Universitat Oberta de Catalunya

uoc.edu

Detecció de bots en xarxes socials per mètodes supervisats

Josep Consuegra Navarrina
Master en Big Data i Data Science
Àrea 5

Julià Vicens Bennassar
Albert Solé Ribalta

01/03/2020

FITXA DEL TREBALL FINAL

Títol del treball:	*Detecció de bots en xarxes socials per mètodes supervisats*
Nom de l'autor:	*Josep Consuegra Navarrina*
Nom del consultor/a:	*Julià Vicens Bennasar*
Nom del PRA:	*Albert Solé Ribalta*
Data de lliurament (mm/aaaa):	*06/2020*
Titulació o programa:	*Màster en Big Data i Data Science*
Àrea del Treball Final:	*5*
Idioma del treball:	*Català*
Paraules clau	*Social Networks, Misinformation, Graphs*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

Aquest projecte neix de la vulnerabilitat de l'opinió pública a través de les xarxes socials, on la presència de bots, principals responsables de la propagació de noticies falses i promotors de la desinformació, actuen amb certa impunitat aprofitant la falta de protocols i directrius de control.

L'objectiu de l'estudi és, principalment, la implementació d'un mètode de catalogació binari d'usuaris de Twitter, per etiquetar-los com a humans o bots, a partir d'un conjunt d'aproximadament sis-cents cinquanta mil tweets obtinguts entre el 24 d'Abril i el 5 de Maig, i d'un conjunt de dades d'entrenament obtingut a través de l'API Botometer.

Per tal fi, s'empren mètodes de classificació supervisats a partir de l'activitat d'aquests usuaris (sense contemplar el contingut dels missatges), obtenint una comparativa dels models estudiats en les qual els classificador MLP i Random Forest semblen genera els millors resultats.

De cara a complementar l'estudi, es classifiquen tots els usuaris del data set inicial de projecte i es genera un graf per tal de visualitzar els resultats, en el qual tots usuari existeix com un node, i totes les interaccions entre usuaris es representen amb una aresta. Addicionalment, s'aplica un algoritme de detecció de comunitats, i es visualitza el graf d'usuaris obtingut a l'aplicació Gephi, observant una polarització dels usuaris i una distribució homogènia de bots en tota la xarxa d'interaccions, en la qual cap comunitat n'està aïllada.

Abstract (in English, 250 words or less):

This project is born as a result of the public opinion vulnerability in regard to social networks, where bot presence, main responsible of fake news propagation and misinformation spread, act with impunity by taking advantage of non-existing or inefficient bot detection (and control) protocols.

The goal of this project is, mainly, to implement a binary classification algorithm for Twitter users, in charge of detecting whether a user is behaving as a bot or not. The algorithm is based on a user activity dataset consisting of 650k tweets downloaded through the Twitter API between April 24th and May 5th, as well as a training dataset obtained by using Botometer API.

Only supervised methods are considered for the implementation, based on the users' activity in Twitter (without considering the contents of the tweet's body), which are afterwards compared, showing that MLP and Random Forest classifiers seem to perform better in this scenario.

For visualization purposes, all users from the original dataset are then classified as a human or a bot, and are added into a graph, where each node represents a user and each edge represents an interaction. Additionally, a community detection algorithm is applied, and the graph is visualized through Gephi tool, showing that there is a polarization of users, and that bots seem to be equally distributed among all communities, meaning they are inherent to the network.

Índex

1. Introducció 1
1.1 Context i justificació del treball 1
1.2 Objectius del Treball 4
1.3 Enfocament i mètode seguit 5
1.4 Planificació del Treball 6
1.5 Breu sumari de productes obtinguts 7
1.6 Breu descripció dels altres capítols de la memòria 8
2. Estat de l'art 9
3. Plantejament del projecte 13
4. Obtenció del conjunt de dades per l'estudi 14
4.1 Configuració de la llibreria tweepy 14
4.2 Estratègia de selecció de contingut 17
5. Catalogació de bots 21
5.1 Botometer 21
5.2 Nou enfoc i alternatives 21
5.3 Solució escollida 22
6. Models predictius supervisats 24
6.1 Normalització i estandardització de les dades 24
6.2 Proporció de categories 24
6.3 Selecció del millor model 25
7. Generació i visualització del graf obtingut 32
7.1 Generació del graf 32
7.2 Detecció de comunitats 32
7.3 Assignació d'etiquetes 33
7.4 Visualització del graf 33
7.5 Usuaris amb més interaccions 36
8. Conclusions 37
9. Glossari 38
10. Bibliografia 39
10. Agraïments 41

Llista de figures

Fig. 1 – Ús d'internet per la població mundial, 2017

Fig. 2 – Interaccions d'usuaris amb internet en un minut, 2019

Fig. 3 – Usuaris de xarxes socials, 2019

Fig. 4 – Planificació inicial del projecte

Fig. 5 – Exemple de model de geometric deep learning per catalogació de noticies

Fig. 6 – Creació d'una app de desenvolupament de Twitter

Fig. 7 – Creació d'una app de desenvolupament de Twitter

Fig. 8 – Credencials de l'app de desenvolupament de Twitter

Fig. 9 – Configuració de les credencials de l'API de Twitter en Python

Fig. 10 – Funció principal en la descàrrega de tweets via l'API de Twitter (Python)

Fig. 11 – Orquestració de les descàrregues de tweets via l'API de Twitter (Python)

Fig. 12 – Representació de la lògica d'orquestració de les descàrregues de tweets via l'API de Twitter (Python)

Fig. 13 – Exemple de dades generades a partir de les timelines d'usuari

Fig. 14 – Esquema de cross-validation amb factor cinc

Fig. 15 – Corba ROC de model LinearSVC obtingut

Fig. 16 – Esquema de capes i neurones per un MLP amb una sola capa oculta

Fig. 17 – Corba ROC de model MLP obtingut

Fig. 18 – Corba ROC de model Bernoulli Naive-Bayes obtingut

Fig. 19 – Corba ROC de model de Regressió Logística obtingut

Fig. 20 – Corba ROC de model de Random Forest obtingut

Fig. 21 – Visualització de comunitats del data set (Gephi)

Fig. 22 – Visualització de bots catalogats a partir del millor model de predicció (Gephi)

1. Introducció

1.1 Context i justificació del treball

Amb l'explosió tecnològica patida durant els darrers vint anys, el dia a dia de l'ésser humà està cada cop més caracteritzat per una connectivitat permanent a internet, comptant a data d'avui amb números que impressionen i no deixen de créixer amb el pas dels anys. Com mostra un estudi de *OurWorldInData*[1] datant del 2016, en aquell any un 46% de la població mundial havia utilitzat internet al menys una vegada en els tres mesos previs a l'estudi, sent les regions més desenvolupades aquelles amb una taxa d'ús d'internet més elevada (Amèrica del Nord – 78%, Europa – 73%). I aquests números només han fet que créixer en els darrers anys, tant per la millora de la infraestructura de comunicacions com per la millor accessibilitat a dispositius mòbils.

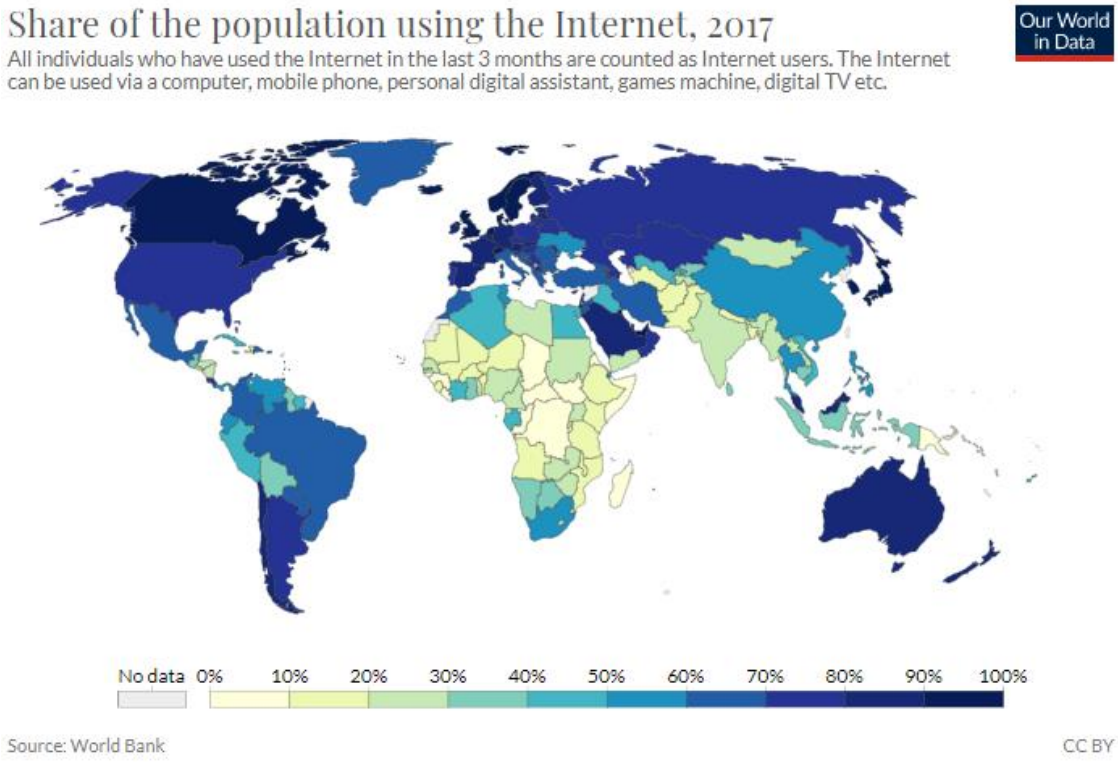

Fig. 1 – Ús d'internet per la població mundial, 2017

Internet ha esdevingut, de fet, la font d'informació més important de la nostra època, contenint més informació de la que podem processar i donant veu a tot i cadascun dels individus que hi participen. Com es pot veure en la infografia de l'empresa DOMO per la setena edició del seu estudi "*Data Never Sleeps*[2]", s'espera que al 2020 hi hagi quaranta vegades més bytes de dades que estrelles a l'univers observable. I la seva infografia, representant les interaccions que fan els usuaris amb internet en un minut, és digne d'estudi:

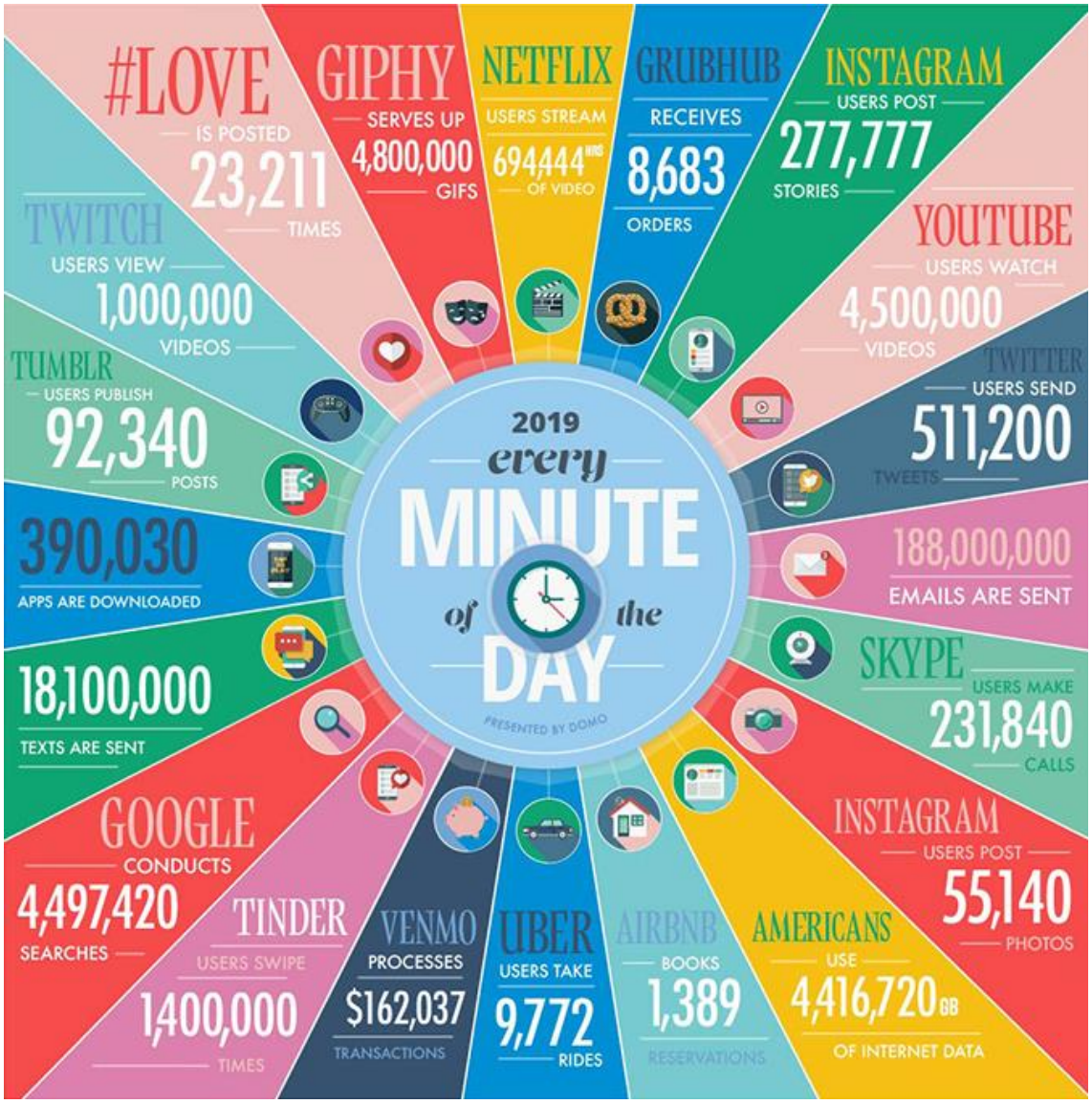

Fig. 2 – Interaccions d'usuaris amb internet en un minut, 2019

Les noves generacions ja han crescut en una cultura de total connectivitat, i veuen en els seus dispositius mòbils un aliat indispensable del seu dia a dia, un element del qual no es poden privar. Lluny queden les enciclopèdies físiques, les biblioteques i les convencions. En el món de la digitalització, quan un usuari necessita informació, no dubta en treure el telèfon mòbil i buscar, ja sigui en un motor de recerca, xarxes socials o fonts d'informació reputades, allò que necessita saber, podent resoldre els seus dubtes en temps real. I no només a l'hora de cercar informació els dispositius mòbils són el millor aliat: la versatilitat d'aquests aparells fa que l'usuari pugui cobrir qualsevol necessitat que tingui des del seu terminal: demanar menjar per emportar, cercar un pis de lloguer, demanar un Über, comprar entrades, etc...

I en aquest context de total connectivitat, les xarxes socials són un motor en l'intercanvi d'informació. Tot i haver estat concebudes inicialment com a eina de comunicació, han esdevingut cada cop més una nova plataforma de publicitat i màrqueting, utilitzada també per premsa i entitats governamentals per difondre informació d'interès general... i també per altres organismes com a font de

desinformació, o com s'anomena comunament, manipulació informàtica o manipulació mediàtica.

La noció de fake (o false) news, com defineixen a l'article "False News On Social Media: A Data-Driven Survey"[3] Francesco Pierri i Stefano Ceri, fa referència a una multitud de conceptes, que engloben principalment tota aquella informació incorrecta o enganyosa que es comparteix, sent el concepte "fake" el terme sovint associat a noticies polítiques. Una noticia falsa ("false new") no va obligatòriament associada a una voluntat d'enganyar el lector, ja que es contemplen possibles errors en la redacció o en el contrast de la veracitat de les noticies; en aquest cas, el concepte empleat es "misinformation". En el cas de la desinformació ("disinformation"), en canvi, la propagació de la noticia té un objectiu enganyós i manipulador.

És important constatar que bona part de les investigacions avaluen en el context actual l'impacte de les fake news en les xarxes socials, principalment per l'increment del consum de noticies per part dels usuaris en aquestes plataformes. Recuperant la infografia de DOMO, es poden observar els ordres de magnitud del volum d'interaccions entre usuaris en certes xarxes socials[2]: a Twitter, per exemple, es publiquen 511200 tweets en tan sols un minut, i a Instagram es publiquen 277777 stories en el mateix interval de temps. Es pot veure també com xarxes socials com Facebook (propietària de Whatsapp), Youtube i Instagram superen el miler de milions d'usuaris mensuals[1].

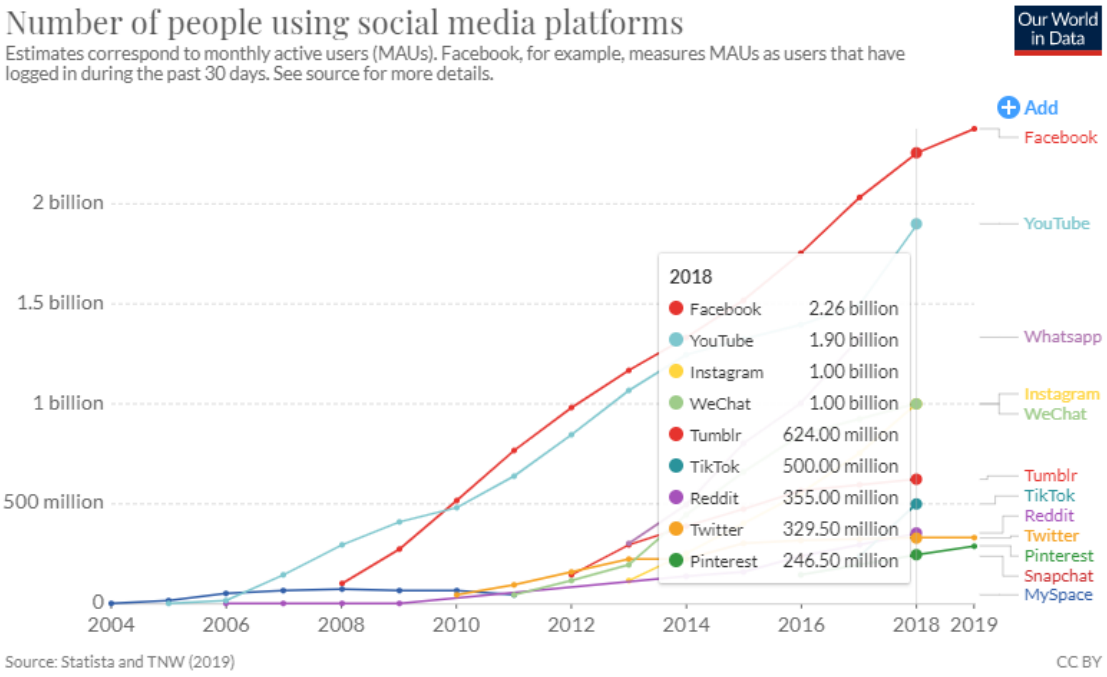

Fig. 3 – Usuaris de xarxes socials, 2019

Considerant que Twitter, per exemple, té 330 milions d'usuaris actius cada més, la informació que s'hi publica pot potencialment arribar a totes aquestes persones. I és aquest fenomen el que ha donat peu a un nou context de desinformació a gran escala. Les noticies falses que es publiquen en aquestes xarxes es propaguen amb molta

rapidesa, i com ja s'ha esmentat prèviament, tenen capacitat d'arribar a milers de milions d'usuaris, sent per tant susceptibles de condicionar l'opinió pública.

Un exemple molt clar i recent és la suposada contractació[4] per part del FC Barcelona d'una empresa que oferia serveis de desprestigi amb comptes falsos a Twitter, promovent noticies falses i dades no acurades per crear una corrent de pensament a favor de l'actual cúpula directiva, i manipulant així la percepció dels socis en temes d'actualitat. Tot i així, la influencia de les fake news en les eleccions presidencials dels Estats Units d'Amèrica durant l'any 2016, on una campanya de desinformació va aconseguir condicionar els resultats electorals dels Estats Units, es probablement el punt d'inflexió en l'estudi de les fake news, degut a la repercussió que aquest fet va tenir a nivell mundial.

Un altre aspecte important en l'estudi de les fake news són els factors de comportament humà que contribueixen a la proliferació de les noticies falses en aquestes xarxes socials. Es parla de dos comportaments principals: el biaix de confirmació (la tendència a recordar i afavorir la informació que confirma les pròpies creences o hipòtesis), i el realisme ingenu (creure que la gent que pensa diferent s'equivoca i la pròpia opinió és l'única correcta). Tots aquest factors participen en la creació de les anomenades "echo chambers", que realimenten la informació rebuda en grups de corrent de pensament similar i amb punts de vista molt polaritzats.

I davant d'aquest nou paradigma, es important educar els usuaris en el rigor i el contrast de la informació que es comparteix, per tal de promoure la informació no contrastada o esbiaixada/tendenciosa. De fet, algunes grans plataformes de difusió de contingut visual, com ara Netflix i HBO, entre d'altres, es comencen a fer ressò d'aquestes temàtiques, i participen en la promoció amb la publicació de documentals, com "After Truth: Disinformation and the Cost of Fake News" a HBO.

En aquesta línia, és important entendre primer com s'estructura aquest fenomen, i quins elements caracteritzen la propagació de noticies falses. ¿Quina es la font, o origen, de tota aquesta informació enganyosa? ¿Quin objectiu persegueixen aquests missatges? ¿Quines estratègies existeixen darrere de la propagació d'aquestes noticies?

1.2 Objectius del Treball

Com ja s'ha comentat ràpidament en l'apartat anterior, el principal objectiu d'aquest treball és entendre com sorgeix el fenomen de desinformació en xarxes socials (en principi limitat a la xarxa social de Twitter) i veure'n qui son els actors principals d'aquest fenomen, intentant principalment descriure l'estructura de la xarxa i les estratègies emprades en aquest nou context.

En aquest sentit, el primer objectiu del treball és obtenir un set de dades adequat per a l'estudi, a priori de la xarxa social Twitter, al voltant d'una temàtica que escaigui amb l'estudi que es vol realitzar. Idealment, es buscarà una temàtica d'interès global (la crisi del coronavirus és una bona candidata) i poder aconseguir d'aquesta manera un volum d'interaccions entre usuaris més elevat.

El segon objectiu que es vol aconseguir és, a partir d'aquest set de dades, poder descriure de manera visual les característiques del data set. Això passa per obtenir un modelat de la xarxa de usuaris que interactuen amb les seves publicacions i respostes, i veure quines característiques presenta, així com quines comunitats s'hi poden trobar, i de quines temàtiques son els missatges que s'intercanvien.

Un tercer objectiu del projecte serà, un cop modelada la xarxa d'usuaris, poder identificar si existeixen bots, i quina funció tenen en aquest ecosistema: Comparteixen informació de qualitat? Són partícips en la propagació de fake news? Quina proporció ocupen en la xarxa, i a qui pertanyen/quina és la seva natura?

Un darrer objectiu serà catalogar tots aquells usuaris que hagin compartit informació poc fiable o no contrastada, i veure visualment com aquesta nova xarxa de desinformació s'integra en la xarxa completa d'usaris, podent veure part del cicle de vida de les fake news. A poder ser, es destacarà també tots aquells usuaris que, per contra, fan accions per mitigar la propagació de noticies falses, desmentint o desacreditant les informacions que no siguin correctes.

1.3 Enfocament i mètode seguit

De cara donar resposta als objectius esmentats prèviament, es procedirà inicialment per la obtenció d'un data set de la millor qualitat possible de la xarxa social Twitter. Existeixen diferents alternatives que permeten realitzar aquesta primera tasca, tot i que la llibreria tweepy sembla la millor candidata. En aquest sentit, es vol disposar d'un data set reduït amb la informació indispensable per a l'anàlisi, que no distarà gaire dels atributs següents:

- Usuari que envia el tweet.
- Tipologia del tweet (retweet, reply, quote). En aquest cas, caldrà mantenir la informació de l'usuari que va fer el tweet original, al qual s'està responent, o al que s'està citant.
- El contingut textual no truncat del missatge, net de caràcters que embrutin el contingut (puntuació i hashtags, entre d'altres). Es conservaran les urls de manera preventiva, per intentar trobar
- La data dels, per analitzar la temporalitat de l'activitat.
- Qualsevol altre dada que sembli rellevant un cop entrat en detall dels camps disponibles en l'extracció de dades.

D'aquesta manera, doncs, i configurant els paràmetres disponibles per executar una extracció controlada de dades, es generarà un data set en llenguatge python per al posterior tractament. Un paràmetre útil per a l'extracció és el poder definir un hashtag concret. En el moment de l'extracció, es cercarà a Twitter quins son els hashtags més emprats en relació amb la temàtica escollida, i s'imposaran com a paràmetre per filtrar les dades extretes.

Un cop consolidat el data set, es vol modelar el graf obtingut de la xarxa subjacent mitjançant una eina de visualització, com per exemple Gephi. D'aquesta manera es podran posar en valor, de manera visual, les diferents característiques de la xarxa i

dels seus usuaris, i realitzar doncs un primer estudi descriptiu del graf, per extreure un primer conjunt de conclusions.

El següent pas de l'estudi serà, mitjançant llibreries de *Python*, poder identificar quines comunitats es discerneixen en el graf, i analitzar d'aquesta manera si hi ha particularitats entre les diferents comunitats, si hi ha algun patró comú, i sobre tot, si existeixen ecosistemes de bots i quines estratègies segueixen aquests bots. Es partirà de la premissa que un bot es tot aquell usuari que comparteix informació de manera molt activa i homogènia, ja que no respondrà doncs als patrons d'activitat d'un usuari real. De totes maneres, aquesta definició es revisarà arribat el moment, i ja s'assumeix que certs usuaris, tot i no ser realment bots, podran quedar catalogats com a tals, desvirtuant mínimament les conclusions.

Finalment, es reflectirà tota aquesta informació en un nou conjunt de visualitzacions en Gephi, on es pugui comprovar l'impacte de les *fake news* en el graf estudiat. Aquestes visualitzacions s'acompanyaran idealment d'un *dashboard* interactiu, on apareixeran característiques descriptives del data set, i potencialment gràfiques temporals amb l'evolució del volum de *tweets* relacionats amb la temàtica.

1.4 Planificació del Treball

Per a la realització del projecte és indispensable disposar d'un portàtil amb una versió actualitzada de *Python* i *Jupyter Notebooks* (a priori en *Anaconda*), i memòria *RAM* suficient per a que el tractament de grans volums de dades no col·lapsi el sistema. Dit això, també és necessari comptar amb un compte de *developer* a *Twitter* de cara a poder fer crides a la *API* i descarregar d'aquesta manera les dades. Aquesta configuració estarà detallada a l'apartat dedicat d'aquesta memòria. Per últim, caldrà disposar de la instal·lació de *Gephi* en el portàtil de treball.

De mateixa manera, serà important disposar d'una planificació de les extraccions de dades a realitzar, així com realitzar un estudi de mercat i estat de l'art per a millorar el producte final o inspirar-se'n i analitzar possibles noves interpretacions de les dades/nous enfocs.

A continuació es descriu detalladament una estimació de les tasques previstes per al projecte:

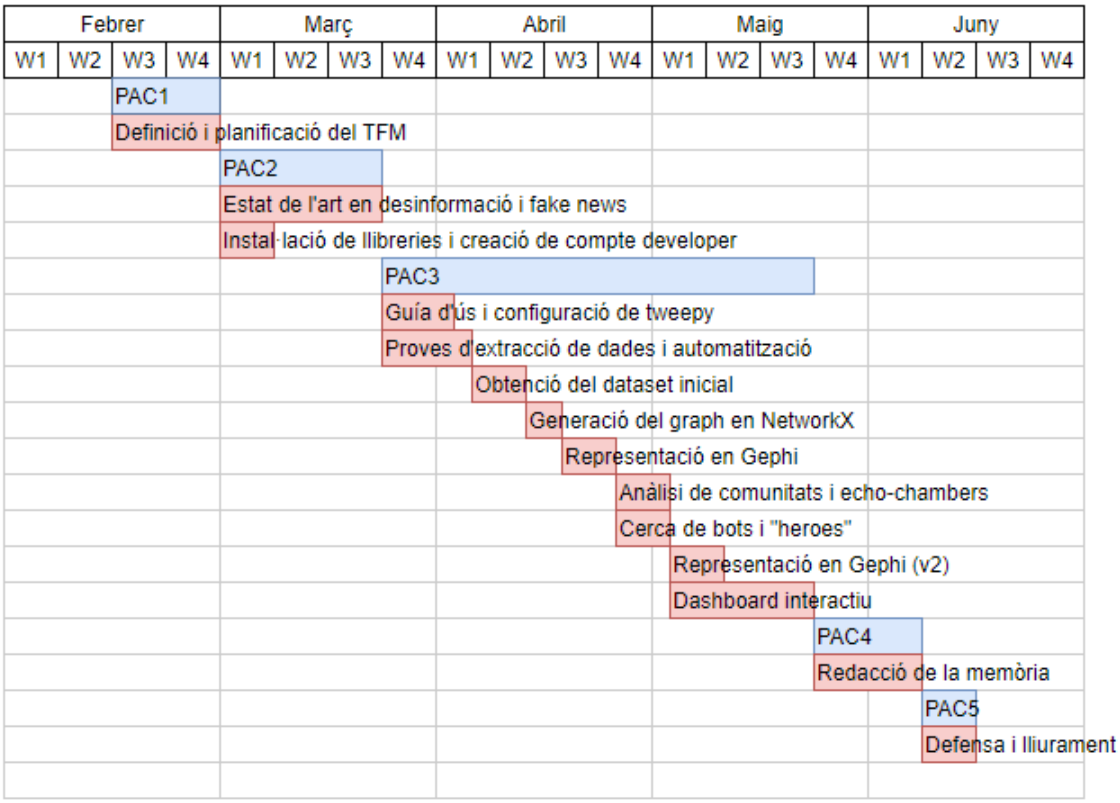

Fig. 4 – Planificació inicial del projecte

1.5 Breu sumari de productes obtinguts

De cara a establir les conclusions del projecte, és indispensable obtenir els següents productes:

- Una guia d'instal·lació i ús de la llibreria *tweepy*, emprada per a la recol·lecció de *tweets*.
- Un data set amb els *tweets* que seran la base d'estudi del projecte, en format adequat i amb les mínimes dades necessàries.
- El codi en *Python* amb el tractament de les dades.
- Una representació en *Gephi* del graf d'interaccions en *NetworkX* amb les seves característiques principals, i configurat per fer aparèixer, visualment, tota la informació rellevant.
- Un estudi de les comunitats detectades en el mateix graf, per detectar eventuals *echo-chambers*, i les seves particularitats (ideologia, orientació política,...)
- Una catalogació dels *tweets* potencialment poc fiables, amb la seva representació en *Gephi* sobre la xarxa complerta (per ressaltar els usuaris que tendeixen a compartir *fake news*).
- Una catalogació dels *tweets* que intenten mitigar les informacions poc fiables i donar veracitat a aquelles informacions que si ho són.
- Un automatització de dades que es pugi visualitzar en un *dashboard*.
- Una memòria del projecte.

1.6 Breu descripció dels altres capítols de la memòria

El contingut del projecte s'ha estructurat en diversos apartats, que tenen per objectiu explicar tots els aspectes imprescindibles del projecte amb la major claredat possible, justificant les decisions preses i les alternatives disponibles, però també explicant raonadament totes les implementacions i productes obtinguts.

En un primer apartat, però, s'analitza el context del projecte arran de la importància del control de *fake news* en la societat contemporània, i totes les noves vies d'investigació que han sorgit arran de l'evolució de les xarxes socials i la seva contaminació en termes de desinformació.

Seguidament, s'estudia el plantejament del projecte, deixant clar quin serà el desenvolupament del projecte, i analitzant els eventuals canvis establerts en l'enfoc inicial, així com la seva justificació.

Entrant en aspectes probablement més tècnics, es tracta posteriorment l'obtenció del conjunt de dades per l'estudi, amb la metodologia i problemàtiques que s'han pogut apreciar en el procés, per entrar de ple en la part central del projecte: l'ús de models predictius supervisats per la catalogació d'usuaris de *Twitter* en funció de la seva activitat, establint dues categories (bots i éssers humans).

A partir d'aquest punt, es plantegen diferents alternatives d'implementació així com les inherents problemàtiques, per finalment realitzar la implementació d'aquestes models i la comparació entre ells de cara a buscar aquell que millors resultats ofereix.

En els darrers apartats s'aprofita aquest model predictiu per etiquetar els usuaris del conjunt de dades estudiat, tot generant un graf que es visualitzarà mitjançant *Gephi*, en el qual es podran apreciar també les comunitats obtingudes a partir d'aquest conjunt d'usuaris.

Un darrer apartat de conclusions durà per objectiu fer un repàs de totes les observacions rellevants realitzades durant el procés, de cara a consolidar el coneixement obtingut i donar peu a noves preguntes per resoldre.

2. Estat de l'art

Les eleccions a la presidència dels Estats Units d'Amèrica de l'any 2016 i la decisió del *Brexit* per part del poble anglès van marcar un abans i un després en l'estudi de les xarxes socials com a eina d'influència de masses, en quant encara a dia d'avui no està del tot clar fins a quin punt va ser determinant l'activitat de milions d'usuaris a *Twitter* en els mesos previs, que tot sembla indicar va condicionar el vot de milions de persones. En el cas de les eleccions americanes, certs estudis[12] indiquen que en els mesos previs a les eleccions, un 25% dels missatges compartits a *Twitter* als Estats Units contenien informacions falses o extremadament esbiaixades, i que en particular els promotors d'aquests missatges eren seguidors o d'ideologia propera al partit de dretes (Donald Trump). Arran d'aquest fet, la gestió de les *fake news* i la desinformació en les xarxes socials ha esdevingut un objecte d'estudi molt interessant i rellevant en el context actual.

Un bon punt de partida en l'anàlisi l'estat de l'art de l'estudi de la propagació de desinformació en xarxes socials passa per analitzar l'estudi de Francesco Pierri i Stefano Ceri sobre *"False News On Social Media: A Data-Driven Survey"*[3]. Aquest estudi, un cop fet un repàs dels conceptes més rellevants en temàtica de *fake news* i *misinformation* (analitzats breument en l'apartat 1.1), analitza les línies d'investigació d'estudis relacionats amb la detecció de *fake news* realitzats durant els anys 2017 i 2018. Aquesta limitació temporal ve marcada, probablement, per la gran repercussió de les eleccions polítiques dels Estats Units al 2016, i com es comenta a l'article, per la limitació tecnològica en la recol·lecció de les dades de xarxes socials.

Segons s'esmenta a l'article, els estudis en matèria de propagació de desinformació en xarxes socials es poden classificar en tres àmbits d'investigació, que serien la detecció de noticies falses, la caracterització de la desinformació que s'hi propaga, i les tècniques de mitigació d'aquestes noticies.

A nivell de detecció de *fake news* en xarxes socials, es poden categoritzar els estudis segons el seu focus, ja sigui en el contingut de la noticia (el text que hi apareix) o bé en el context de la noticia (quins usuaris la comparteixen, quines interaccions hi ha), havent finalment una tercera variant híbrida que contempla tant la versant "contingut" com la versant "context".

a. En termes generals, la detecció de *fake news* en funció del contingut s'ha treballat extensivament tant amb tècniques de *Machine learning* (*Support Vector Machines*) com amb tècniques de *deep learning*, emprant principalment regressions logístiques, xarxes neuronals convolucionals i xarxes neuronals recurrents. En la gran majoria dels casos, es tracta de models supervisats entrenats amb un data set generat manualment.

b. En el cas de la detecció de *fake news* per el context de la interacció, l'estudi gira en torn a les interaccions d'usuaris, com els *likes*, *comments* i *re(tweets)* de *Twitter*. En un cas concret[5], s'analitza la propagació de missatges de contingut maliciós, amb un data set de *tweets* amb noticies catalogades com a

falses o certes, i un classificador (amb xarxes *Long Short-Term Memory*) que analitza el recorregut dels missatges sobre un model de baixa dimensionalitat del graf de la xarxa social.

c. Per la combinació de detecció per contingut i context, les principals conclusions obtingudes amb els diferents estudis afirmen que la incorporació del context en models basats en contingut millora la detecció de noticies falses.

Per la part de caracterització dels models de difusió de les *fake news*, un estudi molt interessant[6] aparegut a la revista *Science* al març de 2018 analitza la difusió de noticies reals (verificades) i falses distribuïdes a *Twitter* entre 2006 i 2017. Per a donar validesa a l'estudi, la classificació de les noticies es contrasta amb sis organitzacions de "*fact-checking*", expertes en la detecció de *fake news* (e.g. Miniver.org a Espanya). Els resultats obtinguts demostren que les noticies falses es propaguen de manera més ràpida, profunda i extensa que les noticies verídiques (en totes les categories d'informació), i que, a més a més, aquestes noticies generen més repercussió sempre que tracten temàtiques de política, terrorisme, desastres naturals, ciència, finances i llegendes urbanes. D'altra banda, es conclou en aquest estudi que la novetat de les noticies fa que aquestes es propaguin més ràpid, i que les noticies falses solen ser sempre més recents, d'aquí la seva ràpida expansió. Tanmateix, sembla ser que la presència de bots propaga noticies independentment de la certesa del contingut, i que son principalment els humans els que fomenten la propagació de les noticies falses.

Finalment, han sorgit al llarg dels anys diferents iniciatives d'intervenció[3] a nivell de mitigació de "*fake news*", distribuïdes en tres versants:

- per una banda, la detecció de bots en funció de l'activitat en les xarxes socials, podent determinar quins usuaris son més susceptibles de ser en realitat bots;
- per una altra banda, la selecció òptima de noticies que es contrastarà amb organitzacions de "*fact-checking*" per a la seva validació, i prevenint així la propagació d'informació incorrecta.
- Finalment, la protecció d'usuaris detectats com a "defensors" de la veritat, que promouen fets contrastats per rebatre noticies falses.

En aquesta línia, diverses xarxes socials, com *Facebook* o *Twitter*, s'encarreguen de posar a disposició dels seus usuaris eines per combatre la desinformació, apel·lant a la saviesa de la multitud. Tanmateix, aquestes plataformes comencen a prendre la iniciativa[8] i censurar aquell contingut que ha estat desacreditat per fonts fiables d'informació, limitant així l'exposició de la resta d'usuaris a dades no fiables.

Com s'ha esmentat anteriorment, la major part d'estudis que analitzen la detecció de les *fake news* s'han enfocat històricament en l'estudi del contingut dels missatges de xarxes socials, per tal de detectar informació falsa o esbiaixada. Les estratègies basades en contingut presenten un problema principal, i és que la interpretació de les noticies requereix de coneixement del context social i de sentit comú, capacitats que les tècniques de NLP actuals no han acabat de dominar.

En aquest sentit, estudis recents han demostrat que la propagació de les noticies segueix diferents patrons de propagació en xarxes socials en funció de la veracitat de

la informació, donant peu als estudis basats en propagació. Així doncs, es comencen a proposar models supervisats[13] sobre data sets amb noticies catalogades com a certes o falses mitjançant tècniques de *geometric deep learning*. En el model estudiat, s'utilitza una arquitectura de xarxa neuronal convolucional de quatre capes, amb dues capes convolucionals i dues capes completament connectades.

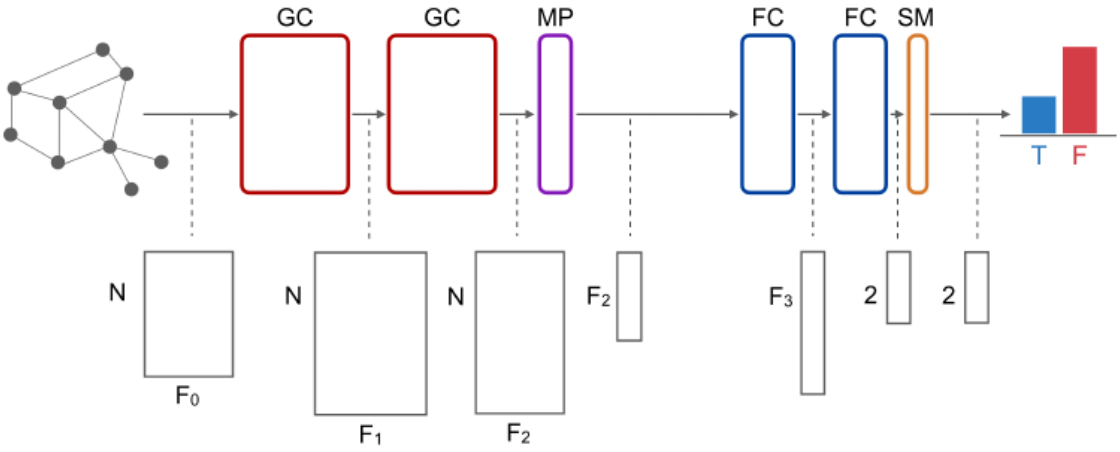

Fig. 5 – Exemple de model de geometric deep learning per catalogació de noticies

D'aquesta manera, la idea és analitzar diferents *tweets* amb enllaços a la mateixa noticia (una *url*), i detectar quins *tweets* la comparteixen, i en quin moment es publiquen, podent detectar així quina propagació ha tingut aquesta *url* a la xarxa social. Una problemàtica derivada d'aquests estudis és entendre si el temps de propagació afecta a la classificació de les noticies, necessitant doncs de diferents models entrenats considerant un temps màxim de vida de la noticia. Aquests estudis segueixen en investigació encara avui, degut a la gran varietat de noves vies d'investigació que sorgeixen arran dels resultats obtinguts.

Altres vies d'investigació, en canvi, fomenten l'ús de models no supervisats, en contrapartida de la gran majoria d'estudis de detecció de *fake news*. En el cas següent[14], i comptant que no es disposa d'informació sobre la veracitat del contingut, es pretén aprofitar l'*engagement* de l'usuari (la seva participació en les xarxes) com a factor d'estudi, atorgant una credibilitat a l'usuari i a les informacions que comparteix.

D'altra banda, val la pena remarcar les iniciatives[7] que es proposen a nivell internacional en forma de competicions, que fomenten i potencien la investigació al voltant de temàtiques com l'estudi de la desinformació en els mitjans de comunicació o la identificació de comportaments en xarxes socials. És el cas de l'ICWSM 2020 DATA CHALLENGE que aquest any proposa un estudi sobre seguretat amb les temàtiques esmentades anteriorment. En ambdós casos, es proporciona un data set recollit i validat especialment per al concurs, i que ha de servir com punt de partida per a que el participant posi a prova diferents estratègies per extreure coneixement: ¿Quines son les tàctiques dels productors de noticies falses, desinformació i propaganda?¿Canvien les noticies falses amb el temps? ¿Es poden dissenyar millors algorismes de *Machine learning* per a detectar desinformació?

Queda finalment posar en valor les iniciatives d'I+D de fundacions privades i altres òrgans (universitats, centres d'investigació, entitats governamentals,...) que pretenen donar una visió més completa i acurada sobre certs fenòmens, basant-se purament en les dades. L'equip de CoMuNe Lab, FBK, en col·laboració amb un professor d'Economia cultural de Harvard i la universitat de Milano, ha preparat un *dashboard* interactiu[9] enfocat a donar resposta, a partir de tècniques de *Machine learning* aplicades a missatges públics de *Twitter*, a tres preguntes: Quin és el sentiment col·lectiu de la població, quina és la pol·lució actual de bots socials a *Twitter* i quina és la fiabilitat dels missatges.

Es important destacar que l'estudi del comportament de bots en xarxes socials ha agafat molta importància en els darrers anys, no només a causa de la evident proliferació del seu ús, sinó també arran de diferents estudis[10] que demostren que l'activitat de bots en les xarxes socials incrementa l'exposició dels usuaris a contingut negatiu i inflamador. En aquest sentit, s'han analitzat en aquest cas milions de tweets per tal de detectar aquells usuaris bots mitjançant un *framework* de *honeypots*[10,11]. Aquesta estratègia consisteix en construir comptes de *Twitter* (bots) dissenyats per no interferir amb usuaris legitimats (només interactuen entre ells i publiquen missatges aleatoris), que s'encarreguen de monitoritzar l'activitat d'usuaris que entren en contacte amb ells (responen, fan *retweet* o mencionen) per iniciativa pròpia. Per mitigar aquests efectes, val a dir que el propi *Twitter* disposa de protocols de detecció i eliminació de bots, i que molts comptes contaminants tenen una durada de vida curta. Tanmateix, *Twitter* disposa de llistats d'usuaris que vulneren els seus termes de servei, i que detecta com a *spammers*.

3. Plantejament del projecte

Tot i que un enfocament inicial del projecte s'orientava cap a l'estudi de la propagació de *fake news* i *misinformation* en la xarxa, es va decidir finalment, degut a la falta de dades sobre la natura dels usuaris, a realitzar una comparativa de models predictius supervisats, capaços de catalogar si els usuaris de un conjunt de dades eren éssers humans o bots.

En aquest sentit, el contingut del projecte s'ha estructurat en certs objectius necessaris per a la consecució d'aquest model:

a. Obtenir en primer lloc les credencials necessàries per a extreure dades de *Twitter*, amb les quals descarregar un conjunt de dades sobre un tema que generi controvèrsia.

b. Obtenir un conjunt de dades d'entrenament amb la catalogació d'usuaris (bots i humans) que serveixi de partida per a establir diversos models predictius.

c. Obtenir el *timeline* dels usuaris del conjunt de dades d'entrenament, per tal de conèixer la seva activitat més recent a *Twitter* i definir així el conjunt complert de dades d'entrenament.

d. Implementar diversos models de classificació supervisada, podent establir models predictius, dels qual s'avaluaran la fiabilitat i la precisió, de cara a obtenir el millor model.

e. Utilitzar aquest model obtingut per tal de classificar la resta d'usuaris del conjunt de dades original, obtenint un etiquetat d'usuaris bots i no bots.

f. Implementar un algoritme de detecció de comunitats amb *Girvan Newmann* i *Louvain,* per tal de detectar conjunts d'usuaris.

g. Representar visualment les dades sobre un graf, per tal de raonar sobre les conclusions que s'extreuen de l'estudi.

4. Obtenció del conjunt de dades per l'estudi

Un cop decidits el passos a seguir, i sent conscients del proper objectiu (aconseguir un data set amb el conjunt de *tweets* que es vulguin considerar per a l'estudi), caldrà configurar el codi de l'aplicació per tal de poder connectar-se amb *l'API* de *Twitter*, i descarregar, de manera seqüencial, aquest conjunt de *tweets*.

En aquest sentit, existeix una llibreria anomenada *tweepy* que facilita substancialment la comunicació amb l'API de *Twitter*. Es tracta de la llibreria de referència en tots els estudis amb dades de *Twitter*, i tot i que no s'entrarà en detall del funcionament d'aquesta llibreria ara, a mesura que sigui rellevant es comentaran els punts que convingui. Com a referència, la documentació existent (c.f. http://docs.tweepy.org/en/latest/) és molt detallada i minuciosa, i per tant pot servir per complementar tota aquella informació que no aparegui explícitament en aquest document.

4.1 Configuració de la llibreria tweepy

De manera indispensable, i per tal de poder treballar amb aquesta llibreria, es necessita disposar d'unes credencials d'accés a la *API* de *Twitter*; aquest accés permetrà autenticar totes les peticions i crides que es realitzin. Per tal d'obtenir aquestes credencials, és necessari disposar d'un compte personal de *Twitter*, amb el qual es sol·licitarà un accés a la secció https://developer.twitter.com/en.

El funcionament d'aquesta secció *developper* es basa en sol·licitar la creació d'una *app*, i en aquesta petició caldrà indicar l'objectiu d'aquesta *app* i quines interaccions s'espera realitzar, de cara a que *Twitter* n'aprovi o en denegui el seu ús. Aquest mecanisme de control permet impedir la creació massiva de bots, així com assegurar l'autenticació dels usuaris, ja que cada *app* creada es vincula amb un compte de *Twitter* real.

Un cop sol·licitada la creació de la app, i aprovada per els mecanismes encarregats de Twitter, es podrà finalment accedir doncs a les noves credencials generades.

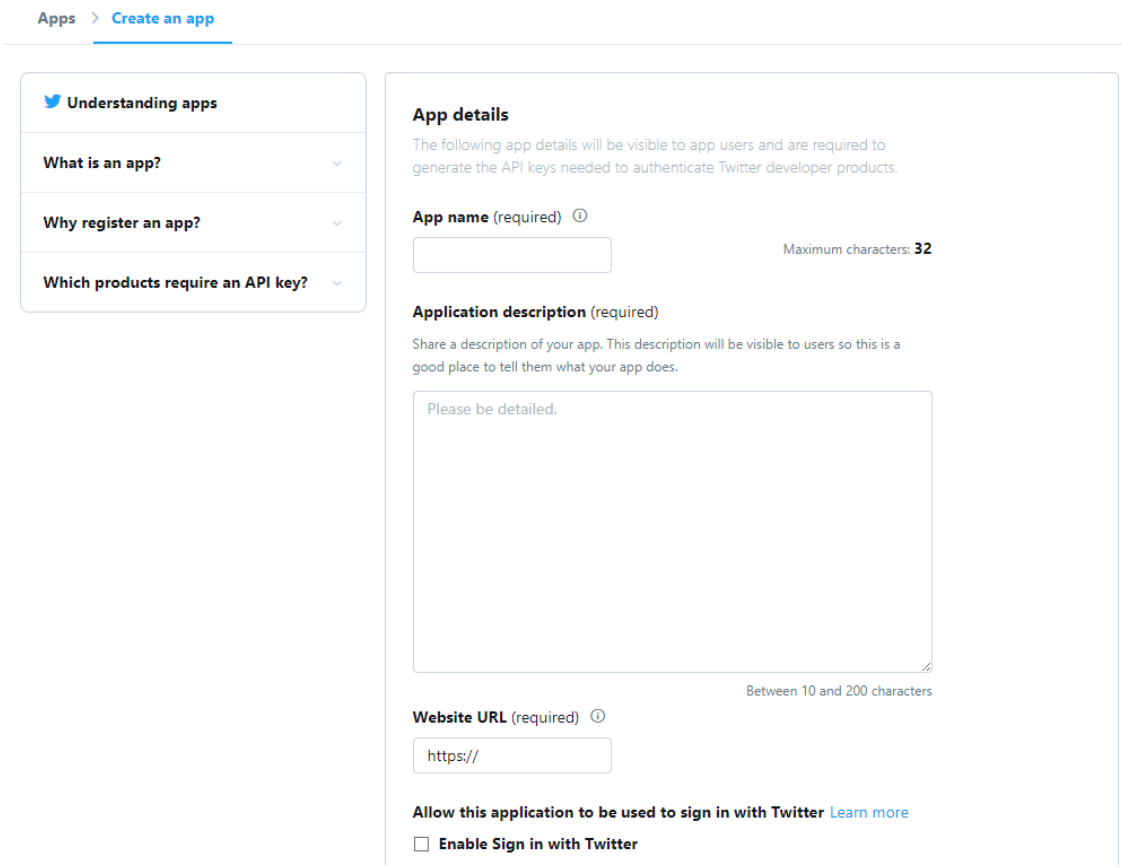

Fig. 6 – Creació d'una app de desenvolupament de Twitter

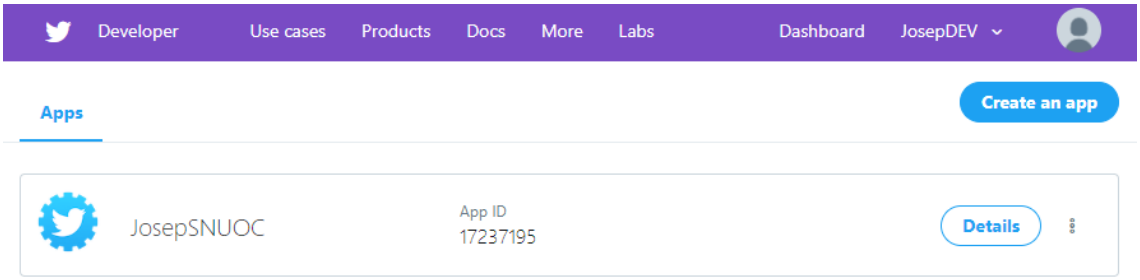

Fig. 7 – Creació d'una app de desenvolupament de Twitter

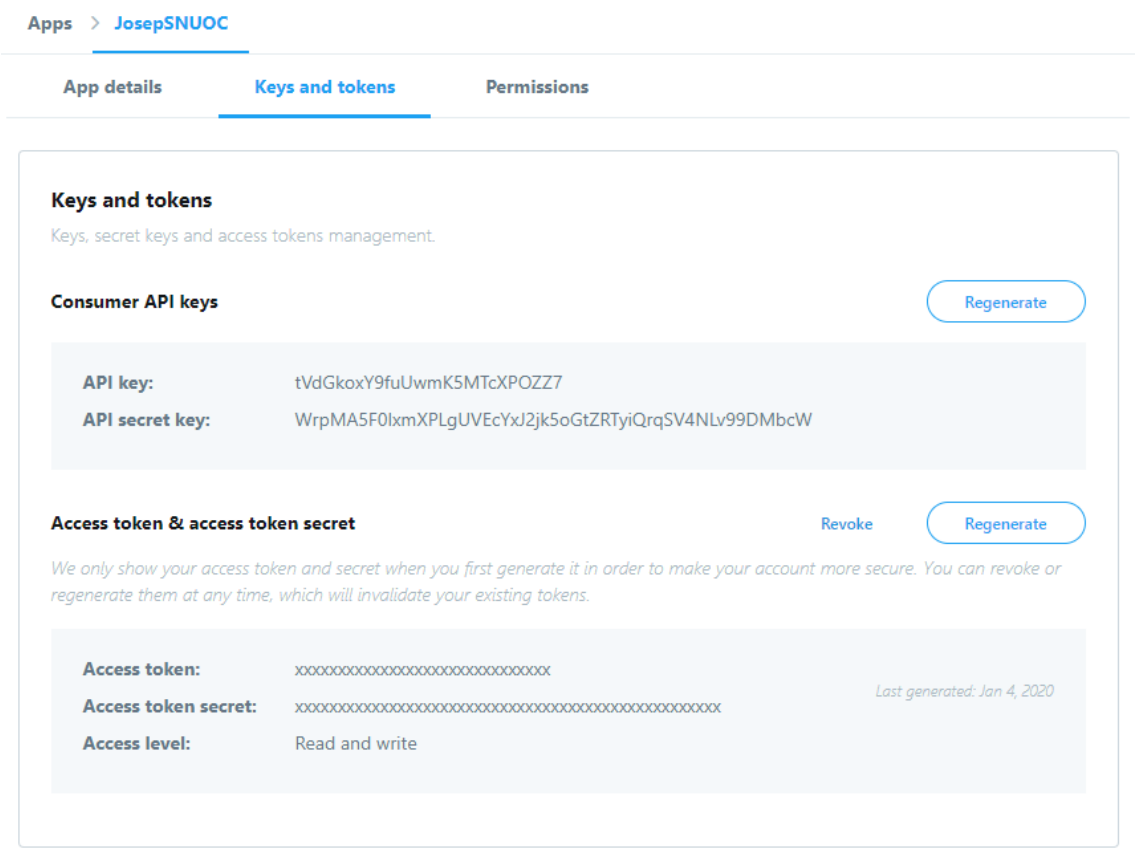

Fig. 8 – Credencials de l'app de desenvolupament de Twitter

Havent obtingut aquestes credencials, simplement quedarà instal·lar la llibreria *tweepy* (segons s'indica a la documentació) i configurar aquestes credencials en el nostre codi. Un altre punt fonamental serà la declaració de l'objecte *tweepy.API* a partir de les credencials generades, que servirà de base per a obtenir les dades de *Twitter*.

```
# Credencials
consumer_key = 'tVdGkoxY9fuUwmK5MTcXPOZZ7'
consumer_secret = 'WrpMA5F0IxmXPLgUVEcYxJ2jk5oGtZRTyiQrqSV4NLv99DMbcW'
access_token = '990678531230306305-2cLLqzIKkVL5UDQv8cgYivdIVquC7QY'
access_token_secret = 'UG37IoTOyhAFIwyealH2KDvsWN5TwcHzL8SEybLvOCBU1'

# Creació d'un objecte api
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth,wait_on_rate_limit=True)
```

Fig. 9 – Configuració de les credencials de l'API de Twitter en Python

La llibreria *tweepy* ofereix diferents procediments per tal de descarregar *tweets*, entre els quals està la cerca a partir d'una cadena de caràcters donada; en aquest sentit, es poden obtenir tots els *tweets* que continguin un conjunt de paraules donats (amb un màxim de set dies d'antiguitat, per limitacions de la pròpia *API* de *Twitter*). Aquesta característica serà molt útil per delimitar el conjunt de missatges analitzats, ja que és imprescindible que sigui un tema d'actualitat, que generi controvèrsia, i que tingui molta activitat a *Twitter*.

4.2 Estratègia de selecció de contingut

Tenint present la crisi mundial del COVID-19 i la controvèrsia generada en el passat amb les eleccions als Estats Units (concretament amb el cas *Russiagate*), es va considerar que una combinació de la paraula "covid" amb altres paraules ("usa", "trump", vaccine", ...) podia ser un bon referent per a obtenir un conjunt de dades interessant (sempre filtrant per idioma, anglès en el cas del projecte). De totes maneres, i un cop realitzades diferents proves, es va concloure que :

- La cerca de hashtags concrets no oferia molts bons resultats, ja que segmentava molt les dades disponibles, i hagués necessitat de diversos processos d'extracció paral·lels per tal d'extreure les dades de tots els hashtags escollits.
- Una combinació de varies paraules, com per exemple "vaccine" i "covid", donava millors resultats, tot i que el volum de dades generat era molt variable en funció de les paraules escollides.
- Com és comprensible, diferents conjunts de dades donaven com a resultat diferents interaccions entre usuaris, amb representacions del graf més o menys polaritzades.
- Certes combinacions de paraules eren molt volàtils, amb un fort impacte durant uns dies i una baixada radical de la seva presència en els tweets. En aquest sentit, la combinació de paraules escollida finalment podria donar resultats clarament diferents en el moment de la lectura d'aquesta memòria.
- Una cerca utilitzant senzillament la paraula "covid" generava un volum no gestionable de dades, degut a les limitacions tecnològiques del projecte, com són la CPU i RAM de l'equip, i la limitació de peticions per dia de les diferents *APIs* emprades.

Arran d'aquestes observacions, es va procedir a descarregar tres conjunts de dades diferents, intentant seleccionar les paraules que poguessin donar peu a una bona presència de *fake news* i amb bots com a part dels usuaris. Els tres *data sets* generats van ser els següents:

- "covid" + "vaccine": En el moment de l'extracció de dades, un tema principal de notícies era la investigació de vacunes per combatre la infecció per covid-19. La incertesa en la tipologia del virus, els mecanismes per combatre'l, la col·laboració internacional i les primeres proves en animals de vacunes van generar una intensa activitat a *Twitter*. Aquest *data set* va acabar contenint al voltant de dos-cents mil *tweets*, tot i que la representació gràfica de la seva xarxa (com es comentarà més endavant) no estava gaire polaritzada, i es va descartar per un exemple més polaritzat.
- "covid" + "trump": En aquest cas, i sent el president dels EUA un dels grans promotors de les paraules "fake news" (sense oblidar el polèmic passat amb l'ús de comptes de Twitter per condicionar el vot americà), es va voler obtenir una mostra al voltant d'un context polític, en el qual es podria veure la polarització dels usuaris. La immensa quantitat de tweets generats amb

aquesta cerca va permetre obtenir al voltant de sis-cents mil *tweets*, amb una interessant distribució dels usuaris.

- "#covidusa": En el moment de l'extracció, aquest *hashtag* semblava un bon candidat ja que el volum de *tweets* generats era interessant, tot i no seguir les premisses definides prèviament (els *hashtags* generen data sets atomitzats i poc relacionats, i no interessava cimentar l'estudi en diversos data sets obtinguts mitjançant cerca de *hashtags*). De totes maneres, el volum de *tweets* generats diàriament es va desinflar ràpidament, i es va descartar finalment utilitzar aquest data set com a part de l'estudi.

Un cop definits els criteris de cerca per la generació del conjunt de dades, caldrà analitzar el funcionament de la llibreria *tweepy* per entendre el codi implementat en el projecte. La funció mostrada a continuació servirà com a punt de partida per a l'anàlisi del procediment d'extracció, donat que és una implementació clau del projecte:

```
def get_covid_tweets(sinceid, maxid, loops, num, query):
    tweets = []
    run = 0
    length_diff = 1
    prev_length = 0

    while len(tweets) < 15000 and length_diff != 0 and run < loops:
        for tweet in tweepy.Cursor(api.search, q=query, tweet_mode = 'extended',
                                   since_id = sinceid, max_id = maxid, lang = 'en').items(num):
            tweets.append(tweet._json)
        run = run + 1
        length_diff = len(tweets)-prev_length
        prev_length = len(tweets)
        maxid = tweets[-1]['id']-1
        print(prev_length, maxid)

    return tweets
```

Fig. 10 – Funció principal en la descàrrega de tweets via l'API de Twitter (Python)

Aquesta funció permet, a partir d'uns *ids* de *tweets* màxim i mínim donats, descarregar tots els *tweets* amb *id* inferior al *id* màxim i superior al *id* mínim (ambdós inclosos), que contenen una combinació de paraules donada ("query") i que estan escrits en anglès (lang = 'en'). Sense gaire misteri, la idea radica en recórrer un cursor de *tweets*, i realitzar un append en una array (inicialitzada nul·la) del camp *.json*, que contindrà tota la informació que volem exportar en un format estructurat (JSON).

Com a comentari addicional, l'API de *Twitter* proporciona per defecte el cos del missatge d'un *tweet* fins a una certa longitud de cadena (140 caràcters, longitud original d'un *tweet*), tot i que recentment es va actualitzar l'API per tal d'incloure el nou format (fins a 280 caràcters) especificant el paràmetre *tweet_mode* amb el valor '*extended*'.

D'altra banda, i per tal de limitar l'impacte que podria ocasionar que l'API retornés una excepció, o que el procés no acabés un cop exportat tot el conjunt de *tweets* disponibles, s'han implementat uns mecanismes de control que limiten el volum de dades que intentem exportar. Aquests mecanismes consisteixen en un màxim d'iteracions (el paràmetre *loops* indica quantes iteracions volem realitzar), crides per blocs (el paràmetre "*num*" indica el nombre de *tweets* exportats a cada iteració), i un control dinàmic de la longitud del data set que s'està generant. Així doncs, si en dues

iteracions es detecta que la longitud del data set no ha canviat, la funció no intentarà descarregar cap *tweet* més, i retornarà el conjunt de *tweets* obtingut fins aleshores.

És important comentar també que l'API de Twitter sempre retorna els tweets en ordre decreixent del seu id, i que per tant s'exporten sempre els més recents primer, amb una persistència de la dada d'una setmana (no es poden obtenir dades més enllà).

Un cop esclarits tots aquests punts, es farà més entenedora la interpretació de l'orquestració de les descàrregues, explicada a continuació:

```
# Inicialització de variables
sinceid = 1258031757036642311
filename = ''
loops = 150
num = 100
tweets = []
today = date.today()
ymd = today.strftime("%Y%m%d")
query = "covid-19 trump"
filepath = str(r'C:\Users\josepconsuegra\Desktop\TFM_extract\covid_trump')

# Obtenció del maxid
for tweet in tweepy.Cursor(api.search, q = query, tweet_mode = 'extended', lang = 'en').items(1):
    tweets.append(tweet._json)
maxid = tweets[0]['id']

# Loop per sinceid < maxid
run = 0
while sinceid < maxid :
    tweets = []
    run = run + 1
    tweets = get_covid_tweets(sinceid, maxid, loops, num, query)
    dt = datetime.now()
    print('maxid = ' + str(maxid) + ', minid = ' + str(tweets[-1]['id']) + ', run nº' +
          str(run) + ', time: '+ str(dt))
    filename = os.path.join(filepath, 'tweets-covid-trump-'+str(ymd)+'-'+str(tweets[0]['id'])
                            +'-'+str(tweets[-1]['id'])+'.json')
    with open(filename, 'w', encoding='utf-8') as f:
        json.dump(tweets, f, ensure_ascii=False)
    maxid = tweets[-1]['id']-1
```

Fig. 11 – Orquestració de les descàrregues de tweets via l'API de Twitter (Python)

Aquesta part del codi es composa de tres blocs ben diferenciats:

- En el primer, definim simplement totes les variables necessàries, i inicialitzem com a nul·les totes les variables que informarem més endavant, com *filename* i *tweets*:
    - *sinceid*: *id* del *tweet* més recent, a partir del qual es vol començar l'exportació.
    - *loops*: nombre d'iteracions màximes que voldrem realitzar en la generació de cada fitxer.
    - *num*: la quantitat de *tweets* que volem exportar en cada trucada a la API.
    - *today+ymd*: genera una referència de data per a anomenar el fitxer de *backup*.
    - *query*: paraules clau que voldrem utilitzar en la cerca de *tweets*, i que es proveiran a la API.
    - *filepath*: ruta on es guardaran els fitxers de *backup* a mesura que es descarreguin els blocs de *tweets*.

- La segona part del codi serveix simplement per obtenir una referència de l'*id* de *tweet* màxim que es considerarà en el bucle. D'aquesta manera, qualsevol nou tweet que entri no afectarà a l'extracció, i es necessitarà per tant d'una extracció diferent per considerar aquests ids superiors a l'id màxim definit.
- La tercera part del codi correspon a l'orquestració de les descarregues, que es basaran en un bucle que iterarà sobre un valor fix de sinceid i un valor dinàmic de maxid.

Per tal de generar fitxers de backup gestionables, s'ha imposat un límit de quinze mil tweets per cada exportació, sent per tant possible que a partir del valor fix de sinceid i el valor inicial de maxid hi hagi més de quinze mil tweets per descarregar. En aquest cas, i donat que l'exportació es realitza per ordre decreixent d'id, caldrà que es reajustin els valors de la variable maxid després de cada iteració, mantenint constant el valor original de sinceid. A títol d'exemple, si es volen exportar els tweets amb ids entre maxid=100000 i sinceid=0, caldrà que, un cop acabada la primera iteració del bucle, s'actualitzi el valor de maxid per 85000.

Tanmateix, els ids de tweets són compartits entre tota la base de dades de Twitter, i per tant no segueixen una progressió unitària (entre dos tweets consecutius retornats a la consulta es poden haver generat milers). Es necessari doncs assignar a la variable maxid el valor de l'id del darrer tweet exportat, i restar-li una unitat. Val a dir que, un cop finalitzada una primera exportació, i de cara a obtenir els tweets de dies posteriors, és imprescindible fixar el valor de sinceid amb l'id màxim (més una unitat) del darrer tweet exportat.

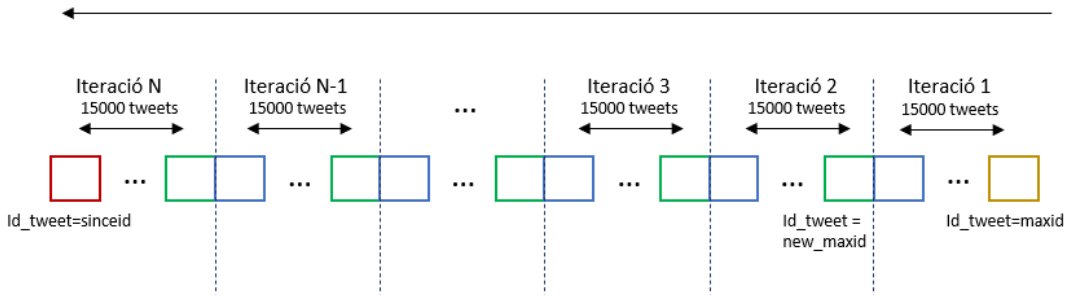

Fig. 12 – Representació de la lògica d'orquestració de les descàrregues de tweets via l'API de Twitter (Python)

5. Catalogació de bots

Un dels principals objectius d'aquest projecte era, inicialment, poder catalogar a partir d'una font externa tots aquells usuaris amb una alta probabilitat de ser bots, i poder analitzar-ne així la presència en una xarxa d'usuaris real (i veure'n l'impacte i l'activitat que hi tenen). Per assolir aquest objectiu, s'ha fet ús d'una eina anomenada *Botometer*[15] que permet, a partir de *l"screen name'* de l'usuari (o del seu *id* d'usuari de *Twitter*), obtenir una puntuació que indica la probabilitat de que l'usuari sigui un bot.

5.1 Botometer

Aquesta eina s'esmenta en molts estudis relacionats amb la detecció de bots en xarxes socials, i tot i ser una dada amb un cert grau de fiabilitat, no deixa de ser una estimació a partir, principalment, de l'activitat (publicació de *tweets*, respostes, *retweets*, seguidors, ...) de l'usuari a *Twitter*, com es descriu en la pròpia documentació de l'eina. Per aquest motiu, és important contemplar que els resultats obtinguts poden no ser totalment fiables.

Tornant a l'eina en si, *Botometer* és fonamentalment una API que respon a peticions en funció d'un model de finançament *freemium*. Això vol dir que, per gaudir de les màximes capacitats, és necessari assumir la càrrega econòmica que suposa. Limitar-se a les capacitats gratuïtes suposa un límit d'aproximadament disset mil consultes diàries, tot i que la realitat és que, amb una latència de gairebé sis segons (en el moment de la descàrrega de les dades), només es poden realitzar al voltant de catorze mil. Comptant que es disposa de gairebé tres-cents mil usuaris diferents en el data set obtingut en el primer apartat, això requeriria dedicar un ordinador a descarregar, durant vint-i-un dies seguits, *l"score'* de *Botometer* de tots i cada un dels usuaris.

Com a comentari addicional, val la pena esmentar també que és necessari proveir unes credencials vàlides de *Twitter* per a poder realitzar les consultes a l'API de *Botometer*, cosa que pot fer desconfiar l'usuari al valorar la privacitat i seguretat de les seves credencials.

Tot i així, i tornant a l'aspecte pràctic, val a dir que aquest procediment ha resultat inviable sense cap mena de dubte, pertorbant així l'abast i l'objectiu inicial (i principal) del projecte, com a mínim amb l'enfoc inicial d'obtenir la catalogació dels tres-cents mil usuaris.

5.2 Nou enfoc i alternatives

Per resoldre aquesta problemàtica, es va decidir realitzar una predicció de la categoria dels usuaris (bot o no bot) utilitzant mètodes supervisats de *Machine learning*.

Bàsicament, els mètodes supervisats empren dades ja catalogades per tal d'entrenar un model predictiu (o classificador), i utilitzar aquest model amb dades futures per a poder estimar (amb un cert risc) la categoria que els hi correspondria. En aquest sentit, és necessari disposar igualment d'un data set per, en primer instancia, entrenar el

model (model que servirà per catalogar tots els usuaris del data set), i per tal fi s'han valorat dues opcions:

- L'obtenció d'un data set extern ja catalogat, generat per exemple arran d'una competició en l'àmbit de la detecció de bots en xarxes socials.
- Aprofitar la infraestructura ja definida en la prova de concepte amb *Botometer* per obtenir el data set d'entrenament (amb la descarrega de, per exemple, quaranta mil usuaris catalogats).

La primera opció sembla a priori molt interessant, tot i que un cop contactats els responsables de diverses competicions, s'ha pogut comprovar que els conjunts de dades compartits es basaven exclusivament en el contingut dels *tweets* (anomenat *body*) i no pas en l'activitat dels usuaris.

Segons Sneha Kudugunta i Emili Ferrara[16], realitzadors de gran nombre d'estudis en l'àmbit de la detecció de bots, la combinació de dades a nivell d'usuari i a nivell de *tweet* (a més del contingut d'aquest) els ha permès d'obtenir els millors resultats. En aquest sentit, les variables que es suggereixen que permeten discernir millor entre usuaris humans i bots són les següents:

- A nivell d'usuari:
  - Statuses Count
  - Followers Count
  - Friends Count
  - Favorites Count
  - Listed Count
  - Default Profile
  - Geo Enables
  - Profile Uses Background image
  - Verified
  - Protected

- A nivell de *tweet*:
  - Retweet Count
  - Reply Count
  - Favorite Count
  - Number of Hashtags
  - Number of URL
  - Number of Mentions

5.3 Solució escollida

Valorant tots els punts esmentats anteriorment, s'ha contemplat dissenyar el model classificador d'usuaris seguint les pautes següents:

- Obtenint un conjunt reduït d'usuaris classificats per *Botometer* (al voltant de trenta-cinc mil usuaris), que serviran com a conjunt d'entrenament per el nostre model.

- Descarregant mitjançant la llibreria *tweepy* i l'*API* de *Twitter* l'activitat més recent (darrers 20 *tweets*) de tots els usuaris que apareixen en el data set obtingut en l'apartat 4. Aquestes dades s'emmagatzemaran en un *DataFrame*, obtenint totes les possibles variables explicatives esmentades en el punt anterior:
    - El nom d'usuari.
    - El nombre mig de caràcters per *tweet*.
    - El nombre mig de *hashtags* per *tweet*.
    - El nombre mig de símbols per *tweet*.
    - El nombre mig de mencions a usuaris per *tweet*.
    - El nombre mig de *urls* per tweet.
    - El nombre mig de paraules per *tweet*.
    - El nombre de *followers* de l'usuari.
    - El nombre d'amics de l'usuari.
    - Si el compte de *Twitter* ha estat verificat (en binari).
    - El nombre de *statuses_count*.
- Unint la catalogació de *Botometer* a les dades obtingudes del punt anterior, i descartant per ara aquells usuaris dels quals no es disposa de cap catalogació.

El resultat de seguir les pautes esmentades prèviament és, doncs, un conjunt de dades d'aproximadament trenta-cinc mil usuaris, que servirà d'entrenament per un model predictiu de bots sobre el conjunt inicial d'usuaris.

6. Models predictius supervisats

Un cop definit el data set d'entrenament, i per tal de definir el model amb la fiabilitat més alta possible, és imprescindible analitzar quins classificadors de Machine learning es poden considerar en l'estudi, i dels que s'adaptin a l'àmbit de l'experiment, comparar els resultats.

De totes maneres, i abans d'entrar en el detall de cada un dels classificadors emprats, és imprescindible tenir presents dos punts que han esbiaixat inicialment els resultats, i que un cop corregits han permès millorar substancialment els resultats obtinguts: la normalització de les dades i la proporció de les categories.

6.1 Normalització i estandarditzación de les dades

Com es pot apreciar en la figura següent, corresponent a un exemple de les dades preparades per a l'entrenament del model, hi ha molta disparitat en els rangs de valors entre cadascuna de les variables.

	avg_length	avg_hashtags	avg_user_mentions	avg_urls	avg_words	followers	friends	statuses_count	verified_num
2	123.55	0.05	2.35	0.00	18.85	76027	75836	258651	0
4	125.50	0.00	1.30	0.05	20.45	63	223	2170	0
6	134.05	0.60	2.15	0.05	19.40	26388	25926	226527	0
15	138.15	0.00	1.40	0.00	21.50	1	0	739	0
19	120.25	0.35	1.00	0.10	18.70	3372	4985	96240	0

Fig. 13 – Exemple de dades generades a partir de les timelines d'usuari

En aquest cas, els models de predicció obtinguts amb aquestes dades queden esbiaixats inherentment per la disparitat entre ordres de magnitud dels factors, i per tant una variació de mil *followers* tindrà més impacte que no pas una diferència de 0.05 *hashtags* per *tweet* entre dos usuaris.

Conceptualment, la normalització radica en convertir cada registre en un vector de factors, i transformar aquests factors de manera que el mòdul del vector tingui valor unitari. L'estandarditzación de les dades, en canvi, es realitza a nivell columnar en el data set. A partir dels valors de desviació estàndard i mitjana, s'aplica una transformació Z per a cada valor (es sostrau la mitjana, i es divideix per la desviació tipus) obtenint així valors en el rang [-1;1].

Aquestes tasques de *preprocessament* es poden realitzar segons diferents estratègies, que en aquest cas s'implementaran gràcies a la llibreria de *sklearn*. Les alternatives considerades dependran exclusivament del cas d'ús, ja que cada model parteix d'unes assumpcions inicials sobre la distribució de les dades, per exemple.

6.2 Proporció de categories

Una de les problemàtiques principals a l'hora de seleccionar el model més adequat (veure apartat següent) ha estat la diferència en la proporció de dades de cada una de les categories dins el data set generat per a l'entrenament i test. En aquest sentit, dels aproximadament trenta-cinc mil usuaris obtinguts, només un baix percentatge (al

voltant de quatre-mil cinc-cents) estaven catalogats com a bots. La proporció de bots en el data set resulta doncs d'un pobre 12,8%.

No haver considerat inicialment un data set més proporcionat ha esbiaixat tots els resultats, generant models molt pobres amb valors del ROC rondant el 0.5, i, per tant, aleatoris.

Per corregir aquest fet ha estat necessari equilibrar les proporcions de les dues etiquetes en el data set. S'ha considerat que un bon punt de partida era agafar tots els usuaris catalogats com a bots en el data set, i agafar un nombre equivalent d'usuaris catalogats com a humans del mateix data set. D'aquesta manera, i amb una proporció de 50-50, els resultats obtinguts en els següents apartats han millorat substancialment.

6.3 Selecció del millor model

De cara a analitzar l'adequació de diferents models per a l'obtenció del millor predictor, s'han realitzat diferents proves amb una sèrie de mètodes supervisats, pels quals es detalla breument en cada cas una breu introducció a la metodologia i principi matemàtic. Seguidament, s'ha cercat mitjançant la llibreria *GridSearchCV* de *sklearn* la millor configuració de paràmetres, mostrant-ne els resultats obtinguts més significatius.

Com a apunt final d'aquesta introducció, destacar que, per mitigar un biaix dels resultats i millorar la precisió de tots els models, s'ha escollit realitzar una validació creuada amb un factor de divisió del data set d'entrenament de tres. En aquest sentit, es divideix doncs el conjunt d'entrenament en tres blocs de dades, i es realitzen tres entrenaments considerant totes les combinacions de subconjunts, de dos en dos, obtinguts a partir del conjunt d'entrenament.

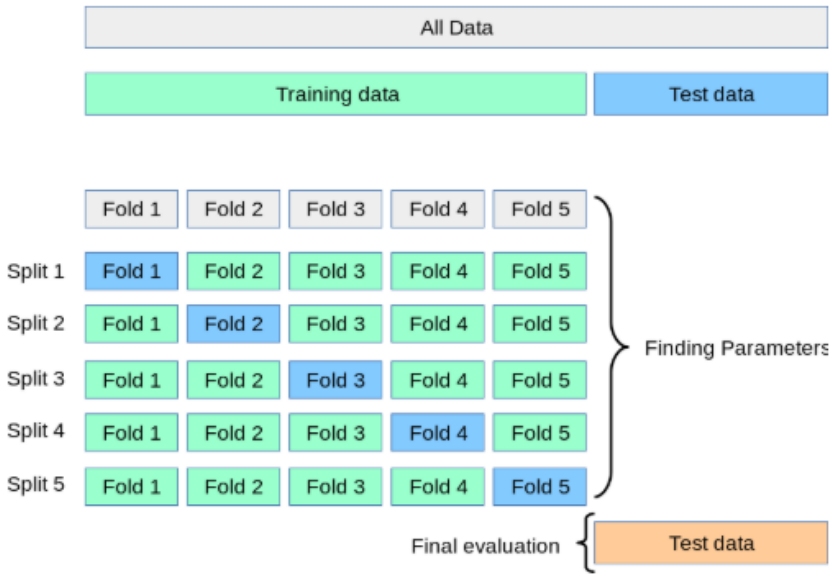

Fig. 14 – Esquema de cross-validation amb factor cinc

6.3.1 LinearSVC

El model *LinearSVC* correspon a un cas específic de *Support Vector Machines* per a classificació lineal. Els *SVM* construeixen un hiperplà o conjunt d'hiperplans en un espai de dimensionalitat elevat, utilitzats per tasques de classificació, regressió i fins i tot detecció de valors anòmals o extrems (*outliers*).

Aplicant aquest model per al cas definit en el projecte, s'obté que la millor combinació de paràmetres és la següent:

Paràmetres	Valor
C	5
Loss function	Squared_hinge
Tolerance	0.0001

El model presenta un *ROC* de 0.609, corresponent a l'àrea sota la corba que es pot apreciar a la gràfica següent. Es pot considerar que aquest model no és un bon predictor de bots, degut principalment al baix valor de ROC. Com a valor de referència, un bon predictor tindria un valor de ROC proper a 1.

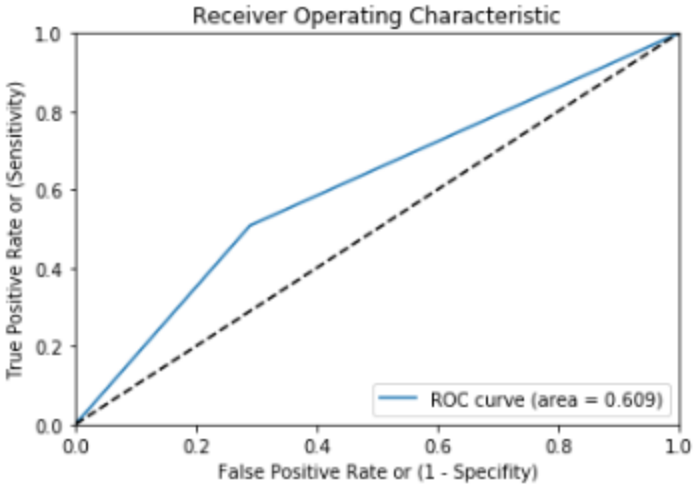

Fig. 15 – Corba ROC de model LinearSVC obtingut

6.3.2 Multilayer Perceptron

El *multilayer perceptron* és un algoritme d'aprenentatge supervisat que, donades unes dimensions d'entrada i de sortida, pot aprendre una funció d'aproximació no-lineal per tasques de classificació o regressió. El model es construeix amb un seguit de capes intermèdies (*hidden layers*) composades de neurones, que transformen els inputs de manera lineal i ponderada, i apliquen una funció d'activació no lineal, per seguidament enviar el output cap a la següent capa.

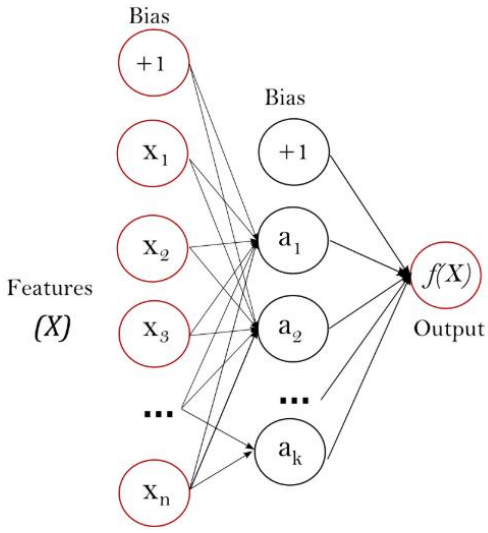

Fig. 16 – Esquema de capes i neurones per un MLP amb una sola capa oculta

En el cas del model Multilayer Perceptron, la millor combinació de paràmetres,i per tant aquella amb la que s'obtenen els millors resultats (un ROC de 0.665), és la següent.

Paràmetres	Valor
Hidden layer sizes	(100,10,5)
Activation function	Relu
Solver	Adam
Alpha	1e-5

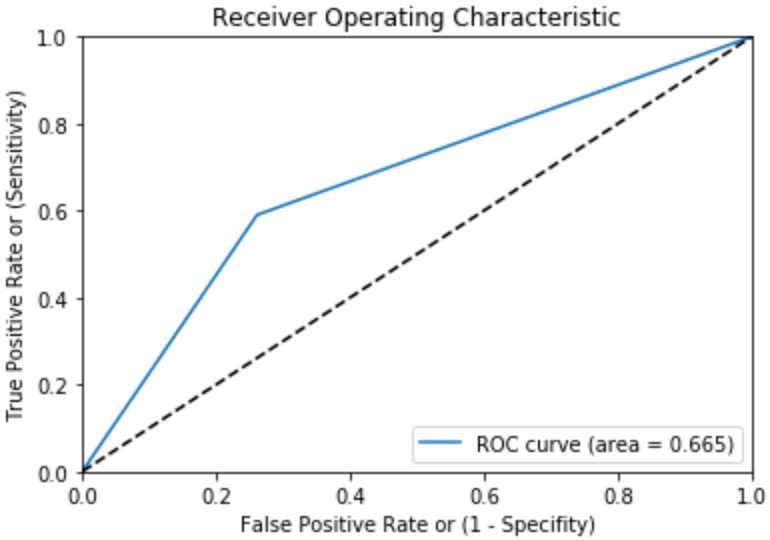

Fig. 17 – Corba ROC de model MLP obtingut

6.3.3 Bernouilli NB

Els classificadors de *Naive Bayes* son algoritmes supervisats que apliquen el teorema de *Bayes* de manera "naiva", assumint la independència entre qualsevol parell de variables donada una classificació. Existeixen diferents classificadors de *Bayes*, que parteixen cadascun d'hipòtesis diferents sobre la distribució que segueixen les dades d'entrada. En concret, aquest model treballa partint de la base que les dades d'entrada segueixen una distribució de *Bernoulli*, segons el qual totes les variables d'estudi haurien de tenir un valor binari (0 o 1) o valors corresponents a dues categories. És possible, de totes maneres, implementar aquest model amb dades contínues, donat que la pròpia llibreria s'encarrega de transformar les dades d'entrada en dades binaries.

Un sol paràmetre es pot configurar amb aquest procés, havent obtingut els millors resultats amb la següent configuració:

Paràmetres	Valor
Alpha	1

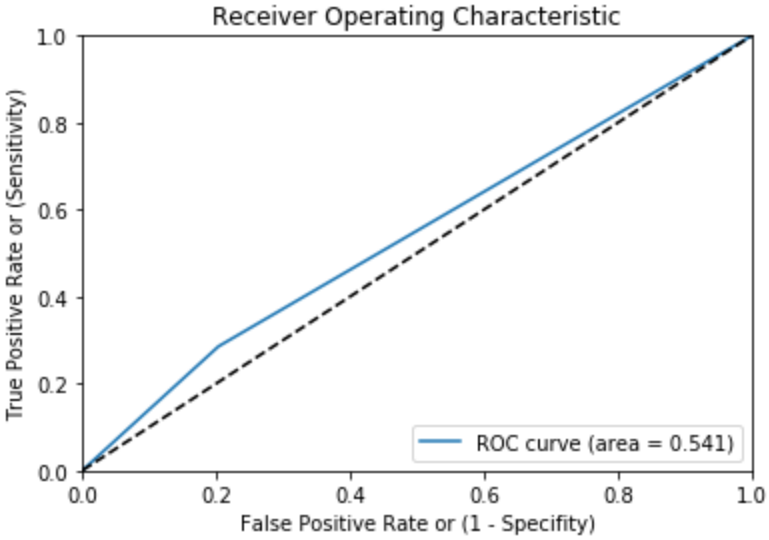

Fig. 18 – Corba ROC de model Bernoulli Naive-Bayes obtingut

6.3.4 Regressió Logística

Els classificadors per Regressió Logística són classificadors dedicats principalment a la predicció de classes binàries per models lineals. El seu nom prové de la funció logística (sigmoide), que de mateixa forma que per a regressions lineals, és la funció a la que s'intenta adaptar el model (al contrari de la regressió lineal, on el model intenta adaptar-se a una recta).

Per tal d'obtenir els millors resultats possibles, és recomanable evitar la presència de valors anòmals (*outliers*), que perjudiquen molt a la precisió del model, així com evitar que les dades d'entrada continguin variables dependents entre elles. Val a dir que,

finalment, els millor resultats s'obtindrien si les dades seguissin una distribució de Gauss.

Com en els casos anteriors, la millor combinació de paràmetres obtinguda ha estat la següent, arribat a un valor de ROC de 0.603, no gaire satisfactori a nivell de fiabilitat del model.

Paràmetres	Valor
Penalty	'entropy'
Tol	20
C	Auto
Solver	4
Max_iter	100

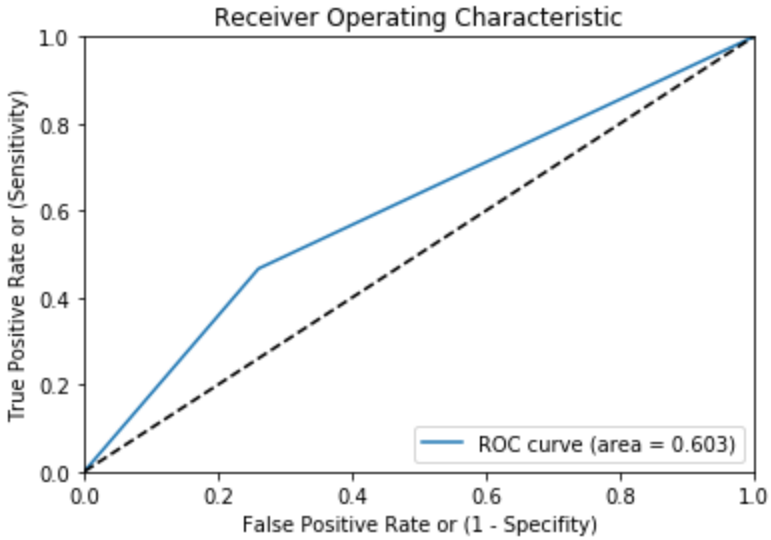

Fig. 19 – Corba ROC de model de Regressió Logística obtingut

6.3.5 Random Forest

Els classificadors per *Random Forest* són considerats classificadors molt robustos, donat que en el procés d'entrenament es seleccionen conjunts de mostres aleatòries amb les quals es defineixen diferents arbres de decisions. De cara a establir una predicció sobre les dades de test, es realitza una votació entre les prediccions de cada un dels arbres generats, i per tant la predicció més votada és la resultant. D'altra banda, aquests classificadors no pateixen sobreentrenament, ja que el vot de la majoria contraresta els biaixos.

En aquest cas, s'han obtingut els millors resultats amb la combinació de paràmetres següent, obtenint el màxim valor de ROC de totes les proves, amb un prometedor 0.71

Paràmetres	Valor
Criterion	'entropy'

Max_depth	20
Max features	Auto
Min samples split	4
N_estimators	100

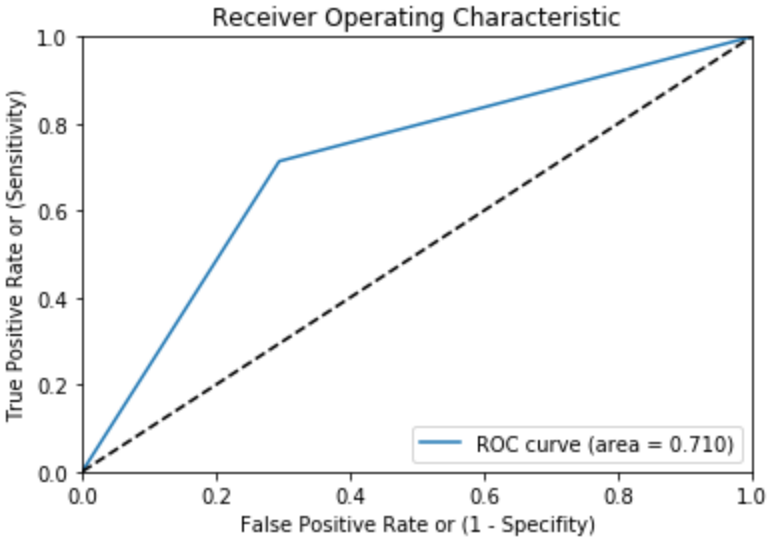

Fig. 20 – Corba ROC de model de Random Forest obtingut

6.3.6 Resum de resultats

A títol de resum, s'han recapitulat els valors obtinguts de ROC dels models obtinguts, com es pot apreciar en la taula següent:

Model	ROC
Linear Support Vector Machine	0.609
Multilayer Perceptron	0.665
Bernouilli Naive Bayes	0.541
Regressió Logística	0.603
Random Forest	0.710

Tot i que els models són millorables, amb un valor de ROC llunyà del valor ideal (1), sembla que les xarxes neuronals (*Multilayer Perceptron*) o els boscos aleatoris (*random forests*) demostren resultats més prometedors, ja que amb un ROC proper a 0.7 es pot afirmar que no es tracten de models aleatoris.

6.4 Predicció de la resta d'usuaris

Un cop escollit el millor model amb totes les variants estudiades, i de cara a poder enriquir la visualització del graf de la xarxa, s'ha volgut predir, a partir del model obtingut, la categoria de tots els usuaris restants del data set. Val a dir que, donat que la precisió del model no és òptima, i que tanmateix el número de falsos positius i falsos negatius tampoc es negligible (mentre no s'obtingui un millor model), cal prendre amb cura aquests resultats. De totes maneres, la visualització del graf amb aquestes categories predites associades a cada usuari pot ser un primer pas en l'estudi de la presencia de bots en la xarxa obtinguda.

Com a darrer apunt d'aquesta secció, és important recalcar que, de cara a obtenir resultats coherents, cal aplicar les mateixes transformacions (concretament, l'estandarditzación de dades realitzada inicialment) al nou conjunt d'usuaris a predir. És important que es tracti de la mateixa transformació, i que no s'estandarditzin les dades de manera aïllada per al nou conjunt, ja que idealment cal que la transformació aplicada a les dades sigui la mateixa en ambdós casos. En cas contrari, la precisió del model es podria veure perjudicada.

7. Generació i visualització del graf obtingut

El paquet de NetworkX per Python presenta moltes utilitats tant per la creació, manipulació i funcionament de gran diversitat de grafs, com en el seu estudi i representació gràfica. Dins de l'abast del projecte, aquesta llibreria ha permès la generació d'un graf representatiu de les interaccions d'usuaris, que s'ha anat enriquint tant amb dades generades en apartats anteriors, com amb noves dades que han aportat un grau més de detall en l'estudi.

7.1 Generació del graf

Existeixen diverses alternatives per tal de generar un graf amb la llibreria de NetworkX, sent la més bàsica la definició d'un graf buit en el qual anar afegint nodes i arestes segons interessi. En aquest sentit, cal precisar que la generació d'un graf es pot realitzar mitjançant simplement la definició d'arestes, donat que en aquest moment es creen automàticament els nodes que no existeixin.

Conceptualment, aquest graf pretén representar les interaccions entre els usuaris, partint de la base que la importància del graf radica en quins usuaris estan connectats entre ells, i no pas en quin d'ells realitza l'acció. Per aquest motiu, s'ha definit un graf no dirigit, en el qual s'han generat arestes entre nodes sempre i quan un dels usuaris hagi realitzat una de les principals accions considerades en l'abast del projecte: *retweet* d'un missatge d'un altre usuari, citació d'un usuari, i resposta directa a un *tweet*.

Ha estat necessari recórrer tot el data set de *tweets* obtingut en els primers apartats del projecte, definint per tant arestes per cada un dels *tweets* que corresponguessin a una de les situacions anteriors, i generant així

7.2 Detecció de comunitats

Un altre punt que s'ha conservat de l'abast inicial del projecte ha estat la detecció de comunitats, de cara a donar més visibilitat sobre el conjunt de dades estudiat. En aquest cas, la pròpia llibreria de NetworkX proporciona funcions per al càlcul de comunitats un cop definit un graf donat.

Existeixen dues alternatives per a la realització d'aquest agrupament d'usuaris, que serien per una banda aplicar el mètode de *Girvan Newmann*, o bé emprar algoritmes basats en l'optimització de la modularitat del graf, més adequat per a grans xarxes amb milions de nodes.

El procés de l'algoritme de *Girvan Newmann* es basa en quatre etapes, en les quals es calcula principalment el grau d'intermediació de totes les arestes d'un graf, descartant a cada pas aquelles arestes amb major grau d'intermediació (i que per tant connecten comunitats). Amb les arestes restants, es recalculen els nous graus d'intermediació, fins que no queden més arestes. El resultat d'aquest algoritme es un dendrograma, en el qual es pot escollir el nombre de comunitats desitjades en funció de l'alçada a la que es realitzi el tall.

D'altra banda, els mètodes d'optimització de la modularitat d'un graf tenen per objectiu, com el seu nom indica, maximitzar la modularitat, que es calcula com a la densitat d'arestes en un graf. Un exemple d'aquests mètodes és l'algoritme de *Louvain*, amb el qual, partint de que cada node es una comunitat, es calcula iterativament l'increment de modularitat ocorregut si un node deixa de pertànyer a la seva comunitat i passa a pertànyer a la comunitat d'un veí.

Tot i haver realitzat l'estudi de comunitats amb aquests dos algoritmes, la grandària de la xarxa i les limitacions dels equips emprats en el projecte han suposat que el cost d'execució del primer mètode no fos assumible (més de 24 hores sense resultats, col·lapsant l'ordinador). Amb una durada d'aproximadament unes dues hores, l'algoritme de *Louvain* ha donat com a resultat 8216 comunitats diferents, de les quals les cinc més grans engloben el 50% dels usuaris, de la següent manera:

Comunitat	*Usuaris*	*Percentatge sobre el total*
3	55481	19.48%
4	34616	12.15%
8	20243	7.11%
35	17874	6.27%
17	16109	5.65%
Suma total	144323	50.66%

7.3 Assignació d'etiquetes

Un cop calculades les comunitats del graf, i disposant de dades sobre la natura dels usuaris (bot o ésser humà), és imprescindible afegir aquestes dades en el graf, de cara a una posterior visualització. Així doncs, és necessari recuperar la catalogació dels usuaris, així com la comunitat a la qual pertanyen, i assignar aquestes dades al node pertinent. Per poder realitzat tal acció, cal tenir present que l'identificador de cada node és el seu '*screen_name'* de *Twitter*.

Un cop actualitzat el graf amb totes les dades, cal exportar aquest graf en un format concret anomenat '*graphml'*. Afortunadament, la llibreria de *NetworkX* permet realitzar aquesta exportació a partir del graf ja generat. De cara a importar aquest graf en una eina de visualització (tal com *Gephi*, que serà l'eina open *source* escollida en aquest cas), cal simplement obrir l'aplicació i importar el fitxer generat, obtenint un núvol de punts sense forma concreta que caldrà processar per donar sentit.

7.4 Visualització del graf

De cara a donar forma a totes les dades obtingudes en el desenvolupament del projecte, falta realitzar un processament de les dades a l'eina de visualització *Gephi*. És important entendre que *Gephi* té un defecte important, i es tracta de la impossibilitat de desfer alguna acció un cop realitzada. Per aquest motiu, es recomanable guardar tots els avenços en diferents fitxers, i en cas de necessitar-ho, tornar a carregar un fitxer d'un estat anterior.

Deixant aquest punt de banda, Gephi proporciona eines per tal de personalitzar la visualització en la mesura del possible, donant llibertat a l'usuari per escollir com vol mostrar les dades, jugant sempre amb la posició dels nodes, la seva mida, els colors de nodes i arestes, i l'etiquetat dels atributs dels nodes, entre d'altres.

Per tal de mostrar de manera senzilla tota la informació de la que es disposa en aquest punt del projecte, i de cara donar bona visibilitat sobre el producte obtingut, s'han realitzar diverses accions de caire estètic per tal de mostrar tanta informació com sigui possible. Aquestes accions consisteixen en:

- Una reorganització dels nodes, agrupant-los al voltant dels nodes amb major grau (és a dir, els més interrelacionats) mitjançant l'algoritme Force Atlas 2.
- Un redimensionament dels nodes en funció del seu grau d'interrelació. Els nodes amb més interaccions es veuran doncs clarament, sent la seva mida proporcional al seu grau.
- Una assignació de colors segons la comunitat a la qual pertanyi el node
- En una visualització posterior, i sense desplaçar els nodes, una distribució dels usuaris bots superposada sobre la visualització anterior, destacant-los en color vermell.

El resultat de totes aquestes accions es pot veure a continuació, on es defineixen clarament dos grans conjunts de nodes, donant peu a un graf polaritzat.

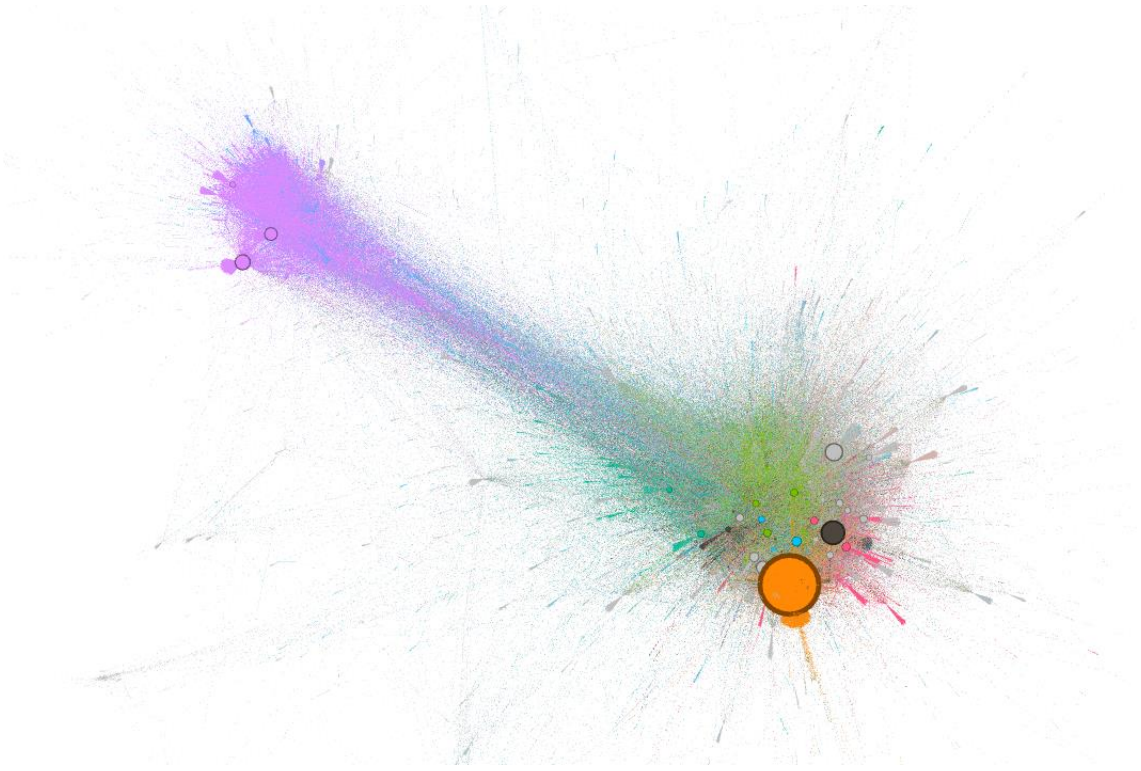

Fig. 21 – Visualització de comunitats del data set (Gephi)

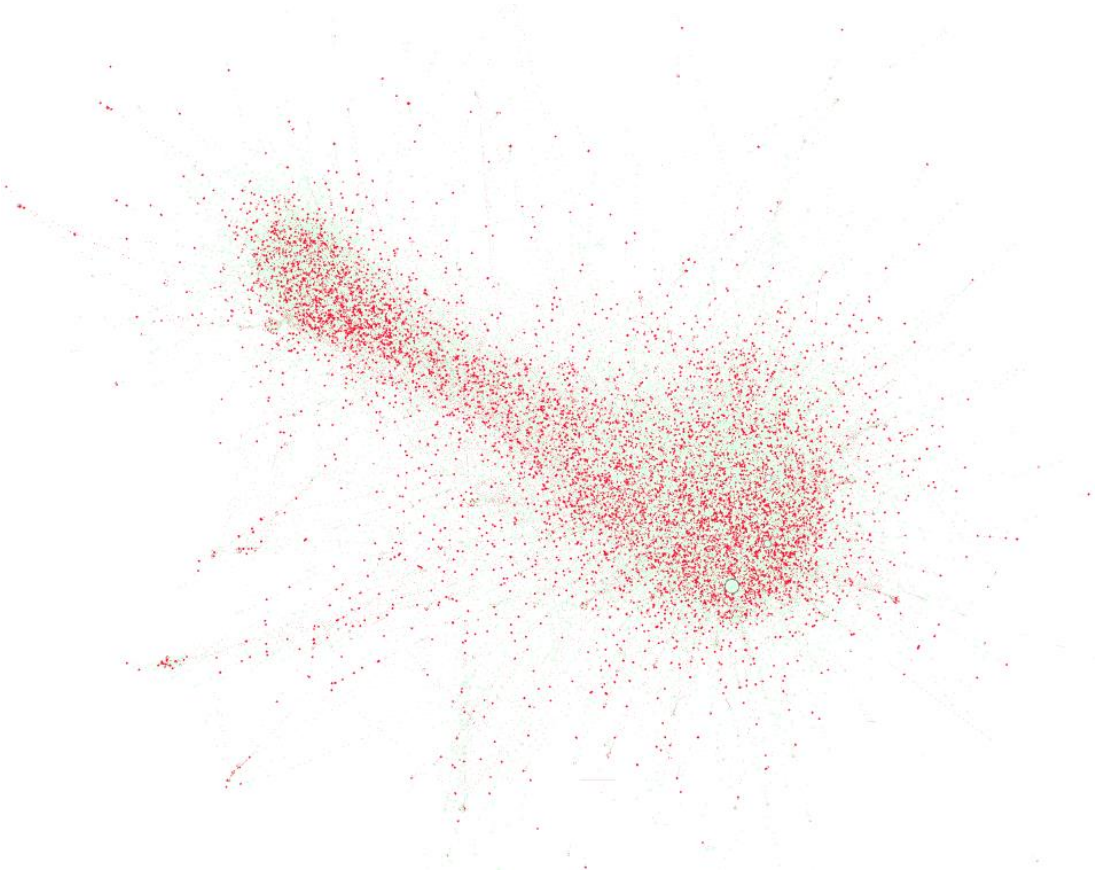

Fig. 22 – Visualització de bots catalogats a partir del millor model de predicció (Gephi)

7.5 Usuaris amb més interaccions

Com a punt final de l'anàlisi, s'ha volgut destacar aquells usuaris amb més interaccions, de cara a detectar patrons en la tipologia d'usuaris que es troben en el conjunt de dades extret:

Usuari	Nom	Interaccions	Professió
@JoeBiden	Joe Biden	40057	Advocat, ex vicepresident del EUA.
@tribelaw	Lawrence Tribe	15673	Advocat dels EUA.
@SethAbramson	Seth Abramson	11694	Advocat i columnista polític dels EUA.
@benshapiro	Ben Shapiro	10170	advocat conservador, comentarista polític i conductor radiofònic.
@mitchellvii	Bill Mitchell	8540	Ex-membre republicà de la cambra de representants de Illinois.
@ShannonFreshour	Shannon Freshour	8453	Candidata al quart districte del Congrés d'Ohio
@SenSchumer	Charles Ellis Schumer	6349	Senador de Nova York
@TheDemCoalition	The Democratic Coalition	6237	Grup anti-Trump de protesta
@ASlavitt	Andy Slavitt	5675	Ex-Obama health care head.
@mmpadellan	Majid M. Padellan	5212	Autor de "The Liddle'est President

Com es pot apreciar en la taula anterior, els deu usuaris amb més interaccions corresponen tots a un patró molt semblant, sent persones físiques reals amb una participació important en el discurs polític americà molt, generalment contraris a la política del president del EUA, Donald J. Trump.

8. Conclusions

Com a cloenda del projecte, valdria la pena realitzar una reflexió crítica sobre la validesa de les observacions extretes. En aquest sentit, i a partir de l'activitat dels usuaris a Twitter, s'ha aconseguit, mitjançant mètodes supervisats, definir un model predictor de bots amb una probabilitat d'èxit en la classificació del 70%, que tot i ser llunyana del valor ideal (1), representa un primer pas d'un eventual model més elaborat. Aquest model ha permès categoritzar tots els usuaris que han publicat missatges a Twitter en un període temporal donat (aproximadament uns deu dies) i que contenien unes paraules clau específiques. D'aquesta manera, s'ha pogut visualitzar, sobre un graf, la presència de bots en un ecosistema a Twitter.

En aquest sentit, s'ha pogut comprovar que, a diferència de la creença inicial en començar el projecte, els bots no son partidaris de cap sector, sinó que existeixen en aquesta xarxa social, i es relacionen amb qualsevol individu de la xarxa social. No hi ha doncs zones de densitat clares, però si una homogeneïtat en la seva presència a la representació gràfica del conjunt de dades que sembla, si més no, curiosa.

De totes maneres, queden punts de millora que es podrien arribar a implementar. Per exemple, les hipòtesis de partida han pogut limitar la qualitat del model, donat que, per exemple, s'han considerat només variables d'activitat d'usuari, sense entrar a analitzar el contingut dels tweets. D'altra banda, un conjunt de dades d'entrenament de més qualitat, sense dependència d'una estimació de tercers, podria ser un punt important en la millora de les prediccions, sense comptar que l'ús de models no supervisats seria també plantejable.

La realització d'aquest projecte ha posat de manifest, però, que la realitat d'un projecte d'aquest caliu suposa efectivament un alt percentatge del temps a la preparació de les dades, amb intents infructuosos, proves i infinitat de problemes tècnics, deixant un temps molt reduït, finalment, per a l'explotació de resultats.

En aquest sentit, s'ha pogut comprovar de primera mà la magnitud de dades que es generen diàriament a internet. El conjunt de dades estudiat conté 650K tweets extrets en un rang de 10 dies, tractant-se només d'aquells tweets que coincidien amb la cerca implementada. Només aquests 650K tweets ja han suposat un mal de cap a l'hora d'executar certs processos, arribant a trigar fins a 24 hores en les que l'ordinador estava única i exclusivament dedicat a l'extracció i tractament de les dades, amb càlculs molt pesats que requeririen de maquinària més potent.

Tot i així, i probablement degut al canvi d'enfoc del projecte, algunes de les preguntes inicials més enfocades a la propagació de fake news i misinformation no han quedat resoltes, però els resultats obtinguts han obert una caixa de noves preguntes i idees que poden donar peu, si més no, a moltes noves línies d'investigació.

9. Glossari

Twitter: Servei de microblogging mundialment conegut, amb l'emblema d'un ocell blau.

Follower: Es diu dels usuaris que segueixen a altres a Twitter, sent un indicador de la popularitat d'un perfil

Hashtag: Conjunt de caràcters precedits per el caràcter #, servint per identificar o etiquetar missatges en webs de microblogs.

Tweet: nom amb el que s'anomenen aquells missatges enviats a través de Twitter.

Retweet: publicació d'un missatge escrit per un altre usuari a dintre de Twitter.

Honeypot: Literalment pot de mel. En aquest context, tipologia de bot de Twitter que atrau l'activitat d'altres bots, permetent-ne la detecció, com si d'abelles buscant mel es tractés.

API: Interfície de programació d'aplicacions, que engloba el conjunt de subrutines, funcions i procediments que ofereix una llibreria com a capa d'abstracció, de cara a ser utilitzat per un altre software.

SVC: Acrònim de Support Vector Classification, relacionat amb el mètodes supervisats de Supper Vector Machines.

MLP: Acrònim de Multilayer Perceptron.

Python: llenguatge de programació multiparadigma (orientació a objetes, programació imperativa, i en menor mesura, programació funcional) utilitzat en el projecte com a base del projecte.

EUA: Estats Units d'America.

10. Bibliografia

[1] Max Roser, Hannah Ritchie and Esteban Ortiz-Ospina (2020) - "Internet". Published online at OurWorldInData.org. Obtingut de: https://ourworldindata.org/internet [Recurs en línia]

[2] DOMO, “Data every minute”, Obtingut de: https://www.domo.com/learn/data-never-sleeps-7 [Recurs en línia]

[3] Francesco Pierri, Stefano Ceri (2020), “False News On Social Media: A Data-Driven Survey”.

[4] Redacció de La Vanguardia, “El Barça rescinde el contrato con I3 Ventures al descubrir su vinculación con las cuentas difamatorias”, Obtingut de: https://www.lavanguardia.com/deportes/fc-barcelona/20200218/473649822700/bartomeu-barcelona-i3-ventures-rescinde-contrato.html [Recurs en línia]

[5] L. Wu and H. Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pages 637–645. ACM, 2018.

[6] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. Science, 359(6380):1146–1151, 2018

[7] l’ICWSM 2020 DATA CHALLENGE, https://sites.google.com/view/icwsm2020datachallenge

[8] EFE, https://www.lavanguardia.com/tecnologia/20200322/4825671008/las-redes-sociales-contra-la-desinformacion-sobre-el-coronavirus.html [Recurs en línia]

[9] Manlio de Domenico, Riccardo Gallotti, Pierluigi Sacco, Nicola Castaldo, Francesco Valle (2020), ”COVID19 Infodemics Observatory”, https://covid19obs.fbk.eu [Recurs en línia]

[10] Massimo Stella, Emilio Ferrara, Manlio De Domenico, “Bots increase exposure to negative and inflammatory content in online social systems”, Proceedings of the National Academy of Sciences Dec 2018, 115 (49) 12435-12440; DOI: 10.1073 /pnas.1803470115. Obtingut de: https://www.pnas.org/content/115/49/12435 [Recurs en línia]

[11] Kyumin Lee, Brian David Eoff and James Caverlee, “Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter” (2011), Texas A&M University. Obtingut de: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2780/3296 [Recurs en línia]

[12] Bovet, A., Makse, H.A. Influence of fake news in Twitter during the 2016 US presidential election. *Nat Commun* **10,** 7 (2019). Obtingut de: https://www.nature.com/articles/s41467-018-07761-2 [Recurs en línia]

[13] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, Michael M. Bronstein, "Fake News Detection on Social Media using Geometric Deep Learning", Obtingut de: https://arxiv.org/abs/1902.06673 [Recurs en línia]

[14] Shuo Yang, Kai Shu, SuhangWang, Renjie Gu, Fan Wu, Huan Liuz, "Unsupervised Fake News Detection on Social Media: A Generative Approach"

[15] https://rapidapi.com/OSoMe/api/botometer-pro/details

[16] Sneha Kudugunta, Emilio Ferrara, "Deep Neural Networks for Bot Detection" (2018), arXiv:1802.04289v2 [Recurs en línia]

[17] Documentació de NetworkX, https://networkx.github.io [Recurs en línia]

[X] Documentació sobre la llibreria tweepy per la descarrega i processament de tweets en Python, http://docs.tweepy.org/en/latest/index.html.

[X] Documentació de Scikit Learn, https://scikit-learn.org/stable/supervised_learning.html, [Recurs en línia]

11. Agraïments

Agrair especialment a en Julià Vicens, el meu tutor, per el seu suport durant aquests mesos, i la seva proactivitat i interès a l'hora de compartir informació rellevant per al projecte.

Agrair també a en Paolo Rosso (Universitat Politècnica de València) per compartir un dades d'entrenament (relacionades amb la competició PAN19 Author Profiling: Bots and gender profiling).