Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

Detecting Journalism in the Age of Social Media: Three Experiments in Classifying Journalists on Twitter

Li Zeng, Dharma Dailey, Owla Mohamed, Kate Starbird, Emma S. Spiro
University of Washington
Seattle, Washington, USA

Abstract

The widespread adoption of networked information and communications technologies (i.e. ICTs) blurs traditional boundaries between journalist and citizen. The role of the journalist is adapting to structural changes in the news industry and dynamic audience expectations. For researchers who seek to understand what, if any, distinct role journalists play in the production and propagation of breaking news, it is vital to be able to identify journalists in social media spaces. In many cases, this can be challenging due to the limited information and metadata about social media users. In this work, we use a supervised machine learning model to automatically distinguish journalists from non-journalists in social media spaces. Leveraging Twitter data collected from three crisis events of different types, we examine how profile information, social network structure, posting behavior and language distinguish journalists from others. Additionally, we evaluate how the performance of the journalist classification model varies by context (i.e. types of crisis events) and by journalism outlets (i.e. print versus broadcast journalism), and discuss challenges in automatic journalist detection. Implications of this work are discussed; in particular we argue for the value of such methods for scaling analysis in journalism studies beyond the capacity of human coders. Employing classification methods in this context allows for systematic, large-scale studies of the role of journalists online.

Introduction

In recent decades, the rapid development and widespread adoption of networked information and communication technologies has dramatically altered news production and consumption (Matheson 2004; Thurman 2008; Mitchelstein and Boczkowski 2009). For instance, Twitter, one of the most widely used social media platforms, provides an "ambient" news environment where ordinary people report on current events or provide new evidence during unfolding situations much like what professional journalists have done for decades (Hermida 2010). Yet journalists are also present in great numbers on the platform. In spite of many examples of citizen's successfully tackling contemporary news work (e.g. (Lee, Lancendorfer, and Lee 2005), (Flew and Wilson 2010)), it now appears that the aspiration that those "formerly known as the audience" (Rosen 2006) can fully take on the role traditionally played by journalists has stalled (Gillmor 2006; Bruns 2016). Placing the work of journalists within the larger social media crowd remains an ongoing project for journalism scholars.

Many studies focus on the practice of journalism on Twitter in order to understand how journalists adjust to structural changes in the news industry and dynamic audience expectations (Bagdouri 2016; Brems et al. 2017; Cammaerts and Couldry 2016). A vital component of these studies on digital journalism is the ability to distinguish journalists from non-journalists. In a context of changing behavior, changing structures, and changing institutions, the line between journalist and non-journalist is not always clear. Many scholars have struggled to single out journalists in these environments (Dahlgren 2016; Hanitzsch and Vos 2017). Large-scale computationally assisted analyses can add clarity to these blurred lines. Traditional methods, such as identifying journalists via manual coding, are restrictive – forcing researchers to spend significant resources to obtain relatively small study populations. Developing and testing methods for automatic classification, using provided metadata about social media users, has the potential to enhance studies in digital journalism. This paper aims to contribute to this goal, providing researchers with tools for identifying journalists in social media spaces.

The distinction between journalists and non-journalists is particularly important in crises. News organizations and those who work for them have historically played a vital role in getting timely information out to the public. In countries like the United States, journalists have been integrated into government response plans for disseminating information to the public for over a century. For example, broadcasters are required by the government to provide particular kinds of information during crises (Federal Communications Commission 2017). Understanding if legacy actors, practices, and institutions continue to play an outsized role in crisis communication is an important policy question. To answer it, we need an accurate picture of the interactions between citizens, government organizations, and journalists with respect to crisis communication.

The public increasingly utilize social media platforms to search for, consume and distribute information during crisis events (Chen and Sakamoto 2013; Vieweg et al. 2010; Starbird and Palen 2011). Indeed, social media are vital in-

formation sources during non-routine situations despite the fact that the accuracy of the information shared through these platforms is often unclear. Lacking professional gatekeepers to check content and traditional markers to determine source credibility (Brashers et al. 2000), information consumers on social media platforms are left to make decisions on information credibility themselves (Westerman, Spence, and Van Der Heide 2014). Previous work links perceived source credibility to the degrees to which a perceiver believes a sender to know the truth, to be willing to tell the truth, and have their best interests at heart (Cronkhite and Liska 1976). But to begin, it is crucial for social media consumers to be able to distinguish information sources (e.g. official emergency responders, journalists, members of the public, etc.) in order to evaluate credibility of information.

Our work is motivated by recognition of the significance of identifying journalists during crisis events. Moreover, journalists are likely to participate in reporting of events of interest, like crisis events (Kovach and Rosenstiel 2014) (giving us a study context where we can a priori expect to find many participants of interest). In this work, we classify Twitter accounts that participate in information-sharing about crisis events as either journalist or non-journalist. Our supervised machine learning methods enable automatic identification of journalists, which could save substantial human labor and cost. With an eye towards generalizability and potential for models to be applied across different contexts, rather than relying exclusively on Twitter posts in this task, our work looks at how users' profile, social networks, posting behavior and description reveal their identities.

We find that random forests models achieve the highest accuracy across all experimental settings and they significantly outperform other models in certain settings that heavily involve rumors. Further, we analyze the most predictive features to generate insights on potential cues for journalist identification. We show that *profile description* is most predictive of journalist accounts; we not only find that words used by journalists to describe their accounts vary from non-journalists, but also that *the number of user mentions in profile description* to be a good cue for journalists. Further, our work finds the *listed count* – an indicator of a "power user" – which captures the number of public lists that an account is a member of has greater predictive power than another more commonly used "power user" features such as the ratio of follower to friends in this task. These findings can contribute to future improvements in automatic journalist identification. We also provide empirical evidence on how particular characteristics vary by types of journalists, showing a future direction to classify journalists of different types based upon the results of this work.

The remainder of this paper is organized as follows. We start by reviewing existing studies on classifying social media user attributes and characterizing journalists on Twitter. Next, we describe the data to be used for the task of journalist classification in this study. We then present the methodology for this work, describing our machine learning experimentation with different conditions in order to evaluate the applicability and generalizability of models. Then we perform an empirical analysis looking closer at different types of journalists to examine difference between journalism outlets. We finally discuss and summarize our findings, pointing out limitations of the work and directions of future work.

Previous Work

Machine Learning for User Classification

A large body of work has been focused on how to infer users attributes or classify users into certain category according to their explicit or implicit characteristics. Previous work has looked at identifying user's age (Rao et al. 2010), gender (Fink, Kopecky, and Morawski 2012), ethnicity (Pennacchiotti and Popescu 2011), regional origin (Huang, Weber, and Vieweg 2014; Rao et al. 2010), political affiliation (Cohen and Ruths 2013; Pennacchiotti and Popescu 2011), interest in particular articles (Carreira et al. 2004), roles in a conversation (Tinati et al. 2012), and etc.

Machine learning methods for user classification make use of various features, but have focused primarily on user profile information, message textual content, and the social graph of relationships between users. For example, (Chu et al. 2010) demonstrated that the ratio of URLs in a tweet, the tweeting device makeup and the followers to friends ratio are predictive features to detect bots. (Carreira et al. 2004) used a Bayesian classifier and content-based filtering to determine users' interests in articles according to their profile information. (Pennacchiotti and Popescu 2011) also demonstrated the predictive power of user profile information along with linguistic content of user messages, user social network features and user tweeting behavior to infer political affiliation, ethnicity and affinity for a particular business of a user. Although these studies focus on classifying/inferring different users attributes than the work described presently, they offer insights into potential feature sets to be verified and enriched in our journalist classification work.

Classifying Journalists on Twitter

A few studies have attempted to systematically categorize journalists on Twitter. The work (Bagdouri and Oard 2015) and (Bagdouri 2016) use "seed" sets of pre-identified journalists combined with journalism keywords to identify additional journalists who share common characteristics with the initial set of journalists. This approach uses social network relationships and mentions of pre-identified journalists to identify potential journalists. Linguistic similarities including journalism keywords are then used to categorize journalists among the potential candidates. This approach has merit, for example, it may be helpful for identifying candidates for "white lists" of potentially credible news sources. However, this method may not help an average Twitter user to distinguish a journalism account from among the crowd because it requires not only an initial set of known journalists but also knowledge of the network of relationships among the seed set and others on Twitter. It may also have limited applicability, as in our case, where we wish to characterize the work of many kinds of journalism across a large social media data set. To use this approach we would have to start with a pre-identified set of journalists to which we could compare. Moreover, because it is based on existing

social network connections this approach is likely limited in reaching out to journalists who newly enter the online crowd and are without existing connections.

Work by (De Choudhury, Diakopoulos, and Naaman 2012) also systematically classifies journalists on Twitter. They make an important distinction between organizations and individuals within this categorization. They further refine the individual category by distinguishing between "journalists/media bloggers" and "ordinary individuals." These are helpful and meaningful distinctions, but as researchers we would like to distinguish between journalists and media bloggers as these monikers suggest different kinds of people may be performing the work traditionally taken on by journalists. Additionally, since this work was published (2012), there has been a rise in the prominence of news of questionable provenance; aspiring journalists, click-bait profiteers, and even disinformation actors become increasingly difficult to distinguish from traditional journalists. Thus, it is important and timely to revisit the issues of how journalists are classified. Importantly, none of the prior studies that take on journalism classification squarely address the issue of what journalism is, an issue we found ourselves recursively revisiting in our qualitative coding process.

Data

Crisis-related Twitter Collection

This research focuses on Twitter accounts that participated in information creation and dissemination during and about crisis events. These data reflect two different kinds of information that propagates in times of crises: rumors and non-rumors. While both data occur during crisis events, prior work has highlighted the unique role journalists may play in rumor spread (Kovach and Rosenstiel 2014). We construct two datasets - a Rumor Set and a Non-Ruomr Set, aiming to compare between cases where rumors were widespread and those in which they were not. The Rumor Set is derived from two distinct crisis events, the Sydney Siege event[1] and the Paris Attacks[2]. For these events, we relied on real-time data collection via the Twitter Streaming API using a set of event-related keywords and hashtags curated by researchers monitoring each event. Reviewing tweets, news coverage and other sources, researchers then identified event-related rumors. Additional rumors were identified post hoc through combined qualitative, quantitative and visual analyses (Maddock et al. 2015). Once rumors were identified, researchers iteratively refined search terms to generate a comprehensive, low noise corpus of tweets for each rumor. Next, each distinct tweet was read by at least three researchers. Using a "majority rule" decision criteria tweets deemed non-rumor related or uncodable were excluded. Researchers identified five false rumors that widely circulated. These comprise the Rumor Set used here. In all, the Rumor Set contains content from 12,614 distinct accounts and 15,225 accounts, respectively for these events. Additional methodological details and empirical analysis are provided in prior publications (Starbird et al. 2018).

The Non-Rumor Set comprises all users and tweets that employed either of two hashtags pertaining to the Oso Landslide[3]. These tweets are from a corpus purchased from Twitter (compared to data collected via the Streaming API). The same process of tweet selection and analyses described above was used to attempt to identify rumors in tweets with prominent hashtags associated with the slide: #530slide and #OsoStrong. However, substantive rumors were not identified among these tweets. The Non-Rumor Set consists of 78,409 tweets and 20,440 accounts.

Coding Procedure

Researchers pursued a grounded theory approach for categorizing journalists. Drawing on several previous rounds of inductive categorizing of Twitter accounts, five researchers established provisional rules based on previous literature and consultation with a former journalist. Over the course of one month of full time work, three undergraduates and one graduate student refined the coding scheme used here in regular consultation with another researcher who is also a former journalist. Categorization heuristics were modified based on issues encountered during coding. To determine ground truth for each journalism account in the training set, researchers reviewed available trace data for Twitter accounts including LinkedIn profiles, bylines on new sites and personal websites. As a result, researchers derived an agreed upon set of cues for identifying journalists using the account name, account description, and account statistics (Starbird et al. 2018). Table 1 includes the number of journalists and non-journalists categorized for each crisis-related dataset.

Applying the categorization schema developed, Twitter accounts were categorized as *journalist*, *non-journalist*, or *ambiguous*. Two coders independently performed the coding task, and a third researcher would arbitrate if the two coders were unable to determine the categorization or if there was disagreement. During the coding process, each coder reviewed the Twitter user's profile information as it appeared at the time of collection. This includes the user name, user description, number of posts, number of followers, number of following, URL, and geographic location. Coders assigned the journalist label to Twitter accounts if that the account self-associated itself with the professional community of practice as outlined by (Wenger 1998). This is inclusive of current and former journalists, student journalists and educators, news producing organizations, and professional journalism associations. In this way, we capture the range of journalistic activity visible on Twitter pertaining to crisis events. According to (Wenger 1998), membership in a community of practice is learned and earned by associating with other members of the community. By this definition, we determined that an "aspiring journalist" is not part of the

[1]The Sydney Siege event: a gunman held hostage ten customers and eight employees of a Lindt chocolate cafe at Martin Place in Sydney, Australia on 15-16 December 2014.

[2]The Paris Attacks: a series of coordinated terrorist attacks that occurred on Friday 13 November 2015 in Paris, France and the city's northern suburb, Saint-Denis.

[3]The Oso Landslide: a mass fatality landslide occurred 4 miles east of Oso, Washington, United States, on March 22, 2014

community of practice of journalism, but an individual who writes for a student newspaper is part.

Non-journalists are inclusive of any account that made no claim of journalistic affiliation as well as the many twitter accounts that claim to be news simply because they retweet news stories. Likewise, bloggers, vloggers and other media makers are "non-journalists" unless researchers could identify them in some manner as members of the journalistic community of practice. In a small number of cases, manual categorization was inconclusive. These cases were labeled "ambiguous" and are excluded from the analyses presented herein. In this category are accounts for which trace data ambiguous, insufficient to make a determination, or conflicted. For example, if an account claimed to be a journalist, but researchers did not deem the claim credibly associated with the community of practice of journalism, it was marked ambiguous.

Coders noted that it was easier to identify a journalist when the account associated itself with a journalistic organization. To explore this further, coders made additional annotations for journalists in the Non-Rumor Set. If the Twitter account belonged to a journalism organization, for example @CNN, it was coded as `organizational`. If the account belonged to an individual who is affiliated with an organization, coders labeled this account as `affiliated individual`. Individuals are affiliated with an organization if they indicated they worked at the organization. If the account belonged to a journalist who did not affiliate themselves with any organization, coders labeled the account `unaffiliated individual`.

Similarly, coders observed that journalist accounts were easier to identify qualitatively when accounts associated themselves with legacy media outlets. To explore this further, researchers annotated the type of affiliation journalists made in the Non-Rumor set: print or broadcast. Organizations and affiliated individuals were labeled as `print` if the news outlet identified itself with a newspaper, magazine, or other print publication. This is inclusive of accounts that were formerly print and now online only. Accounts were sub-coded as `broadcast` if they associated with radio or TV. There were some news outlets in the dataset that are neither print nor broadcast, and these were labeled as `other`.

Data Set	Num. of journalists	Num. of non-journalists	Num. of tweets
Non-Rumor Set	2237	18203	78409
Rumor Set	1535	11079	15225
Total	3772	29282	93634

Table 1: Journalism Category Counts and Tweet Counts

Methods

Feature Generation

Extending related work and building from cumulative cues derived in the coding process, we developed a rich set of features (as shown in Table 2) designed to capture the following aspects of a Twitter account: (1) completeness of user profile; (2) power user status; (3) user posting behavior, and (4) language used in self-description. We evaluate the extent to which such features are predictive of a particular user account on Twitter belonging to a journalist. Moreover, we explore how predictive features are independent of event context so that models can be applied to identify journalists among new unseen accounts and in a general contexts. We detail the motivation and construction of model features for each of these four concepts below.

First, we consider the completeness of a user's profile. That is, how much information does a user provide in their public profile? User profiles often serve as the primary outlet for presentation of self on social media platforms. Prior work has studied how journalists brand themselves or their affiliated organizations through textual and visual profile content (Hanusch and Bruns 2017; Lough, Molyneux, and Holton 2017). Our empirical data analysis suggests that journalists tend to associate themselves with a journalism outlet by including an associated URL in the profile (Starbird et al. 2018).They are less likely to leave a user description empty and to use a default profile image. Moreover, the verified status has been found to be a good indicator of journalists as well as other "power users" (Bagdouri 2016; De Choudhury, Diakopoulos, and Naaman 2012). Therefore, we extract the following features: whether an account profile provides a URL and user description, uses a non-default profile image and whether it is a verified account.

Next, we look at the "power user" characteristics of an account. Prior work suggests that being a journalist is correlated with high power user status in terms of *the number of followers* and *ratio of followers to friends* (e.g. (Starbird et al. 2018), (Zubiaga et al. 2016)). These two measures could be predictive of journalist classification, but it is also necessary to go beyond simple audience measures. Two other potential indicators of power user status we consider are *listed count* (i.e. the number of public lists that an account is a member of) and *favorites count* (i.e. number of tweets that other Twitter user likes).

We also capture posting behaviors of Twitter accounts. Previous studies find that journalists are professional tweeters in terms of post volume (e.g. (Starbird et al. 2018), (Bagdouri 2016)). We consider average *number of daily tweets* which is calculated by the total number of tweets posted by account divided by the length of account tenure on Twitter. Moreover, we would expect that journalists engage earlier in event-related communication after an event occurred. Therefore, we use *waiting time for the first tweet* to measure the number of hours elapsed after an event occurred until an account engages in tweeting. We also consider *ratio of retweets and direct replies* posted by accounts as journalists and non-journalists may adopt different communication strategy and target. Lastly, we include *average number of user mentions, hashtags and URLs* per tweet by accounts to characterize the behavior of incorporating Twitter features and external sources in tweets.

Finally, we extract textual-base features from the user description to capture differences in language by ac-

counts. We include the number of user mentions, hashtags and URLs listed in the user description. We then extract bag of words features, after removal of stop words and word stemming. We also tag part-of-speech in favor of a syntactic language analysis using GATE Twitter part-of-speech tagger (Derczynski et al. 2013). Based on the parsed part-of-speech, we recognize number of named `Person`, `Location`, `Organization`, `Numeric` and `Temporal` entities present in the description text using Stanford CoreNLP.

Machine Learning Experiment Design

Our goal in the machine learning experiment is to build a reliable and accurate classifier to categorize Twitter accounts as either a journalist's account or not, given Twitter profile information, power user status, posting behaviors and language used in the user description. Detailed descriptions of these feature are presented in Table 2.

Our consider three settings of journalist classification that aim to answer different research questions. In the first setting, **single set prediction**, we control for the context of journalist classification in terms of misinformation circulation. We train models on the Rumor Set and the Non-Rumor set, respectively. The comparison of model performance for these two different contexts can help us understand how a rumor-related context will affect journalist classification.

In the second setting, **pooled set prediction**, we merge together the two data sets to construct a pooled set to train models. By this design, context of specific rumors and events might become blurred and classifying journalists might become more challenging. If a classifier can retain satisfactory performance in this setting, it suggests generalizability to some extent.

Lastly, we design a **cross set prediction** of journalists as the third setting. In the first round of cross set prediction, we select one set (i.e. the Rumor Set or the Non-Rumor set) as the training set and test fitted models using the other set. In the next round, we swap the training set and test set to repeat the model training and evaluation. We average the model performance from the two rounds. The design of this experiment poses the greatest challenges to model generalizability among the three settings. The motivation behind the design is the recognition that manual coding of social media content can be very expensive. Therefore, such models will be extremely valuable for scaling analysis and automatically categorizing accounts that appear in new unseen events.

Machine Learning Experiment Setup

In the settings of single set prediction and pooled set prediction described in the previous section, we construct a training set which contains 80% of the entire data in a given setting and a test set which is the remaining 20% of the data. For cross set prediction, we perform two rounds of machine learning experiments. First, we choose one set from either the Rumor Set or the Non-Rumor Set as the training set; the remaining set becomes the test set. Then in the next round we swap the two sets. Training sets in all settings are used for model tuning and training, and we "lock" the test sets in all experiments until the final model evaluation process.

To tune model parameters, we apply five-fold cross validation with the training data. Cross validation divides the training set into five folds of equal size and iterates over each. For each iteration, a model with a set of candidate parameters is trained on the tuning set which contains four folds of the training set and then evaluate on the validation set which is the remaining fold of the training set. Once we determine the best-tuned model, we train the model on the entire training set. Finally, we evaluate and report the performance of the models of each algorithm using the "held-out" test set.

When it comes to the choice of algorithms for journalist classification, we want to balance both model complexity and interpretability. There are many possible models for performing this classification task; as such we trained three types of classification algorithms: Logistic Regression (LR), k Nearest Neighbors (kNN) and Random Forests (RF). Logistic Regression has been widely used to predict categorical outcomes. Even though LR takes account only of a linear relationship, it has proved to be very useful in its simplicity to train as well as for the high interpretability of results. The previous work by (De Choudhury, Diakopoulos, and Naaman 2012) found that kNN was the best performing classifier in their work of categorizing Twitter users. kNN learns a non-linear classifier, however its computational cost can increase dramatically for a large-scale data set. RF is an ensemble classifier that consists of a series of decision trees. In practice, the RF algorithm runs efficiently on large data sets and produces highly accurate classifiers (Breiman 2001; Hurtik, Burda, and Perfilieva 2013; Zeng, Starbird, and Spiro 2016). We consider these set of models in order to identify which performs best in this setting, not only in terms of predictive ability but also applicability and interpretability in real world contexts. While determining which model performs best is a component of the work, it is not in of itself the aim. Our goals is instead to demonstrate the value of these methods and offer an explanation of the features that might distinguish journalists from non-journalists.

Model Evaluation

In each setting, we evaluate the performance of the three different models used in this work - the k-Nearest Neighbors, Logistic Regression and Random Forests. First, we compare based on the *accuracy* which is a ratio of correctly predicted instances to the total instances in the test set, as well as the F_1 score which is the harmonic average of precision and recall.

Additionally, we perform approximate statistical tests for determining whether one machine learning algorithm outperforms another in a particular experimental setting. While several statistical tests are available for this purpose, the McNemar's test and the 5×2 cross-validation test are recommended and widely used due to their low type I errors (Dietterich 1998). Our cross set prediction setting does not support the process of the 5×2 cross-validation test which repeats randomly even splitting (50% training and 50% test data) five times. Therefore, we perform the McNemar's test (McNemar 1947) which is computationally efficient and has

Category	Feature
Profile Completeness	Verified account, presence of user url, presence of user description, presence of user location, default profile image
Power User	Number of followers, ratio of followers to friends, number of lists and favorites
Posting Behavior	Number of daily tweets, ratio of retweets and direct replies, average number of user mentions, hashtags and URLs, waiting time for the first tweet
Description Language	number of user mentions, hashtags, URLs, part of speech tags and named entities in user description

Table 2: Feature categories and features generated for journalist classification

acceptable type I error to compare classifiers.

The McNemar's test, sometimes also called "within-subjects chi-squared test", is a paired nonparametric or distribution-free statistical hypothesis test. In the context of comparing two binary machine learning models, we can use the McNemar's test to compare whether two models disagree in the same way. It is important to point out that this test cannot determine which model predicts more or less accurately than another. However, when predictive accuracy of two models are close to each other, this test can be particularly helpful to answer the question: is better performance due to a statistically significant difference in the models or just due to statistical chance? The null and alternative hypotheses of the McNemar's test are stated as follows:

$$H_0 : \text{two models disagree to the same amount}$$

$$H_1 : \text{two models disagree in different ways}$$

The continuity corrected version of the McNemar's test statistic which is the more commonly used today (Edwards 1948) can be computed as follows:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$$

where b is the number of instances in the test set when the first model classified correctly but the second model classified incorrectly, and c is the count when the first model misclassified while the second model correctly classified.

Results

Classifying Journalists

Table 3 shows the accuracy and F_1 scores of the kNN, LR and RF algorithms in the three different machine learning experimental settings. Receiver operating characteristic (ROC) curves for each of the three models are shown in Figure 1. There exists an imbalance between the number of journalists and non-journalists in both the Rumor Set and Non-Rumor Set; in this case if machine learning models only predict the majority class, the accuracy will still be quite high, despite very poor performance on the minority class. Therefore, we compare performance against a baseline equal to the proportion majority class - non-journalists in each setting. One wants to see accuracy of the trained models higher than this baseline. To the best of the authors' knowledge, the highest accuracy of classifying journalists is 92.41% as reported by (De Choudhury, Diakopoulos, and Naaman 2012). We also want to compare our results to this prior work.

We learn from these assessment measures that all three algorithms in each setting are able to outperform the majority class baseline. The LR and RF models are also able to beat the "state-of-art" accuracy of 92.41% across all experimental settings. We see that the RF models remain the most accurate with the highest accuracy and F_1 scores across all experimental settings, while the kNN models have the lowest accuracy and F_1 scores in all cases. As the RF classifiers only slightly outperform the LR classifiers, it is necessary to conduct statistical significance tests to evaluate statistical confidence in the differences in model performance.

Table 4 shows the results of McNemar's tests comparing between each pair of the three models in all experimental settings. With a significance level of $\alpha = 0.05$, we see significant differences in the disagreement between the kNN and the LR models as well as between the kNN and the RF models across all experimental settings. Further, the LR and the RF models have statistically significant differences in performance when training and testing on the Rumor Set in the setting of the cross-set prediction that trains (or tests) models on the Rumor Set. Interestingly, we do not find statistically significant differences in model performance between the two models trained and tested on the Non-Rumor Set and the Pooled Set. While the pooled set includes the Rumor Set, it might be dominated by the Non-Rumor Set which is larger in size than the Rumor Set. Comparisons between model performance may indicate that simple classification methods like the LR can be used and are able to achieve fairly good performance. Once information environments in which we aim to identify journalists get more complex such as during the case of dynamic rumor propagation, a more flexible model (e.g. the RF) reveals advantages over a simple classification method. This is an interesting direction for future work.

As we discussed in the section Machine Learning Experimental Design, hypothetically the challenge of accurate journalist classification levels up as the experimental setting moves from single set prediction to pooled set prediction,

and finally to cross set prediction. Given the same type of algorithm, model performance drops slightly comparing the single set prediction on the non-rumor set to the pooled set or the pooled set prediction to the cross set prediction. This indicates that the trained models achieve a reasonable level of generalizability and the applicability to be used for journalist classification in a new, unseen event context. We also notice that the performance of models increases from the single set prediction on the Rumor Set to the pooled set. This could be due to data size, the Rumor Set is relatively small and therefore limits the amount of valuable information fed into the single set prediction models for journalist classification.

Data Set	Baseline	Model	Acc.	F_1
Rumor Set	0.878	kNN	0.898	0.883
		LR	0.949	0.946
		RF	0.958	0.955
Non-Rumor Set	0.896	kNN	0.915	0.901
		LR	0.968	0.966
		RF	0.970	0.968
Pooled Set	0.888	kNN	0.909	0.896
		LR	0.963	0.961
		RF	0.965	0.962
Cross Set Prediction	0.888	kNN	0.902	0.890
		LR	0.956	0.930
		RF	0.960	0.958

Table 3: Results of journalist classification. Models for the Rumor Set, the Non-Rumor Set, the Pooled Set and the Cross Set Prediction are shown, along with model accuracy and F_1-scores.

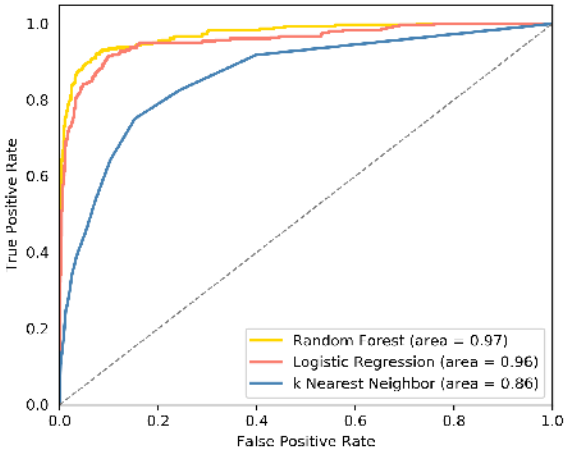

Figure 1: Receiver operating characteristic curves for the k Neareast Neighbor, Logistic Regression and Random Forest Models classifying journalists in the pooled set prediction.

Analyzing Feature Importance

Table 5 shows the top 10 important features for the random forest models built in the single set prediction along with the pooled set prediction. Most of them are bag of words (BoW) features extracted from the user description, indicating that language used in the user description provides cues for journalist classification on Twitter. Unsurprisingly, journalistic terms such as report, journalist, news, guideline are most predictive of journalists. For the single set prediction, we found that the top bag-of-words features in the Rumor Set are distinct from those in the Non-Rumor Set, suggesting journalists who participated in rumor specific or non-rumor specific communication may differ in terms of vocabulary of self-description. In addition to bag of words, number of user mentions included in user description are found to be predictive across the three cases.

Interestingly, we find that listed count is more predictive than other measures of power user status such as number of followers and ratio of followers to friends as suggested in the previous work (Starbird et al. 2018). This empirical finding suggests that listed count is a good candidate for s power user measure. We also observe that verified status is helpful for identifying journalists supporting results from prior studies (e.g. (Bagdouri 2016; Hanusch and Bruns 2017)).

Examining Types of Journalists

Next, we take a closer look at further distinctions between journalists in the non-rumor set: (1) print versus broadcast journalism and (2) organizational versus individual journalism. Note that these two sub-code categories are not mutually exclusive. We examine the capacity of the RF model trained previously to assign the correct *journalist* label to various types of journalists. In this analysis we ask if predicting whether a particular Twitter account belong to a journalist is easier for some types of journalists, compared to others.

Table 6 shows the proportion of misclassifications by the RF model per sub-code category. We only have one account coded as unaffiliated individual journalist in the broadcast outlet. Therefore, we avoid drawing any conclusions on this case. Among the rest, we observe that the classifier finds it challenging to correctly classify unaffiliated individual journalists in the broadcast outlet (i.e. B-U) with the highest error rate = 0.167, but tends to be more accurate when classifying affiliated individual journalists in the broadcast outlet (i.e. B-A) with the lowest error rate = 0.041.

We move on to analyze how power user status, profile and language features vary by type of journalists. Figure 2 shows the distributions of features included in the power user category by types of journalist along with the group of non-journalist (N-J) to serve as a baseline. Again, the only one instance of a print, unaffiliated individual journalist is not included in the following analysis. A glance at Figure 2 tells one the distributions of the three power user measures vary by journalist type. The baseline of non-journalist is "far away" from journalists of all types. We also perform t-tests to confirm that the mean difference of the feature of interest between categories is statistically significant. First, we

	kNN v.s. LR	kNN v.s. RF	LR v.s. RF
Rumor Set	< 0.001***	< 0.001***	0.008**
Non-Rumor Set	< 0.001***	< 0.001***	0.402
Pooled Set	< 0.001***	< 0.001***	0.456
Cross Set Prediction	< 0.001***	< 0.001***	0.038*

Table 4: P-values of the McNemar's hypothesis tests for model performance with a significance level of $\alpha = 0.05$. Failure to reject the null hypothesis ($p \geq \alpha$) indicates that the two classifiers have a similar proportion of errors on the test set, while rejection of null hypothesis ($p < \alpha$) indicates that the two classifiers have a different proportion of errors on the test set.

Rank	False Rumor Set	Non-rumor Set	Pooled Set
1	guidelin (BoW)	report (BoW)	journalist (BoW)
2	verified (Profile)	news (BoW)	report(BoW)
3	listed (Power User)	journalist (BoW)	news (BoW)
4	newspap (BoW)	listed (Power User)	listed (Power User)
5	follower/friends (Power User)	verified (Profile)	verified (Profile)
6	# of mentions in udescr	editor (BoW)	editor (BoW)
7	dc (BoW)	anchor (BoW)	produc(BoW)
8	followers (Power User)	produc (BoW)	anchor(BoW)
9	mix (BoW)	# of mentions in udescr	# of mentions in udescr
10	guardian (BoW)	follower/friends(Power User)	follower/friends (Power User)

Table 5: Feature importance of the random forest classifier in single set prediction and pooled set prediction.

	Affiliated Individual	Unaffiliated Individual	Organizational
Print	0.048	0	0.11
Broadcast	**0.041**	**0.167**	0.063

Table 6: Error Rate of Journalist Classification by Types of Outlets

find that given the same type of print or broadcast, organizational journalists have significantly larger number of listed count ($p < 0.001$), ratio of followers to friends ($p < 0.001$) and follower count ($p < 0.001$) than unaffiliated individual or affiliated individual journalists. Second, we compare the difference between unaffiliated individuals and affiliated individuals. We find that affiliated individual journalists in the broadcast outlet have significantly higher listed count ($p < 0.05$) and follower count ($p < 0.05$) than broadcast, unaffiliated individual journalists.

Next we examine how profile features vary by journalism outlets. Among the five features (see the section Feature Generation for more detail), we find that the proportion verified for different types of journalists differ (Figure 3). B-O enjoys the largest proportion of verified accounts while B-U has the smallest proportion. This may suggest that organizational accounts tend to be easier to determine as being an account of public interest. Moreover, we observe a difference between print and broadcast medium where journalists associated a print outlet have overall a low proportion of verified accounts.

Finally, we look at how types of journalists may differ in the language present in their user description. Empirical evidence of language differences might suggest a direction for model improvement. We perform an exploratory data analysis focusing on the most challenging case, B-U, and the least challenging case in terms of error rate from the RF classifier, as we would expect a larger language difference between these two types than others.

Number of user mentions present in the user description is one of the top 10 features from the trained model. Therefore, we conducted a t-test on the mean of user mentions for the two journalist types. The result shows that broadcast, unaffiliated individual journalists include significantly more user mentions in their user description than broadcast, unaffiliated individual journalists ($p < 0.001$). Moreover, Figure 4 shows the average count of named entities used in user description for B-A, B-U and N-J. Incorporating with the results of t-tests, we find that B-A uses significantly more named location entities ($p < 0.01$) and less named numeric entities ($p < 0.001$) than B-U journalists in their user description. Note we do not observe a consistent pattern of using named entities by different types of journalists. Take, for example, the named location entity. B-A significantly uses more location entities than non-jounalists ($p < 0.001$), however, this is not the case for the B-U. This also suggests the value of potentially separating journalist types for identifying journalists among the crowd - a study we would like to focus on in the future.

Discussion

Automated classification of journalists on social media has many advantages; it also offers pathways for future research. Part of the motivation for this work was the recognition that manual coding of social media accounts can be very expensive, prohibiting researchers from analyzing complete

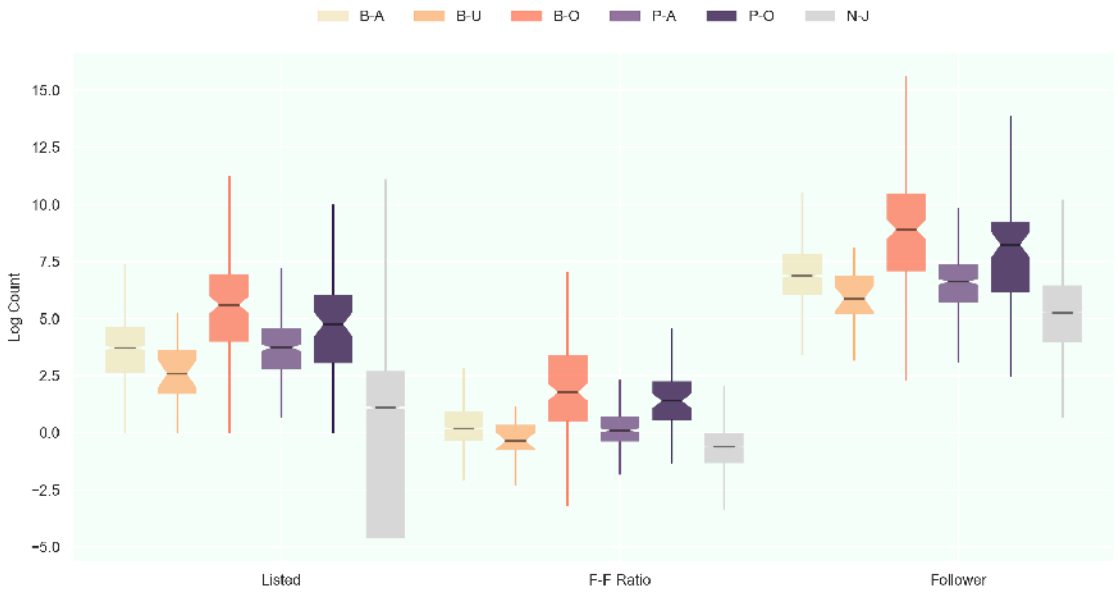

Figure 2: Power user status varies by types of outlets (the group of non-journalists included as baseline).
B-A: Broadcast - Affiliated Individual; B-U: Broadcast - Unaffiliated Individual; B-O: Broadcast - Organizational;
P-A: Print - Affiliated Individual; P-O: Print - Organizational; N-J: Non-Journalist

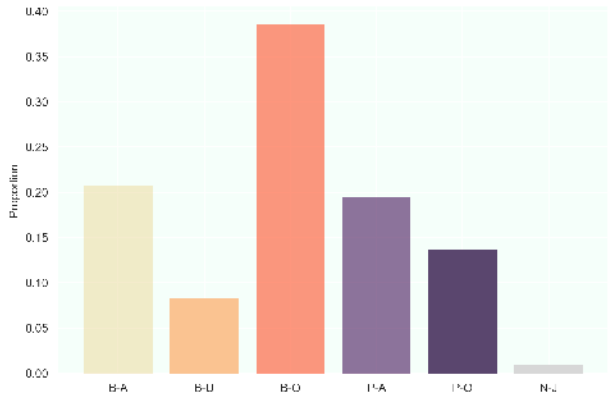

Figure 3: Proportion of verified account for each journalist category in our data (the group of non-journalists included as baseline).

Figure 4: Average count of named entities included in user description varies by types of journalists (the group of non-journalists included as baseline).

data and limiting the generalizability of research findings. Despite these limitations, there are great opportunities in work that augments manual coding tasks with automated classification. If human coders can categorize a small sample of collected data, a learning classifier is able to build from this foundation and finish the coding task. These hybrid approaches could save considerable resources, including time, effort and monetary cost associated with human coding. Moreover, careful interpretation of automated classification could cast insights into discovering previously hidden cues for journalist identification.

In the work presented herein, we focus the analysis of language use on Twitter profile description. Findings show that user descriptions that contain journalistic terms (eg. report, news, editor, anchor) and user mentions (eg. associated organization) are more likely to belong to journalists. Previous work on online journalist identification mostly relies on language in Twitter users' post (De Choudhury, Diakopou-

los, and Naaman 2012; Dahlgren 2016) or some simplistic features that can be directly extracted from the user description. This work suggests the potential to assist with journalist classification by utilizing syntactic and semantic features derived from a fairly short description text. Moreover, journalists are a type of power user on Twitter. In our trained machine learning model, listed count ranks at the top of power user feature categories. Therefore, our findings not only support follower count and ratio of follower to friend being good measures of power users, but also provide empirical evidence on how listed count can be a more powerful indicator of power users.

This analysis also provides empirical evidence on how particular features vary by type of journalists, which in turn points to future directions in classifing journalists. For the machine learning models trained in this work, it is easier to identify affiliated individual journalists than unaffiliated individuals. This may suggest that affiliated individual journalists may make unambiguous claims as to their journalist identity. In future work, we can view these results as baselines to build upon, particularly with respect to more extensive feature engineering in favor of identifying more challenging or ambiguous journalist types. Overall, in this task of binary journalist classification, machine learning models are able to achieve good predictive performance across all experimental settings, exhibiting a level of model generalizability in the crisis context as well as pointing to promising directions for future work on classifying journalists of different types and applying journalist classification to other contexts.

Limitations and Future Work

This work seeks to identify journalists on the Twitter platform. The study is motivated by recognition of the importance of identifying journalists during crisis events and the increased use of social media platforms for news distribution during crises. The ability to identify journalists enables a better understanding of the role that journalists play in crises, which in turn has important implications on emergency response and management. Therefore, we aim to develop machine learning models to automatically identify journalists utilizing data consisting of Twitter accounts who participated in three distinct crisis events - the Sydney Siege event, the Paris Attacks and the Oso Landslide. While this analysis is cross hazard event, one limitation concerns how the trained models would apply to other non-crisis events. Importantly we choose not to include crisis-specific or event-specific features in training the models. It is therefore possible for the models perform well in other events; this is a question left for future work. In addition, a follow up study could further evaluate model performance, comparing journalists who participated in crisis communication versus routine communication.

Future work can also extend the cases considered, while still focus on crisis-related communication. As with many studies of rumoring, the data available for this study are comprised of communication related to false rumors or misinformation. A consequence of this is that true rumors were excluded (or unavailable). It is possible that journalists who participate in true rumor reporting are distinct. Perhaps journalists who only post after verifying information can be found. Further, expanding the Non-Rumor Set as well could diversify the cases drawn from in identifying journalists; this might allow for geographic level variation for example.

This research pursued a grounded theory approach to define journalists. While this approach has many advantages, including the rich articulation of users in relation to the community of practice of journalism, it nonetheless relies on data sampled at a particular point in time. The validity and reliability of the machine learning models developed rely upon the definition constructed at this particular time and therefore may be subject to shifting definitions and roles of journalists in informal online communication over time. Additionally, features extracted in this work are static; yet the changes made in users' profile information might be illuminating information to aid in journalist classification as well. For example, journalists may update their user description fields to note where they are currently located to communicate what regions they are reporting on.

Despite these limitations, the work here offers new findings about the cues that users can identify and use to find journalists on social media. It also offer many avenues for future work – both empirical and methodological.

Conclusion

Our work develops a supervised machine learning approach to automatically classify Twitter accounts as journalist or not. Utilizing crisis-related Twitter collections, our models are able to accurately identify journalists across all experimental settings, suggesting generalizability of the trained models. Analyses of feature importance provided interesting insights. First, we find that journalists and non-journalists differ by language used in user description. Such features suggest potentially useful cues for journalist identification. Second, we show that power user status is a good indicator of journalists. Among all proposed power user measures in this work, we find that the listed count is most predictive.

References

Bagdouri, M., and Oard, D. W. 2015. Profession-based person search in microblogs: Using seed sets to find journalists. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 593–602. ACM.

Bagdouri, M. 2016. Journalists and twitter: A multidimensional quantitative description of usage patterns. In *ICWSM*, 22–31.

Brashers, D. E.; Neidig, J. L.; Haas, S. M.; Dobbs, L. K.; Cardillo, L. W.; and Russell, J. A. 2000. Communication in the management of uncertainty: The case of persons living with hiv or aids. *Communications Monographs* 67(1):63–84.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Brems, C.; Temmerman, M.; Graham, T.; and Broersma, M. 2017. Personal branding on twitter: How employed and free-

lance journalists stage themselves on social media. *Digital Journalism* 5(4):443–459.

Bruns, A. 2016. 'random acts of journalism'redux: News and social media. In *News Across Media: Production, Distribution and Consumptio*. Routledge. 42–57.

Cammaerts, B., and Couldry, N. 2016. *Digital journalism as practice*. SAGE Los Angeles, CA.

Carreira, R.; Crato, J. M.; Gonçalves, D.; and Jorge, J. A. 2004. Evaluating adaptive user profiles for news classification. In *Proceedings of the 9th international conference on Intelligent user interfaces*, 206–212. ACM.

Chen, R., and Sakamoto, Y. 2013. Perspective matters: Sharing of crisis information in social media. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, 2033–2041. IEEE.

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2010. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, 21–30. ACM.

Cohen, R., and Ruths, D. 2013. Classifying political orientation on twitter: It's not easy! In *ICWSM*.

Cronkhite, G., and Liska, J. 1976. A critique of factor analytic approaches to the study of credibility. *Communications Monographs* 43(2):91–107.

Dahlgren, P. 2016. Professional and citizen journalism: Tensions and complements. *The crisis of journalism reconsidered* 247–262.

De Choudhury, M.; Diakopoulos, N.; and Naaman, M. 2012. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 241–244. ACM.

Derczynski, L.; Ritter, A.; Clark, S.; and Bontcheva, K. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics.

Dietterich, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10(7):1895–1923.

Edwards, A. L. 1948. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika* 13(3):185–187.

Federal Communications Commission. 2017. Emergency alert system (eas).

Fink, C.; Kopecky, J.; and Morawski, M. 2012. Inferring gender from the content of tweets: A region specific example. In *ICWSM*.

Flew, T., and Wilson, J. 2010. Journalism as social networking: The australian youdecide project and the 2007 federal election. *Journalism* 11(2):131–147.

Gillmor, D. 2006. *We the media: Grassroots journalism by the people, for the people*. " O'Reilly Media, Inc.".

Hanitzsch, T., and Vos, T. P. 2017. Journalistic roles and the struggle over institutional identity: The discursive constitution of journalism. *Communication Theory* 27(2):115–135.

Hanusch, F., and Bruns, A. 2017. Journalistic branding on twitter: A representative study of australian journalists' profile descriptions. *Digital Journalism* 5(1):26–43.

Hermida, A. 2010. Twittering the news: The emergence of ambient journalism. *Journalism practice* 4(3):297–308.

Huang, W.; Weber, I.; and Vieweg, S. 2014. Inferring nationalities of twitter users and studying inter-national linking. In *Proceedings of the 25th ACM conference on Hypertext and social media*, 237–242. ACM.

Hurtik, P.; Burda, M.; and Perfilieva, I. 2013. An image recognition approach to classification of jewelry stone defects. In *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint*, 727–732. IEEE.

Kovach, B., and Rosenstiel, T. 2014. *The elements of journalism: What newspeople should know and the public should expect*. Three Rivers Press (CA).

Lee, B.; Lancendorfer, K. M.; and Lee, K. J. 2005. Agenda-setting and the internet: The intermedia influence of internet bulletin boards on newspaper coverage of the 2000 general election in south korea. *Asian Journal of Communication* 15(1):57–71.

Lough, K.; Molyneux, L.; and Holton, A. E. 2017. A clearer picture: Journalistic identity practices in words and images on twitter. *Journalism Practice* 1–15.

Maddock, J.; Starbird, K.; Al-Hassani, H. J.; Sandoval, D. E.; Orand, M.; and Mason, R. M. 2015. Characterizing online rumoring behavior using multi-dimensional signatures. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 228–241. ACM.

Matheson, D. 2004. Weblogs and the epistemology of the news: Some trends in online journalism. *New media & society* 6(4):443–468.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.

Mitchelstein, E., and Boczkowski, P. J. 2009. Between tradition and change: A review of recent research on online news production. *Journalism* 10(5):562–586.

Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to twitter user classification. *Icwsm* 11(1):281–288.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 37–44. ACM.

Rosen, J. 2006. The people formerly known as the audience. PressThink: 2006-06-27.

Starbird, K., and Palen, L. 2011. Voluntweeters: Self-organizing by digital volunteers in times of crisis. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1071–1080. ACM.

Starbird, K.; Dailey, D.; Mohamed, O.; Lee, G.; and Spiro, E. S. 2018. Engage early, correct more: How journalists participate in false rumors online during crisis events. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 105. ACM.

Thurman, N. 2008. Forums for citizen journalists? adoption of user generated content initiatives by online news media. *New media & society* 10(1):139–157.

Tinati, R.; Carr, L.; Hall, W.; and Bentwood, J. 2012. Identifying communicator roles in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, 1161–1168. ACM.

Vieweg, S.; Hughes, A. L.; Starbird, K.; and Palen, L. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1079–1088. ACM.

Wenger, E. 1998. *Communities of practice: Learning, meaning, and identity*. Cambridge university press.

Westerman, D.; Spence, P. R.; and Van Der Heide, B. 2014. Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication* 19(2):171–183.

Zeng, L.; Starbird, K.; and Spiro, E. S. 2016. # unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Tenth International AAAI Conference on Web and Social Media*.

Zubiaga, A.; Liakata, M.; Procter, R.; Hoi, G. W. S.; and Tolmie, P. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11(3):e0150989.