Predictive Models for Early Detection of Hoax Spread in Twitter

Didier Henry, Erick Stattner

Predictive Models for Early Detection of Hoax spread in Twitter

Didier Henry
LAMIA
Université des Antilles
Pointe-A-Pitre, France
didier.henry@univ-antilles.fr

Erick Stattner
LAMIA
Université des Antilles
Pointe-A-Pitre, France
erick.stattner@univ-antilles.fr

***Abstract*—Nowadays social media are widely used daily to access to news. Indeed, the social media network allows a fast and wide spread of news. Unfortunately these platforms used by millions of people are not immune to misinformation because everyone can be a source of information. Rumors of celebrities death on social media spread very widely in a short time and are hardly verifiable. These kinds of rumors can lead to worrying or stressful situations, and may also have economic or political repercussions. In this work, we have addressed the problem of death hoax diffusion on the social media Twitter. We have collected data related to 25 rumors (false and true) of the death of well-known celebrities on Twitter. Then, we have observed temporal differences and commonalities between true and false rumors in terms of diffusion dynamic, messages and user characteristics. From these empirical observations, we have trained several models to classify early true rumors and hoaxes. We have obtained a rate of correct classification of 0.9 from 20 minutes after the beginning of the diffusion.**

***Index Terms*—social media, fake news, data analysis, predictive model, death hoax**

I. Introduction

Social media are widely used daily to access news. Unfortunately these platforms, used by millions of people, are not immune to misinformation and regularly see the emergence of rumours. Because of strong consequences (social, political, economical, sanitary, etc.) that rumours can have on individuals and society, many works have attempted to study these phenomena. For instance, Dayani et al. [1] have performed a retrospective analysis on 5 rumors and have noted that rumor detection seems no have a correlation with users based features. Chen et al. [2] have observed that a convolutional neural network is more appropriate than a recurrent neural network for rumor detection task and obtain accuracy near 0.7. Zhao et al. [3] have shown that decision tree outperforms SVM with accuracy above 0.7. Poddar et al. [4] have proposed a neural approach to detect rumor veracity. Their prediction model has achieved accuracy near 0.798. Recently, Liu et al. [5] has reached accuracy about 0.85 by using recurrent and convolutional networks for early detection of fake news on Twitter. One of the most common cases of social media rumours is the announcement of the celebrity death. In this particular case regarding false information about death announcement, we often talk about *death hoax*. In this work, we have addressed the problem of death hoax diffusion on the social media Twitter. Unlike the vast majority of works that focus on rumour diffusion by proposing more and more complex models, in this work, we have adopted a predictive modelling approach in order to detect quickly after the beginning of the diffusion process, death hoax cases. For this purpose, we have collected real data on several well-known celebrities deaths announced on Twitter, including real and hoax cases. Thus, our objective is to propose a methodology for identifying quickly hoax cases in social media. We start by describing the methodology we propose to collect data on both real and hoax deaths on Twitter. Unlike other works that focus on rumor detection, we extract here three kinds of original attributes, related to (i) the diffusion process, (ii) the content of the tweets published and (iii) the users who publish messages on the targeted event, which are likely to have a predictive value for the purpose of death hoax classification. In a second step, we analyze all the features extracted for highlighting temporal differences between true and false death rumours according to these three kinds of features. Finally, from these empirical observations, we have selected relevant features and trained several predictive models to classify death hoaxes. The results obtained demonstrate the good predictive value of the features identified since good performances are obtained from the 20th minute after the beginning of the diffusion. This paper is organised as follows. Section II describes the methodology we propose to collect data and extract features related to users and tweets. Section III is devoted to the analysis work we have conducted to highlight differences and commonalities between real and fake celebrity death rumours. Section IV describes the results we obtained with different predictive models. Finally, Section V concludes and presents our future works.

II. Methods

A. Data collection

Our aim is to observe and understand the differences between the true and false rumors of death of celebrities at diffusion, tweets and users level. The main challenge we have encountered was having access to old tweets related these rumors. In order to collect old tweets, we have used the Twitter search engine[1] by specifying the celebrity name, the date of

[1]https://twitter.com/search-advanced

TABLE I
DATASETS CHARACTERISTICS.

Date	Celebrity	Tweets	Users	Rumor
20/04/2018	Avicii	262119	219853	True
26/12/2016	Britney Spears	18126	12299	False
27/01/2012	Cher	21009	16131	False
20/07/2017	Chester Bennington	252573	218031	True
10/06/2016	Christina Grimmie	126214	88142	True
13/07/2013	Cory Monteith	470479	373762	True
05/06/2016	Jack Black	8285	5383	False
29/03/2011	Jackie Chan	119872	94644	False
24/01/2010	Johnny Depp	10344	8920	False
16/11/2017	Lil Peep	125640	103583	True
04/03/2019	Luke Perry	120373	108302	True
07/09/2018	Mac Miller	456700	410069	True
02/01/2013	Megan Fox	25344	23124	False
03/09/2012	Michael Duncan	10379	10050	True
17/02/2013	Mindy McCready	50943	32711	True
16/12/2010	Morgan Freeman	10471	7839	False
29/03/2012	Patrick Dempsey	3964	3339	False
30/11/2013	Paul Walker	104355	92837	True
07/06/2016	Roger Goodell	7227	5359	False
26/02/2012	Rowan Atkinson	38155	31986	False
20/06/2011	Ryann Dunn	229679	182106	True
21/06/2018	Sophie Grador	7629	6070	True
08/09/2015	Terry Gilliam	2808	2326	False
16/03/2014	Wayne Knight	4232	3054	False
18/06/2018	XXXTentacion	301287	233325	True

the rumor and the English language in the query. By following this method, we have got a large number of tweets related to 25 rumors related to death of celebrities of which 13 real (i.e. the celebrity is really dead) and 12 fake (i.e. it is a hoax about the death of the celebrity). Although there are more cases of celebrity death rumors[2] we have selected these 25 rumors because we have been able to recover a significant amount of messages for these cases. Indeed, in the majority of the cases of false news the messages are often deleted by the users who note that it was a false information. Therefore, very few cases on the false rumor on this subject can be studied. In addition, we chose to balance the dataset with as many truths as false rumors, it is for all these reasons that the study covers a relatively small number of examples. Then, we have collected account information for each user such as his current number of tweets posted on Twitter, his current number of followers, followings, and his account date creation. The characteristics of datasets obtain are summary in the Table I. Moreover, we have also collected tweets in Spanish and French languages by using the same method into to observe if rumors spread or not in different languages.

B. Data analysis

For each rumor, a human expert has noted the date of the first tweet recounting the celebrity death. We have considered this date as the beginning of the rumor diffusion. In a first step, we have taken into account several diffusion characteristics:

- *Nb_Tweets_n*: The number of tweets posted at time t_n,
- *Nb_Retweets_n*: The number of retweets posted at time t_n,
- *Percent_Tweets_n*: The percentage of tweets posted at time t_n defines as $\frac{Nb\_Tweets\_n}{nbTotalMessages_n}$, where $nbTotalMessages_n$ is the total number of messages posted at time t_n,
- *Percent_Retweets_n*: The percentage of retweets posted at time t_n defines as $\frac{Nb\_Retweets\_n}{nbTotalMessages_n}$, where $nbTotalMessages_n$ is the total number of messages posted at time t_n,
- *Percent_New_Messages_n*: The percentage of new messages posted at time t_n defines as $\frac{nbNewMessage_n}{nbTotalMessages_n}$, where $nbTotalMessages_n$ is the total number of messages posted at time t_n,
- *MultiLanguage*: a boolean equal to true if the rumor spread also in French and Spanish in within one hour.

In a second step, we are interested in message characteristics. Firstly, we perform a content analysis of messages to extract the following attributes:

- *Tw_Length_n*: The average number of characters of messages at time t_n,
- *Tw_URL_n* and *Tw_URL_D_n*: The percentage of messages containing URL and distinct URL (pointing to sites other than Twitter) respectively at time t_n,
- *Tw_RIP_n*: The percentage of messages containing the acronym RIP in the tweet (several forms have been taken into account) at time t_n,
- *Tw_HoaxOrHack_n* and *Tw_Alive_n*: The percentage of messages containing "hoax" or "hack" and those containing "alive" at time t_n,
- *Tw_SadSmiley_n*: The percentage of messages with at least a sad smiley at time t_n,
- *Tw_Age_n*, *Tw_Exclamation_n*, *Tw_Question_n*, *Tw_Mentions_n* and *Tw_Hashtags_n*: The percentage of messages containing the age of the celebrity, a least an exclamation, a least a question mark, a least an user mention, a least a hashtag respectively at time t_n.

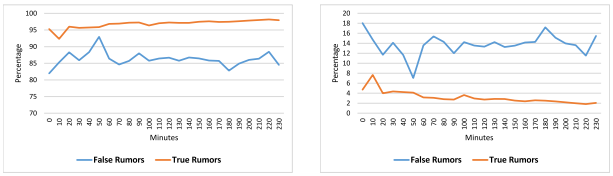
Secondly, we have used TextBlob[3] an API in Python language in order to extract the sentiments of tweets in terms of polarity and subjectivity. We have considered three classes for the message polarity and three classes for the message subjectivity. Then, we have extracted for each t_n the messages distribution in each class. Thirdly, we used an emotion recognition tool proposed by Colnerič and al. [6] to extract the emotions expressed in the tweets. This tool[4] returns the relative emotion of a tweet (joy, fear, sadness, anger, surprise or disgust). Then, we have extracted for each t_n the messages distribution in each class. In a third step, we are interested in user characteristics. Given that we have collected messages dating from several months or years and that the twitter network is very scalable, we have chosen to extract the following attributes:

- *TweetsPerDay_n*: The average ratio of the number of tweets posted on the number of elapsed days since the account creation at time related to Twitter users having posted a least a message at t_n,
- *RatioFF_n*: The average ratio of the number of followers on the number of followings defines as $\frac{nbfollowers}{nbfollowers+nbfollowings+1}$ related to Twitter users having posted a least a message at time t_n,

[2]thewrap.com/celeb-reported-dead-celebrity-death-hoax-barbara-bush-hillary-clinton-jack-black

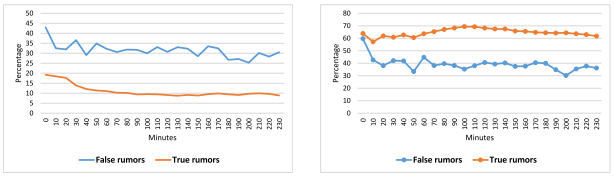
[3]http://textblob.readthedocs.io/en/dev/

[4]https://github.com/nikicc/twitter-emotion-recognition



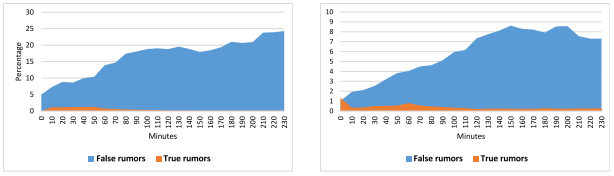
(a) (b)

Fig. 1. Evolution of the average percentage of tweets (a) and retweets (b) over time.



(a) (b)

Fig. 2. Evolution of the average percentage of tweets containing question mark(s) (a) and sad tweets (b) over time.



(a) (b)

Fig. 3. Evolution of the average percentage of tweets containing "hoax" or "hack" (a) and those containing "alive" (b) over time.

- *RatioTF_n*: The average ratio of number of tweets posted on the number of followers defines as $\frac{nbtweets}{nbfollowers+1}$ related to Twitter users having posted a least a message at time t_n,
- *RatioTP_n*: The average ratio of the number of photos or videos posted on the number of tweets posted as $\frac{nb\_photos\_and\_videos}{nbtweets+1}$ related to Twitter users having posted a least a message at time t_n,
- *Verified_n*, *ProfileImage_n*, *BannerImage_n*: The percentage of Twitter users having a verified account, a profile image different from the default, a banner image different from the default respectively at time t_n.

III. Data Exploration

A. Diffusion dynamics and other diffusion languages

Firstly, we have focused on diffusion dynamics of rumors. By observing the evolution of the average number of messages posted we have noted that true rumors seem to spread more largely than false rumors. In addition, we have noticed that in both cases Twitter users tend to diffuse information through tweets (see Figure 1 (a)) but their retweet more when it comes to false rumors (see Figure 1 (b)). We have observed that the average percentage of new messages posted is more weak for real rumors that is to say users tend to diffuse tweets already posted rather than post their own message. Moreover, we have noticed that 59% of false rumors have crossed the language barrier within one hour against 100% for true rumors.

B. Tweets content

Secondly, we were interested in differences related to the tweets content. We have noted that tweets related to false rumors and true rumors are similar in terms of the number of characters not matter the moment of diffusion. In addition, we have remarked that the percentage of tweets containing a least a sad smiley is also quite similar since the beginning of the diffusion. These characteristics don't seem suitable in prediction perspective. Then, we have noticed that users seem more suspicious about false rumors because they post more tweets contained at least a question mark since the beginning of the diffusion and several minutes after (see Figure 2 (a)).

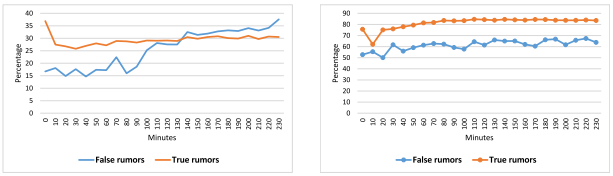
Next, we have remarked that the average number of tweets containing the celebrity age for the true rumors of death is largely greater than false. Indeed, in average more than 5% of tweets contain it even after one hour for the true rumors against only 2% for the false ones. By observing the number of tweets posted with the RIP acronym, we have noted that the percentage of tweets containing this acronym is very close until the first 30 minutes. Then, we have noted that the percentage of messages containing a least a hashtag is greater for real rumors than fake while the percentage of messages containing a least a mention is greater for fake rumors than real. Finally, we have remarked that differences are clear between fake and real rumors since the first moment for tweet including the word "hoax" or "hack" and only 10 minutes after for those including the word "alive" (see Figure 3).

C. Tweets sentiments and emotions

Thirdly, we were interested in commonalities and differences related to the tweets sentiments and emotions. We have observed that the average percentage of subjective tweets is higher for true rumors after 30 minutes. Indeed, it seems the first tweets concern the announcement of the death, next people paying homage to a celebrity by expressing their own feelings. In addition, we have noted that it seems that there is little difference concerning the polarity of tweets. Moreover, we have noted that the number of tweets identify as sad is higher also for true rumors (see Figure 2 (b)). We may suppose that there are less sadness tweets posted as it is about false rumors because users discover that information is a hoax.

D. Users characteristics

Finally, we were interested in commonalities and differences related to the user characteristics. By considering the user distribution according to the ratio of the number of followers on the number of followings we have noted that there are not clear differences at the first moments of diffusion. Thus, this user feature does not seem relevant to differentiate between real death rumors and fakes at the first moments. In contrast, the ratio of the number of tweets posted on the number of followers seem suitable to distinguish real death rumors and



(a) (b)

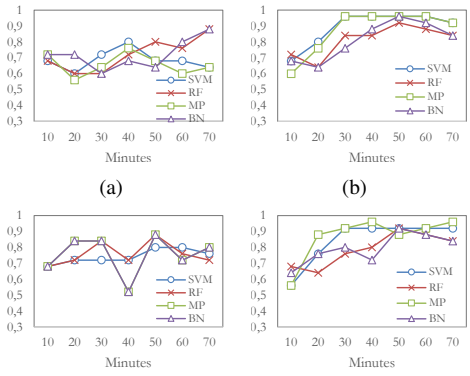
Fig. 4. Evolution of the average percentage of users having a *RatioTF* $\in [0-10[$ (a) and users posting between 0 and 25 tweets on average (b) per day over time.

fake at the first moments (see Figure 4 (a)). In addition, we have remarked that the average percentage of low prolific users is higher for true rumors than false since the beginning of the rumor diffusion and more than 3 hours later (see Figure 4 (b)).

IV. Predictive Modeling of Death Hoax

The trends observed in the previous section suggest that the identified attributes could have a predictive value to classify quickly death hoax or true information after the beginning of the broadcast process. Thus to check the predictive value of the attributes extracted from data, we adopt in this section a predictive modelling approach in order to identify the hoax over time. The objective is to identify the most relevant attributes for the predictive objective and to identify if the prediction is possible sufficiently early. For this prediction purpose, we use cross validation and four different kinds of classifiers have been compared in this study: Bayesian network (BN), Random Forest (RF), Support Vector Machine (SVM) and Multilayer Perceptron (MP). The goal is to compare the time required to identify death hoax according to the classifiers used and the type of attributes used. Figure 5 presents the results obtained when attributes used concern (a) diffusion, (b) tweets, (c) users and (d) all.

We can globally observe that the correct classification rate is relatively high and increases with time (see Figure 5 (b) and (d)), which confirms the predictive value of the identified attributes. Indeed, in the first 10 minutes after the beginning of the diffusion process, all models used provide a correct classification rate higher than about 50%. Moreover, this rate is higher when tweet attributes are used, and even higher when all attributes are exploited. Indeed, when diffusion or user attributes are used (see Figure 5 (a) and (c)) the correct classification rate is always lower than 90%. However, when tweets attributes are used the correct classification rate is about 90% from 30 minutes, while it reaches 90% from 20 minutes if all attributes are used. If we observe in detail the performances of the models, the results are very different according to the type of attributes used. For instance, when diffusion or user attributes are used independently, all models provide approximately the same performances. However, when tweets attributes are considered Multilayer Perceptron and SVM provide the best results and reaches about 90% correct classifications from the 30th minute after the beginning of the diffusion.



(a) (b) (c) (d)

Fig. 5. Evolution of the average rate of correct classifications over time when attributes used concern (a) diffusion, (b) tweets, (c) users and (d) all

V. Conclusion

In this paper, we have addressed the detection problem of celebrity death hoaxes in social media by collecting and analyzing data related to 25 rumors of the death of well-known celebrities on Twitter. We have observed several differences and commonalities between true and false rumors in terms of diffusion dynamic, messages and users characteristics. Then, we have used four different kinds of predictive approaches by selecting suitable features to classify rumors. We have obtained a true positive rate to 0.9 only 20 minutes after the rumor diffusion with a Multilayer Perceptron. As perspectives, we plan to apply your method on a large number of datasets. In addition, we will observe differences between real and fake celebrity death rumors in other languages in order to propose suitable prediction models. As long-term perspectives, we intend to take into account other characteristics such as the number of suspicious users account and the spread dynamic of rumors through the network to improve prediction results. Furthermore, we plan to study the phenomena on other social media platforms and propose new predictive models.

References

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

[2] Y.-C. Chen, Z.-Y. Liu, and H.-Y. Kao, "Ikm at semeval-2017 task 8: Convolutional neural networks for stance detection and rumor verification," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 465–469.

[3] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1395–1405.

[4] L. Poddar, W. Hsu, M. L. Lee, and S. Subramaniyam, "Predicting stances in twitter conversations for detecting veracity of rumors: A neural approach," in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2018, pp. 65–72.

[5] Y. Liu and Y.-F. B. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] N. Colneriĉ and J. Demsar, "Emotion recognition on twitter: Comparative study and training a unison model," *IEEE Transactions on Affective Computing*, 2018.