Fact-Checking Platform and Instant Local Call Alert System for Saving most Vulnerable, less Protected, and most affected Segment during Disaster using Crowdsourcing-powered Machine Learning application

Anonymous Author(s)
Affiliation
Address
email

Abstract

In order to prepare for or respond to a humanitarian disaster, crisis, and emergency, public officials and media organizations often must disseminate large amounts of technical information in a short amount of time. As the result, misinformation can circulate within or outside the affected community, and such misinformation can be particularly deadly during disaster scenarios. Therefore, it becomes a challenge for public safety agencies and organizations to reduce or eliminate the spread of misinformation on social media. The modern era of Artificial Intelligence (AI) based fact-checking models relies on machine learning (ML) models to detect misinformation using sophisticated algorithms. However, most of these ML approaches are limited by the data used to train them and they are over-dependent on being accessed via smartphone app interface which may be insufficient in low-income countries, where more than eighty percent of the population do not have smartphones. Hence in this paper, we propose an integrated fact-checking system that relies on a large network of independent and crowdsourced volunteer "checkers" who collect, verify and upload any fake messages into an app, which also has functionalities to offer anyone the ability to verify any message they have received. The app can receive messages for verification in form of short message system (SMS), app, and chatbot. In order to address the challenge of high illiteracy level in low-income countries, this platform is also able to support basic featurephones via an SMS-based interface where "verifiers" can send any suspicious message to a dedicated phone number as SMS and receive instant call alert which confirms the veracity or otherwise of the message. The app can also read messages of featurephone users based on predefined levels of permission-based access, especially for those who cannot read. The platform relies on the shared intelligence of the "crowd" to train the machine learning models for higher degree of accuracy, while also offering an instant automated call service, which is available as pre-recorded messages in twenty local languages.

1 Introduction

Natural or technological threats linked to resource loss, environmental degradation, financial damage, health effects, and societal disruptions can result in disaster or undermine societal functionality [4, 11, 3]. Disasters have been occurring more frequently over the past few decades as a result of

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

recent changes in global land use, population growth, and climate change [13]. According to the most recent IPCC study, severe events like drought, floods, forest fires, tsunamis, and tornadoes will become more intense globally over the next decades. According to new research, climate change has also increased the subsequent growth of viral, bacterial, and protozoan epidemics in various parts of the world as well as pandemics like COVID-19 [25, 16], heightening the likelihood of disease in the wake of natural disasters.

The four key phases of a catastrophe are prevention, preparation, reaction, and recovery, and effective communication during a crisis can help to lessen the effects of a disaster [3]. During and after disasters, risk communication is vital. This used to take the form of one-way communication to the people from the government [19]. However, it has frequently been observed that government risk communication to the public is insufficient because "People might panic," "People do not need to know," and "Speculation might worsen the disturbance". Experts claim that accurate dissemination of information, including speculative risk and worst-case scenarios, is essential for preventing misunderstandings in the public. Experts claim that accurate dissemination of information, including speculative risk, is essential for preventing misunderstandings in the public and reducing harm in the context of natural disasters [7]. The impact of misinformation disseminated through duplicate news for the purpose of obtaining multiple benefits has dramatically expanded in the age of social media and the internet, throwing noise into this essential channel of communication.

Given the increasing use of social media to disseminate and consume news, online social engagement has become one of the major vehicles for circulating such misinformation. Without a doubt, the harm caused by fake news is growing, and it is having a negative impact on social cohesion. For the security and sustainability of society in a contemporary and open internet environment, it is increasingly necessary to recognize fake news, grasp its characteristics, and learn how it spreads. The community-wide responsibility for the stability and sustainability of society in a contemporary and open internet environment rests with identifying fake news and understanding how it propagates. The detection of such fake news requires a comprehensive methodology that considers several predictive characteristics of fake news, including interactions with other users, traits of the disseminator's user profile, and the propagation of the false news event, in addition to the false news texts or content-based information [20].

Recently, there has been a wealth of research attempts to model fake news detection using the advancement of ML. One of the most simple and popular methods to detect fake news is to extract linguistic features using n-grams from text and then train multiple predictive ML models in an ensembling methodology. Examples include Decision Trees (DT), K-nearest neighbors (KNN), Stochastic Gradient Descent (SGD), and Logistic Regression (LR) [2]. On the other hand, theory-driven methods have proven effective. Shu et al. [21] achieve better accuracy by incorporating textual features with auxiliary data, including user social engagements on social media. The authors also addressed how to identify fake information online using social and psychological theories. There are several data mining approaches for extracting predictive features. However, apart from traditional classifier models, one method deep learning-based: incorporating textual features and metadata for training deep neural network models (DNN) such as Convolutional Neural Network (CNN) [24]. Such models capture complex dependencies in the text which is then mapped using recurrent neural network text embeddings with softmax output activation to obtain the final prediction. Combining both linguistic features with recent DNN models has given the state-of-the-art performance in fake news detection.

However, while algorithmic methods can be modestly accurate, the gold standard is human annotation where it can be made available. Hence, in this work, we propose an integrated AI-assisted fact-checking system that relies on a large network of independent and crowdsourced volunteer "checkers", who regularly update all fake messages they receive in daily basis. This rich body of knowledge enriches the "training" capability of the machine learning platform to better help other users of the platform to validate and verify any fake message they receive with higher degree of accuracy. In addition, the platform is able to proactively act when it confirms misinformation via short message system (SMS) by initiating an incoming mobile phone call to the mobile phone number where it has detected the inaccurate message. The automated mobile phone call ls are pre-recorded in twenty local languages to serve the target populations in the low resource African country where it has been deployed.

2 Background: Misinformation

Misinformation is false information that, although it may appear to be true at first, can deceive and have negative repercussions on both the individual and the community [15]. According to research by Motta et al. [14], the public's perception of the COVID19 pandemic were influenced by misinformation provided by right-leaning media, which ultimately contributed to a climate of distrust in media. Additionally, they noted that "even seemingly innocent [misinformation] from trusted media sources may either give people a false sense of security or cause others to disregard official recommendations." In a sense, this directly or indirectly harms the particular society or community. How can misinformation cause harm in the context of humanitarian emergencies? Agrafiotis et al. [1] studied the harms stemming from misinformation within organizations, developing a useful taxonomy of potential harms including economic harms, reputational harms, physical or digital harms, psychological harms, and social harms that can extend to other domains.

Social media is a crucial component of crisis management. Authorities use it to report breaking news and headlines about developments that are happening in real-time. Public awareness of social media as a crisis communication tool has grown. Due to the pervasiveness of misleading information, social media's triumph has, however, been fleeting. The world's recent experience with misinformation during the COVID19 pandemic is illustrative of the problem. The public's confidence in vaccination has been harmed by anti-vaccination misinformation situations that focus on unproven risks and side effects or the immune system's inability to respond to viruses and bacteria. This has led to a decline in vaccination rates and allowed the community to become exposed to diseases like measles-mumps-rubella, hepatitis B, and H1N1 [17]. Additionally, during the 2016 Zika virus outbreak, myths about the virus' severity (Zika virus symptoms are similar to seasonal flu), cause, immunity, and prevention complicated to combat the serious infectious disease, putting people's health at risk [8].

Gupta et al. [10] looked into how misinformation-filled messages spread after natural disasters like hurricane Sandy. They concluded that the majority of these messages were shared messages and that there were very few original messages. Rajdev and Lee [18] looked at the activities of harmful users who posted disinformation, finding that tweets from malicious users received fewer likes than non-malicious accounts. Similar to this, false information about the 2006 Louisiana floods spread through Facebook messages and posts overwhelmed FEMA and the American Red Cross in March 2016. Recent studies of Twitter users have addressed specific kinds of false information and its negative effects, such as the use of household cleaners as COVID-19 viral treatments [5]. Such factors not only pollute the information media but also has a serious impact on the people and their surroundings.

The dangers of misinformation are further compounded in less developed countries, where official channels of communication and digital literacy are not as robust. Hence in this paper, we tackle the impact of misinformation in a community with a less protected but deeply affected segment during disaster and crisis in Nigeria.

3 Fact Checking System using Crowdsourcing and Instant Call Alert

Here, we explain the overall architecture of our model in detail along with a brief explanation of each component used in the model. The proposed framework not only predicts misinformation but also designs a full system from (1) data collection to (2) delivering the service and (3) user engagement and intervention, suitably applicable for crisis and disaster management. Our architecture composes three components: namely (1) Multi-modal data extraction, (2) ML modeling for fake news detection, and (3) an Engagement/Intervention of the model with the user as shown in Fig. 1.

3.1 Multi-modal data extraction

The first component of the proposed framework deals with the data collection associated with several modes such as text, audio-video, and images. Here we intend to collect data from all possible sources into standard plain text. The sources of these texts are taken to be from SMS, WhatsApp message, social media posts, comments, and hashtags. In addition to this, the text in the images is extracted using Optical Character Recognition (OCR) [23]. OCR is a technique that identifies printed or handwritten text characters inside digital images of real-world documents, including scanned paper documents. OCR's fundamental procedure entails reading text from a document and turning the

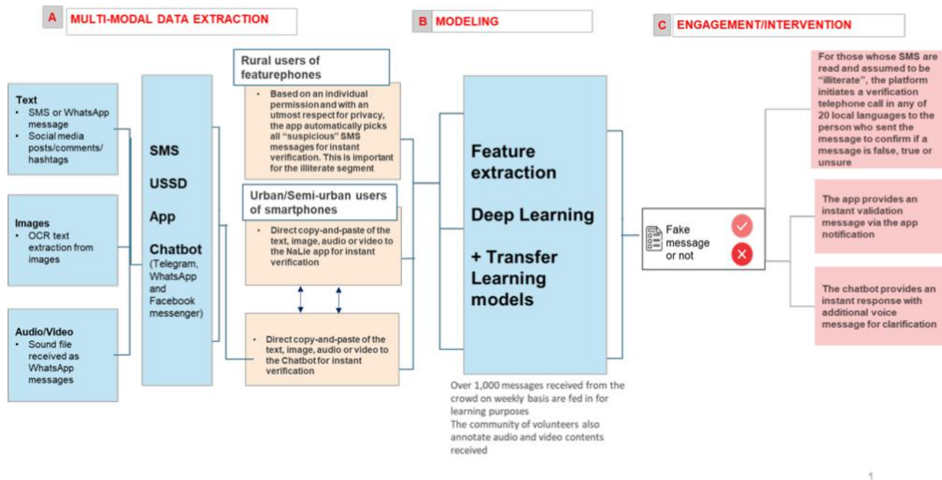

Figure 1: Framework of proposed fact-checking model with instant call alert.

characters into a form of code that may be utilized to process data. At last, we also process the message in audio and video using crowdsourcing where they capture the meaning and represent it in the text format, through app-user notation.

3.2 ML-based modeling

This subsection of the system consists of fact-checking and detection of misinformation. Various models have achieved state-of-the-art performance in fake news detection. One of the models described by Hochreiter and Schmidhuber [12] transforms the Twitter data into variable length representation using RNNs and LSTMs [9] with 1 layer of hidden units for rumor detection. These experiments show that RNN architectures have advantages over conventional ML models. Another variant of RNN apart from LSTM is GRU, which is simpler and more computationally efficient [6]. For the same Twitter data, it has been observed that the accuracy and F-measure performance of their GRU models was strong, though the precision and recall metrics for their classical support vector machine (SVM) models were marginally stronger, suggesting that much simpler models without deep learning will do in this task setting.

As of now there are several fake news detection model that uses sophisticated neural network models integrated with linguistic features extraction techniques. Beyond News Contents: The Role of Social Context for Fake News Detection [22] is a model based on the interaction of people that uses summary data such as the quantity of tweets they post. User credibility is determined by the size of their individual cluster, and the bigger the cluster, the less credible the user is. A hybrid model introduced in [20] uses increasingly sensitive data related to the user's profile. This effort makes it possible to determine a user's reputation, which is defined through the formulation of a score. It also offers the user's social participation as a substitute for obtaining this score. In this instance, a user's score is determined by their social network activities, such as the quantity of likes. A similar architecture, DistrustRank, uses the similarities between questionable websites and the presence of contentious issues to discredit websites that disseminate fake news. The level of the website's reputation is utilized to identify fake news, not the user, and the information used found from the url for the news link using PageRank [26]. However, these models are mostly neural network based architecture that are complex and black-box in nature (low interpretability). Since, our approach extends the traditional ML model using crowdsourcing work as transfer learning, the model has to be computationally efficient for quicker inference. Hence, we experimented the using various traditional models such as KNeighbors, LightGBM, XGBoost, and Random Forest. It has been observed that performance of these models are on par with bigger models with limited computational complexity.

In addition to this, we design proactive and predictive measure that focuses on the use of a volunteer crowd of professionals and non-professionals across the country who share as many as over 1,000 fake messages per day. They share the fake messages and validate the model for fact-checking which in turn improves accuracy over time by augmenting the training data available.

3.3 Engagement/intervention

After running the fact-checking ML models, we engage users into an intervention loop. The model here has two sub-parts: one part dealing with the most important aspect of the illiterate community known as rural smartphone users, and another for urban or semi-urban smartphone users. The system can receive content for verification purposes as an SMS, USSD, App, or chatbot on Telegram, WhatsApp, and Facebook Messenger. For urban and semi-urban users, the concept is pretty straightforward, where they can directly copy and paste the text, image, audio, or video to NaLie App for instant verification. Another approach is to copy and paste the text, image, audio or video to the Chatbot for instant verification. The main context lies with deployment in vulnerable communities, such as users from rural areas where illiteracy and lack of awareness are major challenges that the deployed app aims to address. People in underprivileged areas need constant assistance against such misinformation during crises and disasters. Hence, we tackled this situation differently. We used the individual's permission with the utmost respect for privacy so that the app automatically picks up all the suspicious SMS messages as rated by the predictive system for instant verification.

Unlike app-based verification, SMSs are read and the user assumed to be illiterate, and the platform initiates a verification telephone call in any of the 20 local languages to determine whether the person who sent the message is a valid or invalid source, or it the user is unsure. We have prerecorded messages for each respective case in 20 different local languages, using crowdsourcing.

4 Outcomes

We have implemented our proposed framework at the root level so that it reaches a particular community of underprivileged people during COVID-19, internal displacement, and flood crises. In particular, we have used it in Nigeria for providing potentially lifesaving interventions among above mentioned vulnerable communities The proposed platform has over 2,000 volunteers who submit an average of 1,000 false messages per week, most of which are similar and are related to the various misinformation issues in the country.

The solution and its unique approach, especially for the users of featurephones were recognized by the National Emergency Management Agency as one of the indigenous and homegrown solutions that can complement the government's effort during the humanitarian intervention. The non-profit that is providing this public service is currently engaging mobile telephone service providers to make the services free on their network as free call services (zero-rated calls) to render the service sustainable.

References

[1] Ioannis Agrafiotis, Jason R. C. Nurse, Michael Goldsmith, Sadie Creese, and David M. Upton. A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *J. Cybersecurity*, 2018.

[2] Hadeer Ahmed, Issa Traoré, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *ISDDC*, 2017.

[3] Declan T. Bradley, Marie McFarland, and Michael Clarke. The effectiveness of disaster risk communication: A systematic review of intervention studies. *PLoS Currents*, 6, 2014.

[4] GM Burnham. Chapter 1: Disaster definitions. In *The Johns Hopkins and Red Cross and Red Crescent Public Health Guide in Emergencies*. International Federation of Red Cross and Red Crescent Societies, 2008, 2 edition.

[5] Michael A. Chary, Daniel L Overbeek, Alexandria Papadimoulis, Adina Sheroff, and Michele Burns. Geospatial correlation between COVID-19 health misinformation and poisoning with household cleaners in the greater Boston area. *Clinical Toxicology*, 59:320–325, 2020.

[6] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *SSST@EMNLP*, 2014.

[7] Pablo M. Figueroa. Risk communication surrounding the Fukushima nuclear disaster: An anthropological approach. *Asia Europe Journal*, 11:53–64, 2013.

[8] Amira Ghenai and Yelena Mejova. Catching Zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on Twitter. In *IEEE International Conference on Healthcare Informatics (ICHI)*, page 518–518, 2017.

[9] Alex Graves. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850, 2013.

[10] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In *Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[11] George Haddow and Kim Haddow. *Disaster communications in a changing media world*. Butterworth-Heinemann, 2008.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

[13] Herbert E. Huppert and R. Stephen J. Sparks. Extreme natural hazards: population growth, globalization and environmental change. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 364:1875–1888, 2006.

[14] Matthew Motta, Dominik A. Stecuła, and Christina E. Farhart. How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the U.S. *Canadian Journal of Political Science/Revue Canadienne De Science Politique*, pages 1–8, 2020.

[15] Natalie Lee-San Pang and Joshua Ng. Misinformation in a riot: a two-step flow view. *Online Information Review*, 41:438–453, 2017.

[16] Sara H. Paull, Daniel E. Horton, Moetasim Ashfaq, Deeksha Rastogi, Laura D. Kramer, Noah S. Diffenbaugh, and Auston M Kilpatrick.

[17] Patrick Peretti-Watel, Jocelyn Raude, Luis Sagaon-Teyssier, Aymery Constant, Pierre Verger, and François Beck. Attitudes toward vaccination and the H1N1 vaccine: poor people's unfounded fears or legitimate concerns of the elite? *Social Science Medicine*, 109:10–18, 2014.

[18] Meet Rajdev and Kyumin Le. Fake and spam messages: Detecting misinformation during natural disasters on social media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 1, pages 17–20, 2015.

[19] Barbara J Reynolds. Crisis and emergency risk communication. *Applied Biosafety*, 10:47–56, 2005.

[20] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A hybrid deep model for fake news. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge*, 2017.

[21] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19:22–36, Sept. 2017.

[22] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019.

[23] Sargur N. Srihari, Ajay Shekhawat, and Stephen W. Lam. Optical character recognition (ocr). In *Encyclopedia of Computer Science*, page 1326–1333. John Wiley and Sons Ltd., 2003.

[24] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*, 2017.

[25] C Rogers Williams, Gina Mincham, Helen M. Faddy, Elvina Viennet, Scott A. Ritchie, and David Harley. Projections of increased and decreased dengue incidence under climate change. *Epidemiology and Infection*, 144:3091–3100, 2016.

[26] Vinicius Woloszyn and Wolfgang Nejdl. Distrustrank: Spotting false news domains. In *Proceedings of the 10th ACM Conference on Web Science*, 2018.