Using multi-modal approach to detect false information in tweets during disasters

Anonymous Author(s)
Affiliation
Address
email

Abstract

With growing role of social media as critical components of information, emergency preparedness, response and recovery during disasters, an increased attention is needed to identify and counter the spread of false information and rumors in times of such public emergency. Almost all major disaster events in past few years, viz. Hurricane Sandy, Nepal Earthquake, Chennai floods etc. have witnessed rumors being spread virally on various social media platforms. Such false and incorrect information can lead to chaos and panic among people on the ground and have serious detrimental outcomes for public safety. While prior works have proposed automated techniques to detect false information online, these techniques primarily fail to look beyond the textual content. In this work, we propose a multi-modal model that can detect the credibility of a post by effectively capturing the semantics of the text along with the features of associated piece of multimedia in it.

1 Introduction

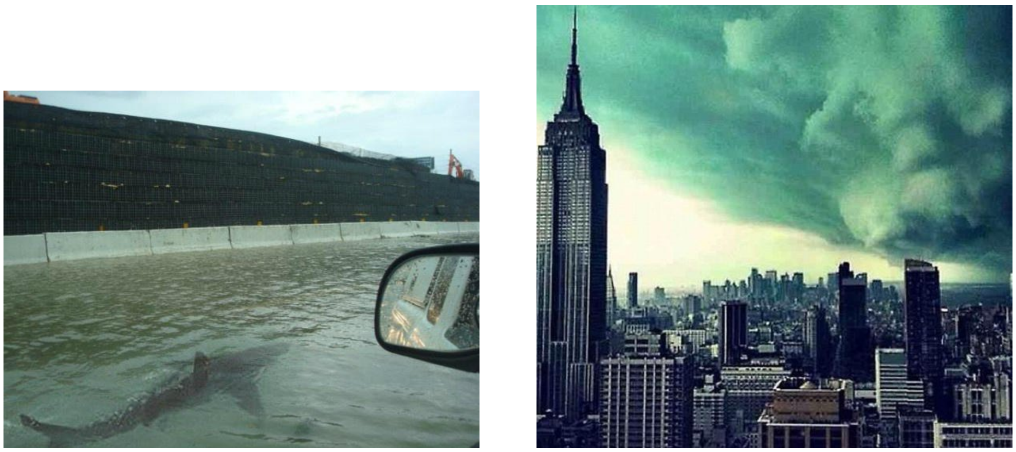
In times of public emergency like natural disasters, there's a natural desire to seek as much information as possible in order to take the best possible decision. In recent days, when social media has become primary resource of news and information during such events [2], rumors and false information can spread fast on social media that risk sowing panic. Results of various studies that investigate the impact of rumors on social media during disasters, suggest that such activities can prove to be dangerous since they can put people at risk and negatively impact the efforts of emergency workers and aid organizations [3].

An example of fake multimedia posted during Hurricane Sandy can be seen in Fig.1. where digitally manipulated image of shark in streets and stormy New York skyline were spread virally which aggravated the panic amongst the mass. In some cases, the consequences of fake content reaching a very large part of the population can be quite severe. For instance, fake images became popular on social media after the Malaysia Airlines passenger flight disappeared on 8th March. During the investigation of the plane trace, false alarms that the plane was detected came up. Taking into account how sensitive the case was, the circulation of this content deeply affected the people directly involved in it, such as the families of the passengers, causing emotional distress.

Therefore, the need for debunking hoaxes and false information during public emergencies is inevitable. The primary step in this process requires detecting and tracking false information online. Previous works done on the same focus on linking user credibility to false information [14] [15] [16]. While user credibility can be an important source of information to detect scams etc., according to a recent study, it has been found, that people spreading false information during disasters have no idea that they are fake. Viral hoaxes spread through well-intention community members as well. In addition, most of the current methods are trained primarily using textual features [5] [6] [7] [8] [13]. These works tend to overlook a very important aspect of fake tweets during such events-

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

multimedia content. Disaster sociologists reveal that social media gets flared up with photos during public emergencies since they aid the job of giving out situational information and helps people take effective decisions to save their lives [4]. It is exactly this setting, when the risk of fake content becoming widely disseminated is the highest. Various incidents in the past, for e.g. Nepal Earthquake, Indonesian Tsunami etc. have witnessed the spread of fake images through social media which have resulted in an increase in panic and chaos amongst people in the affected region.



(a) A spliced image of shark in the streets

(b) A fake image of stormy New York Skyline

Figure 1: Examples of fake multimedia on twitter during Hurricane Sandy

Therefore, in this work we propose a multi-modal architecture which analyzes the tweet text as well as the images associated with it to detect the credibility of the information. Our work focuses on the problem of single post verification, i.e. classifying an individual content item as being fake or real. We also experiment with previous approaches and compare the results with our proposed model.

2 Related Work

Various solutions to filter out spam and malicious content from tweets have been studied and proposed. Existing studies aim at developing machine learning-based classifiers to automatically detect if a post viral in a social media environment is fake based on a variety of post characteristics. [5] makes use of a set of linguistic features such as special characters, emoticon symbols, sentiment positive/negative words, hashtags, etc., to classify a news story as fake or true. Beyond these features named entities are adopted in [6] and swear words and pronouns are examined in [7].

Besides text content, characteristics of source, users have also been explored by several studies. [5] utilizes a set of user characteristics on Twitter, e.g., number of followers, number of friends, registration age to detect fake news. [8] explores a similar set of user characteristics on Sina Weibo, the most popular social media site in China. [10] utilizes recurrent neural networks that capture temporal-linguistic features from a sequence of user comments to detect fake news.

Another path of this research focuses on extracting temporal features from propogation paths/trees of stories in a social network [9] [17]. There are also hybrid approaches that content-based, user based, and propagation-based features to detect fake news [18].

Unlike these works, [6] has focused particularly on examining the tweets during disasters. Inspired by this, we utilize a combination of image and text features to assess the credibility of information during public emergencies.

3 Dataset

Our dataset consists of tweets collected and aggregated from various different sources.

Table 1: **Descriptive Statistics for tweets used from Media Eval dataset. For each event, we report the total number of unique tweets(T_T) and the distribution of real and fake tweets (T_R, T_F)**

Event	T_T	T_R	T_F
HURRICANE SANDY	10206	4664	5542
BOSTON BOMBING	498	153	334
COLUMBIAN CHEM.	180	0	180
MA FLIGHT 370	501	0	501
BRING BACK OUR GIRLS	131	0	131

Table 2: **Descriptive Statistics for additional tweets extracted.**

Event	T_T	T_R	T_F
KERALA FLOODS 2018	675	382	293
AMAZON RAIN-FOREST	984	597	387
CHENNAI FLOODS	479	456	23
INDONESIA TSUNAMI	442	442	0

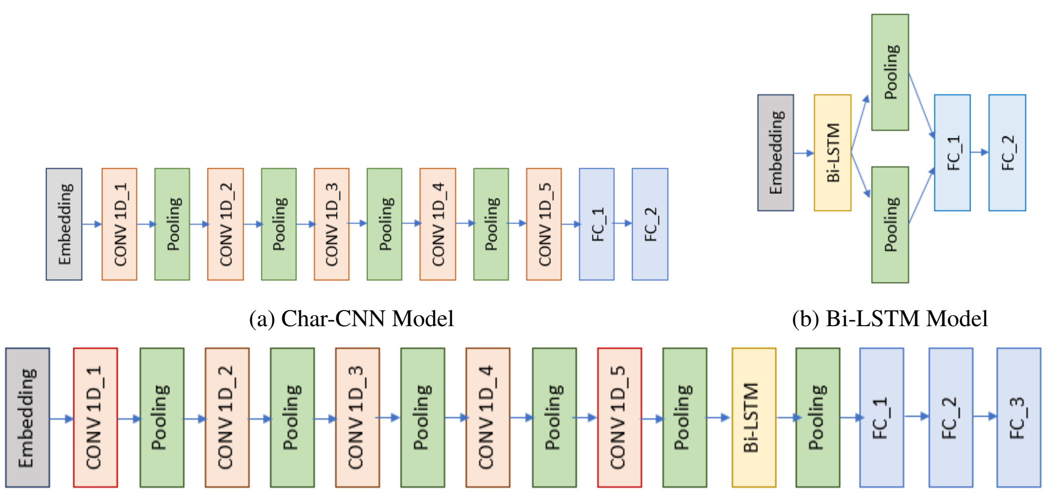
3.1 Publicly available dataset

The conducted experiments were partially based on the benchmark dataset released by MediaEval for their task of Verifying Multimedia Use. It consists of tweet IDs and image URLs of tweets collected around a number of widely known events or news stories. The tweets contain fake and real multimedia content that has been manually verified by cross-checking online sources (articles and blogs).

Out of the entire dataset, we extracted the tweets for the events associated with public emergencies like natural disaster, terrorist attack etc. Since a few tweets mentioned in the dataset have been removed, we were able to extract the content for 21919 IDs in total consisting of 10409 fake (48%) and 11510 real (52%) tweets. Around 11516 unique tweets were filtered out of these with Table 1 gives the descriptive statistics of the dataset.

3.2 Data collection

We further expand the existing data set by collecting additional data to add more capabilities to our modelling .We extracted multimedia tweets associated with recent natural calamities. Using defined keywords for each of the events mentioned in Table 2, we extracted about 1500 tweets for each of those and filtered out the unique and identified the ones associated with fake multimedia using a method similar to [1]. A total of 2580 tweets were extracted with 1877 real and 703 fake tweets extracted in this process.



(a) Char-CNN Model

(b) Bi-LSTM Model

(c) Bi-LSTM + CharCNN Model

Figure 2: Illustration of Deep-Learning (Text-only) models

4 Models and Evaluation

In this section, we describe our models and the results of our experiments with different models. The data includes tweets, the images associated with them, author attributes and features extracted from the tweets, for instance the number of hashtags, urls, etc. We make use of various combination of data for our experiments.

Table 3: Results of the experiments

Model	**Accuracy**
TRADITIONAL METHODS: TEXT + USER + TWEET BASED	
LR	0.81
SVM	0.58
TRADITIONAL METHODS: USER + TWEET BASED	
LR	0.68
SVM	0.67
DEEP LEARNING METHODS: TEXTUAL	
CHAR-CNN	0.91
BI-LSTM	0.88
BI-LSTM + CHARCNN	0.93
DEEP LEARNING METHODS: TEXTUAL+IMAGE	
VGG16+ BI-LSTM+CHARCNN	0.97

4.1 Baselines

Our baseline approach uses traditional models like SVM and Logistic Regression. We use TF-IDF (Char) based approach to generate features combined with tweet based features (number of hashtags, likes etc.) and the author attributes. We also test a simplified model by eliminating the TF-IDF Vectors from our set of features.

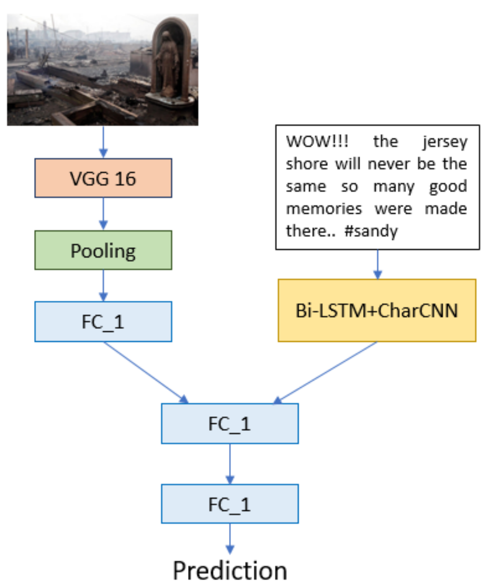


Figure 3: Illustration of Multi-Modal Model

4.2 Deep Learning Methods: Text Only

We experiment with a variety of architectures, the best of which are mentioned in Table 3. Our architectures used a Dropout of 0.5 with Adam Optimizer and Binary Cross-entropy as the loss function. An illustration of our deep-learning models can be seen in Fig. 2.

4.3 Deep Learning Methods: Combining Image and Textual Features

Though the experiments with other deep learning architectures returned good results, we applied a fusion strategy to enhance the accuracy of the model. Let us denote out dataset as T = (t_1, i_1), (t_2, i_2), ..., (t_n, i_n) where each tuple consists of a tweet text t_i and image corresponding to it denoted by i_i. Defining the input as x_i, our model then can be represented as:

$$\mathrm{f}(x_i) = \mathrm{g}(\mathrm{C}(t_i), \mathrm{V}(i_i))$$

where C and V are neural architectures extracting features from tweet text and image respectively. This is then passed through another network which fuses the features and gives prediction as the output.

We use a VGG 16 architecture, pre-trained on Imagenet dataset. The image features extracted are then passed through a dense layer so that the model can learn more complex functions and classify for better results. For the text side, we used the same BiLSTM+CharCNN model (as mentioned in section 3.2) for feature extraction.

These features are then combined and passed through two another fully connected layers with a Dropout of 0.5 in between. As can be seen in Table 3, this architecture increased the accuracy by a large amount. Fig. 3 illustrates the overall architecture of the multi-modal model.

5 Conclusion

In this work, we developed a multi-modal model to automatically detect hoax content and rumor on twitter during public emergencies. Motivated by the need to incorporate multi-media information for false-information detection during natural disasters and public emergencies, we experiment with different feature extraction strategies for different modalities of data. We find that rumors and false information during such events can be detected more accurately when incorporating image data associated with a tweet. Our future work would expand the current study to incorporate other multimedia elements.

References

[1] Boididou, Christina, et al. "Verifying Multimedia Use at MediaEval 2015." *MediaEval*. 2015.

[2] Aiello LM, Petkos G, Mart ın CJ, Corney D, Papadopoulos S, Skraba R, Goker A, Kompatsiaris, I, Jaimes A (2013) Sensing trending topics in twitter. IEEE Trans Multimedia 15(6):1268–1282.

[3] SMWG Countering False Info Social Media Disasters Emergencies *Social Media Working Group for Emergency Services and Disaster Management* 2018

[4] When disaster strikes, count on tweets and data. https://www.aljazeera.com/indepth/opinion/2014/11/how-tweets-algorithms-can-sav-20141130142519956906.html

[5] Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In Proceedings of the 20th International Conference on World Wide Web, 675–684. ACM.

[6] Gupta, A.; Kumaraguru, P.; Castillo, C.; and Meier, P. 2014. Tweetcred: Real-time credibility assessment of content on twitter. *In Proceedings of the International Conference on Social Informatics*, 228–243. Springer

[7] Takahashi, T., and Igata, N. 2012. Rumor detection on twitter. *In Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems, and the 13th International Symposium on Advanced Intelligence Systems,* 452–457. IEEE

[8] Yang, F.; Liu, Y.; Yu, X.; and Yang, M. 2012. Automatic detection of rumor on sina weibo. *In Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 13:1– 13:7. ACM

[9] Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, 708– 717

[10] Ma, J.; Gao, W.; Wei, Z.; Lu, Y.; and Wong, K.-F. 2015. Detect rumors using time series of social context information on microblogging websites. *In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1751–1754. ACM.

[11] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. In Proceedings of the 20th international conference on World wide web. ACM, 665–674.

[12] David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. Communication theory 6, 3 (1996), 203–242.

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

[14] Stringhini G, Kruegel C, Vigna G (2010) Detecting spammers on social networks. In: Proceedings of the 26th annual computer security applications conference. ACM, pp 1–9

[15]Starbird K, Muzny G, Palen L (2012) Learning from the crowd: collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions. *In: Proceedings of the 9th international conference on information systems crisis response management Iscram*

[16] Canini KR, Suh B, Pirolli PL (2011) Finding credible information sources in social networks based on content and social structure. In: Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom). IEEE, pp 1–8

[17] Wu, K.; Yang, S.; and Zhu, K. Q. 2015. False rumors detection on sina weibo by propagation structures.*In Proceedings of the 31st IEEE International Conference on Data Engineering.*

[18] Kwon, S.; Cha, M.; and Jung, K. 2017. Rumor detection over varying time windows.

[19] https://www.theverge.com/2017/8/29/16221600/hurricane-harvey-hoaxes-viral-social-media-sharks-delta-planes