Understanding Health Information Credibility across UGC Platforms: Varying Influences of Credibility Features and Prior Knowledge

사용자 생성 콘텐츠(UGC) 플랫폼에서 건강 정보 신뢰도 이해: 신뢰도 요인과 사전 지식의 영향

Heejun Kim (김희준)*
Bogeum Choi (최보금)**
Jaime Arguello (하이메 알구엘로)***

ABSTRACT

Assessing the credibility of online health information has become increasingly complex as the volume of user-generated content (UGC) increases. This study investigates the predictive modeling of credibility in two distinct types of UGC platforms—Yahoo! Answers and Yelp—by exploring the impact of feature categories and the role of assessors' prior knowledge. A total of 2,000 labeled instances were collected through crowdsourcing, using a rigorously validated credibility instrument and qualification process. Eighty-four features were developed and grouped into categories informed by the Elaboration Likelihood Model (ELM), and feature ablation studies were conducted independently on both datasets. Results indicate that *content informativeness* was the most discriminative factor for Yahoo! Answers, while *sentiment* and *content informativeness* were significant for Yelp. Interestingly, prior knowledge had a platform-dependent effect: it reduced model performance in Yahoo! Answers, likely due to overconfidence and limited domain expertise, but improved performance in Yelp, where lived experience aligned with subjective content. These findings emphasize the importance of tailoring credibility assessments and feature sets to the type of platform and the nature of the content.

초 록

온라인 건강 정보의 신뢰성을 평가하는 일은 사용자 생성 콘텐츠(UGC)의 양이 증가함에 따라 점점 더 복잡해지고 있다. 본 연구는 두 가지 유형의 UGC 플랫폼인 Yahoo! Answers와 Yelp를 대상으로, 신뢰성 예측 모델링을 수행하고 다양한 범주의 특성(feature)과 평가자의 사전 지식이 미치는 영향을 분석하였다. 신뢰성 평가 도구와 엄격하게 검증된 자격 요건을 적용하고, 크라우드소싱을 통해 총 2,000개의 신뢰성 라벨을 수집하였다. 정교화 가능성 모델(Elaboration Likelihood Model, ELM)에 기반하여 84개의 특성을 개발하고 이를 범주로 구분하였으며, 두 데이터셋에 대해 각각 독립적으로 특성 제거(feature ablation) 실험을 수행하였다. 연구 결과, Yahoo! Answers에서는 콘텐츠 정보성(content informativeness)이 가장 영향력 있는 요인으로 나타났으며, Yelp에서는 감정(sentiment)과 콘텐츠 정보성이 모두 중요한 요인이었다. 흥미롭게도, 평가자의 사전 지식은 플랫폼에 따라 상반된 효과를 보였다. Yahoo! Answers에서는 과신과 제한된 분야 전문성으로 인해 모델 성능이 저하된 반면, Yelp에서는 주관적인 콘텐츠와 일치된 생활 경험의 효과로 인해 성능이 향상되었다. 이러한 결과는 신뢰성 평가와 특성 선택이 플랫폼의 유형과 콘텐츠의 성격에 맞게 조정될 필요가 있음을 시사한다.

Keywords: Credibility, Health Information, User-Generated Content, Prior Knowledge, Machine Learning
신뢰성, 건강정보, 사용자 생성 콘텐츠, 사전지식, 기계학습

* University of North Texas, Assistant Professor (heejun.kim@unt.edu) (제1저자, 교신저자)
** University of North Carolina at Chapel Hill, Ph.D. Student (choiboge@live.unc.edu) (공동저자)
*** University of North Carolina at Chapel Hill, Professor (jarguell@email.unc.edu) (공동저자)
논문접수일자 : 2025년 5월 23일 논문심사일자 : 2025년 5월 23일 게재확정일자 : 2025년 6월 5일
한국비블리아학회지, 36(2): 183-209, 2025. http://dx.doi.org/10.14699/kbiblia.2025.36.2.183

1. Introduction

User-generated content (UGC) platforms, including social media, especially, have been a fast-growing source of health information in conjunction with traditional media. In UGC platforms, social capital grows much more rapidly than the number of participants. The total number of potential relationships increases quadratically with the number of participants, thereby expanding the range of possible information sources exponentially as the number of participants grows (Katz & Rice, 2002). In recent years, over 60% of health consumers have reported seeking health information online at least once per week (Maon et al., 2017). Patients who face complex health problems or have difficulties in finding relevant information tend to rely on interactive media in which users can post questions and answers, interact with others, and form collaborative knowledge (Finney Rutten et al., 2019).

Finding credible information on UGC platforms is becoming increasingly difficult despite the vast availability of information. Unlike mainstream media, UGC platforms lack editorial checks, making them more vulnerable to the circulation of low-quality or potentially misleading content. Such content may be shared unintentionally by users who are unaware of its reliability (Lim et al., 2017). This phenomenon can have various adverse effects on our society. The impact caused by unverified or low-credibility content in health contexts is especially concerning compared to general information use, as it can be detrimental to users' health and lead to dangerous consequences. Ordinary users with limited health domain expertise tend to rely on personal knowledge, experiences, and heuristic cues rather than objective verification when evaluating the credibility of health information. Additionally, credibility is a subjective perception that can vary from user to user. This variability underscores the need to understand how credibility is perceived, especially on UGC platforms, which host vast amounts of unverified information. The concept of credibility and its operational definition will be covered in the related work.

To facilitate the communication of credible online information, previous studies have developed theoretical models (Petty & Cacioppo, 1986; Sundar, 2008b), examined factors that influence the credibility of information (Freeman & Spyridakis, 2004; Rieh, 2002), and applied predictive modeling (Bayani et al., 2023; Castillo et al., 2011; 2013). The characteristics of information vary across different types of media platforms, topics, and users, and so does the credibility of information. However, existing work has not investigated the importance of varying effects of features across different platforms (Ma & Atkin, 2017). Previous studies have found that predictive models are hardly generalizable to other events and domains (Liu et al., 2013). Predictive features can be highly dependent on the types of information shared on a given platform and the domain knowledge of its users. As such, these contextual variables should be considered when designing features for credibility prediction. Our research contributes to this effort.

In this paper, we aim to: (1) evaluate a range of machine learning features for predicting the credibility

of health information on UGC platforms, (2) examine varying effects of those features across different types of UGC platforms, and (3) investigate the effects of prior knowledge on credibility judgments. We collected 2,000 credibility labels through crowdsourcing. This process went through a number of validation procedures, such as a pre-test, between-test, and comparisons with experts' labels to ensure the quality of the crowd-sourced credibility labels. The prior knowledge of crowd workers was collected through self-report. Next, to further analyze feature contributions, we conducted feature ablation studies. To do so, we first created a total of 84 features based on prior work on information credibility as a construct, as well as studies on identifying quantifiable, predictive features of perceived credibility. These features were organized into hierarchical categories (e.g., *content informativeness*, *sentiment*, and *source*), allowing us to map each category to a distinct factor influencing perceived credibility. Because different feature categories were relevant to different types of UGC platforms, we performed the feature ablation studies independently for each platform.

2. Related Work

The focus of this study is the credibility of information. Non-credible information is referred to by various terms in the literature, including "fake news," "misinformation," and "rumor," each with nuanced distinctions. Given this variation in terminology, we review a broad spectrum of related work that, while differing in focus, offers valuable insights for the task of predicting the credibility of information.

2.1 Definition of Credibility

While credibility has been examined across various disciplines, there is no clear agreement about the definition yet (Hilligoss & Rieh, 2008). Overall, there are two contrasting perspectives on the nature of credibility across several disciplines: objectivity and subjectivity. For instance, in many scientific fields such as information science, credibility has been considered an objective attribute of "*information quality*" (Wang, 1996). In contrast, in the fields of communication and social psychology, credibility has been regarded as the subjective perception of the information receiver (Flanagin & Metzger, 2008). Despite these different perspectives, the central theme across various definitions of credibility is believability. Many scholars (Flanagin & Metzger, 2008; Rieh, 2010) consider credibility to be the believability of a source, message, or media that consists of two main dimensions of trustworthiness and expertise. Believability refers to "people's confidence in the truth of information without having some form of absolute proof" (Rieh et al., 2014). Trustworthiness refers to people's belief in the information source (Rieh et al., 2014). Expertise refers to an ability with which an agent or group of agents can provide accurate information (Rieh et al., 2014).

With the emergence of Web 2.0 and social media, an expanded set of criteria besides relevance needs to be considered for accepting or rejecting retrieved

information (Rieh & Belkin, 1998). The concept of *information quality* has been widely used to distinguish between correct and incorrect information by focusing on verifying information content. However, this approach is not feasible in the current information environment. Insufficient sources of information are available for verification, considering the vast quantity of information. Recently, the concept of *trustworthiness* has also been used to judge the believability of information, and many social media sites have used source-based characteristics such as friendships and peer ranking as a proxy for trustworthiness. However, these characteristics are simple cues to estimate the quality of information by evaluating the source rather than the content itself. Thus, achieving a balance between quality and trustworthiness is important. This is especially true for health-related information shared on UCG platforms, which inherently host a lot of subjective content (e.g., personal experiences). At the same time, health information must maintain a certain level of quality to avoid causing harm. For this reason, the notion of credibility should be applied to both the information content and its source, encompassing both objective and subjective aspects in its evaluation. Credibility is a criterion that is widely used by actual users. Perceived credibility, which is measured using a credibility instrument, is the target of prediction in this study.

2.2 Elaboration Likelihood Model (ELM)

ELM (Petty & Cacioppo, 1986) is a theory of persuasion and was first developed in the field of psychology. This theory assumes that readers tend to judge the credibility of text either based on arguments of the text or external cues such as the type of publication (Freeman & Spyridakis, 2004). The "*elaboration likelihood*" refers to the probability that an individual will critically evaluate the arguments in a message. It ranges from low to high and increases under conditions such as high motivation and the ability to engage in deep thinking. There are two routes of persuasion in ELM: a central route and a peripheral route. The central route is used to assess information logically by evaluating the claims of the content and requires a substantial cognitive effort. The peripheral route relies more on simple and heuristic cues (e.g., how many friends the source has). When factors related to the persuasion enhance the elaboration likelihood, the central route occurs. When factors related to the persuasion weaken the elaboration likelihood, the peripheral route occurs (Petty & Cacioppo, 1986).

Critical thinking largely depends on two primary factors: motivation and ability. Verification of content requires significant effort, so only individuals with high motivation and/or high ability are likely to engage in it. The central route appears to have longer-lasting and more impactful effects than the peripheral route (Petty, 1977; Petty & Cacioppo, 1986). The ELM supports the definition of credibility in this study and has multiple implications. First, the dimensions of credibility related to the expertise (e.g., accuracy) align with the central route, while the dimensions related to the trustworthiness (e.g.,

reputation) align with the peripheral route. Second, the ELM introduced factors that influence credibility judgments, such as users' familiarity with the subject, the degree of involvement, and their cognitive ability. Thus, we used ELM as a theoretical framework to develop and examine factors that influence credibility judgments in this study.

2.3 Predictive Modeling

Prior work used different machine learning features to predict the credibility of information, and those features were based on content, language, user, propagation, and other metadata. For instance, Nagura et al. (2006) used the commonality among articles to evaluate the credibility of those articles. They measured the commonality by the cosine similarity between sentences in the collected articles from different news publishers. Lim et al. (2017) used features extracted from web search results to determine whether claims on microblogs are credible or not. Other work used particular linguistic cues such as anomalous patterns of part of speech (e.g., pronouns and conjunctions) (Feng & Hirst, 2013; Markowitz & Hancock, 2014) and the inflated number of swear words (Gupta et al., 2014). Some previous work (Castillo et al., 2011; Ciampaglia et al., 2015; Ma et al., 2015) assessed the trustworthiness of the information source by using link-based measures. Recently, Ma et al. (2016) and Shang et al. (2018) used deep learning models to learn latent predictive relationships between features representing credibility. Many of these studies tend to focus on a single platform, rather than examining multiple platforms with distinct characteristics. As a result, they may lack the nuance needed to understand how different credibility features vary in their effects across platforms.

2.4 Contextual Factors

The factors used to evaluate credibility will vary depending on the context in which information is used. One of the main goals of this study was to examine the varying effects of features influenced by the type of UGC platforms and the user's prior knowledge. Depending on the type of UGC platforms, the purpose and environment for using these online platforms are different. Sundar (2008a) asserted that different communication platforms (e.g., websites, blogs, and social media) can affect perceived credibility because the platform acts as a gatekeeper for the quality of online information. However, prior studies have not yet provided evidence that factors affecting the credibility of health information vary from platform to platform (Ma & Atkin, 2017).

Several previous studies have examined the role of users' knowledge in the context of information behavior. Hilligoss and Rieh (2008) found that users mainly use personal knowledge to interact with the content of information while assessing the credibility of information. From an online user study, Yamamoto and Tanaka (2011) found that participants with little to no knowledge thought accuracy was the most critical factor in judging credibility, whereas it varied depending on topic category for knowledgeable

participants. Wathen and Burkell (2002) also found that users combined the evaluation of the information source with his/her domain knowledge while evaluating the content credibility of a website. Although these studies suggested the effects of prior knowledge on credibility judgments, they were not able to examine the varying effects of features in a data-driven method.

3. Methods

The main goal of this study was to build a predictive model of credibility for health information and investigate the varying effects of features depending on contextual factors. Overall, there were three steps to achieve this goal. First, this study created credibility labels for health information. Second, we created features that operationalize the credibility factors discovered by an inductive content analysis in our previous work (Kim & Choi, 2018). Finally, we developed predictive models and evaluated them using the ground-truth data generated in the first step. Feature ablation studies were used to find the most discriminative features for predicting the credibility of health information and examining varying effects according to different contexts.

3.1 Data

To investigate the credibility of health information on user-generated content (UGC) platforms, we selected two widely used and publicly available datasets: Yahoo! Answers and Yelp. These two platforms exemplify different modes of UGC—information-seeking and experience-sharing—and offer rich contexts for studying credibility judgment in health-related content. Importantly, both datasets are openly accessible to the research community without institutional or licensing barriers, allowing reproducibility and broader academic use.

3.1.1 Yahoo! Answers Dataset

We utilized a dataset provided through the Webscope™ Program (*Webscope | Yahoo Labs*, n.d.). Yahoo! Answers, once one of the most frequently consulted reference sites after Wikipedia (Fichman, 2011), served as a leading platform for collaborative information seeking and knowledge sharing (Chua & Banerjee, 2015; Jin et al., 2013). Although the service has been discontinued, similar Q&A platforms (e.g., Quora and Reddit) are actively used by health consumers. For this study, we used the Yahoo! Answers Comprehensive Questions and Answers version 1.0, which contains 4,483,032 questions with corresponding answers, category/sub-category labels, and other metadata (e.g., language, country, and date of posting). This dataset includes limited user information compared to the Yelp dataset.

We used the “health” category to primarily select health-related questions and answers. Then, we excluded questions written not in English and in countries other than the United States. Prior research has shown that question quality is a critical factor influencing the perceived credibility of corresponding answers (Agichtein et al., 2008). To minimize researchers’

bias and ensure topic relevance, we reviewed 500 randomly selected questions and developed a set of exclusion criteria (Table 1) to filter out questions unrelated to health. A significant number of excluded questions consisted of low-quality or joke submissions, which were difficult to identify through manual inspection. To address this, we created a list of spam-related keywords (Table 2) and excluded any questions containing one or more of these terms. Then, the dataset of questions and answers was randomly shuffled, and filtering was applied manually until a final set of 1,000 question-answer pairs was obtained based on the exclusion criteria.

3.1.2 Yelp Dataset

We utilized the dataset made available through the Yelp Dataset Challenge (*Open Dataset | Yelp Data Licensing*, n.d.). Given its extensive volume and wide usage, Yelp is regarded as one of the most representative platforms for analyzing online service reviews, particularly in the context of consumer feedback on local businesses (Racherla & Friske, 2012). The dataset contained online reviews written between March 2005 and January 2017. There are 2,685,065

Table 1. Exclusion Criteria

Criteria	Sample question	Rationale
Conversational	"Does anyone know what would be a real good comeback to someone that insults you about your weight?"	Answers to this type of question would not include health information.
Emotional	"My mom has had bone cancer for over a year, [...] the cancer has spread in all her body [...] I cry day and night. Plz pray for my mom."	This question is rather emotional and only can expect subjective and emotional feedback which is hard to judge credibility.
Out-of-topic	"How many more meltingicebergs, hurricanes, disasters, before we act on global warming?"	These questions are not relevant to health topic.
Trash jokes/silly question	"I'm so hot that I have to get drunk all the time just to deal with myself what should I do?"	The questioner asks a very personal question which is hard and silly to answer.
Scientific inquiry	"Why the suprarenal gland is the first organ that affected by the secondary metastasis in cancer cases?"	This is a question that can be answered by scientists. MTurk workers would not have enough expertise to answer.
Personal affair	"How many of you are Lactose Intolerant?"	The answer will depend on the person who answers. There is no right, credible answer.

Table 2. A List of Spam Words

Topic	Spam words
Cigarette/drug	Smoking, smoke, smoker, cigarette, cigar, smoker, and drug
Sex	Kissing, masturbation, Viagra, gonad, saliva, sex, pantiliner, condom, and virginity
Others	Hiccup and bourbon

reviews written by 686,555 reviewers for 85,950 local businesses. Unlike the Yahoo! Answers dataset, the Yelp dataset has rich network-related attributes (e.g., list of friends) that can be used for network analysis.

The dataset was filtered by using category (e.g., *doctors, dentists, dermatologists*) only to focus specifically on healthcare service-related reviews. The final 1,000 reviews were randomly selected for analysis. Due to the limited availability of source-related features in the Yahoo! Answers dataset, we conducted two separate sets of experiments: (1) those using the Yahoo! Answers dataset and (2) those using the Yelp dataset. This approach allowed us to assess the predictive value of source-related features in evaluating the credibility of health information while minimizing potential confounding effects arising from differences in feature availability across datasets.

3.2 Creation of Credibility Labels

We used the message-level credibility scale, as the primary unit of evaluation presented to assessors consisted solely of either an individual answer or a review text. We employed the message-level credibility scale developed by Rains and Karmikel (2009), consisting of four items—*believable, trustworthy, accurate*, and *complete*—rated on a 7-point Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*). This instrument is well-suited to the context of this study involving health information on UGC platforms. To generate "ground-truth" credibility labels, we utilized annotations from both trained researchers and crowd workers. We used credibility labels from trained researchers: (1) to do quality control during the crowd-sourced annotation phase and (2) to measure agreement between trained researchers and crowd workers as a proxy for the quality of credibility labels by crowd workers.

Regarding trained researchers' annotations, two coders (the lead author and a Ph.D. student) made credibility evaluations on 1,000 random samples (500 for each dataset) independently using the credibility instrument. 630 data instances of these random samples overlapped with the 2,000 final datasets. Weighted Kappa agreement was used for measuring coding quality, in which certain disagreements (1 vs. 7) are more severe than others (3 vs. 5). The annotations by trained researchers were done in our previous work (Kim & Choi, 2018), and the details, including training, are available in that work.

Credibility label creation by trained researchers is time-consuming and expensive. Therefore, crowdsourcing is a promising alternative for acquiring more data for machine learning. We used Amazon's Mechanical Turk (MTurk) for crowdsourcing. Before beginning the actual Human Intelligence Task (HIT), MTurk workers were required to complete a qualification test consisting of five HITs, with item and task order randomized. A sample Yahoo! Answers task is shown in Figure 1. Workers judged each item independently. The five qualification tasks, drawn from researcher-labeled data with the highest inter-rater agreement, were identical for all participants. To qualify, workers needed to correctly label at least four of the five tasks.

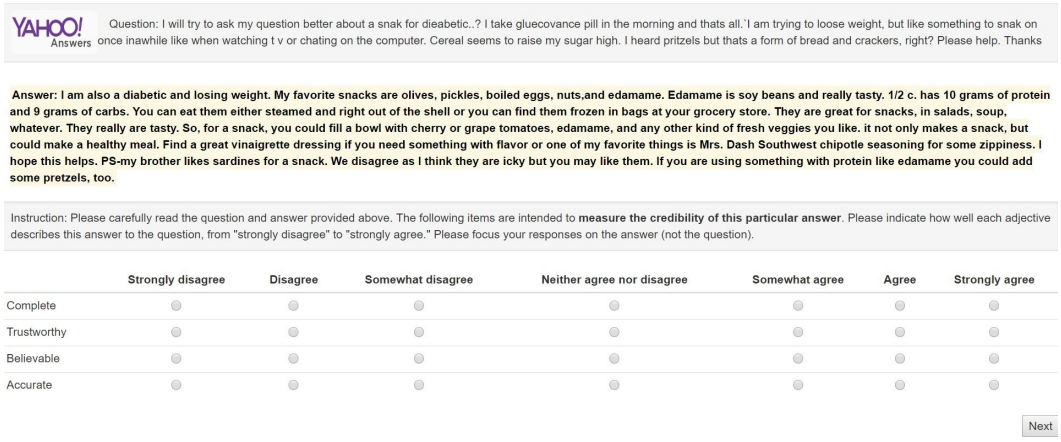

Fig. 1. Labeling Task Interface with Yahoo! Answers Data

To ensure continued work quality after the qualification test, we introduced random checks during the actual HITs. Every 10 HITs, workers received a test item drawn from researcher-labeled data. Those who incorrectly labeled more than three test items were no longer able to do HITs. Additionally, attention check questions (e.g., "I am answering these questions carefully.") appeared every three HITs. Workers failing more than two of these should discontinue HITs. All participants received the same compensation ($0.20 per HIT) regardless of performance. Each HIT also included questions for self-assessing prior knowledge (Table 3).

According to Snow et al. (2008), using as few as four crowdsourced labels for a majority vote could produce labels of similar quality to those made by experts. We collected three annotations for each answer or review and combined them into a final label using a majority vote. Although we collected one fewer label per each data instance than Snow et al. proposed, we applied additional quality control mechanisms such as the qualification test and between-HITs tests as explained above. Additionally, creating final labels with an odd number of annotations helped us avoid ties in the majority vote. In total, this study collected 6,000 annotations (3 redundant HITs per answer/review for 2,000 data instances). Fleiss' Kappa (Kf) was used to determine

Table 3. Prior Knowledge Questionnaire

Item	• I am knowledgeable about the topics (the types of services) covered in this question and answer (review). • I am familiar with the topics (types of services) covered in this question and answer (review). • I have enough background knowledge to judge the credibility of this answer (review). • I understand all the important ideas covered in this question and answer(review).
Response	1 (strongly disagree) to 7 (strongly agree)

the coding quality of MTurk workers. The Cohen's Kappa agreement between MTurkers' majority vote and the trained researchers' label was also measured.

3.3 Machine Learning Feature Generation

Features utilized in this study were grouped into four categories (*content informativeness, sentiment, presentation*, and *source*) at the top-level. These top-categories except *sentiment* had sub-categories. These categories were designed to represent a potential factor influencing the perceived credibility of health information. We utilized 84 features in total. The overall hierarchy of the feature categories is presented in Table 4, with the full details of machine learning features presented in Appendix A. The features used in this study were derived from prior research and supplemented by insights gained through inductive content analysis in our earlier work (Kim & Choi, 2018).

3.4 Evaluation Method

We trained Logistic Regression classifiers and tested their performance in predicting the binary classification of credibility of each answer or review. While more recent algorithms, such as deep learning, may offer improved predictive performance, their complex structure, including convoluted nodes in the hidden layers, limits interpretability, making it challenging to identify which features drive credibility evaluations. A stepwise feature selection method was applied to remove features that are not discriminative and to reduce potential multicollinearity between features. To prevent overfitting, 20% of the data (200 for each dataset) was used in the feature selection, and the remaining 80% (800 for each dataset) was used in feature ablation studies. Logistic regression classifiers based on the generalized linear model were implemented using the R Caret package, and the stepAIC function of the R Mass package was used with both forward and backward selection procedures.

Table 4. Hierarchy of Feature Categories

Top-level	Middle-level	Availability	
		Yahoo! Answers	Yelp
Content Informativeness	Plausibility	Yes	No
	Relevance	Yes	No
	Comprehensiveness	Yes	Yes
	Specificity	No	Yes
	Currency	Yes	Yes
Sentiment	-	Yes	Yes
Presentation	Readability	Yes	Yes
	Spelling	Yes	Yes
Source	Expertise	Yes	Yes
	Reference	Yes	Yes
	Credence	No	Yes

Credibility models were trained and evaluated using 10-fold cross-validation. Given the class imbalance in the credibility labels, we used Average Precision (AP) as the primary evaluation metric. Reported performance metrics refer to the average of APs across all 10 cross-validation iterations. Among the two classes (*credible* and *not_credible*), we report performance metrics specifically for the *not_credible* class for two main reasons: (1) it represents the minority class, and (2) it reflects a realistic use case in which users want to avoid encountering information perceived as not credible. An extensive set of feature ablation studies was iterated through all feature categories, including sub-categories, to evaluate the impact of different feature groups.

To ensure valid statistical inference, paired samples from the held-out test sets were evaluated using the non-parametric bootstrap-shift test (Noreen, 1989), which accounts for the non-independence and potential non-normality of cross-validation samples. To mitigate the risk of Type I error from multiple comparisons, Holm's step-down procedure (Holm, 1979)—a refinement of the Bonferroni correction—was applied, adjusting the significance threshold based on the rank of each p-value. Instead of using the modified significance level $a' = a/m$ (where m is the number of tests) for every comparison, Holm suggested modifying the denominator m to m-(k-1), where m is the number of tests and k is the rank of p-value of the comparison in ascending order.

Additionally, we conducted one more feature ablation study to examine how credibility factors are influenced by prior knowledge, particularly as it relates to assessor qualification. We created two additional sets of features: (1) prior knowledge and (2) interaction features that represent interactions between basic features and prior knowledge. Prior knowledge labels of MTurk workers were created by the majority vote based on their self-reported assessments (see Table 3). The summed 7-point Likert scores were converted into binary class (*prior_knowledge* or *no_prior_knowledge*) by the mean value (72). The mean score sum was 71.163 (SD = 8.088). Scores equal to or above 72 were categorized as *prior_knowledge*; those below were labeled as *no_prior_knowledge*.

3.5 Ethical Considerations

To safeguard the rights and privacy of both users of UGC platforms and MTurk workers, we followed rigorous ethical standards throughout the study. All personally identifiable information was anonymized to prevent the disclosure of user identities. When presenting excerpts from user comments, we exercised additional caution by selecting only portions of comments and replacing any potentially sensitive content with pseudorandom text for illustrative purposes. Data collection, storage, and analysis were conducted using a secure, access-controlled research computing environment. A limited subset of comments (1,000) was transferred to password-protected computers for qualitative analysis by research team members and subsequently deleted upon completion. No data from this study were made publicly available. The study protocol was reviewed and approved by

the Institutional Review Board at University of North Carolina at Chapel Hill (IRB-18-0142).

4. Results

4.1 Credibility Labels

264 MTurk workers completed 6,000 assignments, resulting in 2,000 credibility labels by the majority vote. We calculated the average score across four credibility instrument items (*believable, trustworthy, accurate,* and *complete*) for each worker. Then, these scores were averaged across three MTurk workers and converted into a binary class (*credible* vs. *not_credible*). The midpoint of the scale (48) was used as the threshold. Specifically, if the aggregated credibility score was greater than or equal to 48, the answer or review was labeled as *credible*; otherwise, it was labeled as *not_credible*. The distribution of credibility labels is summarized in Table 5. In both datasets, there were a greater number of instances in the *credible* class than in the *not_credible* class. In particular, the Yelp dataset exhibits a stronger skew toward the *credible* class. The second and third columns of the table report the number of instances for each platform and class, with the corresponding percentages shown in parentheses.

The Fleiss' Kappa and Cohen's Kappa for each dataset are presented in the fourth column and the last column of Table 5. Fleiss' Kappa (K_f) was 0.342 for Yahoo! Answers and 0.18 for Yelp. These levels of agreement can be considered *fair* and *slight*. Assessors had a higher agreement in Yahoo! Answers than in Yelp. Also, Cohen's Kappa coefficient (K_c), the agreement between the majority vote of MTurk workers and the trained experts' label, was calculated to measure the coding quality of the credibility labels by the MTurk workers. The agreement was 0.548 for Yahoo! Answers and 0.364 for Yelp. These levels of agreement can be considered *moderate* and *fair*. Again, Yahoo! Answers demonstrated stronger inter-rater reliability than Yelp.

4.2 Feature Ablation Study

4.2.1 Yahoo! Answers

Before starting the feature ablation study, we applied the stepwise feature selection. Compared to the full model (0.638), AP was increased by 0.079 after the feature selection (0.717). Given that only 33% of the data belonged to the negative class, the classifier after the feature selection showed a fairly decent performance. The final set of selected features is presented in Table 6. As there were five feature categories selected, five corresponding treatment

Table 5. The frequency of Credibility Labels by MTurk Workers

UGC platform	*Credible*	*Not_Credible*	Fleiss' Kappa	Cohen's Kappa
Yahoo! Answers	681 (68.1%)	319 (31.9%)	0.342	0.548
Yelp	921 (92.1%)	79 (7.9%)	0.180	0.364

Table 6. Final Features Selected for Feature Ablation Study (Yahoo! Answers)

Top-Category	Sub-Category	Features
Content Informativeness	Relevance	UMLS-based weighted cosine similarity
	Comprehensiveness	Word count and UMLS concepts count
Presentation	Readability	FleschKincaidGradeLevel, GunningFogIndex, ColemanLiauIndex, LIX, and RIX
Sentiment	-	Positive emotion, joy, calmness, positive_hope, negative_emotion, and sadness
Source	Expertise	Count of UMLS concepts related to jargons

models were created and compared to the control model.

We compared the performance between classifiers using the Bootstrap-Shift test (Table 7). The row labeled "All" represents the performance of the control model, which includes all selected features. Rows labeled "-X" show the performance of treatment models in which features in the category "X" were excluded from the selected features. These rows are ordered by descending performance drop. The letter in parentheses indicates whether the excluded feature category "X" is a top-level (t) or sub-level (s) within the feature hierarchy. The "Percent change" column presents the relative increases (+) or decreases (-) in performance compared to the control model. In all rows except the first row, a decrease in performance indicates that the excluded feature category contributed discriminative power to predicting the credibility of health information. The most discriminative feature group at each hierarchy level (top-level and sub-level) is marked in bold. The final column presents p-values from the paired Bootstrap-Shift tests. Holm's adjustment was applied to control for multiple comparisons, and adjusted p-values were calculated by multiplying the raw p-values by their rank-based correction factor. If the adjusted value exceeded 1, it was capped at 1 for clarity. One-tailed tests were used, and statistically significant differences at the $\alpha = 0.05$ level are marked with an asterisk (*).

Among the top-level feature categories, the *content informativeness* features proved to be more effective and influential in predicting credibility than

Table 7. Feature Ablation Study Results (Yahoo! Answers)

Feature group	AP	Percent change	P-value
All	0.717		
- Content Informativeness (t)	**0.589**	**-17.85%**	**(0.000*)**
- Comprehensives (s)	**0.593**	**-17.29%**	**(0.000*)**
- Expertise (s)	0.628	-8.9%	0.304
- Writing (s)	0.645	-7.2%	1.000
- Relevance (s)	0.653	-6.4%	1.000
- Sentiment (t)	0.654	-6.3%	1.000

presentation, sentiment, and *source* features. Only the *content informativeness* features had statistically significant effects. Among the sub-level feature categories, *comprehensiveness* features were only influential and statistically significant in predicting credibility. Compared to the model with all the *content informativeness* category features excluded, the importance of *comprehensiveness* features is further supported because there is little difference in percent drop and p-value between the two models. Interestingly, another sub-level category, *relevance*, within the *content* category, did not show a meaningful contribution.

4.2.2 Yelp

As in Yahoo! Answers, we applied the stepwise feature selection before starting the feature ablation study. Compared to the full model (0.289), AP was increased by 0.104 after the feature selection (0.393). Given that only 8% of the data belonged to the negative class, the classifier after the feature selection showed a reasonable performance. The final set of selected features is presented in Table 8. We conducted feature ablation studies for the Yelp dataset as in the previous section. As there were six feature categories selected, six corresponding treatment models were created and compared to the control model.

We compared the performance between classifiers using the Bootstrap-Shift test (Table 9), following the same row, column, and symbol conventions as in Table 7. The results revealed two statistically significant differences in the means of AP across six paired test sets. Among the top-level feature categories, the *sentiment* category was the most effective and influential, though the *content informativeness* category also demonstrated statistically significant effects. No sub-level feature categories were influential and statistically significant in predicting credibility. In contrast to the results from Yahoo! Answers, the *comprehensiveness* features in Yelp were not statistically significant on their own. Their contribution to credibility prediction was only observed when combined with *specificity* under the broader category of *content informativeness*.

4.3 Effects of Prior Knowledge

The objective of this experiment was to examine the impact of the prior knowledge of users on the credibility of health information. Two treatment

Table 8. Final Features Selected for Feature Ablation Study (Yelp)

Top-Category	Sub-Category	Features
Content Informativeness	Comprehensiveness	Word count, lexical diversity, and document entropy
	Specificity	Count of UMLS concepts, average counts of named entities and UMLS concepts, and count of named entities
Writing	Readability	ARI, FleschKincaidGradeLevel, LIX, and RIX
Sentiment	-	Positive_emotion, sadness, polarity, and general_dislike
Source	Expertise	Count of UMLS concepts related to jargons

Table 9. Feature Ablation Study Results (Yelp)

Feature group	AP	Percent change	P-value
All	0.393		
- Sentiment (t)	**0.253**	**-35.62%**	**(0.002*)**
- Content Informativeness (t)	0.277	-29.52%	0.017*
- Comprehensiveness (s)	0.348	-11.45%	0.724
- Specificity (s)	0.371	-5.6%	0.744
- Expertise (s)	0.392	-0.25%	1.000
- Presentation (t)	0.414	+5.34%	1.000

models were developed: (1) a model incorporating the basic features along with a prior knowledge indicator, and (2) a model that included the basic features, the prior knowledge indicator, and interaction features representing the product of each basic feature and the prior knowledge variable. The basic features used were those selected through the stepwise selection method described in the previous section for each dataset (refer to Tables 6 and 8). To evaluate the impact of the additional features, each treatment model was compared to a control model using the Bootstrap-Shift test.

4.3.1 Yahoo! Answers

Results from the feature ablation study are summarized in terms of AP and corresponding p-values (Table 10). The table layout and interpretation conventions are consistent with those in Tables 7 and 9, with the exception that the final two columns report p-values for pairwise comparisons, adjusted using Holm's procedure. Interestingly, the inclusion of prior knowledge and interaction features resulted in a reduction in AP. Notably, the treatment model incorporating both prior knowledge and interaction features showed statistically significant differences compared to both the control model and the treatment model with only the prior knowledge feature. However, the difference between the control model and the treatment model with only the prior knowledge feature was not significant. These findings suggest that assessors' prior knowledge interacts significantly with credibility-related features, influencing the overall predictive performance of the model.

4.3.2 Yelp

Results from the feature ablation study are presented in Table 11, which follows the same structure and conventions as Table 10. Among all comparisons, only the inclusion of prior knowledge yielded a statistically significant improvement over the control model. The other two pairwise comparisons showed no statistically significant difference between classifiers. These findings suggest that, for the Yelp dataset, assessors' prior knowledge does not interact meaningfully with other credibility-related features. However, incorporating prior knowledge alone enhanced the model's performance, indicating its independent contribution to predicting credibility.

Table 10. Feature Ablation Study Results by Prior Knowledge (Yahoo! Answers)

Model	AP	Percent change	Pairwise comparison	
			Control + Prior knowledge	Control + Prior knowledge + Interaction
Control	0.717		0.15114	0.022*
Control + Prior knowledge	0.712	-0.7%	-	0.016*
Control + Prior knowledge + Interaction	0.684	-4.1%	-	-

Table 11. Feature Ablation Study Results by Prior Knowledge (Yelp)

Model	AP	Percent change	Pairwise comparison	
			Control + Prior knowledge	Control + Prior knowledge + Interaction
Control	0.393		0.028*	0.416
Control + Prior knowledge	0.433	+10.18%	-	0.11
Control + Prior knowledge + Interaction	0.397	+1.02%	-	-

5. Discussion

Our research makes several contributions. First, methodologically, we demonstrate an effective approach to improving the reliability of crowdsourced annotations. To ensure quality, we implemented a two-step process: first, annotations by trained researchers were used to construct qualification and the between-HIT tests; second, final credibility labels were created by the majority vote of MTurk workers' annotations. Inter-rater agreement between MTurk workers was fair for Yahoo! Answers (K_f = 0.342), while agreement between experts and workers was moderate (κ = 0.548), with similar patterns observed in the Yelp dataset. These results suggest that a small number of experts' annotations can be effectively used to control the quality of crowd workers' labeling efforts. Prior work has explored similar challenges. Alonso et al. (2014) emphasized the importance of selecting appropriate intercoder agreement metrics (e.g., Fleiss's Kappa, Krippendorff's Alpha) to account for subjectivity in labeling, proposing a human-centered framework for improving reliability. Kazai et al. (2013) demonstrated that agreement among workers alone can be a useful proxy for label quality, even in the absence of a gold standard. Bhuiyan et al. (2020) extended this line of inquiry by comparing expert and crowd judgments in the

context of news credibility and identifying factors contributing to disagreement. Our study integrates these perspectives by combining researcher-derived gold standards with inter-worker agreement analysis, thereby contributing to the development of robust crowdsourcing strategies for subjective annotation tasks.

Second, the findings of this study demonstrate that a combination of carefully selected features and machine learning techniques can effectively predict the credibility of health information on UCG platforms. Overall, our binary classifiers showed fairly good performance (Yahoo! Answers: 0.717, Yelp: 0.393) in AP, considering the small amount of data in the target class, with only 33% and 8% of instances labeled as *not_credible* in the respective datasets. It is important to note that inter-rater agreement was also lower for Yelp (K_f = 0.18) than for Yahoo! Answers (K_f = 0.342). Yahoo! Answers' question-answer format may facilitate clearer judgments by providing contextual cues, while the less structured nature of Yelp reviews may lead to greater ambiguity and a more imbalanced label distribution. While direct comparison with prior work is limited due to differences in data distribution and performance metrics, our results are competitive. For instance, Shah and Pomerantz (2010) reported 83.83% accuracy in classifying answer quality, but their dataset was heavily skewed (80% low-quality answers), limiting the relative performance gain. Similarly, Ma et al. (2016) achieved 0.881 accuracy in rumor detection using a gated recurrent neural network; however, their unit of analysis was the event, not individual posts, which allowed their model to capture aggregated contextual signals. Furthermore, Shah and Pomerantz (2010) found that no single feature made a statistically significant contribution to performance, supporting our approach of categorizing features conceptually to enhance interpretability and predictive strength. These findings highlight the promise of using carefully structured feature sets and appropriate machine learning methods for credibility assessment in varied UGC contexts.

Third, our findings underscore the importance of considering the platform type when assessing and predicting the credibility of health information on UGC platforms, as it shapes the nature of the information shared and, by extension, influences the mechanisms of credibility evaluation. While prior work has considered a range of factors involved in credibility assessment, including information source (Sundar, 2008a), to our knowledge, no previous study has directly compared two distinct platforms in the context of health information to examine how the influence of credibility factors may vary, that is, how some factors play a more prominent role than others in shaping credibility judgments. Additionally, "source" in prior work has typically been defined as the author of the content, which is different from the platform as a broader information environment—one that offers different affordances and invites different types of information and interactions. The most discriminative feature categories were found to be different in Yahoo! Answers and Yelp. In Yahoo! Answers, only the *content informativeness* feature category demonstrated statistically significant dis-

criminative power (p = 0.000). In contrast, both the *sentiment feature* category (p = 0.002) and the *content informativeness* category (p = 0.017) were statistically significant predictors of credibility in the Yelp dataset. For Yahoo! Answers, it is likely that detailed content and extensive coverage are more valuable in judging responses to specific questions. On the other hand, explanations of the positive or negative experiences of a reviewer can resonate with readers, and thus *sentiment* can play an important role in evaluating credibility, while *content informativeness* still plays a significant role. This difference across the two datasets highlights the need to develop platform-specific features that reflect the nature of the information shared on different UGC platforms.

Fourth, the effect of prior knowledge on credibility prediction varied between platforms, revealing nuanced interactions between assessors' prior knowledge and information type. In Yahoo! Answers, incorporating prior knowledge decreased model performance, whereas in Yelp, it increased model performance. One possible explanation is that Yahoo! Answers requires more domain-specific and expert information, so superficial or non-expert prior knowledge may cause overconfidence and subjective judgments, ultimately introducing noise to the labeling process. In contrast, Yelp reviews consist of general, experience-based information where prior knowledge may align more closely with relatable, lived experiences, thereby aiding in the accurate evaluation of others' experiences. These results raise concerns about whether crowd workers possess the domain expertise necessary for evaluating specialized health information, especially when confidence may exceed competence, as described by the Dunning-Kruger effect (Dunning, 2011). Our finding suggests the need to account for the role of intermediaries on certain platforms, such as a social Q&A platform, particularly in the context of health-related information. In this study, we used the same prior knowledge questionnaire across platforms for consistency; however, future studies may need to adopt platform-specific questionnaires to assess prior knowledge, as the type of information and the kind of knowledge required to evaluate it accurately can vary significantly across platforms. For instance, in our case, it might have been more appropriate to refine the questionnaire for Yahoo! Answers to capture professional training and working experience in healthcare.

Furthermore, in Yahoo! Answers, potential interactions between prior knowledge and other feature categories were observed. In contrast, no such interactions were found in the Yelp dataset. In the case of Yahoo! Answers, assessors may have attempted to evaluate credibility by considering multiple aspects captured by various features, such as content *informativeness, presentation* quality, and *sentiment.* However, this multidimensional assessment likely introduced noise into the labeling process due to their limited subject-matter expertise. Without adequate domain knowledge, users might misinterpret or overvalue certain indicators of credibility, resulting in inconsistent judgments. By contrast, in Yelp reviews, where the information is more experience-based and less technical, users are better posi-

tioned to assess credibility based on their lived experiences with healthcare services. As a result, agreement among assessors may improve, and credibility judgments can be more consistent, even without formal expertise (Flanagin & Metzger, 2007; Metzger et al., 2010). This contrast highlights the need to calibrate credibility assessments according to the user's prior knowledge and the type of information being evaluated.

Lastly, our results support and reinforce the ELM as a useful framework for examining credibility factors in user-generated content. Credibility factors corresponding to the central route and the peripheral route were examined, grouped into categories, and applied to the feature ablation studies. According to the results of this study, features related to *content informativeness,* such as *comprehensiveness*, were influential in Yahoo! Answers, where users typically deal with factual, technical information. In contrast, features related to *sentiment* were influential in Yelp, where users mainly deal with personal experiences. These results support the ELM, as answers to specific questions likely engage the central route, requiring logical evaluation of *content informativeness*, while *sentiment* in reviews can be simply processed through the peripheral route, relying more on heuristic cues.

ELM was also further verified by the interaction between features and the prior knowledge of MTurk workers. In the case of Yahoo! Answers, which addresses specific information needs compared to Yelp, prior knowledge has a statistically significant effect on predictive modeling based on the interaction with features, suggesting that assessors engaged in deeper, more analytical evaluations, which are a hallmark of the central route. On the other hand, in the case of Yelp, which usually supports exploratory search with less specific information needs, no interaction between prior knowledge and features was found, indicating that users may rely more on surface-level cues and heuristic shortcuts when judging credibility, consistent with the peripheral route.

6. Conclusion

This study demonstrates that machine learning models, when combined with thoughtfully selected features and informed by a theoretical model, can effectively predict the credibility of health information on UGC platforms. However, the relative importance of credibility factors varies significantly depending on the platform. Yahoo! Answers, which supports factual, question-driven interactions, benefits most from *content-related* features. In contrast, Yelp, which reflects personal narratives and experiences, is influenced by both *sentiment* and *content informativeness*. Similarly, the role of prior knowledge was found to be context-dependent: while it enhanced prediction accuracy in Yelp, it introduced noise and reduced performance in Yahoo! Answers. These contrasting results suggest that it is crucial to tailor feature engineering and evaluator qualifications to the specific characteristics of each platform for accurate credibility modeling. Our findings also further validate the Elaboration Likelihood Model (ELM), demonstrating that users engage in

different cognitive routes depending on the platform and the type of information presented, particularly in the healthcare context.

Despite these contributions, the study has a few limitations. First, credibility is inherently subjective and may vary depending on the individual making the evaluation as well as the context of the information. While this subjectivity limits the ability to establish an objective ground truth, credibility remains a central concept in how users assess and select information in everyday decision-making, which is the motivation of this study. Second, the dataset used for machine learning consisted of 2,000 annotated instances, which is relatively modest by current standards. However, the complexity of credibility assessment required rigorous human annotation, making large-scale labeling challenging. Lastly, the datasets used in this study are not from recent years. However, our focus was not on capturing the most recent trends on credibility evaluation but rather on understanding the varying influences of credibility features and prior knowledge on UGC platforms in the context of health information. Future work should explore adaptive annotation strategies and domain-sensitive feature design to further enhance the reliability and applicability of credibility assessment tools.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. Proceedings of the 2008 International Conference on Web Search and Data Mining, 183-194. https://doi.org/10.1145/1341531.1341557

Alonso, O., Marshall, C., & Najork, M. (2014). Crowdsourcing a Subjective Labeling Task: A Human-Centered Framework to Ensure Reliable Results. Availavle: https://www.microsoft.com/en-us/research/publication/crowdsourcing-a-subjective-labeling-task-a-human-centered-framework-to-ensure-reliable-results/

Aronson, A. R. & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3), 229-236. https://doi.org/10.1136/jamia.2009.002733

Batagelj, V. & Mrvar, A. (1998). Pajek-program for large network analysis. Connections, 21(2), 47-57.

Bayani, A., Ayotte, A., & Nikiema, J. N. (2023). Automated credibility assessment of web-based health information considering health on the net foundation code of conduct (HONcode): Model Development and Validation Study. JMIR Formative Research, 7(1), e52995. https://doi.org/10.2196/52995

Bhuiyan, M. M., Zhang, A. X., Sehat, C. M., & Mitra, T. (2020). Investigating differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. Proceedings of the ACM on

Human-Computer Interaction, 4(CSCW2). https://doi.org/10.1145/3415164

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. Proceedings of the 20th International Conference on World Wide Web, 675-684. https://doi.org/10.1145/1963405.1963500

Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. Internet Research, 23(5), 560-588. https://doi.org/10.1108/IntR-05-2012-0095

Chua, A. Y. K. & Banerjee, S. (2015). Measuring the effectiveness of answers in Yahoo! Answers. Online Information Review, 39(1), 104-118. https://doi.org/10.1108/OIR-10-2014-0232

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. PloS One, 10(6), e0128193. https://doi.org/10.1371/journal.pone.0128193

Dunning, D. (2011). The Dunning-Kruger effect: on being ignorant of one's own ignorance. Advances in Experimental Social Psychology, 44, 247-296. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Feng, V. W. & Hirst, G. (2013). Detecting deceptive opinions with profile compatibility. Sixth International Joint Conference on Natural Language Processing, 338-346.

Fichman, P. (2011). A comparative assessment of answer quality on four question answering sites. Journal of Information Science, 37(5), 476-486. https://doi.org/10.1177/0165551511415584

Finney Rutten, L. J., Blake, K. D., Greenberg-Worisek, A. J., Allen, S. V., Moser, R. P., & Hesse, B. W. (2019). Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Reports, 134(6), 617-625. https://doi.org/10.1177/0033354919874074

Flanagin, A. J. & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. New Media and Society, 9(2), 319-342. https://doi.org/10.1177/1461444807075

Flanagin, A. J. & Metzger, M. J. (2008). The credibility of volunteered geographic information. GeoJournal, 72(3-4), 137-148. https://doi.org/10.1007/s10708-008-9188-y

Freeman, K. S. & Spyridakis, J. H. (2004). An examination of factors that affect the credibility of online health information. Technical Communication, 51(2), 239-263.

Friedman, D. B. & Hoffman-Goetz, L. (2006). A systematic review of readability and comprehension instruments used for print and web-based cancer information. Health Education & Behavior, 33(3), 352-373. https://doi.org/10.1177/109019810527732

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). TweetCred: real-time credibility assessment

of content on Twitter. https://doi.org/10.1007/978-3-319-13734-6_16

Hilligoss, B. & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: construct, heuristics, and interaction in context. Information Processing & Management, 44(4), 1467-1484. https://doi.org/10.1016/j.ipm.2007.10.001

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.

Jin, X. L., Zhou, Z., Lee, M. K. O., & Cheung, C. M. K. (2013). Why users keep answering questions in online question answering communities: a theoretical and empirical investigation. International Journal of Information Management, 33(1), 93-104. https://doi.org/10.1016/j.ijinfomgt.2012.07.007

Katz, J. E. & Rice, R. E. (2002). Social Consequences of Internet Use: Access, Involvement, and Interaction. Cambridge: MIT press.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments. Information Retrieval, 16(2), 138-178. https://doi.org/10.1007/S10791-012-9205-0/FIGURES/10

Kim, H. & Choi, B. (2018). A comparative examination of factors that affect the credibility of health information on social media. Proceedings of the Association for Information Science and Technology, 55(1), 843-844. https://doi.org/10.1002/pra2.2018.14505501141

Landauer, T. K., Folt, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2), 259-284. https://doi.org/10.1080/01638539809545028

Lim, W. Y., Lee, M. L., & Hsu, W. (2017). iFACT: an interactive framework to assess claims from Tweets. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17, 787-796. https://doi.org/10.1145/3132847.3132995

Liu, X., Nielek, R., Wierzbicki, A., & Aberer, K. (2013). Defending imitating attacks in web credibility evaluation systems. Proceedings of the 22nd International Conference on World Wide Web, 1115-1122. https://doi.org/10.1145/2487788.2488131

Loria, S. (2018). TextBlob (0.15). Available: https://app.readthedocs.org/projects/textblob/downloads/pdf/dev/

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 3818-3824. https://doi.org/10.5555/3061053.3061153

Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15, 1751-1754. https://doi.org/10.1145/2806416.2806607

Ma, T. & Atkin, D. (2017). User generated content and credibility evaluation of online health information: a meta analytic study. Telematics and Informatics, 34(5), 472-486. https://doi.org/10.1016/j.tele.2016.09.009

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. ACL (System Demonstrations), 55-60. https://aclanthology.org/P14-5010.pdf

Maon, S. N., Hassan, N. M., & Seman, S. A. A. (2017). Online health information seeking behavior pattern. Advanced Science Letters, 23(11), 10582-10585. https://doi.org/10.1166/ASL.2017.10107

Markowitz, D. M. & Hancock, J. T. (2014). Linguistic traces of a scientific fraud: the case of Diederik Stapel. PLoS ONE, 9(8), e105937. https://doi.org/10.1371/journal.pone.0105937

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. Journal of Communication, 60(3), 413-439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41. https://doi.org/10.1145/219717.219748

Nagura, R., Seki, Y., Kando, N., & Aono, M. (2006). A method of rating the credibility of news documents on the web. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '06, 683. https://doi.org/10.1145/1148170.1148316

Noreen, E. W. (1989). Computer-intensive methods for testing hypotheses. Wiley. Available: https://www.jeffreycjohnson.org/app/download/764734156/cimeth.pdf

Open Dataset | Yelp Data Licensing. (n.d.) Available: https://business.yelp.com/data/resources/open-dataset/

Petty, R. E. & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In Communication and Persuasion: Central and Peripheral Routes to Attitude Change (pp. 1-24). Springer. https://doi.org/10.1007/978-1-4612-4964-1_1

Petty, R. E. (1977). The importance of cognitive responses in persuasion. ACR North American Advances, 4, 357-362.

Racherla, P. & Friske, W. (2012). Perceived 'usefulness' of online consumer reviews: an exploratory investigation across three services categories. Electronic Commerce Research and Applications, 11(6), 548-559. https://doi.org/10.1016/j.elerap.2012.06.003

Rains, S. A. & Karmikel, C. D. (2009). Health information-seeking and perceptions of website credibility: examining web-use orientation, message characteristics, and structural features of websites. Computers in Human Behavior, 25(2), 544-553. https://doi.org/10.1016/j.chb.2008.11.005

Rieh, S. Y. & Belkin, N. J. (1998). Understanding judgment of information quality and cognitive authority

in the WWW. Proceedings of the 61st Annual Meeting of the American Society for Information Science, 35, 279-89. https://doi.org/10.1002/asi.10017

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. Journal of the American Society for Information Science and Technology, 53(2), 145-161. https://doi.org/10.1002/asi.10017

Rieh, S. Y. (2010). Credibility and cognitive authority of information. In M. Bates & M. N. Maack (Eds.), Encyclopedia of Libraryand Information Sciences, (pp. 1337-1377). Oxfordshire: Taylor and Francis Group, LLC.

Rieh, S. Y., Jeon, G. Y., Yang, J. Y., & Lampe, C. (2014). Audience-aware credibility: from understanding audience to establishing credible blogs. Proceeding of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM 2014), 436-445.

Shah, C. & Pomerantz, J. (2010). Evaluating and predicting answer quality in community QA. Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10, March 2008, 411-418. https://doi.org/10.1145/1835449.1835518

Shang, J., Shen, J., Sun, T., Liu, X., Gruenheid, A., Korn, F., Lelkes, A. D., Yu, C., & Han, J. (2018). Investigating rumor news using agreement-aware search. Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM '18, 2117-2125. https://doi.org/10.1145/3269206.3272020

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 254-263.

Strapparava, C. & Valitutti, A. (2004). WordNet affect: an affective extension of WordNet. LREC, 4, 1083-1086.

Sundar, S. S. (2008a). Self as source: agency and customization in interactive media. 72-88. https://doi.org/10.4324/9780203926864-12

Sundar, S. S. (2008b). The MAIN model: a heuristic approach to understanding technology effects on credibility. Digital Media, Youth, and Credibility, 73-100. https://doi.org/10.1162/dmal.9780262562324.073

Wang, R. Y. (1996). Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-34. https://doi.org/10.1080/07421222.1996.11518099;CTYPE:STRING:JOURNAL

Wathen, C. N. & Burkell, J. (2002). Believe it or not: factors influencing credibility on the Web. Journal of the American Society for Information Science and Technology, 53(2), 134-144.

https://doi.org/10.1002/asi.10016

Webscope | Yahoo Labs. (n.d.). Available: https://webscope.sandbox.yahoo.com/

Yamamoto, Y. & Tanaka, K. (2011). Enhancing credibility judgment of web search results. Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11, 1235. https://doi.org/10.1145/1978942.1979126

Appendix A. Description of Machine Learning Features (numbers in parentheses indicate the number of features in each category)

Top-level	Middle-level	Description
Content Informativeness (15)	Plausibility (2)	measures similarity between answers and documents searched by Google based on the assumption that information similar to Google search results would be plausible, considering its reputation and performance. We used questions from Yahoo! Answers as search queries in Google Custom Search Engine (CSE) and considered returned websites as comparable documents. We confined sources to eight highly reliable sites (e.g., WebMD, Mayo Clinic, and MedlinePlus). The similarity was calculated using cosine similarity. Term frequency (TF) and term frequency-inverse document frequency (TF-IDF) were used to reflect the weighting of each term. The features we created were: (1) cosine similarity based on TF and (2) cosine similarity based on TF-IDF.
	Relevance (3)	measures similarity between a question and the corresponding answer. As traditional comparisons using bag-of-words representations often fail to account for synonyms (e.g., "cancer" vs. "neoplasm") or word variants (e.g., "run," "ran," "running"), we mapped raw text to Unified Medical Language System (UMLS) medical concepts, using MetaMap (Aronson & Lang, 2010). To incorporate term importance, we used three weighting schemes: TF, TF-IDF, and Latent Semantic Analysis (LSA) (Landauer et al., 1998), which captures semantic associations among terms. The resulting features were: (1) UMLS-based cosine similarity using TF, (2) UMLS-based cosine similarity using TF-IDF, and (3) UMLS-based cosine similarity using LSA.
	Comprehensiveness (6)	measures the extent to which an answer addresses all or most aspects of a question, or a review reflects various facets of a medical facility. In particular, a list of keywords (e.g., pros/cons, +/-, and plus/minus) that indicate whether a review is structured and discusses both positive and negative attributes was created for the Yelp dataset. The features we created were: (1) word count, (2) lexical diversity, (3) document entropy (Manning et al., 2014), (4) the ratio of UMLS concepts to total words, (5) presence of structural information (e.g., pros/cons format), and (6) presence of ratings or comments addressing specific service aspects (e.g., taste, service, amenities).
	Specificity (3)	measures the degree to which a review provides concrete and vivid examples to support its claims. Specific references, such as pricing, locations, or dates, enhance the informativeness of a review. We extracted these indicators by counting named entities like "Money," "Location," and "Date" using Stanford CoreNLP (Manning et al., 2014), and by identifying relevant semantic types such as "Clinical Drug," "Health Care Activity," and "Quantitative Concept" through MetaMap (Aronson & Lang, 2010). The derived features include: (1) count of named entity tags corresponding to specific details, (2) count of UMLS medical concepts associated with specificity, and (3) frequency of price-related terms.
	Currency (1)	captures whether an answer or review includes expressions indicating it was recently written. Given the rapidly evolving nature of medical information and the dynamics of healthcare services, we defined that "within the past six months" indicates recency in this study. These expressions were identified through a manual analysis of 500 randomly selected samples and subsequently detected using regular expression patterns.

Top-level	Middle-level	Description
Sentiment (55)	-	We utilized a hand-crafted lexicon of emotion-related terms derived from WordNet-Affect (Strapparava & Valitutti, 2004), an extension of the WordNet database (Miller, 1995), which organizes words into synsets representing specific emotional concepts. For instance, *liking* is one of the positive emotional concepts, and *fondness*, *preference*, and *admiration* are terms for that concept. For negation detection, we used Stanford CoreNLP, which achieves up to 81.8% accuracy. Detected negated expressions were reverse-coded (e.g., negated "*liking*" became *not_liking*). Additionally, we used the Python TextBlob library (Loria, 2018) to extract polarity and subjectivity scores. In total, 55 sentiment-related features were generated for analysis.
Presentation (7)	Readability (6)	We calculated various readability scores using a Python Readability library. Due to space limitations, we refer readers to the systematic review of readability scores by Friedman and Hoffman-Goetz (Friedman & Hoffman-Goetz, 2006) for detailed explanations of these metrics. In addition, The features we created were: ARI, FleschKincaid Grade Level, Gunning Fog Index, Coleman Liau Index, LIX, and RIX.
	Spelling (1)	We assessed spelling accuracy by comparing tokenized, non-stopword text against three benchmark English corpora (NLTK, Brown, and Reuters). After tokenization and stopword removal, words not found in these three corpora were counted and normalized by the total word count to derive a spelling accuracy metric.
Source (7)	Expertise (1)	captures the level of domain knowledge presented by the author, as individuals with greater expertise are generally perceived as more credible. In this study, we operationalized expertise by counting the number of Unified Medical Language System (UMLS) concepts that belong to categories indicative of specialized knowledge of the author. Examples of concept categories include "Anatomical Structure," "Genetic Function," and "Pharmacologic Substance."
	Reference (2)	measures the presence of cited authoritative sources. We compiled a list of 21 reputable health information sources—such as WebMD, Healthline.com, and Health.com—based on recommendations from the University of Michigan Health Service, the National Institutes of Health, and Refseek.com. We then counted the number of citations in each answer or review that matched entries on this list. Additionally, hyperlinks with top-level domains such as ".gov," ".org," or ".edu" were also considered indicators of authoritative references.
	Credence (5)	measures social network metrics of each user in a social network, as proxies for credence. These features were extracted only from the Yelp dataset, as it uniquely included network-relevant attributes such as friend lists. We used the Pajek tool (Batagelj & Mrvar, 1998) to compute standard social network measures, including degree centrality, eigenvector centrality, clustering coefficient, betweenness centrality, and closeness centrality.