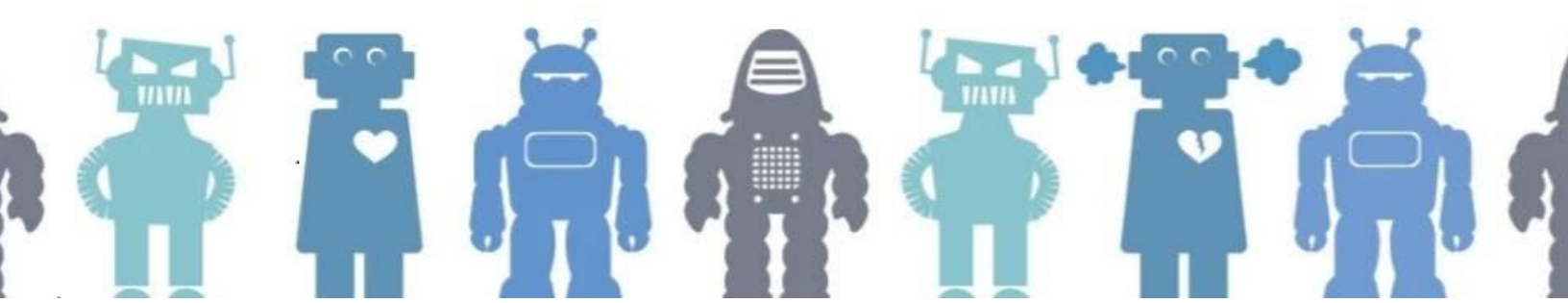


Twitter Bots & COVID-19

Final project CAPP 30254
Machine Learning for Public Policy (Spring 2020)
Link to Git repository

Rukhshan Arif Mian - rukhshan
Macarena Guzman - macaguzman
Yue Kuang - ykuang

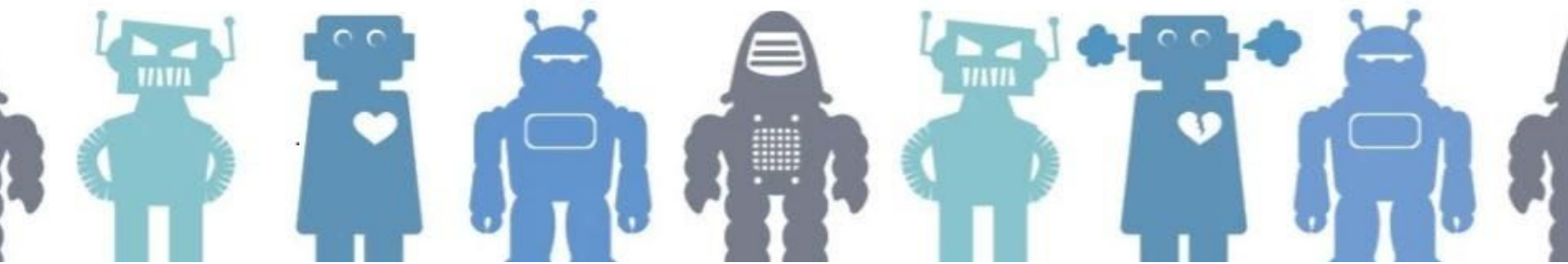


Table of contents

Executive Summary

Given the current pandemic of COVID-19 affecting the world is crucial for the people, to get accurate and timely information. Twitter, one of the most important social media platforms, is working assiduous spreading reliable and trustworthy facts related to the pandemic. However, as the pandemic is spreading, the misinformation is spreading as fast as the disease itself. Bots accounts on twitter are responsible for more than half of misinformation spreading.[1]This inaccurate information includes topics as "unproved cures", "conspiracy theories", "ease quarantine restrictions" and "reopening America". As bots are an important part combating misinformation, our project aims to identify bots accounts tweeting about COVID-19. This project relies on machine learning approaches, training Naives Bayes, Decision Tree and Random Forest models in order to label accounts as bots giving some common features.

Background and Overview:

Policy problems and importance:

Our project aims to address possible misinformation related to COVID-19 on Twitter. This is an important policy problem as many of the twitter audience use the platform to be informed and educated about global news and research. This is an important policy problem as a majority of the Twitter audience use the platform to be informed and educated about global news and research. Receiving misleading or incorrect information from Twitter can spread misinformation quickly. A study from the Massachusetts Institute of Technology in 2018 found that "false news spreads more rapidly on the social network Twitter than real news does"[2]. The reason, according to the study is that "untrue statements inspire strong feelings such as fear, disgust, and surprise"[3].

The spread of misinformation about COVID-19 can be potentially harmful. This could break institutional reputations, making the people less likely to follow scientifically informed government measures needed to reduce the spread of the virus. This is further discussed in the next sub-section.

Bots, which automatically create tweets without direct human oversight, are a source of spreading misinformation on twitter. According to Bot Sentinel, on March 26th bot accounts were responsible for 1,626 counts of #coronavirus, #COVID19 or #Coronavirus hashtags within 24 hours. Thus, having a tool that can predict whether or not an account is a bot becomes more important and crucial for policies that aim to educate people regarding a sensitive topic like COVID-19.

---

[1] https://www.cbsnews.com/news/bots-account-for-nearly-half-of-twitter-accounts-spreading-coronavirus-misinformation-researchers-say/

[2] https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

[3] ibid

Solution and use of Machine Learning*:*

The solution to this problem from our perspective is identifying reliable accounts tweeting through Twitter covid related information. Given that, the definition of bots is crucial: "A Twitter bot is a type of bot software that controls a Twitter account via the Twitter API. The bot software may autonomously perform actions such as tweeting, re-tweeting, liking, following, unfollowing, or direct messaging other accounts.[4]". According to Carnegie Mellon University, "Nearly half the "people" talking about the coronavirus pandemic on Twitter are not actually people, but bots" and under that bots they "identified more than 100 false narratives relating to coronavirus worldwide.[5]"

Machine learning techniques, as classification models, will be used on this project to predict if a twitter user can be classified as bot or human.

Audience and actions that could be taken based on the results:

The audience of our report will be: Local authorities, to create policies related with misinformation on social media (Especially on twitter); Journalist or news agencies, to inform the reliable sources of information; and, all active users of social media who want to identify if the information that they are reading about COVID-19 on twitter is created by a categorized bot fake COVID machine source.

The actions that could be taken based on the implementation of our project are:

- Create a policy that bans and punishes economically all the accounts classified by a bot that creates fake COVID-19 information.
- Identify patterns, behaviors, and trends related to tweeter's COVID-19 misinformation.

Data

Identifying Data Sources:

Our main goal is to use a previously classified dataset of bots to apply our Random Forest, Decision Trees and Naive Bayes models. Additionally, we aim to test our models to identify bots that tweeted about COVID-19 on 18th March 2020. We utilize data from this date as we assume that bot activity does not vary across days. That is, bots tweeting on Monday will have similar activity on Wednesday[6].

---

[4] https://en.wikipedia.org/wiki/Twitter_bot

[5] https://www.vice.com/en_us/article/dygnwz/if-youre-talking-about-coronavirus-on-twitter-youre-probably-a-bot

[6] The US was also implementing Stay-At-Home orders during this time period as well. Additionally, this was consulted with Professor Feamster. He mentioned that it is completely fine if we go ahead with data from one day as bot activity is very unlikely to vary.

For this purpose, we considered 2 types of raw datasets:

1. Classified dataset of bots (Training and Testing)
2. COVID-19 related tweets from 18th March, 2020 (Only testing)

Classified dataset of bots:

This data was taken from an open source Bot Repository — a centralized platform to share annotated datasets of bots on Twitter. There are a total of 15 classified datasets available on the platform of which we utilized the 11 datasets mentioned in Appendix A1. Consolidating these datasets provides us with 63264 rows and 16 columns. Some of the columns we consider for our models include a bot identifier (equal to 1 if bot, 0 if human), number of followers/following (friends) and number of tweets. A complete list of columns and their descriptions is shown in Appendix A2. These columns (variables) allow us to broadly monitor account activity. Literature suggests that accounts with primarily fake followers may have a smaller number of tweets[7]. Additionally, the following-to-follower ratio is a factor that existing literature looks at[8].

Approximately 80% of the accounts in our consolidated dataset are pre-classified as bots whereas the remaining 20% are identified as humans. We realize that there is an imbalance in our dataset and this is addressed in the modelling section of this report. Additionally summary statistics for our consolidated data are shown in Appendix A3.

COVID-19 related tweets from 18th March, 2020:

The next dataset we consider is of COVID-19 related tweets. We use this dataset to extract user related information — that is, we intend on knowing more about the characteristics of **users who tweet about COVID**. We are able to extract usernames of 22,000 users who tweeted about COVID after which we utilize Python's *twint*[9] library to extract their characteristics. These are in-line with what we have in Appendix A2. A limitation to using the *twint* library was that we were unable to extract information on *default_profile* and *listed_count*. Summary statistics are mentioned in Appendix A4. To summarize, we apply, train and test our models on the pre-classified consolidated human/bots datasets and we only test our model on the users who tweeted about COVID-19. Thus, we have 1 training and 2 testing datasets.

[7] https://scholar.smu.edu/cgi/viewcontent.cgi?article=1019&context=datasciencereview
[8] ibid
[9] https://github.com/twintproject/twint

Machine Learning and Details of Solution

Machine Learning problem

Our machine learning problem can be categorized as a classification technique with a binary label, having that said our goal is to train a model that can identify whether a twitter account is a bot or human through user profile features. Previous research related to bot identification has implemented a wide range of machine learning techniques to differentiate bot's behavior from human.

In terms of features selection, researchers such as Kantepe and Ganiz incorporated various types of features including user features, tweet, features, and periodic features to identify bots from all perspectives (Ganiz, Kantepe, 2017). Researchers also focus on twitter accounts' dynamic behavior data or tweets' content data (Chavoshi et al., 2016; Chu et al., 2012). The data used for machine learning in bot detection are often collected through Twitter API or public sources (Alothali et al., 2018). Most of the researches in this field implemented multiple supervised classification methods. Tree-based method and Bayesian method tend to perform the best regarding the confusion matrix (Alothali et al. 2018). Besides supervised methods, unsupervised learning methods such as clustering and deep learning based on convolutional neural networks have also achieved precision scores higher than 88% in bot detection (Cai et al., 2017; Chavoshi et al., 2016).

Applied Models and Features Set:

Since the data we collected from Bot Repository is labelled and has features engineered, supervised classification methods including tree-based methods, logistic regression, Naïve Bayes, and SVM are best candidates for our research. Among these methods, we train our data with a tree-based model (decision tree and random forest) and a Bayesian model (Naïve Bayes). These two models are widely used in the literature of machine learning in bot detection (Alothali et al. 2018). Tree-based methods have high interpretability and performance accuracy. Bayesian models are easy to implement and have good performance despite collinearity among features. To alleviate the overfitting tendency of decision trees, we used random forest, a bagging method that generates multiple trees to predict and is thus less susceptible to variances.

We train two different sets of features using Decision Tree, Random Forest and Naïve Bayes classifiers with various hyperparameters:

Features(1)	Features(2)	Naïve Bayes (1)	Naïve Bayes(2)
tweets	tweets	cat_tweets	cat_tweets
followers	followers	cat_followers	cat_followers
likes	likes	cat_likes	cat_likes
geo_enabled	geo_enabled	geo_enabled	geo_enabled
default_profile	-	default_profile	-
listed_count	-	cat_listed_count	-
has_description	has_description	has_description	has_description
len_en	-	len_en	-
friends	friends	cat_friends	cat_friends
verified	verified	cat_verified	cat_verified

In the first set of features (features(1)), we included 10 features. These 10 features are all the features we have obtained from the bot repository's data. We believe that more features included allow us to differentiate the behavior of bots from that of humans.

We also need models for a second set of features (Features(2)) because the COVID-19 dataset does not share all the features that the bot repository dataset has. This feature discrepancy results from the limitation of the twitter scraping tool (Twint) we used to collect data for the COVID-19 dataset.

Notice that the case for Naïve Bayes is more complicated because sklearn's Naïve Bayes model only allows us to assume one type of feature distribution for all features. There are two types of data in our bot-repository data: binary variables and continuous variables. Since a binomial distribution is just a special case of multinomial distributions, we decided to transform continuous variables in Features(1) and Features(2) into categorical variables using the bins shown in Appendix B1. With a unified feature type, we assumed multinomial distribution for features in our naïve Bayes model.

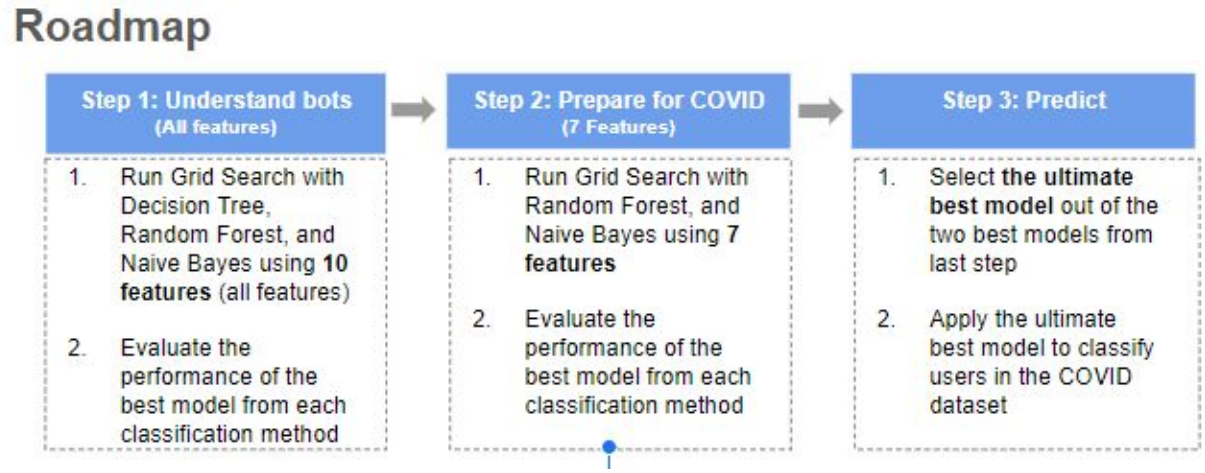
For each method, we also selected multiple parameters to tune the model based on different values of alpha, fit_prior, class_prior in the case of Naïve Bayes and criterion, max_depth and N_estimators in the case of Decision Trees and Random Forest. The exact parameters, values and descriptions for each model are depicted in Appendix B2.

In terms of evaluation metrics, we choose to use precision, recall, and the balance of the two – F1 scores of our model on the test data to determine a model's performance. The bot-repository dataset is an imbalanced dataset with approximately 80% bots and 20% human users. As a result, precision, recall and F1 score should provide us a more reliable measurement of the performance of the model than accuracy score. We use precision as our primary performance evaluation metric because we are more concerned with Type I error where we misidentify humans as bots.

We perform a 10-fold cross validation for both Random Forest and Naïve Bayes. This splits the training data into 10 subsets and then iterates 10 times. During each iteration, one of the 10 subsets will be held out as a validation set and we only use the remaining 9 subsets to train the model. We average the result 10 test performance score to generate the performance of the model.

Summarizing our Machine Learning process

In order to summarize our proposed machine learning process we constructed the following roadmap. Here we define 3 steps that will be developed in order to get the ultimate best model:



Results of implemented models:

Step 1: Understand bots through pre-classified bot data (Training + Testing):

After running the Naive Bayes model, Decision Tree and Random forest with 10 features, we got the results summarized on the following table:

Best Models, 10 features								
			Testing labeled dataset			Testing unlabeled dataset		
	Best Model	Best Feature	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Naives Bayes Appendix B3	Alpha: 0.1 class_prior: None fit_prior: False	cat_tweets ~20%	93.0%	94.0%	97.0%	93.2%	96.6%	94.7%
Decision Tree Appendix B4	max_deepth: 3 criterion: Gini	tweets ~ 82%	93.9%	97.7%	94.5%	94.1%	98.0%	94.5%
Random Forest Appendix B5	max_deepth: 5 criterion: Gini	tweets ~28%	95.6%	97.6%	96.8%	100%	100%	100%

Under step 1 we explore the features of the datasets and we find that under both models, the number of tweets posted by an account is the most important feature. In addition, we realize that we get higher scores for metrics of accuracy, precision and recall. Therefore, we can conclude that our dataset meets the conditions to continue searching for the best model in the next step.

Step 2: Prepare models for COVID-19 dataset:

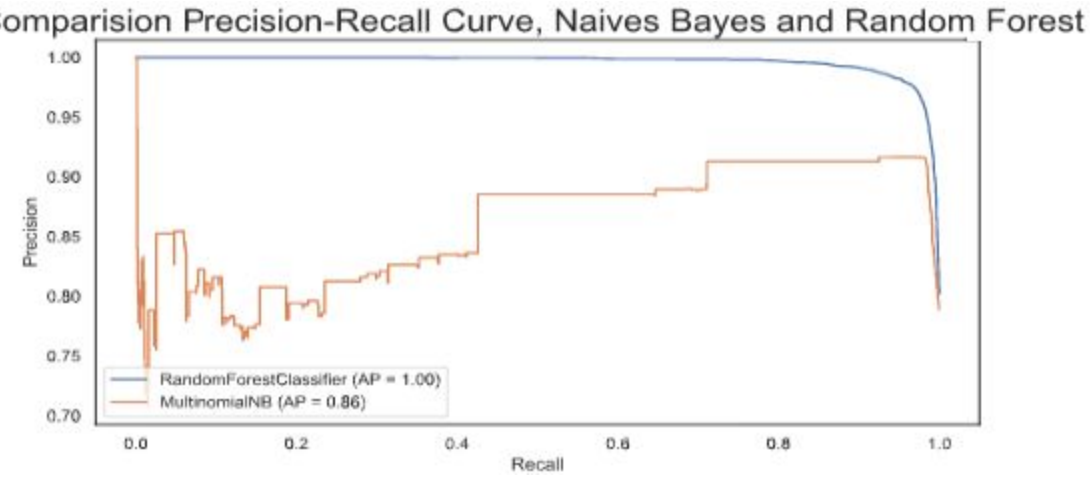
For step 2, we runned the Naive Bayes and Random Forest models with 7 features, we got the results summarized on the following table:

Best Models, 7 features								
			Testing labeled dataset			Testing unlabeled dataset		
	Best Model	Best Feature	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Naives Bayes Appendix B6	Alpha: 0.1 class_prior: None fit_prior: False	cat_tweets ~29%	91.2%	91.4%	97.9%	91.4%	91.5%	98.1%
Random Forest Appendix B7	max_deepth: 5 criterion: Gini	tweets ~37%	95.1%	97.2%	97.3%	99%	99%	99%

After the step 2, we realize that even with less features the models got great scores. In addition, the True values predicted by the models are high, which is one of the crucial characteristics among the models that we are looking for. The feature importance graphs and confusion matrices are shown in the Appendix (From Appendix B3 till B7).

Step 3: Predict:

In order to decide which models will predict better our labels, we plot a comparison of the precision recall curve for different thresholds and we got the following results:



It is important to analyze this curve in the case of our model because our training dataset is imbalanced. That is, we have more bot labels than human labels, so the precision recall curve will work better. It summarizes the trade-off between the true positive rate and the positive predictive value for this model using different probability thresholds. The ROC curve is preferred on balanced label datasets so it would not be useful for our analyses.

In addition, the precision recall curve calculates these metrics after excluding the true negative values, as we are interested in the correct prediction of the minority class human and less interested in predicting the bot label correctly.

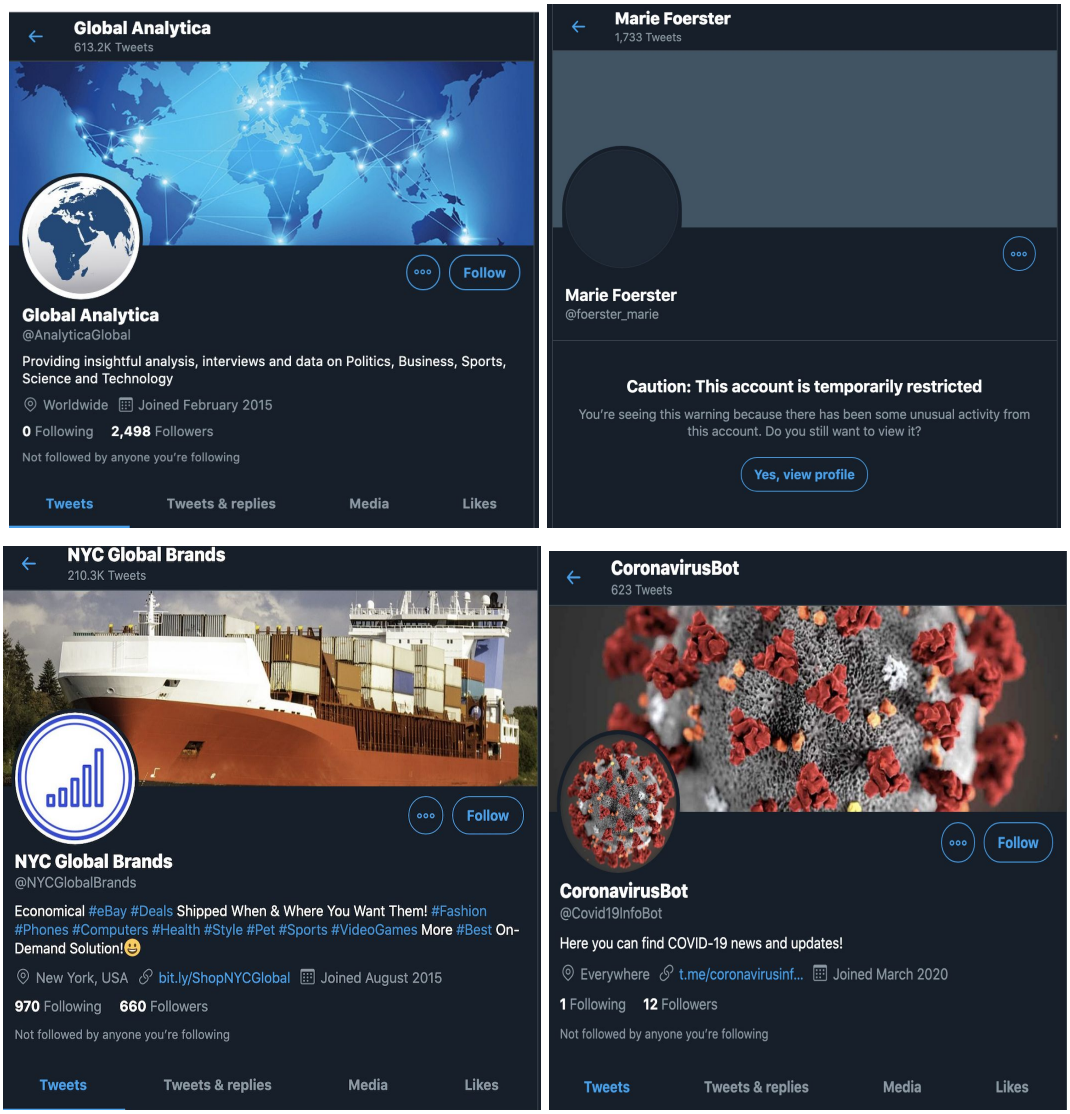
We find that the Random Forest model with 7 features serves as our best model with its most important feature being the number of tweets. We understand that it is likely that accounts that have most recently been made may be misidentified as bots (even though they are humans). To further study this, we take into accounts that are at most 3 months old[10]. For a small percentage of these, we find that our model fails to predict 0. We address this concern by creating a confusion matrix (Appendix B8) for accounts that are less than or equal to 90 days old. We find that for 38 of the total 7193 such accounts, our model identifies a human account as a bot. For 67 accounts, our model identifies bots as humans and for 7078 accounts are correctly identified as bots[11]. We can safely deduce that our model does an effective job at predicting accounts that are "younger"[12]. Thus, we move forward with our analysis using this model. This is now applied on the COVID-19 dataset.

COVID user prediction: Descriptive Statistics

We have now identified bots and humans in our COVID-19 dataset using the Random Forest model. Applying our training model to the testing model to identify bots tweeting about COVID, we find that approximately 11% of the accounts tweeting about COVID-19 are bots whereas the remaining 89% are humans (Appendix C1).

To move forward, we look at descriptive statistics and visualizations of the identified bot accounts. Appendix C2 shows the bot accounts that had high frequencies of tweets. We show profiles of some of these accounts to (qualitatively) validate our model:

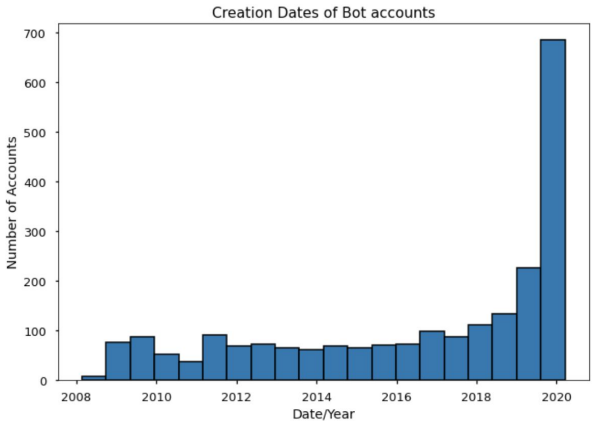
[10] Account Age = Date Scraped - Date Created. In this case, Account Age <= 90 days old.
[11] This is a higher number because our dataset is imbalanced
[12] Less than or equal to 90 days old



We manually observe bot-like characteristics in the profiles of the accounts labelled as bots (through our model). One of these has been flagged by Twitter for unusual activity and for the others, we observe a large number of tweets in a short amount of time. Additionally, we observe that certain accounts self-identify themselves as bots. In Appendix C3, we show a word-cloud to broadly observe what such accounts (identified bots through our model) are tweeting about. For this purpose, we remove words such as *coronavirus, covid-19* etc.

The size of each word indicates its frequency or importance. It allows us to highlight the most tweeted about topics amongst bots. Examples of commonly occurring terms are *testing, quarantine, outbreak* and *social distancing*. Appendix C4 visualizes the frequency of hashtags being tweeted about (excluding naturally high frequency hashtags such as #coronavirus, #covid19 etc).

Lastly, we look at the dates when (identified) bot accounts were created. We show this in the following graph:



We find a higher number of bot accounts being created near the end of 2019 and in early 2020 (as COVID was on the rise). This provides us with useful information in understanding the role of bots on social media during a pandemic. We do find a peak near the start and during COVID. However, we understand that since the number of tweets is the most important feature in our model, it could be the case that human accounts with a lower number of tweets are being labelled as bots. This is further addressed in the Limitations section of this report.

Policy Recommendations

Bots are large contributors to Fake News on social media. Additionally, bots tend to impact and at times influence political discussion. This can be seen from the role of Russian bot accounts that tweeted about the United States elections in 2016. Misinformation in the time of a pandemic can potentially lead to harmful outcomes which is why our project aims to identify the first step in tackling such problems. Our models allow us to identify bot activity on Twitter and thus, further allow us to observe COVID-related activity. Additionally, this provides us with additional avenues of potential research. For example: through conducting rigorous text analysis, we can effectively analyze the content that bot accounts are tweeting about in more detail. An example of such analysis could be trying to identify the type of misinformation, if any, bots are spreading on Twitter. A newspaper article highlights the role of bots in tweeting about easing quarantine restrictions[13]. This is potentially harmful misinformation especially during a pandemic. Identifying bots using such methods could allow organizations such as Twitter (and similar organizations) to flag or remove bot accounts for human consumers. Twitter does

[13] https://www.businessinsider.com/trolls-bots-flooding-social-media-with-anti-quarantine-disinformation-2020-4

occasionally remove bot accounts however an additional step in this could be flagging information being tweeted about bots (on all topics and not just COVID-19 related topics). This can be extended to other social media platforms as well and can thus, limit and reduce the spread of misinformation.

Ethics

Our project idea could lead to ethical concerns when we use personal Twitter accounts to train our models. Additionally, if we mistakenly feed our model with bad accounts (incorrect data that is not captured), or if the pre-classified Bot data contains erroneous information, the model can mistakenly categorize accounts as bot when they are not. Additionally, human error with the internal programming of the aforementioned models could prove to be harmful to the general public. We could end up doing exactly what we are trying to avoid: misinforming the audience.

Limitations, Caveats, Suggestions for Future Work

It is important to note that not all bots tweets false information. For example, some political bot accounts tweet official information about guidance or actions that will be taken on a state or region. A limitation with respect to our models relates to the lack of use of content-related features. These include analyzing tweets made by bots, *types* of followers, types of retweets/likes that an account has made. In this report, we are looking at profile-related features (*number* of tweets, likes etc). Due to capacity constraints, we were unable to gather all tweets made by accounts and perform rigorous text analysis. This may be thought of as potential work on this topic. Including content-related analysis would allow us to make our model more robust as it adds an additional dimension to the type of analysis we are performing. Even though imbalance in the previously classified bots/human data is an additional limitation, we address it by considering precision-recall as our principal metric. However, having a balanced dataset of humans and bots could allow us to make our model more accurate. Suggestions for future work would include a classified dataset on bots that are tweeting about COVID-19. If available, this can be used as an additional training dataset for the purpose of our models. Lastly, it is likely that human accounts with a lower number of tweets are being labelled as bots when we use our model to identify bots in the COVID-19 data. We address this concern when predicting this model. Furthermore, we believe that having data available on classified bot accounts that are tweeting about COVID-19 will allow us to make our approach more robust. Given these limitations, we believe that our models effectively identify Twitter bots and thus allow us to perform broad analysis of what type of COVID-19 related information these bots are tweeting about.

Bibliography

2018 International Conference on Innovations in Information Technology (IIT), Innovations in Information Technology (IIT), 2018 International Conference On, IEEE, pp. 175–80. 2018.

C. Cai, L. Li, and D. Zengi, “Behavior enhanced deep bot detection in social media,” in Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on. IEEE, 2017, pp. 128-130.

E. Alothali, N. Zaki, E. Mohamed, H. Alashwal, “ Detecting Social Bots on Twitter: A literature Review”

M. Kantepe and M. C. Ganiz, “Preprocessing framework for twitter bot detection,” in Computer Science and Engineering (UBMK), 2017 International Conference on. IEEE, 2017, pp. 630-634.

N. Chavoshi, H. Hamooni, and A. Mueen, “Debot: Twitter bot detection via warped correlation.” in ICDM, 2016, pp. 817-822.

Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, “Detecting automation of twitter accounts: Are you a human, bot, or cyborg?” IEEE Transactions on Dependable and Secure Computing, vol. 9, no. 6, pp. 811-824, 2012.

Appendix

Appendix A1

Dataset	Description	Observations	Year
midterm_18	Manually labeled human and bot accounts from 2018 US midterm elections. Labels and processed user objects.	50538	2018
social_spambots2	Spammers of paid apps for mobile devices	3457	2014
gilani_2017	Manually annotated human and bot accounts. Labels and user objects	2503	2017
verified_2019	Verified human accounts. Labels and user objects.	1987	2019
traditional_spambots_4	Group of automated accounts spamming job offers	1128	2009
traditional_spambots_1	Training set of spammers	1000	2009
social_spambots1	Retweeters of an Italian political candidate	991	2012
cresci-rtbust-2019	Manually annotated bot and human accounts. Labels and user objects.	693	2019
social_spambots3	Spammers of products on sale at Amazon.com	464	2011
traditional_spambots_3	Automated accounts spamming job offers	403	2013
traditional_spambots_2	Spammers of scam URLs	100	2014

Appendix A2

Feature	Description
bot	Dummy variable = 1 if bot, 0 if human
len_en	Default language is English
verified	Dummy variable = 1 if verified, 0 if not
followers	Number of followers
geo_enabled	Dummy variable = 1 if location has been enabled for this account, else 0
friends	Number of following
defualt_profile	Dummy variable = 1 if twitter's default profile is used, else 0
listed_count	Count of lists an account is part of
has_description	Dummy variable = 1 if account has a biography
likes	Number of likes
tweets	Number of tweets

Appendix A3

Mean		
Feature	**Bot**	**Human**
Tweets	3843.20668	20338.7416
Followers	19041.9892	203560.847
Following	776.796692	3219.27459

Median:

Median		
Feature	**Bot**	**Human**
Followers	0	888
Following	1	614
Number of Tweets	22	7281
Likes	0	4482

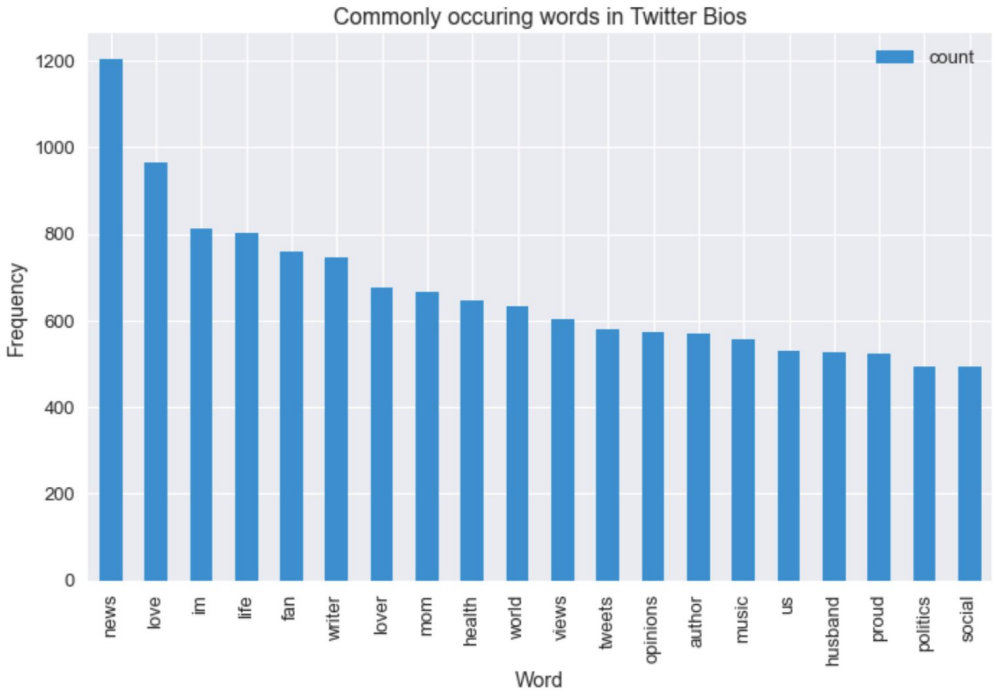
Appendix A4

Summary Statistics:

Mean	
Tweets	22683.68
Followers	21872.26
Friends/Following	1909

Median	
Tweets	5221
Followers	539
Friends/Following	603

We notice a difference between the averages of followers and following primarily because we have verified news accounts in our dataset. However, this is not a problem as we normalize these features before using this as a testing dataset. Next, we look at the keywords appearing in Twitter biographies for these users:



Appendix B1

bin	Value(categorical)
0 x 10	1
11 < x 100	2
100 < x 1000	3
1000 < x 10000	4
X > 10000	5

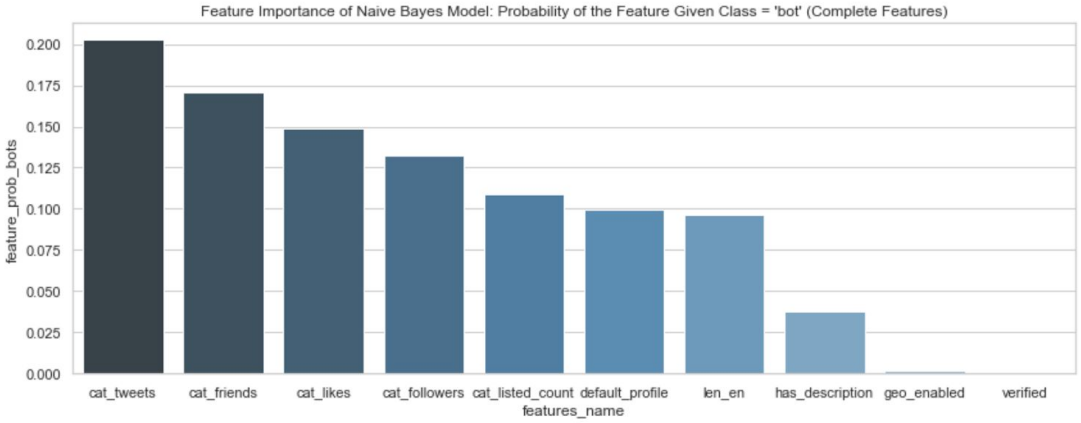
Appendix B2

Naïve Bayes (Multinomial): Parameters		
Parameter	**Values**	**Description**
Alpha	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0	Additive (Laplace/Lidstone) smoothing parameter
Fit_prior	True, False	Whether to learn class prior probabilities or not. If false, a uniform prior will be used.
Class_prior	None, [.1,.9],[.2, .8]	Prior probabilities of the classes. If specified the priors are not adjusted according to the data.

Random Forest & Decision Tree: Parameters		
Parameter	**Values**	**Description**
Criterion	'entropy', 'gini'	The function to measure the quality of a split.
Max_depth	1,3,5	The maximum depth of the tree.
N_estimators (random forest only)	100,1000,5000	The number of trees in the forest.

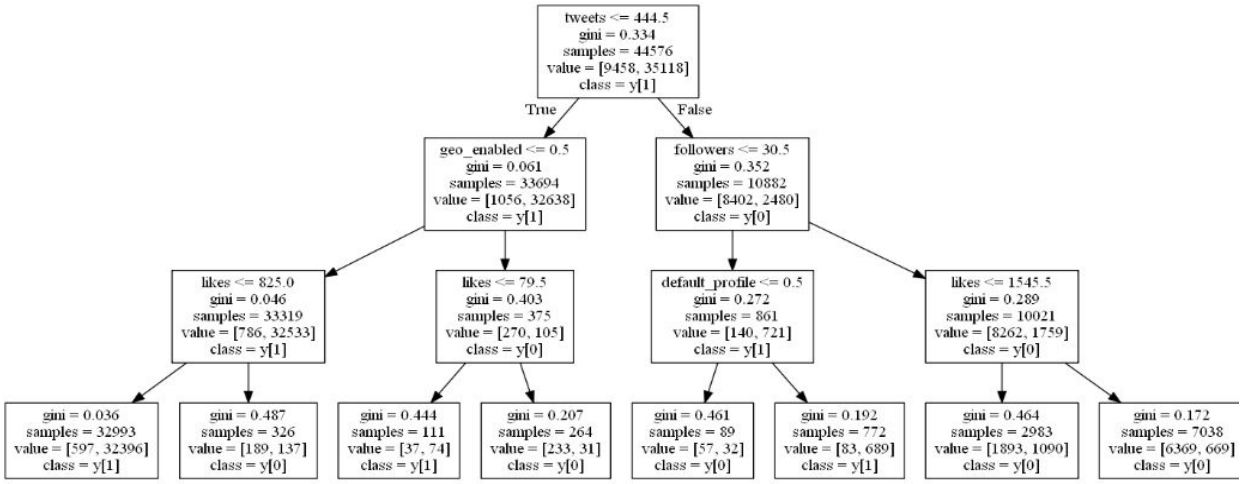
Appendix B3

The most important feature in this model is the number of tweets posted by an account.

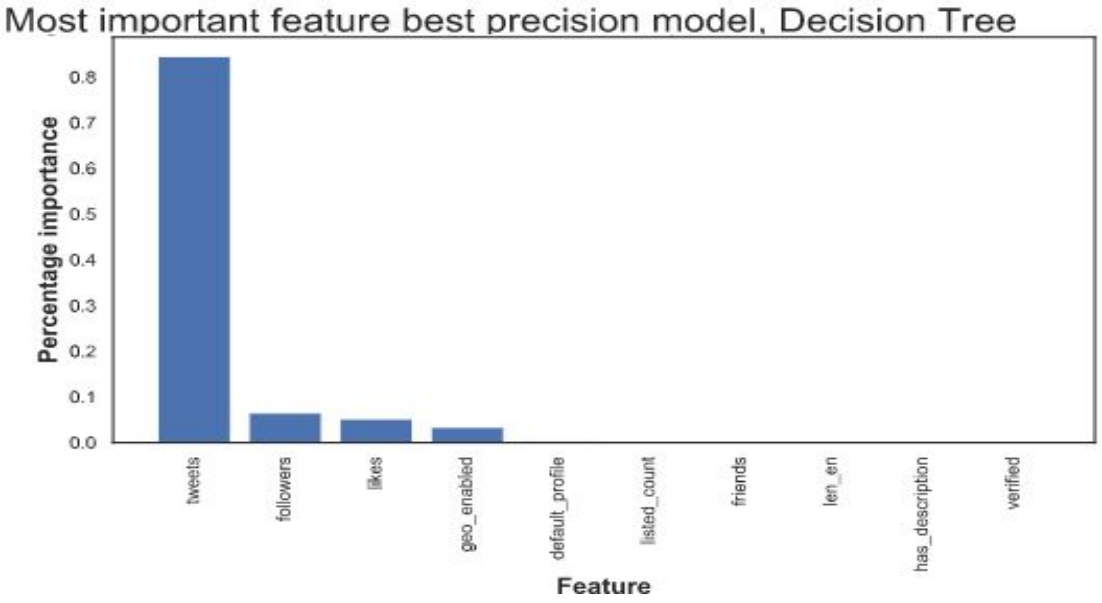


Appendix B4

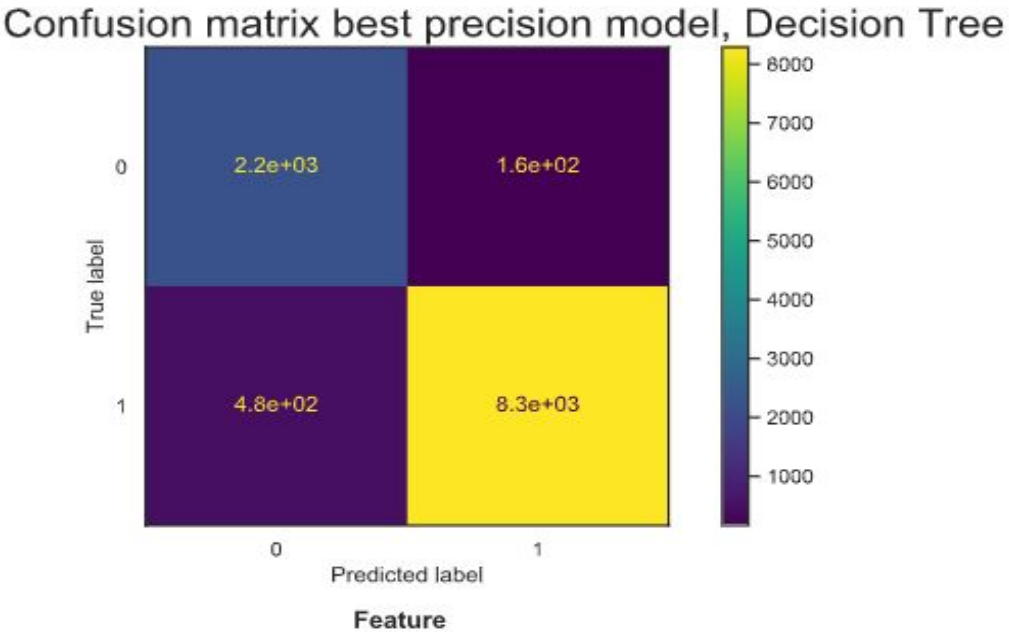
Best Model Decision Tree



The most important feature in this model is the number of tweets posted by an account.

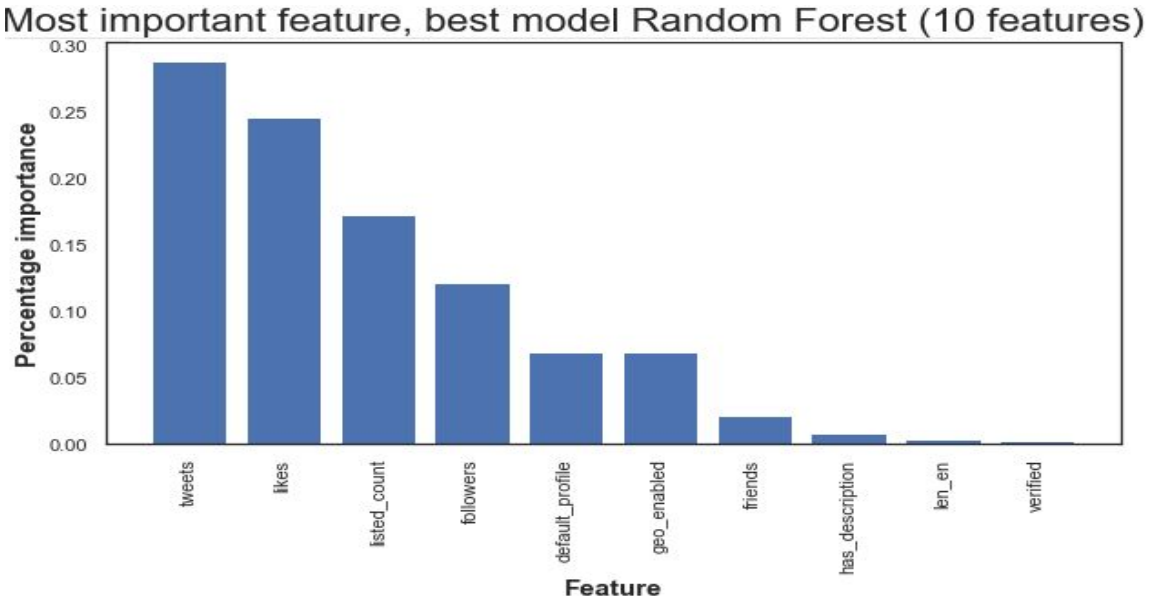


Under this model we got a total of 10,497 true positive and true negatives values, and a total of 648 false positive and false negatives. The following confusion matrix visualizes this values:

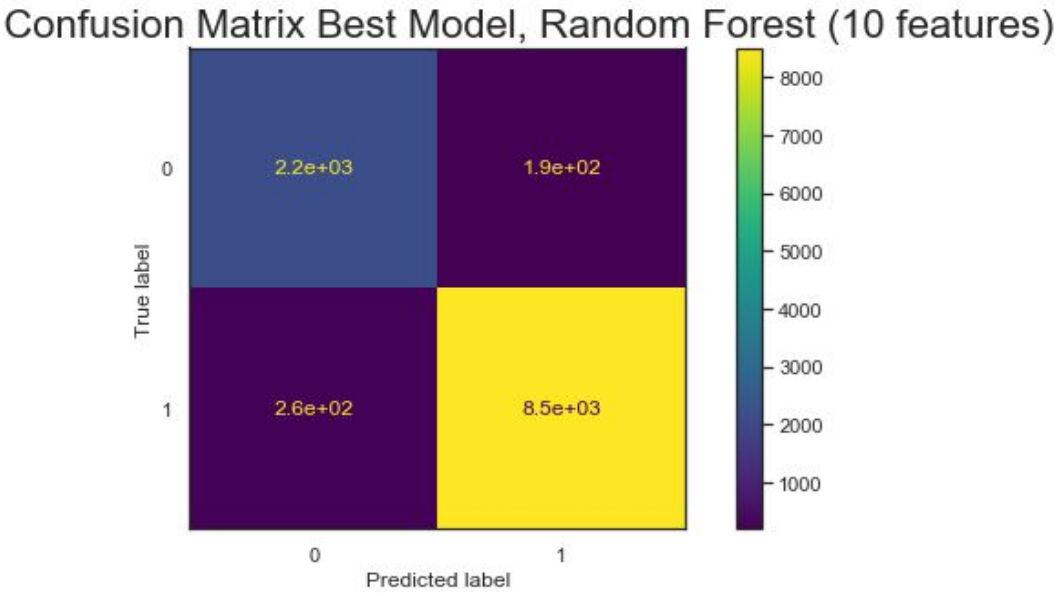


Appendix B5

The most important feature in this model is the number of tweets posted by an account.

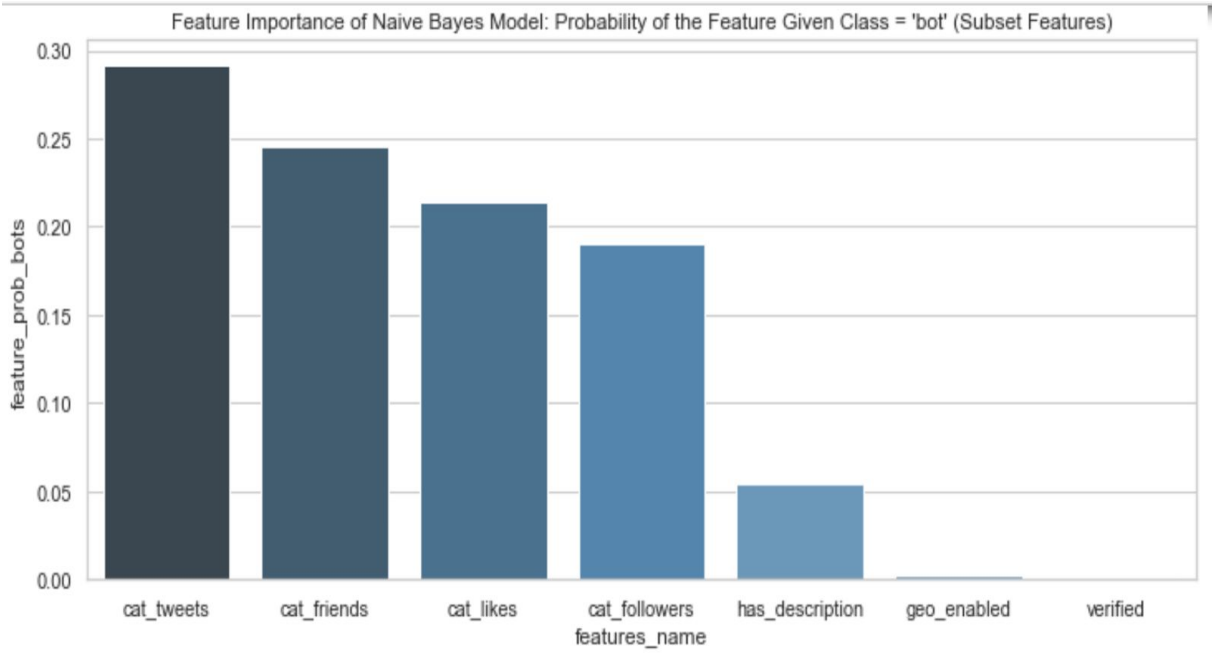


Under this model we got a total of 10,698 true positive and true negatives values, and a total of 447 false positive and false negatives. The following confusion matrix visualizes this values:



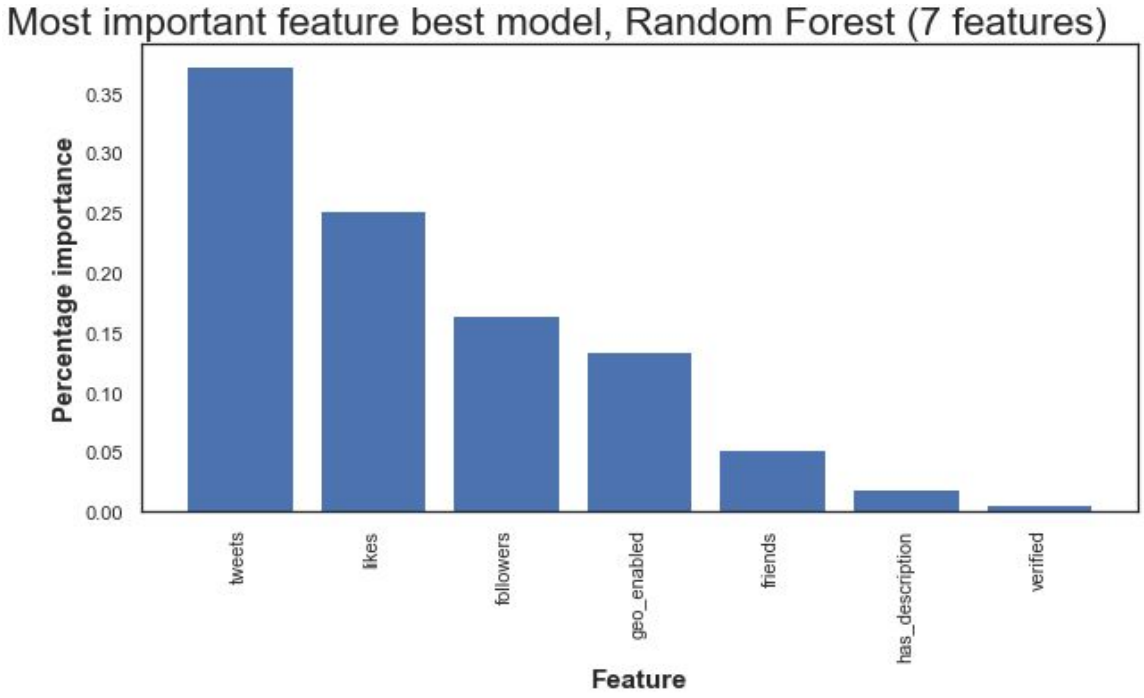
Appendix B6

The most important feature in this model is the number of tweets posted by an account.

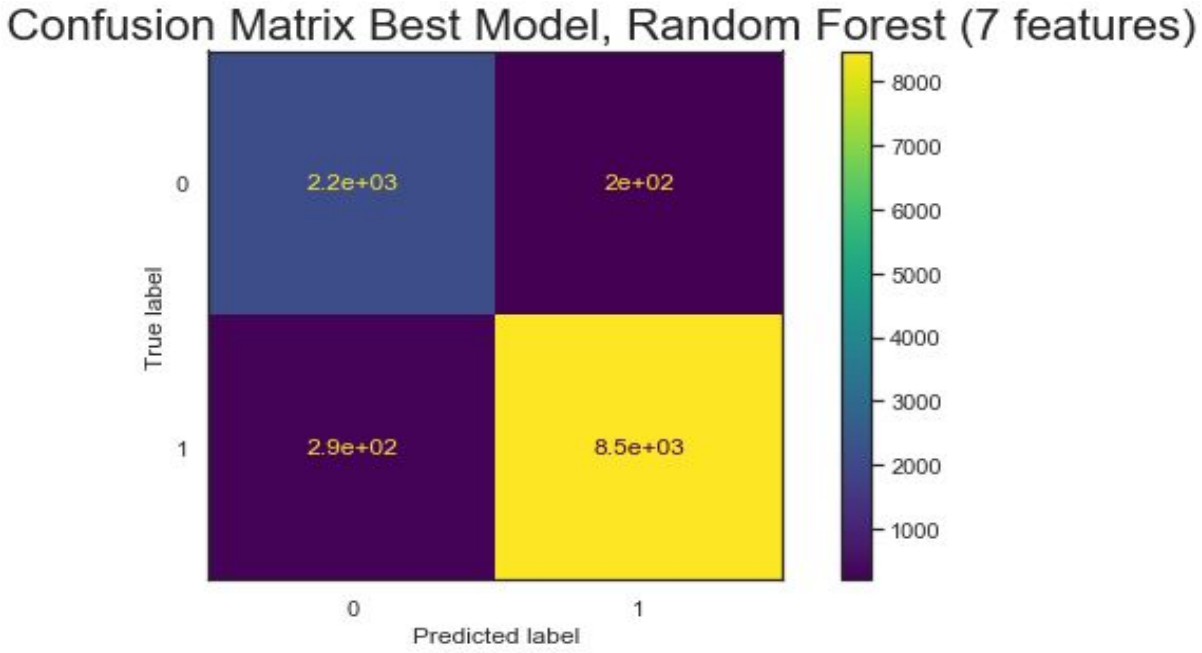


Appendix B7

The most important feature in this model is the number of tweets posted by an account.



Under this model we got a total of 10,647 true positive and true negatives values, and a total of 498 false positive and false negatives. The following confusion matrix visualizes this values:

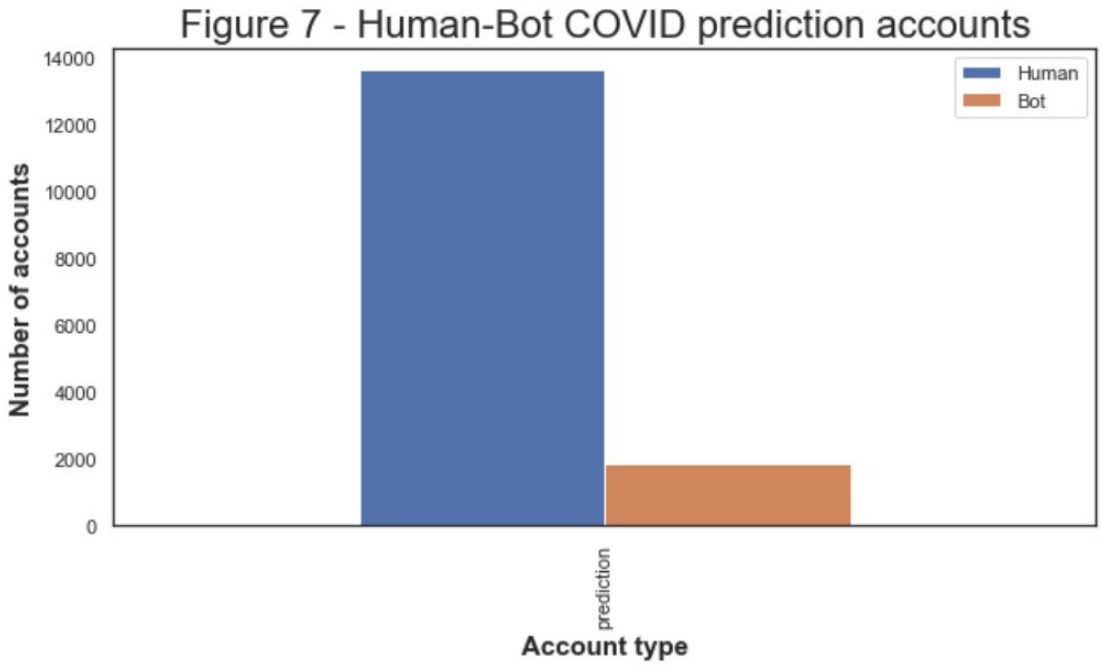


Appendix B8

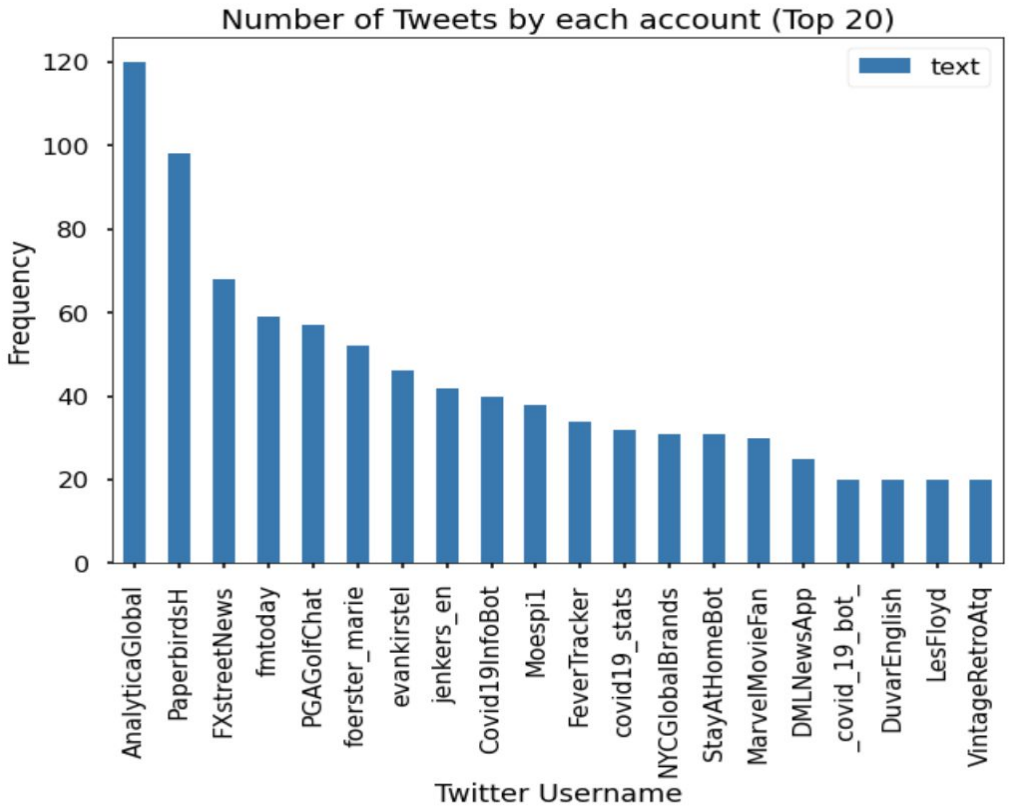
Confusion Matrix for Twitter accounts less than or equal to 90 days old.

Predicted	0	1
Actual		
0.0	10	38
1.0	67	7078

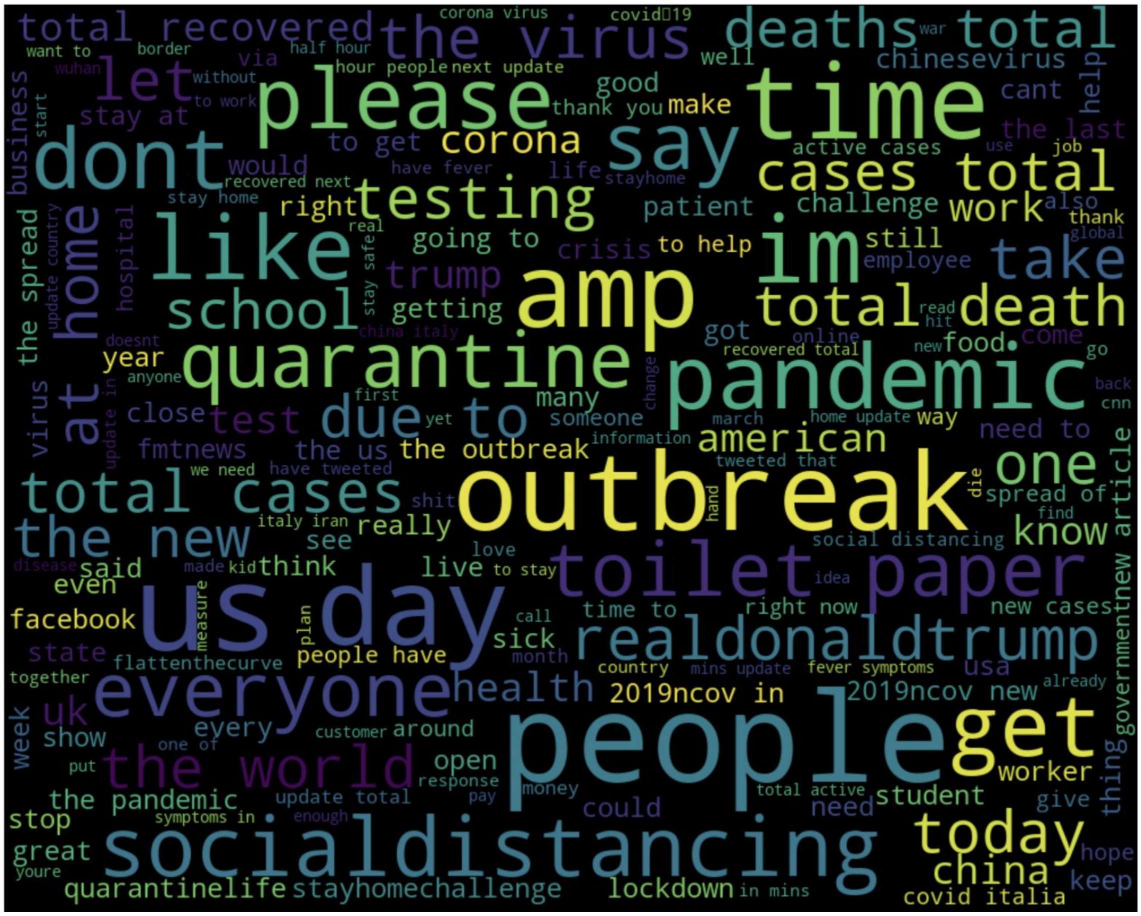
Appendix C1



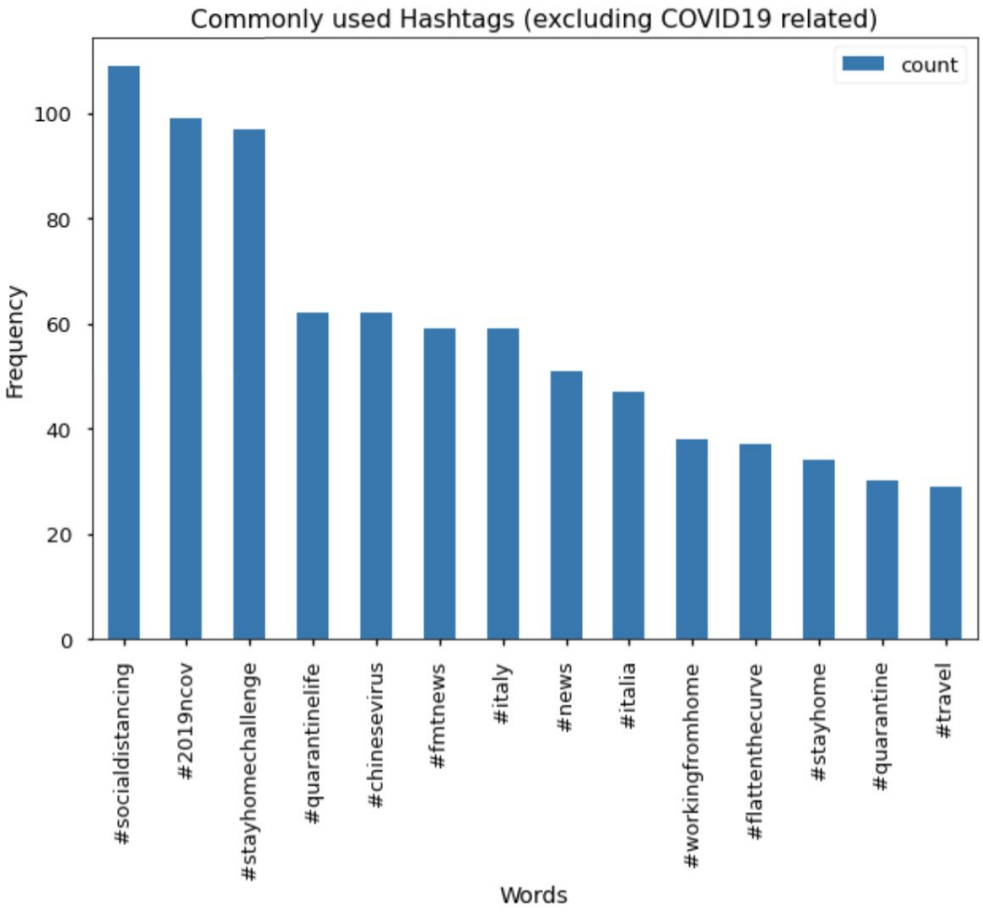
Appendix C2



Appendix C3


total recovered
the virus
deaths
total
please
time
dont
say
corona
cases total
like
testing
im
amp
take
school
total
death
quarantine
pandemic
at home
due to
outbreak
total cases
the new
toilet paper
us
day
realdonaldtrump
everyone
people
get
socialdistancing
today
china
the world
health
2019ncov in
2019ncov new
lockdown
quarantinelife
stayhomechallenge
covid italia
let
chinesevirus
work
one
know
american
the outbreak
test
trump
crisis
going to
getting
right
would
good
make
cant
thank you
stay at
to get
business
the spread
virus
hospital
year
close
fmtnews
the us
many
someone
need to
spread of
article
facebook
said
think
live
sick
even
state
flattenthecurve
people have
uk
week
show
every
around
open
worker
student
the pandemic
stop
great
need
could
give
hope
keep
new cases
right now
time to
usa
government
new
thing

Appendix C4


Commonly used Hashtags (excluding COVID19 related)
count
Frequency
0
20
40
60
80
100
#socialdistancing
#2019ncov
#stayhomechallenge
#quarantinelife
#chinesevirus
#fmtnews
#italy
#news
#italia
#workingfromhome
#flattenthecurve
#stayhome
#quarantine
#travel
Words