IET Networks

Research Article

ISSN 2047-4954
Received on 10th November 2018
Revised 12th March 2019
Accepted on 2nd May 2019
E-First on 3rd October 2019
doi: 10.1049/iet-net.2018.5208
www.ietdl.org

Fuzzy-based fake information detection algorithm to define the user trust on the content of social networks

Munirathinam Nirmala[1], *Madda Rajasekhara Babu*[2] ✉
[1]*School of Information Technology and Engineering, VIT University, Vellore, Tamil Nadu, India*
[2]*School of Computer Science and Engineering, VIT University, Vellore, Tamil Nadu, India*
✉ *E-mail: rajababu.m1@gmail.com*

Abstract: Across the globe, right from people who are technologically naive to people who are with high-technical knowledge, use, share, spread, and connect with each other through online social networks. People start believing and sharing the social network content without any proof of its authenticity. In several cases, the reliability of the information that gets shared between the users remains questionable due to the anonymity of the information creators. In this work, a fuzzy detection algorithm is proposed to identify the trustable content in social media. The proposed methodology is evaluated on Twitter social network platform. By computing the predictive measures, the efficiency of the proposed approach is well established.

1 Introduction

Online social networking platforms create avenues for users to get connected with each other through the World Wide Web. Moreover, the affordable smart devices pave the way for millions of people to start using social network platforms irrespective of their technical background. People use social networks not only to connect with one another but also to share various information and digital media ranging from news to health tips. Statistics portal Statista reports that currently there are >2.5 billion social media users and it will reach more than three billion users in the next three years [1, 2]. So, it is apparent that more number of people are going to get connected through social media in the near future. Given this scenario, there is a vital need to scrutinise the media content that gets shared between the users. Even though social media gets utilised mostly for advantageous purposes, there is a high-risk factor over the circulation of fake information in the social media [3–6]. Manipulation of fake news using photos and other rumours circulated among the social media circles are even spread by news agencies without any verification [7].

Ranging from trending news to contacting customer services, people have started to seek information through social media platforms [8–10]. In addition to the real-time capabilities, the easy access provision of social media acts as an added advantage for users [4, 11–13]. Therefore, it clearly appears that in coming years more people will continue to follow and use social media for their information needs. The reliability of the information and the user trust on the information that is being circulated in the social media remains questionable [4, 14–16]. Currently, there is an alarming concern in detecting the fake news that keeps spreading in social media because spreading of misinformation not only affects the social media users but also affects the general public. Recently, in India mob lynching of >20 innocent people happened because of the fake news spread by WhatsApp, an arm of social media giant Facebook [17]. Telegraph website reported that Twitter bots affected the financial service by spreading fake news [18]. From plagiarising election campaign to pulling down stock prices, spreading of fake news plays a major role. Although social media is a boon to the people, the trending work of spreading misleading information raises concern over the usage of social media. Famous news website the guardian reported that the social media giant Facebook accepted that mob violence was caused due to spreading of misleading information and urged to take new policies to control the spreading of misleading information [19]. Although social media plays a saviour during natural disasters [20–23] due to quickness in viral spreading the trending news, [24] there is a compelling need for proper filtering mechanisms to identify and point the misinformation to the users.

Therefore, controlling the spreading of misleading information and fake news remains as one of the major challenges in social media. So, in this work, a fuzzy-based trust detection methodology is proposed to identify the reliability of social media posts. This paper is classified into the following sections: Section 1 discusses the introduction, in Section 2 a comparative study is made on similar works that are carried to find the reliability of information, and Section 3 elaborates the proposed approach. In Section 4 the proposed fuzzy-based trust detection algorithm is evaluated and the results are compared with other state-of-art approaches and finally, in Section 5, the outcome of the work and its future aspects are discussed.

2 Review on similar works

The fake news in the web world has a distasteful impact on society and gets accretive every day. In this current scenario, researchers started to analyse, control, and classify rumours based on their impact. In the year 1944, Robert Knapp classified the rumours into the following categories namely, wedge-driving, bogy and pipe dream rumours [25]. Based on the current scenario Zubaiga *et al.* classified fake news into two categories namely, fake news that starts emerging during trending news and fake news that is getting spread for long periods [3]. They trained the detection system based on these rumours to classify the misleading information from the original source. When there is a past history of information regarding the fake news or rumours, then the trained classifier will be able to detect the fake news. Based on the trained classifier techniques, many methods came into existence to detect the existing rumours in the system [26, 27]. Even though the trained classifier methods were able to detect the already existing fake information, it failed in detecting new fake information.

Zhao and others detected the new rumours by analysing the user comments which asked for the authenticity of the news [28]. A specific topic based-rumour detection methodology was proposed by Gupta and others [29]. Feature selection methodology was used by Tolosi *et al.* authors to detect fake news [30]. Carlos Castillo used feature extraction methodology on trending topics to classify the tweets as reliable or unreliable [31]. McCreadie *et al.* got the response from multitude of users on news to find about its reliability [32]. Wide range of topic-based detection methods, features selection, and comments based methodologies gives lesser accuracy due to the abundance and availability of vast information. In some cases, biasing of users is so high so that, in such cases, false news becomes true due to less number of genuine users. When the multitude of users support misinformation, then it continues to be system failure. Liu and others proposed a rumour detection algorithm utilising past models and journalistic verification [33]. Even though journalistic verification is one of the best-known approaches, there are some cases where fake news has come as a news article based on social media information [34].

Ennals *et al*. identified the dispute topics or pages and then conducted a phrase dictionary-based analysis on it to find its credibility [35]. In the phrase dictionary methodology, the detection rate was much lesser than the spreading rate. The accuracy based on phrase dictionary approach was less and results were inaccurate when there were a sizeable number of emoticons. Qazvinian *et al*. detected misinformation in Twitter by building different Bayseian classifiers and by learning the linear function of retrieval and classification [36]. In this work, Qazvinian *et al*. made a test on 10 K tweets on five different controversial topics and found the misinformation based on Bayesian classifier. Even though they considered several features of the tweet, they did not analyse the retweets of the tweets. Contradiction-based detection system was also utilised to detect the credibility of messages in social media [37, 38]. Finding contradiction among the information through identified tweets and considering all the contradicting news as fake news will decrease the accuracy of the system in finding the misinformation.

Truthy web service system created by Ratkiewicz *et al*. and others based on mapping and classification, identified the truthful memes in the Twitter [39]. Petter Bae Brandtzaeg and others made a detailed study on the verification practices and concluded that there is a need for new practices and tools to identify the rumours through verification models [34]. Considering some features of the message or analysing the user reaction on the message will not yield the truthfulness of the information. This is because, when a biased user reacts for misinformation, he might like the information and will endorse it without verifying the creditability of the information. Also, some malicious users create bots to increase the endorsing methodology of information to make the information appear like truthful information. If the news is a newly broke news, then analysing the news based on journal verification method fails due to unavailability of information about the news, even if analysing all the features of the information returns possible creditability information about the news. The reliability of the information features becomes questionable in certain cases where multitude of biased users spread hatred over a minority of user or a single user. All these factors led us to propose a fuzzy trust factor-based detection methodology to detect the misinformation in social media.

3 Proposed user trust detection approach

In this paper, we have proposed a fuzzy-based approach in which user trust factor UT is first computed by considering all the posts made by the user. When the user is considered to be trustworthy, then the posts made by the user gets classified as reliable posts. As a further verification strategy, the post and the replies along with endorsements made to the current post are corroborated to conclude the post as reliable information. This algorithm is classified into two modules: one is to detect the user trust factor and other is the valuation of the post. The effectiveness of the proposed algorithm is proved using data set extracted from the social networking site Twitter.

3.1 User trust detection

In order to detect the trust of the user, posts created by the user are analysed and scrutinised. News or information can be blindly followed or tweeted repeatedly by users who want unreliable information to be true. In other cases, unidentified and fake users will start spreading the misinformation for their personal or political gain. One cannot ignore all the news of the user as fake when a genuine user is sharing the information. So in order to validate the user, user trust is computed for each user in the system. User reaction on the reliable posts is fetched and analysed based on the fuzzy decision analysis to find the trust factor of the user. Consider that the user u has *n* number of posts P. The user trust factor UT is computed by finding the trust function of each post and this is depicted in (1). Where µT stands for the user trust function and PS is the score for each post *i*.

The trust function μT is computed based on the fuzzy linguistic decision analysis. For each post-PiS, the score is computed by analysing the presence of positive and negative words using the dictionary-based method. In order to compute the score, the words present in the tweet are classified and computed as seen in (2). Here NP stands for number of positive items and TNW represents total no of items. The total number of positive items NP is computed as shown in (3) where Nw and Ne are the number of positive words and the number of emoticons. Each word carries 0.1 score, whereas emoticon takes double the value of word, i.e. 0.2. Likewise, as in (2), NN, i.e. the total number of negative items for the post is computed, this is shown in (4) and (5) where NN stands for the number of negative items. Finally, the post score is computed by summing up the +ve and –ve as shown in (6).

$$U^T = \sum_{i=0}^{i=1} \left(\mu^T : \left(P^S\right)_i \rightarrow [0, 1]\right) \tag{1}$$

$$+\text{ve} = \frac{\text{NP}}{\text{TNW}} \tag{2}$$

$$\text{NP} = \text{Nw} + \text{Ne} \tag{3}$$

$$-\text{ve} = \frac{\text{NN}}{\text{TNW}} \tag{4}$$

$$\text{NN} = \text{Nw} + \text{Ne} \tag{5}$$

$$P^S = +\text{ve}\ (+) - \text{ve} \tag{6}$$

Then, a fuzzy membership function μA is defined for further analysis and the definition for the fuzzy membership function is depicted in Fig. 1. A is a set containing Z items that belongs to the analysis which is depicted in (7). The membership function on set A, μA is defined as in (8). The membership function over the Z items is defined in (9) and the details of the function that tends to the item are given in (10).

$$A = \left\{ \begin{array}{c} \text{nothing}(n), \text{very low}(\text{vl}), \\ \text{low}(l), \text{medium}(m), \\ \text{high}(h), \text{very high}(v), \ \text{perfect}(p) \end{array} \right\} \tag{7}$$

$$\mu_A = \left\{ \begin{array}{c} \mu_{\text{very low}}, \mu_{\text{low}}, \mu_{\text{medium}}, \\ \mu_{\text{high}}, \mu_{\text{very high}} \end{array} \right\} \tag{8}$$

$$\mu_A : Z \rightarrow [0, 1] \tag{9}$$

$$\left. \begin{array}{c} \text{very low} = \mu_{\text{very low}} : X \rightarrow [0, 0.125] \\ \text{low} = \mu_{\text{low}} : X \rightarrow [0.125, 0.375] \\ \text{medium} = \mu_{\text{medium}} : X \rightarrow [0.375, 0.6] \\ \text{high} = \mu_{\text{high}} : X \rightarrow [0.6,\ 0.85] \\ \text{veryhigh} = \mu_{\text{very high}} : X \rightarrow [0.85,\ 1] \end{array} \right\} \tag{10}$$

Computed post score is converted into linguistic value and user trust factor is computed based on the majority of the linguistic values. In the algorithm titled detecting the reliability of the tweets, step by step procedure of computing is explained in Fig. 2. If a user has post scores mostly of μ_{verylow}, then the bound value of user trust factor U^T is defined as very low. The user trust factor of very low and low are considered to be spreading the misinformation more. Medium trust factor defines that the user is in dilemma and is not inclined to either accepting or objecting the information. Accordingly, high and very high user trust factor indicates that it is

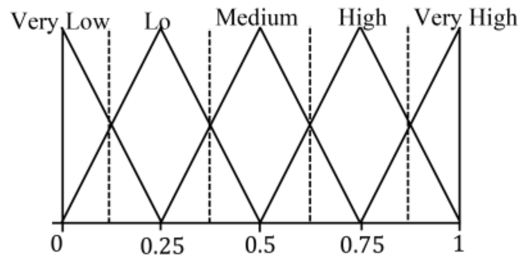

Fig. 1 *Fuzzy membership function definition*

```
U→Users, T→Trust, μT→Trust Function, P→User Post,
VP→Post Verdict, UT→User Trust Factor, PS→Score Post,
PR→Post Replies
MP(P)→ Maximum positive polarity of P , MN(P)→Maximum
negative polarity of Pi
Calculate User trust factor  UT
        Initialize Trust Function μT
          μT: PS→[0,1]
      where μT={ μVL, μL, μM, μH, μVH }
      Compute Post Score PS
        FOR all the post Pi belongs to User Ui
          Fetch the PR belong to Pi
                FOR all the PR belong to Pi
                  Compute Polarity for each PR
          Count positive and negative for each PR
          IF Positive Polarity > Negative Polarity
            Then Assign Positive polarity to Pi
          ELSE
            Assign Negative polarity to Pi
          EndIF
        EndFor
        Count Polarity of P
         Get the MP(P)
         Get the MN(P)
        IF count of  MP(P)  > MN(P)
             PS= MP(P)  / Total No of P
        Else
             PS= MN(P)  / Total No of P
        EndIF
        Convert PS into membership function
        Based on μT Detect UT
```

Fig. 2 *Detecting reliability of the Tweets*

a reliable post and shows that the user is genuine towards reliable information and then the user is termed as a reliable user.

3.2 Tweet valuation

After finding the user trust, the information that is under scrutiny is evaluated to find its reliability. In order to evaluate the tweet, the user trust factor of the user who created the Tweet is analysed. If the user has good trust factor, then the replies made to the tweets and trust factor of the user are considered to define the valuation of the tweet. If most of the users with positive trust factor reacted positively to the post then the tweet is considered as reliable information. If the users with good trust factor are negative towards the post then the post is considered to be not reliable. For further verification and validation of the proposed algorithm, real-time tweets are fetched and misinformation and fake profiles were created to find the predictive measures of the proposed algorithm.

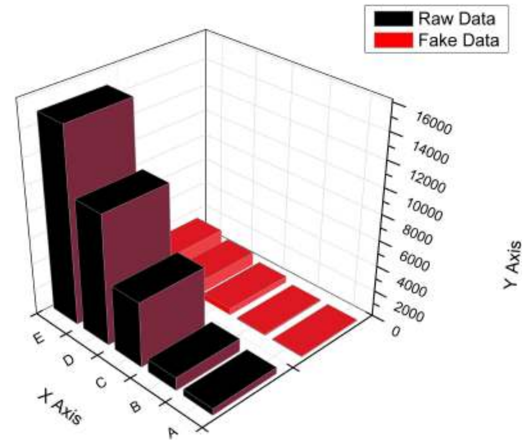

Fig. 3 *Raw data and fake information*

4 Evaluation

In order to evaluate the proposed algorithm, one of the kaggle Twitter data set is utilised [40]. The data set consists of more than fourteen thousand (14 K) tweets and in the tweets, biased users and fake information were manually inserted to find the accuracy of the proposed algorithm. The data set is divided into the following test cases namely A, B, C, D, and E for the evaluation process. Test case A consisted of 550 tweets in which 50 tweets were fake; in test case B, there were totally 1100 tweets in which 100 tweets were fake, C consisted of 5500 tweets in which 500 tweets were manually inserted fake information. Test case D consists of 11,000 tweets where 1000 are fake and test case E consisted of >15,000 tweets in which 1500 were fake information and the same is shown in Fig. 3.

The proposed algorithm is evaluated on the test cases and the accuracy of the algorithm in finding the fake information of the test cases is depicted in Fig. 4. From Fig. 4, it can be visualised that the algorithm tries to find more than half of the inserted fake information. For further analysis, the accuracy percentage of the algorithm is calculated and depicted in Fig. 5.

From the values in Fig. 5, it is clear that the algorithm has an overall accuracy of >65% in detecting the fake information. For further verification, predictive measures like precision, recall, and the mean of precision and recall $f1$ scores are computed. Precision is the ratio between retrieved data and the total data in the data set. The recall is the ratio between the accurate data in the retrieved data and the total data in the data set.

Four test data namely $t1$, $t2$, $t3$, and $t4$ were created for calculating predictive measures. Each test data consisted of > 3750 tweets out of which 1000 fake tweets were induced. Precision score

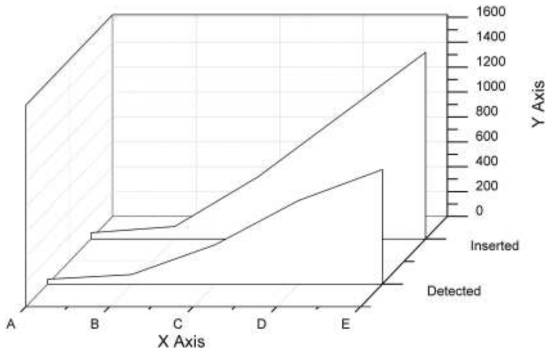

Fig. 4 *Detecting fake information*

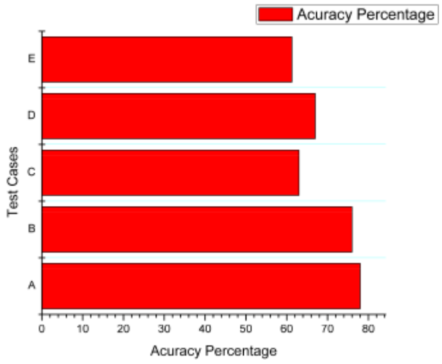

Fig. 5 *Accuracy percentage of the algorithm*

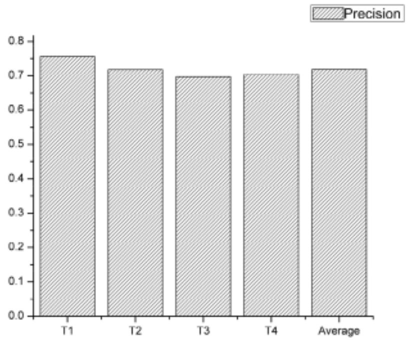

Fig. 6 *Precision score on the test data*

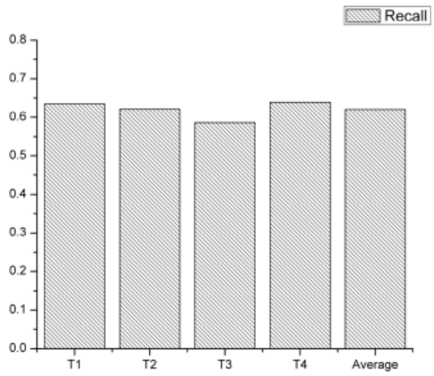

Fig. 7 *Recall score on the test data*

on the test data is depicted in Fig. 6, from the figure it can be visualised that the scores of $t1$, $t2$, $t3$, and t4 are >0.7. The recall score which is depicted in Fig. 7 shows that the proposed algorithm has a recall score of >0.6 and the average of the score is 0.62. Further, $F1$ score is computed as seen in Fig. 8, from which it can be conceived that the proposed algorithm has a mean predictive score of >0.65. The predictive measures and accuracy evaluation of the algorithm show that it has better accuracy in detecting fake information.

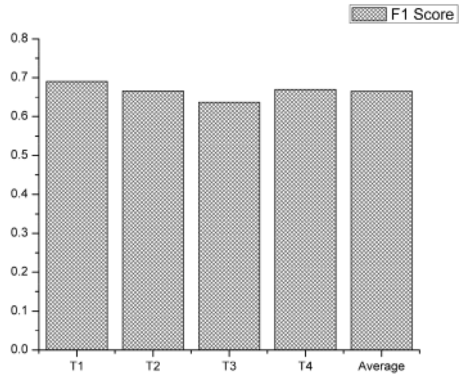

Fig. 8 *F1 score on the test data*

5 Conclusion

Detecting fake information remains as one of the toughest problem of the information era. With a huge number of information accumulating each data, it remains complex to understand and distinguish reliable and fake information. So in the proposed approach, the reliability of the information is detected by analysing the source of the information. The algorithm proved to achieve better efficiency in detecting fake information. As a future enhancement, the algorithm will be executed on real-time data and the prediction accuracy of the algorithm will be detailed. Identifying trustable content in social media is of paramount importance in the interest of society at large. Towards this end, the proposed trust identification and fake information detection strategies will be of great help in regulating social media.

6 References

[1] Statista: 'Number of social media users worldwide from 2010 to 2021', 17 July 2018. Available at https://www.statista.com/markets/424/topic/540/social-media-user-generated-content/

[2] L.D, D., Raj, E.: 'Flocking based evolutionary computation strategy for measuring centrality of online social networks', *Appl. Soft Comput. J.*, 2017, **58**, pp. 495–516

[3] Zubiaga, A., Aker, A., Bontcheva, K., *et al.*: 'Detection and resolution of rumours in social media: A survey', *ACM Comput. Surveys (CSUR)*, 2018, **51**, (2), pp. 1–32

[4] Phuvipadawat, S., Murata, T.: 'Breaking news detection and tracking in twitter'. Int. Conf. on Web Intelligence and Intelligent Agent Technology, Toronto, Canada, 2010

[5] Procter, R., Vis, F., Voss, A.: 'Reading the riots on twitter: methodological innovation for the analysis of big data', *Int. J. Soc. Res. Methodol.*, 2013, **16**, (3), pp. 197–214

[6] Webb, H., Burnap, P., Procter, R., *et al.*: 'Digital wildfires: propagation, verification, regulation, and responsible innovation', *ACM Trans. Inf. Syst.*, 2016, **34**, (3), pp. 1–15

[7] Schifferes, S., Newman, N.: 'Verifying news on the social web: challenges and prospects'. Proc. 22nd Int. Conf. on World Wide Web, Rio de Janeiro, Brazil, 2013

[8] Lazer, D., Pentland, A.S., Adamic, L., *et al.*: 'Life in the network: the coming age of computational social science', *Science*, 2009, **323**, (5915), p. 721

[9] Shah-Mansouri, V., Moradi, A.: 'Opportunistic content dissemination in mobile social networks via adjustment of user selfishness', *IET Netw.*, 2018, **8**, (2), pp. 126–137

[10] Hayat, K., Bickler, G., Irfan, R., *et al.*: 'Survey on social networking services', *IET Netw.*, 2013, **2**, (4), pp. 224–234

[11] Raj, E., Babu, L.: 'A model fuzzy inference system for online social network analysis'. 2015 Int. Conf. on Computing and Network Communications, CoCoNet 2015, Trivandrum, India, 2016

[12] Raj, E., Babu, L.: 'A survey on topological properties, network models and analytical measures in detecting influential nodes in online social networks', *Int. J. Web Based Commun.*, 2017, **13**, (1), pp. 137–156

[13] Raj, E., Babu, L.: 'An enhanced trust prediction strategy for online social networks using probabilistic reputation features', *Neurocomputing*, 2017, **219**, pp. 412–421

[14] Zubiaga, A., Liakata, M., Procter, R., *et al.*: 'Analysing how people orient to and spread rumours in social media by looking at conversational threads', *PloS one*, 2016, **11**, (3), p. e0150989

[15] Sharon Moses, J., Babu, L., Raj, E.: 'Evaluation of similarity based recommender using fuzzy inference system'. 2015 Int. Conf. on Computing and Network Communications, CoCoNet 2015, Trivandrum, India, 2016

[16] Sharon, M., Babu, L.: 'A locality centred recommendation system combining CNM clustering technique with fuzzy preference tree-based ranking algorithm', *Int. J. Bus. Intell. Data Mining*, 2016, **11**, (1), pp. 63–84

[17] Elliott, J.K.: 'India WhatsApp killings: Why mobs are lynching outsiders over fake videos', Global News, 16 July 2018. Available at https://globalnews.ca/

news/4333499/india-whatsapp-lynchings-child-kidnappers-fake-news/, Accessed 18 July 2018
[18] Wallace, T.: 'Twitter bots manipulating stock markets as fake news spreads to finance', 31 March 2018. Available at https://www.telegraph.co.uk/business/2018/03/31/twitter-bots-manipulating-stock-markets-fake-news-spreads-finance/, Accessed June 20 2018
[19] Solon, O.: 'Facebook's plan to kill dangerous fake news is ambitious – and perhaps impossible', 19 July 2018. Available at https://www.theguardian.com/technology/2018/jul/19/facebook-fake-news-violence-moderation-plan, Accessed 20 July 2018
[20] De Longueville, B., Smith, R.S., Luraschi, G.: 'Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires'. Proc. 2009 Int. Workshop on location based social networks, Seattle, Washington, USA, 2009
[21] Vieweg, S.: 'Microblogged contributions to the emergency arena: discovery, interpretation and implications'. Computer Supported Collaborative Work, Savannah, Georgia, USA, 2010, pp. 515–516
[22] Vieweg, S., Hughes, A.L., Starbird, K., *et al.*: 'Microblogging during two natural hazards events: what twitter may contribute to situational awareness'. Proc. SIGCHI Conf. on human factors in computing systems, Atlanta, Georgia, USA, 2010, pp. 1079–1088
[23] Hughes, A.L., Palen, L.: 'Twitter adoption and use in mass convergence and emergency events', *Int. J. Emerg. Manag.*, 2009, **6**, (3), pp. 248–260
[24] Poulsen, K.: 'Firsthand reports from California wildfires pour through Twitter', 2007
[25] Knapp, R.H.: 'A psychology of rumor', *Public. Opin. Q.*, 1944, **8**, (1), pp. 22–37
[26] Qazvinian, V., Rosengren, E., Radev, D.R., *et al.*: 'Rumor has it: identifying misinformation in microblogs'. Proc. Conf. on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, 2011
[27] Hamidian, S., Mona, D.: 'Rumor identification and belief investigation on twitter'. Proc. 7th Workshop on Computational Approaches to Subjectivity, San Diego, California, USA, 2016
[28] Zhao, Z., Resnick, P., Mei, Q.: 'Enquiring minds: early detection of rumors in social media from enquiry posts'. Proc. 24th Int. Conf. on World Wide Web, Florence, Italy, 2015
[29] Gupta, A., Lamba, H., Kumaraguru, P., *et al.*: 'Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy'. Proc. of 22nd Int. Conf. on World Wide Web, Rio de Janeiro, Brazil, 2013
[30] Tolosi, L., Tagarev, A., Georgiev, G.: 'An analysis of event-agnostic features for rumour classification in twitter'. Tenth Int. AAAI Conf. on Web and Social Media, Colonge, Germany, 2016
[31] Castillo, C., Mendoza, M., Poblete, B.: 'Information credibility on twitter'. Proc. 20th Int. Conf. on World wide web, Hyderabad, India, 2011
[32] McCreadie, R., Macdonald, C., Ounis, I.: 'Crowdsourced rumour identification during emergencies'. Proc. 24th Int. Conf. on World Wide Web, Florence, Italy, 2015
[33] Liu, X., Nourbakhsh, A., Li, Q., *et al.*: 'Real-time rumor debunking on twitter'. Proc. 24th ACM Int. on Conf. on Information and Knowledge Management, Melbourne, Australia, 2015
[34] Brandtzaeg, P.B., Lüders, M., Spangenberg, J., *et al.*: 'Emerging journalistic verification practices concerning social media', *Journalism Pract*, 2016, **10**, (3), pp. 323–342
[35] Ennals, R., Byler, D., Agosta, J.M., *et al.*: 'What is disputed on the web?'. Proc. 4th Workshop on Information Credibility, New York, USA, 2010
[36] Qazvinian, V., Rosengren, E., Radev, D.R., *et al.*: 'Rumor has it: identifying misinformation in microblogs', 2011
[37] Ritter, A., Downey, D., Soderland, S., *et al.*: 'It's a contradiction–no, it's not: a case study using functional relations'. Proc. Conf. on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, USA, 2008
[38] Marneffe, M.-C., Rafferty, A.N., Manning, C.D.: 'Finding contradictions in the text'. Proc. ACL-08: HLT, Columbus, Ohio, USA, 2008
[39] Ratkiewicz, J., Conover, M., Meiss, M., *et al.*: 'Detecting and tracking the spread of astroturf memes in microblog streams', arXiv preprint:1011.3768, 2010
[40] Rathee, A.: 'Demonetization in India twitter data', kaggle, December 2016. Available at https://www.kaggle.com/arathee2/demonetization-in-india-twitter-data, accessed October 2018