analysis

Coefficient	Estimate	Std. Err.	t-value	p-value	
(Intercept)	2.890	0.039	74.395	< 2.2e-16	***
Length	12.091	0.783	15.451	< 2.2e-16	***
Length (poly 2)	-4.730	0.694	-6.813	5.18e-11	***
Length (poly 3)	2.930	0.710	4.124	4.82e-05	***
Specificity	1.752	0.721	2.429	0.016	*
Readability	2.333	0.708	3.297	0.001	**
Spelling	-1.847	0.814	-2.269	0.024	*

Note: ***p < 0.001; **p < 0.01; *p < 0.05

Residual Standard Error: 0.683; R-Sq. (adj.): 0.554; F(6,302): 64.8; p-value: < 2.2e-16

It shows that the length (t = 15.451; SD = 0.783; p = < 2.2e-16) and the readability (t = 3.297; SD = 0.708; p = 0.001) of a contribution are highly significant indicators for its quality. Both features are positively correlated to the quality of the contribution. Interestingly, as indicated by the polynomials, we observe a diminishing marginal utility effect associated with number of words in a contribution, which seems conceptually reasonable. Writing 55 instead of 5 words benefits the contribution more than extending it from 150 to 200 words. Regardless of this effect, our results still support H_1 which states that the length of a textual contribution is positively associated with the quality of the contribution. The model also supports H_3 and shows that the readability of a textual contribution is positively associated with the quality of the contribution. Similarly, the specificity of the terms (t = 2.429; SD = 0.721; p = 0.016) and the number of spelling mistakes (t = -2.269; SD = .814; p = 0.024) in a contribution are significant indicators for its quality. The former is positively correlated to the quality of the contribution. The latter is negatively correlated to the quality of the contribution. These results support H_2 and H_4. The model reveals a high value for R^2 and explains the quality of the contributions significantly well. We found no evidence that potential effects between the individual contributors and their contributions affect our results. We also examined the residuals and found our model to be sound. There are no signs of heteroscedasticity nor autocorrelation. The residual show to be normally distributed. We can conclude that the proposed variables explain the quality of crowdsourced contributions at statistically significant levels. We find support for our four hypotheses and will use these findings as the foundation for predictive modeling.

5.2 Predictive Modeling

Based on the previously analyzed variables, we train and evaluate a classifier that is capable of predicting the quality of the contributions and automatically filter them. A binary classification allows for a clear selection rule [2] that decides on whether the contributions fulfill the quality requirements and are thus eligible to be forwarded to the organization for further consideration or whether they are of poor quality and should be filtered out to not induce unnecessary workload. Hence, we represent the evaluation of the contributions as a classification problem. We set the threshold for separating high quality from low quality feedback to 3.5, which is comparable to previous studies conducted for product reviews [11], and labeled the contributions. As a result, 83 contributions were classified as high-quality contributions, whereas 226 contributions were classified as low-quality contributions. This distribution is consistent with findings documented in previous studies on the quality of crowdsourced contributions [11, 2]. We tested different classification algorithms and compared the performance of Logistic Regression, Naïve Bayes, k-Nearest Neighbor, Decision Trees, and Random Forest for this study. We found the Random Forest algorithm to perform substantially better in classifying the contributions compared to the other approaches – both regarding the accuracy and the receiver operating characteristic. Our findings are consistent with comparative experiments conducted for similar classification tasks [11]. Thus, we focus on the results of the Random Forest algorithm.

The Random Forest algorithm [38] builds a large number of decision trees with different combinations of the given variables. These decision trees are internally trained and evaluated using random subsets of the same data. The Random Forest model then averages the decision trees. In this vein, it reduces the variance that comes with individual decision trees, provides information about the importance of the variables for the classification, and overcomes the risk of overfitting [39].

We use 100 decision trees for our Random Forest model and set the cutoff for the model's probability estimates at the standard value of 0.5. To build and evaluate the classifier, we followed the widely used k-fold cross-validation approach with 5 folds. That is, we randomly split our data set in a stratified manner into 5 subsets. 4 subsets are used to train the classifier with the given labels. The remaining subset does not include the quality labels and is used evaluate the performance of the classifier by comparing the labels predicted by the Random Forest algorithm to the actual labels provided by the experts. We measure the accuracy, the sensitivity, the specificity[1], and the receiver operating characteristic [40]. This procedure is repeated until each split of the data set has been used to train and evaluate the classifier.

The results of the cross-validation reveal an accuracy of 80.03% on average for our Random Forest model. Thus, by only using the four variables based on our proposed textual features, the algorithm is able to automatically predict the quality of the crowdsourced contributions and correctly classify them in over 80% of the cases. The classifier shows a very high specificity of 87.73%, indicating that it performs exceptionally well at recognizing and filtering low quality contributions. As suggested by the slightly lower sensitivity measure (60.27%), it is more difficult for the algorithm to achieve a high true positive rate. The sensitivity of the classifier can be increased by adjusting the cutoff for the probability estimates. Lowering the cutoff by 20% increases the classifier's sensitivity to 75.30%. Naturally, however, this comes at the expense of reducing its specificity to 76.56%.

[1] It is important to note that, in this context, specificity refers to a statistical measure that describes the true negative rate.

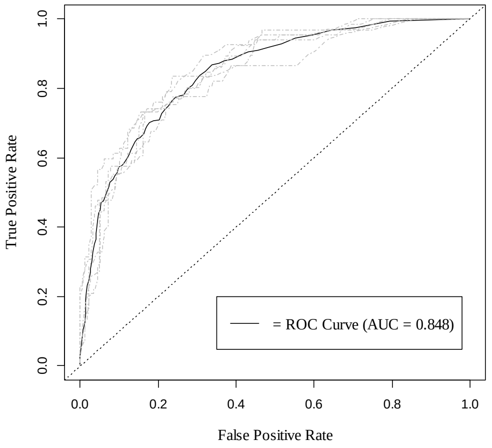

Figure 1. Receiver Operating Characteristic

The curve of the classifier's receiver operating characteristic (ROC) is depicted in **Figure 1**. It plots the true positive rate against the false positive rate [40]. The diagonally plotted line represents the strategy of randomly guessing the quality of the contributions. A classifiers that reaches the upper triangular region of this line exploits information in the data and performs better than the random classification strategy [40]. The area under curve (AUC) is equivalent to "the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance" [40], making it also equivalent to the Wilcoxon test of ranks. Here, the AUC reveals a high value of 0.848. Hour classification algorithm performs very well.

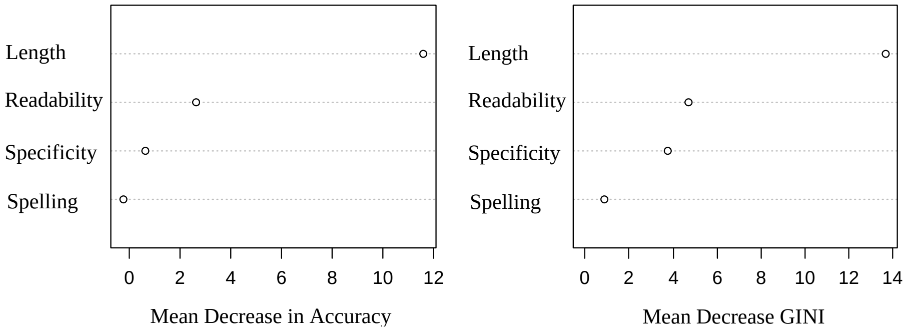

Figure 2. Variable Importance Plots

Finally, **Figure 2** displays the importance of the four proposed variables, measured by the mean decrease in accuracy and the mean decrease in node impurity (i.e., Gini index) for each variable [41]. All variables were used by the Random Forest algorithm and have predictive power. The length of the contribution is by far the most important variable for the classification. When aiming for a sparse prediction model, the variable "Spelling" may be omitted without risking much worse results.

6 Discussion and Implications

The models and results presented in the previous section yield two important findings. First, we find support for our hypotheses and show that the length of a contribution, the specificity of the terms, the readability of the text, and the number of spelling errors are all associated with contribution quality in crowdsourcing at statistically significant levels. Therefore, in a second step, we used the textual characteristics in combination with an expert-based baseline measure for contribution quality to train and evaluate an algorithm capable of predicting the contribution quality and classifying the data. Even for a small data set of 309 contributions, our Random Forest classifier achieves an accuracy of 80.03%. The algorithm has shown to perform especially well at recognizing and filtering low quality contributions. It outperforms random classification substantially and also achieves a much higher accuracy compared to a naïve classifier that would always predict the category with the majority of the ratings (i.e., 73.14%). Thus, our Random Forest algorithm proves to be very reliable. These findings have valuable implications for both researchers and practitioners alike.

6.1 Theoretical Implications

To the best of our knowledge, we are the first to show that it is possible to reliably explain and predict the quality of contributions in crowdsourcing based on textual features of the data alone. We provide empirical evidence for the relationship between both contextual and representational characteristics of contributions and their quality in crowdsourcing. This indicates that well elaborated and precise solutions, ideas, or suggestions are vital for companies trying to leverage the information submitted by a crowd. Furthermore, our results suggest that companies require textual contributions to be presented in a clear and easily interpretable manner to fully benefit from them.

Moreover, we contribute a set of models and variables to operationalize these contextual and representational characteristics. The models and variables proposed in our study have been shown to work well with algorithms capable of automatically assessing and classifying textual contributions. In this vein, we provide the foundation for partially automating the evaluation of textual data in crowdsourcing, which has frequently been requested by related literature. Kittur et al. [10] emphasized that, while "quality control is improving for tasks with a closed set of possible answers, we still have few techniques for open-ended work and highly skilled tasks" (p. 7-8). The authors specifically asked for studies to analyze potential metrics and propose feasible approaches to predict output quality. In crowdsourced software testing in particular, related work has expressed the need for efficient mechanisms to assess the quality of crowdsourced contributions and automate the evaluation of the data [7]. With our study, we close this gap and extend existing research that already uses machine learning and text mining algorithms to cluster the variety of topics covered in crowdsourcing projects [6, 18–20] by providing both the appropriate variables and models for an automated evaluation of high quality and low quality contributions in the potentially large sets of textual data. Regarding the importance of different variables, we found the length of a contribution to be the most effective indicator for explaining and predicting its quality. Both the readability of the contributions and the specificity of the terms are positively associated with the quality of the contributions at highly significant levels but reveal only moderate predictive power for classification algorithms. Interestingly, spelling errors have shown to be the least important feature for the classification and may even be omitted for sparse models. Therefore, our findings may help researchers in selecting variables for predictive modeling in crowdsourcing.

6.2 Practical Implications

Our proposed Random Forest classifier allows companies to substantially reduce the amount of information that needs to be reviewed manually. It shows that the classification algorithm is capable of automatically identifying high quality contributions in large data sets and removing those that do not fulfill the quality requirements defined by the companies or platforms. In this study, we set the threshold to only include the top 30% of the contributions. Hence, the algorithm can make the evaluation of the results submitted to crowdsourcing projects much more efficient and offers both time and cost savings. It is possible to incorporate the algorithm directly as a filter mechanism on the platforms or in tools for companies retrieving data from these platforms.

We also show that the sensitivity and specificity of the Random Forest algorithm can be adjusted to fit the preferences of practitioners. As both measures are inherently linked to each other, the decision to increase one measure will always come with the trade-off of decreasing the other. If the costs of wrongfully rejecting a high quality contribution is higher than the cost of wrongfully including a low quality contribution in the evaluation process, this is a trade-off that should potentially be considered.

Finally, our automated machine learning and text mining approach also contributes to practitioners in the domain of software testing. Related work already proposes algorithms that can be used to evaluate technical bug reports more efficiently. For example, it is possible to automatically assess the severity of the bug reports [42] and detect duplicates in the data sets [43]. As our data stem from crowdsourced software testing, we extend these findings and provide developers with an approach to facilitate the evaluation of test reports obtained in user acceptance testing, user experience testing, or usability testing. These contributions are typically submitted in a free text format and entail a high workload for the developers [7]. Our proposed classifier may help developers in evaluating these types of test reports more efficiently.

6.3 Limitations and Future Research

As with any research, our work does have its limitations. First, the manually assigned quality labels used for our data set are inherently dependent upon the rating scales and the subjective judgements of the experts. We attempted to address this issue by using scales that have been developed specifically for crowdsourced contributions as analyzed in this study [14]. Furthermore, we let two experts independently review the contributions. The Cohen's Kappas indicate an intersubjective agreement between the experts. Second, the data set stems from a crowdsourcing project in the field of software testing. We aimed to provide as much generalizability as possible by choosing a user acceptance test setting that yields contributions similar to other crowdsourcing contexts that are based on textual data, such as ideas, feedback, or reviews.

The findings presented in this study may encourage future efforts to analyze the performance of the proposed features or models in different crowdsourcing settings and expand on our initial results. There is still great potential in making the algorithms cost-sensitive and studying the optimal trade-off between sensitivity and specificity in crowdsourcing. Furthermore, as we focused on the textual characteristics of a contribution, future work may also examine the role of non-textual characteristics and analyze features such as the experience or the expertise of the individuals who submitted the contributions. Finally, text mining and machine learning methods benefit from large data sets. Hence, we need scalable concepts for labeling crowdsourced contributions and training algorithms with more data. Addressing these issues would pave the way for leveraging the full potential of machine learning in crowdsourcing.

7 Conclusion

The process of manually reviewing and filtering large volumes of textual contributions has been a longstanding challenge in crowdsourcing. Given the unstructured format of textual data and the diversity of inputs submitted by a crowd, identifying valuable inputs and separating them from low quality contributions that cannot be used by the companies is very time-consuming and cost-intensive. In this study, we propose an approach based on the principles of text mining and machine learning to partially automatize this process. Our results indicate that it is possible to explain the quality of crowdsourced contributions purely based on textual features, such as the length of a contribution, the specificity of the words, the readability of the text, and the number of spelling errors. We use these textual features in combination with an expert-based baseline measure to train and evaluate a classification algorithm that is capable of reliably predicting the quality of the contributions and automatically filtering them for companies.

References

1. Howe, J.: The Rise of Crowdsourcing. Wired Mag. 14, 1–5 (2006).
2. Blohm, I., Leimeister, J.M., Krcmar, H.: Crowdsourcing: How to Benefit from (Too) Many Great Ideas. MIS Q. Exec. 12, 199–211 (2013).
3. Afuah, A., Tucci, C.L.: Crowdsourcing as a Solution to Distance Search. Acad. Manag. Rev. 37, 355–375 (2012).
4. Jeppesen, L.B., Lakhani, K.R.: Marginality and Problem-Solving Effectiveness in Broadcast Search. Organ. Sci. 21, 1016–1033 (2010).
5. Leimeister, J.M., Huber, M., Bretschneider, U., Krcmar, H.: Leveraging Crowdsourcing: Activation-Supporting Components for IT-Based Ideas Competition. J. Manag. Inf. Syst. 26, 197–224 (2009).
6. Barbier, G., Zafarani, R., Gao, H., Fung, G., Liu, H.: Maximizing Benefits from Crowdsourced Data. Comput. Math. Organ. Theory. 18, 257–279 (2012).

7. Zogaj, S., Bretschneider, U., Leimeister, J.M.: Managing Crowdsourced Software Testing: A Case Study Based Insight on the Challenges of a Crowdsourcing Intermediary. J. Bus. Econ. 84, 375–405 (2014).
8. Chen, H., Chaing, R.H.L., Storey, V.C.: Business Intelligence and Analytics: From Big Data to Big Impact. MIS Q. 36, 1165–1188 (2012).
9. Debortoli, S., Müller, O., Junglas, I.A., vom Brocke, J.: Text Mining for Information Systems Researchers: An Annotated Tutorial. Commun. AIS. 1–30 (2016).
10. Kittur, A., Nickerson, J. V., Bernstein, Michael, S., Gerber, E.M., Shaw, A., Zimmermann, J., Lease, M., Horton, J.J.: The Future of Crowd Work. In: Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work, CSCW 2013. pp. 1–17. ACM, San Antonio (2013).
11. Ghose, A., Ipeirotis, P.G.: Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. IEEE Trans. Knowl. Data Eng. 23, 1498–1512 (2011).
12. Lukyanenko, R., Parsons, J., Wiersma, Y.F.: The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-Generated Content. Inf. Syst. Res. 25, 669–689 (2014).
13. Wang, R.Y., Strong, D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. J. Manag. Inf. Syst. 12, 5–34 (1996).
14. Blohm, I., Riedl, C., Füller, J., Leimeister, J.M.: Rate or Trade? Identifying Winning Ideas in Open Idea Sourcing. Inf. Syst. Res. 27, 27–48 (2016).
15. Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H.R., Bertino, E., Dustdar, S.: Quality Control in Crowdsourcing Systems: Issues and Directions. IEEE Internet Comput. 17, 76–81 (2013).
16. Riedl, C., Blohm, I., Leimeister, J.M., Krcmar, H.: The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities. Int. J. Electron. Commer. 17, 7–36 (2013).
17. Ghose, A., Ipeirotis, P.G., Li, B.: Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. Mark. Sci. 31, 493–520 (2012).
18. Walter, T.P., Back, A.: A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests. In: Proceedings of the 46th Hawaii International Conference on System Sciences, HICSS. pp. 3109–3118. IEEE, Waikoloa, Hawaii (2013).
19. Feng, Y., Chen, Z., Jones, J.A., Fang, C., Xu, B.: Test Report Prioritization to Assist Crowdsourced Testing. In: Proceedings of the 10th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2015. pp. 225–236. ACM, Lombardy (2015).
20. Rogstadius, J., Vukovic, M., Teixeira, C.A., Kostakos, V., Karapanos, E., Laredo, J.A.: CrisisTracker: Crowdsourced Social Media Curation for Disaster Awareness. IBM J. Res. Dev. 57, 1–13 (2013).
21. Otterbacher, J.: “Helpfulness” in Online Communities: A Measure of Message Quality. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI ’09). pp. 955–964. ACM, Boston (2009).
22. Jeon, J., Croft, W.B., Lee, J.H., Park, S.: A Framework to Predict the Quality of Answers with Non-Textual Features. In: Proceedings of the 29th Annual International ACM SIGIR Conference. pp. 228–235. ACM, Seattle, Washington (2006).
23. Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., Zhou, M.: Low-Quality Product Review Detection in Opinion Summarization. In: Proceedings of the 2007 Join Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 334–342. , Prague (2007).

24. Kim, S.-M., Pantel, P., Chklovski, T., Pennacchiotti, M.: Automatically Assessing Review Helpfulness. In: Proceedings of the 2006 Conference of Empirical Methods in Natural Language Processing (EMNLP 2006). pp. 423–430. , Sydney (2006).
25. Weimer, M., Gurevych, I.: Predicting the Perceived Quality of Web Forum Posts. In: Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP). pp. 643–648. , Borovets, Bulgaria (2007).
26. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding High-Quality Content in Social Media. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. pp. 183–193. ACM, Palo Alto (2008).
27. Coleman, M., Liau, T.L.: A Computer Readability Formula Designed for Machine Scoring. J. Appl. Psychol. 60, 283–284 (1975).
28. Amabile, T.M.: Social Psychology of Creativity: A Consensual Assessment Technique. J. Pers. Soc. Psychol. 43, 997–1013 (1982).
29. Nørgaard, M., Hornbæk, K.: Exploring the Value of Usability Feedback Formats. Int. J. Hum. Comput. Interact. 25, 49–74 (2009).
30. Cohen, J.: Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. Psychol. Bull. 70, 213–220 (1968).
31. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics. 33, 159–174 (1977).
32. Barki, H., Pinsonneault, a.: Small Group Brainstorming and Idea Quality: Is Electronic Brainstorming the Most Effective Approach? Small Gr. Res. 32, 158–205 (2001).
33. Blohm, I., Bretschneider, U., Leimeister, J.M., Krcmar, H.: Does Collaboration Among Participants Lead to Better Ideas in IT-based Idea Competitions? An Empirical Investigation. In: Proceedings of the 43rd Annual Hawaii International Conference on System Sciences, HICSS 2010. pp. 1–10. IEEE, Honolulu (2010).
34. Gallupe, R.B., Dennis, A.R., Cooper, W.H., Valacich, J.S., Lana, M., Nunamaker, J.F.: Electronic Brainstorming and Group Size. Acad. Manag. J. 35, 350–369 (1992).
35. Hotho, A., Nürnberger, A., Paaß, G.: A Brief Survey of Text Mining. LDV Forum - Gld. J. Comput. Linguist. Lang. Technol. 20, 19–62 (2005).
36. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge (2007).
37. Porter, M.F.: An Algorithm for Suffix Stripping. Program. 14, 130–137 (1980).
38. Breiman, L.: Random Forests. Mach. Learn. 45, 5–32 (2001).
39. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2009).
40. Fawcett, T.: An Introduction to ROC Analysis. Pattern Recognit. Lett. 27, 861–874 (2006).
41. Liaw, A.: Package “randomForest,” https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.
42. Lamkanfi, A., Demeyer, S., Soetens, Q.D., Verdonckz, T.: Comparing Mining Algorithms for Predicting the Severity of a Reported Bug. In: 15th European Conference on Software Maintenance and Reengineering, CSMR 2011. pp. 249–258. IEEE, Oldenburg (2011).
43. Runeson, P., Alexandersson, M., Nyholm, O.: Detection of Duplicate Defect Reports Using Natural Language Processing. In: Proceedings of the 29th International Conference on Software Engineering, ICSE 2007. pp. 499–508. IEEE, Minneapolis (2007).