I Want to Believe: Journalists and Crowdsourced Accuracy Assessments in Twitter

Cody Buntain
University of Maryland
College Park, Maryland 20742
cbuntain@cs.umd.edu

Jennifer Golbeck
University of Maryland
College Park, Maryland 20742
golbeck@cs.umd.edu

ABSTRACT

Evaluating information accuracy in social media is an increasingly important and well-studied area, but limited research has compared journalist-sourced accuracy assessments to their crowdsourced counterparts. This paper demonstrates the differences between these two populations by comparing the features used to predict accuracy assessments in two Twitter data sets: CREDBANK and PHEME. While our findings are consistent with existing results on feature importance, we develop models that outperform past research. We also show limited overlap exists between the features used by journalists and crowdsourced assessors, and the resulting models poorly predict each other but produce statistically correlated results. This correlation suggests crowdsourced workers are assessing a different aspect of these stories than their journalist counterparts, but these two aspects are linked in a significant way. These differences may be explained by contrasting factual with perceived accuracy as assessed by expert journalists and non-experts respectively. Following this outcome, we also show preliminary results that models trained from crowdsourced workers outperform journalist-trained models in identifying highly shared "fake news" stories.

CCS CONCEPTS

•Information systems →Web and social media search; •Human-centered computing →Empirical studies in collaborative and social computing;

KEYWORDS

misinformation, data quality, fake news, twitter

1 INTRODUCTION

While accuracy and credibility have been well-studied across disciplines [4, 11, 12], the introduction of fast-paced social media and a new generation of "citizen journalists" [7] and "digital gatekeepers" [2] has necessitated a re-examination of these issues. Recent forays into this re-examination explore rumor propagation on Twitter [8, 16, 21], the sharing of fake images of disaster aftermath [5], and politically motivated "astroturfing" [17], but rumors and "fake news" seem to be increasingly important despite this attention. Some researchers are now suggesting these factors played a significant (and potentially deliberate) role in the 2016 US presidential election [24], and recent investigations have identified strong monetary incentives for publishing fake news stories [22]. While one may expect the new generation of young "digital natives" to be more adept at separating these "fake news" stories from truth, recent research has shown this expectation is far from true [20, 21].

To support users in these judgements, recent efforts have made significant progress in identifying features of credible social media stories [3, 8, 15]. The ground truth for this research, however, is often acquired by asking either journalists or crowdsourced workers to evaluate social media stories. These works implicitly assume a population of crowdsourced workers are valid substitutes for journalists with fact-checking expertise, with little research investigating whether these populations assess stories differently. Given the lack of preparation users have in making these critical judgements and their positive bias towards believing online stories [14, 21], understanding these differences is crucial for evaluating whether crowdsourced assessments are valid proxies for journalistic vigor. This paper investigates these differences by comparing the factors that predict a professional journalist's accuracy assessments on Twitter versus accuracy assessments produced by crowdsourced workers. We use these factors to train learning models to predict both journalist and crowdsourced accuracy assessments and test whether these models produce statistically correlated results. These models are then evaluated against a new data set of potential "fake news" stories that were highly shared on Twitter. This third data set provides a means to validate our models and determine which population model is best suited for identifying these stories.

This research leverages two existing sets of Twitter data: the PHEME rumor scheme data set [25] and the CREDBANK data set [14]. PHEME is a curated data set of conversation threads about rumors in Twitter replete with journalist annotations for truth, and CREDBANK is a large-scale set of Twitter conversations about events and corresponding crowdsourced accuracy assessments for each event. We first align CREDBANK's structure with PHEME's and develop a new data set using features inspired by existing work on credibility in Twitter [3, 8]. A feature selection study then determines which features contribute the most to journalist accuracy annotations and shows these features differ from those that contribute to crowdsourced accuracy annotations. We then apply our PHEME models to CREDBANK and vice versa to predict journalists' labels for CREDBANK threads and crowdsourced workers' labels for PHEME threads respectively. We show these models are not equivalent but do produce statistically correlated results. We theorize this divergence between crowdsourcing and journalism is rooted in the difference between **credibility** and **accuracy**: accuracy is only one aspect of credibility, and while journalists in PHEME are establishing accuracy by fact-checking their stories, crowdsourced workers are evaluating their stories' believability or how credible the authors and stories appear.

Lastly, after applying both models to a third set of potential "fake news" stories shared on Twitter and sourced from BuzzFeed

News [19], our findings show crowdsourced models outperform journalistic models in classifying true and false stories.

This work makes the following contributions:

- Establishes feature importance for predicting accuracy labels from journalists and crowdsourced workers,
- Demonstrates models of journalistic and crowdsourced accuracy assessments produce statistically correlated but otherwise divergent results, and
- Evaluates journalist and crowdsourced accuracy models against a set of "fake news" articles, showing the model of crowdsourced assessment better separates truth from fiction.

2 RELATED WORK

Social media's explosions in popularity has enabled research into credibility in the online context, especially on microblogging platforms. Several previous efforts have proposed methods for evaluating the credibility of a given tweet [16] or user [6] while others have focused more on the temporal dynamics of rumor propagation [8]. Most relevant to our paper, however, is the 2013 Castillo et al. work, which provides a comprehensive examination of credibility features in Twitter [3]. This study was built on an earlier investigation into Twitter usage during the 2010 Chile earthquake, where Twitter played a significant role both in coordination and misinformation [13]. The later study developed a system for identifying newsworthy topics from Twitter and leveraged Amazon's Mechanical Turk (AMT) to generate labels for whether a topic was credible, similar to CREDBANK but at a smaller scale. Castillo et al. developed a set of 68 features that included characteristics of messages, users, and topics as well as the propagation tree to classify topics as credible or not. They found a subset of these features, containing fifteen topic-level features and one propagation tree feature, to be the best performing feature set, with a logistic regression model achieving an accuracy of 64% for credibility classification. Given general users have difficulty judging correct and accurate information in social media [20, 21], however, crowdsourced credibility assessments like these should be treated with caution. The investigation presented herein builds on this past work by evaluating whether crowdsourced workers (as used in both CREDBANK and Castillo et al.) are valid accuracy assessment sources.

3 DATA SET DESCRIPTIONS

This work leverages three data sets to compare and evaluate crowdsourced and journalist-sourced accuracy assessments. The two primary data sets with which this work is concerned are the PHEME rumor scheme data set and the CREDBANK crowdsourced data set. The third data set consists of a set tweets of headlines provided by BuzzFeed News on "Fact-Checking Facebook Politics Pages."

3.1 The PHEME Rumor Data Set

The PHEME rumor scheme data set was developed by the University of Warwick in conjunction with Swissinfo, part of the Swiss Broadcasting Company [25]. Swissinfo journalists, working with researchers from Warwick, constructed the PHEME data set by following a set of major events on Twitter and identifying threads of conversation that were likely to contain or generate rumors. A "rumor" in this context was defined as an unverified and relevant statement being circulated, and a rumor could later be confirmed as true, false, or left unconfirmed. Relating "rumor" to this paper's research, an unconfirmed rumor is a thread of conversation with unknown accuracy; a rumor can then be confirmed as true or false or left undetermined.

PHEME's events included the social unrest in Ferguson, MO in 2014 following the shooting of Michael Brown, the Ottawa shooting in Canada in October 2014, the 2014 Sydney hostage crisis and Charlie Hebdo attacks, and the Germanwings plane crash. The data set also contained conversations around four known rumors: conversations about a footballer, Michael Essien, possibly contracting ebola (later confirmed as false); a rumored secret concert performance in Toronto, Canada (later confirmed as false); a rumor about a museum in Bern, Germany accepting a Nazi-era art collection (later confirmed as true); and rumors about Russian president Vladimir Putin going missing in March 2015 (later confirmed as false).

During each of these events, journalists selected popular (i.e., highly retweeted) tweets extracted from Twitter's search API and labeled these tweets as rumor or non-rumor. This construction resulted in a set of 493 labeled rumorous source tweets. For each tweet in this labeled set, the authors then extracted follow-up tweets that replied to the source tweet and recursively collected descendant tweets that responded to these replies. This collection resulted in a tree-like set of conversation threads of 4,512 additional descendant tweets. In total, the currently available version of PHEME contains 4,842 tweets.

The Swissinfo journalists labeled source tweets for each of these threads as true or false. Once this curated set of labeled source tweets and their respective conversation threads were collected, the PHEME data set was then made available to crowdsourced annotators to identify characteristics of these conversation threads. This crowdsourced task asked annotators to identify levels of support (does a tweet support, refute, ask for more information about, or comment on the source tweet), certainty (tweet author's degree of confidence in his/her support), and evidentiality (what sort of evidence does the tweet provide in supporting or refuting the source tweet) for each tweet in the conversation. Past work found disagreement and refutation in threads to be predictive of accuracy [3], and these annotations of whether a tweet supports or refutes the original tweet help quantify this disagreement, which we leverage later.

3.2 The CREDBANK Data Set

In 2015, Mitra and Gilbert introduced CREDBANK, a large-scale crowdsourced data set of approximately 60 million tweets, 37 million of which were unique. The data set covered 96 days starting in October of 2014, broken down into over 1,000 sets of event-related tweets, with each event assessed for accuracy by 30 annotators from AMT [14]. At a high level, CREDBANK was created by collecting tweets from Twitter's public sample stream, identifying topics within these tweets, and using human annotators to determine which topics were about events and which of these events contained accurate content. Then, the systems used Twitter's search API to expand the set of tweets for each event.

CREDBANK's initial set of tweets from the 96-day capture period contained approximately one billion tweets that were then filtered for spam and grouped into one-million-tweet windows. Mitra and Gilbert used online topic modeling from Lau et al. [9] to extract 50 topics (a topic here is a set of three tokens) from each window, creating a set of 46,850 candidate event-topic streams. Each potential event-topic was then passed to 100 annotators on AMT and labeled as an event or non-event, yielding 1,049 event-related topics (the current version of CREDBANK contains 1,377 events). These event-topics were then sent to 30 AMT users to determine the event-topic's accuracy.

This accuracy annotation task instructed users to assess "the credibility level of the Event" by reviewing tweets returned from searching for the event's keywords on Twitter's website (see Figure 5 in Mitra and Gilbert [14]). After reviewing a selection of relevant tweets, annotators were asked to provide an accuracy rating on a 5-point Likert scale of "factuality" (adapted from Sauri et al. [18]) from $[-2, +2]$, where -2 represented "Certainly Inaccurate" and $+2$ was "Certainly Accurate" [14]. Annotators were required to provide a justification for their choice as well.

Once these tweets, topics, event annotations, and accuracy annotations were collected, this data was published as the CREDBANK data set.[1] Data provided in CREDBANK includes the three-word topics extracted from Twitter's sample stream, each topic's event annotations, the resulting set of event-topics, a mapping of event-topics' relevant tweets, and a list of the AMT accuracy annotations for each event-topic. One should note that CREDBANK does not contains binary labels of event accuracy but instead has a 30-element vector of accuracy labels.

In CREDBANK, the vast majority ($> 95\%$) of event accuracy annotations had a majority rating of "Certainly Accurate" [14]. Only a single event had a majority label of inaccurate: the rumored death of Chris Callahan, the kicker from Baylor University's football team, during the 2015 Cotton Bowl (this rumorous event was clearly false as Callahan was tweeting about his supposed death after the game). After presenting this tendency towards high ratings, Mitra and Gilbert thresholds for majority agreement and found that 76.54% of events had more than 70% agreement, and 2% of events had 100% agreement among annotators. The authors then chose 70% majority-agreement value as their threshold, and 23% of events in which less than 70% of annotators agreed were "not perceived to be credible" [14]. This skew is consistent with Castillo et al. [3], where authors had to remove the "likely to be true" label because crowdsourced workers labeled nearly all topics thusly. We address this bias below.

3.2.1 *Aligning CREDBANK with PHEME.*

Despite both data sets focusing on accuracy annotations, CREDBANK and PHEME are significantly different in their construction. This section describes the steps needed to convert CREDBANK into a form comparable to PHEME, which includes inferring accuracy labels from CREDBANK's annotations, constructing threads of conversation from CREDBANK data, and classifying tweet content for disagreement.

The first major difference between these sets is the labeling scheme: While PHEME contains truth labels for each thread of conversation, CREDBANK instead contains a collection of annotator accuracy assessments. The first alignment task is then to convert the sets of annotator accuracy assessments into binary labels comparable to PHEME's "truth" labels. Given annotator bias towards "certainly accurate" assessments and the resulting negatively skewed distribution of average assessments, a labeling approach that addresses this bias is required.

Since majority votes are uninformative in CREDBANK, we instead turn to statistics and compute the mean accuracy rating for each event and quartiles for the set of all mean ratings. First, the grand mean of CREDBANK's accuracy assessments is 1.7, but the median is 1.767, and the 25th and 75th quartiles are 1.6 and 1.867 respectively; since the median is closer to the 75th quartile, we know this distribution is negatively skewed. We then theorize that an event whose mean rating is towards the minimum or maximum mean ratings in CREDBANK are more likely to be classified false or true respectively (though these events may not be true in the real world but are only perceived to be true). To construct "truth" labels from this data, we look at the top and bottom 15%, so events with average accuracy ratings more than 1.9 become positive samples or less than 1.467 become negative samples. Events between these values are indeterminate and unlabeled. This labeling process results in 203 positive events and 185 negative events.

The second major difference between CREDBANK and PHEME is the form of tweet sets: in PHEME, topics are organized into threads, starting with a popular tweet at the root and replies to this popular tweet as the children, whereas CREDBANK simply contains all tweets that match the three-word topic query. To better compare these two data sets, we adapt CREDBANK's tweet sets into threads by identifying the most popular tweet in each event (we define popularity as the number of retweets) and use PHEME's thread capture software to construct threads of conversation from these roots. We then discard any CREDBANK thread that has no reactions, leaving a final total of 115 positive samples and 95 negative samples. In this manner, we are able to compare threads of conversation and predict their annotations from both journalists and crowdsourced workers.

Inferring Disagreement in Tweets. Finally, one of the more important features suggested in Castillo et al. is the amount of disagreement or contradiction present in a conversation [3]. PHEME already contains this information in the form of "support" labels for each reply to the thread's root, but CREDBANK lacks these annotations. To address this omission, we developed a classifier for classifying tweets that express disagreement using a combination of the support labels in PHEME and the "disputed" labels in the CreateDebate data set of version 2 of the Internet Argument Corpus (IACv2) [1]. We merged these two datasets into a single ground-truth set which we then used to train this disagreement classifier; this augmented set was necessary as relying only on PHEME's data was unable to achieve acceptable accuracy (area under the receiver operating characteristic curve of 72.66%).

Tweet and IACv2 forum texts were featurized into bags of unigrams and bigrams. After experimenting with support vector machines, random forests, and naive Bayes classifiers, we found stochastic gradient descent to be the best predictor of disagreement and disputed labels. 10-fold cross validation of this classifier

[1] Available online http://compsocial.github.io/CREDBANK-data/

achieved a mean area under the receiver operating characteristic curve of 86.7%. We then applied this classifier to the CREDBANK data set to assign disagreement labels for each tweet. A human then reviewed a random sample of these labels. While human annotators would be better for this task, an automated classifier was preferable given CREDBANK's size.

3.3 BuzzFeed News Fact-Checking Data Set

In late September 2016, journalists from BuzzFeed News collected over 2,000 posts from nine large, verified Facebook pages (e.g., Politico, CNN, AddictingInfo.org, and Freedom Daily) [19]. Three of these pages were from mainstream media sources, three were from left-leaning organizations, and three were from right-leaning organizations. BuzzFeed journalists fact-checked each post, labeling it as "mostly true," "mostly false," "mixture of true and false," or "no factual content." Each post was then checked for engagement by collecting the number of shares, comments, and likes on the Facebook platform. In total, this data set contained 2,282 posts, 1,145 from mainstream media, 666 from right-wing pages, and 471 from left-wing pages [19].

Many of these posts were links to articles on the Facebook page owner's website, all of which have a presence on Twitter as well. To align this data with PHEME and CREDBANK, we extracted the ten most shared stories from left- and right-wing pages and searched Twitter for these headlines (we selected a balanced set from both political sides to avoid bias based on political leaning). We then kept the top three most retweeted posts for each headline and extracted the conversation threads from each of these posts using the same code we used for recovering CREDBANK threads. This process resulted in 35 conversation threads with journalist-provided labels, 15 of which were mostly true, and 20 were mostly false.

4 EXPERIMENTAL DESIGN

Motivating this work is the question of whether crowdsourced accuracy assessments are a valid substitute for journalistic expertise. This objective decomposes into a set of testable hypotheses centered around our theory that these groups maintain distinct models for constructing their accuracy labels. At a high level, we explore three hypothetical components of this theory by developing models to predict accuracy assessments for both populations. These models lead to our first hypothesis **H1**: the features crowdsourced workers use to assess accuracy will necessarily differ from those used by journalists. We test this hypothesis via a feature selection experiment to identify the top ten most important features in each model and an evaluation of each model's performance.

Regardless of whether these populations leverage different features when determining the accuracy of a social media thread, an oft-made assumption is that crowdsourced accuracy assessments should still be correlated to journalists' in some way. This assumption leads to our second hypothesis **H2**: accuracy labels produced by models of journalists and models of crowdsourced workers should be similar. Assuming the models perform well in their native contexts, if the labels produced by each model are statistically independent of the true labels, then no relationship exists between crowdsourced and journalists' evaluations, and crowdsourcing should not be used as a proxy for journalists. Alternatively, if these models do produce correlated results, then leveraging crowdsourced workers for accuracy assessment may be a valid approach despite known issues with lay-evaluations.

Results from these hypotheses are valuable for future evaluations and designing informative interfaces, but these models could be especially timely for differentiating between true and fake news stories. We theorize that rumor-sharing patterns on social media are similar to those governing how fake news stories are shared, motivating our final hypothesis **H3**: models for predicting accuracy for rumorous threads should also predict the truth of potentially fake news stories based on how they are shared in Twitter. Our final experiment tests this theory by applying our most performant CREDBANK and PHEME models to the BuzzFeed News fact-checking data set. Evaluating these models in this new context adds an additional layer of model validation and, if these models perform well in this new context, this work can support future research into identifying "fake news."

The following subsections present the details for testing these hypotheses.

4.1 Features for Predicting Accuracy

Prior to discussing methods for modeling and comparing PHEME and CREDBANK, we first describe 45 features we use across four types: structural, user, content, and temporal. This feature set includes fourteen of the sixteen most important features found in Castillo et al., omitting the two features on most frequent web links since those features do not lend themselves to the streaming context. Structural features capture Twitter-specific properties of the tweet stream, including tweet volume and activity distributions (e.g., proportions of retweets or media shares). User features capture properties of tweet authors, such as interactions, account ages, friend/follower counts, and Twitter verified status. Content features measure textual aspects of tweets, like polarity, subjectivity, and agreement. Lastly, temporal features capture trends in the previous features over time, e.g., the slopes of the number of tweets or average author age over time. As mentioned, many features were inspired by or reused from Castillo et al. [3] (indicated by ⋆).

4.1.1 Structural Features. Structural features are specific to each Twitter conversation thread and are calculated across the entire thread. These features include the number of tweets, average tweet length⋆, thread lifetime (number of minutes between first and last tweet), and the depth of the conversation tree (inspired by other work that suggests deeper trees are indicators of contentious topics [23]). We also include the frequency and ratio (as in Castillo et al.) of tweets that contain hashtags, media (images or video), mentions, retweets, and web links⋆.

4.1.2 User Features. While the previous set focuses on activities and thread characteristics, the following features are attributes of the users taking part in the conversations, their connectedness, and the density of interaction between these users. User features include account age⋆; average follower-⋆, friend-, and authored status counts⋆; frequency of verified authors⋆, and whether the author of the first tweet in the thread is verified. We also include the difference between when an account was created and the relevant tweet was authored (to capture bots or spam accounts).

This last user-centric feature, network density, is measured by first creating a graph representation of interactions between a conversation's constituent users. Nodes in this graph represent users, and edges correspond to mentions and retweets between these users. The intuition here is that highly dense networks of users are responding to each other's posts and endogenous phenomena. Sparser interaction graphs suggest the conversation's topic is stimulated by exogenous influences outside the social network and are therefore more likely to be true.

4.1.3 Content Features. Content features are based on tweets' textual aspects and include polarity* (the average positive or negative feelings expressed a tweet), subjectivity (a score of whether a tweet is objective or subjective), and disagreement*, as measured by the amount of tweets expressing disagreement in the conversation. As mentioned in PHEME's description, tweet annotations include whether a tweet supports, refutes, comments on, or asks for information about the story presented in the source tweet. These annotations directly support evaluating the hypothesis put forth in Mendoza, Poblete, and Castillo [13], stating that rumors contain higher proportions of contradiction or refuting messages. We therefore include these disagreement annotations (only a binary value for whether the tweet refutes the source). Also borrowing from Castillo et al., we include the frequency and proportions of tweets that contain question marks, exclamation points, first/second/third-person pronouns, and smiling emoticons.

4.1.4 Temporal Features. Recent research has shown temporal dynamics are highly predictive when identifying rumors on social media [8], so in addition to the frequency and ratio features described above, we also include features that describe how these values change over time. These features are developed by accumulating the above features at each minute in the conversation's lifetime and converting the accumulated value to logarithmic space. We then fit a linear regression model to these values in log space and use the slope of this regression as the feature's value, thereby capturing how these features increase or decrease over time. We maintain these temporal features for account age, difference between account age and tweet publication time, author followers/friends/statuses, and the number of tweets per minute.

4.2 Feature Selection and Model Evaluation

The previous section presents the features we use to capture structures and behaviors in rumorous conversation threads on Twitter. To test **H1**, we perform a recursive feature elimination study and evaluate how well these features predict CREDBANK and PHEME labels.

To evaluate feature and model performance, we measure the area under the receiver operating characteristic curve (ROC-AUC) for each model. The ROC metric varies the decision threshold for classification and calculates the true and false positive rate for each threshold value to construct a performance curve; the area under this ROC curve then characterizes how well the model performs on a scale of 0 to 1 (a random coin toss would achieve a ROC-AUC of 0.5). The ROC-AUC metric is also more resistant to class imbalance than F1 scores. For each feature set, we perform thirty instances of 10-fold cross-validation using a 100-tree random forest classifier (an ensemble made of 100 separate decision trees trained on a random feature subset) to estimate the ROC-AUC for that feature set.

With the classifier and evaluation metric established, our feature selection process recursively removes the least performant feature in each iteration until only a single feature remains. The least performant feature is determined using a leave-one-out strategy: in an iteration with k features, k models are evaluated such that each model uses all but one held-out feature, and the feature whose *exclusion* results in the *highest* ROC-AUC is removed from the feature set. This method demonstrates which features hinder performance since removing important features will result in losses in ROC-AUC score, and removing unimportant or bad features will either increase ROC-AUC or have little impact. Given k features, the process will execute $k - 1$ iterations, and each iteration will output the highest scoring model's ROC-AUC. By inspecting these $k - 1$ maximum scores, we determine the most important feature subset by identifying the iteration at which the maximum model performance begins to decrease.

4.3 Cross-Context Evaluation

With performant features and models developed for CREDBANK and PHEME, we then evaluate **H2** and determine whether and to what extent crowdsourced workers' and journalists' accuracy assessments correlate with each other. We measure this correlation after transferring PHEME models to CREDBANK and CREDBANK models to PHEME and evaluating each model in these alternate contexts. This cross-context examination uses two metrics: the thirty-round repeated ROC-AUC measurement described above, and a chi-squared test for independence between actual and predicted labels. The ROC-AUC metric yields insight into model accuracy, and the chi-squared test captures whether accuracy assessments from journalists influence crowdsourced workers and vice versa.

Regardless of whether a connection exists between journalists and crowdsourced workers, parallels between accuracy assessment of potential rumors and "fake news" identification suggest PHEME or CREDBANK models could also identify these stories, as we suggest in **H3**. As above, testing this final hypothesis is also a cross-context evaluation task, so we test this hypothesis using the same metrics used for **H2**.

5 RESULTS

5.1 Feature Selection

Recursively removing features from our models and evaluating classification results yielded significantly reduced feature sets for both PHEME and CREDBANK, the results of which are shown in Figure 1. The highest performing feature set for PHEME only contained seven of the 45 features: proportions and frequency of tweets sharing media; proportions of tweets sharing hashtags; proportions of tweets containing first- and third-person pronouns; proportions of tweets expressing disagreement; and the slope of the average number of authors' friends over time. The top ten features also included account age, frequency of smile emoticons, and author friends. This PHEME feature set achieved an ROC-AUC score of 0.7407 and correctly identified 66.93% of potentially false threads.

CREDBANK's most informative feature set used 12 of the 45 features: frequencies of smiling emoticons, tweets with mentions, and tweets with multiple exclamation or question marks; proportions of tweets with multiple exclamation marks, one or more question marks, tweets with hashtags, and tweets with media content; author account age relative to a tweet's creation date; average tweet length; author followers; and whether the a thread started with a verified author. Proportions of tweets with question marks and multiple exclamation/question marks were not in the top ten features, however. This feature set achieved an ROC-AUC score of 0.7184 and correctly identified 70.28% of potential false threads.

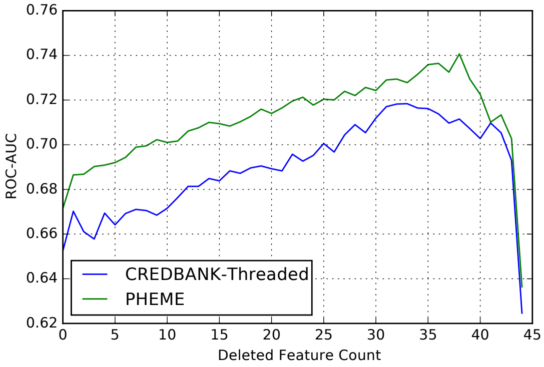

Figure 1: Feature Elimination Study

Of these feature subsets, only three features are shared by both crowdsourced worker and journalist models (frequency of smile emoticons and proportion of tweets with media or hashtags). Our first hypothesis **H1** theorized these assessors would use different features when making their judgements, and these results support this hypothesis.

These results are consistent with the difficulty in identifying potentially fallacious threads of conversation in Twitter discussed in Castillo et al. [3]. Furthermore, both PHEME and CREDBANK's top ten features contain five of the 16 best features found in Castillo et al. [3]. Despite these consistencies, our models outperform the model presented in this prior work (61.81% accuracy in Castillo et al. versus 66.93% and 70.28% in PHEME and CREDBANK). These increases are marginal, however, but are at least consistent with past results.

5.2 Crossing CREDBANK and PHEME Models

Given these ten features and classifiers trained on all the data in each data set, our next hypothesis **H2** stated that accuracy labels produced by models of journalists and models of crowdsourced workers should be similar. To test this hypothesis, we first constructed the ROC curves of applying our top-ten-feature model from CREDBANK to PHEME and vice versa, as shown in Figure 2. This cross-context application between data sets yielded poor classification performance: the classifier with the best ten features from PHEME applied to CREDBANK (P ⟶ C) scored a ROC-AUC of 57.80%, and the best CREDBANK classifier scored 55.31% on PHEME (C ⟶ P). Looking at accuracy, the PHEME model classified CREDBANK threads correctly 57.62% of the time, and the CREDBANK model captures PHEME threads correctly 51.53% of the time, which was only slightly better than random guessing.

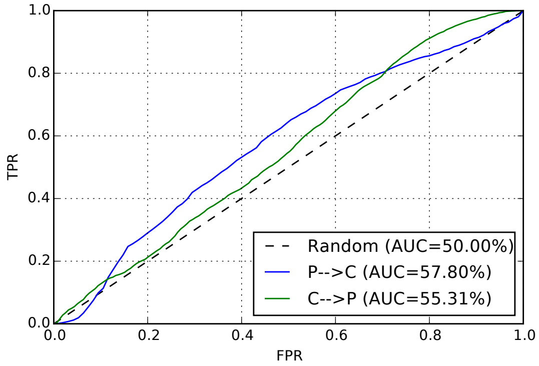

Figure 2: Cross-Evaluating Accuracy

While these models were inaccurate with respect to classification, they may still produce statistically correlated results. We tested this using a chi-squared test for independence between the actual and predicted labels for PHEME-to-CREDBANK and vice versa, the contingency tables for which are shown in Table 1. The null hypothesis in these tests was that no relationship exists between the actual and predicted labels, and we used $p < 0.05$ for significance. For PHEME to CREDBANK, this test showed a significant relationship between actual and predicted labels with $\chi^2(1, N = 210) = 5.637, p = 0.01758$. For CREDBANK to PHEME, however, this test showed no significant relationship, with $\chi^2(1, N = 330) = 0.2143, p = 0.6434$, a result on which we elaborate later.

Table 1: Cross-Evaluation Contingency Tables

(a) PHEME ⟶ CREDBANK

	Predicted		
	False	True	Total
Actual False	52	43	95
Actual True	43	72	115
Total	95	115	

(b) CREDBANK ⟶ PHEME

	Predicted		
	False	True	Total
Actual False	97	62	159
Actual True	99	72	171
Total	196	134	

5.3 Predicting BuzzFeed Fact-Checking

While previous results show only limited cross-context power for crowdsourced and journalistic accuracy labels, as mentioned in

H3, these models may still have predictive power for identifying fake news stories shared in Twitter. We tested this hypothesis by applying the both CREDBANK and PHEME models to our BuzzFeed News data set and calculating the ROC-AUC scores for each model, as shown in Figure 3. From this graph, CREDBANK-based models applied to BuzzFeed News performed nearly equivalently to their performance in their native context, achieving an ROC-AUC of 74.56% and accuracy of 64.57%. PHEME-based models, however, performed poorly, only achieving an ROC-AUC of 36.98% and accuracy of 34.43%. Neither data set's results were statistically correlated with the underlying actual labels either, with CREDBANK's $\chi^2(1, N = 35) = 2.803, p = 0.09409$ and PHEME's $\chi^2(1, N = 35) = 2.044, p = 0.1528$.

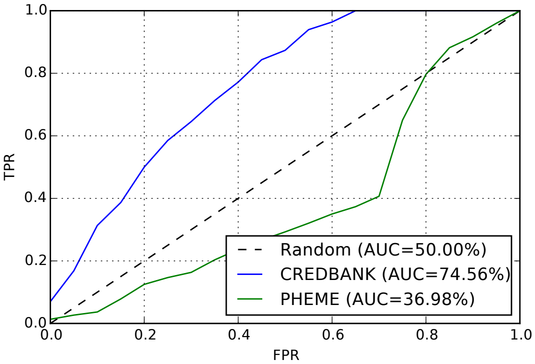

Figure 3: Adapting to Fake News Classification

6 FINDINGS

In analyzing the results presented above, we identify two main findings that require additional discussion. First, the distinction between crowdsourced and journalist accuracy models and the unidirectional relationship between them may result from assessing two fundamentally different qualities: Journalists evaluate factual accuracy while crowdsourced assessors measure credibility (or the appearance of accuracy). Second, the crowdsourced model's higher performance in identifying fake news may be a consequence of this unidirectional influence of accuracy on credibility: a headline story's accuracy drives users' perceptions of credibility in the Twitter conversation, which influences how they respond to and share this information. We detail these findings below.

Regarding contrasting accuracy models, we see diverging feature sets and poor cross-context performance between PHEME and CREDBANK. While both data sets are built on "accuracy" assessments, we theorize this question captures two separate qualities: for PHEME's journalists, "accuracy" is objective or factual truth, whereas CREDBANK's crowdsourced workers equate "accuracy" with credibility, or how believable the story seems. In PHEME, journalists evaluated the factual accuracy of conversation threads after "a consensus had emerged about the facts relating to the event in question" and after reviewing all the captured tweets relevant to that event [25]. CREDBANK assessors, however, did not necessarily have the benefit of hindsight as they were asked to review tweets rapidly and only saw those that were most current on the Twitter platform as part of CREDBANK's "real-time responsiveness" [14]. Furthermore, PHEME journalists could use any resource available to them to corroborate or contradict a given story, whereas the task design for CREDBANK workers focused their attention on the few tweets they saw during their task. Given these procedural differences and the monetary motivations of Mechanical Turk workers, CREDBANK assessors likely only had time to make a subjective judgement rather than a more thorough, objective fact-check like the journalists in PHEME. This distinction would also explain assessors' significant bias towards rating threads as accurate, which was present in both CREDBANK and Castillo et al. [3], since readers are pre-disposed to believe online news [10, 21].

While CREDBANK and PHEME seem to measure different aspects of information quality, these aspects are likely still related. The connection between accuracy and credibility is well-known in many disciplines, from journalism to psychology to human-computer interaction [4, 11, 12]. Our results are consistent with this link as well, as shown with **H2** in which we show a one-way statistically significant relationship exists between PHEME predictions and actual crowdsourced labels. As discussed in Maier, accuracy directly influences whether a reader finds a source credible [11]. Whether a story seems true (i.e., is credible), however, may have little impact on its actual truth. The lack of a bidirectional relationship between PHEME and CREDBANK in **H2** is consistent with this connection between accuracy and credibility.

Having established the link between PHEME and CREDBANK, we now discuss the potentially unintuitive result of **H3**, in which models of crowdsourced credibility assessment significantly outperform journalistic models in identifying journalist-fact-checked news stories. Since the BuzzFeed News data set was fact-checked by journalists, one might expect the PHEME model to perform better in this context. We propose an alternate explanation: when a thread in Twitter starts with a story headline and link, the story's underlying accuracy influences the thread's credibility, but the thread's credibility dictates how the story is shared. Stated another way, the underlying story's accuracy drives users' perceptions of credibility in the Twitter conversation, which influences how they respond, and the CREDBANK model captures this perception better than the PHEME model. Furthermore, our CREDBANK model is more rooted in the Twitter context than PHEME since CREDANK assessors were asked to make their judgements based solely on the tweets they saw rather than the additional external information PHEME journalists could leverage. From this perspective, CREDBANK models may be more appropriate for a social media-based automated fake news detection task since both rely primarily on signals endogenous to social media (rather than external journalistic verification). Finally, given the commensurate performance CREDBANK and PHEME exhibit in their native contexts, PHEME's poor performance for fake news suggest some fundamental difference between how endogenous rumors propagate in social media and how fake news is perceived and shared, but more work is needed here.

6.1 Limitations

While the results discussed herein suggest major differences between journalists and crowdsourced workers and between rumor and fake news propagation, limitations may influence our results. This work's main limitation lies in the differences between PHEME and CREDBANK's construction and structure. Regarding construction, differences in assessors' labeling tasks may impact the validity and comparability of our labels; since we had to infer single truth labels from CREDBANK's 30-judgement, skewed sets to create binary labels, the threshold between what a crowdsourced worker would doubt versus what is factual may be inconsistent. We addressed this issue by discarding the middle 70% of the data and focusing only characterizing the extreme threads, but the skewed nature of CREDBANK assessments implies more variability exists in the negative samples than in the positives. In Castillo et al. [3], this skew was addressed by altering the labeling scheme to remove the "likely to be true" label, an approach future work could adopt to create a more balanced set. Our CREDBANK model's performance on the fake news data set at least partially validates the approach outlined herein.

Beyond construction, structural differences between CREDBANK and PHEME could also affect validity. If the underlying distributions that generated our samples are significantly different, differences in feature sets or cross-context performance could be attributed to structural issues rather than assessors. We checked this possibility using a Hotelling's T^2 test, and results showed the data sets were significantly different ($p < 0.05$), but in reviewing which features deviated between CREDBANK and PHEME, the majority of the top ten features in both models were not statistically divergent. In future work, this limitation could be addressed by constructing a single data set of potential rumors and fake news threads and using both crowdsourced and journalist assessors to evaluate the same data. This new data set would obviate any issues or biases introduced by the alignment procedure we employed herein.

7 CONCLUSIONS

This work compares accuracy assessments of conversations in Twitter from journalists in PHEME versus crowdsourced workers in CREDBANK. The goal is to determine whether these two populations are appropriate proxies for each other and whether they can inform the recently salient "fake news" identification task. After aligning these two data sets, we compare 45 new and prior work-inspired features that might predict accuracy labels within these data sets, compare the most important features from each model, and evaluate how well these models perform across contexts. Our findings show first that with just a few features, one can automatically predict journalist- or crowdsourced worker-labels more than 2/3 of the time. These small feature sets exhibit little overlap, however, with only two of the top ten features appearing in both models. In a cross-context evaluation of CREDBANK and PHEME models, neither model performs significantly better than random guessing in predicting the other, but the PHEME model does exhibit a statistically significant relationship to CREDBANK labels. These divergent models and this unidirectional link may stem from different interpretations of "accuracy" by journalists and crowdsourced workers, where CREDBANK labels capture credibility rather than factual truth with truth being one facet of credibility. We also show that models of CREDBANK credibility assessments significantly outperform PHEME models in identifying fake news stories shared on Twitter, which may result from how the underlying story's accuracy drives perceptions of credibility among non-journalist Twitter users.

This work is notable in its potential to identify authors who regularly share false, misleading, or non-credible rumors and stories on Twitter. Such a system could be valuable to social media users by augmenting and supporting their own credibility judgements, which would be a crucial boon given the known weaknesses users exhibit in these judgements. These results may also be of value in studying propaganda on social media to determine whether propaganda stories follow patterns similar to rumors and fake news, which is becoming increasingly important as users continue turning to social media for news and information.

8 ACKNOWLEDGEMENTS

This research was supported by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at the University of Maryland, College Park, administered by Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence.

9 DATA AVAILABILITY

Data and analyses are available online at: https://github.com/cbuntain/CREDBANK-data

REFERENCES

[1] Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A. Walker. 2015. Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. (2015), 4445–4452.

[2] Peter Bro and Filip Wallberg. 2014. Digital gatekeeping. *Digital Journalism* 2, 3 (2014), 446–454. DOI:http://dx.doi.org/10.1080/21670811.2014.895507

[3] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Research* 23, 5 (2013), 560–588. DOI:http://dx.doi.org/10.1108/IntR-05-2012-0095

[4] B J Fogg and Hsiang Tseng. 1999. The Elements of Computer Credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM, New York, NY, USA, 80–87. DOI:http://dx.doi.org/10.1145/302979.303001

[5] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. *Proceedings of the 22nd ...* (2013), 729–736. DOI:http://dx.doi.org/10.1145/2487788.2488033

[6] Byungkyu Kang, John O'Donovan, and Tobias Höllerer. 2012. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (IUI '12)*. ACM, New York, NY, USA, 179–188. DOI:http://dx.doi.org/10.1145/2166966.2166998

[7] Yeojin Kim and Wilson Lowrey. 2014. Who are Citizen Journalists in the Social Media Environment? *Digital Journalism* 0811, December 2014 (2014), 1–17. DOI:http://dx.doi.org/10.1080/21670811.2014.930245

[8] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. *Proceedings - IEEE International Conference on Data Mining, ICDM* (2013), 1103–1108. DOI:http://dx.doi.org/10.1109/ICDM.2013.61

[9] JeyHan Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line Trend Analysis with Topic Models: #twitter Trends Detection Topic Model Online. *International Conference on Computational Linguistics (COLING)* 2, December (2012), 1519–1534. https://www.aclweb.org/anthology/C/C12/C12-1093.pdf

[10] Jenn Burleson Mackay and Wilson Lowrey. 2011. The Credibility Divide: Reader Trust Of Online Newspapers And Blogs. *Journal of Media Sociology* 3, 1-4 (2011), 39–57.

[11] S. R. Maier. 2005. Accuracy Matters: A Cross-Market Assessment of Newspaper Error and Credibility. *Journalism & Mass Communication Quarterly* 82, 3 (2005), 533–551. DOI:http://dx.doi.org/10.1177/107769900508200304

[12] Amina A Memon, Aldert Vrij, and Ray Bull. 2003. *Psychology and law: Truthfulness, accuracy and credibility.* John Wiley & Sons.

[13] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter Under Crisis: Can We Trust What We RT?. In *Proceedings of the First Workshop on Social Media Analytics (SOMA '10)*. ACM, New York, NY, USA, 71–79. DOI: http://dx.doi.org/10.1145/1964858.1964869

[14] Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. *International AAAI Conference on Web and Social Media (ICWSM)* (2015). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10582

[15] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is Believing ? Understanding Microblog Credibility Perceptions. *CSCW '12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (2012), 441–450. DOI: http://dx.doi.org/10.1145/2145204.2145274

[16] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor Has It: Identifying Misinformation in Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1589–1599. http://dl.acm.org/citation.cfm?id=2145432.2145602

[17] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2010. Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. *CoRR* abs/1011.3 (2010). http://arxiv.org/abs/1011.3768

[18] Roser Sauri and James Pustejovsky. 2009. *Factbank: A corpus annotated with event factuality.* Vol. 43. 227–268 pages. DOI: http://dx.doi.org/10.1007/s10579-009-9089-9

[19] Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. 2016. Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate. (oct 2016). https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis

[20] Stanford History Education Group. 2016. *Evaluating information: The cornerstone of civic online reasoning.* Technical Report. Stanford University, Stanford, CA. 29 pages. https://sheg.stanford.edu/upload/V3LessonPlans/ExecutiveSummary11.21.16.pdf

[21] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M. Mason. 2014. Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. *iConference 2014 Proceedings* (2014), 654–662. DOI: http://dx.doi.org/10.9776/14308

[22] Laura Sydell. 2016. We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned. (nov 2016). http://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs

[23] Chenhao Tan, Cristian Danescu-niculescu mizil, Vlad Niculae, Cristian Danescu-niculescu mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 613–624. DOI: http://dx.doi.org/10.1145/2872427.2883081

[24] Craig Timberg. 2016. Russian propaganda effort helped spread fifake news' during election, experts say. (nov 2016). https://www.washingtonpost.com/business/economy/russian-propaganda-effort-helped-spread-fake-news-during-election-experts-say/2016/11/24/793903b6-8a40-4ca9-b712-716af66098fe

[25] Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, Rob Procter, and Peter Tolmie. 2015. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLoS ONE* (2015), 1–33. DOI: http://dx.doi.org/10.1371/journal.pone.0150989 arXiv:1511.07487