Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 142 (2018) 294–300

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), November 17-19 2018, Dubai, United Arab Emirates

Detecting rumors in social media: A survey

Samah M. Alzanin[a], Aqil M. Azmi[a],*

[a]*Department of Computer Science, College of Computer & Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia*

Abstract

With recent development of technology, especially mobile devices has made the social networks accessible 24/7. Information spreading has become faster than ever, regardless of the credibility of this information. This brings unparalleled challenges in ensuring the reliability of the information. Misinformation spreading has a strong relation especially in the context of breaking news, where the information released gradual, often starting as unverified information. Automatically identifying rumors from online social media especially micro-blogging websites is an important research. Recent research in detecting rumors automatically on social networks have addressed many languages. In this article, we provide an overview of the research into rumors detection in social media which we divided into three groups: supervised based approaches, unsupervised based approaches, and hybrid approaches based on the type of the machine learning used in each approach.

Keywords: Rumour detection; Machine learning; Misinformation; Social media; Survey.

1. Introduction

It has been less than two decades since the modern social media came into existence during which it attracted a huge following. One of the basic forms of social media is social networks, which are sites that allow users to create personal web pages and then connect with different friends to share content and interact with each other. In 2003, LinkedIn was launched, and this was followed by Myspace and Facebook in the year 2004. YouTube was launched in 2005 followed by Twitter, followed by others such as Blogging, Google+, Instagram, WhatsApp, and Snapchat [1].

* Corresponding author. Tel.: +966 11 467-6574.
E-mail address: aqil@ksu.edu.sa

10.1016/j.procs.2018.10.495

The rapid development of technology in recent years, especially in mobile phones made social networks available at all time and information spreading has become faster than ever [2]. The nature of social networks allowed a rapid spreading of information in real-time, regardless of the credibility of this information, which created unparalleled challenges in information reliability assurance. Misinformation spreading relates specially to breaking news, where the information is released gradual, often starting as unverified information. Allport and Postman [3] defined rumor as "propositions of faith on specific (or current) topics that pass from person to person, usually by word of mouth, without any evidence of their truth". DiFonzo and Bordia [4] define rumor as "unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger or potential threat, and that function to help people make sense and manage risk". Another definition is presented by Dunn and Allen [5] "A rumor is a hypothesis offered in the absence of verifiable information regarding uncertain circumstances that are important to those individuals who are subsequently anxious about their lack of control resulting from this uncertainty." Recently, there has been many rumors, covering all parts of life. According to the statistics of the Anti-Rumors council†, about 88 rumors were documented in the period from January to April 2017. This large number of rumors in such a short time span can have damaging results for individuals as well as the society. One of the most popular online social networks is Twitter, which is used to share information with other users. It was designed so that each user can send information as short messages of length up to 140 characters, these messages called 'tweets'. The relationship of 'follower – following' in Twitter enables the rapid spreading of information, where each user can subscribe to receive messages from other users by becoming their 'followers', in the same time those other users become following for the first user. This feature of easy information spreading may constitute a thread for rumors dissemination in Twitter social network.

In recent years, many of researches have been done for dealing with the problem of rumors detection in the field of social media. This paper surveys the different approaches used to detect rumors in social media, which we classified into three categories: supervised, unsupervised, and hybrid based approaches.

2. Supervised based approaches

The analysis of information credibility spread through social media networks by Castillo et al. [6] represented the starting point for many of the later works. Castillo et al. [6], who pioneered in engineering features for credibility assessment on Twitter [7], build their work upon that there is some of indicators in the social media environment that enable users to assess information credibility. A dataset was built for studying credibility on Twitter. Then, each topic is labeled by a group of human assessors according to NEWS, which reports a fact that can be of interest to others, and CHAT, which is a message that is based on personal opinions and/or conversations among friends. Each item of the former class is assessed on its level of credibility by another group of judges. Next, a set of features is extracted to build a classifier that attempts to automatically determine if a topic corresponds to a newsworthy information/event, and then to automatically assess its level of credibility. The features are categorized into four groups, message-based features, user-based features, topic-based features, and propagation-based features. The features that have more contribution to the main task is selected using a best-first selection method. For the training dataset. A random sample of the dataset was obtained using sampling with replacement considering a uniform distribution for the probability of extracting an instance across the three classes (NEWS, CHAT and UNSURE). A set of supervised learning methods were adopted including Support Vector Machines (SVM), decision trees, decision rules, and Bayes networks. The best result was achieved by a J48 decision tree method. This classifier achieves good results with accuracy equal to 86%. Castillo et al. [6], concluded that newsworthy topics tend to include URLs and to have deep propagation trees. In addition, credible news is propagated through authors that have previously written a large number of messages, originate at a single or a few users in the network, and have many re-posts.

Also, an early study on rumor analysis and detection on Sina Weibo, a popular Chinese social media site, was addressed by Yang et al. [8]. The problem was formulated as a classification problem. The microblogs matching keywords in the topics published by the rumor busting account were collected from March 1, 2010, to February 2,

† http://www.nonrumors.net/

2012. The function was done by annotators. Label each collected microblog with "1" if the orientation of the microblog endorses the rumor, and with "-1" otherwise. This study extended some of features that have been studied in earlier works [6, 9, 10] which are content based, account based, and propagation based features, with two new features. The two new proposed features are client program used and event location. By analyzing the collected dataset, it found that about 71.8% of false information is posted by non-mobile client programs. In addition, about 56.1% of the events occurred abroad. A statistical analysis was carried using Pearson's chi-squared test to perform the test of independence between the client program feature and the truthfulness; the same test is done for the event location feature as well. The results of this test show that the client program feature has a relationship with the truthfulness, also the test showed that the event location is not independent of the truthfulness. Two sets of experiments were conducted at the feature level. In the first experiment, SVM with the Radial Basis Function (RBF) kernel was trained using a specific subset of the previously proposed features (content based, account based, propagation based). In the second set of experiments, the impact of including the two newly proposed features was studied. The results of the first experiment show that propagation-based features alone do not perform as well as using the other two subsets of features. In general, the results of the last three grouped features are acceptable, with an average accuracy of 72.5% for each. Furthermore, the results of adding the two new features in this study provide a brief improvement of 5.5% in accuracy (from 72.5% to 77%).

The data set published by Qazvinian et al. [10] is an annotated Twitter data set for five different 'established' rumors. This data set was used by Dayani et al. [11] and Hamidian and Diab [12] for their research on rumors detection within the context of Twitter social media. Dayani et al. [11] expanded the dataset in [10] with more information about the users who have posted these tweets. K-Nearest neighbor (KNN) was applied on user based features, and Naive Bayes classifier (NB) was applied on content-based features for rumors detection. To apply Naive Bayes algorithm, preprocessing of tweets was carried before applying the classifier. The results showed that low prediction accuracy is achieved with KNN classifier, the authors attributed this to the fact that rumor detection has no correlation with the user-based features. But this conclusion conflicts with other studies like [6] that showed good results by including user based features in the selected features. The results for applying NB classifier showed that rumor detection had been improved after applying preprocessing algorithm with accuracy of 86% for both rumor endorsed, and rumor denied, and 74% for rumor questioned. This result is another conflict also with [6] who see that URL has important rule for determining newsworthy topics. In addition, the authors reported that word frequencies play a key role in rumor detection.

Hamidian and Diab [12] addressed the problem of rumor detection and classification within the context of microblog social media. The used data set was published by [10]. The effectiveness of multistep with various sets of features and preprocessing tasks versus a single step detection and classification approach was investigated. In the single step classification, rumors detection and classification were performed simultaneously as a 6-way classification task among the six classes in the labeled data (Not Rumor, Undetermined tweet, Endorsing the Rumor, Denys the Rumor, Questions the Rumor, and Neutral about the Rumor). In the multi-step classification, an initial 3-way classification task is performed as following (Not Rumor, Undetermined tweet, and the compound) labels. This is followed by a 4-way classification step as following, (Endorsing the Rumor, Denys the Rumor, Questions the Rumor, and Neutral about the Rumor). For experiment the authors used the J48 decision tree implemented within the Weka platform. All the features proposed in [10] were included in this study, in addition to developing pragmatic attributes as well as additional network features. For network and Twitter features, two features were added, Replay and Time when a tweet posted (Busy day or Regular day) and for pragmatic features, some of features were proposed as Named-Entity Recognition (NER), Sentiment, and Emoticon. The experiments were carried with three data sets of [10]. The results showed that multistep rumor detection and classification outperformed single rumor detection and classification with the best results achieved on Obama dataset with 71% vs. 85% F-measure for SRDC and TRDC respectively. In addition, the results showed that the set of features added in this study achieved a brief enhancement to the overall performance in certain rumors dataset.

Another study about the Trustiness of tweets during crises, Sahana et al. [13] build their data set based on the rumored tweets posted during the London riots in. A set of features based on tweet content and user accounts were selected based on previous studies. The authors performed Information Gain Attribute Evaluator with Ranker search method to find the most significant ones. A J48 decision tree with a 10-fold cross validation and 10 iterations was performed for classification. According to this study, the content-based features showed important rule for the

detection of rumors and conversely user-based features other than the status count play considerably a smaller part, From the point of view of the author, this is because the dataset used to train the classifier is populated by retweets and contains very few original tweets. In addition, the author clarifies that a user who has a high status count is possible to retweet rumors as active users on Twitter tend to retweet new information regardless their credibility. The achieved accuracy of correctly classified instances was 87.9%. This study is consistent with the previous research in [38] that user based features have played an insignificant role for the detection of rumors, With the variation in the interpretation of the reason. Furthermore, in term of accuracy, the results of two studies were convergent.

A recent study of rumors detection was presented by Kwon et al. [14] by examining rumor characteristics over different observation time windows. User, linguistic, network, and temporal features were adopted in this study in order to identify the significant differences between rumors and non-rumors for the first 3, 7, 14, 28 and 56 days from the initiation, The variable selection process proposed in [15] was used and repeated for selecting the user, linguistic, network, and temporal features in order to compare the prediction effort for each feature. Accordingly, a set of features were selected for the interpretation and prediction task. Two new algorithms were proposed for rumor classification. The first algorithm takes every feature into account, and the second algorithm is outfitted for early rumor classification and considers user and linguistic features. The authors reported that structural and temporal features distinguished rumors from non-rumors over a long-term window. However, user and linguistic features have better performance when the task is to detect rumour as early as possible.

Arabic rumors identification was presented by Floos [16], who focus on detecting rumors in Arabic tweets. The authors identified two types of tweets Rumors and News, then TF-IDF was computed for each term in the tweets. To classify the new tweet. The dot product was applied to the tweet vector and the news vector. The same steps are repeated for the rumors vector, the largest dot product result is used as pointer to classify the new tweet.

3. Unsupervised based approaches

One of the studies that used unsupervised approach in rumors detection is presented by Takahashi and Igata [17] based on an investigation of several characteristics of rumors. One of the investigated characteristics is retweet ratio, from the experiments, the authors concluded that retweet ratio tells no difference between normal tweets and rumors, but for large sample size, it can get useful results. Also, the difference of word distribution characteristic appears to have an important rule for rumors detection.

A clustering-based approach for political rumors detection on Twitter was proposed by Chang et al. [18]. Their study identified particular kind of malicious users, called extreme users, tend to post false news on Twitter. The trustiness of news tweets has been decided based on the number of the extreme users included, some of the features were outlined related to extreme users such that a large number of following, a high tweeting frequency, and show enthusiasm about the target topic, use of extreme keywords in description or tweets. Two steps have been included: cluster the tweets that contain the same URL link and merge similar clusters that discuss the same news into one cluster based on cosine similarities. The experiments were carried on two sets of tweets about Hillary Clinton in August 2015, and Barack Obama in September 2015. The authors reported that the best rule differs from dataset to another. So, there is no definite way for identifying the rules, many rules may be constructed as a combination of parameters, where it does not give a clear picture of how to improve the results. Also, labeling the clusters as rumors or not was done manually which is consuming effort and need to be improved.

Chen et al. [19] treated false rumor detection as an anomaly detection problem in Sina Weibo. A set of features were proposed which are divided into three categories, crowd wisdom, content features, and posting behavior. To predict whether a suspected Weibo w is a rumor or not, the set of recent k Weibos of the user who posted this suspected Weibo is extracted in addition to a set of needed features to these Weibos. Then, Factor analysis of mixed data FAMD is performed on the previous set, and the number of extracted features is reduced. After that, these Weibos are ranked according to their degree of deviation using two strategies: Euclidean Distance and Cosine similarity. To predict the label of a suspected Weibo, if it ranks top in either of Euclidean Distance or Cosine similarity, it will be regarded as a rumor, otherwise non-rumor. K-means method has been chosen as the baseline for the comparison with the proposed method in this study. The two methods were carried out on the same collected data. The results show that the proposed method as an unsupervised approach is better than K-means for rumors. Some accounts in social networks are

dedicated to rumors spreading, such this method may not have an important rule for detecting rumors as it based on that the style of presentation of a rumor post is different from a normal message for the individual.

A recent study was presented by Jain et al. [20] based on the premise that verified News Channel accounts on Twitter would provide more credible information than the unverified account of user. Four main steps have been defined for detecting rumors on Twitter. First, extracting tweets corresponding to Twitter trends and identify topics being talked about in each trend based on clustering using hashtags and then collect tweets for each topic. Second, isolate the tweets for each topic based on whether its tweeter is a verified news channel or a general user. The tweets from a verified news channel are categorized as news tweets, all the other tweets fall into public tweet set. Third, calculate and compare the contextual and sentiment mismatch between tweets of the same topic from verified Twitter accounts of News Channels and other unverified (general) users using semantic and sentiment analysis of the tweets. And finally, label the topic as a rumor based on the value of mismatch ratio, which reflects the degree of conflict between the news and public on that topic. If the mismatch ratio is greater than a threshold value, the topic is labeled as a rumor, otherwise, it is labeled as not rumor. According to the authors, the results was better when the extracted tweets were more objective and less subjective, thus, a larger number of tweets are categorized as False Positives/ Negatives due to sentiment score above the threshold.

4. Hybrid Approaches

Hashimoto et al. [21] proposed a framework to detect the rumor information in social media. word-of-mouth messages were collected from online bulletin board kakaku.com which is the most popular marketing site in Japan. morphological analysis including RIDF, LSA, and TF-IDF was used to extracts nouns, verbs, adjectives, and adverbs from one document. Then the score of an individual keyword is calculated. To show rumor information structure, directed graphs were constructed using the concept graph. Visualization tool was developed, it provides time series interface using a slider in which users can recognize structure changes in time-series variation. To detect the rumor information, graph topology-based distance [22] was employed for measuring changes in concept graph topology over time. When the concept graph structure becomes larger, it shows the rapid spread of topic. Such topic is considered as candidates of rumor. Thus, the reliable sources such as TV programs and newspapers is verified for each rumor information candidate. There are no results reported in this study.

Some of the studies combine the two approaches, supervised and unsupervised learning in their studies, one of these studies was presented by Cai et al. [23] who focus on identifying rumors on Sina Weibo by extracting features from the texts of retweets and comments. Different words were selected as feature sets using Chi-square, even including some special symbols or words. To determine whether stop words and punctuations will be useful to be included in the feature set to enhance the results of rumors detection, cluster analysis was carried using the methods of Hierarchical Cluster [24] to analyze the sample dataset and Jaccard coefficient [25] to calculate the similarity between these texts of retweet and comment. SVM algorithm was used to train a classification model for rumor detection. The authors reported that feature set containing some special characters with the training set consists of samples with lower similarity helps to get more efficient results.

Another study was addressed by Liu and Xu [7] based on observing rumors propagation patterns in social media environment. The rumor detection task is addressed as a classification problem with two classes of microblogs, rumors, and credible messages. A new information propagation model of a given microblog over a social network was suggested based on that propagation patterns of rumors and credible messages are distinct from each other in social media. The study was carried on Chinese social media site Sina Weibo. To enhance the accuracy of classification, the proposed method clustered all users appearing in the propagation pathway into k groups based on user-status-sensitive features. The EM clustering algorithm [26] was used for implementation, in the algorithm's training stage, the log-likelihood that the observed propagation pathway was generated by the proposed information propagation model under the credible message mode and the rumor mode is calculated through some equations. If the log-likelihood that the observed propagation pathway was generated by the proposed information propagation model under the credible message mode greater that the log-likelihood that the observed propagation pathway was generated by the proposed information propagation model under the rumor mode, then the microblog is determined as a credible message, otherwise is determined as a rumor. The proposed algorithm was compared with other existing works and the authors reported that the proposed algorithm outperformed the other works employed for the performance comparison.

5. Conclusion and future work

With the proliferation of social media, it is easy to access information for all users. Under the broad and rapid deployment of information and the absence of strategies to ensure the reliability of such information, many of the research, which aim to develop systems to detect rumors emerged. Some languages have had a great interest in detecting rumors in social media and some still lack such research, so there is still an urgent need to expand the search to include many languages.

References

[1] I. Oyza and A. M. Edwin, (2016) "Effectiveness of Social Media Networks as a Strategic Tool for Organizational Marketing Management," *Journal of Internet Banking and Commerce (JIBC),* Jan 2016, vol. 21, no. S2.

[2] Y. Zhang and C. Chen, (2014) "A rumor spreading model considering latent state," *Proceedings of the Eighth International Conference on Management Science and Engineering Management*, pp. 155-162.

[3] G. W. Allport and L. Postman, (1947) *The psychology of rumor*, New York: Henry Holt.

[4] N. DiFonzo and P. Bordia, (2007) "Rumor, gossip and urban legends," *Diogenes,* vol. 54, pp. 19-35.

[5] H. B. Dunn and C. A. Allen, (2005) "Rumors, urban legends and internet hoaxes," in *Proceedings of the Annual Meeting of the Association of Collegiate Marketing Educators*, p. 85.

[6] C. Castillo, M. Mendoza, and B. Poblete, (2011) "Information credibility on twitter," *Proceedings of the 20th international conference on World wide web*, pp. 675-684.

[7] Y. Liu and S. Xu, (2016) "Detecting rumors through modeling information propagation networks in a social media environment," *IEEE Transactions on Computational Social Systems,* vol. 3, pp. 46-62.

[8] F. Yang, Y. Liu, X. Yu, and M. Yang, (2012) "Automatic detection of rumor on Sina Weibo," *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, p. 13.

[9] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, (2012) "Tweeting is believing: understanding microblog credibility perceptions," *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 441-450.

[10] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, (2011) "Rumor has it: Identifying misinformation in microblogs," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1589-1599.

[11] R. Dayani, N. Chhabra, T. Kadian, and R. Kaushal, (2015) "Rumor detection in twitter: An analysis in retrospect," *2015 IEEE International Conference on Advanced Networks and Telecommuncations Systems (ANTS),* pp. 1-3.

[12] S. Hamidiain and M. Diab, (2015) "Rumor detection and classification for twitter data," in *The Fifth International Conference on Social Media Technologies, Communication, and Informatics,* pp. 71-77.

[13] V. Sahana, A. R. Pias, R. Shastri, and S. Mandloi, (2015) "Automatic detection of rumoured tweets and finding its origin," *2015 International Conference on Computing and Network Communications (CoCoNet),* pp. 607-612.

[14] S. Kwon, M. Cha, and K. Jung, (2017) "Rumor Detection over Varying Time Windows," *PLOS ONE,* vol. 12, art. e0168344.

[15] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, (2010) "Variable selection using random forests," *Pattern Recognition Letters,* vol. 31, pp. 2225-2236.

[16] A. Y. M. Floos, (2016) "Arabic Rumours Identification By Measuring The Credibility Of Arabic Tweet Content," *International Journal of Knowledge Society Research (IJKSR),* vol. 7, pp. 72-83.

[17] T. Takahashi and N. Igata, (2012) "Rumor detection on twitter," *2012 Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS),* pp. 452-457.

[18] C. Chang, Y. Zhang, C. Szabo, and Q. Z. Sheng, (2016) "Extreme User and Political Rumor Detection on Twitter," *Proceedings of 12th International Conference on Advanced Data Mining and Applications (ADMA 2016), Gold Coast, QLD, Australia, December 12-15, 2016*, pp. 751-763.

[19] W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, (2016) "Behavior deviation: An anomaly detection view of rumor preemption," *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON),* pp. 1-7.

[20] S. Jain, V. Sharma, and R. Kaushal, (2016) "Towards automated real-time detection of misinformation on Twitter," *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI),* pp. 2015-2020.

[21] T. Hashimoto, T. Kuboyama, and Y. Shirota, (2011) "Rumor analysis framework in social media," *IEEE Region 10 Conference TENCON 2011*, pp. 133-137.

[22] H. Bunke, (1997) "On a relation between graph edit distance and maximum common subgraph," *Pattern Recognition Letters,* vol. 18, pp. 689-694.

[23] G. Cai, H. Wu, and R. Lv, (2014) "Rumors detection in chinese via crowd responses," *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* pp. 912-917.

[24] S. C. Johnson, (1967) "Hierarchical clustering schemes," *Psychometrika,* vol. 32, pp. 241-254.

[25] J. Han, J. Pei, and M. Kamber, (2011), *Data mining: concepts and techniques*: Elsevier.
[26] X. Jin and J. Han, (2011), "Expectation maximization clustering," in *Encyclopedia of Machine Learning.* Springer, pp. 382-383.