Rumors, Fake News and Social Bots in Conflicts and Emergencies: Towards a Model for Believability in Social Media

Christian Reuter, Marc-André Kaufhold, René Steinfort
University of Siegen, Institute for Information Systems
{christian.reuter, marc.kaufhold, rene.steinfort}@uni-siegen.de

ABSTRACT

The use of social media is gaining more and more in importance in ordinary life, but also in conflicts and emergencies. The social big data, generated by users, is partially also used as a source for situation assessment, e.g. to receive pictures or to assess the general mood. However, the information's believability is hard to control and can deceive. Rumors, fake news and social bots are phenomenons that challenge the easy consumption of social media. To address this, our paper explores the believability of content in social media. Based on foundations of information quality we conducted a literature study to derive a three-level model for assessing believability. It summarizes existing assessment approaches, assessment criteria and related measures. On this basis, we describe several steps towards the development of an assessment approach that works across different types of social media.

Keywords

Social media, believability, measurement.

1. INTRODUCTION

The increasing use of social media promotes, and even requires, new kinds of cooperation and collaboration between authorities, citizens and media in exceptional situations such as emergencies or large-scale crises. On the one hand, citizens demand current situational information by authorities and media, but citizen-generated content might also contain valuable information for the formal process of crisis management (Palen, Vieweg, & Anderson, 2010, p. 2f). In any case, the believability and reliability of used information play a crucial role to prevent the propagation or processing of misinformation such as fake news and rumors (Gupta, Lamba, Kumaraguru, & Joshi, 2013). With the emergence of "social bots", the topic becomes even more important: These bots are used in social media such as Facebook, Google+ or Twitter to publish information using specific hashtags or retweeting information of specific authors, and are often disguised as regular users with profile pictures, posts and followers. For instance, political bots are used to manipulate the public opinion (Ferrara, Varol, Davis, Menczer, & Flammini, 2016), which can also be used to carry out conflicts between countries.

A body of work already examined the assessment of information on the internet (Friberg, Prödel, & Koch, 2010; Ludwig, Reuter, & Pipek, 2015; Reuter, Ludwig, Kaufhold, & Pipek, 2015; Wang, Strong, & Guarascio, 1996), whereby *believability* is always mentioned as an important component of information quality. The aim of this paper is to explore existing approaches and criteria for assessing information believability in social media, to analyze their potential uses and finally to combine appropriate methods to realize a preferably reliable assessment of the believability of citizen-generated information. Such an approach is ideally supposed to be automatable to enable a consistent and quick assessment and to facilitate a manual screening of (mass) information. To achieve these goals, the paper introduces the basic context and terms (section 2) and presents, rates and analyzes existing approaches and criteria for assessing information believability within a systematic literature study (section 3). These approaches and criteria were used to develop a novel assessment method for information believability (section 4). Finally, the closing discussion outlines limitations and potentials of assessing the believability of citizen-generated content in social media (section 5).

2. FOUNDATIONS: QUALITY CRITERIA OF CITIZEN-GENERATED INFORMATION

Many papers have already considered most of the different criteria and partial aspects of information quality. Since the focus of this article is on a partial aspect of information quality, namely the believability of information, only a rough overview is given. The terms data quality and information quality are used synonymously here because information is only gained through data processing (Huang, Lee, & Wang, 1999, p. 7), but it is not assumed that citizens distribute data consciously to provide as many people as possible with information. Furthermore, the believability of data is considered, supposing that other quality criteria are met to enable an accurate further use concerning information. Accordingly, implausible information is equivalent to implausible and thus useless data.

Wang et al. (1996) define believability as follows: “The extent to which data are accepted or regarded as true, real, and credible”. To this date, there is no uniform definition of the term believability, but it is broadly recognized that believability results from a combination of the sender’s, message’s and recipient’s attributes (Wathen & Burkell, 2002, p. 2). The measuring of the quality criteria of information is a challenging task since it must be as exact as possible and simple to apply in practice. These goals are conflictive (Naumann & Rolker, 2000) so that a compromise between both conditions must be found. Thereby, the relation in which the specific condition has to be fulfilled can vary (e.g. exact acquisition vs. saving costs).

Although this paper focuses on citizen-generated content in social media such as Facebook, Twitter and Google+ (Boyd & Ellison, 2007) believability of information plays a crucial role in every field since only credible information are taken into account by the user consider the assurance of believability and trust an important challenge while using social media. They also emphasize the development of an automated assessment mechanism for separating credible and implausible information as a necessary future step while using information from social media, especially related to disaster situations and emergency service employments. Reuter, Ludwig, Ritzkatis and Pipek (2015) provide a “tailorable quality assessment service (QAS) for media content” that allows the different criteria to be adjusted according to the situation’s and user’s requirements. Besides believability, a close research field examines the detection of rumors. The PHEME project (https://www.pheme.eu), for instance, published a preliminary annotation scheme for social media rumours by analyzing the source tweet (polarity, modality, evidentiality, and author type) and response tweets (modality, evidentiality, author type, and response type) with the goal of training a model for automatic rumour detection., for instance, published a preliminary annotation scheme for social media rumours by analyzing the source tweet (polarity, modality, evidentiality, and author type) and response tweets (modality, evidentiality, author type, and response type) with the goal of training a model for automatic rumour detection.

Since the application of social media also increases consistently in disaster situations and even emergency services count more and more on the use of social media to spread or get information (St.Denis & Hughes, 2012), the use of fake news, which comprise fabricated, imposter, manipulated and misleading content, e.g. consciously through click baiting (Chen, Conroy, & Rubin, 2015), as well as satire, parody, false connections and false contexts, can have serious consequences in such situations (Wardle, 2017). Moreover, the time factor plays a major role in disaster situations (Palen et al., 2010) so that there is often no time for a manual screening and assessment of social media posts. To reduce the risk of misjudging information, this paper is supposed to provide a possibility to sort or screen posts by means of their believability by combining various, existing approaches and criteria for the assessment of the believability of user-generated information.

3. LITERATURE STUDY

3.1 Methodology

We conducted a literature study to identify and analyze, in terms of their suitability concerning the study’s aims, concepts and criteria that serve for the assessment of believability to encourage the development of an assessment approach. Using the terminology and framework of vom Brocke et al. (2015), we approached the field with a systematic literature review which is “a systematic, explicit, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers, scholars, and practitioners” (Fink, 2010). We applied the *sequential* process comprising the steps of input (searching), processing (analyzing, synthesizing) and output (writing). Our sources contained *bibliographic databases* (University Library, Google Scholar and Bielefeld Academic Search Engine (BASE)) and known *key publications*, which shaped the process of the literature review. In terms of coverage, our aim was not to collect a comprehensive sample of publications in the field of information quality, but to focus on *seminal works* on the believability of social media content. Furthermore, we applied the techniques of *keyword search*, using a keyword list for the search for German and English-speaking literature, and *backward search* to obtain further relevant works from references of the collected papers.

We used the following German and English words: *Authentizität; authenticity; believability; bürgergenerierte Informationen; citizen generated information; collective intelligence; credibility; Eigentrust; Glaubwürdigkeit; intrinsic information quality; intrinsische Informationsqualität; kollektive Intelligenz; reliability; Social Media; Soziale Netzwerke; trustworthiness; user-generated-content; Validität; validity; Web 2.0*

The paper of Palen et al. (2010) turned out to be one of the most important publications. None of the papers within the literature study provided an appropriate cross-platform assessment approach. Therefore, this paper is supposed to be the first step towards a uniform assessment approach for citizen-generated information in social media. It intends to give an overview of the challenges of its practical realization.

3.2 A Three-Level Model for Assessing Believability of User-Generated Information

The analysis of existing papers reveals that the presentation of the landscape of assessment approaches and criteria for user-generated information takes place on three levels. Therefore, we developed a three-level model to get a better overview of the existing ideas and tools and to create a clear foundation for the following sections (Figure 1). First, there are existing practical applications that implement assessment models for certain social media which build the highest level of existing assessment approaches (level 1). These models, in turn, make use of specific assessment criteria, which can have an influence on the believability of information (level 2). The lowest level comprises possibilities for examining criteria of level 2 and plays a rather subordinate role since the aim of this work is a combination of various criteria and approaches for the assessment of believability and not the assurance for the correctness of the considered aspects (level 3).

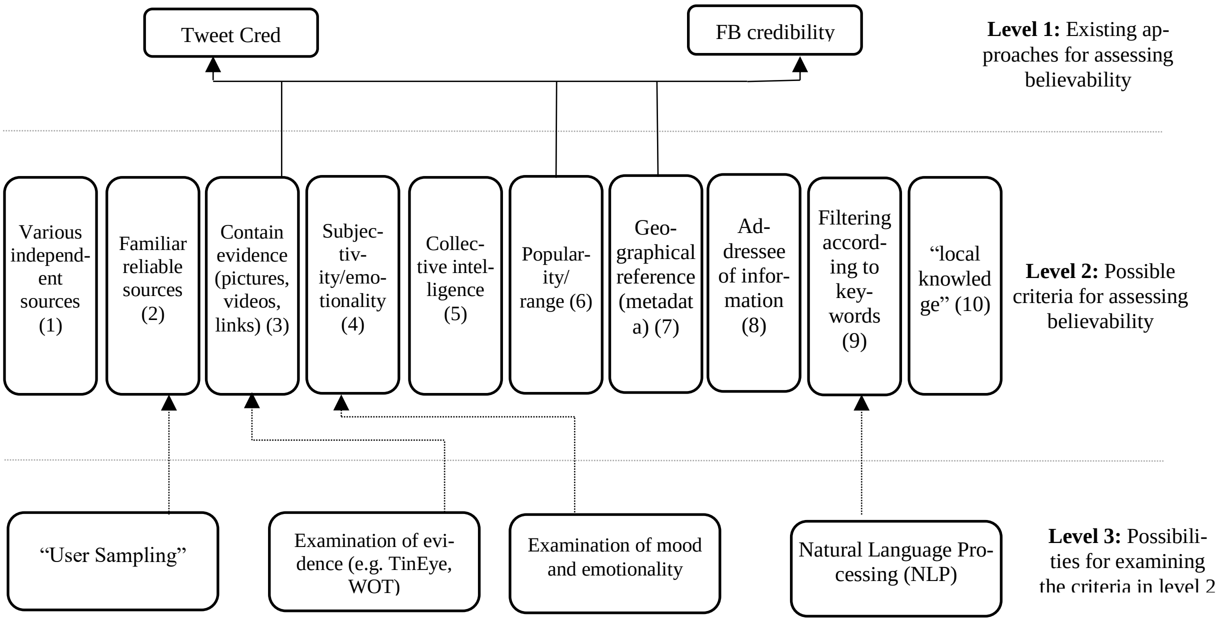
Tweet Cred
FB credibility
Level 1: Existing approaches for assessing believability
Various independent sources (1)
Familiar reliable sources (2)
Contain evidence (pictures, videos, links) (3)
Subjectivity/emotionality (4)
Collective intelligence (5)
Popularity/range (6)
Geographical reference (metadata) (7)
Addressee of information (8)
Filtering according to keywords (9)
"local knowledge" (10)
Level 2: Possible criteria for assessing believability
"User Sampling"
Examination of evidence (e.g. TinEye, WOT)
Examination of mood and emotionality
Natural Language Processing (NLP)
Level 3: Possibilities for examining the criteria in level 2

Figure 1. Three-level model of existing approaches for assessing believability

3.2.1 Level 1: Existing Assessment Approaches

Some of the identified criteria are already part of existing assessment approaches for the believability of social media posts (level 1, Figure 1). *TweetCred*, for instance, is a tool for real-time assessment of believability which also allows the examination of the trustworthiness of content on Twitter (Gupta, Kumaraguru, Castillo, & Meier, 2014). The system is based on the concept of semi-supervised learning and initially was trained through the manual classification of tweets concerning their situation reference and their believability. The SVM-rank algorithm for assessing the believability of tweets considers 45 characteristics in seven different categories in total. The access to the used tweets is realized with the help of the Twitter API. There are some parallels between the ten criteria visualized earlier (Figure 1) and the criteria named by Gupta et al. (2014).

A similar tool, *FB credibility*, is also available as a browser plug-in and was developed by Saikaew and Noyunsan (2015). It evaluates the believability of Facebook posts by means of the following criteria: *(1) number of likes, (2) number of comments, (3) number of shared posts, (4) number of contained links, (5) number of pictures, (6) number of hashtags, (7) number of videos, (8) availability of geographical metadata (location information).* Seven of these eight criteria, which were collected in the literature study, are covered; only the number of hashtags is added. First, the applied algorithm needs to be trained with the help of manually assessed posts. Since the Facebook API does not grant a sufficient access to the required data, the collection is realized with an own JavaScript code. For the assessment, a scale from 1 (implausible) to 10 (believable) is used. TweetCred and FB credibility further allow users to indicate their agreement on the computed believability value.

In a study, 81% of the users of FB credibility agreed to its assessment of believability (Saikaew & Noyunsan, 2015). While Saikaew and Noyunsay (2015) used eight criteria for their assessment of believability and got 81% agreement by the users, Gupta et al. (2014) took 45 criteria into account and got 40% agreement even though the information were more believable than they were rated. This shows that the use of more criteria does not necessarily result in a more reliable assessment of believability. Nevertheless, it makes sense to improve the named criteria since the examination of the respective criteria cannot take place consistently across media anyway.

3.2.2 Level 2: Assessment Criteria

Based on the literature review and on existing approaches, assessment criteria were analyzed and selected in terms of their relevance for believability in the context of social media. Overall, ten specific criteria constitute the second level of the three-level model:

1. **Various independent sources**: Several sources provide the same information (Palen et al., 2010; Rieh & Hilligoss, 2008, p. 14f): examination through so-called *cross-checking*, i.e. search for additional sources of similar content.
2. **Famililar and reliable sources:** The source is known and trustworthy (e.g. emergency services, Trusted Volunteers, transitive trust or the like) (Kamvar, Schlosser, & Garcia-Molina, 2003; Naumann & Rolker, 2000; Palen et al., 2010): Examination of trustworthy persons can be made through a database or *transitive trust*, i.e. the classification of trustworthy senders through reliable individuals.
3. **Contain evidence:** The source contains evidence such as pictures, videos, links to official sources or the like (Castillo, Mendoza, & Poblete, 2011, p. 7f; Oh, Kwon, & Rao, 2010, p. 12; Palen et al., 2010): Examination on content-related correctness about the attached evidence.
4. **Subjectivity/emotionality:** Filtering according to keywords – determination of subjectivity and emotionality (Castillo, Mendoza, & Poblete, 2013, p. 12; Oh, Agrawal, & Rao, 2013, p. 3f): Check text quality, because many orthographic or grammatical mistakes speak for a low believability, for example.
5. **Collective intelligence:** Correction of information through collective intelligence (Mendoza, Poblete, & Castillo, 2010; Palen et al., 2010; Vieweg, Palen, Liu, Hughes, & Sutton, 2008): Information might be examined through comments, corrections of the author, or questions of other observers, reducing the spread of misinformation (reference to criterion 6).
6. **Popularity/range:** Popularity and range of information – e.g. number of “likes”, “shares” (Facebook) or “Followers” (Twitter) (Castillo et al., 2011; Palen et al., 2010): The believability of contained information increases due to the pre-filtering through subjective assessment of disseminating users.
7. **Geographical reference**: Geographical proximity or another personal reference to the content (Gupta et al., 2014, p. 8; Oh et al., 2013; Palen et al., 2010): Believability results from the (personal) interest in trustworthy information by the author (e.g. search for help) due to physical proximity or belonging to organizations, communities or the workplace.
8. **Addressee**: Is information addressed to the public or particular persons? Existing concepts “imply that a rumor is more likely to spread within a community (i.e. particular persons) that is sustained by affective trust and strong social ties” (Oh et al., 2013, p. 412).
9. **Filtering keywords and signs**: Is specific information often questioned or disputed? Are there special symbols often used such as question marks or exclamation marks? (Castillo et al., 2013; Mendoza et al., 2010). Questions, doubts, or positive statements (combination of 4 & 5) influence believability positively or negatively. The use of many special signs, such as question marks or exclamation marks, can be a sign for lower believability.
10. **Existing “local knowledge”**: Local details, which outsiders maybe do not know and which signal the own consternation and thereby serious interest in misfortune (Palen et al., 2010): The believability results from the high geographical reference (e.g. knowledge of recent incidents, special buildings), the examination possibly proves to be problematic.

Within these assessment criteria, a conflict can result between the criteria 6 and 8 if popular members of social media spread generally accessible posts, for example. Furthermore, Oh et al. (2013) state that there is a restriction in situations, in which the author or the propagator of messages is worried or anxious. That questions the effectiveness of the collective intelligence (criterion 5) as well as the validity of the range and popularity of information (criterion 6) in connection with their believability in specific and time-sensitive situations (Oh et al., 2013). Naumann and Rolker (2000), for example, take the believability of information for not automatically assessable since the assessment of believability is always subjective and depends on the respective user of the data. Therefore, subjective quality criteria, such as believability, can only be measured for an individual person. This implicates that the assessment mechanism must be individualized for every user profile. This restriction for automating information assessment shows that a universal solution for automated assessment of believability is hardly realizable and an individualization should be allowed.

3.2.3 Level 3: Examination of Assessment Criteria

The work of Naumann and Rolker (2000) states three options for assessing subjective criteria: (1) *User Experience* (cf. approach 2), (2) *User Sampling* (from experiences of previous assessment as well as several articles of particular unknown sources), (3) Continuous User Assessment. Starbird et al. (2012) provide another example with their work about the determination whether an article was written on-site or by an outsider. This can be helpful insofar that the proximity to the affected area speaks for a higher believability than a spatial distance to the place of action. Using *SentiWordNet*, texts can be evaluated concerning their mood or predominant opinion (Pang & Lee, 2008, p. 111). Thereby, the three categories “positive”, “negative”, and “objective” are distinguished. With *OpinionFinder*, Wilson et al. (2005) present a system which consists of four components for analyzing subjectivity in texts. With that, opinions, moods and other subjective aspects can be identified.

Besides the examination of emotionality, texts can also be filtered for certain thematic orientations and keywords using *Natural Language Processing (NLP)*. Moreover, spelling and syntax mistakes can be realized and corrected. The possibility of multilingual filtering of texts is also interesting for the cross-platform social media application (Chowdhury, 2003, p. 22f). The examination of the quality of linked websites can be facilitated with the help of Web of Trust (WOT), for instance. This platform for assessing websites shows users to what extent previous users of this website take it for trustworthy. The assessment, in turn, is based on the believability criterion of collective intelligence (criterion 5).

4. TOWARDS THE DEVELOPMENT OF A CROSS-PLATFORM BELIEVABILITY ASSESSMENT APPROACH

4.1 Preliminary Considerations and Selection of Assessment Criteria

The previous section shows the complexity of developing a suitable algorithm for believability assessment due to plenty of factors: First and foremost, believability is subjective and users follow different interests in the exploitation of information and have different views on information. There are different claims for believability values (absolute trust required - e.g. emergency services - versus lower claims for the information search - e.g. private research with no time pressure and the possibility to examine believability) and users may consciously or unconsciously spread misinformation. Moreover, there are many assessment criteria, which are partly difficult to measure, context-dependent and difficult to weight. In this regard, different platforms with various conventions (hashtag-syntax, followers, likes, friends, focus on pictures/videos, etc.) and different numbers of users and posts constitute challenges in the definition of an algorithm. Also, there are differences between the availability of data; for instance, the Twitter API provides broad access to public data, but the Facebook Graph API offers only reduced access due to the larger proportion of private structures. Thus, also data protection issues must be considered (saving of user data of reliable sources, use of photos, videos and user information).

Because not every discussed criterion is applicable to every social medium, it is obvious that one should use platform-independent criteria. However, the notions and specifications of the used assessment criteria differ in some aspects so that an apparent commonality cannot be determined for certain. Furthermore, the existing tools are limited to two social media platforms, which have some parallels so that the type of contents (text, pictures, videos), the structure of social contacts (friends, followers, subscribers, etc.), and the formal conditions of the respective platform (character limit, required syntax) additionally should be considered for contents for further platforms. Therefore, it will not be possible to determine the criteria exactly, which can be applied to all social media equally. Hence, the selection of the suitable criteria is made in general terms and must be inspected for concrete possibilities for the implementation.

Based on the literature study and the presented existing tools, the three most mentioned criteria for the assessment are the range/propagation of the posts, the availability of reliable sources and evidence (pictures, videos, URLs) and geographical or content-related reference. These criteria can be fulfilled cross-platform and partly collected in many social media. Further criteria are the emotionality or mood of the posts (4) and the filtering by relevant and significant keywords (9). Using NLP, one can also identify content-related parallels (Brill & Mooney, 1997, p. 5) between posts to find several sources to proof the believability of the texts (1). One possibility which can be applied to all social media is the management of known and reliable sources. Although correcting false information by collective intelligence (5) seems to have a positive effect on the believability, it cannot be measured actively as it is a passive mechanism. The availability of *local knowledge* is not considered as an independent criterion since the mentions of actual true local circumstances cannot be examined reliably. The aspect is covered by the local connection (7) regarding the non-geographical reference and is proved insofar that the posts, which contain the same information (1), confirm it. Figure 2 reflects the selected criteria for the further procedure.

4.2 Modelling the Possible and Negative Effects of the Assessment Criteria

Figure 2 shows the potential positive ("+") and negative ("-") effects of the selected criteria. While a high range and reputation (e.g. likes) always affect believability positively, most of the other criteria can have both positive and negative effects. Existing proofs for the spread contents can have negative effects if they are dubious, old or implausible. Especially examining the quality of photos and videos is difficult since the content is not made automatically. Therefore, three possibilities are distinguished. Known helpful data can have positive effects on believability since they present a (positive) content-related overlap with other sources (VII) and contain believable evidence. If a photo turns out to be known but implausible or too old, the believability is effected in a negative way. If none of these two situations arises, one deals with an unknown evidence, which must be proved manually. Additionally, the geographical proximity signals interest in this location and therefore raises the believability.

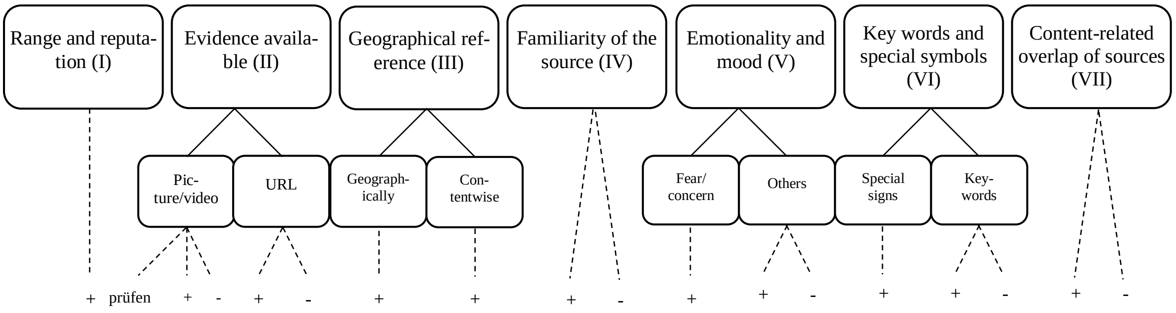
Range and reputation (I)
Evidence available (II)
Geographical reference (III)
Familiarity of the source (IV)
Emotionality and mood (V)
Key words and special symbols (VI)
Content-related overlap of sources (VII)
Picture/video
URL
Geographically
Contentwise
Fear/concern
Others
Special signs
Keywords
+
prüfen
+
-
+
-
+
+
+
-
+
+
-
+
+
-
+
-

Figure 2. Criteria for assessing believability in social media and effects

The publicity of a source cannot only have positive effects on the believability of contents, but also save bad experiences with certain users in social media, e.g. if someone distributes dubious posts regularly. However, there should be regular examinations of the classification of believability for trustworthy and implausible sources since the assessment may change. Other persons can use user accounts; accounts can be hacked. While the literature study showed that fear and anxiety of the authors of citizen-generated posts in social media could have negative effects on the believability, other emotions and moods can have positive effects (Castillo et al., 2013). Moreover, a frequent mention with same content can have a positive effect on the believability of information whereas contradictions can lead to a downgrade of believability.

4.3 A Proposal for the Cross-Platform Assessment of Believability

The assessment of believability might be realized by weighting the identified criteria (Figure 2). The level of influence on believability is different between the individual criteria for various reasons. First, these are measurable cross-platform to different degrees. However, more important is the scaling of the individual measurement results. A geographical reference in a post can either exist or not, range and reputation can be available to different degrees. Further points of reference are the reliability for the measuring and its practicability. If the post contains a URL or a picture, it regularly speaks for high believability. On the contrary, it is not clear how great the impact of emotionality and mood is in terms of believability and therefore must be examined and adapted by unknown factors implementing further examinations.

These considerations determine the ranking of the individual assessment criteria. According to that, the familiarity of the source (IV) has the greatest impact on believability. In the second place, the examination of evidence follows, especially containing URLs, which ideally enable a verification of the information. Additionally, it has synergies to criterion VII since, once examined, evidence is saved and its reliability can be determined directly in case of recurrence if posts show parallels when using evidence. Subsequently, range and reputation of a post follow in the ranking of impact on the believability value, as they are easy to measure, compare and scale. The content-related overlap of sources regarding mutual confirmation or contradictions is ranked fourth. The geographical reference is ranked below since not every post refers to a location. Therefore, a lower relevance of this criterion serves the comparability of posts of various platforms. If one classified a tweet about a catastrophe as significantly more credible because it contains a suitable geographical reference, it would potentially for no reason be classified as more credible than another post from a platform, whose content does not have a geographical reference. Finally, the criteria follow which evaluate texts by emotionality, mood, keywords and special signs. On the one hand, the possibility to evaluate such criteria is unequally distributed on social media since a few platforms are specialized on the sharing of pictures or videos (YouTube, Flickr), while others rather focus on text messages (Twitter) or support different information carriers (Facebook). Moreover, the exact consequences and their scaling on believability are not as clear as the ones of other criteria of believability. Thus, based on the study and our argumentation, the following list of criteria results:

1. Familiar sources evaluated concerning their believability (IV)
2. Information is proved with evidence such as pictures, videos or URLs (II)
3. Range and reputation of a post (I)
4. Content-related compliance between posts (VII)
5. Geographical reference (III)
6. Filtering by keywords, emotionality, mood and special signs (V, VI)

Another point is the assessment of absent criteria. It is obvious that the absence of positive criteria cannot be evaluated positively. However, the lack of evidence does not mean that the information is automatically implausible. Therefore, an assessment is suggested, which uprates posts fulfilling positive criteria; however, the assessment of these posts, which do not fulfill the criterion, should not suffer. Thus, the absence of positive criteria does not affect the assessment of believability. Consequentially, posts, which do not meet any of the criteria, whether positive or negative, receive a neutral assessment of believability.

To evaluate the practical suitability of the developed cross-platform approach concerning the evaluation of believability in social media, the following steps must be executed: Initially, there should be a possibility to collect and unify relevant posts for the evaluation of believability in various social media so that they can be evaluated similarly. Existing SVM algorithms are suitable for adapting the evaluation of believability. After realizing technical aspects, such as the possibility to manage familiar sources and to implement various criteria from level 3 (see Figure 2), the algorithm's realization (e.g. scoring of criteria, behavior on absent criteria) and evaluation can be executed. While the two presented existing evaluation approaches are implemented in the form of browser plug-ins, and TweetCred is additionally offered as a web-based application and API, the question regarding technical potentials and the meaningful implementation of such an approach arises.

5. CONCLUSION AND OUTLOOK

This paper has shown that assessing believability, especially in cross-platform fields, has hardly been addressed yet and therefore requires several developments until a suitable assessment mechanism is realized. An important first step in this direction is the conception of a possibility to collect and standardize data cross-platform to enable the filtering of data for a believability assessment using a scoring algorithm. Furthermore, the implementation of the assessment approach implies obstacles, for instance, technical challenges such as the realization of the individual believability criteria and the procedures for their examination. Moreover, such an algorithm should support the selection of individual assessment criteria and the adjustment of their weightings to enable an individualization of the assessment of believability, which – especially during conflicts and emergencies – depends on the situation, purpose, persons, and role. Furthermore, the proper procedures for the examination of believability criteria must be selected and developed. Regarding the technical implementation, further aspects have to addressed, for example, the question concerning the data protection or the liability for wrong decisions (St.Denis & Hughes, 2012).

Currently, there are several developments in the field of assessing believability in online communities, which indeed are limited to certain platforms, but still could gain many findings in the field of assessing believability in social media. These findings represent a part of the preparatory work for developing a cross-platform assessment approach for the believability of citizen-generated information in social media. This paper shows a further step in this direction through conceptualizing a general approach for cross-plattform believability in social media and provides a first theoretical recommendation for action for the realization of such a solution. However, to implement and refine such an approach, it must be carfully synchronized with social trends and research proceedings in the fields of information quality (Shankaranarayanan & Blake, 2017) and validation, rumor detection (Mendoza et al., 2010) and social bots (Ferrara et al., 2016).

BIBLIOGRAPHY

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230.

Brill, E., & Mooney, R. J. (1997). An Overview of Empirical Natural Language Processing. *AI Magazine*, *18*(4), 13–24.

Brocke, J., Simons, A., Riemer, K., Niehaves, B., & Plattfaut, R. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the AIS*, *37*(1).

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web - WWW '11* (p. 675). New York, New York, USA: ACM Press.

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

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading Online Content: Recognizing Clickbait As "False News." In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection* (pp. 15–19). New York, USA: ACM.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89.

Christofzik, D., & Reuter, C. (2013). The Aggregation of Information Qualities in Collaborative Software. *International Journal of Entrepreneurial Venturing (IJEV)*, *5*(3), 257–271.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The Rise of Social Bots. *Communications of the ACM*, *59*(7), 96–104.

Fink, A. (2010). *Conducting research literature reviews: from the internet to paper*. California: Sage Publications Ltd.

Friberg, T., Prödel, S., & Koch, R. (2010). Analysis of information quality criteria in crisis situation as a characteristic of complex situations. In *Proceedings of the 15th International Conference on Information Quality*. Little Rock, USA.

Gräfe, G., & Maaß, C. (2015). Bedeutung der Informationsqualität bei Kaufentscheidungen im Internet. In K. Hildebrand, M. Gebauer, H. Hinrichs, & M. Mielke (Eds.), *Daten- und Informationsqualität* (pp. 169–191). Wiesbaden: Springer Fachmedien Wiesbaden.

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). TweetCred: Real-Time Credibility Assessment of Content on Twitter. In L. M. Aiello & D. McFarland (Eds.), *Social Informatics* (Vol. 8851, pp. 228–243). Springer International Publishing (Lecture Notes in Computer Science).

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy. In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 729–736). New York, NY, USA: ACM.

Hiltz, S. R., Diaz, P., & Mark, G. (2011). Introduction: Social Media and Collaborative Systems for Crisis Management. *ACM Transactions on Computer-Human Interaction (ToCHI)*, *18*(4), 1–6.

Huang, K.-T., Lee, Y. W., & Wang, R. Y. (1999). *Quality information and knowledge*. Upper Saddle River, N.J.: Prentice Hall PTR.

Kamvar, S. D., Schlosser, M. T., & Garcia-Molina, H. (2003). The Eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the twelfth international conference on World Wide Web - WWW '03* (p. 640). New York, New York, USA: ACM Press.

Ludwig, T., Reuter, C., & Pipek, V. (2015). Social Haystack: Dynamic Quality Assessment of Citizen-Generated Content during Emergencies. *Transactions on Human Computer Interaction (ToCHI)*, *21*(4).

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* (pp. 71–79).

Naumann, F., & Rolker, C. (2000). Assessment Methods for Information Quality Criteria. In *Proceedings of the International Conference on Information Quality (IQ)* (pp. 148–162). Cambridge, United Kingdom.

Oh, O., Agrawal, M., & Rao, R. (2013). Community Intelligence and Social Media Services: A Rumor Theoretic

Analysis of Tweets during Social Crises. *Management Information Systems Quarterly*, *37*(2), 407–426.

Oh, O., Kwon, K. H., & Rao, H. R. (2010). An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010. In *Thirty First International Conference on Information Systems* (pp. 231–245). St. Louis, Missouri, USA.

Palen, L., Vieweg, S., & Anderson, K. M. (2010). Supporting "Everyday Analysts" in Safety- and Time-Critical Situations. *The Information Society*, *27*(1), 52–62.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Reuter, C., Ludwig, T., Kaufhold, M.-A., & Pipek, V. (2015). XHELP: Design of a Cross-Platform Social-Media Application to Support Volunteer Moderators in Disasters. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)* (pp. 4093–4102).

Reuter, C., Ludwig, T., Ritzkatis, M., & Pipek, V. (2015). Social-QAS: Tailorable Quality Assessment Service for Social Media Content. In *Proceedings of the International Symposium on End-User Development (IS-EUD). Lecture Notes in Computer Science*.

Rieh, S. Y., & Hilligoss, B. (2008). College Students' Credibility Judgments in the Information-Seeking Process. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 49–72). Cambridge, MA: MIT Press (The John D. and Catherine T. Macarthur Foundation series on digital media and learning).

Saikaew, K. R., & Noyunsan, C. (2015). Features for Measuring Credibility on Facebook Information. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *9*(1), 174–177.

Shankaranarayanan, G., & Blake, R. (2017). From Content to Context: The Evolution and Growth of Data Quality Research. *ACM Journal of Data and Information Quality*, *8*(2), 9:1--9:28.

St.Denis, A. L., & Hughes, A. L. (2012). Trial by Fire: The Deployment of Trusted Digital Volunteers in the 2011 Shadow Lake Fire. In L. Rothkrantz, J. Ristvej, & Z. Franco (Eds.), *Proceedings of the Information Systems for Crisis Response and Management (ISCRAM)*. Vancouver, Canada.

Starbird, K., & Palen, L. (2012). (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW)*. Bellevue, WA, USA: ACM Press.

Vieweg, S., Palen, L., Liu, S. B., Hughes, A. L., & Sutton, J. (2008). Collective Intelligence in Distaster: Examination of the Phenomenon in the Aftermath of the 2007 Virginia Tech Shooting. In F. Friedrich & B. Van de Walle (Eds.), *Proceedings of the Information Systems for Crisis Response and Management (ISCRAM)* (pp. 44–54). Washington D.C., USA.

Wang, R., Strong, D., & Guarascio, L. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems (JMIS)*.

Wardle, C. (2017). Fake news. It's complicated. Retrieved March 26, 2017, from https://firstdraftnews.com/fake-news-complicated/

Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology*, *53*(2), 134–144.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., … Patwardhan, S. (2005). OpinionFinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations* - (pp. 34–35). Morristown, NJ, USA: Association for Computational Linguistics.