Twitter Bot Detection in Ukraine War Context

Florian, Julius, Matteo, Oliver & Triyan

An introduction to social bots on Twitter

Politics and definition issues

Automated Diffusion? Bots and Their Influence During the 2016 U.S. Presidential Election

Olga Boichak, Sam Jackson, Jeff Hemsley & Sikana Tanupabrungsun

Conference paper | First Online: 15 March 2018

4613 Accesses | **9** Citations | **11** Altmetric

Computational Propaganda in Brazil: Social Bots during Elections

Dan Arnaudo, University of Washington

[Submitted on 1 Jul 2017]

Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election

Emilio Ferrara

Rec
exp

[Submitted on 25 May 2018]

Effects of Social Bots in the Iran-Debate on Twitter

Andree Thieltges, Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, Simon Hegelich

2018 started with massive protests in Iran, bringing back the impressions of the so called "Arab Spring" and it's revolutionary impact f
Maghreb states, Syria and Egypt. Many reports and scientific examinations considered online social networks (OSN's) such as Twitter o

Social bots who distort political discussions on Twitter have been a popularly discussed theme over the last year, most notably since the 2016 US election.

News > World > Europe

Revealed: Putin's army of pro-Kremlin bloggers

Hundreds of workers are paid around £500 a month and required to write at least 135 comments per day - or face immediate dismis

Paul Gallagher • Friday 27 Ma

Twitter bans over 100 accounts that pushed #IStandWithPutin

Russia-linked manipulation, but little of the Kremlin's

Laptop generals and bot armies: The digital front of Russia's Ukraine war

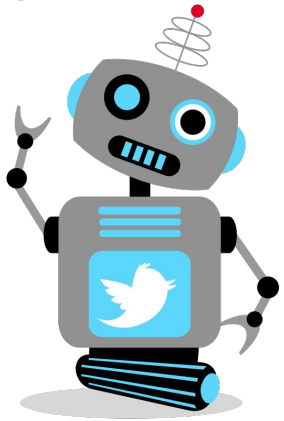
Analysis Digital technology plays a key role in the armed conflict in Ukraine – as a tool for cyberattacks and digital protest, and as an accelerator for information and disinformation.

1 March 2022 by Sabine Muscat and Zora Siebert

Over the last years, state controlled disinformation campaigns have often been associated with the Kremlin. In the Ukraine war, traditional military warfare is accompanied by an information and cyber war.

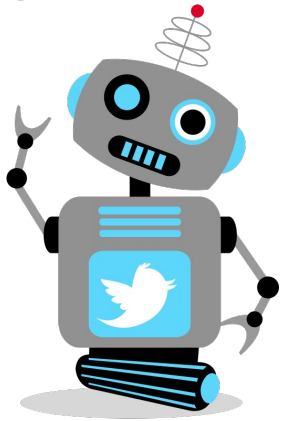
Bot or not?

- Automated accounts controlled by software, often aiming to mimic human users



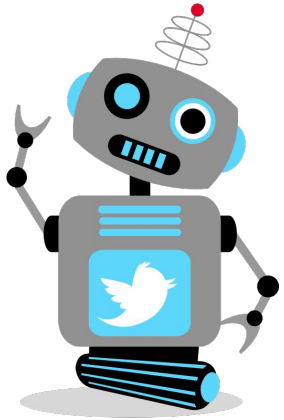
Bot or not?

- Automated accounts controlled by software, often aiming to mimic human users
- Bots are employed for a broad set of tasks:
  - Transparent automation (e.g. for a news outlet)
  - Commercial advertising or scamming
  - Trolling/political disinformation



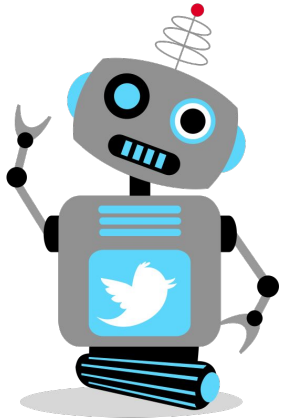
Bot or not?

- Automated accounts controlled by software, often aiming to mimic human users
- Bots are employed for a broad set of tasks:
  - Transparent automation (e.g. for a news outlet)
  - Commercial advertising or scamming
  - Trolling/political disinformation
- Bot/Human might be a grey area:
  - Hybrid forms (Human-in-the-loop)
  - Bot tag almost impossible to verify from the outside



Bot or not?

- Automated accounts controlled by software, often aiming to mimic human users
- Bots are employed for a broad set of tasks:
  - Transparent automation (e.g. for a news outlet)
  - Commercial advertising or scamming
  - Trolling/political disinformation
- Bot/Human might be a grey area:
  - Hybrid forms (Human-in-the-loop)
  - Bot tag almost impossible to verify from the outside
- Automated social media accounts as a phenomenon of data-driven societies



Our research questions

Where do automated bot accounts appear in the discourse on the Ukraine war on Twitter?

Our research questions

Where do automated bot accounts appear in the discourse on the Ukraine war on Twitter?

- What *topics and positions* do they try to propel? What *reach* do they gain?

Our research questions

Where do automated bot accounts appear in the discourse on the Ukraine war on Twitter?

- What *topics and positions* do they try to propel? What *reach* do they gain?
- Do bot accounts *interact*? Can we identify *bot networks*?

Bot Detection

Dataset | Preprocessing | Predictions

Dataset

Ukraine dataset
(26 Mar - 31 Mar)

Dataset

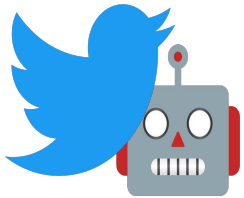
Ukraine dataset
(26 Mar - 31 Mar)

- User id, name
- Nb of followers/friends
- Nb of tweets
- Tweets (text, posting time, nb of retweets & favorites…)

Datasets

Ukraine dataset (26 Mar - 31 Mar)

- User id, name
- Nb of followers/friends
- Nb of tweets
- Tweets (text, posting time, nb of retweets & favorites…)



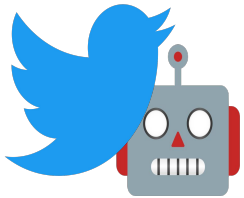
Twibot dataset

- User id, name
- Nb of followers/friends
- Nb of tweets
- Tweets (text, posting time, nb of retweets & favorites…)
- Is user verified?
- Has user a profile image?
- Which domains are usually tackled?
- Is profile protected?
- **Is user a bot?**

Datasets

Ukraine dataset (26 Mar - 31 Mar)

- User id, name
- Nb of followers/friends
- Nb of tweets
- Tweets (text, posting time, nb of retweets & favorites…)



Twibot dataset

- User id, name
- Nb of followers/friends
- Nb of tweets
- Tweets (text, posting time, nb of retweets and & favorites…)
- **Is user a bot?**

Preprocessing (1/2)

Preprocessing (1/2)

Preprocessing (1/2)

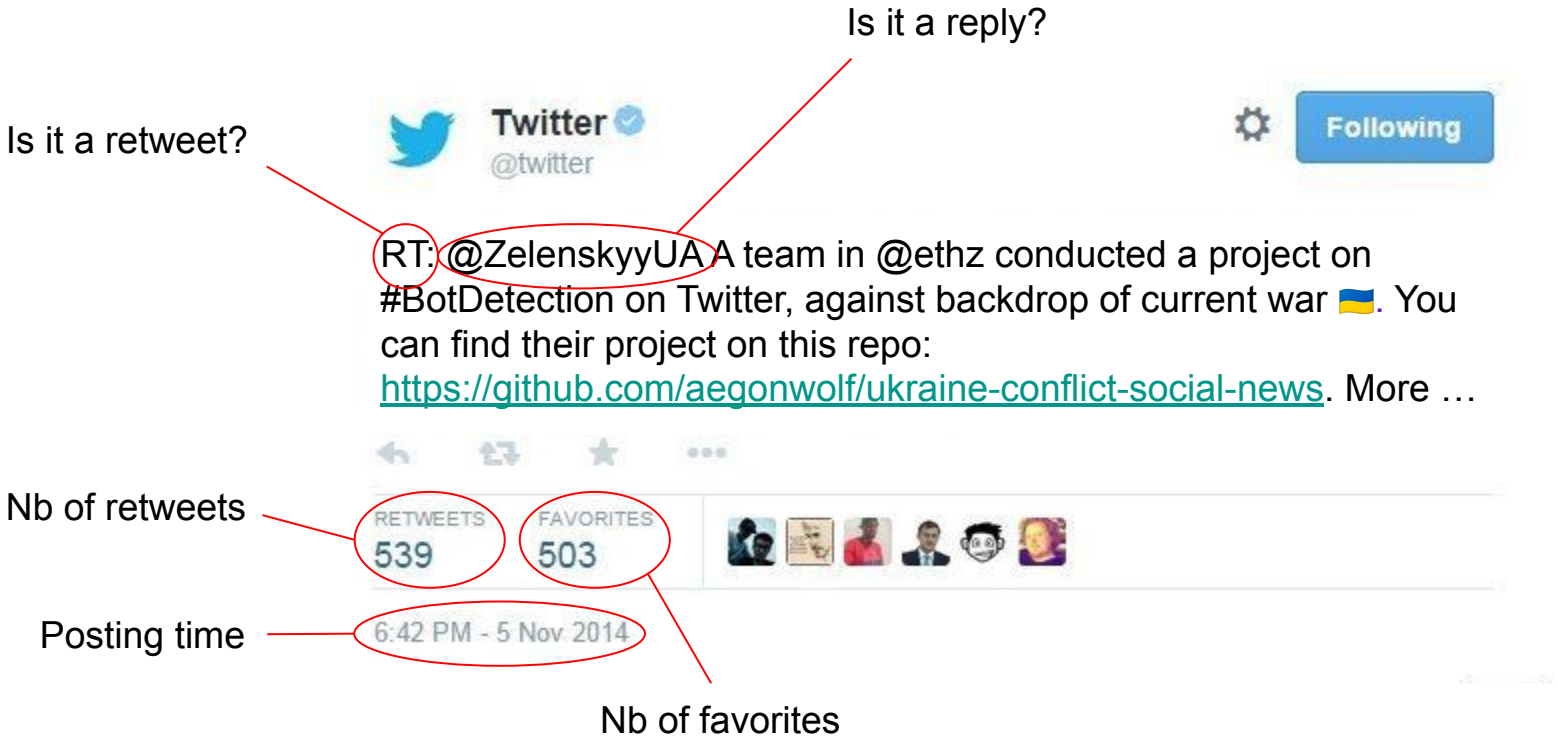
Is it a retweet?

Nb of retweets

Posting time

Nb of favorites

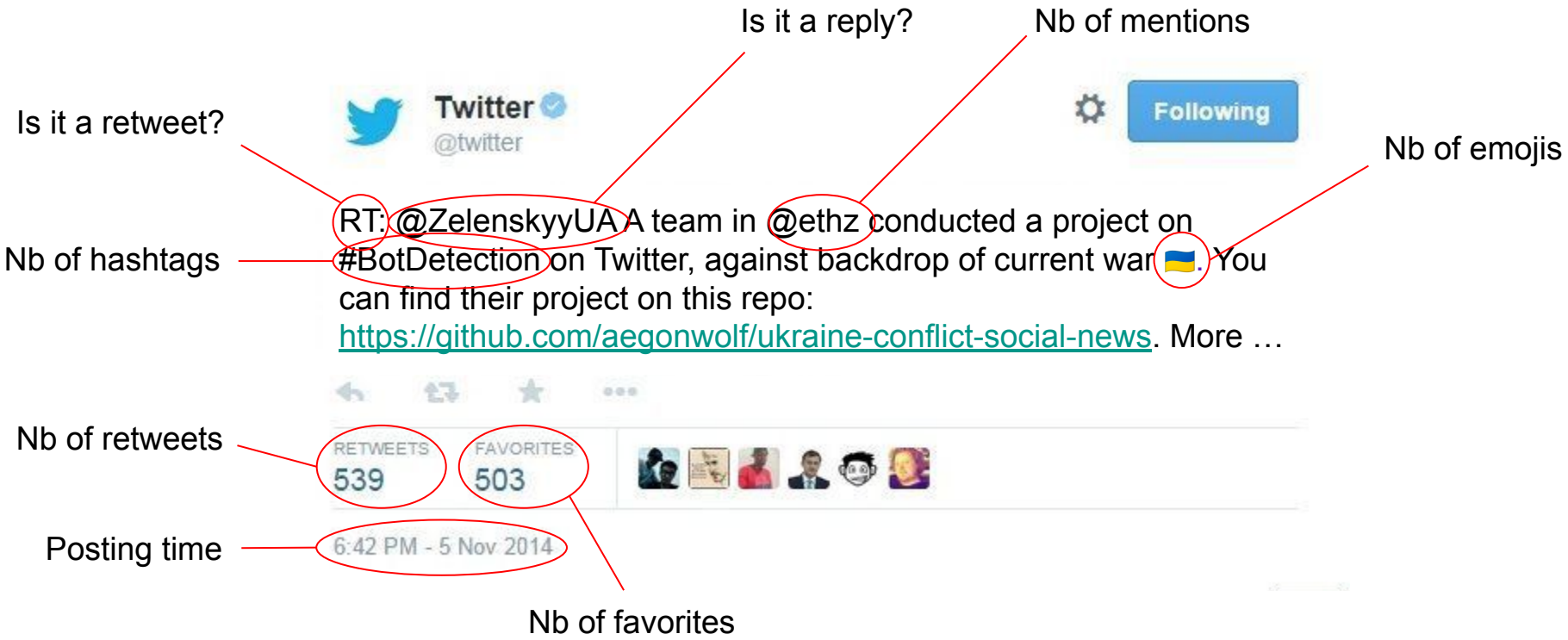
Preprocessing (1/2)



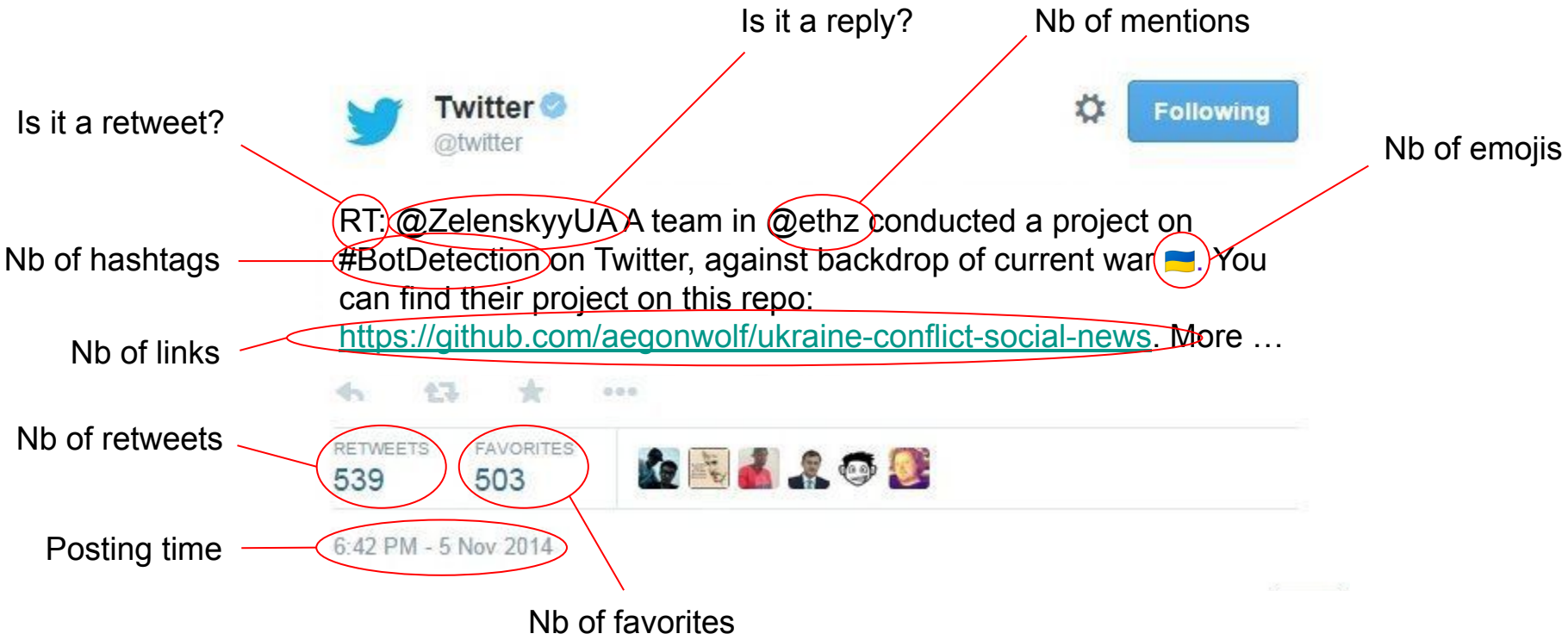
Preprocessing (1/2)

Preprocessing (1/2)

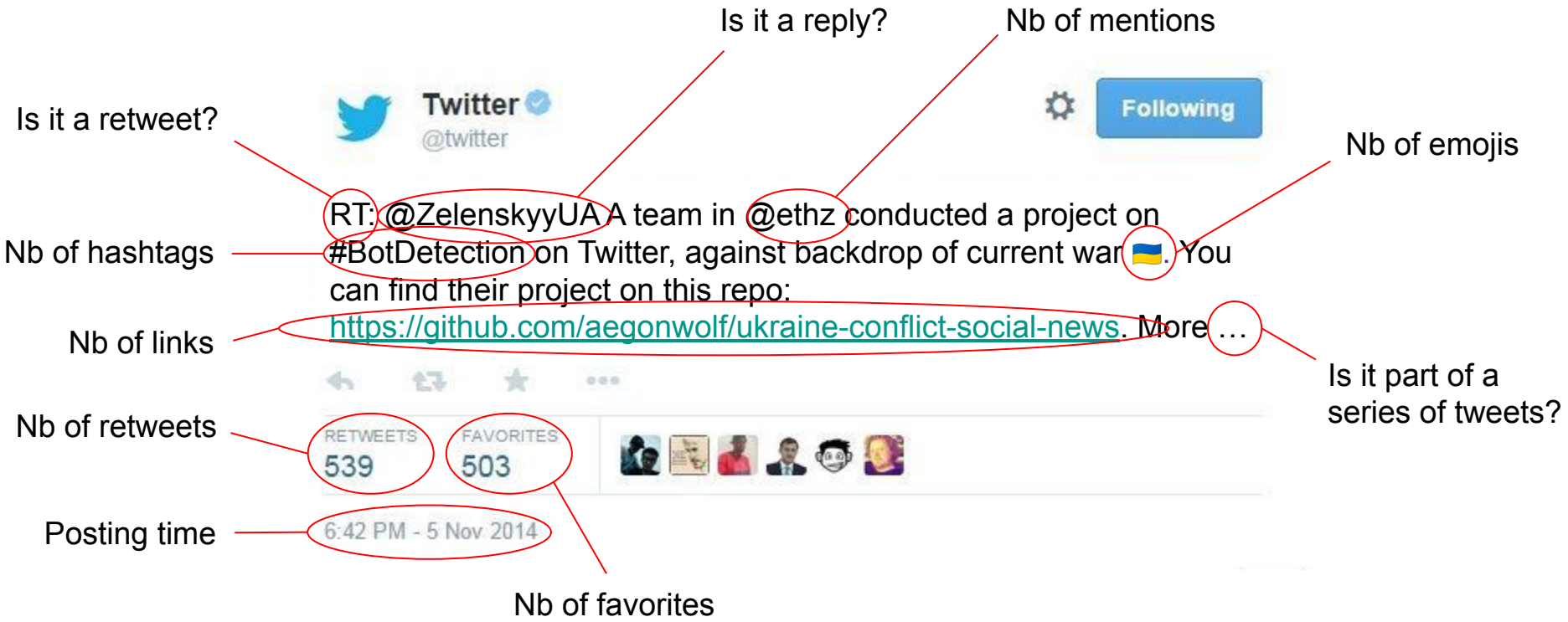
Preprocessing (1/2)



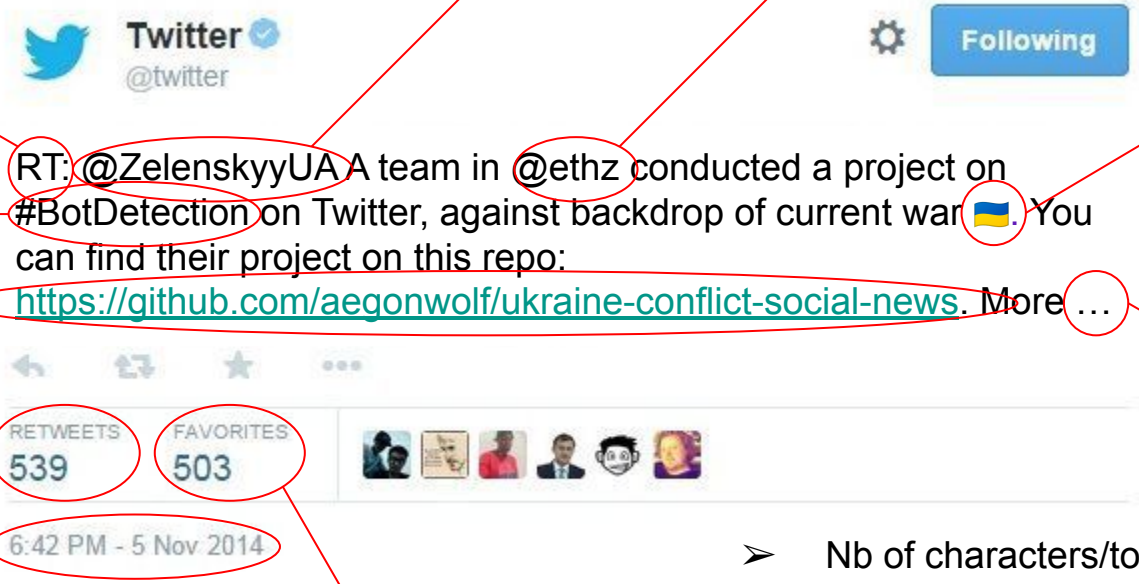
Preprocessing (1/2)



Preprocessing (1/2)



Preprocessing (1/2)



- Nb of characters/tokens
- Average length of tokens
- Nb of punctuation signs, verbs, etc.
- …

Preprocessing (2/2)

User-level metadata

- Nb of followers
- Nb of friends

Tweet-level metadata

- Is modified?
- Is a reply?
- Has an ellipsis?
- Saturation (nb of characters / 240)
- Ratio of unknown characters
- Nb of cashtags
- Nb of hashtags
- Nb of links
- Nb of mentions
- Nb of emojis

Preprocessing (2/2)

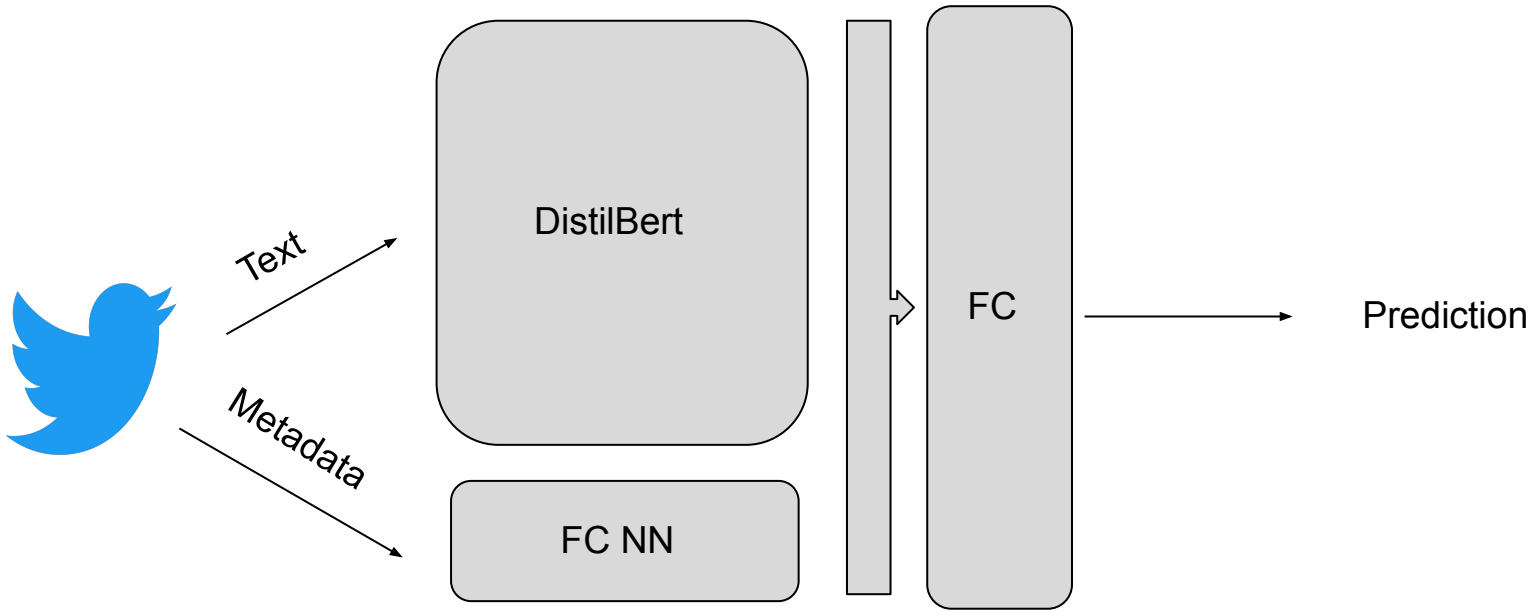
User-level metadata

- Nb of followers
- Nb of friends

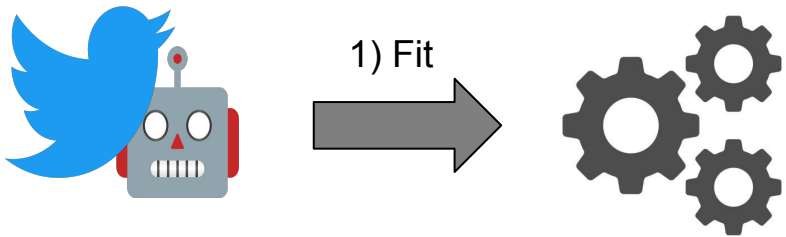
Tweet-level metadata

- Is modified?
- Is a reply?
- Has an ellipsis?
- Saturation (nb of characters / 240)
- Ratio of unknown characters
- Nb of cashtags
- Nb of hashtags
- Nb of links
- Nb of mentions
- Nb of emojis

Basic Model Structure

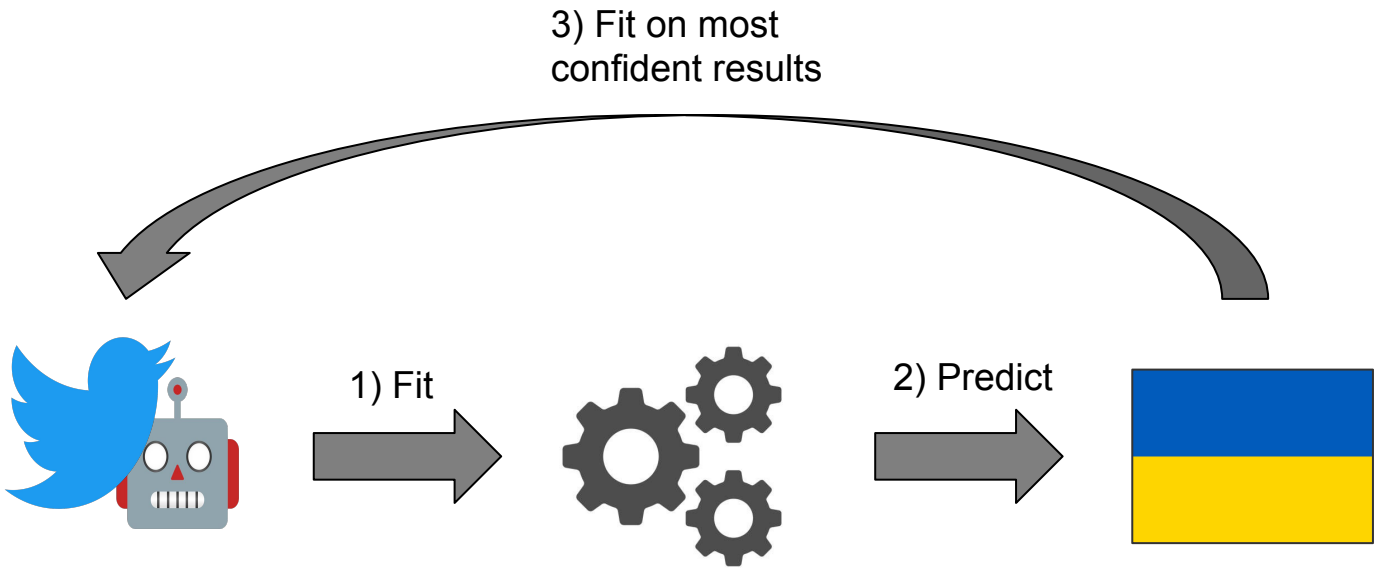


Predictions

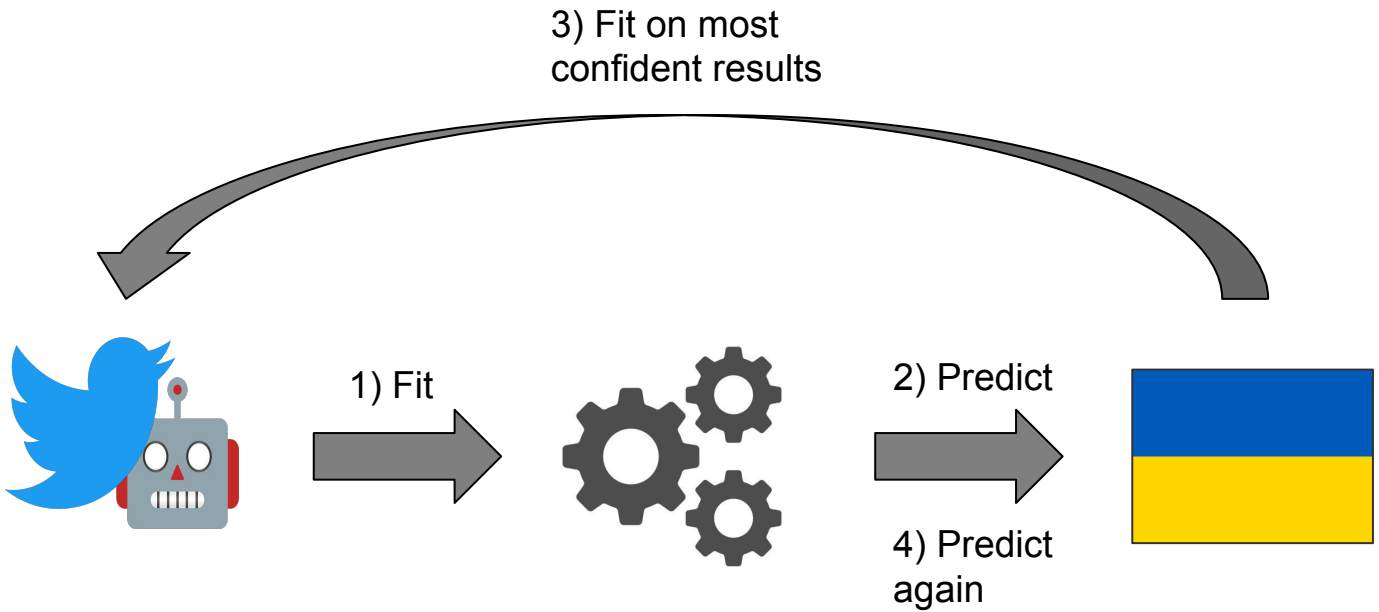


Predictions

Predictions



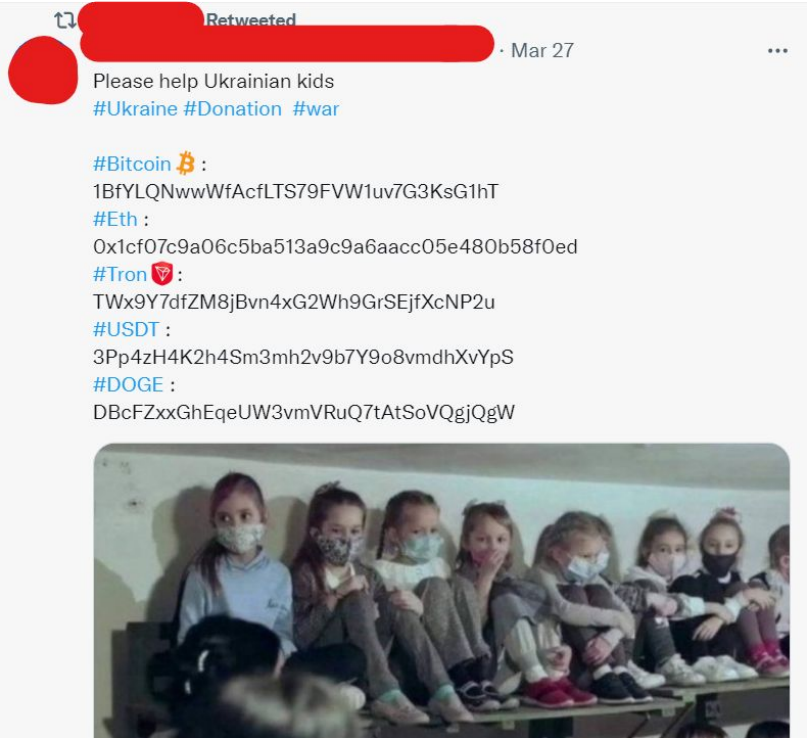
Predictions



Results

Let’s look at some bots!

Crypto bots everywhere!



- Bot confidence: 97.8%
- Part of a bot network posing as charities
- Possibly created in response to Ukraine's official call for crypto donations
- Always exact same tweet (except first line), posted multiple times a day by multiple accounts
- The addresses seem to have no transactions
- False negatives: not all members of the network could be identified
- We decided to filter out crypto bots

Your typical spam bots

...

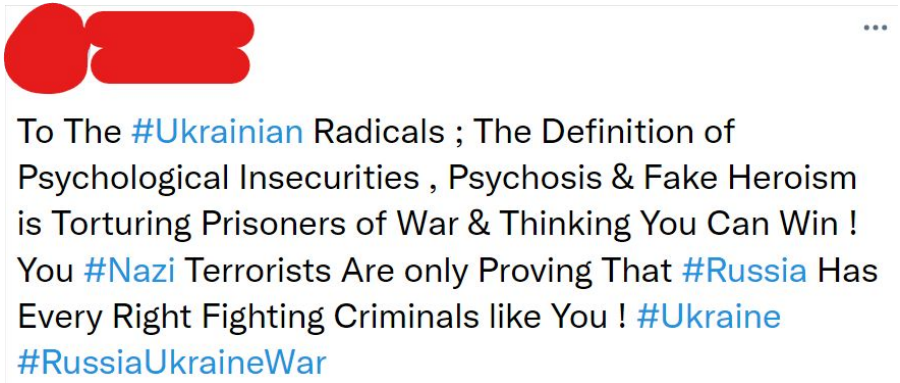
@SianMorg @TUIUK #UkraineUnderAttack #boycottTUI #Billionaire #Alexey_Mordashov, ardent supporter of war criminal #Putin, has simply handed over his shares in #TUI, largest tourism company in the world, to his wife. TUI bookings supports murder of innocent people in the #UkraineWar

...

#Ukraine #UkraineRussiaWar #russia #Russians #RussianWarCrimes #RussiaInvadesUkraine Breaking News: Just Now Ukraine vs Russia War: Watch Video 😭😭😭😭😭😭😭😭😭😭😭😭😭😭 https://t.co/LcYfdqcKim

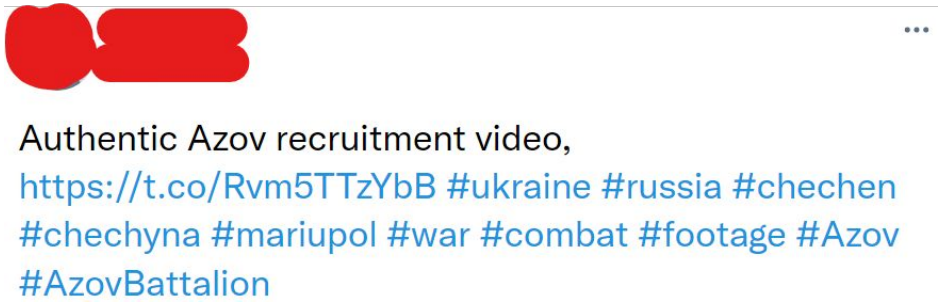
- Bot confidence: 92.0% and 97.5%
- On the order of 100 tweets over the course of a few days in burst fashion (precisely every 6-7s)

Pro Russian stance



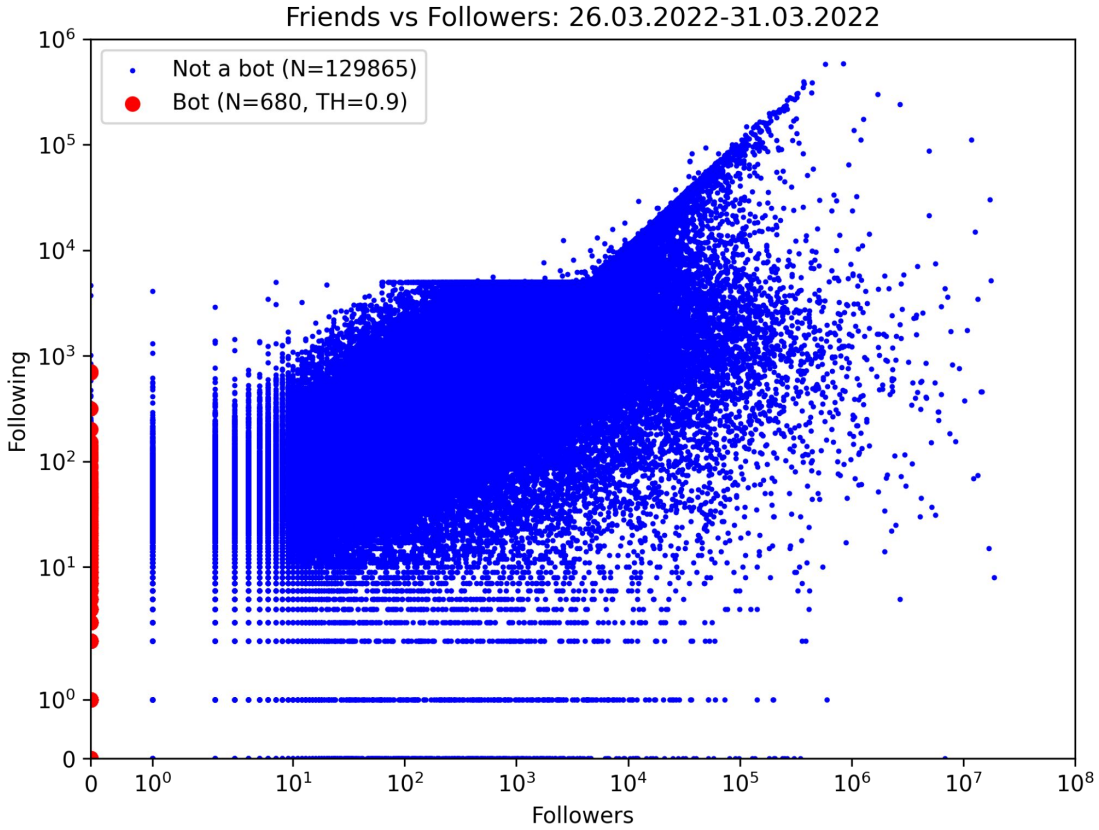
- Bot confidence: 98.1% and 94.1%
- Many bots resort to the “denazification” narrative of the Russian government
- Account suspended

False positive

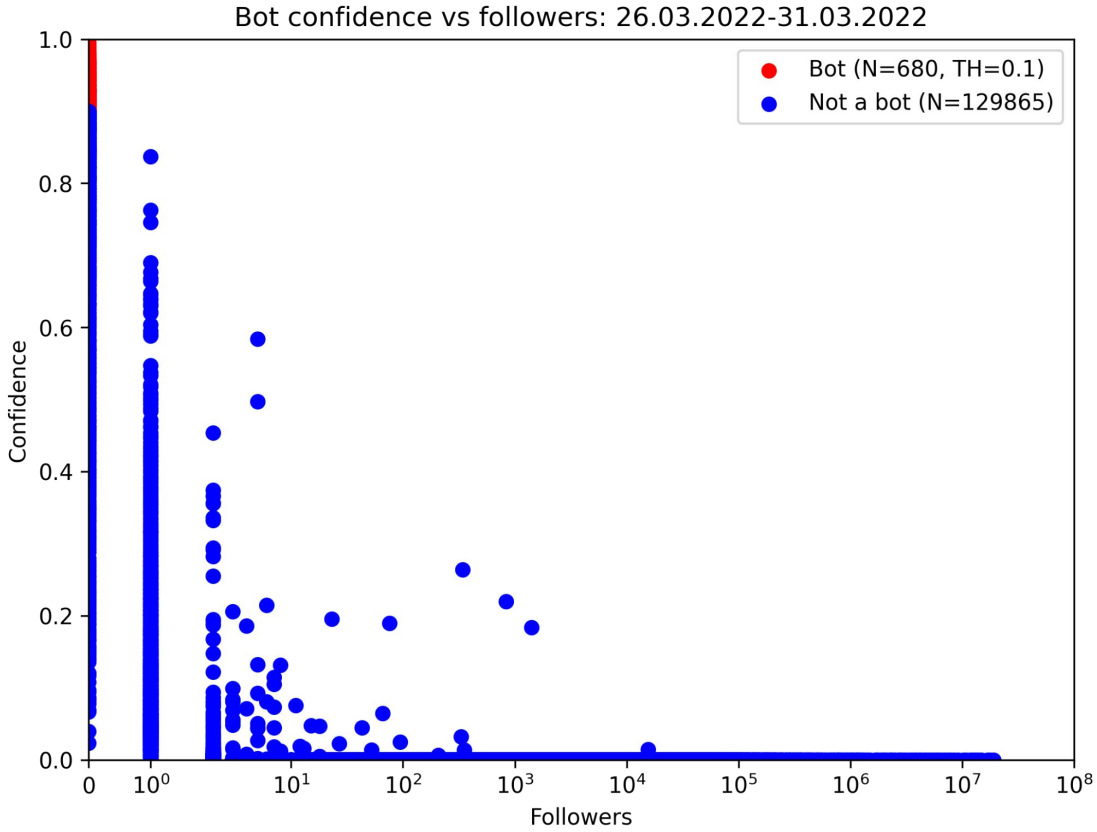


- Bot confidence: 97.8%
- It has the hashtags, and a URL
- User has no followers/friends
- Account belongs to an American YouTuber independently reporting on/analysing the Russia-Ukraine events

Analysing Bot Activity - Friends/Followers



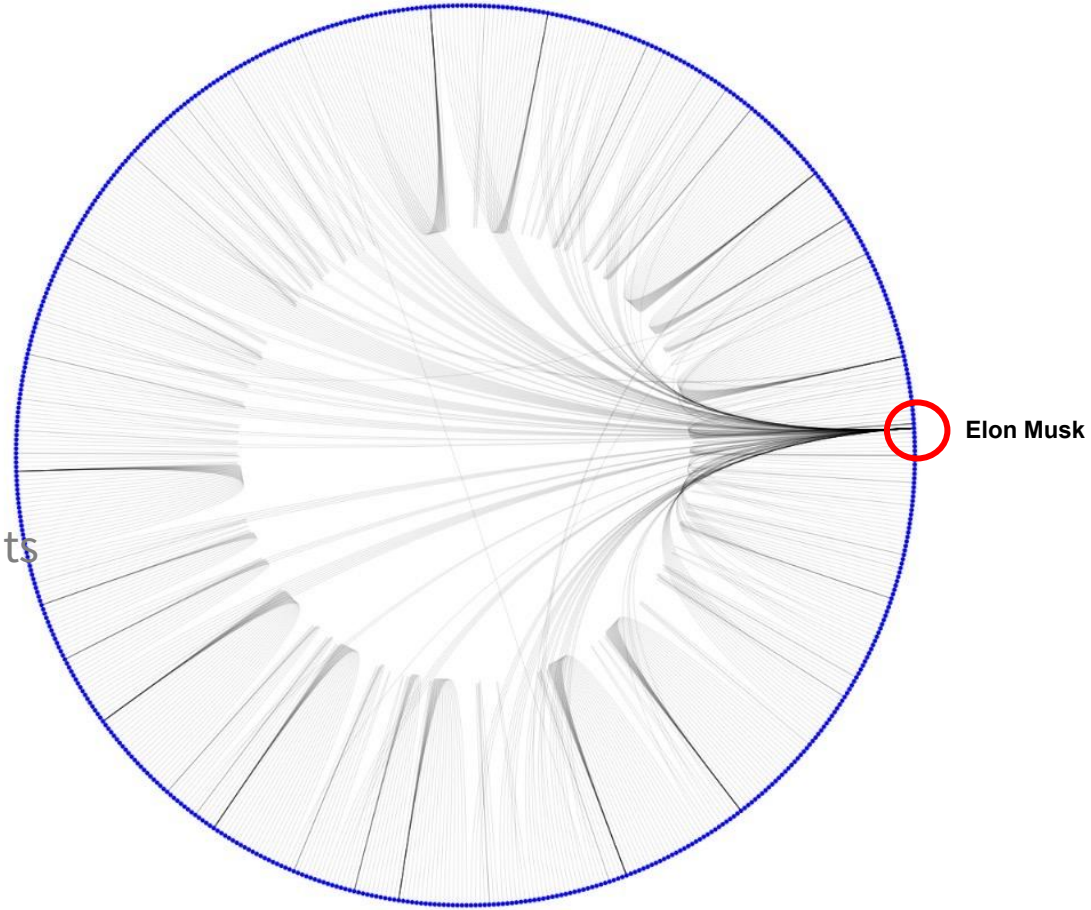
Analysing Bot Activity - Confidence/Followers



Who do bots follow?

Bots are not followed by many other users

Bots tend to follow large-scale accounts (e.g. Elon Musk, Zelenskyy, POTUS)

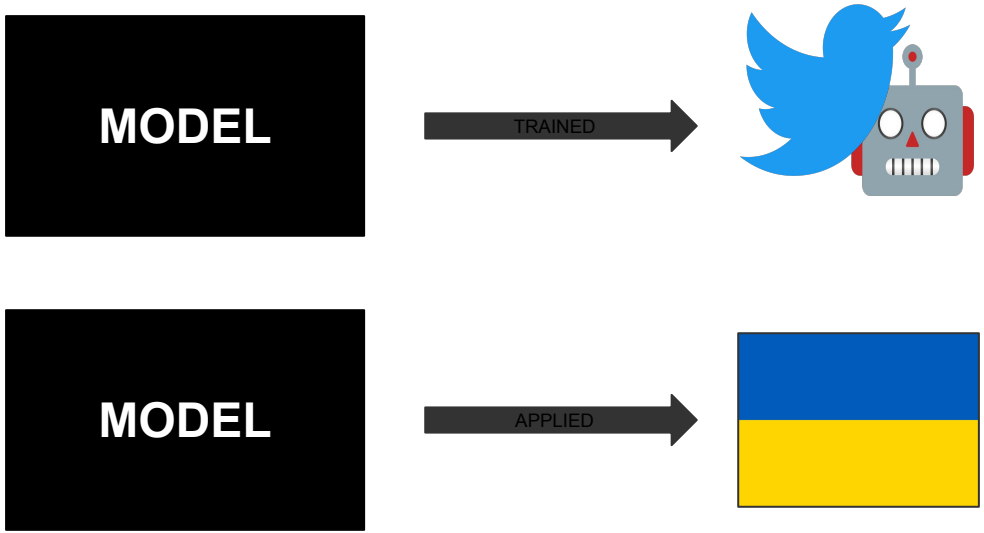


Graph visualization of accounts followed by bots

Trustworthiness of our results

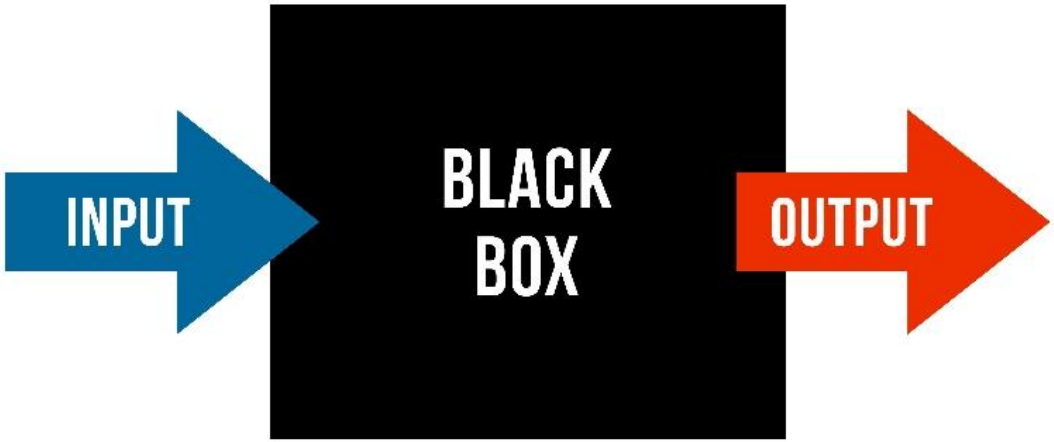
Is the dataset reliable? (1/3)

Two datasets: Twibot (labelled), Ukraine dataset (unlabelled)

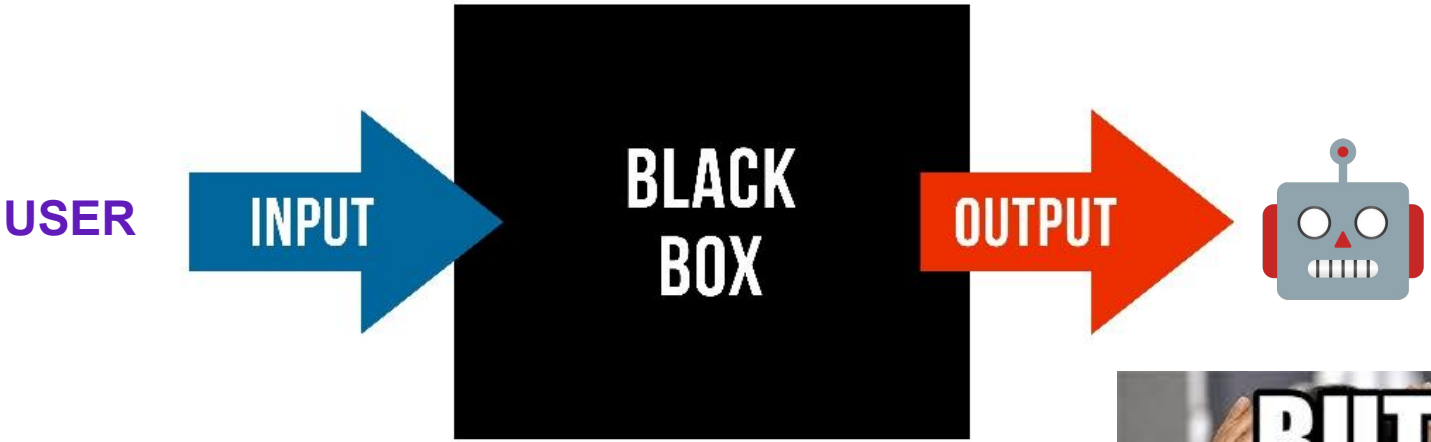


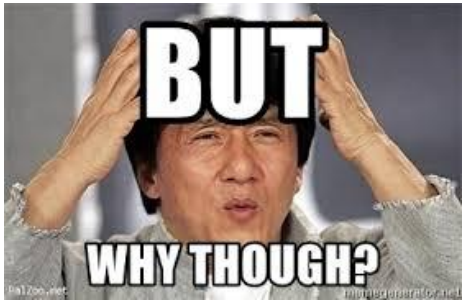
Interpretability in Machine Learning

High performance ML models are often complex and hard to interpret



Why do we need interpretability?





Is the dataset reliable? (2/3)

- Five annotators are then assigned to each user in TwiBot-20 to determine whether it is operated by bot or not.
- In order to identify potentially ambiguous cases, annotators are permitted to report ‘undecided’ regarding a specific user.
- Test questions where also mixed with real annotations tasks to assess annotators performance.
- Annotators who are more than 80% correct on standard questions are considered to be trustworthy and their annotation is adopted.

Is the dataset reliable? (3/3)

- User labelled accordingly if ⅘ annotators agree upon
- Twitter's direct message feature was used to send out simple questions in natural language, the annotation was then performed manually
- Remaining undecided users were manually examined within the research team. Disputed cases were discarded and only users where a consensus was reached were annotated.

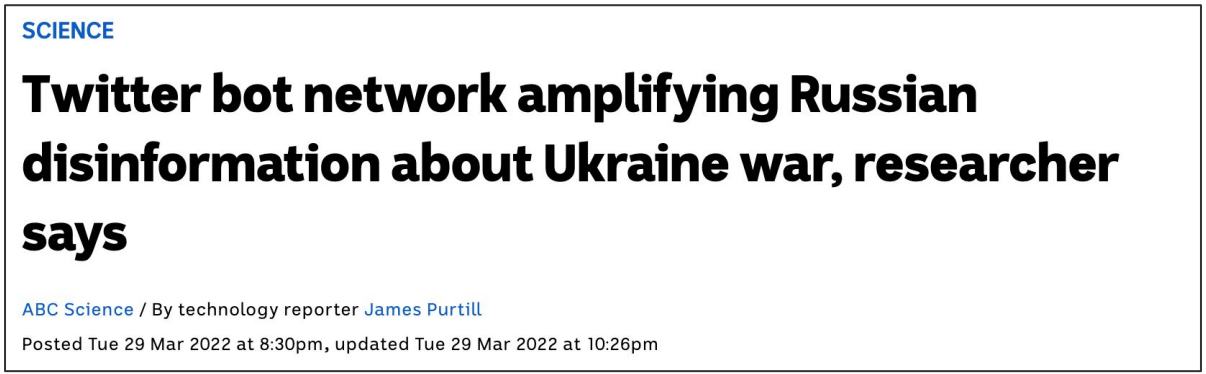
Conclusion (1)

- Within our dataset from 26 - 31 Mar, we labelled bots with various purposes and across the political spectrum in the Ukraine conflict – alongside some false positives.
- Because of the nature our topic, the reliability of our labels is hard to evaluate.
- None of the bots gained significant attention. Interaction was only observable on a very limited scale.
- Many of these accounts have been blocked by Twitter in the meantime.

Conclusion (2)

This is not the end of the line. For further research, it would be worthwhile…

- … to increase our timescale. Other research has already identified bot networks that operated for some time.
- … to compare disinformation attempts by humans and automated accounts.



SCIENCE

Twitter bot network amplifying Russian disinformation about Ukraine war, researcher says

ABC Science / By technology reporter James Purtill

Posted Tue 29 Mar 2022 at 8:30pm, updated Tue 29 Mar 2022 at 10:26pm

Thanks for your attention!

Elon Musk
@elonmusk

If our twitter bid succeeds, we will defeat the spam bots or die trying!

8:53 PM · Apr 21, 2022 · Twitter for iPhone

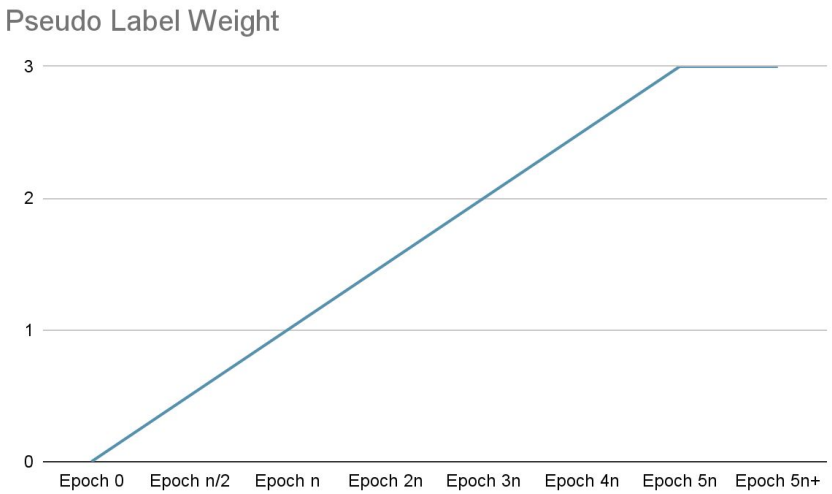
78.9K Retweets **12.8K** Quote Tweets **901.2K** Likes

Supplementary Material

Pseudo-Labelling with Neural Networks

Semi-Supervised Learning => Bootstrapping

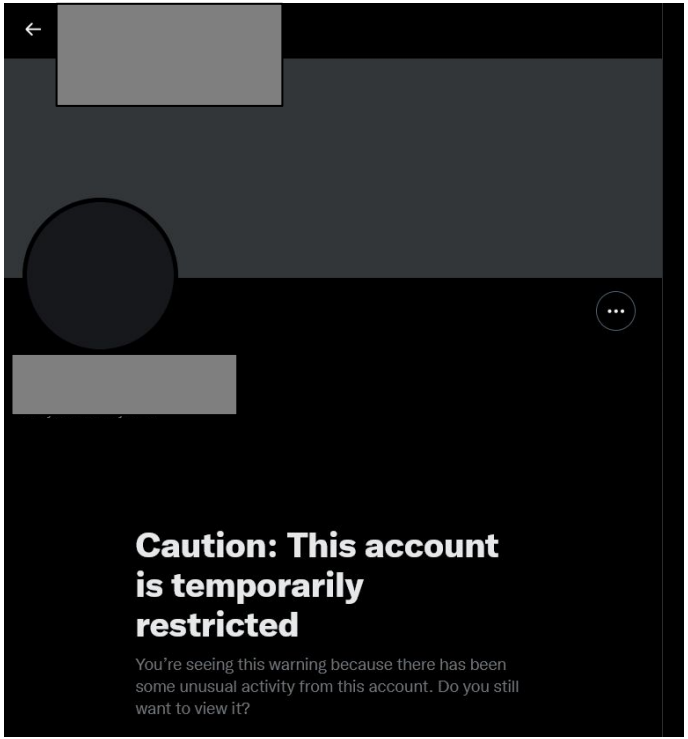
Basic Idea:
Loss = Labeled Loss + weight * Unlabeled Loss

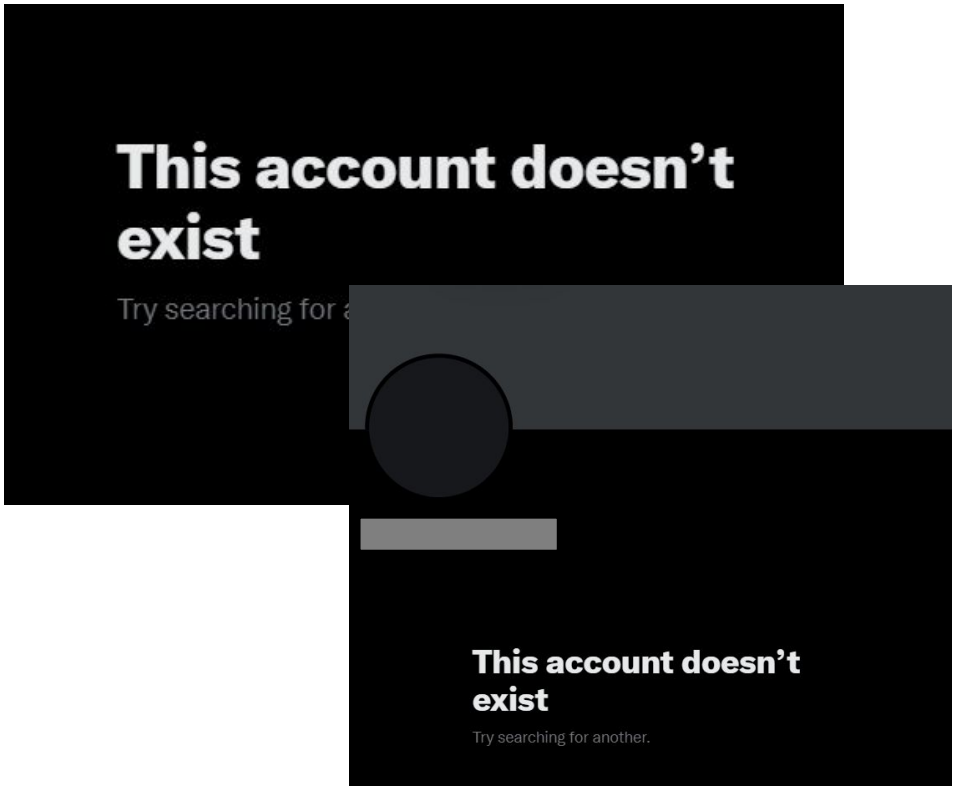


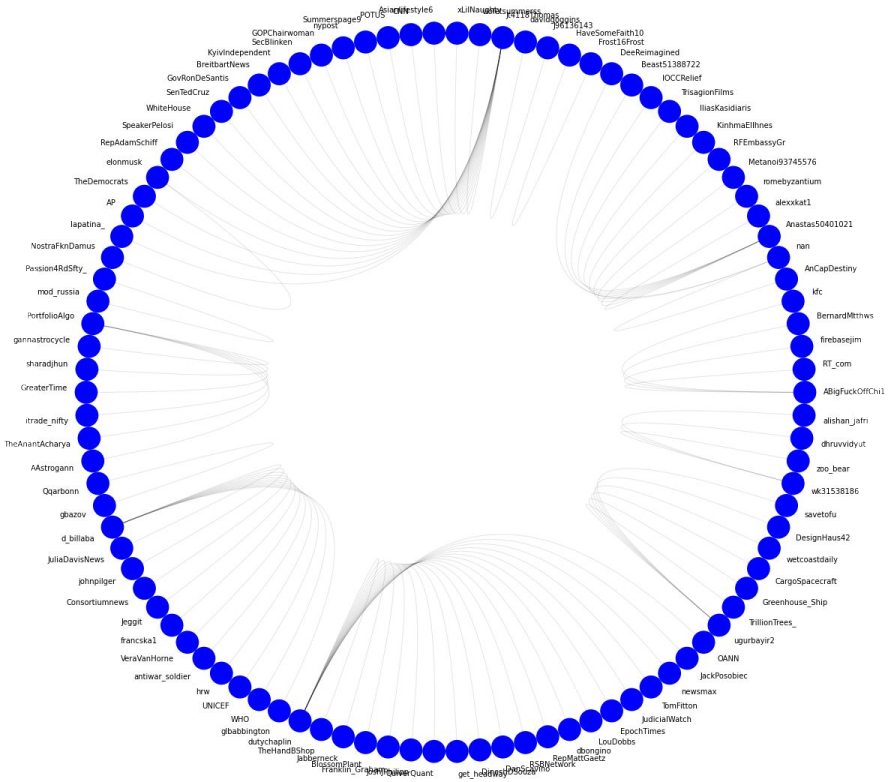
- Also possible per batch
- Different Learning Rates

- Empirically proven
- Dangers of overfitting

Deleted, suspended or restricted - a common theme

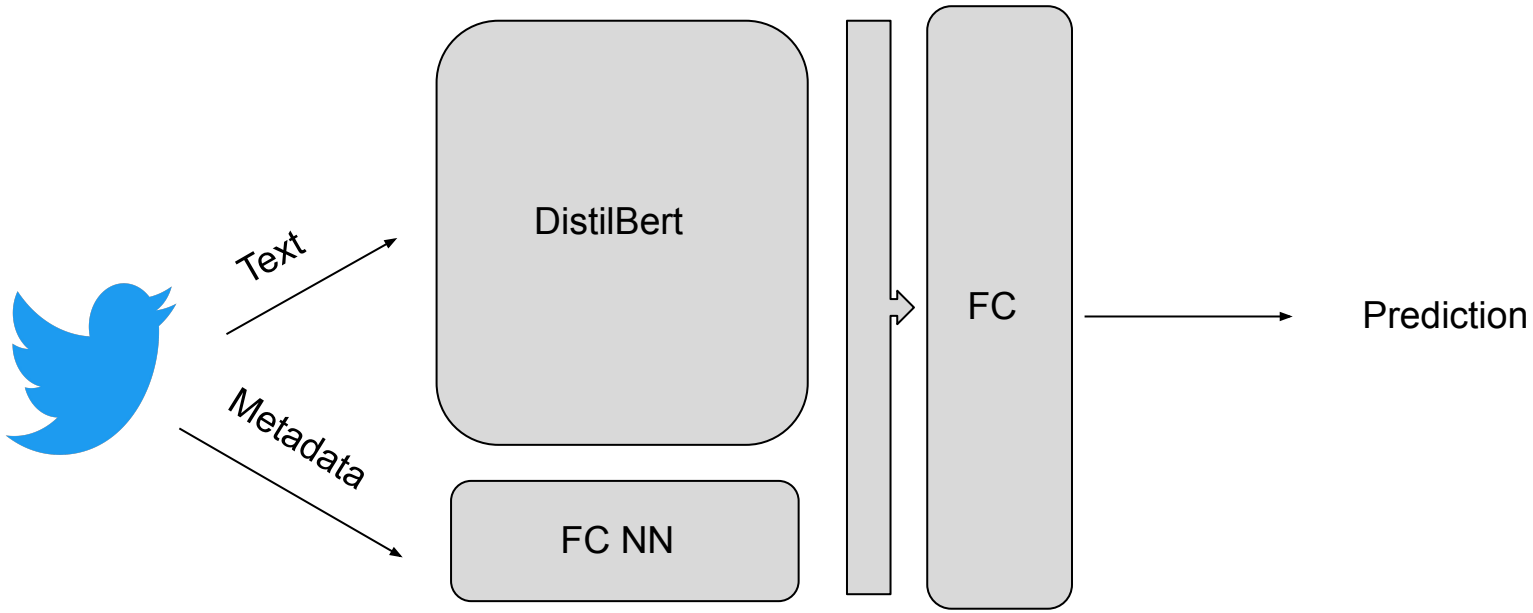




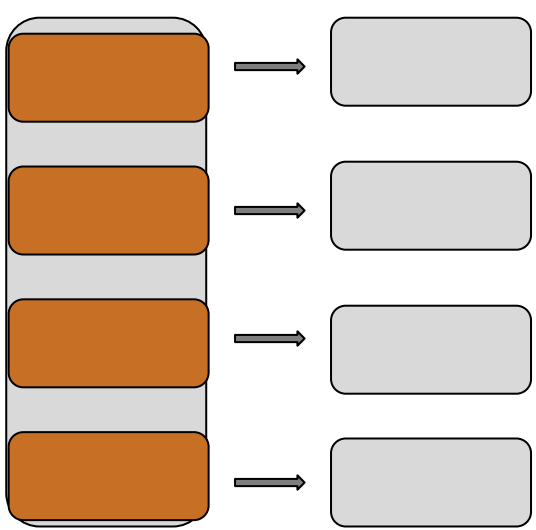


Summerspage9
nypost
POTUS
GOPChairwoman
SecBlinken
KyivIndependent
BreitbartNews
GovRonDeSantis
SenTedCruz
WhiteHouse
SpeakerPelosi
RepAdamSchiff
elonmusk
TheDemocrats
AP
lapatina_
NostraFknDamus
Passion4RdSfty_
mod_russia
PortfolioAlgo
gannastrocycle
sharadjhun
GreaterTime
itrade_nifty
TheAnantAcharya
AAstrogann
Qqarbonn
gbazov
d_billaba
JuliaDavisNews
johnpilger
Consortiumnews
Jeggit
francska1
VeraVanHorne
antiwar_soldier
hrw
UNICEF
WHO
glbabbington
dutychaplin
TheHandBShop
jabberneck
BlossomPlant
Franklin_Graham
JoshuaQuinn
Quiver Quant
get_headway
DiscussDaily
DoscoSchmid
RSBNetwork
RepMattGaetz
dbongino
LouDobbs
EpochTimes
JudicialWatch
TomFitton
newsmax
JackPosobiec
OANN
ugurbayir2
TrillionTrees_
Greenhouse_Ship
CargoSpacecraft
wetcoastdaily
DesignHaus42
savetofu
wk31538186
zoo_bear
dhruvvidyut
alishan_jafri
ABigFuckOffChi1
RT_com
firebasejim
BernardMtthws
kfc
AnCapDestiny
nan
Anastas50401021
alexxkat1
romebyzantium
Metanoi93745576
RFEmbassyGr
KinhmaEllhnes
IliasKasidiaris
TrisagionFilms
IOCCRelief
Beast51388722
DeeReimagined
Frost16Frost
HaveSomeFaith10
j96136143

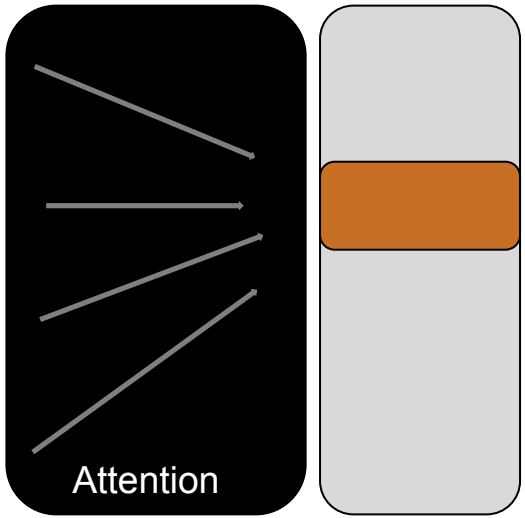
Basic Model Structure



The Transformer



Encoder



Attention + Decoder

Hidden State for each input from the encoder

Each output assigns weight (attention) to generate output state in the decoder
Extension self-attention
Basically the same also for the encoder

What do they tweet about? (in May)

General

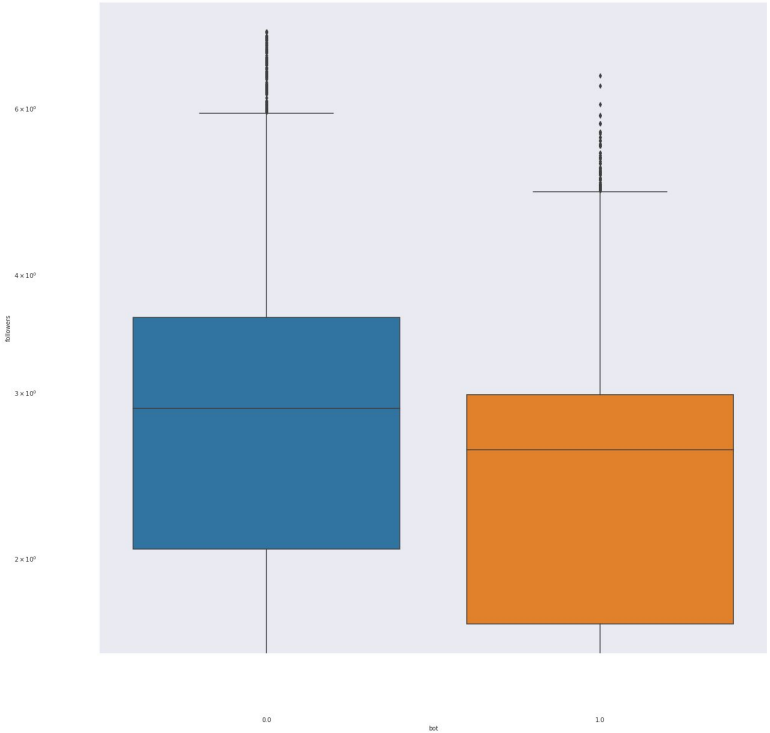
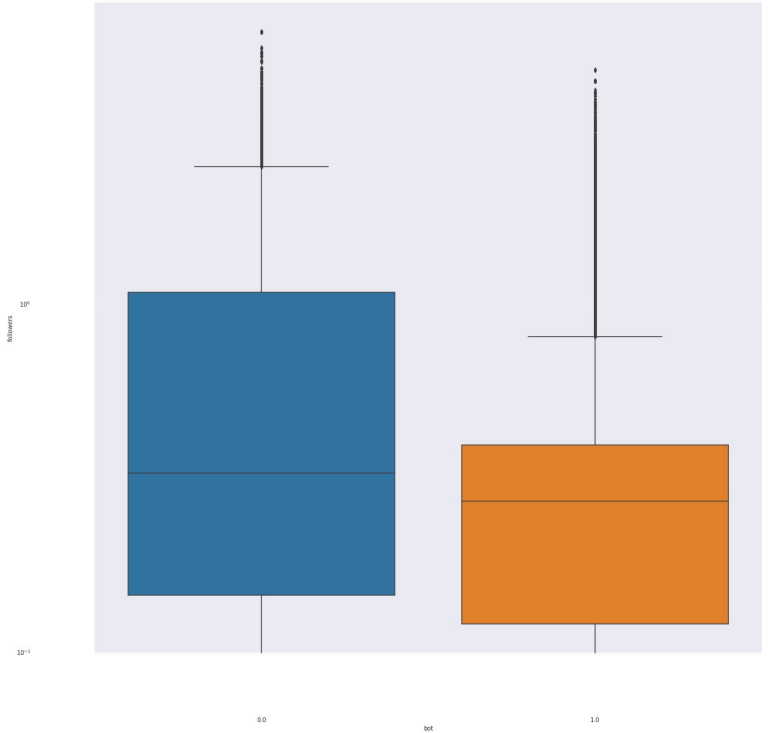
Bots

Bot or No Bot by text only Training Twibot

- 200 tweets per user in Twibot
- Train/validation/test split via userid, i.e. no leak via same user/different tweets
- In the Ukraine dataset: With validation and pseudolabelling this is not done, but learning rate has been lowered massively to reign in overfitting

Followers and Following Distributions

Followers Human/Bot



Ratio Human/Bot