Accurate, Fast, But Not Always Cheap: Evaluating "Crowdcoding" as an Alternative Approach to Analyze Social Media Data

Article *in* Journalism & Mass Communication Quarterly · December 2019

DOI: 10.1177/1077699019891437

Lei Guo
Boston University
44 PUBLICATIONS **1,144** CITATIONS
SEE PROFILE

Kate Mays
Boston University
12 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Sha Lai
Boston University
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Mona Jalal
Boston University
13 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Journal & Mass Communication Quarterly
1–24
© 2019 AEJMC
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1077699019891437
http://journals.sagepub.com/home/jmq
SAGE

Accurate, Fast, But Not Always Cheap: Evaluating “Crowdcoding” as an Alternative Approach to Analyze Social Media Data

Lei Guo[1], Kate Mays[1], Sha Lai[1], Mona Jalal[1], Prakash Ishwar[1], and Margrit Betke[1]

Abstract
Crowdcoding, a method that outsources “coding” tasks to numerous people on the internet, has emerged as a popular approach for annotating texts and visuals. However, the performance of this approach for analyzing social media data in the context of journalism and mass communication research has not been systematically assessed. This study evaluated the validity and efficiency of crowdcoding based on the analysis of 4,000 tweets about the 2016 U.S. presidential election. The results show that compared with the traditional quantitative content analysis, crowdcoding yielded comparably valid results and was superior in efficiency, but was more expensive under most circumstances.

Keywords
content analysis, crowdcoding, crowdsourcing, Twitter, sentiment analysis

Analyzing content is central to communication research because all human communication involves messages, or “content.” Quantitative content analysis (QCA) has been proposed as a robust way to analyze many types of content because it enables a reproducible analysis when appropriately implemented (Krippendorff, 2004; Lacy et al., 2015). Although a powerful methodology for traditionally sized data sets, QCA is

[1] Boston University, MA, USA

Corresponding Author:
Lei Guo, Division of Emerging Media Studies, Boston University, 704 Commonwealth Ave., 302D, Boston, MA 02215, USA.
Email: guolei@bu.edu

limited in its efficiency and cost for "big data"—data sets that are extreme in volume, velocity, variety, dimension, and scope (Beyer & Laney, 2012). An advanced sampling strategy may help reduce data size, but it is not always easy to create a sound sampling frame for big social data (Riffe et al., 2014). Computational methods such as dictionary-based analysis and machine learning offer alternative solutions; still, ample human coding is needed to train and evaluate computer models (Guo et al., 2016).

This study evaluates an emerging approach, different from QCA, that can potentially be used to analyze content in communication research: *crowdcoding* (Haselmayer & Jenny, 2017), which outsources "coding" tasks to a large pool of annotators online. In crowdcoding, each analytical unit is annotated independently by more than one crowdworker and their judgments are suitably aggregated to make decisions. With the large pool of internet workers available, crowdcoding uses human reasoning to annotate a large amount of verbal or nonverbal content in a short time period.

Crowdcoding has been widely used in computer science to train, evaluate, and interpret algorithmic output. Recently, the approach has also begun to attract attention in political science (Benoit et al., 2016; Haselmayer & Jenny, 2017). However, researchers have not yet explored the applicability of this method in journalism and mass communication research, with few exceptions (e.g., Lind et al., 2017). In particular, the validity and reliability of using crowdcoding to annotate social media data have not been explored.

This study evaluates the crowdcoding approach for analyzing the Twitter public's sentiment toward politicians. The analyses were conducted on two major crowdsourcing platforms, Amazon Mechanical Turk and Figure Eight. Crowdworkers on each platform annotated the same 4,000 tweets about the 2016 U.S. presidential campaign, and their results' validity was assessed based on ground truth labels provided by domain experts. The crowdcoding approach was also compared with QCA, in terms of validity and the time and cost to complete the analysis.

QCA

Defined as "a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use" (Krippendorff, 2004, p. 18), QCA usually involves drawing a representative sample of textual or visual data, training two or more human coders on a coding protocol to identify differences in content, and measuring intercoder reliability (ICR) between the coders' results. QCA is an important and ubiquitous method in communication research because it allows researchers to analyze human communication in a non-obtrusive manner and can be applied to answer a variety of research questions (Riffe et al., 2014).

In particular, strong coding protocols and systematic ICR testing help check human coders' potential bias and provide a basis for replicating studies (Lovejoy et al., 2016). The ICR test measures how much two or more human coders agree in their application of a coding protocol to categorize content units. Coders cannot start analyzing the entire data set until they have reached a certain degree of ICR because high reliability often indicates high validity of the coded results.

As with any research method, QCA has its limitations. Foremost, the approach is time-consuming and labor-intensive, as the work is done entirely "by hand" and only by a few human coders (Krippendorff, 2004). Although this limitation can be overcome to some extent with a sampling strategy when analyzing structured data (e.g., newspaper articles), it is not always applicable when analyzing social media data such as tweets. It can be difficult to create representative samples because the population of tweets is unknowable and inherently unstable over time (Riffe et al., 2014), and the validity and reliability of user level data such as demographic information and geolocation are not guaranteed (Kim et al., 2013).

Computer-assisted programs based on built-in dictionaries can help automate the content analysis, but the accuracy of computer-generated results remains questionable. For example, SentiStrength is a lexicon-based classifier that uses an existing set of terms and additional linguistic information to detect sentiment in short informal English text (Thelwall et al., 2012). In the creators' own assessment, SentiStrength's accuracy predicting the positive and negative sentiment of six types of social media data (including tweets) only averaged 59.2% and 66.1%, respectively (Thelwall et al., 2012). One limitation of a lexicon-based approach is its confinement to a fixed set of words; on social media, new expressions and jargons emerge regularly.

Supervised machine learning (SML), another kind of computer-assisted analysis, requires a large amount of manual labeling and evaluation. Collingwood and Wilkerson (2012) found that, for an SML model to reach a 0.79 accuracy in categorizing U.S. congressional bills across 20 topics, they had to hand-code 1,000 bills per topic for constructing training and testing sets. QCA is one of the most used methods for generating manual labels for SML, but its lack of efficiency remains a concern.

There have been other problems in how researchers have employed the QCA methodology. Not all researchers have implemented ICR tests (Neuendorf, 2017; Riffe et al., 2014), and when they have, some have failed to meet ICR standards (Lovejoy et al., 2016). Robust ICR can be difficult to achieve because of certain coding challenges, such as a lack of a common frame of reference, language skill differences, or simply fatigue and boredom among coders (Riffe et al., 2014). Furthermore, post-ICR reviews are rare; once an acceptable ICR is achieved, researchers usually have their coders independently annotate the data without any more check-ins, which may leave potential misinterpretations unnoticed. Moreover, while establishing ICR helps readers evaluate the results' validity, high reliability is a necessary but not a sufficient condition for validity (Lovejoy et al., 2016).

In sum, QCA can provide high-quality results but may be limited in several aspects. Recently, crowdcoding has emerged as an alternative approach for annotating data. This study examines whether crowdcoding may be able to address some of QCA's shortcomings.

From Crowdsourcing to Crowdcoding

Crowdsourcing has been defined as taking a task, once performed by an employee or a firm, and outsourcing it to a generally large group of people through an open call, typically over the internet (Howe, 2008). A relatively recent concept, crowdsourcing has applicability across diverse fields, including urban planning, public policy-making, and journalism.

Crowdcoding[1] (Haselmeyer & Jenny, 2017) is one specific use of crowdsourcing, which relies on nonexpert, lay people to annotate content through online platforms such as Amazon Mechanical Turk (MTurk) and Figure Eight (F8). On these platforms, individuals or businesses can post a task, and workers can choose tasks based on their interests. For crowdcoding, researchers invite two or more crowdworkers to annotate the same piece of content (e.g., a tweet) independent of each other; usually, different groups of crowdworkers annotate different pieces of content. To make the final decision on a piece of content's coding, researchers implement an aggregation approach, combining crowdworkers' judgments in certain ways, like taking the majority vote. To clarify, researchers also conduct surveys or online experiments on crowdsourcing platforms, and the goal is often to examine respondents' characteristics and the relationships between them. Crowdcoding, on the other hand, is employed to obtain objective annotations from crowdworkers based on the instructions provided. In other words, crowdcoding can be used as an alternative approach to QCA to annotate communication content.

In computer science, crowdcoding has emerged as a popular tool to facilitate the design and tests of new algorithms in a variety of natural language processing (NLP) and computer vision tasks by providing human annotations for SML. In NLP, for example, crowdcoding has been used to code consumer sentiment about certain products (Mellebeek et al., 2010), detect emotions in news headlines (Snow et al., 2008), and identify the names of drugs and treatments in clinical trial announcements (Zhai et al., 2013). Researchers generally agree that crowdworkers' labels are a viable and cost-effective alternative to expert annotations. Social scientists have also begun to use this approach to analyze text documents. In political science, crowdcoding generated results comparable to expert decisions in different settings when analyzing political manifestos and a multilingual debate in the European Parliament (Benoit et al., 2016). Crowdcoding has also been employed in political communication, to estimate the sentiment in press releases, minutes from parliamentary debates, and media reports on Austrian election campaigns (Haselmayer & Jenny, 2017), and for analyzing the latent content of news texts in political actor evaluations (Lind et al., 2017). Both studies suggested that the group of lay coders effectively replicated the expert data.

Unlike well-structured text data such as news articles or press releases, social media data are an extreme version of informal text and presents a significant challenge to content analysis. For example, tweets are often messy, truncated, and contain irony or sarcasm (Guo et al., 2016). Studies exploring the use of crowdcoding in analyzing tweets have found inconsistent results. Finin et al. (2010) recruited MTurk crowdworkers for identifying the names of people, companies, and locations mentioned in tweets. Each tweet was annotated by two MTurk workers and compared with the expert-generated labels, and the crowdworkers' annotations were not as accurate. Following Finin et al.'s (2010) work, Fromreid et al. (2014) raised the concern of quality control on MTurk. After manually examining 2,974 tweets from Finin et al.'s (2010) project, they found incidences of both spammers with random annotations, and annotators who did not understand the tweets' context. More recently, Vargas et al. (2016) asked three crowdworkers on CrowdFlower (now F8) to label tweets' sentiment about three crisis events and found the degree of agreement between crowdworkers was fair to moderate, though they did not assess the crowdcoded data's accuracy.

Given the inconclusive findings about using crowdcoding to analyze social media, this study provides a systematic evaluation of the method's validity and efficiency and focuses on one important variable in communication research: political sentiment on Twitter. Sentiment about politicians in news coverage or public discourse has been extensively analyzed in communication research to test theories such as agenda setting and framing. More recently, research on the public's sentiment toward political candidates on social media platforms such as Twitter sheds light on media polarization, echo chamber effects, and other emerging political communication phenomena. Considering the significant research interest in analyzing political sentiment in unprecedentedly large social media data sets, an evaluation of crowdcoding such data is a timely and potentially impactful research endeavor.

Validity of Crowdcoded Data

This article's central question is whether crowdcoding is a valid method for analyzing social media data. Potter and Levine-Donnerstein (1999) argue that determining how to set the standard for the validity of coded data depends on the type of content. For manifest content (e.g., the number of words in a tweet), the standard exists on the content's surface. Examining the validity of latent content—the underlying meaning of a message—is a more challenging task. There are two types of latent content: *Pattern content* analysis assumes that the content has an objective pattern that all coders should uncover by sorting through symbols, whereas *projective content* analysis is more subjected, relying more on coders' judgment to discern the content's meaning through their pre-existing mental schema (Potter & Levine-Donnerstein, 1999). They analyze gender stereotypes on television as an example: Measures of body type and fitness are pattern content, whereas a character's attractiveness is projective content. To examine pattern content's validity, experts must set the standard because they create the coding rules and thus should best understand their correct application. With projective content, the societal norm becomes the standard.

The difference between pattern and projective latent content is not a clean dichotomy, and one may argue that the interest of this analysis—Twitter sentiment—can be put in either category. However, as Potter and Levine-Donnerstein (1999) contend, most content analyses following social science principles assume an expert standard and objective coding protocol to establish the results' validity. This also aligns methodologically with most crowdcoding research, which examines the validity of crowdcoded data based on domain experts' coding results, termed as the "ground truth" (e.g., Benoit et al., 2016; Haselmayer & Jenny, 2016). For this analysis, we also consider

Twitter sentiment a type of pattern latent content and labels provided by domain experts the standard. We ask the following question:

> **Research Question 1 (RQ1):** To what extent are the crowdcoded data valid in analyzing Twitter sentiment toward political candidates?

Crowdcoded Data Collection and Aggregation

In crowdcoding, the term "crowd" refers to a large set of internet workers, numbering in hundreds or thousands, whose participation is facilitated by a crowdsourcing platform. Each unit of analysis (e.g., a tweet), however, is annotated by only a relatively small number of crowdworkers, and what that number should be is a pertinent research question.

Previous research has demonstrated that increasing the number of crowdworkers improves accuracy (e.g., Hara et al., 2013; Snow et al., 2008) because doing so will better "account for natural variability of human performance [and] reduce the influence of occasional errors . . ." (Sorokin & Forsyth, 2008, p. 2). Researchers have also observed that the accuracy gains due to decision aggregation will diminish in magnitude as group size grows.

A theoretical basis for these phenomena can be found in Theorem 2 of Sameki et al. (2019). Consider a simple scenario with binary decisions for each item (i.e., crowdworkers only have to choose one of two possible options). Suppose that the crowdworkers have a basic level of competency—the chance (probability) that their decision will match the ground truth is at least better than 0.5 (i.e., the likelihood of a match due to pure chance). If the decisions made on individual items are statistically independent across different crowdworkers, then combining the decisions of more crowdworkers via a majority vote will increase (or more precisely, not decrease) the probability of matching the ground truth. Moreover, a "law of diminishing returns" comes into effect as an increasing number of independent, competent decisions are aggregated, meaning that the improvements from the majority vote diminish with each additional competent decision.

Given the law of diminishing returns and the additional cost yielded by increasing the number of crowdworkers, researchers have compared the crowdcoding results based on differently sized groups of crowdworkers per unit. Hara et al. (2013) concluded that using more than five crowdworkers might not be worth the additional cost. Benoit et al. (2016) also recommend five independent judgments per unit of analysis. Still, the cost-effectiveness of using different group sizes may vary depending on the task, which we explore here as follows:

> **Hypothesis 1 (H1):** Increasing the number of independent, competent crowdworkers per tweet will increase the validity of the crowdcoded data, but the gains will diminish when more crowdworkers are included per tweet. If so,
>
> **Research Question 2 (RQ2):** What is the threshold beyond which further increasing the number of crowdworkers does not lead to increased coding validity?

Another critical question beyond group size is how to aggregate the crowdcoded responses to determine a single judgment on each unit of analysis. Researchers have used a variety of aggregation techniques (e.g., Benoit et al., 2016; Hara et al., 2013; Haselmayer & Jenny, 2017). Among all methods, a simple majority vote that aggregates each judgment independently is most straightforward and most widely used in crowdcoding research. As previously discussed, this aggregation method can be effective if the individual crowdworker decisions for any sample are independent, and each crowdworker has a basic level of competency. That is, this aggregation method would be problematic if a large number of crowdworkers are not competent. Therefore, a more sophisticated aggregation method that estimates individual crowdworker's competence could, in theory, improve over the simple majority rule. In practice, however, empirical evidence about the relative accuracy of different aggregation methods is mixed (Hung et al., 2013; Irshad et al., 2015). More complicated aggregation methods may not necessarily outperform simpler methods because the former approach typically requires larger amounts of information to be collected about individual crowdworkers or estimated from the observed data. If the necessary information cannot be collected (e.g., crowdworker biases on similar tasks) or the data are too "noisy" or messy for estimation, then it becomes a challenge to add meaningful information that will contribute toward improving the aggregation method's overall accuracy. Irshad et al. (2015) found that an aggregation method that accounts for crowdworkers' trust levels—annotation accuracy assessed in prior projects—did not translate to the present study because it might require a different set of knowledge. Furthermore, implementing more complicated aggregation techniques often requires additional cost and computation time (Hung et al., 2013). Together, research about the differences between various aggregation methods is inconclusive; additionally, the effectiveness of different aggregation approaches may vary depending on the coding tasks. Therefore, this study examines the extent to which different aggregation methods influence the crowdcoding results; specifically, those based on a simple majority vote and two weighted aggregation techniques:

> **Research Question 3 (RQ3):** To what extent does the validity of the crowdcoded data vary using different aggregation approaches?

Reliability of Crowdcoded Data

In QCA, ICR usually measures the degree of agreement between a pair of coders across analytical items. In crowdcoding, as mentioned earlier, the analytical items are usually coded by different groups of individuals. Therefore, ICR between two coders about their coding decisions across data, as assumed in Cohen's kappa, is not applicable here. Other more generalized ICR measures such as Fleiss' kappa (κ) and Krippendorff's alpha (α) have been created for instances when different groups of coders code different sets of subjects. Therefore, these reliability measures can be used to evaluate inter-*annotation* reliability (IAR) in the context of crowdcoding research.

Different from ICR, which examines the agreement among a fixed set of coders, IAR evaluates the agreement among annotations provided by different coders.

Another reliability measure, Gwet's AC_1, has been proposed to address older reliability statistics' deficiencies, underestimating reliability when the prevalence of traits is extremely low or high (Gwet, 2008). Given the controversy over appropriate reliability measures, Lacy et al. (2015) suggest that researchers calculate at least two measures of reliability—percent agreement and Krippendorff's alpha. In addition, Gwet's AC_1 should be calculated when the data have high levels of percent agreement but a low alpha. Following this suggestion, we examine IAR among crowdworkers using three reliability measurements—percent agreement, Krippendorff's alpha, and Gwet's AC_1:

> **Research Question 4 (RQ4):** To what extent do crowdworkers agree on the coding of Twitter sentiment toward political candidates?

Here, the meaning and standard of reliability for QCA and crowdcoding is worth further discussion. In performing QCA, a low ICR may suggest—among other reasons—that one or more coders cannot follow the coding protocol. Therefore, relying on them to code data independently, as often practiced in QCA, would be problematic. Unlike QCA, a key feature of crowdcoding is that a low IAR due to individual workers' coding errors can be compensated by the aggregation method. Because each analytical item's annotation is based on a group's decision rather than that of a single coder, low IAR does not necessarily mean low validity.

Still, IAR is important to crowdcoding because a low IAR may indicate other problems. For example, Alonso et al. (2014) found an extremely low IAR (as low as 0.013 α) among crowdworkers evaluating the interestingness of tweets, suggesting that the task may be too subjective to be systematically analyzed.

Finally, it is important to note that because the variety of coders is much higher than in QCA, reliability of crowdcoding is expected to be lower than that of QCA. Researchers have applied reliability measures to crowdcoding and suggested that an IAR coefficient of ~0.3 based on Krippendorff's alpha or Fleiss' κ is "fair" (Vargas et al., 2016, p. 697) or at least "tolerable" (Lind et al., 2017, p. 198). However, a "standard" for IAR in crowdcoding has not been formally examined, which remains a direction for future research.

Comparing Crowdcoding and QCA

To better understand the performance of crowdcoding and establish a benchmark for reference, this article compares crowdcoding and QCA. Although both approaches rely on human reasoning, crowdcoding and QCA as research procedures are distinct in numerous aspects. First, crowdworkers' long-term commitment to a project is not assured; again, a data sample is often annotated by a large number of crowdworkers with different groups of crowdworkers coding each unit. Second, crowdworkers are not traditionally trained (e.g., in face-to-face meetings) and extended coding instructions are uncommon (Lind et al., 2017). Third, while QCA relies on coders' independent judgments after they pass the ICR test, every single unit of data in crowdcoding is annotated by more than one crowdworker and their judgments are aggregated. Given these and other differences in research procedure, this study examines how crowdcoding and QCA may differ in terms of validity, as well as time and cost:

> **Research Question 5 (RQ5):** How does the crowdcoding approach perform as compared with QCA in terms of validity, and time and cost to complete the tasks of coding Twitter sentiment?

Research Design

The research is based on an analysis of tweets collected during the 2016 U.S. presidential campaign. Tweets mentioning the two final major-party nominees—Donald Trump and Hillary Clinton—during the second and third presidential debates were retrieved through Twitter's public streaming API. A simple random sample of 2,000 tweets about each candidate was drawn, so the complete data set contains 2,000 tweets mentioning Trump and 2,000 mentioning Clinton.[2] The tweets were labeled by three groups of coders: (a) domain experts for establishing ground truth labels, (b) student coders for conducting QCA, and (c) crowdworkers recruited from MTurk and F8 for implementing crowdcoding.

Ground Truth Labels

Two communication researchers developed a coding scheme to annotate Twitter sentiment toward the two political candidates. Based on previous literature, the "sentiment" variable was operationalized as each tweet's overall attitude toward the given candidate with four options: (a) positive, (b) neutral, (c) negative, and (d) N/A (not applicable, i.e., not about the candidate). Specific coding rules and examples were provided to operationalize each option (Supplemental Appendix A).

These researchers coded all 4,000 tweets to establish expert labels for ground truth. Specifically, they first independently coded 200 tweets outside of the coding sample and reached an ICR of 0.98 α for coding sentiment toward Trump, and 0.87 α toward Clinton, achieving agreement well above the 0.70 α threshold for a robust ICR (Lacy et al., 2015). The high ICR suggests that the two "expert" researchers agree on the operational definition of the variable and that their coded data are valid. The two researchers discussed their coding discrepancies on the 200 tweets thoroughly and then independently coded the sample of 4,000 tweets. Their coding results were compared and all disagreements (less than 5%) were discussed and adjudicated to reach one expert-coding sample. Having two experts make judgments independently adds another quality check to ensure the ground truth labels' validity.

QCA

Two communication graduate students were recruited to code a sample of 1,000 tweets mentioning Trump and another 1,000 tweets mentioning Clinton from the data set. Following QCA's standard procedure, the two communication researchers provided a codebook (Supplemental Appendix A) and training to instruct the two students on analyzing the tweets' sentiment. The student coders then independently coded 100-tweet samples (drawn outside of the coding sample) about Trump and Clinton, but they did not reach acceptable ICR in the first round of coding. In a second training session, the coding rules were further clarified and the students' disagreements discussed, and then they independently coded another 100-tweet sample per candidate and ICR was re-calculated. Ultimately, student coders reached an acceptable ICR for tweets about Clinton (α = .84) after three iterations and about Trump (α = .71) after six iterations. The students then independently coded 1,000-tweet samples for each candidate, so that there were two student samples of coding results for comparison.

Crowdcoding

Crowdcoding was conducted on MTurk and F8, the two most widely used crowdsourcing platforms, to compare and provide insight on the affordances of each. The task included two questions: (a) the tweet's sentiment toward Trump or Clinton, and (b) the coder's confidence in his or her judgment, which is used for one of the tested aggregation methods.

The same codebook used in QCA was provided, with some adjustments for the web presentation. See Supplemental Appendixes B and C for screenshots of the web interface, which was programmed identically on both platforms. For each task (coding one tweet), seven crowdworkers were invited to make judgments independently, and each crowdworker was requested to complete at least 10 tasks.[3] Two different projects were created on each platform to separate the crowdcoding tasks for tweets mentioning Trump or Clinton. On both platforms, two batches of 1,000 tasks per candidate were distributed per day.

Quality control. On MTurk, researchers can winnow their participating workers based on a number of criteria including prior activities on MTurk, demographic information, and media consumption habits. Based on previous literature (Sameki et al., 2016), we selected workers who completed at least 100 tasks on MTurk and had an approval rating of 92% or above. Because the task is about U.S. political communication, we also specified that workers should be from the United States.

Compared with MTurk, the selection criteria option on F8 are more limited. We specified on F8 that only U.S. workers could participate. In addition, F8 has a feature that rigorously checks the workers' coding decisions. Like other researchers (Vargas et al., 2016), we used this feature and required that workers keep their accuracy levels at or above 70% to remain on the project. To check this accuracy level, F8 required ground truth labels for 50 tweets per candidate, 10 for "quiz" and 40 for "test" tweets.

To participate in our project, a crowdworker needed to correctly label at least seven "quiz" tweets beforehand. F8 then interspersed the "test" tweets throughout the remaining tweets so that, as they were completing tasks, the crowdworkers would not know which ones were the "test" tweets. If their accuracy on these tweets dropped below 70%, their judgments were considered "untrustworthy" and removed from the analysis. This feature also provides more training for crowdworkers: when they incorrectly answer "quiz" or "test" questions, they are able to see the correct answer and the reasoning behind it.

Payment. Crowdsourcing can be exploitative when unpaid or underpaid labor replaces paid work. Ross et al. (2010) found that the U.S. MTurk worker's average wage was only US$2.30 per hour; it has recently increased to US$5.55 (Berg, 2016), but that remains less than the federal minimum wage (US$7.25). Although it may be true that some crowdworkers would have few employment alternatives beyond crowdsourcing work, it has been argued that responsible researchers should pay crowdworkers at least the minimum wage of their location (Silberman et al., 2018). We followed this ethical guideline for fair payment, first estimating the time it would take to complete each task based on student coding in QCA and a small-scale test on MTurk. We based payment on the minimum wage in the U.S. state where the research was conducted, calculating that we needed to pay crowdworkers US$0.10 for completing each task.

Data Aggregation

Three aggregation techniques were used to determine the sentiment toward a candidate in a tweet: simple majority vote and two different weighted approaches. In a simple majority vote, the final decision for each tweet was made based on the crowdworkers' most popular choice of sentiment. If there was a tie between two or more sentiment labels, then a random decision was made from these labels.

The other two aggregation approaches were based on weighted voting, taking into account crowdworkers' confidence in their answers or their "trust level" on F8. Given that sentiment is more difficult to determine in some tweets than in others, and that the ability to interpret the sentiment of certain types of tweets might vary across individual crowdworkers, we asked crowdworkers to indicate the confidence of their sentiment assessment for each tweet on a 5-point scale (1 = *not confident at all*, 5 = *very confident*). For example, if a tweet involves sarcasm (e.g., ". . . One of the great things about Donald Trump are the specifics he brings to his speeches"), some crowdworkers who notice the sarcasm would be more likely to make a correct decision, and presumably, they would be more confident about the decision. The following formula was used to generate the aggregated label for each tweet:

$$\textit{aggregated label} = \arg\max_{j \in J} \sum_{i \in I} 1\left(S_i = j\right) \times C_i$$

In this formula, S_i refers to the sentiment label decided by worker *i, J* refers to the set of sentiment labels (1 = *positive*, 2 = *neutral*, 3 = *negative*, 4 = N/A), C_i refers to the confidence score of worker i, and I refers to the set of worker indices (i.e., the crowdworker group size). "1" is an indicator function, which takes value 1 when the event holds true (i.e., S_i is indeed equal to j), and the value 0 otherwise. The notation "arg max" outputs j (i.e., the sentiment label), written under the "max" notation, that is associated with the maximum sum. The rationale behind the formula is straightforward: Each sentiment label is assigned a score, which is the sum of the confidence scores indicated by the crowdworkers who selected the given label. The label that received the highest score is determined as the final decision.

Alternatively, F8 provides a way to aggregate data by incorporating crowdworkers' "trust scores" (see figure-eight.com), which reflect their overall past performance on F8. The formula for aggregating data is similar to the one described above, with the trust score replacing the confidence level.

Data Analysis

Using the expert labels as the ground truth, the crowdcoded data's validity was assessed by calculating accuracy, precision, recall, and F score. In addition, Krippendorff's alpha and Gwet's AC_1 were used to compare the agreement between crowdcoding and the ground truth.

To examine to what extent the crowdworkers' group size influences the results, we adhered to the following protocol. Each tweet was coded by seven crowdworkers. To produce results for group sizes of two, three, four, five, and six crowdworkers, we conducted simulations. We simulated five crowdcoding experiments with fewer numbers of crowdworkers, say, x workers, by taking all possible subsets of cardinality x of the seven judgments. For example, to analyze the validity of crowdcoding based on a simple majority vote among five workers, all combinations of five workers ($N = 21$) were considered for a majority vote and the other two weighted aggregation approaches and all the validity scores were averaged.

Results

A total of 656 crowdworkers from MTurk participated in our project. Of these, 492 completed at least 10 tasks; they completed a median of 33 tasks, ranging from 10 to 431. The distribution of crowdworkers' accuracy scores is badly skewed to the left, with a median of 74% ranging between 0% and 100%. On F8, 211 crowdworkers coded Twitter sentiment, completing a median of 110 tasks, ranging from 10 to 1,430. The distribution of accuracy scores on F8 is nearly normal ($M = 0.72$, $SD = 0.10$).

Validity of Crowdcoded Data

Tables 1 and 2 detail the crowdcoded data's validity in analyzing Twitter sentiment toward Trump and Clinton, respectively. Overall, using crowdcoding to annotate tweets reaches a considerable level of accuracy, ranging from 66% to 83% for both politicians using varied numbers of crowdworkers per judgment and different

Table 1. Validity of Crowdcoded Data for Tweets About Clinton.

Platform	Aggregation	# workers	Accuracy	Precision	Recall	*F* score	α	AC_1
MTurk	Majority vote	7	0.81	0.81	0.81	0.81	.70	0.77
		6	0.81	0.81	0.81	0.80	.69	0.76
		5	0.81	0.80	0.81	0.80	.69	0.76
		4	0.80	0.79	0.80	0.79	.67	0.74
		3	0.79	0.78	0.79	0.78	.66	0.73
		2	0.74	0.73	0.74	0.73	.58	0.67
	Weighted by confidence	7	0.81	0.81	0.81	0.80	.70	0.77
		6	0.81	0.81	0.81	0.80	.70	0.76
		5	0.81	0.80	0.81	0.79	.69	0.76
		4	0.80	0.79	0.80	0.78	.67	0.74
		3	0.79	0.78	0.79	0.77	.66	0.73
		2	0.75	0.74	0.75	0.74	.60	0.69
F8	Majority vote	7	0.80	0.80	0.80	0.80	.68	0.74
		6	0.78	0.79	0.78	0.78	.65	0.72
		5	0.77	0.78	0.77	0.77	.64	0.71
		4	0.76	0.77	0.76	0.76	.62	0.69
		3	0.74	0.75	0.74	0.74	.60	0.67
		2	0.68	0.70	0.68	0.69	.51	0.60
	Weighted by confidence	7	0.80	0.80	0.80	0.79	.68	0.74
		6	0.78	0.79	0.78	0.78	.66	0.72
		5	0.78	0.78	0.78	0.78	.65	0.72
		4	0.76	0.77	0.76	0.76	.63	0.70
		3	0.74	0.75	0.74	0.74	.60	0.67
		2	0.70	0.70	0.70	0.70	.53	0.62
	Weighted by trust	7	0.79	0.80	0.79	0.79	.68	0.74
		6	0.78	0.79	0.78	0.78	.65	0.72
		5	0.77	0.78	0.77	0.77	.64	0.71
		4	0.75	0.76	0.75	0.75	.61	0.69
		3	0.74	0.75	0.74	0.74	.59	0.67
		2	0.67	0.69	0.67	0.68	.49	0.58

Note. The analysis was based on 1,789 tweets mentioning Clinton. Tweets that had at least one worker who did not complete a minimum of 10 tasks were removed from the analysis.

aggregation methods (**RQ1**). On MTurk, when five or more crowdworkers coded each tweet, the levels of accuracy all exceeded 80%. The precision and recall scores show similar patterns, indicating that crowdcoding tweets lean neither toward false positives nor false negatives. The Gwet's AC_1 coefficients for agreement between crowdcoded data and ground truth labels range from 0.59 to 0.79. With five or more crowdworkers, the AC_1 coefficients are all above 0.70, which is considered acceptable for communication research (Lacy et al., 2015). Agreement in terms of Krippendorff's alpha is slightly lower, perhaps because of the imbalance of the data (Gwet, 2008). According

Table 2. Validity of Crowdcoded Data for Tweets About Trump.

Platform	Aggregation	# workers	Accuracy	Precision	Recall	*F* score	α	AC$_1$
MTurk	Majority vote	7	0.83	0.83	0.83	0.82	.66	0.79
		6	0.81	0.82	0.81	0.81	.63	0.78
		5	0.81	0.81	0.81	0.80	.62	0.77
		4	0.79	0.79	0.79	0.79	.59	0.75
		3	0.78	0.78	0.78	0.77	.56	0.73
		2	0.71	0.73	0.71	0.71	.46	0.64
	Weighted by confidence	7	0.82	0.82	0.82	0.81	.64	0.79
		6	0.82	0.82	0.82	0.81	.64	0.78
		5	0.81	0.81	0.81	0.80	.62	0.77
		4	0.80	0.80	0.80	0.79	.60	0.76
		3	0.78	0.78	0.78	0.78	.57	0.73
		2	0.74	0.75	0.74	0.74	.49	0.68
F8	Majority vote	7	0.79	0.78	0.79	0.78	.58	0.75
		6	0.78	0.77	0.78	0.77	.55	0.73
		5	0.77	0.77	0.77	0.77	.55	0.73
		4	0.76	0.75	0.76	0.75	.52	0.71
		3	0.75	0.75	0.75	0.74	.50	0.69
		2	0.69	0.71	0.69	0.70	.41	0.63
	Weighted by confidence	7	0.79	0.78	0.79	0.78	.58	0.75
		6	0.78	0.78	0.78	0.78	.56	0.74
		5	0.77	0.77	0.77	0.77	.55	0.73
		4	0.76	0.76	0.76	0.76	.53	0.72
		3	0.75	0.75	0.75	0.74	.50	0.70
		2	0.70	0.71	0.70	0.70	.42	0.64
	Weighted by worker trust	7	0.78	0.78	0.78	0.78	.56	0.74
		6	0.76	0.76	0.76	0.76	.53	0.71
		5	0.76	0.76	0.76	0.76	.53	0.72
		4	0.74	0.74	0.74	0.74	.48	0.68
		3	0.74	0.74	0.74	0.74	.49	0.69
		2	0.66	0.68	0.66	0.67	.36	0.59

Note. The analysis was based on 1,655 tweets mentioning Trump. Tweets that had at least one worker who did not complete a minimum of 10 tasks were removed from the analysis.

to the ground truth labels, there are twice as many negative tweets (49.8%) than positive (21.0%) mentioning Clinton; an even larger portion of tweets about Trump are negative (65.6%), whereas only 8.3% of tweets are positive.

In comparing validity levels based on the number of crowdworkers per judgment (**H1**), Cochran's Q test was used to investigate the overall difference for each aggregation method and McNemar's test was used as a post hoc test to examine the pairwise difference, with the *p* values adjusted with a Bonferroni correction to prevent Type I error. The results show that increasing from two to three crowdworkers per tweet

Table 3. IAR of Crowdcoding Based on Seven Crowdworkers per Tweet.

Candidate	Platform	Percent agreement	α	AC_1
Clinton	MTurk	0.93	.50	0.63
	F8	0.91	.41	0.51
Trump	MTurk	0.91	.33	0.60
	F8	0.93	.34	0.56

Note. The analysis was based on 1,655 tweets mentioning Trump, and 1,789 tweets mentioning Clinton. Percent agreement refers to the percentage of tweets that achieve simple majority vote in sentiment annotation. To calculate alpha and AC_1 for crowdcoding, reliability data were tabulated in an M-by-N matrix where *M* is the number of crowdworkers and *N* is the number of tweets. Each entry at the position [i, j] is the *N* value (e.g., sentiment label) crowdworker *i* has assigned to unit *j*, or N/A if the coder has not annotated the unit. IAR = inter-*annotation* reliability.

significantly increases the annotation accuracy across the board (see Supplemental Appendix D). Further increasing the number of crowdworkers from three to five also significantly boosts the coding accuracy regardless of the aggregation method and platform used. However, only under certain circumstances does accuracy further increase when more than five crowdworkers per tweet are used. **H1** was supported. To answer **RQ2**, the results show that five crowdworkers per unit is the threshold of diminishing returns in most cases.

In answering **RQ3**, the three aggregation methods used—simple majority vote, weighted voting by confidence, and weighted voting by trust scores (F8)—produce similar results. According to McNemar's test, there is no significant difference between the aggregation methods on MTurk. On F8, simple majority vote and weighted aggregation by confidence are comparable, and both methods significantly outperform weighted aggregation by trust in annotation accuracy except for the case of three crowdworkers per judgment.

Reliability of Crowdcoded Data

In answering **RQ4**, three reliability coefficients were calculated to examine IAR (see Table 3). In addition, Supplemental Appendixes E and F visualize the number of tweets with a varied number of crowdworkers agreeing on the same label. In terms of percent agreement, the results show a high degree of homogeneity among crowdworkers in annotating tweets. For example, looking at the MTurk results (Supplemental Appendix E), a majority of crowdworkers (i.e., at least four out of seven) agreed on the same label for most tweets (93% for Clinton, 91% for Trump), and there was unanimous consensus for 36% of Clinton tweets and 24% of Trump tweets. The F8 results are similar (Supplemental Appendix F). IAR coefficients in terms of Krippendorff's alpha are lower: 0.50 α (MTurk) and 0.41 α (F8) for tweets mentioning Clinton, and 0.33 α (MTurk) and 0.34 α (F8) for tweets about Trump. Again, in light of the high percentage of agreement, the low alpha values may be caused by the data imbalance. When considering Gwet's AC_1, the IAR coefficients are higher and more stable, reaching 0.60 and 0.63 for crowdcoding tweets mentioning Trump and Clinton on MTurk, respectively.

Comparing Crowdcoding and QCA

To address **RQ5**, a sample of 1,000 tweets about each politician was drawn to compare crowdcoding and QCA in terms of coding validity, and time and cost to complete the analysis (see Table 4). Although both student coders went through the extensive training sessions and passed the ICR tests, they differ in their coding accuracy. Student 2 outperforms both Student 1 and the aggregated crowdcoding decisions, reaching 87% accuracy (0.74 α, 0.84 AC_1) for tweets about Trump, and 86% (0.78 α, 0.82 AC_1) about Clinton. Student 1's coding accuracy is slightly lower; notably, the crowdcoded data's coding validity surpasses student 1's when certain aggregation decisions are employed. Taken together, the findings show that crowdcoding can produce annotations of comparable quality to that of student coders in QCA, and in some circumstances can outperform QCA, which is contingent on the veracity of a single coder's judgment.

Crowdcoding is unequivocally more efficient than QCA. It took 2 to 7 hr on both MTurk and F8 to annotate a batch of 1,000 tweets. The project time is reported as 1 day because we posted one batch of 1,000 tweets each day. Comparatively, the student coders spent 22 and 17 (paid) hr to code tweets about Trump and Clinton, respectively. The entire project spanned 2 months: iterative training sessions and ICR tests took 6 weeks, and the actual coding lasted 2 weeks. Other research teams may be able to shorten the QCA procedure by reducing the time between training sessions or requiring coders to finish their work faster. Regardless, it is safe to say that crowdcoding was superior to QCA in terms of the coding speed, especially considering schedules in an academic setting.

Finally, the cost comparison shows that crowdcoding is generally more expensive than QCA. On MTurk, crowdcoding is cheaper than QCA when four or fewer crowdworkers code tweets about Trump, and three or fewer crowdworkers code tweets about Clinton. F8 charged more than MTurk because on F8 untrustworthy judgments—annotations by crowdworkers who fail 30% of the test questions at any point—were excluded from the analysis but their prior work was compensated. We incurred the extra cost on F8 from 982 untrustworthy judgments in tweets about Trump, and 667 untrustworthy judgments about Clinton.

Discussion

Crowdcoding has emerged as a popular approach for annotating texts and visuals in computer science, but its performance for analyzing social media data in journalism and mass communication research has not been systematically assessed. This study evaluated the validity and efficiency of crowdcoding based on the analysis of 4,000 tweets about the 2016 U.S. presidential election. The results show that crowdcoded data reach a considerable level of validity, providing annotations of Twitter sentiment

Table 4. A Comparison of Crowdcoding and QCA.[a]

			Time						
			Paid hours	Project length	# workers	Cost	Accuracy range	Alpha range	AC_1 range
Trump	Crowdcoding	MTurk	7 hr	1 day	7	US$840	0.81–0.84	0.62–0.67	0.77–0.81
					6	US$700	0.80–0.84	0.61–0.66	0.76–0.80
					5	US$600	0.80–0.83	0.60–0.65	0.75–0.80
					4	US$480	0.77–0.82	0.57–0.63	0.72–0.78
					3	US$360	0.75–0.80	0.54–0.60	0.70–0.76
					2	US$240	0.68–0.77	0.43–0.54	0.61–0.72
		F8	2 hr	1 day	7	US$1,073	0.77–0.80	0.53–0.61	0.73–0.76
					6	US$880[c]	0.75–0.80	0.48–0.61	0.70–0.76
					5	US$733	0.75–0.79	0.48–0.59	0.70–0.75
					4	US$587	0.72–0.78	0.43–0.57	0.67–0.73
					3	US$440	0.73–0.76	0.45–0.55	0.67–0.71
					2	US$293	0.65–0.71	0.32–0.47	0.57–0.65
	QCA	Student 1	22 hr[b]	2 months		US$510	0.82	0.66	0.80
		Student 2	22 hr	2 months		US$510	0.87	0.74	0.84
Clinton	Crowdcoding	MTurk	2 hr	1 day	7	US$840	0.81–0.82	0.69–0.71	0.76–0.77
					6	US$700	0.81–0.82	0.69–0.70	0.76
					5	US$600	0.81	0.68–0.69	0.76
					4	US$480	0.80	0.66–0.68	0.64–0.75
					3	US$360	0.79–0.80	0.65–0.67	0.73–0.74
					2	US$240	0.73–0.76	0.58–0.62	0.65–0.70
		F8	2 hr	1 day	7	US$1,033	0.76–0.83	0.61–0.73	0.69–0.78
					6	US$880	0.74–0.81	0.59–0.71	0.68–0.76
					5	US$733	0.74–0.81	0.58–0.70	0.67–0.75
					4	US$587	0.72–0.80	0.55–0.67	0.64–0.74
					3	US$440	0.71–0.78	0.53–0.66	0.63–0.72
					2	US$293	0.63–0.74	0.43–0.60	0.54–0.67
	QCA	Student 1	17 hr	2 months		US$360	0.81	0.70	0.76
		Student 2	17 hr	2 months		US$360	0.86	0.78	0.82

Notes. [a]This comparison was based on the analysis of 915 tweets mentioning Clinton, and 834 tweets mentioning Trump. Tweets that had worker(s) who completed less than 10 tasks were removed from the analysis. [b]The paid hours of QCA included time spent on training sessions, coding tweets for ICR tests, and the actual coding. [c]The cost needed to run crowdcoding on F8 based on two to six workers per judgment was estimated by F8. The actual cost might be slightly different depending on the number of untrustworthy judgments. QCA = quantitative content analysis; ICR = intercoder reliability.

comparable to QCA in certain scenarios. Crowdcoding is superior to QCA in the time it takes to complete the tasks; however, it can be more expensive to achieve a similar degree of accuracy.

Previous research demonstrates that crowdcoding can provide high-quality labels for sentiment analysis and named entity recognition in formal text like news articles and medical documents (e.g., Snow et al., 2018; Zhai et al., 2013). Using the standard in journalism and mass communication research, our study shows that crowdcoding is also capable of providing valid annotations of user sentiment in tweets, a type of informal text. With appropriate instructions, about one fourth of the crowdworkers applied the coding protocol as competently as a student coder trained for conducting QCA. While there certainly are crowdworkers who produce noise due to a lack of training or expertise, individual workers' poor judgments are offset largely by aggregation methods. In particular, when a sufficient number of crowdworkers annotate each tweet, the aggregated, crowdcoded data are as accurate as at least some trained student coders (e.g., Student 1 in our analysis). To provide more points of comparison, we used two popular sentiment analysis tools—SentiStrength and LIWC—to examine the sentiment of tweets in our data set, and both programs' average accuracy scores are no higher than 50% (see Supplemental Appendix G). Although sentiment analysis software may be faster and cheaper than crowdcoding, it has low accuracy. Hence, we would recommend crowdcoding over sentiment analysis software.

Notably, using different aggregation methods does not always make a significant difference in annotation accuracy. Consistent with Benoit et al. (2016), our study shows that a simple majority vote produces similar results as the other two more complicated approaches. This suggests that achieving high validity does not necessarily require a complicated research design. In particular, this study reveals that, in some cases, the aggregation method incorporating crowdworkers' trust scores on F8 is less accurate than simple majority vote, aligning with Irshad et al.'s (2015) finding that crowdworkers' coding competence does not always transfer across projects. Alternatively, the confidence scores were measured by asking for crowdworkers' self-assessment of how confident they were about their annotation for each tweet (see Supplemental Appendix C). We found that this aggregation method does not generate results significantly better than simple majority vote, suggesting that additional information about crowdworkers is not necessarily reliable enough to predict overall coding accuracy.

In terms of the number of workers per judgment, the study shows that increasing crowdworkers significantly boosts annotation accuracy in most scenarios but not all. In particular, increasing the group size to more than five does not always help, which may be attributed to a law of diminishing returns for aggregation methods. It can be mathematically proved that the probability that a simple majority of independent and identically distributed decisions agrees with the ground truth is a (discrete) concave function of the number of decisions being aggregated (see Theorem 2 in Sameki et al., 2019). Considering this finding and the cost involved, we concur with other scholars (Benoit et al., 2016; Hara et al., 2013) and recommend that researchers use five crowdworkers to annotate each tweet. Future research could explore approaches that dynamically assign the number of workers per judgment based on the nature of tweets (Sameki et al., 2016). Another fruitful research avenue could be allowing workers to exchange justifications for their judgments, which may also help improve crowdcoding's accuracy (Drapeau et al., 2016).

Our study also calls attention to considering the reliability of crowdcoded data. The crowdcoded data's IAR scores in this study would not be considered robust using the standard in communication research (e.g., an α of .70 or higher). Nevertheless, we find high validity in our crowdcoded data, despite the relatively low IAR scores. Again, this speaks to the unique nature of crowdcoding, in which multiple workers code the same content and individual coders' poor or divergent judgments (thus low IAR) are counteracted by aggregation. However, the question "how low is too low?" remains. To explore the relationship between reliability and validity in crowdcoding, we conducted some preliminary analysis using Monte Carlo computer simulations (see Supplemental Appendix H). We observed that IAR (measured in AC_1) increases very slowly with accuracy, until accuracy becomes quite high. Thus, even with a small AC_1 value (low reliability), we can achieve high accuracy (high validity) by combining crowdworker labels via a majority vote. Most importantly, the simulation analysis demonstrates that, for equivalent accuracy levels, different ground truth label prevalence distributions and different crowdworker capabilities can produce varied IAR scores. In other words, an "acceptable" level of IAR for crowdcoding may vary by different tasks and crowdsourcing platforms, which should be further explored in future research.

This research was motivated in part to determine whether crowdcoding would prove a more time- and cost-effective approach for content analysis than QCA. Our study shows that crowdcoding—which finished the coding tasks in a matter of days—is clearly much more efficient than QCA, which lasted two months. Thanks to the emerging crowdsourcing platforms such as MTurk and F8, numerous internet workers can be recruited to complete coding tasks; this large number of coders cannot be imagined in a QCA context. Crowdcoding's efficiency advantage is highly pertinent for making timely assessments of public opinion in ongoing political and public events. In practice, crowdcoding can provide quick annotations for daily tasks. For example, with the use of crowdcoding, mobile applications have been created to assist blind users in extracting information from images and videos in real time (Gurari et al., 2018).

However, crowdcoding can be more expensive than QCA, when crowdworkers are paid according to the local minimum wage. In other words, crowdcoding may only be a cheap solution if crowdworkers are not fairly compensated. Along with Silberman and colleagues (2018), we contend that researchers should conduct crowdcoding responsibly; the crowdsourcing platform should be treated as "an interface of human workers" rather than "a vast computer without living expenses" (p. 39).

To further evaluate crowdcoding, we conducted a post hoc survey on MTurk to understand how individual crowdworkers' personal traits might influence their annotation accuracy (see Supplemental Appendix I). Among other findings, crowdworkers who more frequently consumed news from Twitter appeared to be better at coding Twitter sentiment. Researchers may consider applying additional criteria when selecting crowdworkers on MTurk (keeping in mind that each additional criterion costs US$0.05 to US$1.00 per judgment).

Overall, the study suggests that crowdcoding can provide accurate annotations, and, compared with QCA, completes tasks more efficiently but is not always cheaper. We reiterate here that accuracy is operationalized as the extent to which crowdcoded annotations match those generated by domain experts. The conclusion is based on an analysis of political sentiment on Twitter and may not be generalizable to other tasks analyzing informal text, on Twitter or another social media platform. In particular, crowdcoding may not be applicable to more complex coding tasks such as framing analysis, for which extensive training is usually needed. It should also be noted that while computer scientists use crowdcoding to annotate all kinds of content, communication researchers should use the approach to analyze content with an objective pattern as assumed in quantitative communication research (Potter & Levine-Donnerstin, 1999).

For those who are interested in Twitter sentiment analysis, we recommend the following: (a) researchers conducting a small-scale content analysis with a small budget should use QCA, (b) researchers conducting an analysis with a large number of annotations should consider crowdcoding, which can provide QCA-level annotations in a much shorter time frame, and (c) if the research is only exploratory, researchers could also consider using crowdcoding with a smaller number of crowdworkers per task, rendering the expense similar to QCA and sacrificing a modicum of validity. Researchers should decide on the method and setting based on the nature of their projects, their budget and timeline, as well as their goals. If time constraints and budgets are no issue, researchers could consider triangulating QCA and crowdcoding to obtain high-grade annotations. Supplemental Appendix J presents the similarities and differences between QCA and crowdcoding. In addition, we provide a list of recommendations for best practices in performing crowdcoding (see Supplemental Appendix K). Most importantly, for any future project using crowdcoding, we recommend that researchers conduct a test study to assess the crowdcoded data's validity and reliability before the actual analysis.

To conclude, our study is important because it empirically demonstrates that crowdcoding can be used to annotate not only texts with a formal structure, but also social media data that are often messy and contain sarcasm (Guo et al., 2016). With an appropriate research design and implementation, crowdcoding can produce data as valid as QCA and is more efficient. Our study suggests that communication researchers should consider crowdcoding as a valid alternative to QCA, especially for analyzing big social data (e.g., as training data in building SML models), which presents new avenues for unobtrusively observing a range of human behavior. It is important to emphasize that new methods for big social data analysis should be explored with an eye toward advancing social science theories. As Jungherr and Theocharis (2017) point out, studies utilizing big data so far have not contributed much by way of theory building; rather, they either offer descriptive accounts of digital media usage and behavior or imagine the potential of big data insights. In this light, communication researchers could consider using crowdcoding for examining media messages and public opinion via social media, ultimately testing communication theories such as agenda setting and framing. Considering the potential of this emerging crowdcoding approach, more research should be conducted to test its validity and efficiency in other contexts of journalism and mass communication research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) declared receipt of the following financial support for the research, authorship, and/or publication of this article: This study is supported in part by the U.S. NSF grant 1838193.

ORCID iD

Lei Guo https://orcid.org/0000-0001-9971-8634

Supplemental Material

Supplemental material for this article is available online.

Notes

1. This study follows Haselmayer and Jenny (2017) in terming the approach "crowdcoding" as distinguished from other crowdsourcing applications.
2. Our data set contains 3,657,628 tweets: 1,704,918 mentioned Trump only and 912,461 mentioned Clinton only. Due to the limitations of Twitter's public API, the sampled tweets are not necessarily representative of all tweets about the debates. In this analysis, the Twitter sample should not significantly affect the results because the goal is not to generate any public opinion insight from the sample, but to compare the performance of QCA and crowdcoding, which are used to annotate the same set of tweets.
3. On both MTurk and F8, we instructed crowdworkers to finish at least 10 tasks. However, the default setting on MTurk does not allow for forcing this action. Therefore, tweets annotated by crowdworkers who completed less than 10 tasks were removed from the analysis.

References

Alonso, O., Marshall, C., & Najork, M. (2014). *Crowdsourcing a subjective labeling task: A human-centered framework to ensure reliable results* (Microsoft Research, MSR-TR-2014–91). https://www.microsoft.com/en-us/research/publication/crowdsourcing-a-subjective-labeling-task-a-human-centered-framework-to-ensure-reliable-results/

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295.

Berg, J. (2016). Income security in the on-demand economy: Findings and policy lessons from a survey of crowdworkers. *Comparative Labor Law & Policy Journal*, *37*(3), 543–576.

Beyer, M. A., & Laney, D. (2012). *The importance of "big data": A definition*. Gartner.

Collingwood, L., & Wilkerson, J. (2012). Tradeoffs in accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics*, *9*(3), 298–318.

Drapeau, R., Chilton, L. B., Bragg, J., & Weld, D. S. (2016, September). *Microtalk: Using argumentation to improve crowdsourcing accuracy* [Paper presentation]. *Fourth AAAI Conference on Human Computation and Crowdsourcing*, Austin, TX, United States.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010, June). Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 80-88). Association for Computational Linguistics.

Fromreide, H., Hovy, D., & Søgaard, A. (2014, May). *Crowdsourcing and annotating NER for Twitter# drift. In LREC proceedings* (pp. 2544-2547).

Guo, L., Vargo, C., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. P. (2018, February). Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3608–3617). IEEE.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.

Hara, K., Le, V., & Froehlich, J. (2013, April). Combining crowdsourcing and Google Street view to identify streetlevel accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)* (pp. 631–640). IEEE.

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity*, *51*(6), 2623–2646.

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Crown.

Hung, N. Q. V., Tam, N. T., Tran, L. N., & Aberer, K. (2013). An evaluation of aggregation techniques in crowdsourcing. In X. Lin, Y. Manolopoulos, D. Srivastava, & G. Huang (Eds.), *Web information systems engineering—WISE* (pp. 1–15). Springer.

Irshad, H., Montaser-Kouhsari, L., Waltz, G., Bucur, O., Nowak, J. A., Dong, F., & Beck, A. H. (2015). Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: Evaluating experts, automated methods, and the crowd. *Pacific Symposium on Biocomputing*, *2015*, 294–305.

Jungherr, A., & Theocharis, Y. (2017). The empiricist's challenge: Asking meaningful questions in political science in the age of big data. *Journal of Information Technology & Politics*, *14*(2), 97–109.

Kim, A. E., Hansen, H. M., Murphy, J., Richards, A. K., Duke, J., & Allen, J. A. (2013). Methodological considerations in analyzing Twitter data. *Journal of the National Cancer Institute Monographs*, *2013*(47), 140–146.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. SAGE.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*, *92*(4), 791–811.

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods and Measures*, *11*(3), 191–209.

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2016). Three decades of reliability in communication content analyses reporting of reliability statistics and coefficient levels in three top journals. *Journalism & Mass Communication Quarterly*, *93*(4), 1135–1159.

Mellebeek, B., Benavent, F., Grivolla, J., Codina, J., Costa-Jussa, M. R., & Banchs, R. (2010, June). Opinion mining of Spanish customer comments with non-expert annotations on Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 114–121). Association for Computational Linguistics.

Neuendorf, K. A. (2017). *The content analysis guidebook*. SAGE.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258–284.

Riffe, D., Lacy, S., & Fico, F. (2014). *Analyzing media messages: Using quantitative content analysis in research*. Routledge.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010, April). Who are the crowdworkers? Shifting demographics in Amazon Mechanical Turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 2863–2872). Association for Computing Machinery.

Sameki, M., Gentil, M., Mays, K. K., Guo, L., & Betke, M. (2016, August). *Dynamic allocation of crowd contributions for sentiment analysis during the 2016 US presidential election* [Paper presentation]. *Association for the Advancement of Artificial Intelligence*, Austin, TX, United States.

Sameki, M., Lai, S., Mays, K. K., Guo, L., Ishwar, P., & Betke, M. (2019). *BUOCA: Budget-optimized crowd worker allocation*. https://arxiv.org/abs/1901.06237

Silberman, M. S., Tomlinson, B., LaPlante, R., Ross, J., Irani, L., & Zaldivar, A. (2018). Responsible research with crowds: Pay crowdworkers at least minimum wage. *Communications of the ACM*, *61*(3), 39–41.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). *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* (pp. 254–263). Association for Computing Machinery.

Sorokin, A., & Forsyth, D. (2008, June). Utility data annotation with Amazon Mechanical Turk. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1–8). IEEE.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173.

Vargas, S., McCreadie, R., MacDonlad, C., & Ounis, I. (2016, March). Comparing overall and targeted sentiments in social media during crises. In *Proceedings of Tenth International AAAI Conference on Web and Social Media* (pp.695-698). Association for the Advancement of Artificial Intelligence.

Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., & Solti, I. (2013). Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *Journal of Medical Internet Research*, *15*(4), Article e73.

Author Biographies

Lei Guo (PhD, The University of Texas at Austin) is an assistant professor in the emerging media studies division at College of Communication, Boston University. Her research focuses on the development of media effects theories, computational social science methodologies, and emerging media and democracy in the United States and China.

Kate Mays is a PhD candidate in the emerging media studies division at College of Communication, Boston University. Her research interests include networked publics, technological affordances of digital platforms, and emerging media effects.

Sha Lai is a PhD student in the department of computer science, Boston University. His research interests include machine learning, information retrieval, and digital signal processing.

Mona Jalal is a PhD student in the department of computer science, Boston University. Her research interests include computer vision, computer architecture, and big data systems and frameworks.

Prakash Ishwar (PhD, University of Illinois Urbana–Champaign) is a professor in the department of electrical and computer engineering in Boston University. His current research centers on data science to advance statistical and computational tools for learning and inference problems using both model-based and data-driven methods. He is a recipient of the 2005 NSF CAREER award, a co-recipient of the AVSS'10 best paper award, and a co-winner of the 2010 ICPR Aerial View Activity Classification Challenge.

Margrit Betke (PhD, Massachusetts Institute of Technology) is a professor in the department of computer science at Boston University. Her research interests are in computer vision and human computation.