The Impact of Twitter Features on Credibility Ratings - An Explorative Examination Combining Psychological Measurements and Feature Based Selection Methods

Judith Meinert
University Duisburg-Essen
judith.meinert@uni-due.de

Ahmet Aker
University Duisburg-Essen
a.aker@is.inf.uni-due.de

Nicole C. Krämer
University Duisburg-Essen
nicole.kraemer@uni-due.de

Abstract

In a post-truth age determined by Social Media channels providing large amounts of information of questionable credibility while at the same time people increasingly tend to rely on online information, the ability to detect whether content is believable is developing into an important challenge. Most of the work in that field suggested automated approaches to perform binary classification to determine information veracity. Recipients´ perspectives and multidimensional psychological credibility measurements have rarely been considered. To fill this gap and gain more insights into the impact of a tweet´s features on perceived credibility, we conducted a survey asking participants (N=2626) to rate the credibility of crisis-related tweets. The resulting 24.823 ratings were used for an explorative feature selection analysis revealing that mostly meta-related features like the number of followers of the author, the count of tweets produced and the ratio of tweet number and days since account creation affect credibility judgments.

1. Introduction

Based on the rise of Social Media, online communication has changed fundamentally within the last years. Nowadays, every single user has not only the opportunity to consume content, but also to produce and distribute information [1], [2]. Social Media channels generally bear great potential for users to receive information faster, to connect to people around the world or to public persons, brands, parties and organizations. As an ongoing tendency, users tend to use Social Media not only for private communication purposes, but also as a source of news and political information [3], [4]. On the other hand, public persons like politicians are able to use Social Media as "privately owned publicity channel" [5, p. 40] to directly and reciprocally communicate with potential voters, share and explain political actions and projects without being dependent on mass media. Even organizations and media personas like journalists and mass media journals utilize the new channels to distribute information to the public in a fast-pacing manner [6], for example, in cases of high uncertainty as in crisis situations or during extreme events. Due to contextual factors like real-time communication, short messages and a high distribution, particularly Twitter is predestined for consuming and producing breaking news, political content and updates of emergency communication as well as current events as soon as they happen [7], [8], [9].

However, the opportunity of real-time communication reaching a wide audience within a few seconds are countered by a lack of gatekeepers, filtering options or control for quality standards [1], [10], which raises the question of how credible the published content is. Particularly the area of political communication and news in Social Media recently developed into an environment influenced by distrust, deception and strategically deployed misinformation to reach manipulative, political or financial aims. Especially since the 2016 US presidential election campaign, the term "Fake News" is on everyone's lips and the distribution of false information discrediting presidential candidate Hillary Clinton was assumed to have an actual impact on the election results [10].

Besides intentionally spreading misinformation, accidental errors of reporting occur, especially because news magazines tend to invest less effort in fact checking for their online dissemination of information than for their offline publications [1]. Aggravating this issue, news consumption nowadays often takes place through Social Media without people doublechecking information in traditional media [11], [12].

While the credibility of online information is an almost-universal topic in both media and research, recipient's perspective of how users assess credibility is still understudied. A large body of work either focusses on technical solutions to increase the accuracy of

URI: https://hdl.handle.net/10125/59698
ISBN: 978-0-9981331-2-6

predictions through the comparison of models which are designed to detect the veracity of facts [4], [13], [14], [15] or investigate isolated aspects potentially influencing perceived credibility in lab scenarios with self-reported measurements [16], [17], [18]. This is not to say that research on the veracity of facts is not timely or important, however, we decided to focus on perceived credibility here, as the corresponding psychological mechanisms are understudied.

Therefore, in the current study, we aim to combine psychological credibility assessments with a feature based selection approach in the realm of recipients' Twitter ratings, to overcome boundaries of approaches that solely target veracity and contribute to an understanding of human credibility perception in Social Media environments. To this aim, we analyze the features which impact if a tweet will be perceived as credible or not. This knowledge on human perception can be applied for future applications in terms of interface design and content presentation as well as for user education concerning media competence through highlighting relevant features.

2. Online credibility

In an online environment without any gatekeepers, filtering options or quality control, the importance and the difficulty of valid credibility assessments increases [2]. Yet, the process of credibility assessments in Social Media is not sufficiently understood. There is, for example, only sparse knowledge on which features people base their credibility judgments, which is also be owed to the "dizzying array of credibility cues to choose from" [2, p. 449].

In general, online credibility judgments are said to be more complex than interpersonal evaluations due to the various technological aspects influencing the reception situation [19]. Especially, Twitter provides communication characteristics like a high connectivity of users and fast distribution of information whereby mentioning and referencing each other are common conventions. Caused in that, further research is needed to understand Twitter communication features and their role in the credibility assessment process [19], [20].

Basically, credibility is described as believability of source and message [2]. A crucial factor in this regard is the assumption that credibility is a perceptual variable which is rather subjectively perceived by recipients than objectively attributable [19]. Early research on persuasion defined trustworthiness and expertise of the communicator as key dimensions of credibility [21] which was later extended by further aspects like goodwill [22].

However, most of the research in the field is based on the concept of veracity which refers to a binary distinction of content in true or false [4]. If information can be proven, the message is true, and if not, it will be indicated as wrong or fake. In particular, this is successfully used for classifying content with the core task of identifying the veracity of messages [13] and ensuring the accuracy of online information like news articles based on fact checking methods [1].

But, even if we get to learn which approaches are performing in the best way to eliminate inaccurate online information, binary judgments are not a realistic and applicable representation when it comes to human perceptions and ratings in a reception situation characterized by uncertainty and a fast speed of information [18], [22]. For instance, even if content like satire and parody do not intentionally deceive recipients, it might nonetheless happen, because the content is not clearly identifiable as true or false. As a result, the recipient of the information must be considered as an influencing factor of how information is processed and perceived [2], [15], [18], [19]. In this sense, Wassmer and Eastman [23] differentiate between actual and perceived credibility, whereby actual credibility can be equaled with veracity.

In contrast, we focus on credibility as a multidimensional construct which mainly relies on perceptions of how believable, accurate and trustworthy an information or source is [2]. Until now, perceived credibility of online content is often measured with a single-item question [14], [24], which could be broadened by using multidimensional scales assessing if different aspects of perceived credibility are related to different features or cues.

With the aim of avoiding a gap between system-based measures and human ratings, we aimed to consider users' perceptions in the process of content evaluation. Considering that "message credibility is an individual's judgment of the veracity of the content of communication" [25, p. 63], we want to take an expanded look at the concept of credibility including multidimensional perceptions which seems to be promising to get more insights into people´s actual real-life evaluations of Twitter communication.

3. Challenges to credibility assessments

Confronted with a huge and potentially unlimited amount of information but limited processing capacities [26], users are not always able to examine the credibility for every piece of information in an elaborated way [25]. According to dual process models like the Elaboration Likelihood Model [27] and the Heuristic Systematic Model [28], impressions can be formed through two

different ways of information processing which will be chosen depending on recipients' motivation and ability to process information thoroughly. Thus, the likelihood to scrutinize any given information via the central route is increased for recipients with high involvement or higher need for cognition. In contrast, the peripheral route describes a simplified processing which is based on peripheral cues or heuristic rules. This route is taken when a person is neither willing nor able to process the information in an elaborated way.

Social Media communication in general was found to be processed in a more peripheral way [29]. Due to these contextual preconditions the likelihood to rely on cognitive heuristics for effort reduction is increased for Social Media users, especially in judgment situations under uncertainty [30], [31], [32].

Cognitive heuristics are strategies that do not include all available information in order to minimize the cognitive load [32], are mainly unaware and can (but not necessarily do) lead to biased judgments. Metzger and colleagues [30] already investigated the operation of heuristics in Social Media using focus groups and defined different heuristics used by recipients for credibility ratings. However, considering that the process of heuristic judging takes place automatically, using focus groups or self-reports does not seem to be fully efficient.

Another question reflects on the anchors taken from Social Media communication and used for judging the credibility of content. Since in Social Media no constraints for the publication of content exist, the reliance of information can only be attributed based on implicit factors [33], particularly if recipients are uncertain about the communication source, events and context. Heuristic judgments are found to be triggered by specific aspects of the message, author or interaction situation [34], but which cues or features are potentially able to effect credibility ratings of content, for instance in Twitter communication, is still under investigation.

Nevertheless, some findings regarding the credibility-enhancing effects of Social Media cues or features have been presented. With regard to source-related cues, it has been shown that a communicator who is presented as competent and an expert in the target field, leads to increased credibility perceptions. This effect is described as authority heuristic [31] or reputation heuristic [30] and was demonstrated to be an important factor for the selection of online news articles [35]. Moreover, in Social Media environments recipients tend to be guided by a simple heuristic rule described as bandwagon heuristic, "If others think that something is good, then I should, too", [31, p. 83] which was already found to be influential for ratings and reviews in e-commerce [36].

4. Related work

One of the most defining characteristics of Social Media applications refers to the huge amount of available data [2], [15]. To deal with large data sets, many researchers put the lens on the development of system-based approaches, models or algorithms for efficiently detecting the truth value of information [4], [15], [18].

For instance, Derczynski and colleagues [4] designed a model to identify rumors in online information, defined as unverified information spread through Social Media [37], by integrating the reactions of the community. In this regard, retweets were classified into supporting, denying, querying and commenting. These community interaction patterns turned out to be efficient which supports the relevance of including recipients and their reactions and perceptions into the evaluation of online content.

Further approaches consider user profile meta data like location and topicality of posting behavior to make a prediction of how accurate the author is communicating [38]. Comparing the similarity of words and facts with web content from the same topic domain, is introduced as another possible system-based approach for detecting the veracity of published content [33].

To predict the usefulness of online reviews, Levi and Mokryn [40] evaluated if integrated sentiment, review length and reviewer status are influential factors in four different data sets from Yelp, Amazon and IMDb using a supervised learning paradigm based on a binary classification model. Particularly, the expression of disgusting emotions as well as the number of punctuations and question marks in reviews determine perceived usefulness whereas the number of used adjectives decreased perceived usefulness. Furthermore, the status of the reviewer (e.g. displayed with a badge or 'Top reviewer' label) was found to be an influential feature. Here, the authors reasoned that reviewers who wrote many reviews were perceived as more familiar which further evokes trust. Additionally, content which was perceived as interesting and evoking positive feelings, was shared more often on Twitter, so that interest and sentiment probably serve as indicators for content distribution [41].

Overall, a lot of proposed models to verify online information exist, based on semantic web technologies, external source checking, extracting and highlighting the reputation and experience of the source, comparing information to facts on formal websites or applying symmetry in textual and temporal features as well as data similarity [13]. Scholars already started to compare and rank different models in terms of prediction rates and accuracy. However, a common feature of all models is that they put major effort into the identification of the

"correct value of a fact" [13, p. 228] with a view to providing valid fact checking measures to the users. The recipients' perception of the credibility of information by an empirical investigation is not considered.

Apart from that, only some user studies deal with features coming along with a tweet and its effect on users' evaluation of credible information. Zubiaga and Ji [18] investigated in a controlled experimental setting how factors like authority and plausibility of the message, presentation and corroboration of the tweet affect the probability of how accurate users can rate if a tweet is true or not. They found that information about the communicator like the number of followers, the location and the description in the Twitter profile, mostly leads to higher accuracy of the ratings. However, in the experiment all features were handled in an isolated way, so that assumptions about relations are difficult to make. Additionally, the user test did not ask users how credible they perceive the tweets. Despite that, the authors assume that features like the number of followers and followees as well as location and description of the account holder could possibly support users in assessing believability in a more valid way.

Accordingly, the perception of credibility is assumed to be associated with Twitter features [39]. Particularly, the number of included hashtags, the length of the message, a user mentioned in the tweet, the number of received retweets and if affect is included were found to be influential when users had to estimate the credibility of tweets. A study [8] directly asking users to indicate which features they rely on to rate tweets as credible revealed that an included link, hashtags, retweets, user mentions and the displayed account name influence credibility assessments. Words like 'update' or 'breaking' seemed to serve as credibility-increasing keywords. A further result refers to the finding that even non-objectively observable features might have an effect on credibility ratings. Participants mentioned being influenced by the attitude of the communicator towards the tweet topic which they implicitly derived from words like 'plausible' or 'fact'.

With regard to users' evaluation another study [17] explored users to be generally poor in assessing credibility ratings on Twitter data, independent of the individual level of experience. In addition, the tweet topic was found to be an influential factor with science tweets receiving generally higher credibility ratings than political postings. Regarding the reliance on Twitter features, first, 26 features were selected via think aloud user tests and subsequently, participants had to indicate to which degree they use them for assessing the credibility of tweets. Above all, features related to the author of the tweet such as follower and retweet number, twitter account description, location as well as a Twitter verification symbol resulted in enhanced credibility evaluations. Moreover, tweets including an URL reached higher levels of attributed credibility.

An interesting finding is presented by Aigner and colleagues [16] who conducted a study focusing on how recipients evaluate the believability of news on twitter in the area of refugee related information. They demonstrated that tweets were rated as more credible if they received a higher number of retweets and likes, and that this is even true if the tweets were factually false. As already reported by Morris and colleagues [17], tweets with URL links received higher credibility assessments.

Broadly speaking, Twitter features like author-related, message-related and meta information-related aspects seem to have an influence on users' assessment of the content's credibility, but to date the majority of user studies is based on self-reported data which involves the risk of biased responses concerning suggestions or social desirability. Accordingly, findings of user studies differ somehow, which can be due to different topics and contexts as well as biased user reports. Altogether, the results of the user experiments using artificially varied feature sets should be transferred to a real-life setting to make reliable conclusions. In addition, the evaluation of credibility is often assessed by a single item which can be criticized as not addressing the multidimensionality of perceived credibility in an extensive and fully sufficient way thereby limiting the results. From a methodological viewpoint a diminished reliability of the credibility measurement needs to be considered. All in all, the relation between Twitter features and credibility assessments by recipients needs more systematic and controlled consideration.

Therefore, the present research aims to investigate the impact of Twitter features on users' credibility ratings in a more comprehensive and large-scale way. In this respect, we combined a multidimensional measurement of credibility with an automated feature selection approach to avoid both boundaries of a limited reliability through one-dimensionality and self-reported effects. With our present study, we want to address the following research questions:

RQ1: Which features of Twitter communications affect credibility ratings of recipients?

RQ2: Are different dimensions of message credibility affected by different features?

5. Method

To exploratively investigate which features of a tweet are influencing credibility assessments, we set up an online survey using Figure Eight (https://make.figure-eight.com), a crowdsourcing platform for data annotations and ratings. By using Figure Eight we were able to recruit a large sample consisting of older and more diverse participants compared to common undergraduate samples [42], [43]. Crowd working platforms like Figure Eight are widely and successfully used, especially for tasks with rating or labelling content [44], [45], [46]. To ensure data quality we also asked participants to add an explanatory sentence to their ratings like it was recommended by [14], [47]. Additionally, the platform offers the option to directly embed a huge number of tweets (see figure 1 for an example).

INSERT FIGURE 1 ABOUT HERE

After viewing a tweet, participants were asked to rate its credibility. To overcome boundaries of a binary judgment we used the message credibility scale of Appelman and Sundar [25] asking participants to indicate on a five-point Likert scale (1 = describes it very poorly to 5 = describes it very well) how accurate, authentic and believable the tweet is. We extended the scale by adding the items comprehensible, important, informative and interesting to the questionnaire (α =.94; M = 3.66; SD = 0.89). Like it is described in chapter 2, credibility is a perceptual variable related to trustworthiness, competence [21] and goodwill [22] of the communicator. While the items believable and authentic refer to the trustworthiness dimension, we aim to strengthen competence evaluations (already tackled with the item accurate) with adding the items comprehensible and informative. To assess communicators´ goodwill in the area of event-related Twitter communication we included the items important and interesting.

In addition, we assessed participants' gender, age and educational background. Participants had the possibility to rate as many tweets as they wanted up to a maximum of 40 and they received a fee of $0.02 for every rating.

Data set. The tweets were selected from a publicly available data set provided by Zubiaga and colleagues [48] consisting of real Twitter data tracked during five different crisis events (Charlie Hebdo, Ferguson shooting, Germanwings crash, Ottawa shooting, Sydney siege) and collected from the Twitter streaming API which were manually annotated by journalists to consist either of rumors or non-rumors. We only used source tweets (no retweets) to avoid redundant content. In sum, 828 tweets were evaluated, with every tweet being rated by 30 different raters. Due to technical reasons a few ratings had to be excluded, resulting in a total number of 24.823 ratings.

Sample. 2626 persons older than 18 years participated in the online survey. The sample had a mean age of 33.94 (SD = 10.93) years; 930 participants were female and 1696 were male. Most of them were employees (1264 participants), 945 participants were self-employed and 417 students.

Feature selection. In our analysis we aimed to include author-, message- and meta-informational features, whereas especially meta-related aspects are relatively understudied until now [52]. Author-related features refer to aspects of the account holder e.g. the length of the authors' Twitter account description, message features describe information related to the tweet's text, for instance if it is containing a URL, and meta-informational features include aspects like the number of followers.

In total we included the following features, which turned out to be useful in prior research in the area of stance detection in Twitter communication [53] and were already annotated in the data set: for the author-related features we used authors Twitter account description, length of the account description, and role (refers to the relation between follower and followee number), for the message-related features we took URL included, location included, person included, date included, negation included, Google bad word included (using a dictionary from Google to check if the tweet contains slang words), geo information enabled, average word length, and for the meta-informational features we comprised originality (refers to the number of tweets of a user), number of followers, engagement (refers to the number of tweets related to user account age) and sentiment (describes on a scale ranging from positive to negative the valence of the tweet with an assigned value between 0 and 4).

To analyze what features users associate with credibility, we automatically extracted several features and tested their relevance against the responses the raters gave for each assessment type. The responses were given on a five-point Likert scale to improve the representation of the credibility perception and avoid forcing raters to put their answers in categories, however for the classification needed for the relevance computation, we collapsed the points between 1 and 3 as well as 4 and 5 together to obtain binary decisions. According to Beamish [49] collapsing responses in the way we did, has distinct advantages in terms of capturing trends in the data which is a commonly used procedure for data classification in the realm of feature selection [24], [50]. Furthermore, referring to the analysis of Grimbeek and colleagues [51], the

conversion of Likert scale responses into dichotomized categories does not result in a loss of data richness.

6. Results

In the analysis, we tested the feature significance for each of the integrated features and each item of the credibility scale using chi-squared test in the implementation provided by Python scikit-learn package [54], a method widely used for feature selection based on classification [55], By applying this method, we received a value indicating if the specific feature is a significant indicator to discriminate between the classes of low and high attributed accuracy, for example. Thereby, numbers over 3.84 describe a significant influence on a 95 percent level and values higher than 6.63 refer to a significant effect on a 99 percent level.

We found that author-related features, message-related features as well as meta-informational features seem to be influential, whereby meta-information like the number of followers, the originality (sum of all tweets produced) and the engagement (ratio between number of tweets and active days) of the tweet author seem to have the most impact. As can be derived from table 1, showing an overview of all features and their values from the feature significance test, the follower count as well as the amount of tweets a user has produced effect all seven credibility dimensions on a 99 percent level of significance. The number of followers has the highest value for rating tweets as believable, whereas originality mostly impacts the ratings of authenticity. Furthermore, the engagement of the tweet´s author, described as the ratio of number of tweets and time since the user is active, primarily determined the dimensions informative and authentic.

The length of the authors´ Twitter account description turned out to be a significant indicator for the differentiation between tweets rated as informative and interesting and tweets rated as less informative and interesting.

INSERT TABLE 1 ABOUT HERE

The assessment of believability is influenced by enabled geo-information. If a location or an organization is mentioned, this affects users' ratings of how informative a tweet is, and mentioning a person predicts accuracy perceptions as well as tweets with included negations. Tweets containing dates contribute to evaluations concerning the dimension interesting and the relation between followers and followees of the tweet´s author is connected to ratings of comprehensibility of the tweets. On the contrary, an included URL, the valence of the tweet, Google bad word indicator, the average word length and the description of the Twitter account holder did not show a significant influence on the credibility rating. This pattern of feature effects on credibility ratings was shown for all tweets of the data set, independent of whether the tweets were rumors or non-rumors. Overall, among all survey participants, there was a fair level of agreement concerning the credibility ratings (Krippendorffs' $\alpha = .38$).

7. Discussion

Social Media and Twitter in particular offer a space for producing and spreading large amounts of content. Besides the benefits of receiving information faster and consuming event-related information in real-time, recipients are confronted with the omnipresent question of how credible information is. Due to this, the relevance of valid credibility assessments enhances.

Investigating the impact of Twitter features on multidimensional credibility ratings of crisis-related tweets, which were either non-rumors or rumors, we found that credibility ratings were mainly influenced by the number of followers and the originality score which involves the total number of tweets an author has created. Both features highly impacted all measured credibility items (accurate, authentic, believable, comprehensible, important, informative and interesting). Interestingly, these features both are not visible to the user (neither in the study nor in real-life Twitter settings) but still are better predictors for perceived credibility compared to visible features such as number of words or inclusion of an URL.

Future studies need to scrutinize further by which evident cues people sense that the author has a high number of followers and has written a large number of tweets. Potentially, an author with a higher number of followers communicates in a slightly different way than someone with fewer followers – although the content-related features we assessed did not have a strong influence on credibility ratings. A person who posts a lot of tweets can be assumed to have high experience (probably including high ability to write good, convincing tweets). Similarly, someone with a large number of followers seems to be able to attract numerous people either by his/her authority or his/her tweets´ quality, both which will be recognizable to the reader.

Another possible explanation refers to results derived from former communication studies. It was found that recipients especially tend to perceive information as biased if they estimate the content to be exposed to a large audience [56]. Studies revealed that

people are apparently able to estimate the audience size of an information piece which increases the assumption that others might be convinced more easily by content with a higher reach [57].With regard to the effect of the number of followers on credibility assessments we found, it can be assumed that recipients are able to predict the potential audience size (in a Twitter context the number of followers of the authors account) through the visible information of the tweet.

Most likely, both cues are strongly connected to sensing the quality of the source - or, put differently, are the only cues in our feature list that will be strongly related to the expertise and quality of the source. This would be in line with numerous findings on the importance of the source when assessing the credibility of a message [17], [30], [31], [35]. Alternatively, the effect might be explainable by a bandwagon effect [31]. Tweets of authors with numerous followers will benefit from large amounts of likes and retweets which might also persuade readers of the quality of the posting. This is in line with results of a user study by Aigner and colleagues [16] who found that credibility ratings mainly depend on the number of retweets and likes indicating a bandwagon effect. In this way, likes and retweets can be understood as recommendations of content by other users and might be taken as an anchor for rating something as credible.

Additionally, the authors´ engagement score, the ratio between number of tweets and period the account is active, showed an effect on at least six of the credibility dimensions. This, again, is a non-visible, meta-informational aspect – which might also be related to the quality and subsequent credibility of the source. With regard to prior results showing that recipients' ratings were influenced by the implicitly derived attitude of the author [8], we can assume recipients' ability to use implicit feature information for credibility judgments. However, further investigation is needed to explore these patterns of using implicitly transmitted cues in detail. Future work will have to identify those observable cues that are used by the reader. Following our assumption that source is the relevant variable here, a necessary next step would be to come up with categorizations of different sources.

Regarding author-related features, our results showed an impact of the length of the description stored in the Twitter profile on user ratings how accurate, believable, informative and interesting tweets were perceived. However, the fact whether an author provides a description or not (feature: description) showed no impact on any credibility dimension. This differs slightly from former findings demonstrating that recipients seemed to rely on account descriptions of the author for assessing credibility. This difference probably results from the fact that recipients report to take the description into account [17], [18], whereas the length of the description is actually the decisive factor. In general, information about the author of tweets was found to determine the accuracy of tweets verification ratings [18] as well as user credibility assessments [17] which strengthen the influential impact of author-related aspects.

With regard to message-related features, several aspects turned out to be influential for different aspects of credibility. For instance, enabled geo information in the tweet relates to perceptions of authenticity and believability and the inclusion of an organization or location tends to be a discriminator between informative and not informative content. Furthermore, it was shown, that accuracy perceptions are determined by mentioning a person or including negation. The found impact of negation relates to the findings of Levi and Mokryn [40] who evaluated that especially negative sentiment in online reviews enhanced perceptions of usefulness. Tweets containing a date seem to shape the impressions whether some content is interesting or not which is in line with classic news value assumptions [58].

In contrast to former findings, our analysis showed no impact of the URL, the valence of the tweet, Google bad word indicator, the average word length and the description in the authors' profile on credibility ratings. A possible explanation for this inconsistency might be that in the user studies which explored an influence of URL, affect and user description, users indicated this tendency via questionnaires [16], [17], [39]. Due to the experimental setting solely involving and varying a few features, features probably have been more salient to the recipients. In contrast, our study confronted recipients with all features like in a real-world scenario and the impact of the features were assessed via the automatic extraction based on the categorized user ratings.

Surprisingly, no difference regarding the feature impact occurred between the rumors and non-rumors. Users obviously seem to apply the same rating mechanisms for tweets consisting of true facts and tweets with false facts. In this regard, it would be interesting to examine if the impact of the features underlies a conscious process or if it happens in a more automatic way. Also, future studies should include an explicit dichotomous rating of whether the person believes the tweet to be true or false in order to be able to not only include the objective fact of whether it is a rumor or not but also the recipients' explicit judgment on this.

An important factor to consider is the topic domain of the tweets rated in the current survey. According to Morris and colleagues [17], users tend to react differently depending on the topic of Twitter communication, for example, science related tweets did generally receive higher levels of credibility judgments.

However, we do not expect a large bias here as we took great care to include diverging topics that cover a broad range of events and emergency situations.

8. Conclusion and further work

In sum, we shed light on a wide range of Twitter features and investigated their role in the credibility judgment process. We extended the common use of binary decisions between true or false by incorporating recipients' perceptions and applying a multidimensional credibility measurement. The present findings demonstrate that especially meta-related information like the number of followers, the originality score (count of tweets a user has produced) as well as the engagement ratio (number of tweets related to the time the account is active) influence credibility ratings.

In general, we contribute to a more detailed understanding of which Twitter features play a major role in credibility ratings of online information.
Additionally, in our immediate future work, we aim to extend the set of features we analyzed to also capture non-meta-features such as network related information, tweet content as well as time dimensions.

Especially, the time of a tweet seems to be promising in having an influence, referring to the findings of Levi and Mokryn [40] revealing that the later reviews were posted, the more useful they were rated. This could possibly also emerge for the credibility of information included in a tweet, in particular in the fast-pacing context of crisis-related events.

Next steps will also include turning the results into a supervised classification problem. Since we have the manually annotated data, we can use the significant features to train a machine learning model in order to perform automatic predictions. Only if we learn more about how users assess credibility and which features contribute to this process, we will be able to efficiently support Social Media recipients with technical solutions like highlighting credibility-relevant features [18]. Therefore, we emphasize the necessity to integrate users' perceptions into the investigation to optimize methods and will contribute to this process in the future.

9. Acknowledgements

This work is supported by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media". We also thank our student assistant Birte Högden for supporting us with the data collection.

10. References

[1] Bedolla, T., & Molla, B. (2012). Credibility of sources and the veracity of content. SSRN. doi:10.2139/ssrn.1985671

[2] Metzger, M. J., & Flanagin, A. J. (2015). Psychological approaches to credibility assessment online. The handbook of the psychology of communication technology, 32, 445-466.

[3] Bode, L. (2015). Political news in the news feed: Learning politics from social media. Mass Communication and Society. Advance online publication.

[4] Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., & Zubiaga, A. (2017). SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. arXiv preprint arXiv:1704.05972.

[5] Lee, E. J., & Jang, J. W. (2011). Not so imaginary interpersonal contact with public figures on social network sites: How affiliative tendency moderates its effects. Communication Research, 40, 27-51. doi:10.1177/0093650211431579

[6] Mirbabaie, M., Ehnis, C., Stieglitz, S., & Bunker, D. (2014). Communication Roles in Public Events. In Working Conference on Information Systems and Organizations (207-218). Springer, Berlin, Heidelberg.

[7] Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT? In Proceedings of the first workshop on social media analytics (71-79). ACM.

[8] Shariff, S. M., Zhang, X., & Sanderson, M. (2014). User perception of information credibility of news on twitter. In European conference on information retrieval (513-518). Springer, Cham.

[9] Vieweg, S. (2010). Microblogged contributions to the emergency arena: Discovery, interpretation and implications. In Computer Supported Collaborative Work.

[10] Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, 31(2), 211-236. doi:10.3386/w23089

[11] Mitchell, A., Gottfried, J., & Matsa, K. E. (2015). Millennials and political news. Pew Research Center, 1. Retrieved May 6, 2018, from http://www.journalism.org/2015/06/01/millennials-political-news/

[12] Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics - Challenges in topic discovery, data collection, and data preparation. International Journal of Information Management, 39, 156-168. doi:10.1016/j.ijinfomgt.2017.12.002

[13] AlDoaies, B. H., Ashi, A. M., & Alotaibi, F. S. (2017). Exploring and evaluating the impact of the veracity of big Data sources. International Journal of Computer and Information Technology, 6(5), 227-236.

[14] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In: Proc. WWW, 675-684.

[15] Shariff, S. M., Zhang, X., & Sanderson, M. (2017). On the credibility perception of news on Twitter: Readers, topics and features. Computers in Human Behavior, 75, 785-796.

[16] Aigner, J., Durchardt, A., Kersting, T., Kattenbeck, M., & Elsweiler, D. (2017). Manipulating the Perception of Credibility in Refugee Related Social Media Posts. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (297-300). ACM.

[17] Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012,). Tweeting is believing?: Understanding Microblog credibility perceptions. In S. Poltrock, C. Simone, J. Grudin, G. Mark, & J. Riedl

(eds.), Proceedings of the ACM 2012 conference on computer supported cooperative work (441-450). ACM.

[18] Zubiaga, A., & Ji, H. (2014). Tweet, but verify: epistemic study of information verification on twitter. Social Network Analysis and Mining, 4(1), 163-175. doi:10.1007/s13278-014-0163-y

[19] Choi, W., & Stvilia, B. (2015). Web credibility assessment: Conceptualization, operationalization, variability, and models. Journal of the Association for Information Science and Technology, 66(12), 2399-2414.

[20] Schmierbach, M., & Oeldorf-Hirsch, A. (2012). A little bird told me, so I didn't believe it: Twitter, credibility, and issue perceptions. Communication Quarterly, 60(3), 317-337.

[21] Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. Public opinion quarterly, 15(4), 635-650.

[22] McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. Communications Monographs, 66(1), 90-103.

[23] Wassmer, M., & Eastman, C. M. (2005). Automatic evaluation of credibility on the Web. Proceedings of the American Society for Information Science and Technology, 42(1).

[24] Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. Internet Research, 23(5), 560-588.

[25] Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. Journalism & Mass Communication Quarterly, 93(1), 59-79. doi:10.1177/1077699015606057

[26] Lang, A. (2000). The limited capacity model of mediated message processing. Journal of communication, 50(1), 46-70. doi:10.1111/j.1460-2466.2000.tb02833.x

[27] Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In: Berkowitz, L. (ed.), Advances in Experimental Social Psychology, 19 (123-205). New York: Academic Press.

[28] Chaiken, S. (1987). The heuristic model of persuasion. In: Zanna, M.P, Olsen, J.M., & Herman, C. P. (eds.), Social influence: The Ontario symposium (3–39). Hillsdale, NJ: Erlbaum.

[29] Lee, E. J., & Shin, S. Y. (2012). Are they talking tme? Cognitive and affective effects of interactivity in politicians' Twitter communication. Cyberpsychology, Behavior, and Social Networking, 15, 515-520. doi:10.1089/cyber.2012.0228

[30] Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. Journal of communication, 60(3), 413- 439.

[31] Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In: Metzger, M.J., & Flanagin, A.J. (eds.), Digital media, youth, and credibility (73–100). Cambridge, MA: MIT Press.

[32] Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In: Utility, probability, and human decision making (141-162). Springer Netherlands.

[33] Zhao, L., Hua, T., Lu, C. T., & Chen, R. (2016). A topic-focused trust model for Twitter. Computer Communications, 76, 1-11. doi:10.1016/j.comcom.2015.08.001

[34] Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. Psychological review, 118(1), 97.

[35] Winter, S., & Krämer, N. C. (2014). A question of credibility: Effects of source cues and recommendations on information selection on news sites and blogs. The European Journal of Communication Research, 39, 435–456.

[36] Sundar, S. S., Xu, Q., & Oeldorf-Hirsch, A. (2009). Authority vs. peer: How interface cues influence users. Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI'09), 27, 4231–4236.

[37] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2017). Detection and Resolution of Rumours in Social Media: A Survey. ACM Computing Surveys. Retrieved from https://arxiv.org/pdf/1704.00656.pdf

[38] Bodnar, T., Tucker, C., Hopkinson, K., & Bilén, S. G. (2014). Increasing the veracity of event detection on social media networks through user trust modeling. In C. Aggarwal, N. Cercone, & V. Honavar (eds.), Proceedings of the 2014 IEEE International Conference on Big Data (636-643). IEEE. doi:10.1109/BigData.2014.7004286

[39] Ravikumar, S., Talamadupula, K., Balakrishnan, R., & Kambhampati, S. (2013). RAProp: ranking tweets by exploiting the tweet/user/web ecosystem and inter-tweet agreement. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (2345-2350). ACM.

[40] Levi, A., & Mokryn, O. (2014). The social aspect of voting for useful reviews. In International conference on social computing, behavioral-cultural modeling, and prediction (pp. 293-300). Springer, Cham.

[41] Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In Proceedings of the fourth ACM international conference on Web search and data mining (65-74). ACM.

[42] Follmer, D. J., Sperling, R. A., & Suen, H. K. (2017). The Role of MTurk in Education Research: Advantages, Issues, and Future Directions. Educational Researcher, 46(6), 329-334.

[43] Hirth, M., Hoßfeld, T., Mellia, M., Schwartz, C., & Lehrieder, F. (2015). Crowdsourced network measurements: Benefits and best practices. Computer Networks, 90, 85-98.

[44] Mason, W., & Suri, S. (2010). Conducting behavioral research on Amazon's Mechanical Turk. Behavioral Research Methods, 44, 1–23.

[45] Kittur, A., Chi, E. H., & Suh, B. (2008, April). Crowdsourcing user studies with Mechanical Turk. In Proceedings of the SIGCHI conference on human factors in computing systems (453-456). ACM..

[46] Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008, August). Get another label? improving data quality and data mining using multiple, noisy labelers. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (614-622). ACM.

[47] Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010, June). Annotating named entities in Twitter data with crowdsourcing. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (80-88). Association for Computational Linguistics.

[48] Zubiaga, A., Wong Sak Hoi, G., Liakata, M., & Procter, R. (2016). PHEME dataset of rumours and non-rumours. figshare. Dataset.

[49] Beamish, W. (2008). Consensus about program quality: An Australian study in early childhood special education. VDM Publishing.

[50] Olteanu, A., Peshterliev, S., Liu, X., & Aberer, K. (2013, March). Web credibility: features exploration and credibility prediction. In European conference on information retrieval (557-568). Springer, Berlin, Heidelberg.

[51] Grimbeek, P., Bryer, F., Beamish, W., & D'Netto, M. (2005). Use of data collapsing strategies to identify latent variables in CHP questionnaire data. Stimulating the'Action'as Participants in Participatory Research: Volume 2, 125-139.

[52] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19 (1), 22-36

[53] Aker, A., Zubiaga, A., Bontcheva, K., Kolliakou, A., Procter, R., & Liakata, M. (2017). Stance Classification in Out-of Domain Rumours: A Case Study Around Mental Health Disorders. In G. Ciampaglia, A. Mashhadi, & T. Yasseri (eds.) Social Informatics. SocInfo 2017. Lecture Notes in Computer Science, vol. 10540. doi:10.1007/978-3319-67256-4_6

[54] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

[55] Liu, H., & Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In Tools with artificial intelligence, 1995. proceedings., seventh international conference on (388-391). IEEE.

[56] Gunther, A. C., & Schmitt, K. (2004). Mapping boundaries of the hostile media effect. Journal of Communication, 54(1), 55-70.

[57] Gunther, A. C., & Liebhart, J. L. (2006). Broad reach or biased source? Decomposing the hostile media effect. Journal of Communication, 56(3), 449-466

[58] Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. Journal of peace research, 2(1), 64-90.

Figure 1. Example for an embedded tweet (and all displayed features) in the survey.

Table 1. Feature values indicating a significant impact on the different credibility dimensions (** p < .01; * p < .05).

Twitter features	Credibility dimensions						
	accurate	authentic	believable	comprehensible	important	informative	interesting
Number of followers	400200.**	327400.**	519100.**	13060.**	18760.**	81240.**	6813.**
Originality	51350.**	390500.**	35600.**	83630.**	77000.**	37870.**	23630.**
Engagement	1.01	84.53**	4.56*	33.41**	8.97**	92.20**	8.24**
Length of description	16.**	0.45	20.76**	0.00	1.95	139.7**	116.2**
Geo enabled	0.33	6.80**	4.18*	0.04	0.52	1.06	0.16
Location mentioned	1.59	0.60	1.80	1.48	0.11	9.83**	0.36
Person mentioned	8.29**	1.50	0.021	0.32	0.76	0.02	2.22
Role	1.04	0.35	0.26	7.02**	1.06	1.35	0.03
Negation	5.88*	3.20	1.26	0.07	0.99	0.00	0.52
Organization mentioned	3.45	0.15	1.41	1.88	1.84	4.75*	1.26
Date mentioned	0.35	0.69	0.08	0.15	1.40	2.55	4.31*
URL included	0.39	0.10	2.16	0.16	0.13	0.02	0.04
Sentiment	0.17	3.82	0.40	0.62	2.84	0.12	9.67
Google bad words	0.28	0.07	0.72	0.42	0.05	1.70	0.00
Average word length	0.00	0.00	0.24	0.25	0.42	0.12	0.04
Description	0.02	0.00	0.00	0.01	0.04	0.00	0.00