Anàlisi d'extremistes violents en xarxes socials

Pau Riba Escobar, Autor. Jordi González Sabaté, Tutor.

Resum—Aquest projecte tracta la problemàtica de l'extremisme violent a les xarxes socials. Busca analitzar la conducta que segueixen els extremistes a través dels missatges que aquests mateixos publiquen a les xarxes socials. L'anàlisi se centra en la xarxa social twitter, una xarxa amb una gran quantitat d'usuaris. Alguns només busquen entreteniment i compartir part de les seves vides amb els seus, mentre que d'altres intenten fer de les xarxes socials un lloc hostil, en el que la violència i l'extremisme conviuen i creixen gràcies a la difusió que la xarxa els proporciona. En aquest treball busquem, a partir de *tuits* d'usuaris fidels a ISIS, en primer lloc analitzar d'on provenen aquests missatges, qui els publica, qui els veu i qui son els usuaris més influents dins de la comunitat estudiada. D'altra banda busquem predir a partir d'un model de classificació simple basat en Naive Bayes si un determinat tuit que parla sobre ISIS ha estat realment escrit per un seguidor del grup terrorista radical o és merament una publicació de caire informatiu. Per a fer-ho farem servir una aproximació probabilística a partir de la tècnica *Bag of Words approach* que ens permetrà predir el sentiment d'un tuit a partir de les paraules que hi apareixen.

Paraules clau—Xarxes socials, Twitter, Tuit, Retuit, Anàlisi de clusters, Anàlisi del sentiment, ISIS, extremisme, violència, Naive Bayes, *Bag of Words approach,* Visualització de Dades, anàlisi del sentiment

Abstract— This project addresses the issue of violent extremism on social media. It seeks to analyze the behavior of extremists through the messages they post on social media. The analysis focuses on the social network twitter, a network with a large number of users. Some only seek entertainment and share part of their lives with their own, while others try to make social media a hostile place, where violence and extremism coexist and grow thanks to the spread that the network provides. In this work, we seek, from the tweets of loyal users of ISIS, to first analyze where these messages come from, who publishes them, who sees them and who are the most influential users within the community studied. On the other hand, we seek to predict from a simple classification model based on Naive Bayes whether a particular tweet about ISIS was actually written by a follower of the radical terrorist group or is merely an informative publication. To do this, we will use a probabilistic approach based on the *Bag of Words approach* technique that will allow us to predict the feeling of a tweet from the words that appear in it.

Index Terms—Social Media, Twitter, Tweet, Retweet, Cluster Analysis, Sentiment Analysis, ISIS, Extremism, Violence, Naive Bayes, Bag of Words approach, Data Visualization, sentiment analysis

1 Introducció i presentació del problema

Les xarxes socials han esdevingut un gran aparador per a tot tipus de contingut.Tot i que en molts casos aquest contingut ha sigut o bé neutre o bé positiu en d'altres trobem contingut ofensiu, violent, i en definitiva inapropiat per a un espai que és públic i que pot visitar pràcticament tothom.

Dins d'aquest context de continguts ofensius hi trobem ,de vegades, el comportament que volem analitzar en aquest treball. I, això és l'extremisme. Què vol dir ser extremista doncs?

"L'extremista es una persona les creences de la qual son radicalment diferents a les

creences de la resta de la societat."

L'extremisme normalment es presenta en contextos polítics i religiosos, tot i que també els trobem en altres contextos que no tindrem en compte en aquest treball com per exemple en el context esportiu o, més concretament el futbol. Les xarxes socials estan plagades d'aquests tipus de missatges. Els missatges no son únicament escrits, sinó que també trobem contingut fotogràfic i audiovisual que dóna mostres d'extremisme. Aquí tenim alguns exemples de missatges escrits amb contingut extremista:

"Que caras los de Nigeria!! Messi, por las dudas guarda el celular". Tuit de @marley_ok

"@Renfe me ha robado un panchito de vuestros trenes". Tuit de @SoldadoAtlético

Amb aquests dos exemples podem veure tuits de caire racista. Veiem però que els dos tuits tenen diferències entre ells. El primer tuit veiem que té un context implicit que cal conèixer per saber que el que es diu és un

comentari racista. El context del primer tuit és el següent: es juga un partit entre les seleccions nigeriana i argentina de futbol, els nigerians són étnicament negres. El tuitassocia el fet de ser negres amb el fet de que poden robar-li a Messi amb el comentari "por las dudas guarda el celular". D'altra banda en el segon tuit el contingut és molt més explícit. Una persona reclama a Renfe, una empresa ferroviaria, que un "panchito", referint-se a una persona sudamericana i de ètnia indígena, li ha robat en un dels seus trens.

D'aquests dos exemples podem extreure la conclusió de que el contingut extremista, en aquest cas racista, pot ser presentat a les xarxes socials de
manera explícita o de manera implícita. Aquestes dues formes d'expressar-se les trobem en contingut escrit però també en contingut audiovisual, perquè seguirem tenint paraules que són interpretables. Vist el fet que tenim dos tipus de missatges tant els explícits com els implícits són perillosos per a la societat. El motiu d'això és que de vegades, i especialment en el cas del missatges extremistes i violents ens podem trobar en casos que aquests missatges influeixin a certs grups de persones vulnerables a aquests. D'aquest fet en tenim varis exemples en la nostra societat avui en dia. El terrorisme islàmic és un d'ells, en que en pos de reclutar més fidels a certs grups terroristes, infundeixen en la població un odi indiscriminat contra el món Occidental, odi que moltes vegades és irracional i provocat per la situació desesperada en que es troben aquestes persones.
Un exemple similar són els delictes d'odi contra menors immigrants, segons el Observatorio Español del Racismo y la Xenofobia (OBERAXE). Als mesos de juliol i agost de 2021 segons OBERAXE el 30% dels missatges d'odi volcats en xarxes socials eren contra aquest col·lectiu. En general, com va manifestar el 28 de juliol en roda de premsa el ministre del interior del govern espanyol Grande-Marlaska , els delictes d'odi han augmentat un 9,3% desde 2019. Molts d'aquests delictes comesos en xarxes socials.

Dit això ja tenim identificat el problema: les xarxes socials són un portal públic per a manifestar comportaments extremistes i violents que poden tenir un gran impacte en la societat.

2 Objectius del treball

Els objectius d'aquest treball són dos principalment:

El primer d'ells és el de analitzar el comportament dels usuaris d'ISIS que publiquen tuits violents o d'odi. Qui son els usuaris que fan aquestes publicacions, qui son els usuaris que veuen aquests tuits i els comparteixen amb la resta d'usuaris de twitter (fer Retuit). Quines comunitats es formen entre usuaris de ISIS: quins usuaris només mencionen a altres usuaris o *retuitiegen* els seus tuits i no en reben, els que només reben aquestes mencions i retuits i aquells que son mencionats i retuitejats i a la vegada ells ho fan amb altres usuaris. A banda d'això s'ha volgut fer un anàlisi de les diferències entre els tuits de la base de dades de tuits informatius sobre ISIS i la base de dades de tuits de seguidors de ISIS. Aquest anàlisi es fa en un espectre temporal, de manera que es comparen a nivell temporal com es produeixen tuits informatius i tuits violents.

Havent acabat la part analítica el segon objectiu es ser capaços de classificar els tuits en dues categories:

1. Tuit que parla sobre ISIS de manera informativa
2. Tuit d'un seguidor real d'ISIS

Això es farà amb una aproximació probabilísitca del problema de classificació basada en la classificació mitjançant Naive Bayes, que estipula la independencia condicional. En aquest cas el que estarem dient es que calcularem la probabilitat de que un tuit sigui de la categoria 1 o de la categoria 2 segons l'aparició de les paraules del mateix tuit en altres tuits d'una i altra categoria. De manera que calcularem la probabilitat de que un tuit pertany a una i altra categories i el classificarem com a tuit de la categoria que hagi obtingut una probabilitat més alta.

Aquesta és la tasca principal d'aquest segon objectiu, no obstant això, farem un anàlisi de com de bé es classifiquen cadascuna de les categories i de quines són les millors configuracions d'hiperparàmetres per a una bona classificació i com la variació d'aquests afecta a la mateixa.

3 Metodologia de treball

Per a desenvolupar aquest projecte, que al final es un projecte software el tipus de metodologia ideal es una metodologia àgil com Agile. Aquest tipus de metodologies el que permeten es major flexibilitat a l'hora de desenvolupar el software, ja que funcionen de manera iterativa i incremental, i permeten reaccionar millor davant els canvis. No obstant això aquest tipus de metodologies estan pensades per a treballar amb equips, de poques persones en general, però equips. En aquests equips hi tenim varis rols, que en aquest cas seré jo el que els adopti tots excepte el de client, que en aquest cas serà el meu tutor.

Agile es una metodologia de desenvolupament que neix a principis de segle com a conseqüència dels problemes que portaven les metodologies de desenvolupament del segle passat, com el mètode de desenvolupament en cascada, on els principals inconvenients en el desenvolupament de projectes software eren la poca integració i implicació dels clients(en aquest cas el tutor) en el desenvolupament del projecte. Hi ha una sola fase de captació de requeriments i una sola fase d'entrega del projecte, un cop està ja fet. Això provoca que potser el client no es trobi amb el que espera, ja que no pot modificar res un cop establerts els requeriments , o pot però a canvi d'un gran cost. Per contra les metodologies de desenvolupament àgils de SW com agile es fan varies iteracions del cicle de desenvolupament tradicional on hi afegim un contacte freqüent amb el client a través de reunions periòdiques

per tenir feedback i poder fer canvis amb un cost molt menor que no pas amb una metodologia de desenvolupament tradicional.

4 ESTAT DE L'ART

He pogut llegir varis treballs que segueixen la temàtica del present treball. Per exemple en el cas de [1] podem observar que es un treball on es fa una anàlisi del sentiment com el que fem aquí pero el realitzen amb múltiples sentiments(més de dos, com és el cas d'aquest) i mitjançant dos algorismes de classificació .

En primer lloc amb un classificador basat en vectors de suport anomenat *Linear Support Vector Classifier.* Aquest classificador el podem trobar a la llibreria de python *sklearn* i a diferència d' un SVC(*Support Vector Classifier)* usual de *sklearn* utilitzant un kernel lineal, la llibreria que s'utilitza en la classe que implementa el *Linear Support Vector Classifier*, anomenada *LinearSVC* s'adapta millor a problemes on la classificació és multinomial com en el cas que presenta l'estudi. Amb aquest tipus de classificador s'obtenen molt bons resultats amb accuracys mitjans superiors al 80%.

D'altra banda també a [1] utilitzen com a segona estrategia de classificació una aproximació amb Naive Bayes molt semblant a la que hem aplicat en aquest treball amb la única diferència que en el cas de [1] el problema es de classificació multinomial mentre que en el nostre cas es binomial. Els resultats obtinguts pel classificador multinomial de Naive Bayes que fan servir en [1], que és el de la llibreria *sklearn* , són àmpliament inferiors als resultats obtinguts anteriorment amb *LinearSVC*(accuracy mitjana del 66% amb *Naive Bayes* en front al 82% amb *LinearSVC).*

No obstant això, si parlem de *Naive Bayes* els resultats estan marcats per la poca quantitat de dades, especialment si parlem d'un problema multinomial com aquest on tenim en concret 4 categories diferents a classificar: moderat, neutral, poc extrem, i altament extrem. En aquest cas comptem amb 20000 files i unes 476000 paraules. Tot i que també cal destacar que en ocasions la sobreinformació també es perjudicial per a la classificació en el cas de Naive Bayes. Pot donar-se el cas que una certa paraula x aparegui tant en instàncies del tipus moderat com en instàncies del tipus poc extrem, complicant així la classificació entre aquests dos tipus. Una altra cosa que sorprenent d'aquest treball és la manera de tractar les dades per a ser utilitzades en la classificació. La classificació d'una determinada instancia(fila del dataset) en les categories ja esmentades(moderat, poc extrem, altament extrem) es fa mitjançant una puntuació calculada a partir dels *lexicons*(arrels lèxiques) en els diferents idiomes amb els que treballen, a cadascun d'ells els donen una puntuació basada en el seu nivell d'extremisme. En aquest cas els nivells d'extremisme seran pesos, i en tenim 5: abús(-2),terror(-4),lluita(-3),extraordinari(2),pau(3). A partir dels pesos de cada lexicon es calcula una puntuació per a la fila corresponent que a la vegada serà classificada en els grups primerament esmentats: moderat,neutral,poc extrem, altament extrem.

També m'agradaria comentar un altre treball que ha seguit una aproximació molt diferent a la que jo he seguit però molt interessant a la vegada. És el cas de la feina de *[2] Imane Guelil,Ahsan Adeel, Faical Azouau* on veiem una detecció de llenguatge d'odi en la comunitat aràbiga. En primer lloc ells recullen dades referents a l'entorn polític algerià a partir de comentaris,i respostes als comentaris, de vídeos de Youtube que parlen de la cúpula governamental d'Algèria. Per extreure les dades de manera més ràpida usen un conjunt de paraules usuals referents al govern d'Algèria(concretament 43). Després augmenten aquest número de paraules clau usuals a partir d'algorismes com el *Word2vec*(encara que aquest no és un algorisme ell mateix sinó que és una família d'arquitectures model per a aprendre *embeddings* de paraules) i el *voluminous corpus* que permeten buscar paraules semànticament semblants a les 43 originals. I per últim filtren ells manualment i seleccionen només aquells vídeos més relacionats amb l'entorn polític Algerià.
Per a cada paraula en la llista final de 205 paraules , s'extreuen els 50 vídeos més relacionats amb la propia paraula(on la paraula hi apareix més), i d'aquests vídeos s'extreuen els comentaris i acabem tenint més de 1.200.000 comentaris.

Per a construir el seu dataset d'entrenament (*train*) el que van fer es seleccionar 5000 comentaris aleatòriament i fer que 3 anotadors(3 persones que classifiquen els comentaris) classifiquin els comentaris. Un cop classificats tots els comentaris per tots els anotadors es seleccionen aquells comentaris en que ha coincidit la classificació; una classificació que serà 0(no odi), 1(odi). D'això obtenen un dataset inicial que està desbalancejat que conté molts més comentaris que no són d'odi que els que sí que ho son. Llavors el que es fa es seleccionar de manera aleatoria comentaris del set dels comentaris que no són d'odi fins arribar a la mateixa quantitat de comentaris que en el set de comentaris d'odi.

Un cop construïts els sets es fa una extracció de caracterísitques mitjançant dos algorismes: *Word2vec*(per a algorismes d'aprenentatge computacionals classics) i *fastText*(per a algorismes de *Deep Learning*). Aquests dos algorismes generen dos representacions de les dades:

- *Bag-of-Words*: prediu la probabilitat d'una certa paraula en funció del seu context
- *Skip-Gram*: predicció del context d'una paraula(paraules que l'acompanyen) a partir de la mateixa paraula

Un cop extretes les característiques es passa a la classificació. En aquest estudi es comparen gran nombre de classificadors i els millors resultats s'obtenen amb una aproximació *Skip-Gram* per als classificadors LSVC i MLP.

Un altre treball que és molt interessant analitzar per al present projecte es el de *[3]Safa Alsafari, Samira Sadaoui* i *Malek Mouhoub* que tracten la detecció de llenguatge ofensiu i d'odi en les xarxes socials aràbigues.
D'aquest treball i els anteriors cal destacar el fet de que la base de dades està construïda pels propis investigadors. En aquest cas la base de dades es de tuits, com la que tindrem nosaltres en el present treball. No obstant això la nostra base de dades ja estava més treballada, la seva s'ha fet desde zero. Al treball s'explica com a partir de l'API de *twitter* s'han recuperat tuits segons 3 criteris bàsicament. El primer és el de obtenir tuits que continguin unes certes paraules clau o *keywords*, cosa que ja haviem vist en anteriors treballs.
En el cas d'aquest treball però, també es recuperen tuits a partir de dos criteris més. El primer d'aquests dos criteris és el perfil. En casos d'odi pot haver-hi uns perfils *target* o *objectiu*. Per exemple en el cas d'un discurs d'odi de gènere hi haurà una serie de comptes de figures públiques dona que són potencial objectiu d'aquests atacs d'odi. Aquest cas en que tenim un objectiu de l'odi, aquest tipus de discurs s'anomena dirigit o objectivat. És molt interessant posar exemples d'aquest tipus de discurs en el dataset. En analogia als tuits serien els tuits que estan dotats de mencions, és a dir, una @ seguida d'un nom d'usuari. Ex.:@usuari1.

Abans de que els anotadors classifiquin els tuits, es produeix una neteja dels propis tuits que inclueix eliminar tuits duplicats, tuits amb anuncis, tuits similars(mitjançant mètriques de similaritat com la de Jaccard) i tuits el context dels quals difereix de forma notable del context objectiu del treball.
Doncs els algorismes de classificació funcionen molt millor quan el context sobre el qual es treballa està ben definit. En el cas del paper en qüestió el context és del discurs d'odi escrit en àrab modern estàndard(MSA en anglès) i dialecte del golf mitjà(MGD en anglès). Un cop netejat el dataset es contracten a anotadors que durant 2 mesos classificaran cada una de les dades.

El més interesant d'aquest treball és el fet que es fa un treball de classificació exhaustiu provant tant algorismes de classificació supervisats, com algorismes de deep learning. A més hi trobem la taula de la Fig. 1 on veiem un resum de les tècniques de detecció de llenguatge d'odi.

Labels	Features	Algorithms	Measures
Offensive vs Not Offensive	Unigram, Bigram and	SVM	F-macro
Hate vs Not Hate	Trigram	NB	
Clean vs Abusive	Character-ngrams	DT	Accuracy
Clean vs Obscene	Typed dependencies	RF	AUC
Normal vs Abusive	n-gram	LSTM	
Targeted vs Non	Sentiment scores	GRU	
Targeted	Syntactic features	CNN	
Hate vs Offensive vs Ok	Pos tag	BILSTM	
Normal vs Abusive vs	Linguistic features	Ensembles	
Hate	Average length of		
Hate Target: Individual	word		
vs Group vs Other	Distributional		
Hate Target: six	semantics		
religious groups	Word2Vec, Fasttext		
	Glove, Babylon, Bert		

Fig 1. Taula on veiem les diferents tècniques que es poden aplicar per a la detecció i la classificació del llenguatge d'odi.

Per últim m'agradaria també comentar la feina de *[4] Marzieh Mozafari, Reza Farahbakhsh i Noël Crespi* que fan un estudi sobre la detecció de llenguatge d'odi i mitigació del biaix racial en xarxes socials basat en el model de BERT. El model de BERT és un model de llenguatge pre-entrenat pel processament del llenguatge natural (NLP en anglès) desenvolupat al 2018 per *Google*. En el paper s'indica que ells desenvolupen 2 mòduls, el primer per la detecció de llenguatge d'odi i el segon per la mitigació del biaix que comparteixen el component BERTBASE que es compartit pels dos mòduls. Es pot veure l'arquitectura que es fa servir en la Fig.2

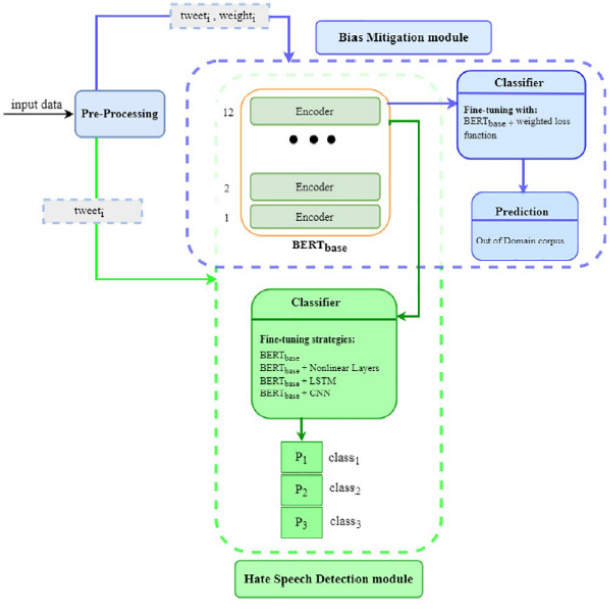


Fig. 2: Arquitectura basada en el model BERT proposada. Veiem que hi han 2 mòduls que comparteixen el BERTBASE. El mòdul de detecció és el que podem veure delimitat per una línia discontinua verda. El mòdul de mitigació del biaix està delimitat per una línia discontinua blava.

5 Desenvolupament del treball

Aquest treball com hem comentat amb anterioritat es descompon en dos parts diferenciades. La primera d'elles és una part en que analitzem els tuits dels usuaris que tenen una relació d'afiliació o pertinença al grup terrorista islàmic ISIS. L'anàlisi es fa sobre una primera base de dades que només conté dades d'usuaris relacionats amb ISIS. Aquest anàlisi serà referent a les comunitats de simpatitzants d'ISIS que es creen a la propia xarxa social de *twitter* , quins son els usuaris que més influència tenen, com es comporten en referencia a la seva activitat; és a dir si tenen un paper diferencial en la comunitat o si no són tan importants entre d'altres aspectes. Això pel que respecta únicament a la base de dades que conté tuits d'usuaris simpatitzants d'ISIS.

Després més tard crearem una base de dades més gran a partir de la que tenim de tuits d'odi d'ISIS que tindrà també tuits sobre ISIS però de caire simplement informatiu. Sobre aquesta segona base de dades també realitzarem un anàlisi referent a la temporalitat dels tuits que es publiquen. Veurem les diferències entre les zones horàries en les que es publiquen tuits d'ISIS i tuits sobre ISIS.
Aquesta segona base de dades més gran és la que ens servirà per a posar a prova el nostre classificador basat en Naive Bayes i un conjunt de característiques respaldat pel model/ aproximació *Bag-of-words*. Tindrem una gran base de dades amb la que esperem obtenir uns resultats en la detecció de tuits d'odi molt bons.

5.1 Bases de dades

Un cop feta una petita introducció sobre en què consisteix el desenvolupament del treball cal explicar com seran les dades amb les que tractarem.

En primer lloc com he dit tenim una primera base de dades de tuits de *fanboys* de ISIS. Aquesta base de dades conté tuits de més de 100 usuaris desde que es varen produir els atacs del Novembre de 2015 a Paris. Aquesta base de dades la podem trobar a *Kaggle* amb el títol de *[5] How Isis Uses Twitter* on veurem també una descripció dels atributs/ dades que hi podem trobar a la mateixa. Són 8 i són els següents:

1. ***name:*** el nom de pila de l'usuari que ha realitzat el tweet.
2. ***username***: el nom d'usuari.
3. ***description***: descripció del tuit. És útil en el cas que apareixi contingut audiovisual en el tuit
4. ***location***:Localització
5. ***followers***:Nombre de seguidors en el moment de la descàrrega del tuit
6. ***numberstatuses***: Nombre de peticions fetes a l'API de *twitter* fins al moment de la descàrrega del tuit
7. ***time***:Data i *timestamp* del tuit. Hora exacta en que es va publicar el tuit
8. ***tweets***:El cos del tuit/contingut del tuit.

En segon lloc tindrem una segona base de dades composta de la base de dades descrita ja(que conté tuits de fanboys d'ISIS) i una altra base de dades que conté tuits que parlen d'ISIS però no en una connotació extremista o violenta si no que són tuits merament informatius. Aquesta base de dades de tuits informatius la trobem a només conté quatre atributs que la base de dades de fanboys d'ISIS té. Per tant la nova base de dades que contindrà aquests 4 atributs coincidents en les dues bases de dades, més un atribut extra que serà 1 si la fila pertanyia a la base de dades de fanboys d'ISIS(per tant si la fila correspon a la informació d'un tuit d'odi) i 0 en cas de que hagués estat a la base de dades de tuits informatius sobre ISIS(per tant tuits amb sentiment merament informatiu, i no violent).

De manera que la segona base de dades contindrà els següents atributs:

1. ***name***: el nom de pila de l'usuari que ha realitzat el tweet.
2. ***username***: el nom d'usuari.
3. ***time***:Data i *timestamp* del tuit. Hora exacta en que es va publicar el tuit
4. ***tweets***:El cos del tuit/contingut del tuit.
5. ***sentimentLabel***: atribut que indica si el tuit té contingut violent extremista(1) o no(0)

5.2 Anàlisi de la comunitat entre els fanboys de ISIS

Bé, la primera part d'aquest treball consisteix en fer un anàlisi de la comunitat de persones que trobem a partir de la primera base de dades. Per a fer això ens hem valgut de gràfiques i representacions per a visualitzar les dades.

5.2.1 Anàlisi de retuits i mencions

El primer que és com es propaguen els missatges entre aquesta comunitat. A *twitter* a l'hora de visualitzar un tuit tens l'opció de donar-li al botó de *retuit*. Clicant aquest botó el que pasa es que es copia el tuit i es publica desde el teu compte un nou tuit que conté el tuit *retuitejat* i una referència al tuit original. Aquesta és la principal manera de difondre tuits, que al final son missatges , a *twitter*.

Abans de fer res hem d'analitzar quins són els objectius d'Estat Islàmic(ISIS). Els objectius d'aquest col·lectiu son com en el cas de qualsevol grup terrorista amedrentar i terroritzar a les persones. En específic com a objectiu tenen el mon Occidental, que segons ells els ha fet tant de mal. Aquest terror l'han aconseguit provocar durant molts anys mitjançant atemptats duts a terme en vàries localitzacions geogràfiques de tot el món i en diferents moments. Es varen produir atacs a París el Novembre del 2015 o, a Barcelona el 17 d'agost de 2017. Per a conseguir que hi hagin persones capaces de perpetrar semblants

actes de terror, completament violents i inhumans, a tot el mon Occidental es necessita d'una gran comunitat. Aquesta comunitat cada cop ha anat creixent més degut en gran mesura per l'auge de les xarxes socials i les eines de difusió que aquestes proporcionen. Eines com el retuit , que en aquest cas permet que molts més usuaris vegin un determinat tuit.

Dit això, en el cas que ens pertoca observem que nosaltres tenim una base de dades de tuits d'odi que conté 17410 tuits. El que volem comprobar és quants d'ells són realment tuits i quants d'ells són retuits d'altres tuits ja publicats. Veiem en la gràfica de la Fig. 3 com la majoria de dades que trobem a la base de dades són realment tuits, però cal recalcar la gran quantitat de retuits que hi trobem. Tenim 11574 tuits reals per 5836 retuits. Si en fem una proporció d'això podem veure que aproximadament per cada 2 tuits en tenim 1 retuit, amb lo qual la tercera part de totes les dades es correspon a retuits. Així doncs això reafirma el caràcter sectari i adoctrinador d'Estat Islàmic.

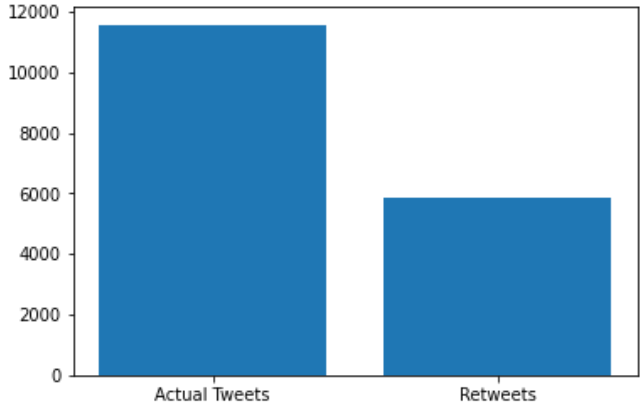


Fig. 3: gràfica on podem observar la quantitat de tuits reals(*Actual Tweets)* en front a la quantitat de retuits en la base de dades de Fanboys de ISIS.

Un cop fet això anem a examinar quins són els autors originals d'aquests tuits. A l'atribut ***tweets*** si el cos és un retuit d'un altre tuit ho tenim indicat a l'inici del tuit amb les lletres ***RT*** . Seguit d'aquestes lletres, justament després tenim una menció a l'usuari que va publicar el tuit original. Aquesta menció té aquest format: ***@usuarituitoriginal*** .
Sabent tot això i tenint en compte que les mencions no tenen perquè anar acompanyades d'un retuit, és a dir que es poden fer mencions també en tuits propis, anem a seguir aquest rastre d'aquestes mencions per a arribar a aquests usuaris. Fent aquest anàlisi obtenim els següents resultats:

- En primer lloc veiem que la majoria de les mencions que es fan son a usuaris que son fora del dataset. Això en principi és simplement una dada curiosa, no obstant si que podem arribar a pensar que al haber-hi tantes mencions a usuaris fora del dataset que els usuaris més influents són aquells que menys publiquen. Ja que en un recull de més de 17000 tuits la majoria de les mencions són a usuaris fora d'aquesta base de dades de tuits. Hem obtingut que de les 12384 mencions només 1619 d'ells son a usuaris presents al dataset, i la resta 10765 són a usuaris fora del dataset. Si filem més prim i de cadascuna d'aquestes dues estadístiques agafem només els valors únics obtenim la gràfica de l'esquerra de la Fig. 4 on veiem que tenim només 64 usuaris que estan al dataset i són referenciats, en front als 3263 usuaris que són referenciats i no són al dataset.

- En segon lloc el que també és interessant analitzar els noms d'usuari que apareixen en les mencions. El que hem fet és veure quants dels usuaris mencionats que es troben al dataset hi ha respecte el total d'usuaris del dataset. Veiem com 64 dels 112 usuaris únics que tenim al dataset apareixen en les mencions. Gràficament això ho podem veure també a la Fig 4. al gràfic de la dreta.

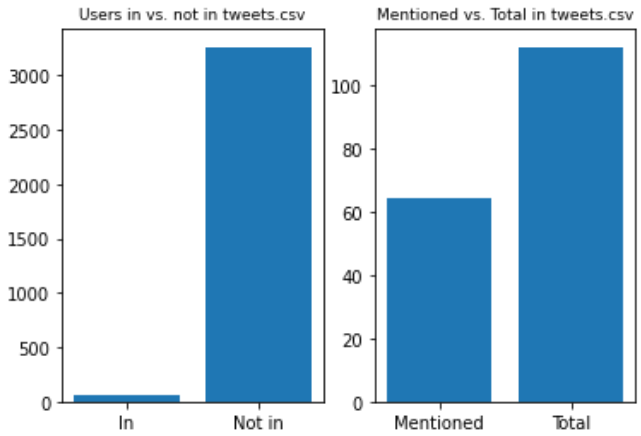


Fig. 4: tenim dos gràfiques, la de l'esquerra correspon al nombre d'usuaris mencionats que es troben al dataset(*in*) en front als que no hi són(*not in*) i a la segona gràfica podem veure el nombre d'usuaris mencionats que són al dataset(*Mentioned)* en front al nombre d'usuaris totals del dataset(*Total*).

5.2.2 Anàlisi de clusters i grups

El següent anàlisi que s'ha fet sobre aquesta base de dades de seguidors de l'Estat Islàmic és el que fa referència a les relacions que hi ha entre els usuaris de la base de dades i no solament de la base de dades sinó que també amb altres usuaris que són mencionats en tuits de la base de dades pero que no hi apareixen ja que no tenen participació directa amb els seus tuits.

Per a fer aquest anàlisi el que hem fet és definir 3 categories d'usuaris:

- **usuaris que fan mencions**
- **usuaris que reben mencions**

- **usuaris que fan i reben mencions**

Per mostrar la comunitat que tenim a partir d'aquesta base de dades hem creat un graf que representa aquesta comunitat i on hi tenim nodes de tres tipus d'acord a les categories d'usuaris ja definides:

- nodes **vermells**, que seran aquells usuaris que només mencionen
- nodes **blaus**, que seran aquells usuaris que només reben mencions
- nodes **verds**, seran aquells usuaris que fam i reben mencions

Amb tot això obtenim el graf de la Fig. 5 on podem observar la comunitat que podem representar a partir dels usuaris i les mencions d'aquest dataset d'Estat Islàmic.

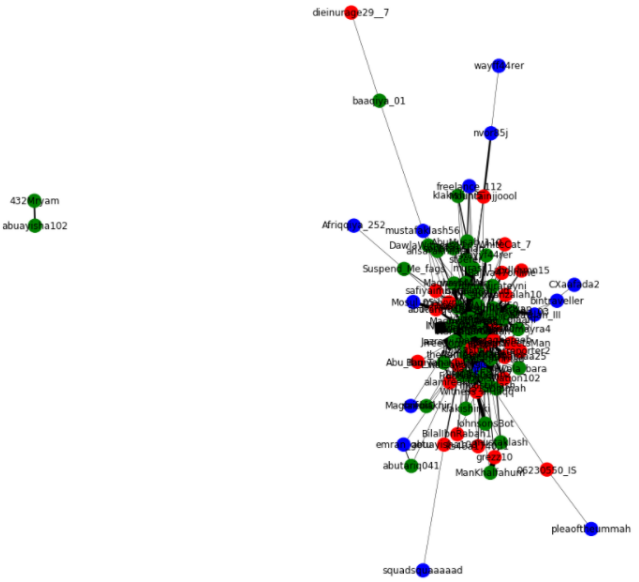


Fig. 5: graf que representa la comunitat d'usuaris simpatitzants d'ISIS que es pot construir a partir del dataset. Cadascun dels nodes és un usuari i el color defineix la seva interacció amb la resta d'usuaris. Veure 5.2.2.

Un cop vista aquesta comunitat a partir del graf el que volem saber és quins són els usuaris que més mencionen i els usuaris que són més mencionats tant de dins del dataset com de fora. Això és el que podem observar a la Fig. 6 on veiem una captura en que es veu precisament el comentat.

```
Top 3 of mentioned users present on the dataset:  [['RamiALLolah', 565], ['Nidalgazaui', 336], ['WarReporter1', 135]]
Top 3 of mentioned users not present on the dataset:  [['WarReporter1', 121], ['7layers_', 116], ['ScotsmanInfidel',
79]]
Top 3 of mentioner users present on the dataset:  [['mobi_ayubi', 393], ['warrnews', 203], ['WarReporter1', 74]]
Top 3 of mentioner users not present on the dataset:  [['Uncle_SamCoco', 1548], ['mobi_ayubi', 685], ['RamiALLolah',
652]]
```

Fig. 6: captura de pantalla d'un output de consola on podem observar els top 3 d'usuaris més mencionats i més mencionadors presents i no presents en el dataset.

5.3 ANÀLISI DE LA TEMPORALITAT DE TUITS D'ISIS VS TUITS SOBRE ISIS

Una altra cosa interessant a analitzar és la temporalitat dels tuits en cadascuna de les dues bases de dades que composen la base de dades per a la classificació. És interessant veure la temporalitat de les publicacions perquè ens dóna informació no solament temporal sinó que també dóna informació sobre la localització des d'on es fa.
Per a simplificar l'anàlisi ens hem limitat a analitzar el rang horari en que es van fer les publicacions, de manera que hem establert que totes elles es van fer el mateix dia, ja que la temporalitat més enllà del rang horari no ens aporta informació sobre la ubicació dels usuaris al moment de publicar.

Dit tot això podem observar les gràfiques de les figures 7, 8 i 9. En la primera d'elles observem les proporcions de tuits publicats segons zona horaria i sentiment(informatiu(0), odi(1)). Veiem com és cridaner el fet de que els tuits informatius d'ISIS es troben concentrats en la primera meitat del día, on d'altra banda hi trobem molts pocs tuits de fanboys d'Estat Islàmic. D'altra banda en la segona meitat del dia, els tuits informatius pràcticament desapareixen i és on hi ha la major concentració de tuits d'odi. Això no només mostra la diferencia en la temporalitat dels tuits d'una i altra base de dades sino que sabent que els temps estan representats en la mateixa zona horaria podem observar que les publicacions es realitzen desde dos localitzacions molt separades entre si.
I si filem més prim podem veure en la Fig. 8 com els tuits informatius es publiquen en la seva gran majoria(més del 50% d'aquests tuits) entre les 8 i les 13 hores. De la mateixa manera a la Fig. 9 podem observar com més de la meitat de tuits publicats per fanboys de ISIS(pràcticament el 60%) es publiquen entre les 15 i les 24 hores.

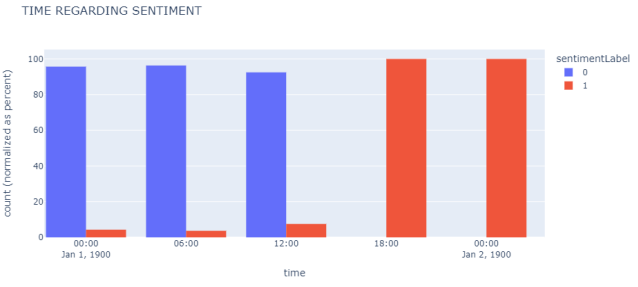


Fig. 7: gràfica on es veu la quantitat de tuits d'una i altra base de dades segons el moment de la publicació i el sentiment de la mateixa.

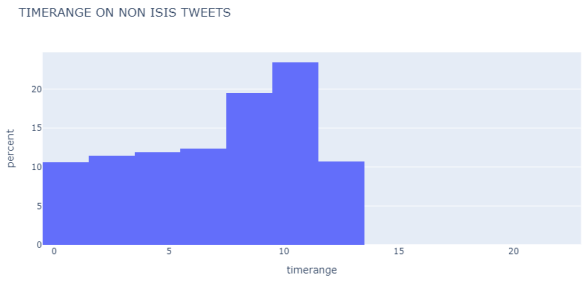


Fig. 8: gràfica que mostra el percentatge de tuits informatius sobre ISIS segons el moment de la publicació

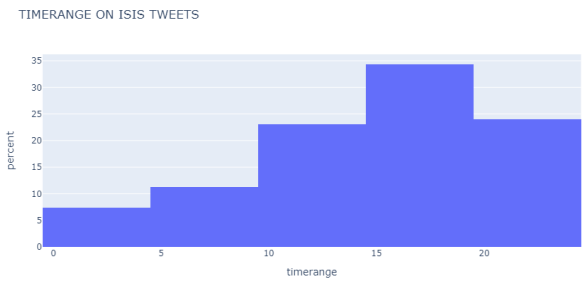


Fig. 9: gràfica que mostra el percentatge de tuits de fanboys d'ISIS segons el moment de la publicació

5.4 Classificació i detecció de missatges violents d'ISIS amb un classificador Naive Bayes i un model de característiques de Bag-of-words

La idea de fer servir aquesta aproximació és demostrar que amb un model simple en el que bàsicament tenim un classificador bayesià que es basa en la idea de independencia condicional ; és a dir que la aparició d'una determinada paraula x no depèn d'una paraula y. Això, tot i no ser cert, permet simplificar molt el problema reduint en gran manera les taules de probabilitats i com veurem més endavant no té un efecte negatiu important en el rendiment. D'altra banda també destacar com a model de característiques l'aproximació Bag-of-words que és bàsicament un diccionari amb dues pàgines on a la primera pàgina hi tenim les paraules que apareixen en tuits la base de dades informativa d'ISIS, per tant que tenen sentiment 0, i el seu nombre d' aparicions i una altra pàgina amb la informació anàloga per la base de dades de tuits de fanboys d'Estat Islàmic.

Cal destacar que aquesta implementació del classificador Bayesià ha estat feta per mi seguint el següent esquema:

1. Carregar dades, barrejar-les i fer particions(train i test)
2. Crear diccionari
3. Realitzar les prediccions
4. Validació

6 Resultats de la classificació i anàlisi

Abans de començar amb l'anàlisi dels resultats cal recalcar el fet que s'ha hagut de balancejar el dataset final ja que tenim moltes més dades de tuits informatius d'ISIS que no pas de tuits de fanboys de l'Estat Islàmic. Per tant el que hem fet és construir un dataset totalment equilibrat on les probabilitats de trobar un tuit informatiu sobre ISIS(amb *sentimentLabel 0)* o un tuit d'un fanboy d'ISIS són les mateixes.

Un cop dit això anem a analitzar els resultats, per tal de assegurar-nos que els resultats fossin vàlids el primer que hem fet és que a l'hora de balancejar els datasets les dades que agafem per construir el nou dataset de tuits informatius a mida siguin agafades de forma aleatòria(*sampling*) del dataset original. Tanmateix a l'hora de fer particions de train i test hem fet un *shuffle*(o barreja) previ de les dades per evitar la localitat d'aquestes i que les particions de train i test tinguin sempre les mateixes dades.

Fet tot això només cal validar els resultats del nostre classificador. En aquest cas no ho hem fet amb un mètode tradicional com la *k-fold cross-validation* sinó que hem volgut veure com varien els resultats segons 2 hiperparàmetres claus en aquesta classificació:

- La mida del diccionari
- La mida del conjunt de train(al variar la mida del conjunt de train varia també obviament la del conjunt de test)

S'han avaluat els resultats segons les 4 mètriques bàsiques següents:

- Accuracy
- Recall
- Precisió
- Especificitat

Així doncs podem veure els resultats que hem obtingut variant la mida del diccionari són el que s'esperaria a la Fig. 10. Totes les mètriques han pujat el seu rendiment substancialment a mesura que s'ha anat augmentant la mida del diccionari. El recall no obstant això ja partia d'un punt molt alt, indicador de que el classificador és capaç de classificar tuits informatius (0) molt bé inclús amb "poques" dades.

En la mateixa línia de resultats estem en el cas de la Fig.11 on fixem la mida del diccionari al 80% i anar variant la mida del conjunt de train. Veiem que totes les mètriques són excel·lents fins i tot quan la mida del train és molt petita. Això indica que la mida del train en aquest cas incideix menys en el rendiment del classificador que no pas la mida del diccionari.

En qualsevol cas el que si que podem afirmar es que el classificador té un rendiment excel·lent i això probablement és degut a la qualitat de les dades que tenim. És a dir, és fàcil de distingir quins tuits són

informatius i quins són ofensius ja que uns i altres presenten grans diferències que el classificador és capaç de detectar mitjançant les paraules.

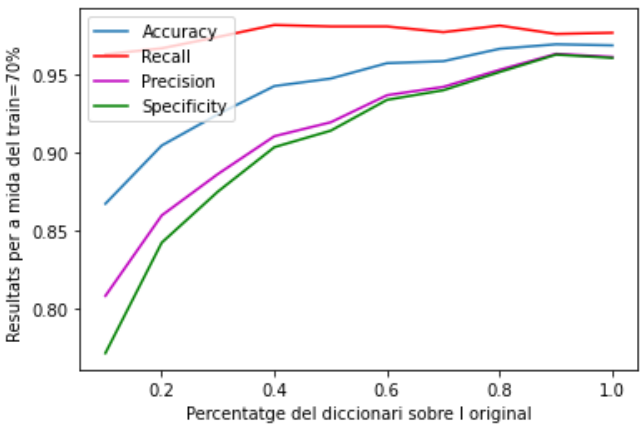
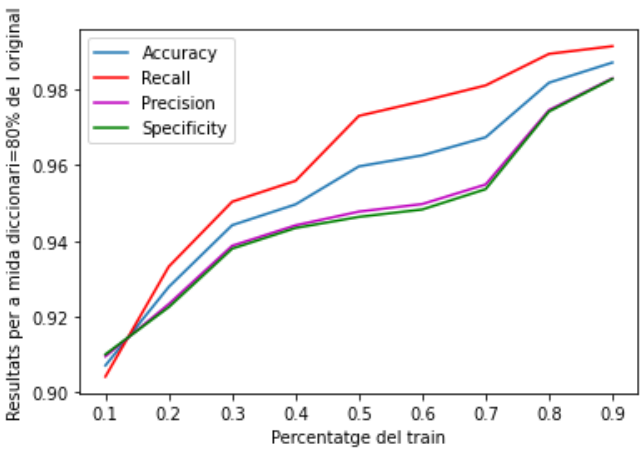


Fig. 10: gràfica on podem observar el rendiment del nostre classificador a mesura que varia la grandària del diccionari mantenint la mida del conjunt de train al 70%.



7 Conclusions

A destacar d'aquest treball podem dir que el col·lectiu que conforma Estat Islàmic és un col·lectiu sectari que busca l'adoctrinament de persones amb fins malèvols, i amb l'única intenció d' atemorir Occident i provocar el terror. Per a crear aquesta comunitat d'extremistes i radicals ISIS recurreix a les xarxes socials on hi troba l'aparador i les eines perfectes perquè persones vulnerables caiguin presos de l'adoctrinament. Twitter és una d'aquestes xarxes socials, i característiques pròpies de la mateixa com les mencions o els retuits permeten a aquests grups violents propagar el seu missatge de forma molt ràpida i eficaç. Un missatge que cala sobretot a l'Orient Mitjà com hem pogut veure a partir de l'anàlisi de la temporalitat en la quan es fan les publicacions. Per últim afegeixo i he demostrat que amb un model de classificació d'allò més simple i unes dades de gran qualitat, que generen un context adequat podem detectar aquest tipus de publicacions violentes i extremistes que tan de mal fan a la societat i que per desgràcia els últims anys han alimentat a persones que després han acabat cometent innombrables atrocitats com els atemptats de Barcelona o París.

8 Agraïments

Com a comiat a aquest treball m'agradaria recalcar la labor del meu tutor, el Jordi que m'ha ajudat una barbaritat sobretot a l'hora de decidir cap a on havia d'encarar el treball i que sempre ha tingut molta paciència amb mi. El mínim que puc fer es mostrar el meu agraïment per haver-me ajudat així que moltes gràcies Jordi.

9 Bibliografia

[1] Muhammad Asifa, Atiab Ishtiaqa, Haseeb Ahmada, Hanan Aljuaidb, Jalal Shahc,Sentiment analysis of extremism in social media from textual information, 2020

[2] Imane Guellil, Ahsan Adeel ,Detecting hate speech against politicians in Arabic
community, 24 de Novembre de 2019

[3] Safa Alsafari, Samira Sadaoui , Malek Mouhoub ,Hate and offensive speech detection on Arabic social media, 2020

[4] Marzieh Mozafari, Reza Farahbakhsh, Noel Crespi,Hate speech detection and racial bias mitigation in social media based on BERT model, 27 d'Agost de 2020

[5]Fifth tribe, How Isis Uses Twitter, Kaggle
How ISIS Uses Twitter | Kaggle

[6] ActiveGalaxy, Tweets Targeting Isis
Tweets Targeting Isis | Kaggle

[7] Alexandra Schofield, Thomas Davidson,Identifying Hate Speech in Social Media, Hivern 2017