A Model for Identifying Misinformation in Online Social Networks

Conference Paper · October 2015

DOI: 10.1007/978-3-319-26148-5_32

3 authors:

Sotiris Antoniadis
Trasys International
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Iouliana Litou
Athens University of Economics and Business
13 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Vana Kalogeraki
Athens University of Economics and Business
233 PUBLICATIONS **0** CITATIONS
SEE PROFILE

A Model for Identifying Misinformation in Online Social Networks

Sotirios Antoniadis[1], Iouliana Litou[2(✉)], and Vana Kalogeraki[2]

[1] Nokia Solutions and Networks Hellas A.E., Athens, Greece
sotiris.antoniadis@nsn.com
[2] Deptartment of Informatics,
Athens University of Economics and Business, Athens, Greece
{litou,vana}@aueb.gr

Abstract. Online Social Networks (OSNs) have become increasingly popular means of information sharing among users. The spread of news regarding emergency events is common in OSNs and so is the spread of misinformation related to the event. We define as misinformation any false or inaccurate information that is spread either intentionally or unintentionally. In this paper we study the problem of misinformation identification in OSNs, and we focus in particular on the Twitter social network. Based on user and tweets characteristics, we build a misinformation detection model that identifies suspicious behavioral patterns and exploits supervised learning techniques to detect misinformation. Our extensive experimental results on 80294 unique tweets and 59660 users illustrate that our approach effectively identifies misinformation during emergencies. Furthermore, our model manages to timely identify misinformation, a feature that can be used to limit the spread of the misinformation.

1 Introduction

Online Social Networks (OSNs) have evolved into major means of communication and information spreading. They enumerate over 1.61 billion users, which corresponds to 22% of the world's population. However, one major challenge is that the information communicated through the network is not always credible. Previous studies confirm the existence of *spam campaigns* in OSNs [1][2]. Spam campaigns are organized attempts towards spreading false or malicious content through the coordination of accounts or other illicit means in the network. It is estimated that, among the messages published on Twitter, 1% of the messages is spam while 5% of the accounts are spammers[1].

Twitter[2] has evolved as one of the most popular microblogging services. Users publish short messages (*tweets*) of at most 140 characters and follow any other

S. Antoniadis—Part of this work was performed when this author was at Athens University of Economics and Business.

[1] http://digital.cs.usu.edu/ kyumin/tutorial/www-tutorial.pdf

[2] https://twitter.com/

C. Debruyne et al. (Eds.): OTM 2015 Conferences, LNCS 9415, pp. 473–482, 2015.
DOI: 10.1007/978-3-319-26148-5_32

registered users to receive status updates. Twitter offers to users the opportunity to report a tweet as *spam*, *compromised* or *abusive*. Other filters to detect spam (e.g the number of followers in regard to followees, random favorites and retweets etc.) are also used. Still, tweets containing misinformation regarding an emergency event may not be identified based solely on the aforementioned mechanisms.

The credibility of images propagated in the network has been the focus of recent work. Zubiaga and Ji [3] and Gupta et al. [4] focus on the credibility of images propagated in the network during emergency events but not in the content of the information. The works closest to ours is that of Castillo et al. [5] and Xia et al. [6]. Both works use supervised learning and Bayesian Network classification to identify credible information propagated in the network. Castillo et al. [5] cluster the instances to newsworthy or chats and later perform credibility analysis on the newsworthy clusters. Xia et al. [6] propose a model to detect an emergency event and identify credible tweets. As we illustrate in our experimental evaluation, our approach performs better than both approaches in identifying misinformative tweets with over 14% higher accuracy.

In this work we suggest a methodology for identifying and limiting misinformation spread in OSNs during emergency events by identifying tweets that are most likely to be inaccurate or irrelevant to an event. Our work makes the following contributions:

- We present a novel filtering process for identifying misinformation during emergencies that is fast and effective. The filters are identified based on an extensive analysis conducted on a large dataset of users and tweets related to the emergency event of Hurricane Sandy. As our experimental results illustrate, the filtering process extracts over 81% of the misinformative tweets, while identifying over 23% of the tweets that contain misinformation.
- We employ a number of supervised learning algorithms that very effectively classify credible or misinformative tweets. Based on the features we propose, classification techniques achieve weighted average accuracy of 77%.
- Our experiments suggest that without considering propagation of tweets, our classification methodology achieves 77.8% weighted average accuracy, offering the ability to timely limit the spread of false news before cascading. Furthermore, the filtering process and the classification algorithms perform in less than 2 seconds, making our methodology appropriate for real-time applications.

2 Problem Description, Parameters and Methodology

Several studies reveal that news spread faster in the network of Twitter compared to traditional news media [7,8]. Yet, a fundamental challenge is the quality of content published in OSNs. Distinguishing between credible and inaccurate information regarding emergency events is important, since misguidance or inability to timely detect useful information may have critical effects.

Fig. 1. Example of misinformation for Hurricane Sandy.

Objective: The objective of our work is to detect misinformation related to emergency events in the Twitter social network. We define as *misinformation* any false or inaccurate information that is spread either intentionally or unintentionally. An example of misinformation concerning the event of Sandy hurricane is presented in Figure 1.

Our Approach: Our approach for solving the problem of misinformation identification follows 3 discrete steps: (i) Given a set of tweets T related to an event and a set of users U that published at least one tweet $t \in T$, we conduct an extensive analysis on characteristics of tweets and users who published them. Our analysis focuses on a number of features and combinations of them and assists in identifying abnormal behaviors of users and characteristics of tweets. (ii) Based on the findings of the analysis, we extract extreme or suspicious behaviors and exploit them to filter tweets that are more likely to constitute misinformation. (iii) We then apply a series of learning algorithms implemented on Weka [9] to identify misinformative tweets, using supervised learning techniques.

2.1 Parameters

Tweet Features: Each tweet $t \in T$ is represented as a feature vector I_t that includes information about the tweet and the user who published it. Thus, each tweet $t \in T$ is characterized by the following information: **(i)** *Number of characters - words:* Short messages may not contain useful information, while long messages may cover unrelated topics. **(ii)** *Number of favorites - retweets - replies:* The popularity of a tweet may be an indication of its content. We expect that tweets of interest will be cascaded in the network and thus be more

retweeted or favorited. **(iv)** *Number of mentions - hashtags - URLs - media:* Features related to the structure of the tweet are considered to draw conclusions about the quality of the tweet.

User Features: For each user $u \in U$ that published at least one tweet $t \in T$ we consider the following characteristics: **(i)** *Number of followers - followees:* Trustworthy users such as news agencies are expected to have many followers [7], while spammers may have more followees. We define as followees the number of users an account follows. **(ii)** *Followers-Followees Ratio (FF-Ratio):* We compute the FF-Ratio of a user u as $FF\_Ratio = followers(u)/(followers(u) + followees(u))$. **(iii)** *Total tweets - Tweets during the event:* We suspect illegitimate users may be more active for a short time (e.g. the time of the event), thus we also consider the number of tweets users publish. **(iv)** *Days Registered:* The days the user is registered in the network before publishing a tweet. Recent users have greater chances of being spammers in contrast to older ones.

Additional Features: Finally, for each tweet we extract the following set of features: **(i)** *URLs to Tweets (UtT) - Media to Tweets (MtT):* For tweets published by a user we compute the ratio of tweets containing URLs and Media separately. We suspect that users frequently publishing URLs are candidate spammers, while media may be irrelevant to the event. **(ii)** *Followers to Replies (FtR) - Retweets (FtRt) - Favorited (FtFav):* Less popular tweets may indicate disapproval from followers. Therefore we consider the ratio of followers to the features indicating the popularity. **(iii)** *Average Tweets per Day (ATpD):* The average number of tweets published by user u that is registered $d\_a(u)$ days in the network is computed as $ATpD = t(u)/d\_a(u)$, where $t(u)$ is the total number tweets u published. **(iv)** *Positive / Negative / Average Sentiment:* We use SentiStrength [10] to extract the positive and negative sentiment rate of the tweet text and compute the average sentiment.

3 Data Analysis

In order to evaluate the performance of our approach for detecting misinformation during emergency events we used a dataset of tweets related to the Sandy Hurricane, a major emergency event that unfolded in 2012, from October 22 to November 2, and severely affected the area of New York City [3]. Tweets related to the event were collected based on the keywords "sandy" and "hurricane", as described in [3]. We then use the findings of the analysis to decide which values constitute a possible indication of misinformation.

Analysis of User Characteristics: In Figures 2 and 3 we present the number of tweets published by users, both total and during the event. We split the number of tweets in buckets of 100, i.e., bucket 0 contains the number of users that published 1 to 99 tweets in total. The Power Low distribution shown in the Figures is in accordance to findings of Bagrow *et al.* in [7]. Most of the users

[3] http://en.wikipedia.org/wiki/Hurricane_Sandy

published tweets related to the event with a frequency of 60 to 1000 seconds, while there are users that published more than one tweet per minute. The number of users' followers and followees are presented in Figures 4 and 5 respectively. The trend is similar for both connection types, with the majority of users having few followers and followees. The peak in the number of users that have up to 2000 followees is due to Twitter policy that limits the users to follow up to 2000 users and is later differentiated based on the followers to followees fraction. In Figure 6 we also present the FF-Ratio. The FF-Ratio approaches a Gaussian distribution with most users having an average ratio of around 0.5, meaning that they have equal amount of followers and followees, although we can observe another peak from 0.9 to 1.

Analysis of Tweet Characteristics: In Figure 7 we present an analysis of the number of words contained in a tweet. The majority of the tweets include 20 to 120 characters and 5 to 20 words. We further consider retweets, favorites and replies of a tweet to determine its popularity. As observed in Figures 8 and 9, the number of retweets and favorites follows a power law distribution. Finally, most of the tweets have fewer than 20 replies, but after a point the tweets containing more than

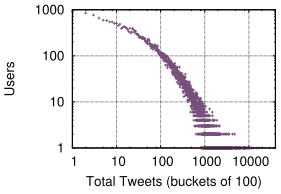

Fig. 2. User total tweets.

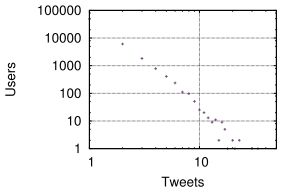

Fig. 3. User Sandy Tweets.

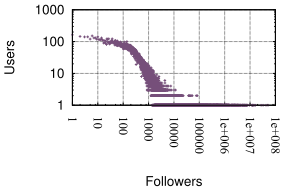

Fig. 4. User Followers.

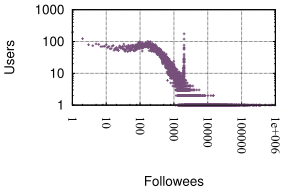

Fig. 5. User Followees.

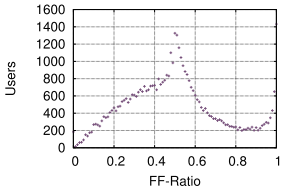

Fig. 6. User FF-Ratio.

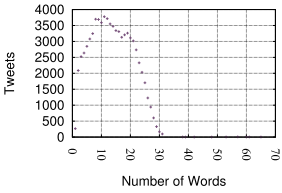

Fig. 7. Words in tweets.

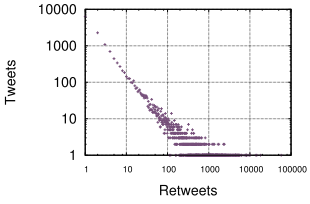

Fig. 8. Retweets received by tweets.

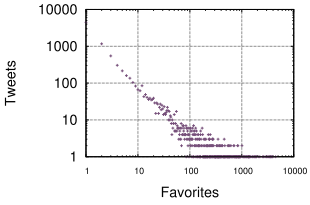

Fig. 9. Favorites received by tweets.

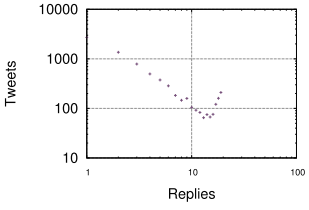

Fig. 10. Replies received by tweets.

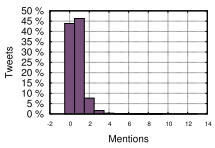

Fig. 11. Tweets with mentions.

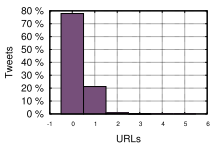

Fig. 12. Tweets with URLs.

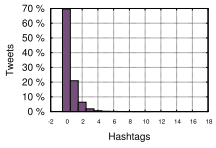

Fig. 13. Tweets with hashtags.

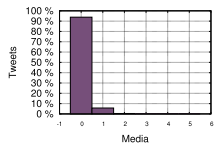

Fig. 14. Tweets with media.

20 replies rises (Figure 10). Figures 11 through 14 depict the number of mentions, URLs, hastags and media presented in a tweet. Most of the tweets contain at most one mentions and the majority of the tweets related to the event contain no link, while there are tweets with over two links. 70% of the tweets contain no hashtags, while the number of tweets containing a media is restricted to less than 10%.

4 Experimental Evaluation

We evaluated the performance of our approach on 80294 tweets related to hurricane Sandy from 59660 users. In the first set of experiments we focus on estimating the performance of the filtering process. By applying the filters of Table 1, 12955 tweets are returned. We manually annotate a sample of 4000 randomly selected tweets among them. Humorous, irrelevant and deleted tweets and accounts are considered as misinformation (assuming they are reported or deleted due to violations [11]). For 176 of the tweets we could not draw conclusions. Out of the remaining 3824 tweets, 898 constitute misinformation. Since tweets are randomly selected, we conclude that over 23% of the filtered tweets are indeed misinformation. We also annotated 4000 random tweets among those that did not meet the filtering criteria. For the 3559 tweets that could be classified, 212 are identified as misinformation, i.e., less than 6%. Overall, 1110 out the total 7383 labelled tweets constitute misinformation. The filtering process captures 898, yielding recall values of over 81%.

Supervised Learning: We exploited a set of different supervised learning algorithms implemented on Weka [9] to evaluate the performance of information identification on the set of features we considered. We use 10-fold cross validation for evaluating the classification results. The labelled dataset of the 3824 filtered tweets is used as input to Weka. In Table 2 we present the classification results. Weighted Average F1 measure indicates that Bootstrap Aggregating has the best performance. Regarding average precision, Random Forest achieves better results, with 0.792 average precision.

Table 1. Filters applied during the filtering process.

Words ≥ 30	Characters $== 140$	Favorites $(\geq 2$ && $\leq 10)$ $\|\|$ (≥ 1100)
Hashtags ≥ 4	Mentions ≥ 4	Retweets $((\geq 2$ && $\leq 10)$ $\|\|$ $(\geq 1000))$
Media ≥ 2	Followees $(\leq 10$ $\|\|$ $\geq 100000)$	Followers $(\leq 10$ $\|\|$ $\geq 200000)$
Replies ≥ 11	URLs ≥ 3	Followers/Followees ≥ 30000
Event tweets ≥ 7	Total Tweets ≥ 500000	Interval $(\leq 300sec$ $\|\|$ $\geq 70000sec)$

Table 2. Summary of Classification using Supervised Learning Algorithms.

	Precision	Recall	F-Measure	Precision	Recall	F-Measure
	Bayes Network			**J48**		
Credible	0,845	0,834	0,839	0,825	0,889	0,856
Misinformation	0,480	0,500	0,490	0,516	0,385	0,44
Weighted Avg.	0,759	0,755	0,757	0,752	0,771	0,758
	k-Nearest Neighbors			**Random Forest**		
Credible	0,800	0,951	0,869	0,821	0,958	0,884
Misinformation	0,586	0,225	0,325	0,699	0,318	0,438
Weighted Avg.	0,750	0,781	0,741	0,792	0,808	0,779
	Adaptive Boosting			**Bootstrap Aggregating**		
Credible	0,839	0,888	0,863	0,828	0,931	0,877
Misinformation	0,549	0,447	0,493	0,622	0,372	0,466
Weighted Avg.	0,771	0,784	0,776	0,780	0,799	0,780

Real-Time Misinformation Identification: The values of retweets, favorites and replies are unknown at the time the tweet is published. Thus, to evaluate the performance of our approach in timely detecting misinformation we conducted another set of experiments ignoring the above attributes and features related to them. The results for Bootstrap Aggregating and Random Forest are presented in Table 3. The table shows that precision and recall drop slightly. Still, weighted average precision is over 0.77, indicating the approach is appropriate for real time misinformation identification. The filtering process requires just 963ms and adding the execution times of the algorithms, less than 2 seconds are needed to efficiently extract tweets containing misinformation, proving that the method is efficient under real-time constraints.

Table 3. Classification with features known at run time.

	Precision	Recall	F-Measure	Precision	Recall	F-Measure
	Random Forest			**Bootstrap Aggregating**		
Credible	0,812	0,949	0,875	0,816	0,951	0,879
Misinformation	0,631	0,286	0,394	0,655	0,301	0,412
Weighted Avg.	0,770	0,793	0,762	0,778	0,799	0,769
Execution time:			0.41 sec			0.89 sec

5 Related Work

Castillo et al. [5] aim at automatically detecting credibility of information in the network of Twitter. They use a number of features related to tweets and supervised learning to distinguish between newsworthy or false news and later perform credibility analysis. Gupta et al. [12] present TweetCred, an extension of the previous work that enables users feedback. Xia et al. [6] also study the

problem of information credibility on Twitter after the event is detected and relevant tweets are retrieved. Bosma et el. [13] suggest a framework for spam detection using unsupervised learning. Anagnostopoulos et al. [14] study the role of homophily in misinformation spread. McCord and Chuah [15] are using traditional classifiers to detect spams on Twitter. Stringhini et al. in [11] aim at identifying spammers in social networks. Identifying spammers on the network of Twitter is also the objective of Benevenuto et al. in [16]. They extract features of the account that may be indication of spamming behavior and use SVM learning model to verify their approach. Zubiaga and Ji in [3] and Gupta et al. [4] focus on the credibility of images propagated in the network during emergency events. They consider a number of features related to the image and the tweet. Budak et al. [17] address the problem of misinformation spread limitation by performing an extensive study on influence limitation. Faloutsos [18] developed a botnet-detection method and a Facebook application, called MyPageKeeper, that quantifies the presence of malware on Facebook and protects end-users. Ghosh et al. [19] examine suspended accounts on Twitter and investigate link farming and finally discourage spammers to acquire large number of following links. Thomas et al. [1] identify the behaviors of spammers by analyzing tweets of suspended users in retrospect. Mendoza et al. [20] focus on cascades of tweets during emergency events and study the propagation of rumours. They conclude that this defers from the propagation of news tweets and it is possible to detect rumors by aggregating analysis on the tweets. Liu et al. [21] propose a hybrid model that utilizes user behavior information, network attributes and text content to identify spams.

6 Conclusions

In this work we presented a methodology for identifying misinformation on social networks during emergency events. As we illustrate in our experiments our approach manages to correctly identify misinformation achieving accuracy of up to 77%. The filtering process suggested in this work identifies over 81% of misinformative tweets. Our approach is fast and effective and timely identifies misinformation, offering the ability to limit the spread in the network.

Acknowledgment. This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program:Thalis-DISFER, Aristeia-MMD, Investing in knowledge society through the European Social Fund, the FP7 INSIGHT project and the ERC IDEAS NGHCS project.

References

1. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect: an analysis of twitter spam. In: Internet Measurement Conference, pp. 243–258 (2011)
2. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B.Y.: Detecting and characterizing social spam campaigns. In: ACM Conference on Computer and Communications Security, pp. 681–683 (2010)
3. Zubiaga, A., Ji, H.: Tweet, but verify: Epistemic study of information verification on twitter (2013). CoRR, vol. abs/1312.5297
4. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Ser. WWW 2013 Companion (2013)
5. Castillo, C., Mendoza, M., Poblete, B.: Predicting information credibility in time-sensitive social media. Internet Research **23**(5), 560–588 (2013)
6. Xia, X., Yang, X., Wu, C., Li, S., Bao, L.: Information credibility on twitter in emergency situation. In: Chau, M., Wang, G.A., Yue, W.T., Chen, H. (eds.) PAISI 2012. LNCS, vol. 7299, pp. 45–59. Springer, Heidelberg (2012)
7. Bagrow, J.P., Wang, D., Barabasi, A.-L.: Collective response of human populations to large-scale emergencies (2011). CoRR, vol. abs/1106.0560
8. Guy, M., Earle, P., Ostrum, C., Gruchalla, K., Horvath, S.: Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. In: Cohen, P.R., Adams, N.M., Berthold, M.R. (eds.) IDA 2010. LNCS, vol. 6065, pp. 42–53. Springer, Heidelberg (2010)
9. Weka. http://www.cs.waikato.ac.nz/ml/weka/
10. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment in short strength detection informal text. J. Am. Soc. Inf. Sci. Technol. **61**(12), 2544–2558 (2010)
11. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: ACSAC, pp. 1–9 (2010)
12. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: Tweetcred: real-time credibility assessment of content on twitter. In: Aiello, L.M., McFarland, D. (eds.) SocInfo 2014. LNCS, vol. 8851, pp. 228–243. Springer, Heidelberg (2014)
13. Bosma, M., Meij, E., Weerkamp, W.: A framework for unsupervised spam detection in social networking sites. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 364–375. Springer, Heidelberg (2012)
14. Anagnostopoulos, A., Bessi, A., Caldarelli, G., Vicario, M.D., Petroni, F., Scala, A., Zollo, F., Quattrociocchi, W.: Viral misinformation: The role of homophily and polarization (2014). CoRR, vol. abs/1411.2893
15. McCord, M., Chuah, M.: Spam detection on twitter using traditional classifiers. In: Calero, J.M.A., Yang, L.T., Mármol, F.G., García Villalba, L.J., Li, A.X., Wang, Y. (eds.) ATC 2011. LNCS, vol. 6906, pp. 175–186. Springer, Heidelberg (2011)
16. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on twitter. In: CEAS (2010)
17. Budak, C., Agrawal, D.: Abbadi, A.E.: Limiting the spread of misinformation in social networks. In: WWW, pp. 665–674 (2011)
18. Faloutsos, M.: Detecting malware with graph-based methods: traffic classification, botnets, and facebook scams. In: WWW (Companion Volume), pp. 495–496 (2013)

19. Ghosh, S., Viswanath, B., Kooti, F., Sharma, N.K., Korlam, G., Benevenuto, F., Ganguly, N., Gummadi, P.K.: Understanding and combating link farming in the twitter social network. In: WWW, pp. 61–70 (2012)
20. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: can we trust what we rt? In: Proceedings of the First Workshop on Social Media Analytics, ser. SOMA 2010, pp. 71–79. ACM, New York (2010)
21. Liu, Y., Wu, B., Wang, B., Li, G.: Sdhm: a hybrid model for spammer detection in weibo. In: 2014 IEEE/ACM International Conference on ASONAM, pp. 942–947, August 2014