Machine Learning for Classification of COVID-19 Vaccine Misinformation on Twitter

Alana Foreman, Dr. Graciela Gonzalez-Hernandez

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA

Introduction and Background

Problem: COVID-19 Misinformation on Social Media

330 million Globally active Twitter users (Tankovska, 2021)

500 million Tweets shared per day (Statista, 2021)

Overabundance of COVID-19–related information on Twitter (“infodemic”)

Extensive spread of misinformation and conspiracy theories about virus and vaccine

Harmful to physical/mental health, increases stigmatization and vaccine hesitancy

Figure 1. Example of Misinformation (left) and Non-misinformation (right) Tweet.

Characteristics of Misinformation

Emotion
- Misinformation contains emotions of fear, surprise, and disgust (Vosoughi et al., 2018)
- Truthful posts contain emotions of anticipation, sadness, joy, trust (Vosoughi et al., 2018)
- Emotion makes information more viral (Vosoughi et al., 2018)

Popularity
- More followers/following/likes/retweets = more likely to be reshared (Avram et al., 2020)

Propagation
- True and false content differ in their patterns of propagation (Rosenfeld et al., 2020)
- False information is 70% more likely to be retweeted (Vosoughi et al., 2018)

The Role of Bots

- 9-15% of Twitter accounts are automated bots (48 million accounts) (Varol et al., 2017)
- 53 - 66% of accounts Tweeting about COVID-19 are bots (Himelein-Wachowiak et al., 2021)
- Mimic human interactions: following users, posting, liking, retweeting (Shao et al., 2018)
- Language choice/high activity may deceive users (Himelein-Wachowiak et al., 2021)

Topic Modeling

- **Before March 2020:** Origin of the virus, its sources, impact on people and the economy, preventing infection (Abd-Alrazaq et al., 2020; Xue et al., 2020)
- **Before May 2021:** Community spread, impact on work and family life, medical advice, politics (Jang et al., 2021; Song et al., 2021; Melton et al., 2021)
- **After May 2021:** ???

Guiding Research Questions

- Can we create a more efficient Twitter COVID-19 vaccine misinformation detection system?
- Does misinformation circulate differently between humans and bots?
- What topics are humans and bots tweeting about?

Objectives

1. Develop a Twitter data collection and preprocessing pipeline, and annotate/describe the data
2. Train six state-of-the-art machine learning models for classification of misinformation
3. Compare and analyze model efficiency and accuracy
4. Perform feature extraction to determine the most predictive features of misinformation
5. Identify hidden topics and patterns in the misinformation, bot, and human subsets
6. Develop prototype to scale findings

A novel, more effective, COVID-19 vaccine misinformation classification system

Data Collection and Preprocessing

Twitter Streaming API
- Tweet collection
- Applied relevant keywords: “COVID vax”, “corona vaccine”, “bill gates vaccine”, “microchip”

Demographer (Dredze et al., 2013)
- Gender
- Individual versus organization

NLTK Preprocessing (Bird, Klein, & Loper, 2009)
- Noise removal
- Tokenization
- Stop word removal
- Stemming/
- lemmatization
- Vectorization

Dataset Description and Annotation

- 1.5 million Tweets collected from October 2020-August 2021
- News Guard Covid Misinformation Tracking Center for ground truth/to deem a Tweet misinformation or not
- 1,000 annotated Tweets by myself and a peer to reduce bias
- Inter-annotator agreement: 0.73 (substantial agreement), Issues resolved

Sentiment Analysis, Emotion Analysis, Bot Detection

- **Text-blob** (Loria, 2018): Determine Tweet sentiment
- **Text2Emotion** (Diaz et al., 2018): Determine Tweet emotions (anger, fear, happiness, sad, surprise)
- **Botometer** (Yang et al., 2020): Determines if profile is bot or not; 56% of collected data

Six Trained Machine Learning Models

Content-based features
- Tweet emotion
- Tweet text (vectorized)
- Hashtags
- Subjectivity
- Tweet sentiment
- URL
- Favorite count
- Retweet count

User-based features
- Followers count
- Friends count
- Statuses count
- Verification status
- Total favorites count
- Gender
- Individual/Organization source

Models: Random Forest; Decision Tree; Naïve Bayes; Support Vector Machine; Logistic Regression; K-Nearest Neighbors

Best Performing Model: Random Forest

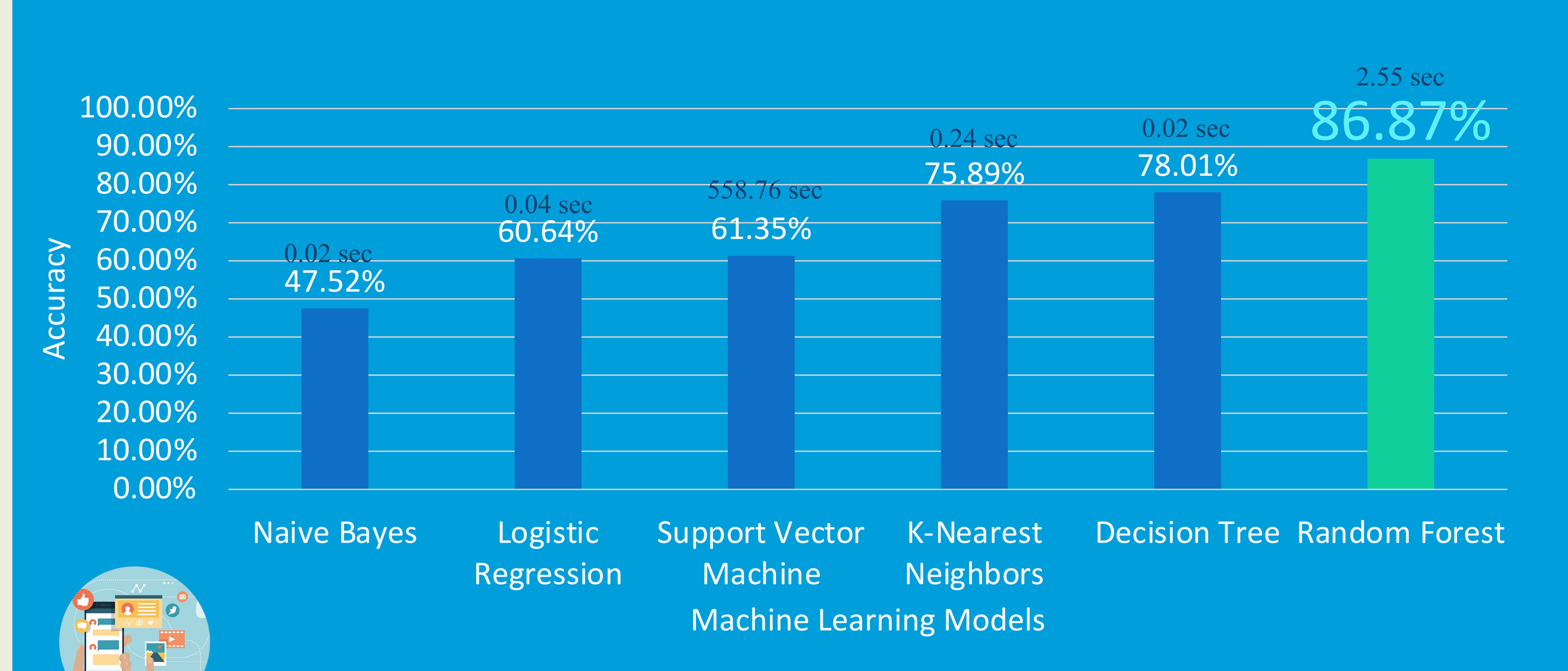

Figure 2. Summary of classifier performance in order of model accuracy.

Feature Selection

Recursive feature elimination to extract the most significant predictors of misinformation.

Emotion and propagation features indicative of veracity, in line with previous studies

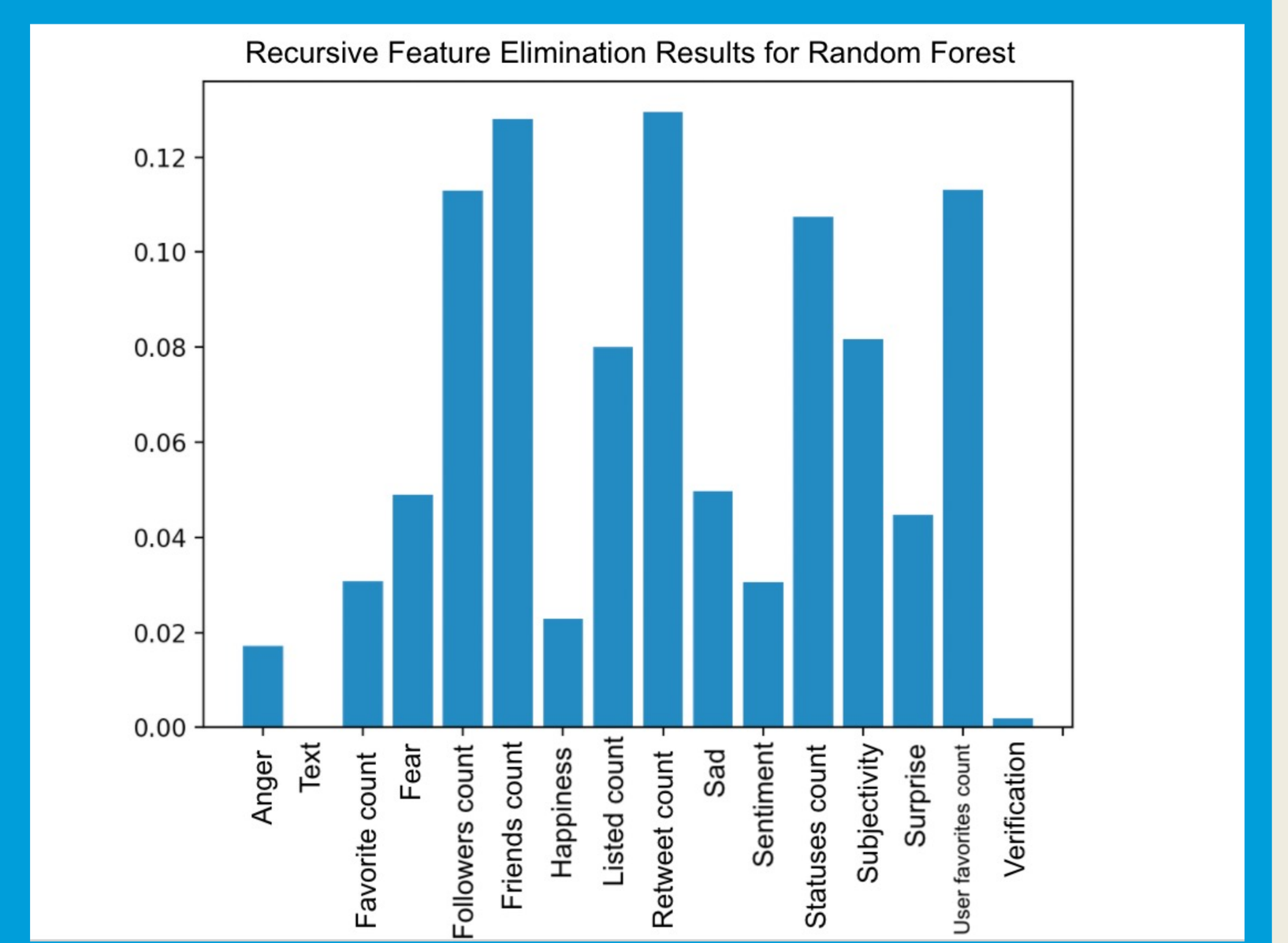

Figure 3. Recursive feature elimination results for random forest.

Bots Amplify the Spread of COVID-19 Vaccine Misinformation on Twitter

41% of bot-authored Tweets (56% of total) were misinformation

32% human-authored Tweets were misinformation

Popular discussion topics within COVID-19 Vaccine-related Tweets

Latent Dirichlet Allocation
1. Extract obscure topics in Tweets
2. Probabilities for words that frequently occur together
3. Word groupings manually assigned to word/phrase

Misinformation
- Contents of the vaccine
- Public policy
- Origin of the vaccine
- Effects of the vaccines
- Alternatives to the vaccine

People
- Return to pre-pandemic life
- General news
- Politics and Media
- Pro-vaccine opinions
- Vaccine-hesitant opinions

Bots
- Misinformation
- News
- Anti-vaccination opinions
- Advocation for vaccination

Future Research and Limitations

- Sarcasm relieved with additional amounts of data
- Topic analysis by age group
- Accuracy increased with more, annotated samples
- Research further into the impact that celebrities have on the spread of misinformation
- Increase prediction power with an ensemble of classifiers/deep learning

References

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. Science 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science. aap9559 arXiv:http://science.sciencemag.org/content/359/6380/1146.full.pdf
Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384.
Larson, H. J. (2018). The biggest pandemic risk? Viral misinformation. *Nature*, *562*(309). https://doi.org/10.1038/d41586-018-07034-4
Prooijen J-W v, Douglas KM. Conspiracy theories as part of history: the role of societal crisis situations. Mem Stud. 2017;10:323–33.
Tankovska, H. “Twitter MAU in the United States 2019.” *Statista*, 27 Jan. 2021, www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/.
Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. (2020, September 23). WHO. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation
Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a Tool for Health Research: A Systematic Review. *American journal of public health*, *107*(1), e1–e8. https://doi.org/10.2105/AJPH.2016.303512
Miller JM. Do COVID-19 conspiracy theory beliefs form a monological belief system? Can J Polit Sci. 2020. https://doi.org/10.1017/S0008423920000517.
Lynas, M. COVID: top 10 current conspiracy theories, https://allianceforscience.cornell.edu/blog/2020/04/covid-top-10-current-conspiracy-theories/ (2020).
Jamison, M. A., Broniatowski, D. A., Dredze, M., Sangraula, A., Smith, M. C., & Quinn, S. C. (2020). Not just conspiracy theories: Vaccine opponents and proponents add to the COVID-19 ‘infodemic’ on Twitter. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-38
Walker, M., & Matsa, K. E. (2021, September 20). *News consumption across social media in 2021*. Pew Research Center's Journalism Project. Retrieved November 1, 2021, from https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/.
Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web. ACM, 675–684. http://dl.acm.org/citation.cfm?id=1963500
Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. AAAI Press, 3818–3824.
Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. In Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI Press, 3901–3907
Rubin, V. (2017). Deception detection and rumor debunking for social media. Handbook of Social Media Research Methods
Yang Liu and Yi-Fang Brook Wu. 2018. 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.
Nan X, Xie B, Madden K. Acceptability of the H1N1 vaccine among older adults: The interplay of message framing and perceived vaccine safety and efficacy. *Health Communication.* 2012;27:559–568. doi: 10.1080/10410236.2011.617243.
Zarocostas, J., 2020. World Report How to fight an infodemic. The Lancet 395, 676. doi:10.1016/S0140-6736(20/30461-X
Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1
Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112. https://doi.org/10.1016/s0022-5371(77)80012-1

Prototype for Deployment

1. Keywords searched
2. Run random forest model on selected features of Tweet — reduces runtime
3. Flag misinformation as banner
4. Keywords of text used to tailor factual information
5. Decrease user engagement and alter attitude toward the Tweet — Prevents the spread of the information (Lanuis, Weber, & MacKenzie, 2021)

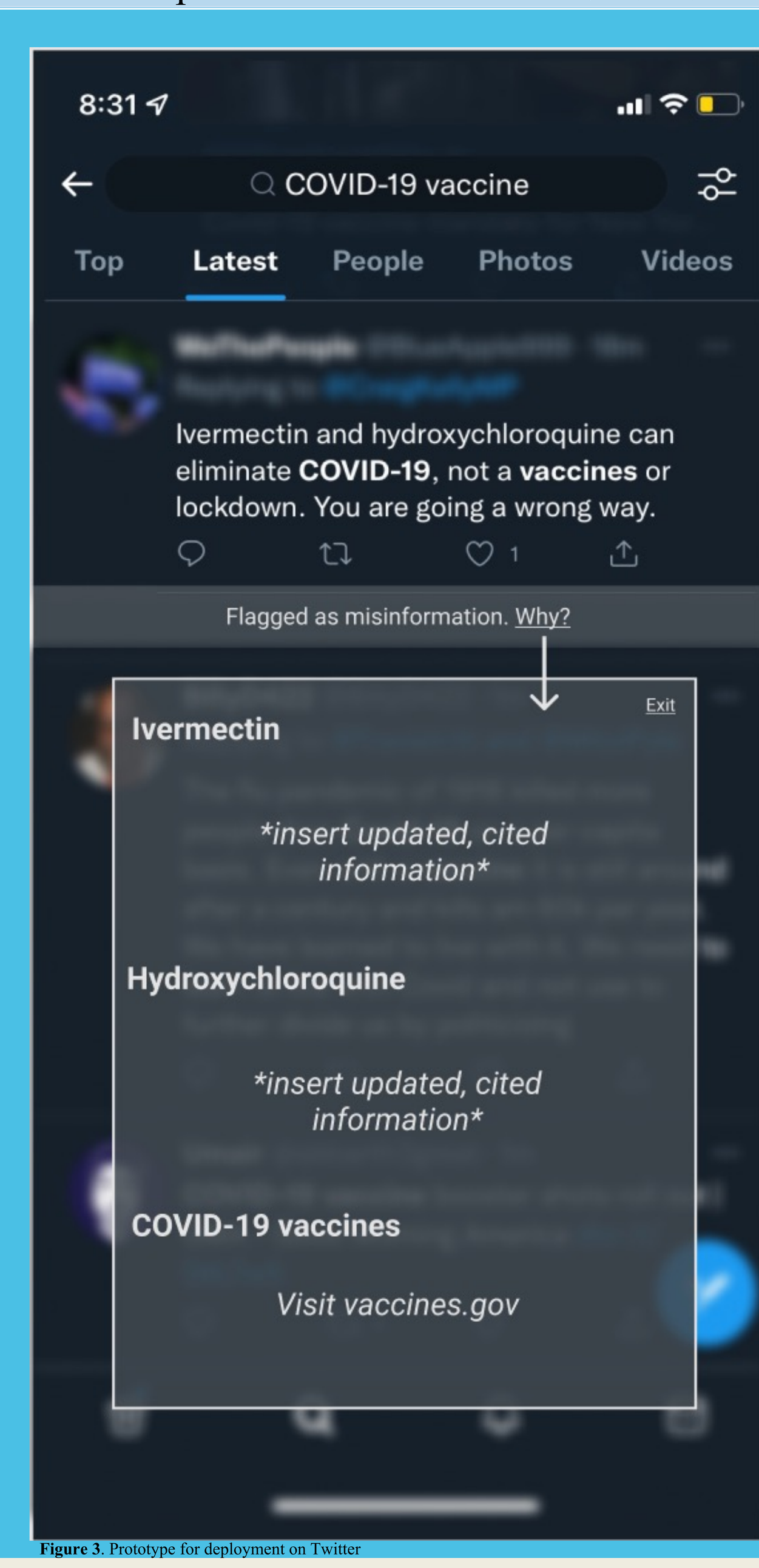

Figure 3. Prototype for deployment on Twitter

Conclusions

- Novel, efficient, scalable misinformation classification system
  Random Forest model (87% accuracy, 2.55 s)
- Propagation, emotion features indicative of veracity
- Bots amplify the spread of COVID-19 vaccine misinformation
- Developed prototype for deployment on Twitter

Call for more open-source social media APIs

Need for increased media literacy in schools and workplaces

Combating misinformation is a crucial component in building a safer Web

Limiting the spread of misinformation will improve the effectiveness of public health measures