Manipulating the Perception of Credibility in Refugee Related Social Media Posts

Johannes Aigner Amelie Durchardt Thiemo Kersting

Markus Kattenbeck David Elsweiler

Chair for Information Science
University of Regensburg
D-93040 Regensburg
{johannes.aigner,amelie.durchardt,thiemo.kersting}@student.ur.de
{markus.kattenbeck,david.elsweiler}@ur.de

ABSTRACT

This paper describes a controlled web-based study (n=126), investigating whether the perception of the credibility of refugee-related Tweets can be influenced by cues already reported in the literature for social media content generally. We provide empirical evidence that both a Tweet's popularity and the presence of links – even neutral links created by URL shortening services – may increase a user's belief that the Tweet contains credible information. This is important because the propagation of false information relating to refugees on social media sites has been well documented.

Keywords

false information, rumours, perception of information, social media

1. INTRODUCTION

Social media has become a prominent information source for breaking news [18], monitoring events [27], as well as sourcing expert opinions [23] and gossip [27]. Community driven, real time posts mean that this kind of media provides quick access to content inaccessible from other sources [10]. Unfortunately, the way social media content is generated means that not all of the information shared is accurate, representative or complete, with literature reporting large proportions of posts to contain "mindless babble" [16], or include spam [2], rumours [3] or even willful deception [6]. Social media can be used to form and shape user opinions or beliefs altering their viewpoints on a variety of topics, including how brands are perceived [24] and the political decisions people take [17].

This is a prominent issue in the context of the refugee situation in western Europe where extensive propagation of false information has been well documented [28, 7]. Research has demonstrated how information presented about refugees can impact on how refugees are perceived, as well as people's emotions and attitudes towards them [11]. False beliefs can, in turn, alter how people behave towards refugees [25].

Whereas past work has investigated the features of social media posts influencing user credibility judgements generally, in this work we focus on judgements about posts regarding refugees by social-network users located in the South of Germany, an area which has recently experienced a large influx of refugees. By means of a controlled study we show that simple and subtle Tweet manipulations can statistically alter how credible readers perceive Tweets to be.

CHIIR '17, March 07 - 11, 2017, Oslo, Norway

ISBN 978-1-4503-4677-1/17/03…$15.00

DOI: http://dx.doi.org/10.1145/3020165.3022137

2. RELATED WORK

Users of social-networks have several means by which they can attain information via these services. Information can be sourced by browsing the timeline i.e. via posts made by contacts in their network [21], by actively searching the database [10], by accessing social media content via search engines [27] or by posting questions for contacts in their network to answer [23].

People are known to be more likely to accept information aligning with their existing viewpoints (confirmation bias) [29] or actions they have taken (selective exposure) [13]. Evidence suggests that the Internet is trusted as a news-source at least as much as traditional media [12], although blogs seem to be viewed as a less credible source [1]. Web page credibility and factors influencing it have been well studied with visual appearance [20], the means of locating the web page [15] and characteristics of the user herself [12] all being shown to be important. More recently similar studies have been published for social media content. Castillo et al. performed a series of investigations with the aim of building classifiers for Tweet credibility [4, 14]. Their work has identified several features of Tweets and Tweet authors, which provide insight into how users will perceive credibility. For example, reputation is important (i.e. who the user is and how well he or she is connected in the network), how a Tweet has been propagated is a cue used (i.e. the number of times a Tweet has been retweeted), as are content based features. Tweets which do not include URLs tend to be judged as

non-credible, whereas tweets containing negative sentiment words are judged as credible. Insights such as these were able to be used to create a practical and helpful real-time system [14].

Morris et al. [22] complemented this work taking a mixed-methods approach. In a first step they collected qualitative observation data whereby participants described the features of Tweets they felt influenced how the Tweet was perceived. The language used (abbreviated language was deemed untrustworthy), the author's account image, as well as the author's connectedness were all important cues. Follow-up experiments in a controlled setting showed that users are poor judges of credibility and are often biased by information like username. Similarly, topical biases exist with Tweets about science receiving a higher mean credibility rating than those about other topics. This finding shows why it is important to study Tweet perception in specific topics of interest, such as the current refugee crisis.

Elsweiler [9] performed a short survey investigating Facebook users' perception of the prominence of false information in Facebook posts. After browsing their Facebook timelines only about one third of respondents did not believe any of the attended to posts contained false information. Many of the posts respondents believed to contain inaccurate information were political, with right and left-wing comments about the refugee crisis being prominent. Two further notable findings from the survey were that 1) a large percentage of posts perceived to contain false information came from friends and 2) the most common source of false information was an external link contained within the post. Neither of these findings align well with the findings of previous work, which shows that people tend to prefer sourcing information from people they know [23] and where links made social media posts seem more credible [4, 22].

3. EXPERIMENTAL DESIGN

We wish to better understand the impact of various Tweet features on the credibility judgements of refugee Tweets. To this end, we conducted an online experiment similar to [22] whereby several properties of Tweets were altered in order to measure their impact on users' credibility assessments.

The study manipulated 6 Tweet properties, which the literature and our own intuition suggested might influence how Tweets about refugees are perceived (see Table 2). We utilised a within-groups design whereby participants each provided 12 (2×6) credibility judgments. One judgment per Tweet (both with feature and without) was collected for each manipulation strategy. To control learning effects and participant fatigue, as well as to ensure no specific strategy was advantaged due to ordering effects, the Tweets judged were rotated in a balanced manner in blocks of 6 participants (see Table 1). The Tweets shown for a condition-strategy combination were chosen at random from a large pool (described below). Participants were shown each Tweet in isolation and provided credibility ratings on a visual analogue scale with the poles being "not credible at all" (0) to "completely credible" (100). No participant was shown the same Tweet for more than one strategy.

Participants. 126 participants (55 female, $M(age) = 26.45$ years, $range = 15 - 55$ years) were recruited via a social-networking marketing campaign and email mailing lists. The recruitment process led to the enlisting of participants living in the South of Germany which reflects our target population well. All of the participants were social-network users with 70% reporting using social networks on a daily basis.

Part.	Strategy-Condition Order	
1	A1 - B1 - C1 - D1 - E1 - F1	A2 - B2 - C2 - D2 - E2 - F2
2	B2 - C2 - D2 - E2 - F2 - A2	B1 - C1 - D1 - E1 - F1 - A1
3	C1 - D1 - E1 - F1 - A1 - B1	C2 - D2 - E2 - F2 - A2 - B2
4	D2 - E2 - F2 - A2 - B2 - C2	D1 - E1 - F1 - A1 - B1 - C1
5	E1 - F1 - A1 - B1 - C1 - D1	E2 - F2 - A2 - B2 - C2 - D2
6	F2 - A2 - B2 - C2 - D2 - E2	F1 - A1 - B1 - C1 - D1 - E1

Table 1: The ordering of strategy(A-F)-condition(1/2) in blocks of six variants.

Sourcing Tweets. We sourced 18 Tweets according to two main criteria: They had to be 1) topically related to refugees in Germany and 2) contain a statement, which was factually incorrect (we checked this manually), but plausible. The idea here was that content provoking doubt would mean that participants rely more on the cues we were manipulating to make their judgements. Moreover, the selection process was performed in such a way that the statements contained in Tweets were balanced in terms of whether they were positive or negative about refugees. Examples of positive Tweets included news that a group of refugees helped in response to a flood and that refugees and local students had been playing football together and enjoying each other's company. Negative Tweets encompassed reports of violence caused by refugees and aggression against people volunteering to help with the refugees situation.

Manipulating Tweets. Each Tweet was manually edited for every strategy (see Table 2 for an overview) so that a pool of Tweets existed with a version of the Tweet for the strategy and a version without. *Popularity* (A) was induced by assigning high re-tweet and favourites to a Tweet. *Profile pictures* (B) were chosen to be non-controversial images (9 male, 9 female), Tweets in condition 2 featured the default Twitter icon. Tweets in the *Verified Account* (C) strategy featured the Twitter verified account badge, which lets users know that an account of public interest is authentic. This is a verification that the Tweet author is who she says she is and is endorsing the content by not making any attempt to anonymise the post. The *Link* (D) strategy modified Tweets with links in the form of a shortened URL-service. This avoided gaining any reliability clues from a domain name. *Emoticons* (E) were hypothesized to be a negative credibility cue and thus condition 1 was without and condition 2 contained at least one smiley. *Retweeted*(F) Tweets were framed as Twitter does when a Tweet enters a user's timeline because it was retweeted by someone the user follows. The pool consisted of 108 (18×6) tweet-pairs in total.

4. RESULTS

Figure 1 offers means and error bars showing how judgements varied across strategies for each condition. We conducted a two-way repeated measures ANOVA based on the two aforementioned factors (strategy and condition) using *GNU R* [26] and its package *ez* [19]. The results show a significant main effect of condition ($F(1, 125) = 19.07$,$p =$

Letter	Strategy	Condition 1 changes	Condition 2
A	Popularity	Add large number of retweets and likes.	low/ no RT / likes
B	Profile Pic	Replace default icon by picture of person	default icon
C	Account	Add "Verified Account" next to nickname	not added
D	Link	Add a shortened URL at the bottom of tweet.	no link
E	Emoticons	No emoticons present	emoticons ≥ 1
F	Retweeted	Frame tweet with retweeting person	no frame

Table 2: A description of changes made in each strategy for each condition. Please note: Condition 2 means no cue present.

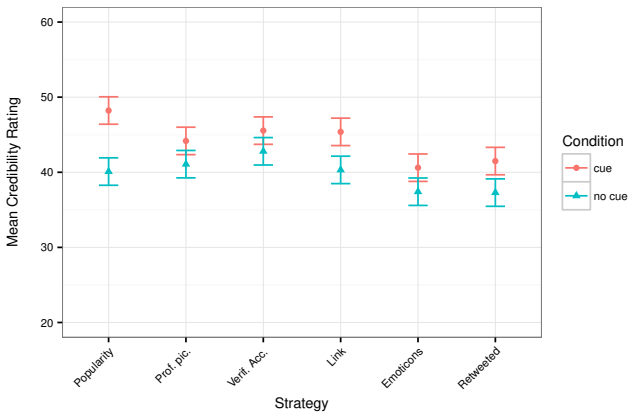

Figure 1: The means and error bars for each condition per strategy.

$2.6e^{-5}$, $\eta^2 = 0.005$), but no significant main effect for strategy or the interaction between these. This shows that it is possible to alter the perceived credibility of Tweets using the manipulations we tested. Pairwise t-tests of conditions per strategy, however, reveal that scores differed only for two strategies (the significance level $\alpha = 0.05$ was Bonferroni-corrected, see [8]). Popular Tweets were given a higher credibility ($M = 48.22$, $SD = 30.75$) than unpopular ones ($M = 40.10$,$SD = 31.02$), $t(125) = 3.56$, $p = 0.00026$, $r = 0.30$. Similarly, Tweets containing a link ($M = 45.38$, $SD = 31.54$) appear to be more trustworthy than those which do not ($M = 40.32$,$SD = 30.48$), $t(125) = 2.48$, $p = 0.00718$, $r = 0.22$. These figures represent a medium sized effect for popularity and a small effect for links according to Cohen's classification [5].

5. DISCUSSION

Our results confirm that it is possible with relatively simple and subtle changes to a Tweet's presentation, to manipulate how the credibility of Tweets relating to the refugee situation in Germany are perceived. While these findings are not surprising given previously published results, they are nevertheless disturbing: Popularity (represented by like and retweet count) does take some effort to manipulate, but it is possible and even likely when one examines the number of likes and retweets factually dubious Tweets receive.

Links created via shortened URL services can, on the other hand, be added with little effort and our findings show these to have a positive effect on user credibility judgements. It is difficult to explain why a neutral shortened external link, which offers no clue as to the source, would increase the credibility-perception of a Tweet. Perhaps this is an automatic, sub-conscious reaction in response to past experiences with social media.

Although reported in earlier studies, we did not find significant effects for profile pic, verified account, and retweets (which also have to do with popularity). This could be because we restricted Tweets to a single, specific topic and these features have less of an influence in this domain. It may also be explained by the presence of other cues in the raw Tweets, for which we did not control (i.e. cues not previously reported in the literature). Examining the judgements for unaltered Tweets (condition 2) shows high variability in the judgements between Tweets. This suggests that other factors may be involved and we plan to examine these Tweets qualitatively to establish other cues, which may be playing a role and can be investigated in future studies. We also observed variability in the judgements for individual Tweets across users. This could mean that individual differences, such as personality, political persuasion or topical knowledge might influence judgements. Future work is required to investigate both intra- and inter-personal factors more carefully. The Tweets judged in our study were all factually incorrect. In future studies we will balance factually incorrect and correct Tweets to examine whether this firstly influences the judgements and secondly to determine how able users are to establish factual accuracy.

The findings should serve as a warning sign both for Twitter users - who should perhaps be made aware of their limitations in judging how credible social media posts are - and for social-network service providers and others such as search engines who distribute content. One option which may wish to be explored based on our findings and the findings of others would be to derive more robust and objective measures of credibility and communicate this to users explicitly.

6. CONCLUSION

In this paper we have presented an initial study investigating user perception of Tweet credibility in the context of Tweets about refugees. We showed empirically that the perception of credibility can be increased by the fact that a URL is present in the Tweet or that the Tweet has been liked or retweeted often by other users. We discussed the findings and suggested other aspects, which could be studied in future work.

7. REFERENCES

[1] *Internet overtakes newspapers as news outlet. http://pewresearch.org/pubs/1066/internetovertakes-newspapers-as-news-source. last accessed on 21.9.2016 2008.*

[2] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida, *Detecting spammers on twitter*, Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), vol. 6, 2010, p. 12.

[3] Alex Burns and Ben Eltham, *Twitter free iran: an evaluation of twitter's role in public diplomacy and information operations in iran's 2009 election crisis*, Record of the Communications Policy & Research Forum (CPRF) 2009, Network Insight Institute, 2009, pp. 298–310.

[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete, *Information credibility on twitter*, Proceedings of the 20th international conference on World wide web, ACM, 2011, pp. 675–684.

[5] Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Erlbaum Associates, 1988.

[6] David M Cook, Benjamin Waugh, Maldini Abdipanah, Omid Hashemi, and Shaquille Abdul Rahman, *Twitter deception and influence: Issues of identity, slacktivism, and puppetry*, Journal of Information Warfare **13** (2014), no. 1.

[7] Lizzie Dearden, *The fake refugee images that are being used to distort public opinion on asylum seekers*, Independent Newspaper online accessed via http://www.independent.co.uk/news/world/europe/the-fake-refugee-images-that-are-being-used-to-distort-public-opinion-on-asylum-seekers-10503703.html, last accessed 11.09.2016 (2015).

[8] Olive Jean Dunn, *Multiple comparisons among means*, Journal of the American Statistical Association **56** (1961), no. 293, 52–64.

[9] David Elsweiler, *False information in social-media platforms, luddite-technologist accessible via http://davidelsweiler.blogspot.de/2016/01/false-information-in-social-media.html, last accessed 15.09.2016*, (2016).

[10] David Elsweiler and Morgan Harvey, *Engaging and maintaining a sense of being informed: Understanding the tasks motivating twitter search*, Journal of the Association for Information Science and Technology **66** (2015), no. 2, 264–281.

[11] Victoria M Esses, Scott Veenvliet, Gordon Hodson, and Ljiljana Mihic, *Justice, morality, and the dehumanization of refugees*, Social Justice Research **21** (2008), no. 1, 4–25.

[12] Andrew J Flanagin and Miriam J Metzger, *Perceptions of internet information credibility*, Journalism & Mass Communication Quarterly **77** (2000), no. 3, 515–540.

[13] Dieter Frey, *Recent research on selective exposure to information*, Advances in Experimental Social Psychology **19** (1986), 41–80.

[14] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier, *Tweetcred: A real-time web-based system for assessing credibility of content on twitter*, Proc. 6th International Conference on Social Informatics (SocInfo). Barcelona, Spain, 2014.

[15] Eszter Hargittai, Lindsay Fullerton, Ericka Menchen-Trevino, and Kristin Yates Thomas, *Trust online: Young adults' evaluation of web content*, International Journal of Communication **4** (2010), 27.

[16] Jonathan Hurlock and Max L Wilson, *Searching twitter: Separating the tweet from the chaff.*, ICWSM, 2011, pp. 161–168.

[17] Matthew James Kushin and Masahiro Yamamoto, *Did social media really matter? College students' use of online media and political decision making in the 2008 election*, Mass Communication and Society **13** (2010), no. 5, 608–630.

[18] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, *What is twitter, a social network or a news media?*, Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 591–600.

[19] Michael A. Lawrence, *ez: Easy analysis and visualization of factorial experiments*, 2015, R package version 4.3.

[20] Jonathan Lazar, Gabriele Meiselwitz, and Jinjuan Feng, *Understanding web credibility: a synthesis of the research literature*, Now Publishers Inc, 2007.

[21] Florian Meier and David Elsweiler, *Going back in time: An investigation of social media re-finding*, Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, ACM, 2016, pp. 355–364.

[22] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz, *Tweeting is believing?: understanding microblog credibility perceptions*, Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, ACM, 2012, pp. 441–450.

[23] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich, *What do people ask their social networks, and why?: a survey study of status message q&a behavior*, Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 2010, pp. 1739–1748.

[24] Rebecca Walker Naylor, Cait Poynor Lamberton, and Patricia M West, *Beyond the "like" button: The impact of mere virtual presence on brand evaluations and purchase intentions in social media settings*, Journal of Marketing **76** (2012), no. 6, 105–120.

[25] Anne Pedersen, Susan Watt, and Susan Hansen, *The role of false beliefs in the community's and the federal government's attitudes toward australian asylum seekers*, Australian Journal of Social Issues **41** (2006), no. 1, 105.

[26] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016.

[27] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris, *# twittersearch: a comparison of microblog search and web search*, Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 35–44.

[28] Renate van der Zee, *Rumours and lies: 'the refugee crisis is an information crisis'*, Guardian Online accessed via https://www.theguardian.com/global-development-professionals-network/2016/aug/18/rumours-and-lies-the-refugee-crisis-is-an-information-crisis, last accessed 11.09.2016 (2016).

[29] Ryen White, *Beliefs and biases in web search*, Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, ACM, 2013, pp. 3–12.