Evaluating Personal Metadata and Credibility Perception in Social Media Streams

BYUNGKYU KANG, University of California, Santa Barbara
TOBIAS HÖLLERER, University of California, Santa Barbara
JOHN O'DONOVAN, University of California, Santa Barbara

Today's Internet users are increasingly relying on user provided content from social media streams for news updates. In many cases, there is a limited window of information upon which an information consumer can make a credibility assessment about a piece of information or its source. To better understand the dynamics of credibility assessment in social media streams, we describe series of experiments to identify and evaluate key factors that influence credibility perception in microblogs. Specifically, we describe a user experiment (N=646) that builds on a previous study of credibility perception in order to answer the following research questions: (1) What are the important cues that contribute to information being perceived as credible. (2) Can we separate these cues from the content and quantify their influence on credibility perception?, and (3) To what extent is such a quantification portable across different microblogging platforms? To answer the third question, we study two popular microblogs, *Reddit* and *Twitter*. Key results include that significant effects of individual factors can be isolated, are partially portable, and that links, profile pictures and image content are the strongest influencing factors in credibility assessments.

Categories and Subject Descriptors: H.1.2 [**Information Systems**]: User/Machine Systems; H.5.2 [**Information Interfaces and Presentation (HCI)**]: User Interfaces

General Terms: Algorithms, Experimentation, Human Factors, Measurement

Additional Key Words and Phrases: Credibility, Microblog, Social Network, Perception, Interaction, User Experiment, User Study, Survey

ACM Reference Format:
Byungkyu Kang, Tobias Höllerer, and John O'Donovan, 2015. Evaluating Personal Metadata and Credibility Perception in Social Media Streams. *ACM Trans. Interact. Intell. Syst.* 9, 4, Article 00 (March 2010), 22 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Microblogging platforms such as Twitter and Reddit are increasingly relied upon for real-time news and a broad range of other information. However, all platforms that support user-provided content, including microblogs contain a large amount of noisy and unreliable content. Microblogs have emerged as serious sources for news at a global level, as indicators and early informers of natural phenomena such as earthquakes and severe weather, and even as mechanisms for financial transactions. For

This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053; The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.
Author's addresses: B. Kang and T. Höllerer and J. O'Donovan, Department of Computer Science, University of California, Santa Barbara, California, 93106, U.S.A.; email:bkang@cs.ucsb.edu.

DOI: http://dx.doi.org/10.1145/0000000.0000000

instance, tweet-to-pay functionality was recently introduced by several banks[1]. Accordingly, the importance of identifying credible information and information sources on microblog platforms continuously increases.

History. Traditionally used as online journals or specialized peer-communication platforms [Java et al. 2007], microblogs have transformed and proliferated into powerful online information sources operating at a global scale in every aspect of society. This is due in part to the advance of mobile technologies, including audio and video streaming, speech recognition and others. These technologies enable people on-location at an event or incident to serve as news reporters [Starbird and Stamberger 2010]. Several recent studies [Mendoza et al. 2010; Burns and Eltham 2009; Vieweg et al. 2010] show how user-provided microblog content is an effective mechanism for understanding crisis situations such as earthquakes, hurricanes or political conflicts. In fact, a recent study of traditional media journalist practice [Lasorsa et al. 2012] shows that they rely heavily on social media for their information. Another study [Willnat and Weaver 2014] reported that 53.8% of all U.S. journalists use microblogs to collect information and to report their stories, which of course raises the potential issues of cyclic dependency and rapid propagation of misinformation. These are issues which bolster the need for better approaches to information credibility on the web.

Evolution. The mass proliferation of microblog usage also brought about a shift in the interaction mechanisms and information flow within the platforms themselves. In particular, a 2013 PEW research report [Holcomb et al. 2013] shows that an increasing number of users search microblogs by keyword or hashtag as opposed to the traditional content stream or message exchange practices. This means that a larger portion of information is coming from complete strangers, accessed via keyword matching than from sources that a user is actively following and are likely to be known by the user. This reduced window of information about the source presents a difficult challenge in assessing credibility of information, and requires a more comprehensive understanding of the components of a microblog message and their potential impact on the information consumer's assessment of information credibility. This is bolstered by the fact that the majority of users in Twitter (52%) and Reddit (60%) treat the system as their primary source of news information [Holcomb et al. 2013].

Understanding. Recently there have been many efforts to study information credibility in microblogs, ranging from automated algorithms to model and predict credibility of users [Kumar et al. 2013; Wagner et al. 2012] and messages [Castillo et al. 2011; Kang et al. 2012] at general [Castillo et al. 2011; Sikdar et al. 2013] and topic-specific [Kang et al. 2012] levels, to visualization and interaction applications such as [Marcus et al. 2011; Rios and Lin 2013; Kang et al. 2013]. However, with the exception of [O'Donovan et al. 2012; Yang et al. 2013], little research has focused on isolating the impact of individual microblog features such as profile images, links and other available metadata on perceptions of credibility—a problem that is increasingly important as a growing portion of news information gets produced by people the information consumer does not know.

Contributions. Now that we have discussed a high level motivation of the problem, including history and common methods to evaluate information credibility in microblogs, we turn to the key contributions of our work, within the above context. In particular, we propose the following three research questions, noting that the first question has been partially answered through our earlier study in [Kang et al. 2015].

[1]Example available at http://www.paywithatweet.com/

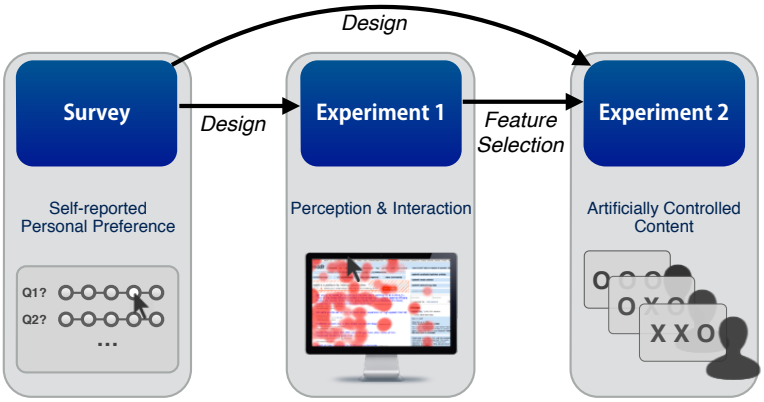


Fig. 1. Overview and dependencies between the initial survey (N=81) [Kang et al. 2015], Experiment 1 (N=102) [Kang et al. 2015] and Experiment 2 (N=646) reported in this paper.

We include the research question from [Kang et al. 2015] and a summary of that experiment since they are necessary for understanding the user experiment [N=646] in this paper.

(1) What are the important cues that contribute to information being perceived as credible on microblogs?
(2) Can we separate them from the content and quantify their impact on a credibility assessment?
(3) To what extent is such a quantification portable across different microblogging platforms?

A high level overview of the three key phases in our studies and the dependencies between them is shown in Figure 1. First, a crowd sourced survey (from [Kang et al. 2015]) of 81 microblog users was performed on Amazon Mechanical Turk (MTurk) to assess patterns in user assessments of information credibility across different topics and for demographic groupings. In this survey, the participants answered to 13 demographic and 17 main questions regarding their microblog usage and perceived credibility on microblog messages. By analyzing a large set of microblog features, a set of important factors was identified for further evaluation.

Second, a user experiment (N=102, from [Kang et al. 2015]) was designed to place users in context across two microblog platforms and elicit a more refined set of salient factors that influenced their judgement of information credibility. To do this, heatmaps were computed from mouse click behavior in the microblog interfaces. To address the question of portability, we perform experiments across two of the most popular microblogs for news-consumption: *Reddit* and *Twitter*.

Third and finally, we conduct a novel experiment (N=646) in which variables from [Kang et al. 2015] are experimentally controlled. This allows us to assess the impact of the each individual variable on credibility assessments in a range of contexts. The main findings in this study is described in Section 5. In summary, by artificially controlling metadata such as number of friends, or the profile image type of an information provider, and gathering human-provided assessments of associated content, we observe a significant shift in reported “information credibility” between the treatments.

An example of a high and low value treatment for the "number of friends" feature is the average number of friends in the 5th and 95th percentiles for a randomly sampled batch of 250,000 users from Twitter and Reddit. By ranking each evaluated feature by the absolute distance between mean reported credibility scores on treatments sampled from large samples of real data, we are able to produce a top-n list explaining the relative influence of metadata features on human assessments of information credibility. Surprisingly, our results show that features that can easily be user controlled, such as profile image or a company logo produce a larger shift in credibility assessment than features that are network-controlled, such as number of shares or number of friends, despite the obvious fact that these features are much more difficult to fake.

1.1. HYPOTHESES

2. RELATED WORK

We have introduced the motivation of our research and provided high level description of our experiments. Now, we will discuss it within the context of relevant related research, before providing details of the experimental setup and results. In this section, we discuss definition of information credibility from literature and define our notion of credibility. This is followed by comparing and contrasting different approaches for modeling information credibility in microblogs, at the computational and user-perception levels.

Credibility Definition. Information credibility is a multi-dimensional concept that has been studied in a variety of disciplines over many decades. There is no universally accepted definition of information credibility, and definitions vary across, and even within research disciplines. In this paper, we have a particular focus on credibility of information within major media, including traditional outlets such as television and radio, to microblogging platforms such as Twitter and Reddit. Again, there is no universally accepted definition of information credibility within this field. Fogg and Tseng, in their study on the elements of computer credibility [Fogg and Tseng 1999], provided a definition that has been frequently cited in subsequent literature related to information creibility in social media. Fogg describes credibility as "(a) a *perceived quality* (b) made up of multiple dimensions, mostly *trustworthiness* and *expertise*." [Fogg and Tseng 1999] define trustworthiness and expertise as "a key element in the credibility calculus defined by the terms well-intentioned, truthful and unbiased", and "another dimension of credibility defined by terms such as knowledgeable, experienced, competent", respectively.

They also claim that credibility does not reside in an object, a person or a piece of information. In this work, we consider credibility as a function of *perception* including the person and the object being perceived. Many studies of credibility (e.g.: [O'Donovan et al. 2009]) find that it is comprised of two primary dimensions–*trustworthiness* and *expertise*.

Credibility models. In recent years, researchers proposed different credibility models to evaluate information on the Web. For example, [Fogg 2003] studied how people evaluate information credibility online and proposed the idea of *Prominence-interpretation theory* which, as the name implies, is comprised of the two attributes: "prominence" (important or distinct entities) and "interpretation" (user assessment of the entities). Other researchers proposed models or algorithms that either measure or predict credibility of users in different contexts. For example, [Kumar et al. 2013] propose a model of location and topic affinity to identify credible, relevant users under crisis situations. [Castillo et al. 2011] propose a model for identifying a group of credible and newsworthy sources of news information based on computational models. [Kang et al. 2012] also

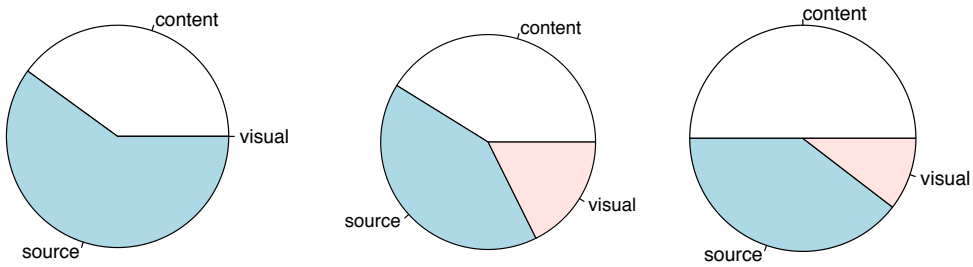
build a model that identifies credible messages on microblogs and evaluate the model on individual message level. Going beyond theory, some researchers focus on empirical evaluation of credibility and the problem of ground truth. For example, [Sikdar et al. 2013] proposed a two pronged approach to gathering ground truth information on the credibility of microblog data by combining manually annotated scores with observed network statistics (e.g: retweets) from the data to achieve a "more stable" estimate of credibility. A recent study [Gupta et al. 2014] proposed a semi-supervised model for computing credibility scores of individual microblog messages in real-time through a large-scale system that employs the model.

Perceived Credibility. While there is no broadly accepted definition of information credibility across different disciplines, most researchers have agreed that "credibility" that is inherent to an entity and perceived credibility of that entity are not necessarily equivalent. The latter could be viewed as a subjective function of the former. Perceived credibility can fluctuate from (inherent) credibility based on the way in which the entity is represented, and based on the characteristics of the person making the credibility assessment. Accordingly, numerous researchers from different disciplines have attempted to identify a set of salient factors that contribute to our perception process. Visual and textual components have been studied in the same vein [Fogg et al. 2003; Morris et al. 2012; Flanagin and Metzger 2000; 2007] to reveal complex relationships between data, metadata and context that inform our perception of information credibility. Other researchers have studied influence at the network level, identifying trends and patterns that lead to tipping points of credibility and popularity [Gladwell 2002] and surges in influence within the network [Cha et al. 2010].

In recent years, as the role of microblogs as a news source becomes more important, researchers have studied how individual features or cues can affect perceived credibility of information on microblogs [Flanagin and Metzger 2007; Yang et al. 2013; Morris et al. 2012]. A common theme in these recent studies is to first select numerous candidate features that are likely to contribute to the assessment processes, and then analyze them in both qualitative and quantitative ways through online surveys or user studies. For example, [Morris et al. 2012] found disparity between the features considered for evaluating information credibility between search engines and Twitter. Morris also reported a controlled experiment that revealed the features through which users assess information credibility on Twitter. Their study also provided insight and suggestions for interface design to improve credibility perception from the end-user perspective. Perceived credibility has also been compared across different cultural settings by [Yang et al. 2013] Their study reports on experimental and survey data that compares and contrasts the impact of several features of microblog updates (authors gender, name style, profile image, location, and degree of network overlap with the reader) on credibility perceptions among U.S.(Twitter) and Chinese(Weibo) audiences. Their goal was to design new user experiences which can maximize both *credibility* (as a property of entity) and *perception of credibility* (an end-user subjective function, to which the entity is an argument) of the contents on social media. Perceived credibility can be impacted by personality characteristics such as those modeled by [Mahmud et al. 2013]

3. DEMOGRAPHIC ANALYSIS

Before our user experiments, we conducted a crowd sourced survey to understand how users evaluate their credibility assessment on microblog content through self-reported responses on our questionnaire. This initial survey provides us a guideline to 1) set up our research questions and hypotheses on important features and 2) better design



(a) Monthly Users (N=15) (b) Weekly Users (N=18) (c) Daily Users (N=48)

Fig. 2. Self-reporting responses to the question about the primary factor for credibility assessment on microblog content across different user groups (usage frequency).

Table I. Primary use of information on Twitter

Business	22(27.2%)	Information Sharing	21(25.9%)
Social Purpose	16(20.0%)	Information Search	15(18.5%)
Serendipitous Search	4(4.9%)	Other	3(3.7%)

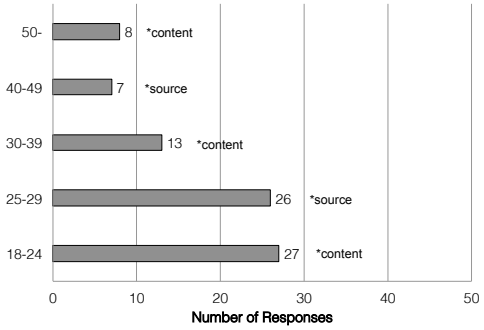
the two pipelined studies that measures quantitative user interactions and artificially control selected features one at each time.

To gather a fair assessment of candidate features to evaluate in our main experiments, and to gather insight about credibility decisions in microblogs, we conducted a crowd sourced study targeting Twitter users in late 2013 using Amazon's Mechanical Turk (MTurk) platform. A total of 81 respondents were asked a series of questions to explore what information from microblogs they mostly consider when they need to search for credible information about particular events. The 59 male and 22 female participants were from different parts of the world, with a majority from the United States and India. Participant age ranged between 18 and 60 with an average of 28. 60% of the subjects used microblog on a weekly (22%) or daily (38%) basis.

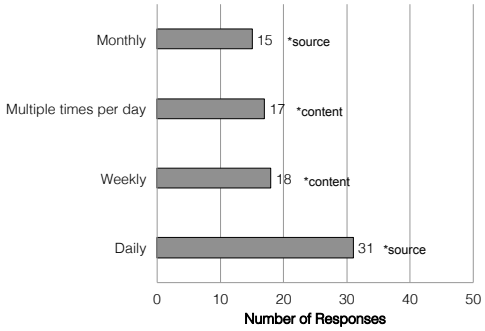
In the demographic questionnaire, participants reported basic information such as Twitter usage, educational and cultural backgrounds and yearly income. For example, 82% of the subjects reported that they hold Bachelor's degree or equivalent for educational background. The majority of the participants reside in either medium-sized or large cities. In terms of the marital status, the subjects are equally distributed between single and married groups. The results are shown in Figure 3.

The overarching goal of the survey was to explore the following general hypothesis through self-reported metrics and to identify the set of Twitter features (E.g: links, profile images etc.) reported as most influential in credibility assessment. The two additional experiments in this paper expand the general hypothesis into 6 additional hypotheses, and were both designed based on analysis of the results from this survey. In the survey, 20% (16/81) of the participants reported that they consider visual cues as a major factor that affects their credibility assessments. Details of the resulting design decisions are discussed in Section 4.

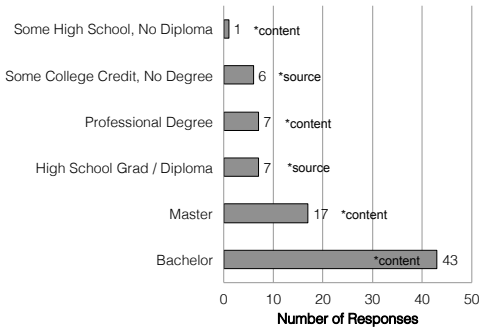
Research Question (RQ) 1: Does the display of metadata in microblogs influence perceived credibility of the associated content.



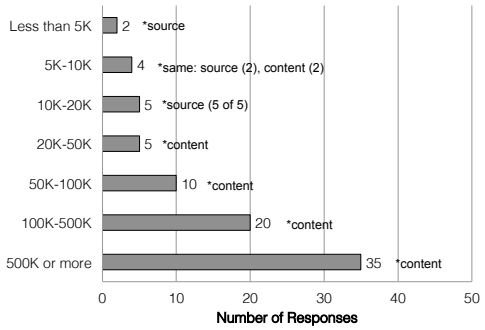
(a) Age



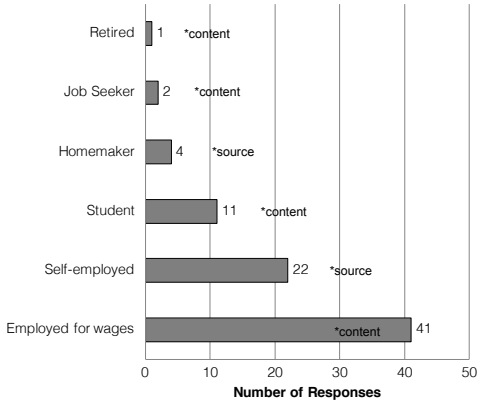
(b) Microblog Usage



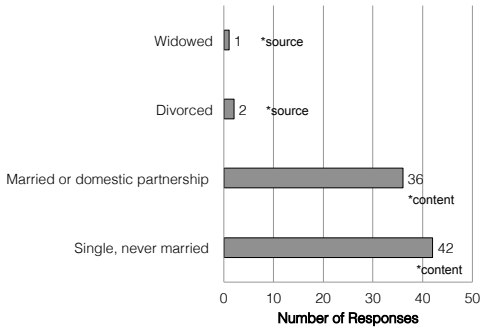
(c) Educational Background



(d) Population of Locality



(e) Employment Status



(f) Marital Status

Fig. 3. Self-reporting responses on the demographic questionnaire from microblog users. `*label` represent most frequent primary factor that survey participants consider for their credibility assessment.

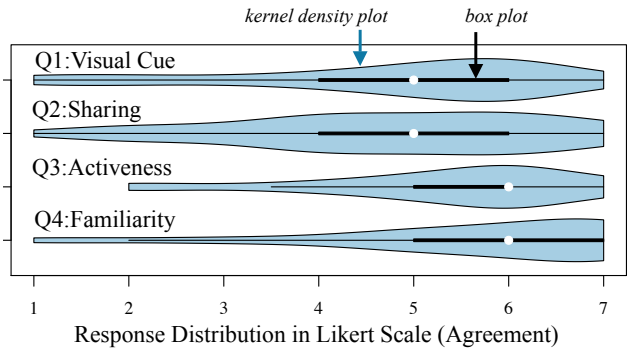


Fig. 4. Survey responses for key questions (marked with $*$ in Table II. Each line and enclosed curve shows the response distribution for questions in (1) a box-whisker-chart indicating medians and quartiles and (2) a kernel density plot, respectively. Responses are provided on a Likert scale (Strongly disagree (1) – Strongly agree (7)).

HYPOTHESIS. *Metadata display (textual or visual) influences perceived credibility of microblog content. The direction of influence is dependent on the specific content displayed.*

To gauge usage patterns and credibility perception, participants were asked 17 questions in a web survey, covering aspects such as activity rate, perceived impact of visual cues, and sharing frequency, among others. A selection of these survey questions are shown in Table II, and response distributions are shown in Figure 4. The results indicate that the majority of the participants consider themselves active information consumers on microblogs (79% for 'Activeness'). These users also share their own content frequently with their followers (57%, 'Sharing'). The most common usage reasons (Table I) were reported as business (27%) including online marketing, information sharing (26%) and social use (20%). Interestingly, 68% of participants reported that visual elements have significant impact on their credibility assessment in the microblog, as indicated by Q1 in Figure 4. Our population exhibited reasonably heavy use of Twitter (38% daily use) and a good standard of education across participants, with 67/81 possessing at least a bachelor's degree.

Credibility Factors. The main section of the survey investigated what kinds of attributes participants consider as a primary factor when they search for credible information on microblogs. We can intuitively expect that both content and information source would be highly ranked and the results indeed support this. However, it is interesting to note that *visual cues* such as design and layout were also reported as influential in the process of credibility assessment. 10% of participants responded that design/layout was the primary factor (20% elected it a major factor) in their credibility assessments.

Correlation Analysis. People consider many different factors during credibility assessment with microblog information. Numerous researchers concluded that, ultimately, credibility can be perceived or measured in different ways based on the given context, cultural background, language, etc. [Fogg 2003]. We also find that many microblog users agree with this statement from pre-study offline interviews. Thus, we designed our questionnaire to find underlying correlation, if any, between demographic background and the question responses. Results are shown in Figure 5. Table II provides a full description of each element in the correlation plot of Figure 5. Some notable correlations include Twitter use and general information use. There was a positive cor-

Table II. Survey Questions. * denotes further detail in Figure 4

Label	Questions
Activeness*	Do you consider yourself as an active online information consumer?
Sharing Freq*	Do you frequently share your information with the people in your network (followers)?
Primary Usage	What is your primary usage of information on microblogs?
Familiarity*	Are you familiar with microblog services?
1st Cred Factor	Which do you consider as a primary factor for measuring information credibility?
2nd Cred Factor	Which do you consider as a secondary factor for measuring information credibility?
Visual Cues *	Do you think that visual cues are important for judging credibility?
Url Relevance	Do you think that the presence of URLs in a tweet, which point to an external information source, can enhance information credibility?
In-Person Friends	About how many of your "friends" on Twitter have you met in person?
Non-Human Friends	About how many companies or organizations do you currently follow on Twitter?
Celebrity Friends	About how many celebrities do you currently follow on Twitter?
Time-On-Others	On Twitter, about how much time do you spend looking at what other users have posted?
Time-On-Me	On Twitter, about how much time do you spend posting tweets about yourself?

relation between employment type and content use –this may have been a result of the number of users who said they used the microblogs for marketing purposes. There was a strong correlation between locality (size of city lived in) and amount of information shared on microblogs. People in larger cities shared more information than those in small cities and towns. Predictably, employment type was positively correlated with primary usage of the microblog. We also find a correlation between gross income / ethnic origin and microblog usage, complementary to [Goel et al. 2012], who found strong correlation between these factors and browsing behavior. Demographic factors (both age and cultural background) correlated with usage rate, and with the impression of visual cues as an information credibility factor –younger people had higher usage rates and were more influenced by visual cues.

In summary, the initial analysis from the survey highlights visual cues as a useful factor for further study, incorporating aspects of content, and metadata about the source/provenance of microblog messages.

3.1. Cultural Background

Both West European and Asian subjects reported that they consider information source (author) as the primary factor during their credibility assessment on microblog content while others consider content as the top priority.

4. EXPERIMENTAL SETUP

To further explore what factors influence credibility perception in microblogs, we designed and conducted two different user experiments guided by insights from the credibility perception survey. Both experiments were conducted on MTurk. In this section, we detail the design of both experiments (`Exp1` and `Exp2`) and discuss a refined set of experimental hypotheses.

4.1. `Exp1`: Perception and Interaction

The initial survey highlighted that content (meaning) and sources (origin) of microblog posts are the most influential factors in credibility assessment. However, the representation of information such as metadata also plays a significant role in user perception of information credibility. According to these findings, we refine the initial research questions to include the following three questions/hypotheses.

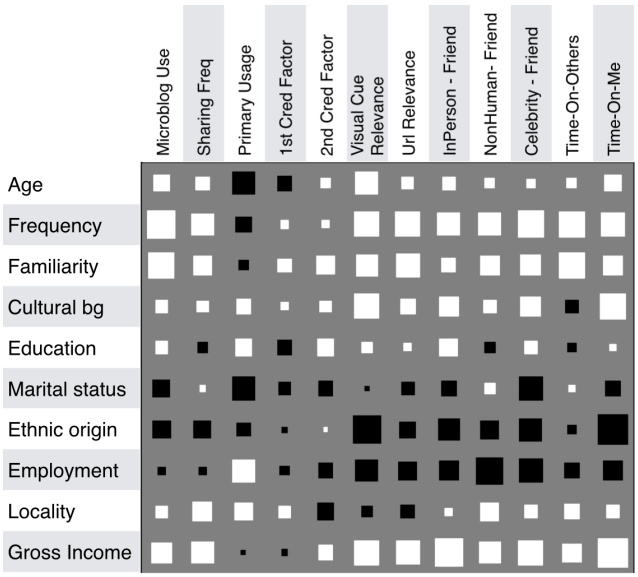


Fig. 5. remove non-nominal features Correlation matrix between the demographic information of the participants and the responses to the survey. The matrix is visualized in a Hinton Map (white and black squares represent positive and negative correlation, respectively. Square size is proportional to the absolute value of the score [0–1].)

RQ2: Do different features influence credibility by different amounts?

HYPOTHESIS. *Features have varying degree of influence over credibility perception*

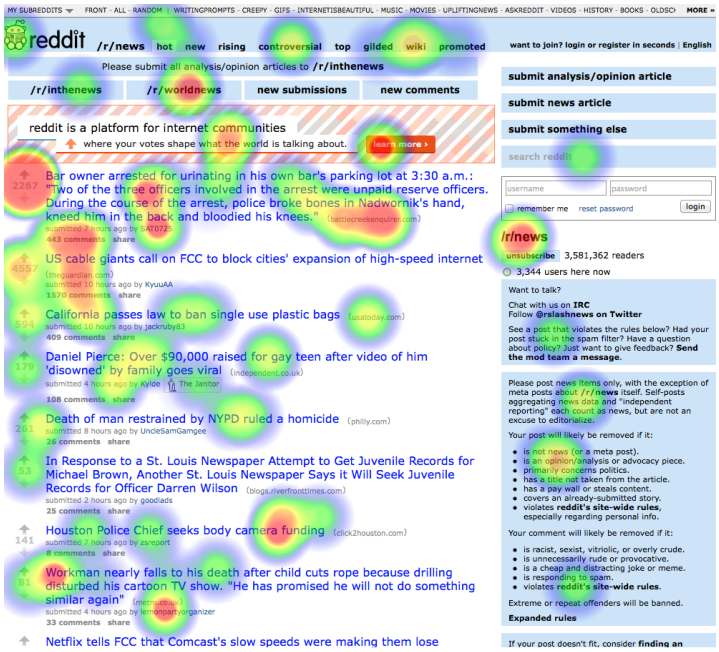
RQ3: What are the effects of different classes of microblog features on perceived credibility?

HYPOTHESIS. *Visual factors will have the most influence, followed by network and content-based (text) factors.*

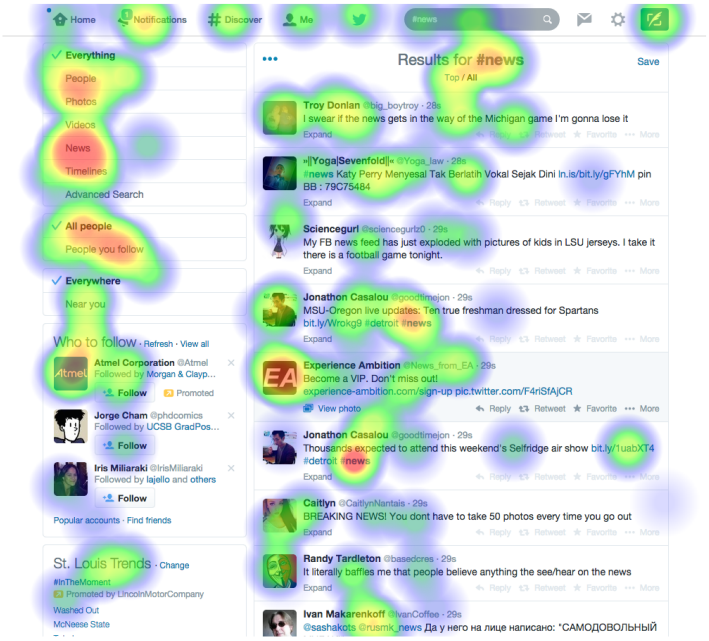
RQ4: Can our models of feature influence be ported successfully to different microblogs?

HYPOTHESIS. *Influence of features is consistent across platforms*

To test these hypotheses, a study was designed to place users in a familiar/typical microblog context and provide them with a simple mechanism to highlight the specific features that they felt had an influence on their perception of content. To address the last hypothesis, the study was designed to be cross-platform, comparing features from Twitter and Reddit. Figure 6 shows two example interfaces from the study (N=102). Users were requested to click on or close to items that they felt had *any* impact on their perception of information, regardless of positive or negative direction, which is evaluated separately in our third experiment. They were given no a-priori information on specific feature lists. This mechanism for identifying influential features was used in an effort to avoid bias from manual or expert selection of a feature set (intended for detailed analysis in Experiment 2). Domains (Twitter and Reddit) were a between-subjects variable, and only participants with significant prior experience with a domain were allowed perform the task.



(a) News subreddit page in Reddit



(b) A search result page with 'news' keyword in Twitter

Fig. 6. Heatmap visualization of the user annotations for Exp1, where clicks on an entity indicate perceived credibility.

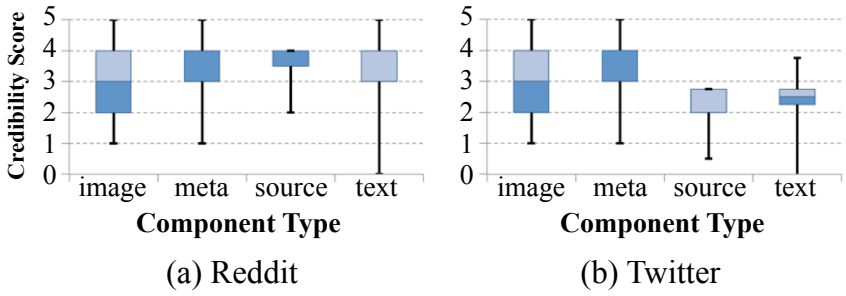


Fig. 7. Click frequency/perceived credibility by UI component type for (a) Twitter and (b) Reddit (Exp1)

Table III. List of features and their independent variables in the second user experiment. (For each treatment, posts from both outlets, NYTimes and The Onion, are presented to the participants.)

Type	Feature	Treatment 1	Treatment 2
Visual	Embedded image	Present	Not present
Visual	Profile image(person)	Present	Not present
Visual	Profile image(logo)	Professional	Unprofessional
Network	# of friends	95th percentile(7,524)	5th percentile(8)
Network	Classification	News	Non-news
Network	# of comments	95th percentile(9,565)	5th percentile(0)
Network	# of shares	95th percentile(933)	5th percentile(0)
Network	Age of message	95th percentile	5th percentile
Content	Sentiment degree	No sentiment	High sentiment(95%)
Content	Sentiment polarity	Negative value(-0.95)	Positive value(+0.95)
Content	Tags	Tags present	No tags
Content	Links	Links present	No links

First, participants were asked 6 general questions about their microblogging practice. Then they were shown 3-4 screenshot images of the microblog (3 for Reddit and 4 for Twitter). To capture the aforementioned features avoiding possible bias, we let participants select three visual elements instead of having them rank an arbitrary selection of features we provide. On each click, a slider selector was shown below the image to record the amount of impact the element that the user clicked on has on her credibility assessment. We collected coordinates of each click and its corresponding score in likert scale (0 for no effect to 5 for major effect).

Feature extraction. In order to extract meaningful features from this experiment, we analyzed the results from a heatmap visualization (Figure 6) and statistical analysis using five-number summaries (Figure 7). As can be seen in Figure 7, there is overall similarity in credibility ratings on different elements for both Reddit and Twitter users. However, users of Reddit express higher priority on both information sources and textual elements for their credibility assessments. This observation may be due to small differences between two social platforms: For example, most of the posts in Reddit are directly connected to external webpages and this makes the source (URLs) more important during credibility assessments. Additionally, posts in Reddit are longer than in Twitter, which could account for the higher text credibility score for Reddit. Although the metadata scores are similar in Figure 7 for both platforms. Twitter does provide a richer set of metadata (e.g. classifications, hashtags, retweet counts etc.) on their page layouts, and this is evident from the increased number of clicks on metadata components in the heatmap (Figure 6(b)).

4.2. Exp2: Artificially Controlled Content

From the previous experiment, we selected a set of features on which to base our experimental evaluation for Exp2. Once again, our research question and hypothesis was refined based on information from the previous studies. By artificially controlling values for each target feature, we can assess the directional effect of metadata content on credibility perception. Furthermore, to avoid topic-specific biases in assessing the stability of feature influence on credibility perception across topics, we incorporated a variety of common topics (e.g.: World, Health, Politics, Entertainment) into the evaluation.

RQ5: How do different treatments of metadata variables influence credibility perception?

HYPOTHESIS. By applying artificially controlled values for metadata from the 5th and 95th percentiles of sampled real world data, we will observe differences in perceived credibility of the associated information.

RQ6: Is the influence of displayed metadata on credibility perception consistent across different topics?

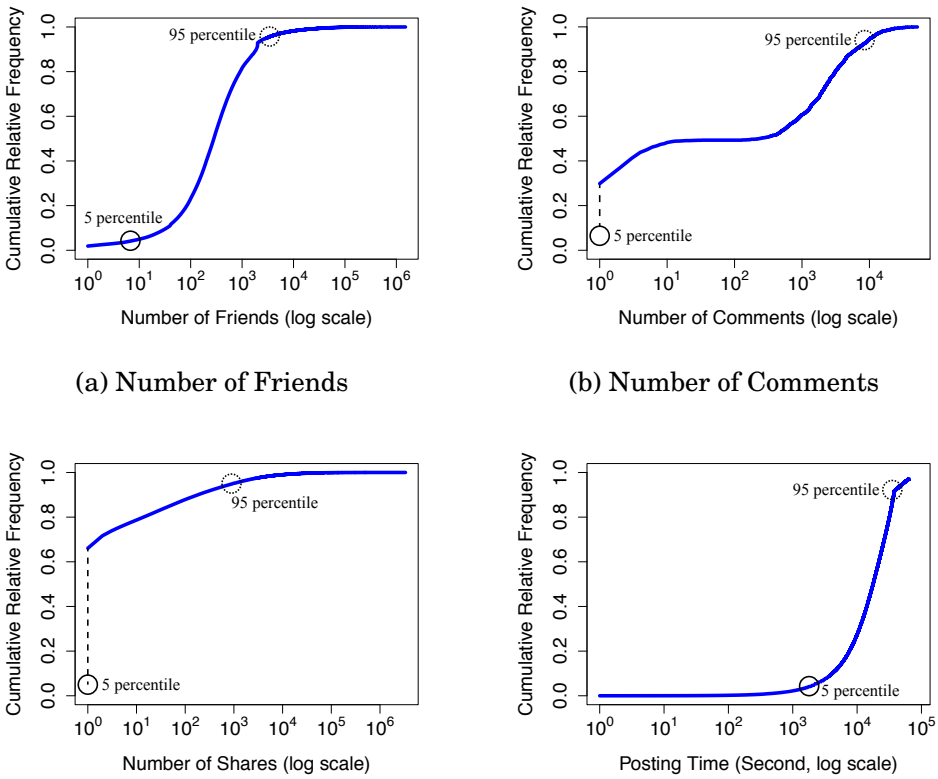
HYPOTHESIS. *Feature influence varies across topics.*

4.2.1. Study design. In this experiment, we aim to test hypotheses 1 through 6 by artificially controlling metadata values for each of 12 salient factors identified in the previous experiment, and eliciting credibility assessments from participants. To achieve this, we construct 12 different formulated lists of microblog postings, controlling one independent variable on each list to capture how much impact that individual factor (e.g: profile image, link, number of friends) has on perceived credibility of information. Figure 9 shows a screen shot of the interface used in the study. The treatment in this case is a default profile image, which is the only controlled variable in this example. A five point Likert scale for feedback on perceived credibility is shown beneath the blog post.

A larger user experiment (N=646) was deployed on MTurk to evaluate the effects of artificially controlled treatments of each feature. In order to determine in which direction each factor impacts on the perceived credibility, we designed the study with two treatments and one baseline for each feature. *Treatment 1* uses 'feature present' in case of binary features and 95th percentiles (high values) in case of numeric features. Correspondingly, *Treatment 2* exhibits 'feature absent' in case of binary features and 5th percentiles (low values) for numeric features. The exception to this choice of percentiles are our two "sentiment" related features. Since we assume that low sentiment indicates more credibility, in view of the fact that objectivity is linearly correlated with credibility, we mapped the 5th percentiles to Treatment 1 and 95th percentile to Treatment 2 here.

Parameter Selection. To estimate reasonable values for the treatments listed in Table III, 1,727,556 Twitter messages and 4,000 Reddit posts including user profiles were crawled using the Twitter Streaming API and Reddit API. A distribution analysis of feature values in this data allowed us to find reasonable thresholds (and extremes) to select values for our experimental treatments. From this dataset, we extracted 5th and 95th percentiles from the feature distributions. The selected values (5th and 95th percentiles) for each numeric feature are shown in Figure 8.

The complete list of features tested in this experiment (Exp 2) are shown in Table III. The leftmost column shows a categorization of each feature into one of three classes:



(a) Number of Friends (b) Number of Comments

(c) Number of Shares (d) Age of Message

Fig. 8. Cumulative Relative Frequency (CRF) on numeric features on which we base to select conditions for the two treatments in Exp2. Please note that we use Twitter dataset (a sample set of 1,727,556 messages) for (a), (c) and (d) and Reddit dataset (a sample set of 4,000 posts) for (b).

— *Visual* This is the set of highly visual factors in the microblog, including profile pictures, attached images, and photos.
— *Network* This is the set of network-based factors, including static features such as number of friends or followers, and dynamic/conversational features such as retweets, votes or mentions.
— *Content* This is the class of features solely based on text, including sentiment terms, hashtags, and links.

Table III describes our selection of treatments for each target feature. While we cannot exhaustively evaluate all values for a particular feature, our aim was to construct a reasonably diverse set of values for each, based in some cases on analysis of real world distributions (e.g., for numeric attributes such as number of friends), and on manual selection for others (e.g: profile image content). The two rightmost columns give examples of controlled metadata treatments for each variable in our study. For instance, in the case of profile images, one treatment was selected from a popular satirical website. A second treatment (showing an image of President Obama giving a speech), was selected from a major US news outlet. The main goal of Table III is to show that differences can be produced in perceived credibility compared to a baseline, by manipulating a single visual feature, and our results show that this is indeed the

case. For data with a range, such as number of friends, scores were taken based on large-scale distribution analysis for each feature, as described in the Random Sampling section below. The majority of examples were straightforward to construct, for example, network features such as number of friends or retweets were represented as low or high values. Recency was evaluated by grouping messages by age of message and displaying 5th or 95th percentile values of the resulting distribution. Sentiment features were more difficult however: for polarity and sentiment degree, a negative value (negative sentiment used) and low value (little sentiment used) were assigned respectively. These were indicated using "score bars", shown in the example column of Table III. For content/text features such as tags and links, a binary value (present or not present) was used. A baseline condition was also used. The baseline consisted of raw text with no associated metadata. Participants were asked to answer the same type of general questions about their demographic information and experience with microblogs as in Exp1. To avoid topic-specific biases and to explore our earlier hypothesis, 5 topics (World, Health, Politics, Entertainment, Business) and 10 posts from both *New York Times* and *The Onion* accounts on Twitter were manually collected. We purposely sampled two outlets that represent two different aspects of online journalism (objectivity and satire) in order to reflect real-world contexts.

5. RESULTS AND DISCUSSION

In this section, we provide the findings from our two main user experiments. This section provides an overview of participant statistics for both studies, and following that, is organized around the six research questions and hypotheses posed earlier.

5.1. Study Participants

Exp1 had 102 participants. The average interaction time for both Twitter and Reddit users was 5 minutes. Users annotated three items each for a total of 306 annotations. 646 users participated in Exp2. Most reported that they were active daily on Twitter and had been active for more than a year. Most used the the official application, on a combination of mobile and desktop platforms. Most users did NOT guess that content was sourced from either the New York Times or the Onion. Facebook and Twitter was the most common response for the two source platforms. Participants spent an average of 9 minutes completing the survey and were paid $0.40. 55% were male and 45% female. They ranged in age between 18 and 60, with the majority between 18 and 29.

5.2. Influence of Metadata

RQ1: Does the display of metadata in microblogs influence perceived credibility of the associated content?

HYPOTHESIS. *Metadata display (textual or visual) does influence perceived credibility of microblog content. The direction of influence is dependent on the specific content displayed.*

All three experiments produced results that reveal an impact of metadata on perceived credibility in microblogs. The subjective results in Survey 1 show a strong (but self-reported) indication that Content of a message and Origin (author) of a message are the strongest influencing factors. This is followed by visual features, including design of the UI and visual components such as profile pictures and other metadata. The discussions that follow here further illuminate and reinforce this basic result.

5.3. Cross-Feature Analysis

RQ2: Do different features influence credibility by different amounts?

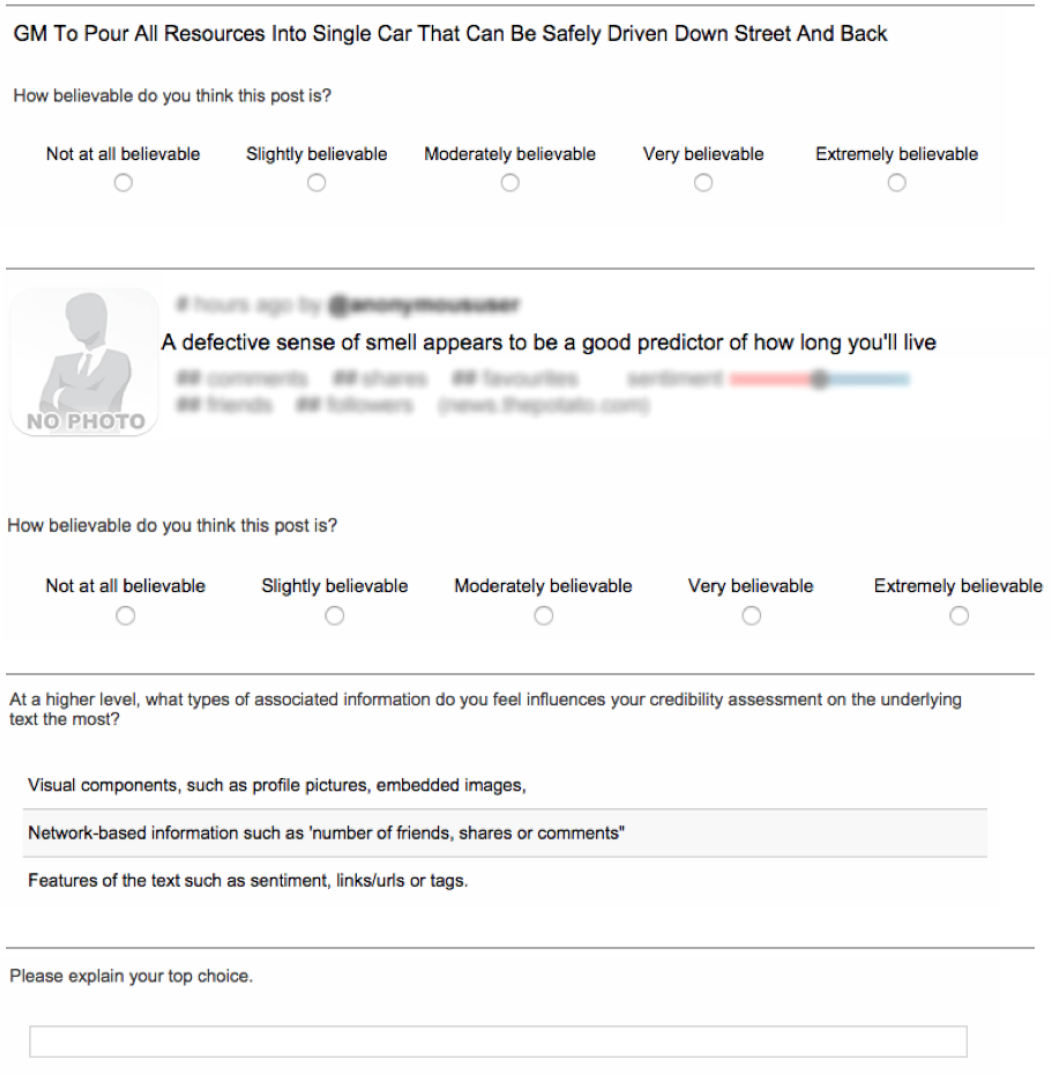


Fig. 9. Screenshot showing part of the microblog interface from Exp2.

HYPOTHESIS. *Features have varying degree of influence over credibility perception*

Figure 10 shows a list of the 12 evaluated features ranked according to the observed difference in influence between the treatments from Table III. The distribution clearly supports our initial hypothesis on RQ2. From this list, the profile picture/logo is the most influential factor on credibility perception, showing a significant increase over other features in terms of the difference in credibility rating between the treatments (ANOVA $p = 7.79e - 9$ and $p < 0.05$ in Tukey post-hoc tests). Factors reinforced by the underlying network (Number of Friends –static, and Number of Shares –dynamic) are next-most influential. On the opposite end of the scale, sentiment polarity and age of message were the only two features where the lower value treatment achieved a better

score than the higher value treatment. It seems that users in this study were not too concerned with recency. The more sentiment a message contained, the more likely it was to be deemed not credible.

Figure 10 shows the individual differences between the treatment 1 (white, top bar), baseline (dark gray, middle bar) and treatment 2 (light gray, bottom bar) for each of the 12 features. Overall, it is clear that the treatment 1 had a much stronger influence compared to the baseline (no controlled metadata) for almost all features. Effects were significant between the treatment 1 and the baseline in most cases but not significant between the treatment 2 and the baseline.

5.4. Influence of Feature Classes

RQ3: What are the effects of different classes of microblog features on perceived credibility?

HYPOTHESIS. *Visual factors will have the most influence, followed by network and content-based (text) factors.*

Many existing credibility models perform feature classification to arrive at a credibility prediction for a given message. To evaluate our simple classification of (Network, Visual and Content)-based factors, we examined the overall group-wide credibility ranking and the results are shown in Figure 11 (a). Our initial hypothesis was that visual factors would be most influential, and it appears from the left column in Figure 11 (a) that this is in fact the case, at least for the high value treatments of the features in the group. No significant effect was shown for the lower value feature differences. The visual features produced a 10% increase over the baseline, compared with 7% for the Network and 3% for the Content group. This result further supports the notion that manipulating visual components can have a large effect in terms of how strangers perceive a microblog profile.

5.5. Cross-Platform Analysis

RQ4: Can our models of feature influence be ported successfully to different microblogs?

HYPOTHESIS. *Influence of features is consistent across platforms*

To recap, `Exp1` evaluated self-reported importance of microblog features across two platforms: Reddit and Twitter, allowing users to place simple clicks in context of the actual interface. Figure 7(a) and (b) show the results for Reddit and Twitter respectively. Our results disprove our initial hypothesis that feature ratings are invariant, since the text features achieved a higher credibility rating in Reddit. This is likely to be related to the fact that there is significantly more text per post allowed in Reddit. Image features appeared to garner similar ratings across the two platforms, which is a meaningful finding, especially coupled with the fact that the Visual/Image-based features are the strongest influences on credibility perception.

5.6. Impact of Different Treatments

RQ5: How do high and low values of metadata contents influence perceived credibility?

HYPOTHESIS. *L*ow values for metadata have a stronger effect than high values

Figure 11(d) shows a significant (7%) improvement across all features and topics for treatment 1 over treatment 2. There is also a significant improvement shown for treatment 1 over the baseline (6%), Specifically, a single factor ANOVA test over the

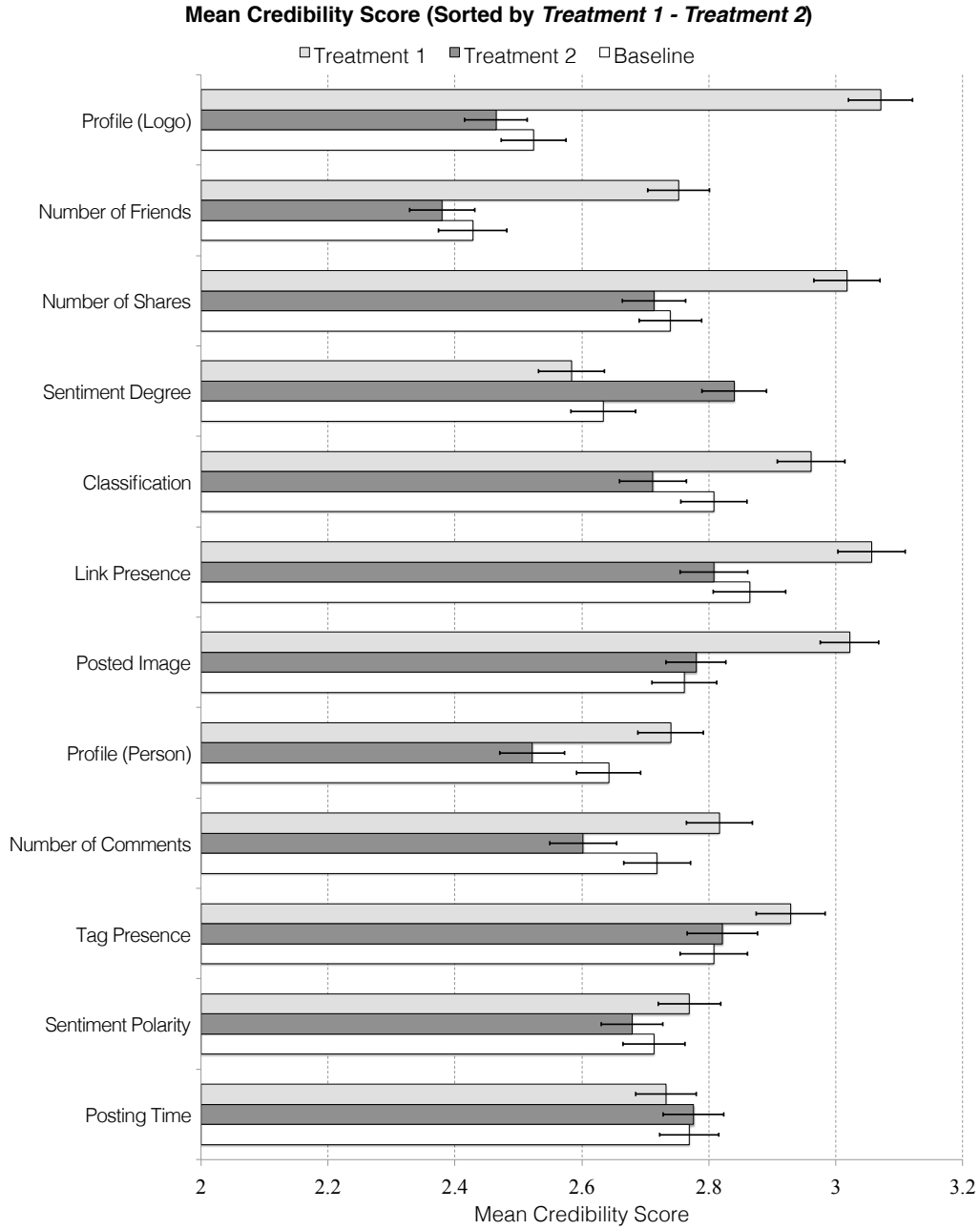


Fig. 10. Mean credibility scores across different features for the three treatments including baseline. Groups are sorted from top to bottom by the difference between the treatment 1 and treatment 2 scores.

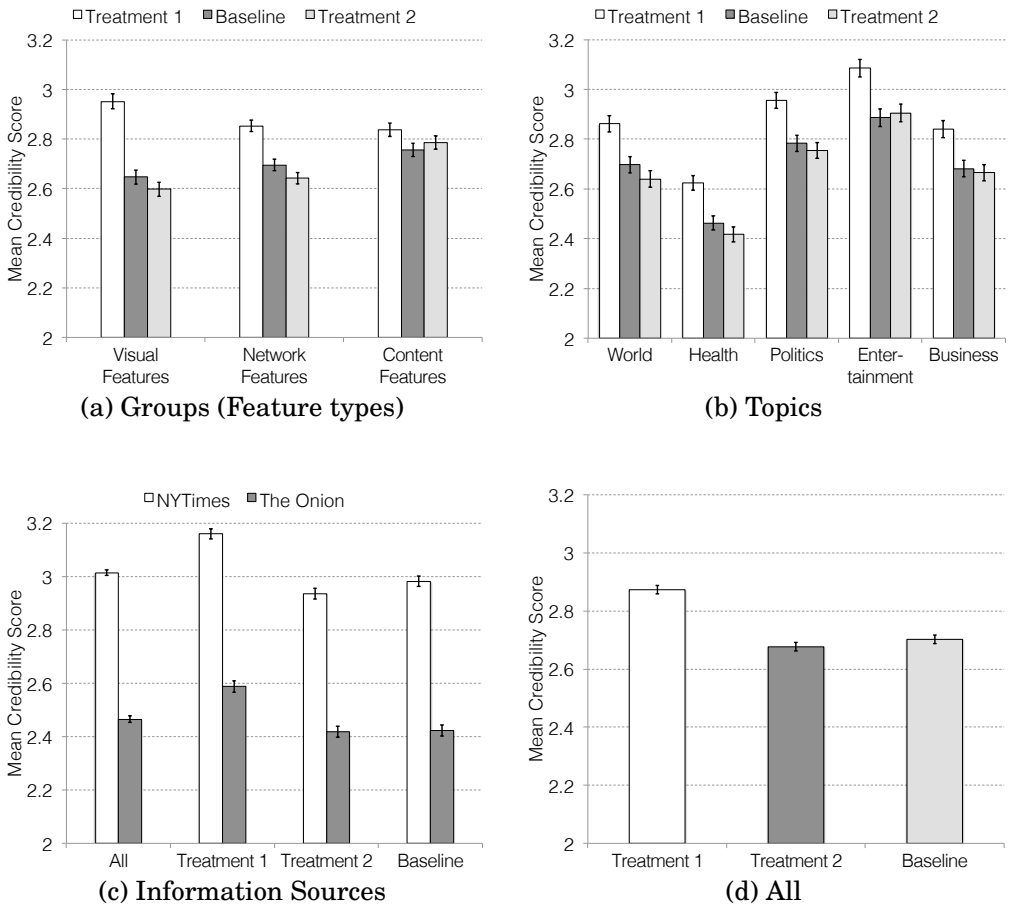


Fig. 11. Mean credibility scores across different treatments

treatments shows the result is statistically significant ($F\ value$:87.54, $Pr(> F) < 2e - 16$).

5.7. Cross-Topic Analysis

RQ6: Is the influence of displayed metadata on credibility perception consistent across different topics within a domain?

HYPOTHESIS. *Influence of the treatments varies across topics.*

To determine whether or not credibility ratings for our feature sets vary across different topics, and also to avoid introducing topic-specific biases in our other evaluations, we examined a collection of 5 diverse topics (World, Health, Politics, Entertainment and Business). This approach could be relevant for a system such as "CatStream" [Garcia Esparza et al. 2013] which classifies Twitter feeds automatically based on topic interests. Messages from two well-known websites were sampled for each topic on a specific date in September 2014. The two websites were the New York Times (News) and The Onion (Satire/Comedy). Figure 11 (b) shows the results of an analysis of our different treatments across each of the five topics individually. From the results, it is immediately clear that the recurring trend of treatment 1 having a stronger effect

than treatment 2 is invariant across all topics. ANOVA tests for differences showed $p < 2e-16$. Another interesting result is that the overall ratings for the health-related topic was significantly lower than all other topics, across all treatments. This might be a reflection of participants' cautious behavior when approaching a serious topic such as health. Conversely, the Entertainment topic produced the highest credibility ratings. Perhaps it is easier to appear credible in a domain that has far fewer constraints.

In addition to the cross-topic analysis, we also examined the overall ratings of articles from "New York Times (NYT)" and "The Onion" across every rating context, broken down by the three treatment conditions. Figure 11 (c) shows the results of this analysis. There is a significant improvement of 25% (ANOVA, $p < 2e-16$) in credibility rating between the two sources, with NYT at the upper end. Over all features, treatment 1 of metadata produces an equivalent increase in both NYT and Onion ratings, keeping the difference at approximately 25%. When comparing treatment 2 to the baseline treatment, we did not observe any significant change.

6. CONCLUSIONS AND FUTURE WORK

To conclude, in this set of three interlinked studies we evaluated ways in which individual components of microblogs can influence end-user perceptions of information credibility. In particular, an initial survey (N=81) provided general insight into a candidate set of factors that influence credibility most; a second study (N=102) examined these factors in the context of two microblog domains, Reddit and Twitter, to allow real users to communally identify the most influential factors. A set of 12 factors were evaluated in detail in a third study that artificially controlled values for each feature and assessed credibility opinions from 646 participants in a crowd sourced evaluation.

Six hypotheses related to the impact of different microblog elements and treatments on human-provided assessments of credibility were tested and the results were discussed in detail. Key findings from the study show that 1) metadata from the high end of observed real-world distributions (e.g. high number of friends, high number of hashtags) have a far stronger effect on credibility ratings than treatments from the lower end of the distribution. 2) Visual factors, in particular, display of a Profile Picture Logo had the most positive impact on reported credibility. 3) Participants in the study did not view recency of posts as an important factor in credibility assessment. 4) Factors that influence perceived credibility did not remain constant across platforms. Text-based features scored higher on Reddit while Visual features did remain constant.

In follow-up studies, the authors plan to further examine some of the findings from this work. For example, what are the exact reasons for 'negative' traits, such as low numbers of friends, comments, and shares, not exhibiting as much of an effect on the mean credibility score as the positive ones across most features and topics? Future work will also apply findings from this experiment to improve prediction models from real world data such as [Castillo et al. 2011; Kang et al. 2012], predicting newsworthiness, credibility, or actions such as retweets, up-votes or shares.

ACKNOWLEDGMENTS

This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053; The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

Alex Burns and Ben Eltham. 2009. Twitter free Iran: An evaluation of Twitter's role in public diplomacy and information operations in Iran's 2009 election crisis. (2009).

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 675–684. DOI:http://dx.doi.org/10.1145/1963405.1963500

Meeyoung Cha, Hamed Haddadi, Fabrcio Benevenuto, and Krishna P. Gummadi. 2010. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Weblogs and Social*.

Andrew J. Flanagin and Miriam J. Metzger. 2000. Perceptions of Internet Information Credibility. *Journalism & Mass Communication Quarterly* 77, 3 (2000), 515–540. DOI:http://dx.doi.org/10.1177/107769900007700304

Andrew J Flanagin and Miriam J Metzger. 2007. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society* 9, 2 (2007), 319–342.

B. J. Fogg. 2003. Prominence-interpretation Theory: Explaining How People Assess Credibility Online. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*. ACM, New York, NY, USA, 722–723. DOI:http://dx.doi.org/10.1145/765891.765951

B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. 2003. How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences (DUX '03)*. ACM, New York, NY, USA, 1–15. DOI:http://dx.doi.org/10.1145/997078.997097

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

Sandra Garcia Esparza, Michael P. O'Mahony, and Barry Smyth. 2013. CatStream: Categorising Tweets for User Profiling and Stream Filtering. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 25–36. DOI:http://dx.doi.org/10.1145/2449396.2449402

Malcolm Gladwell. 2002. *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books.

Sharad Goel, Jake Hofman, and M. Sirer. 2012. Who Does What on the Web: A Large-Scale Study of Browsing Behavior. (2012).

Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. TweetCred: Real-Time Credibility Assessment of Content on Twitter. In *Social Informatics*, Luca Maria Aiello and Daniel McFarland (Eds.). Lecture Notes in Computer Science, Vol. 8851. Springer International Publishing, 228–243. DOI:http://dx.doi.org/10.1007/978-3-319-13734-6_16

Jesse Holcomb, Jeffrey Gottfried, and Amy Mitchell. 2013. News Use Across Social Media Platforms. (2013). http://www.journalism.org/2013/11/14/news-use-across-social-media-platforms/

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07)*. ACM, New York, NY, USA, 56–65. DOI:http://dx.doi.org/10.1145/1348549.1348556

Byungkyu Kang, Tobias Höllerer, and John O'Donovan. 2015. Believe it or Not? Analyzing Information Credibility in Microblogs. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. (in Press).

Byungkyu Kang, George Legrady, and Tobias Höllerer. 2013. Tweetprobe: A real-time microblog stream visualization framework. In *Proceedings of the IEEE Visualization Conference (VIS) Arts Program (VISAP)*. http://visap.uic.edu/2013/#program

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

Shamanth Kumar, Fred Morstatter, Reza Zafarani, and Huan Liu. 2013. Whom Should I Follow?: Identifying Relevant Users During Crises. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT '13)*. ACM, New York, NY, USA, 139–147. DOI:http://dx.doi.org/10.1145/2481492.2481507

Dominic L. Lasorsa, Seth C. Lewis, and Avery E. Holton. 2012. NORMALIZING TWITTER. *Journalism Studies* 13, 1 (2012), 19–36. DOI:http://dx.doi.org/10.1080/1461670X.2011.571825

Jalal Mahmud, Michelle X. Zhou, Nimrod Megiddo, Jeffrey Nichols, and Clemens Drews. 2013. Recommending Targeted Strangers from Whom to Solicit Information on Social Media. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 37–48. DOI:http://dx.doi.org/10.1145/2449396.2449403

Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. 2011. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems (CHI '11). ACM, New York, NY, USA, 227–236. DOI:http://dx.doi.org/10.1145/1978942.1978975

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

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

J. O'Donovan, B. Gretarsson, S. Bostandjiev, T. Hollerer, and B. Smyth. 2009. A Visual Interface for Social Information Filtering. In *Proceedings of the 2009 IEEE International Conference on Computational Science and Engineering, 2009. CSE '09.*, Vol. 4. 74–81. DOI:http://dx.doi.org/10.1109/CSE.2009.26

John O'Donovan, Byungkyu Kang, Greg Meyer, Tobias Höllerer, and Sibel Adalı. 2012. Credibility in Context: An Analysis of Feature Distributions in Twitter. In *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT), and Proceedings in the 2012 International Conference on Social Computing (SocialCom)*. IEEE, 293–301. DOI:http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.128

Miguel Rios and Jimmy Lin. 2013. Visualizing the "Pulse" of World Cities on Twitter. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, Cambridge, MA, USA, July 8-11, 2013*.

Sujoy Sikdar, Byungkyu Kang, John O'Donovan, Tobias Höllerer, and Sibel Adalı. 2013. Understanding Information Credibility on Twitter. In *Proceedings of the 2013 International Conference on Social Computing (SocialCom)*. IEEE, 19–24. DOI:http://dx.doi.org/10.1109/SocialCom.2013.9

Kate Starbird and Jeannie Stamberger. 2010. Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting. (2010).

Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 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 (CHI '10)*. ACM, New York, NY, USA, 1079–1088. DOI:http://dx.doi.org/10.1145/1753326.1753486

C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier. 2012. It's Not in Their Tweets: Modeling Topical Expertise of Twitter Users. In *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT), and Proceedings of the 2012 International Conference on Social Computing (SocialCom)*. IEEE, 91–100. DOI:http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.30

Lars Willnat and David H. Weaver. 2014. *The American Journalist in the Digital Age*. Technical Report. School of Journalism, Indiana University.

Jiang Yang, Scott Counts, Meredith Ringel Morris, and Aaron Hoff. 2013. Microblog Credibility Perceptions: Comparing the USA and China. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 575–586. DOI:http://dx.doi.org/10.1145/2441776.2441841