# COVID-19 and Haters — A User Model Perspective

Soumitra Mehotra, Edoardo Serra, Anna Squicciarini, Younes Karimi

The 9th IEEE International Conference on Data Science and Advanced Analytics

October 13–16, 2022

# Introduction

- Online harassment, cyberbullying and trolling have become increasingly common in recent years, in light of the growing adoption of social media
- The recent COVID-19 health crisis has fueled outbursts of strong negative emotions and intensified such cyber aggression.

# Motivation

- Researchers from various disciplines have addressed online hate primarily via detection and related warning mechanism by analysis of text, focusing on detection of tone, vulgar content, and offensive words
- Previous works focus on hateful or abusive messages rather than offering a user-centered perspective to the problem of online hate
- Studies that focus on aggressors are often centered around selected forms of hate from specific demographic groups (e.g., youth)
- User-focused studies informed by data at scale are lacking
- The effect of external factors in accentuating negative online behavior has not been investigated in depth

# Summary of Contributions

- Develop a user-centered hate forecasting model capable of performing short-term (weekly) and long-term (6 weeks) forecasting.
- Develop a large set of user and time-varying features for forecasting hate probability, and analyze how psychological traits impact the predictive abilities of users' hate propensity and perform hypothesis testing to prove the significance of personality traits
- Estimate users' likelihood to post hateful comments in a given week based on a combination of innovative features

# Dataset

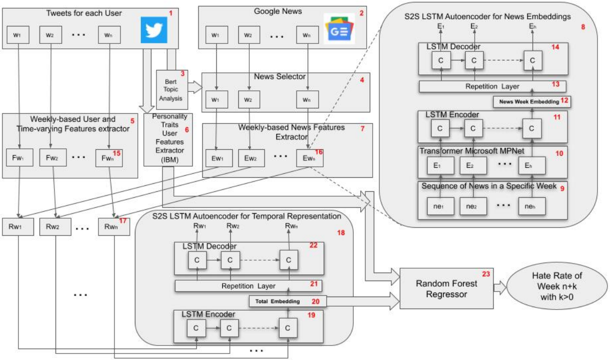
- A sample from a public COVID-specific Twitter repository [1]
- 508 million Tweet IDs collected with COVID-related keywords
- keywords : coronavirus, covid19, covid, corona, vaccine etc.
- Re-hydrated the tweets from the tweet IDs and only kept the original English content and filtered out all retweets and quotes
- The 2000 most active users in the pre-COVID phase (i.e., before March 11, 2020), and 2000 users in the post-COVID phase (i.e., after March 11, 2020)
- We identify 3,269 active users who had an average of 459 tweets, 75% of which tweeted in 15 out of the total of 21 weeks
- We labeled the 2,470,888 tweets in our dataset using a state-of-the-art pretrained model for hate detection
- The model uses a text-based classifier developed using a hand-labeled dataset of 2,400 tweets as the training set

1. E Chen, K Lerman, and E Ferrara. Covid-19 : The first public coronavirus twitter dataset. arxiv 2020

# Model Overview

Components :

- Sliding Windows and M$^2$DHF Training.
- Weekly-based User and Time-varying Features
- Personality Traits
- LSTM unsupervised Training




Multi-Modal Deep Hate Forecast (M2DHF)
Model Overview

# Sliding Windows and $M^2$DHF Training

- Model's training is done by using $m$ weeks.
- We create $win_1, \ldots, win_{m-n-h}$ windows. Each windows $win_i$ ($\forall i \in \{1, \ldots, m-n-h\}$) contains exactly $n$ weeks, i.e. $w_i, \ldots, w_{i+n-1}$ weeks for the model input and the week $w_{i+n-1+h}$ for computing the hate ratio used as the target of the regressor model.
- The hate confidence for each tweet in a week are obtained with a pretrained classifier for hate detection, and we average all the confidence hate ratios
- For each window and each user, we insert an instance in the training set of the random forest regressor.
- The two sequence-to-sequence autoencoders are trained in an unsupervised way by using the data in all the windows of the trainigset and excluding (for each window) the week $w_{i+n-1+h}$.

PennState BOISE STATE

# Weekly-based User and Time-varying Features

- **Vader Score** (Valence Aware Dictionary and Sentiment Reasoner) Vader is well-known a lexicon and rule-based sentiment analysis tool.
- **Readability Score** Among other textual features, we employ Flesch-Kincaid Reading Score to evaluate the reading ease of tweets.

$$206.835 - 1.015\left(\frac{\mathit{words}}{\mathit{sentences}}\right) - 84.6\left(\frac{\mathit{syllables}}{\mathit{words}}\right)$$

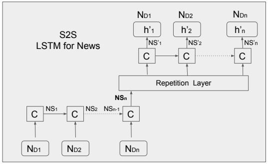
- **Social Network-based features** We use the number of followers and followings as features. In addition, we use the number of hashtags and retweets for a given tweet. The tweets are also encoded as sentence embeddings derived with a BERT-based Sentence transformer.

# Personality Traits

We use the big five personality traits described in the OCEAN :

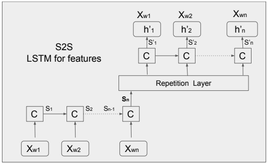
- *Agreeableness* is a person's tendency to be compassionate and cooperative toward others.
- *Conscientiousness* is a person's tendency to act in an organized or thoughtful way.
- *Extraversion*is a person's tendency to seek stimulation in the company of others.
- *Emotional* range, also referred to as Neuroticism or Natural reactions, is the extent to which a person's emotions are sensitive to the person's environment.
- *Openness* is the extent to which a person is open to experiencing different activities

# LSTM unsupervised Training for News




- Each News article in a specific week is treated as a paragraph and goes through a series of text pre-processing steps : removal of extra spaces, hashtags and mentions removal, URL removal, followed by lemmatization and tokenization.
- The output text is vectorized using a pretrained microsoft/mpnet-base sentence transformer model5 and results in 768-dimensional dense vector space for each news piece.
- The news vectors extracted from the transformer inside a week are ordered according to their creation date, padded for consistency, and fed to a Sequence to Sequence Autoencoder LSTM model.

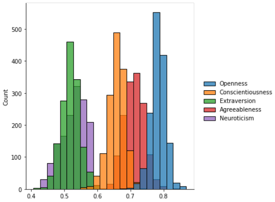
# LSTM unsupervised Training for Features



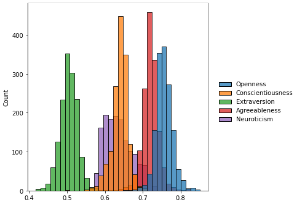

- The weekly based feature extracted from the tweets and the news week embeddings $Ew_1, \ldots, Ew_n$ (module 16) are concatenated in a unique representation $Rw_1, \ldots, Rw_n$ for each week.
- The sequence is aggregated in a unique representation through the use of a sequence-to-sequence LSTM autoencoder

# Personality Clustering

- Cluster 0 has a lower level of Neuroticism : [0.4-0.6]
- Cluster 1 has a higher level of Neuroticism : [0.6-0.7]



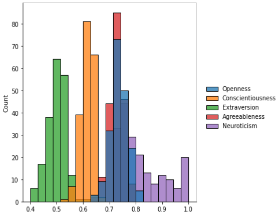

(a) Distribution of users in Cluster 0



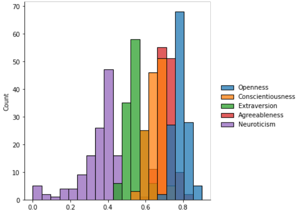

(b) Distribution of users in Cluster 1

# Personality Clustering

- Cluster 2 has the highest level of Neuroticism : [0.8-1.0]
- Cluster 3 has the lowest level of Neuroticism : [0.0-0.4]
- Hateful comments are expected to be coming from users with high Neuroticism, so the lowest proportion of Neuroticism makes it difficult for the model to forecast
- Users in cluster 3 exhibit lower predictive performance both in terms of R2 score and Pearson Correlation




(a) Distribution of users in Cluster 2




(b) Distribution of users in Cluster 3

# Performance

| | w_1 | w_2 | w_3 | w_4 | w_5 | w_6 |
|---|---|---|---|---|---|---|
| **$M^2$DHF with Encoder size=4 weeks** | | | | | | |
| R2 | 0.7066 | 0.6998 | 0.6704 | 0.6901 | **0.7516** | 0.7319 |
| R | 0.8420 | 0.8425 | 0.8244 | 0.8418 | **0.8790** | 0.8761 |
| **LSTM Baseline** | | | | | | |
| **LSTM Uni.** R2 | 0.5039 | 0.4386 | 0.4864 | 0.3543 | 0.3974 | 0.5207 |
| R | 0.7266 | 0.6828 | 0.7226 | 0.6236 | 0.6443 | 0.7548 |
| **LSTM Bi.** R2 | 0.4806 | 0.5183 | 0.5991 | 0.3840 | 0.4276 | 0.6328 |
| R | 0.7006 | 0.7216 | 0.7762 | 0.6209 | 0.6569 | 0.7977 |
| **Random Forest Baseline** | | | | | | |
| R2 | 0.5418 | 0.5656 | 0.6412 | 0.4967 | 0.5155 | 0.5599 |
| R | 0.7365 | 0.7527 | 0.8086 | 0.7051 | 0.7245 | 0.754 |
| **Sentence Transformed Tweets with Encoder size=4 weeks** | | | | | | |
| R2 | 0.673 | 0.533 | 0.521 | 0.584 | 0.6201 | 0.600 |
| R | 0.827 | 0.710 | 0.705 | 0.748 | 0.786 | 0.790 |

$M^2$DHF Short-Term Performance ($w_i$ denotes week $i$ performance, R is Pearson Coeff.)

$M^2$DHF outperforms all the baselines significantly, as we report that the average Pearson correlation and R2 score increase by 10-15% compared to the LSTM and RF baselines.

PennState BOISE STATE

# Performance

| | w_1 | w_2 | w_3 | w_4 | w_5 | w_6 |
|---|---|---|---|---|---|---|
| **$M^2$DHF with encoder size=10 weeks** | | | | | | |
| R2 | 0.6917 | 0.6495 | 0.5905 | 0.6111 | 0.6644 | **0.6493** |
| R | 0.8333 | 0.8101 | 0.7779 | 0.7992 | 0.8317 | **0.8325** |
| **LSTM Baseline** | | | | | | |
| R2 | 0.5039 | 0.3709 | 0.2839 | 0.3020 | 0.2584 | 0.1696 |
| R | 0.7266 | 0.6487 | 0.5967 | 0.6030 | 0.5736 | 0.4791 |
| **Random Forest Baseline** | | | | | | |
| R2 | 0.5418 | 0.3413 | 0.3563 | 0.3137 | 0.3556 | 0.3607 |
| R | 0.7365 | 0.596 | 0.5971 | 0.5612 | 0.5974 | 0.6028 |

$M^2$DHF long-term Prediction (R is Pearson correlation coefficient)

Given the historical data in input, $M^2$DHF predicts the next 6 weeks. We report results for our model along with the two baseline models (LSTM and RF).
We define an encoder of 10 weeks. This allows us to consider a significantly longer contextual memory and increase predictive performance both in terms of R2 score and Pearson Correlation.
Our proposed approach significantly outperforms the other baseline models in all weeks.

# Performance

| RF | w_1 | w_2 | w_3 | w_4 | w_5 | w_6 |
|---|---|---|---|---|---|---|
| R2 | **0.6836** | 0.6347 | 0.5498 | 0.5814 | 0.6281 | 0.6194 |
| R | **0.8291** | 0.7997 | 0.7554 | 0.7749 | 0.8051 | 0.8098 |

$M^2$DHF with encoder size=10 weeks and no news

| RF | w_1 | w_2 | w_3 | w_4 | w_5 | w_6 |
|---|---|---|---|---|---|---|
| R2 | 0.7001 | 0.6937 | 0.6712 | 0.6902 | 0.7499 | 0.7295 |
| R | 0.8397 | 0.8341 | 0.8156 | 0.8744 | 0.8844 | 0.8678 |

$M^2$DHF encoder size=4 weeks, with no news

| RF | w_1 | w_2 | w_3 | w_4 | w_5 | w_6 |
|---|---|---|---|---|---|---|
| R2 | 0.6963 | 0.6868 | 0.6503 | 0.6786 | 0.7406 | 0.7227 |
| R | 0.8353 | 0.8235 | 0.8049 | 0.8261 | 0.8695 | 0.8611 |

$M^2$DHF Performance without Personality : with encoder size=4 weeks + News Vector through S2S + RF (120 epochs)

While we still achieve 0.69 R for several weeks, the performance is consistently lower than our complete model, where we achieve an R2 of 0.7319 (R 0.8761) in w_6.

# Performance

Performance for each cluster group of Personality traits with Features passed through S2S of encoder size=4 weeks + News Vector through S2S + RF

| **Cluster** | w_1 | w_2 | w_3 | w_4 | w_5 | w_6 |
|---|---|---|---|---|---|---|
| 0 - R2 | 0.6901 | 0.7334 | 0.6966 | 0.7145 | 0.7833 | 0.7671 |
| 0 - R | **0.8311** | **0.8566** | **0.8347** | **0.8504** | **0.8915** | 0.8851 |
| 1 - R2 | 0.7191 | 0.6636 | 0.6633 | 0.6825 | 0.7510 | 0.7106 |
| 1 - R | 0.8497 | 0.8186 | 0.8180 | 0.8859 | 0.8804 | 0.8675 |
| 2 - R2 | 0.6966 | 0.6561 | 0.5615 | 0.5836 | 0.6074 | 0.6373 |
| 2 - R | 0.8420 | 0.8209 | 0.7569 | 0.7739 | 0.7834 | 0.8088 |
| 3 - R2 | 0.6224 | 0.6530 | 0.4741 | 0.5472 | 0.6040 | 0.6129 |
| 3 - R | 0.8011 | 0.8118 | 0.6990 | 0.7444 | 0.7812 | 0.7894 |

# Discussion and Limitations

1. We provided a user-centric perspective on the rise of hate trends on Twitter during the COVID-19 crisis.
2. We leveraged different sequential models to create advanced temporal space for short and long-term forecasting over different time intervals.
3. Our results show that our model provides an accurate prediction of hateful tweeters in the short and long term.
4. We observe that external events represented by the likes of news and headlines revolving around COVID-19 play a significant role in boosting up the forecasting performance for longer-term prediction
5. Future work includes further analyzing the role of personality traits and using Graph Neural Networks.

Thank you for your attention !