Refinament de l'anàlisi del sentiment per xarxes neuronals mitjançant tècniques de PLN

Treball final de grau

Miquel Àngel Fraire
Grau d'enginyeria informàtica
Intel·ligència artificial

Dr.David Isern Alarcón

8 de juny de 2021

Aquesta obra està subjecta a una llicència de
Reconeixement-NoComercial-
SenseObraDerivada 3.0 Espanya de Creative
Commons

FITXA DEL TREBALL FINAL

Títol del treball:	*Refinament de l'anàlisi del sentiment per xarxes neuronals mitjançant tècniques de PLN*
Nom de l'autor:	*Miquel Àngel Fraire Ferrer*
Nom del consultor/a:	*David Isern Alarcón*
Nom del PRA:	*Carles Ventura Royo*
Data de lliurament (mm/aaaa):	*06/2021*
Titulació o programa:	*Grau d'enginyeria informàtica*
Àrea del Treball Final:	*Intel·ligència artificial*
Idioma del treball:	*Català*
Paraules clau	*anàlisi del sentiment, xarxes neuronals, processament del llenguatge natural*

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

En l'anàlisi del sentiment hi conflueixen tant models d'aprenentatge automàtic com tècniques bàsiques de processament del llenguatge natural. Aquestes darreres solen aplicar-se fonamentalment a adaptar el corpus textual d'entrada al tipus de dades que exigeix el model, així com per reduir-ne la dimensionalitat sense perdre informació rellevant, fet que permet a l'algorisme classificador guanyar eficiència.

Amb tot, un coneixement més profund de les particularitats lingüístiques del corpus permetria d'emprendre accions que fessin incrementar també el grau de precisió en els resultats de l'anàlisi del sentiment. Particularment, l'etiquetatge morfosintàctic per tal d'assignar més pes a aquells mots amb més càrrega semàntica durant l'entrenament de l'algorisme hauria de traduir-se en una millora dels resultats de classificació.

Amb aquesta finalitat s'han seleccionat tres corpus textuals diferents, se'ls ha sotmès a les operacions habituals en un exercici d'anàlisi del sentiment i se n'han mesurat els resultats. Tot seguit i després d'haver aplicat als corpus originals la tècnica abans esmentada, s'ha tornat a entrenar el model. Posteriorment s'ha procedit a incrementar la precisió en l'etiquetatge morfosintàctic mitjançant tècniques de reconeixement i correcció del llenguatge informal. Finalment s'ha fet el càlcul de distribució de freqüències

per tal de conèixer les diferències de pes semàntic entre corpus, les quals poden explicar resultats incomprensiblement divergents o allunyats, i eventualment corregir aquestes desviacions

La conclusió és que el mètode d'assignació de més pes a mots semànticament rellevants mitjançat l'etiquetatge morfosintàctic és especialment eficaç en corpus tematitzats sense paraules de freqüència totalment esbiaixada cap a una categoria en particular.

Abstract (in English, 250 words or less):

Both machine learning and basic natural language processing techniques come together in the sentiment analysis. The latter are usually applied mostly to adapt the input text corpus to the data types the model requires, as well as to reduce its dimensionality without losing relevant information, which allows the classifying algorithm to gain efficiency.

However, a deeper knowledge of the linguistic particularities of the corpus would allow to carry out some actions in order to increase the degree of accuracy in the sentiment analysis results. Particularly, using the technique of Part-of-Speech tagging to assign more weight to those words with more semantic load during the algorithm training, should turn into improved ranking results.

To that end, three different textual corpora have been selected and subjected to the usual operations in a sentiment analysis exercise, and then the results have been measured. Then, after applying the aforementioned technique to the original corpora, the model has been trained again. Subsequently, the accuracy of Part-of-Speech tagging has also been increased through techniques for recognizing and correcting informal language. Finally, the word frequency distribution has been calculated in order to know the differences in semantic weight amongst corpora, which can explain an incomprehensibly divergent or distant outcome, and occasionally correct these deviations.

The conclusion is that the method of assigning more weight to semantically relevant words by means of Part-of-Speech tagging is especially effective in themed corpora not containing words of totally biased frequency towards a certain category.

Índex

1.Introducció
- 1.1 Context i justificació del treball 1
- 1.2 Enfocament i mètode seguit 2

2.Recerca i tria 8
- 2.1 Models per a l'anàlisi del sentiment 8
  - 2.1.1 Les xarxes neuronals 12
- 2.2 Tècniques per al processament del llenguatge natural 14
  - 2.2.1 Anàlisi textual i tractament de corpus 14
  - 2.2.2 L'etiquetatge morfosintàctic 16
  - 2.2.3 Càlcul de distribució de freqüències 18

3.Instal·lació de les eines necessàries 19
- 3.1 Per a la creació de xarxes neuronals centrades en l'anàlisi del sentiment 19
- 3.2 Per al processament del llenguatge natural 20

4.Evolució del projecte: fase 1 22
- 4.1 Cerca i tria de les dades 22
- 4.2 Anàlisi preliminar de les dades 22
- 4.3 Preprocessament bàsic de les dades 24
- 4.4 Vectorització de les dades lingüístiques 24
- 4.5 Creació de la xarxa neuronal per a l'anàlisi del sentiment 24
- 4.6 Entrenament del model amb els jocs de dades 25
- 4.3 Descripció de les proves efectuades i avaluació de resultats 25

5.Evolució del projecte: fase 2 28
- 5.1 Eliminació del soroll informatiu 28
- 5.2 Etiquetatge morfosintàctic 29
- 5.3 Distribució de freqüències 33
- 5.4 Descripció de les proves efectuades i avaluació de resultats 37

6 Conclusions i futures línies d'investigació 39

Annex A: Relació de lliurables 40

Annex B: Taules de resultats 41

Bibliografia 71

1.Introducció

1.1 Context i justificació del treball

L'anàlisi del sentiment és una subdisciplina a cavall de l'aprenentatge computacional i del processament del llenguatge natural, que malda per classificar les manifestacions humanes segons que aquestes expressin opinions favorables o desfavorables envers un determinat domini. Atesa l'enorme versatilitat d'aquesta manifestacions i, en conseqüència, la impossibilitat d'establir unes regles de construcció que determinin la semàntica de totes les produccions possibles, s'imposa l'anomenat aprenentatge supervisat, el qual a partir d'una mostra significativa d'observacions prèviament classificades, és capaç d'entrenar algorismes que prediguin amb un grau de precisió acceptable la classificació de noves observacions. Arribats a aquest punt, les *xarxes neuronals* són un model matemàtic o computacional de processament de dades que s'ha revelat especialment addient per a l'entrenament supervisat. D'altra banda, com que les opinions solen expressar-se verbalment, el *processament del llenguatge natural* pren una rellevància fonamental a l'hora de preprocessar i analitzar les dades.

En gran part, l'eficàcia de l'aprenentatge supervisat depèn del format i qualitat de les dades d'entrenament. Si bé avui dia l'enregistrament automatitzat ens proveeix una quantitat immensa d'informació en molt àmbits, no tot aquest volum és útil ni aprofitable per a un algorisme classificador, i la part que no ho és contribuiex a minvar-ne tant l'eficacia com la precisió. En conseqüència, cal sotmetre les dades a un tractament previ o preprocessament per reduir-ne la dimensionalitat, això és eliminar tot allò superflu i deixar, en la mesura del possible, el que és útil per a la classificació que es pretén. Exemples d'això són els processos bàsics de neteja prèvia per unificar formats o assumir valors absents, així com tècniques específiques de reducció de la dimensionalitat com ara l'anàlisi de components principals o la descomposició en valors singulars, molt emprats en l'anàlisi estadística i la mineria de dades.

D'altra banda, el format amb què es presenten les dades també imposa certes limitacions, i en el cas que ens ocupa s'han de tenir en compte les característiques propies no només del llenguatge natural (tan divers com ambigu), sinó del registre i grau d'informalitat propis de les opinions expressades en mitjans socials. La disciplina del *processament del llenguatge natural* o PLN ha desenvolupat tècniques d'anàlisi lingüística que poden ser determinants a l'hora de preprocessar adequadament les

dades d'entrada: l'eliminació del soroll informatiu, el reconeixement de les especificitats pròpies del llenguatge en mitjans socials i l'etiquetatge morfosintàctic o *part-of-speech tagging* per destacar aquells mots semànticament rellevants, són les mesures que de bon principi creiem adequades per millorar la qualitat de les dades d'entrada i, amb aquestes, del rendiment i eficàcia de l'algorisme classificador. En aquest sentit, considerem pionera –i inspiradora del present treball– la línia d'investigació iniciada l'any 2017 per un equip d'investigadors que fent ús de l'etiquetatge gramatical va aconseguir d'incrementar la precisió d'un algorisme per a la generació automàtica de subtítols a partir de la descripció verbal d'imatges[1], així com els esforços adreçats a comprendre les especificitats del llenguatge informal a les xarxes socials per part d'analitzadors automàtics[2] (vegeu l'apartat de *Bibliografia*).

1.2 Enfocament i mètode seguit

Basem el nostre treball en la hipòtesi que l'anàlisi i el tractament de dades expressades verbalment mitjançant tècniques de processament del llenguatge natural pot optimitzar l'eficàcia, la precisió i el rendiment en els models d'aprenentatge computacional que pretenen classificar i predir el sentiment[3]. Per tal de provar aquesta hipòtesi, hem estructurat el procés en dues fases clarament diferenciades:

En primer lloc es procedirà a la implementació d'una xarxa neuronal capaç de classificar i predir el sentiment en opinions expressades verbalment en mitjans socials. Caldrà doncs (a) escollir un conjunt de dades que aplegui aquest tipus d'informació, (b) un sistema de vectorització de dades lingüístiques i una arquitectura de xarxa neuronal addients tant al volum i tipus de dades com a l'objectiu proposat i (c) comprovar-ne la precisió. La tria del model de classificació i predicció es fonamentarà en una recerca sobre l'estat de l'art en *anàlisi del sentiment*, però en aquest estadi el tractament previ de les dades serà el mínim imprescindible perquè l'algorisme les pugui processar.

---

[1] "Different from previous works focusing on visual representation extraction, we consider from another perspective using the Part-of-Speech (PoS) tag of each word as an information cue to enhance the learning of LSTM model, which is popular in natural language processing, but never used in image captioning" (He, 2017: 1).

[2] Vegeu: **Thede** (1998), **Nakawaga**, **Kudoh** & **Matsumoto** (2001), **Moré**, **Climent**, **Oliver** & **Taulé** (2005), **O'Connor *et alii*** (2013) i **Derczinski** (2013).

[3] Plantejaments similars, si bé des d'òptiques i metodologies força diferents, han estat seguits per **Das** & **Balanbataray** (2014) i per **Khong, Soon**, **Goh & Haw**, SC. (2018).

En segon lloc es procedirà al tractament previ de les dades aplicant-hi tècniques més avançades de processament del llenguatge natural, de manera que la dimensionalitat d'aquestes pugui quedar significativament reduïda en eliminar-ne el soroll informatiu i la seva qualitat informativa convenientment realçada per haver tingut en compte les característiques pròpies del llenguatge emprat en mitjans socials, així com per haver-ne destacat els mots semàntic ament rellevants. L'objectiu darrer és refinar la capacitat de l'algorisme a l'hora de classificar i predir les dades. Caldrà doncs (a) triar les tècniques de filtratge textual i etiquetatge morfosintàctic addients al domini d'aplicació, (b) implementar-les en la fase adequada, (c) preparar l'algoritme perquè pugui treballar amb dades vectoritzades bo i considerant el plus d'informació gramatical que aporten, i (d) tornar-ne a testar la precisió. La tria de les tècniques de processament del llenguatge natural per aplicar al damunt de les dades es fonamentarà en una recerca sobre l'estat de l'art en PNL. Com que les múltiples tècniques que en formen part permeten un tractament i descripció incrementals de les dades lingüístiques segons que es posi el focus en l'ortografia (mer reconeixement de la forma), la morfologia (formació i tipus de mots), la morfosintaxi (classificació desambiguada de mots), la sintaxi (funció dels elements en el si de la frase) o la semàntica (significat i context), hem concebut aquesta segona fase com un procés cíclic d'aplicació, adaptació i prova que ens ha de permetre comprovar quines d'aquestes tècniques, en quin grau i en quin sentit tenen un efecte rellevant en el comportament final de l'algoritme. Aquest *modus operandi* assegura una certa tolerància a possibles reorientacions davant de l'obtenció de resultats inesperats.

2.Recerca i tria

Aquesta apartat conté la relació, justificació i breu explicació teòrica dels models i tècniques que s'han considerat apropiats per assolir els objectius del present treball en les dues fases més amunt esmentades. La implementació concreta, així com els problemes i consideracions que han sorgit de l'aplicació d'aquestes tècniques i models es descriuran als propers apartats 4 i 5, en què s'explica l'evolució del projecte.

2.1 Recerca i tria de models per a l'anàlsi del sentiment

Les tres tasques fonamentals per tal de dur a terme un exercici d'anàlisi del sentiment són: (i) el *preprocessament* i *tokenització* del joc de dades, (ii) la *vectorització* de les dades lingüístiques i (iii) la creació d'un *model de classificació*. Cada etapa disposa d'un repertori de tècniques pròpies.

(i) *Preprocessament* i *tokenització*[4]: Tot i que de l'anàlisi de cada joc de dades se'n desprèn la conveniència d'aplicar unes tècniques de preprocessament concretes i *ad hoc*, hi ha una sèrie de tècniques bàsiques de preprocessament textual a què cal sotmetre qualsevol cadena de text que es vulgui fer servir per a l'anàlisi del sentiment. Aquestes van adreçades a netejar les dades lingüístiques de qualsevol possible interferència no lingüística o que no tingui valor lingüístic. D'aquesta manera s'intenta impedir que, en el processament d'unes dades lingüístiques, informació aliena a la naturalesa del document, o bé informació redundant, buida de contingut i sense cap valor semàntic, afecti el resultat del processament –en aquest cas, classificació per anàlisi del sentiment–, així com accelerar el mateix procés tot estalviant al model de tractar amb dades innecessàries al nostre objectiu. Així doncs, d'entrada cal (a) convertir totes les lletres a minúscules (atès que la majoria dels llenguatges de programació amb què s'implementen aquests models són *case sensitive*, això és que distingeixen entre lletres minúscules i majúscules com si fossin signes completament diferents, i d'entrada aquesta distinció no afecta la semàntica del missatge), (b) eliminar els possibles caràcters no alfanumèrics (això són altres caràcters que lletres i xifres, inclosos els signes de puntuació) presents a les cadenes textuals i (c) reduir els múltiples espais en blanc (fonamentalment els que serveixen per indicar la frontera entre dos mots consecutius) a un de sol. Amb tot, encara roman el problema de la presència en el text d'errors d'escriptura o transcripció. Algunes actuacions que es poden fer al respecte són eliminar una tercera lletra consecutiva igual que les precedents (per exemple en el mot "annnoying", que passaria a ser "annoying"), o bé eliminar lletres soles entre espais (per exemple en el sintagma "way o home", que passaria a ser "way home"), ambdues situacions contràries a les normes gramaticals. Tanmateix, l'adopció de tècniques que malden per corregir el text original ha de tenir sempre en compte en quina llengua ha estat escrit, i de tota manera la casuística dels errors és tan gran que difícilment es pot garantir una correcció precisa i amb profunditat alhora. Finalment, però sempre en la línia de descartar informació sobrera o irrellevant, hi ha l'eliminació d'allò que en anglès s'anomenen *stopwords*, és a dir els mots més habituals d'una llengua que inevitablement són presents a qualsevol manifestació verbal, com ara articles, determinants, auxiliars i connectors i que tenen més càrrega gramatical que no pas semàntica, atès que malgrat la seva importància en la formació d'estructures lingüístiques no aporten pas gaire coneixement ni distinció d'allò de què parla el text.

[4] Vegeu: **Davydova** (2018).

Quant a la *tokenització*, aquesta tècnica separa cadenes de text en unitats lèxiques per a la seva posterior anàlisi. Aquestes poden coincidir amb paraules –cosa habitual quan es tracta d'NLP o *Processament del Llenguatge Natural*–, però també es podrien considerar altres divisions. El principal problema sol consistir a determinar quan acaba una paraula i en comença una altra, atès que en primer lloc no totes les llengües naturals comparteixen unes mateixes regles de formació, i d'altra banda i enllaçant amb aquesta diversitat gramatical, perquè les llengües flexives i encara més les aglutinants tendeixen a incorporar a la base lèxica dels seus mots partícules que en matisen en un grau o major o menor grau el significat. De tota manera, cal no confondre *tokenització* amb *lematització*, que és la tècnica posterior a tokenitzar que se centra a associar totes les possibles realitzacions d'una mateixa paraula amb el seu lema o forma canònica.

(ii) *Vectorització*: Com que els algorismes només poden processar informació numèrica, abans d'entrenar el nostre model de predicció amb el joc de dades caldrà que representem amb format numèric els mots continguts a cada observació. D'aquest procediment, que se'n diu vectorització, n'hem preseleccionat tres tècniques: *Bag-of-Words*[5], *Tf-idf*[6] i *Word2Vec*[7].

En una primera fase, *Bag of words* crea un llistat de tots els mots únics que concorren en un text juntament amb el seu nombre d'ocurrències. És a dir, la relació només inclou cada mot una sola vegada i no pas sempre que hi surt repetit, de manera que la freqüència queda resumida i especificada pel nombre d'aparicions associat a cada mot. Per tant, l'estructura de dada que retorna és el que en alguns llenguatges de programació (com ara *Python* que hem emprat en aquesta pràctica) es coneix com a diccionari, en què una clau única i irrepetible (això serien els mots) està associat a un determinat valor (això serien les aparicions). A partir d'aquest vocabulari de concurrències, *Bag-of-Words* crea un vector d'una dimensió equivalent al nombre de diferents mots trobats, en què cada posició representa un dels mots inventariats al vocabulari. Aleshores associa un vector a cada frase i en quantifica les aparicions dels mots, bo i adjudicant els valors d'aquestes aparicions en la frase a la posicio corresponent en el vector. Per tal com aquesta tècnica tracta amb mots, és convenient d'aplicar-la amb textos degudament netejats i tokenitzats. Les limitacions d'aquesta

[5] Vegeu: **Dubey** (2018).
[6] Vegeu: **Tripathi** (2018).
[7] Vegeu**: Malik** (2018).

tècnica de vectorització són que (1) prescindeix de cap enfocament semàntic o que tingui en compte el context d'aparició dels mots, això és que només considera quants tipus diferents de mots i amb quina freqüència apareixen en un text donat i (2) en documents extensos el vector final resultant pot tenir unes dimensions que en fan difícil el processament.

Tf-idf és la sigla de les expressions angleses *Term Frequency* i *Inverse Data Frequency*, que expliquen i sintetizen en què consisteix aquesta tècnica de vectorització, la qual informa no només de la freqüencia dels mots en un text, sinó també de la freqüència inversa o, dit d'una altra manera, de quins són els mots que menys hi apareixen. D'aquesta forma es posa remei a la primera de les debilitats explicades a l'anterior apartat en referència a la tècnica del *Bag of words* (vegeu també la nota a peu), per tal com solen ser precisament els mots menys freqüents els que caracterizen el sentit d'un text en vistes a la seva classificació semàntica, i no pas al contrari. Així doncs, si la *Term Frequency* calcula la ràtio d'aparicions d'un mot en un text en proporció al nombre total de mots concurrents, la *Inverse Data Frequency* calcula el pes que hi tenen els mots menys freqüents, essent més gran el seu pes associat com més estranya és la seva aparició. El producte d'aquests dos valors dóna com a resultat el vector final *Tf-idf*.

Finalment, *Word2Vec* és un model de vectorització numèrica de dades textuals basat en les xarxes neuronals de diverses capes. Construït a partir d'una arquitectura *Skip-gram* o bé *CBOW*, en tots dos casos té capacitat predictiva, en el primer per tal de predir els mots de context a partir d'un mot central, en el segon per fer-ho a la inversa, això és predir un mot central a partir d'un context donat. Si suposem que l'entrada són els mots que formen part del context, els seus respectius identificadors formen un vector a la capa d'*embedding* (terme matemàtic que designa la inclusió d'un objecte a dins d'un altre) i tots aquests formen una matriu que identifica una filera de mots contextuals. La propera capa s'encarrega de fer la mitjana d'aquesta matriu de vectors i en resulta un nou i únic vector, el qual té per dimensions la mida del vocabulari i com a valors les probabilitats d'aparició dels mots en relació al seu context.

(iii) *Modelització*: Malgrat que alguns models no basats en xarxes neuronals han assolit notables nivells de precisió a l'hora de fer anàlisi del sentiment en corpus textuals, sobretot en aquells de temàtica general i dimensions reduïdes, la increïble proliferació de models basats en xarxes neuronals ha fet de l'aprenentatge profund el camp

d'investigació més prometedor per a l'anàlisi del sentiment. La versatilitat del model neuronal el fa igualment apte per a l'anàlisi del sentiment en textos de temàtica més concreta i especialitzada, així com per a corpus de qualsevol mida, on assoleix bons nivells de precisió[8]. A banda de l'algorisme per a l'anàlisi del sentiment, les xarxes neuronals també són emprades tant en tècniques de vectorització (com ara *Vec2Word*, que és una de les que aplicarem a les dades textuals) com en tècniques avançades de processament del llenguatge natural (com ara el *Part-of-Speech tagging* o etiquetatge morfosintàctic dels mots, que farem servir a la segona fase del treball). Per tal com les xarxes neuronals són especialment indicades per a l'anàlisi del sentiment i seran una constant en l'evolució d'aquest estudi, els algorismes d'aprenentatge computacional que usarem estaran tots basats en el model de xarxa neuronal.

2.1.1 Les xarxes neuronals[9]

Atesa doncs la rellevància de les xarxes neuronals, farem una breu descripció de llur funcionament intern, malgrat que les funcions amb què les implementarem mantinguin la seva estructura transparent a l'usuari.

Una xarxa neuronal artificial és un model de processament de dades que s'inspira en l'estructura del sistema nerviós dels animals, d'on li prové el nom. Les seves unitats de processament, anomenades neurones, estan disposades en xarxa per tal de potenciar col·lectivament la seva capacitat de càlcul, el qual es troba determinat per les connexions entre les diverses neurones i els seus paràmetres. Cada neurona rep uns valors d'entrada, els processa i propaga a d'altres neurones a través de les connexions de sortida. Com que les neurones estan disposades en capes i una xarxa neuronal pot comptar amb diverses capes de neurones, de tal manera que entre les neurones que reben els primers valors a la capa d'entrada i les darreres que emeten un nou valor a la capa de sortida poden haver-hi diverses capes ocultes de neurones, per això aquest model d'aprenentatge se sol qualificar de "profund". Donats uns valors d'entrada i la solució a un problema, del qual se n'espera un resultat però no es coneix el patró

[8] Vegeu: **Vinodhini & Chandrasekaran** (2016), així com també: **Singh** (2018).

[9] Per a una comprensió cabal de les xarxes neuronals, vegeu el capítol 11 de **Hastie, Tibshirani & Friedman (2001)**, font de la qual en gran part ha estat extreta la informació aquí subministrada.

o fórmula per obtenir-los, una xarxa neuronal pot entrenar-se per aprendre a reconèixer automàticament aquests patrons de classificació i predicció.
Així doncs, el valor que rep una neurona situada en una capa intermèdia o bé a la capa de sortida, ha estat prèviament processat per totes les neurones de la capa precedent. El processament a què les neurones sotmeten el valor rebut abans de propagar-lo cap a les neurones d'una capa posterior consisteix a (1) ponderar cada valor d'entrada segons un pes inicialment aleatori, (2) fer-ne el sumatori, (3) afegir-hi un valor de biaix i (4) aplicar-hi l'anomenada "funció d'activació" amb la qual s'obté un valor final proporcional i ajustat a un rang (per exemple: 0 – 1). Quan els valors originals de la capa d'entrada ja han estat processats per cada neurona i successivament propagats cap a les neurones de les capes següents, s'obté una funció de sortida que genera uns valors que cal comparar amb els valors esperats. A partir de les diferències observades, té lloc un reajustament retroactiu dels pesos que cada neurona aplica al processament dels seus valors d'entrada mitjançant funcions d'optimització. Com que aquest procés es mou en la direcció oposada que el processament dels valors d'entrada, s'anomena *retropropagació*. Aquesta seqüència de propagació i retropropagació representa una iteració sencera en el procés d'entrenament de la xarxa neuronal, que finalitza quan la funció de sortida és capaç de replicar, dins un marge acceptable de variació, els valors esperats. En aquest punt, la xarxa neuronal ja està entrenada i per tant té capacitat per classificar i predir observacions aquest cop desconegudes segons el model que ha extret de l'entrenament.

Les xarxes neuronals són molt versàtils pel que fa a l'estructura, atès que teòricament no hi ha limitació ni al nombre capes de la xarxa ni al nombre de neurones per capa. L'experimentació en aquest camp de l'aprenentatge computacional es dedica a cerca models de xarxa neuronal adequats per a la resolució òptima de problemes concrets. Gràcies a l'estructura en xarxa, els models basats en xarxes neuronals són autoescalables i molt tolerants a fallades.

D'entre la tipologia cada cop més extensa de xarxes neuronals, el model que hem escollit com a algorisme de classificació per a l'anàlisi del sentiment és l'anomenat *perceptró multi-capa*, que respon a les sigles en anglès MLP (*multi-layer perceptron*).

2.2 Recerca i tria de tècniques per al processament del llenguatge natural

Més enllà de les tècniques bàsiques abans descrites per al preprocessament de les dades lingüístiques i que tenen com a finalitat la neteja i preparació del corpus per tal que aquest pugui ser processat per l'algorisme classificador amb bones garanties d'èxit tant pel que fa la precisió com l'eficiència, la PLN o *Processament del Llenguatge Natural* ens forneix un repertori de tècniques més avançades que permeten d'analitzar un corpus textual en les diverses vessants amb què hom sol dividir la lingüística descriptiva: ortografia, morfologia, sintaxi i semàntica. El tractament de les cadenes lingüístiques des d'aquestes òptiques o nivells ens hauria de permetre (a) eliminar el soroll informatiu que encara resti al corpus, (b) reconèixer i preveure el llenguatge informal que es fa servir en les xarxes socials, (c) etiquetar morfosintàcticament els mots per tal de discriminar-ne la importància funcional en el si de la cadena de text i (d) valorar semànticament mots i expressions de manera que es prioritzi el pes d'aquells amb més càrrega de significat per a l'anàlisi del sentiment.

La intenció darrera és augmentar la precisió de l'algorisme classificador per la via de depurar les dades lingüístiques que aquest ha de processar i prioritzar el pes dels indicadors més rellevants per sobre dels altres. Algunes de les tècniques que preveiem útils per a l'assoliment d'aquest propòsit són (a) l'*anàlisi textual i tractament de corpus,* mitjançant la qual es poden localitzar mots o construccions estranyes a les regles de la llengua però pròpies de l'escriptura informal en xarxes socials i, eventualment, tractar-les, (b) el *Part-of-Speech tagging* o etiquetatge morfosintàctic, mitjançant el qual es poden classificar els mots segons la seva funció morfològica i a partir d'aquí assignar-los un pes determinat d'acord amb la seva importància com a configuradors del significat en el si d'una frase qualsevol i (c) el *càlcul de distribució de freqüencies*, mitjançant el qual es pot conèixer la presència d'un determinat mot en un corpus textual.

2.2.1 Anàlisi textual i tractament de corpus.

El soroll informatiu i el llenguatge informal present en els corpus textuals no només fan minvar l'eficiència en els algorismes d'aprenentatge automàtic per tal com es veuen forçats a tractar amb una gran quantitat de dades no rellevants, sinó que fan minvar la precisió en els algorismes d'etiquetatge

morfosintàctic[10] com el que aplicarem a la segona fase del present estudi per tal de detectar els mots amb més càrrega semàntica i atorgar-los un pes complementari.

En part, les tècniques de preprocessament aplicades al joc de dades que aplicarem a la primera fase de l'estudi resolen la presència de soroll informatiu en el corpus. Amb tot, la casuística és massa gran per a una neteja planificada amb expressions regulars. Les errades tipogràfiques en són un bon exemple, tot i que en les equivocacions involuntàries les persones solem establir uns patrons. És gràcies a aquestes pautes d'error que els correctors ortogràfics poden fer la seva feina. Aquests solucionen algunes errades tipogràfiques molt habituals del tipus descuit com ara *bihday* per *birthday* (present al nostre corpus), així com les típiques confusions entre lletres que representen un mateix so o un de semblant, però en d'altres ocasions poden malinterpretar la intenció del parlant. Per tant, com que les solucions automatitzades no són perfectes ni definitives i, a més a més, cada corpus textual té les seves particularitats lingüístiques, convé de fer-hi un cop d'ull per saber de primera mà a què ens enfrontem.

Precisament la característica més destacada d'un corpus textual extret d'un mitjà de comunicació social és la informalitat en el llenguatge. Una paraula que pertany a un argot no pot ser tractada com una errada tipogràfica perquè no ho és, per molt que aquesta paraula no sigui oficialment reconeguda. El problema que sovint comporten aquestes paraules és que poden presentar una gran variabilitat pel que fa a l'escriptura[11], per tal com que encara no han estat recollides als diccionaris i, per tant, fixades ortogràficament. Amb tot, si aquest no és el cas que ens ocupa, aleshores tant hi fa que apareguin o no als diccionaris, perquè al nostre analitzador només li interessa la freqüència amb què apareix, en relació amb quines altres paraules apareix i quina etiqueta hi té relacionada. Podem saber quines paraules estranyes apareixen al nostre corpus (ja sigui perquè estan mal escrites o perquè formen part d'un determinat argot)

[10] "Part-of-Speech tagging information is a pre-requisite in many NLP algorithms. However, Twitter text is difficult to part-of-speech tag: it is noisy, with linguistic errors and idiosyncratic style" (Derczinski, 2013: 198)

[11] "An abundance of slang is a characteristic feature of microblog text, and these words are often incorrectly tagged, as well as being rarely seen due to a proliferation of spelling variations (all incorrect)" (*íbid.*: 201)

consultant-les a un dels grans corpus i diccionaris que inclou la llibreria *nltk*. Aquests corpus tenen l'avantatge que no es limiten a aplegar textos, sinó que han estat tractats, netejats i fins i tot etiquetats tant morfosintàcticament com semàntica, per la qual cosa ens forneixen molta informació dels mots d'una llengua. En el cas que ens ocupa, ens serviran com a referent per a un estàndard lèxic. Un cop haguem detectat quines formes presents al nostre corpus textual se n'aparten, aleshores podrem efectuar-hi intervencions de substitució o correcció automàtica basades en un diccionari propi[12].

2.2.2 L'etiquetatge morfosintàctic[13]

L'etiquetatge morfosintàctic o *Part-of-Speech tagging* és una tècnica de la PNL que malda per classificar els mots segons la seva funció gramatical, això és el tipus de paraula. Normalment, aquesta classificació comprèn substantius, adjectius, pronoms, verbs, adverbis, preposicions, conjuncions i interjeccions, però segons quin autor hi poden haver classificacions amb diferent granularitat, per exemple una que distingeixi entre els adjectius qualificatius i els determinants (al capdavall totes dues classes es caracteritzen per acompanyar, referir-se i, depenent del tipus de llengua, fins i tot concordar gramaticalment amb un substantiu), etc. La forma externa d'una paraula ens proporciona molta informació, però no sempre la suficient per determinar-ne el tipus. Dit en altres paraules: una mateixa forma pot tractar-se d'un o altre mot depenent del context en què aparegui. Per exemple, la forma *sol* tant pot ser un substantiu, com un adjectiu, com un verb. Determinar el tipus de paraula per la forma però tenint també en compte el context d'aparició, és el que diferencia l'etiquetatge morfosintàctic de l'anàlisi morfològica, per tal com aquesta es limita a contemplar totes les possibilitats funcionals d'una forma, mentre que l'etiquetatge morfosintàctic desambigua i classifica el tipus de mot.

[12] "Some kind of automatic correction or expanded lexicon could be employed to either map these back to dictionary words or to include previously-seen spelling variations" (*ibid*.)

[13] Per ampliar la informació relacionada amb el contingut d'aquest apartat, vegeu: Oliver, Antoni (2010): "Etiquetatge morfosintàctic", a: *Lingüística computacional*. Barcelona, FUOC.

La majoria d'algorismes que etiqueten gramaticalment les paraules formen part de la gran família de l'aprenentatge supervisat[14]: mitjançant un corpus textual prèviament etiquetat, es pot entrenar un algorisme perquè pugui reconèixer la funció gramatical dels mots en corpus textuals no etiquetats. Com que ja hem dit que no només la forma determina la funció gramatical d'un mot, sinó que també ho fa el context d'aparició, aquests algorismes solen tenir en compte els *n* mots anteriors i posteriors al mot analitzat. L'etiquetador per *unigrames* només té en compte el mot mateix (és a dir, prescindeix del context d'aparició), mentre que l'etiquetador per *bigrames* també té en compte els primers mots immediatament adjacents, l'etiquetador per *trigrames* els dos mots immediatament adjacents, etc. Tot i que a primera vista podria semblar que l'ampliació del context beneficia el reconeixement de la funció gramatical en un mot, amb tot cal tenir en compte que les paraules no apareixen sempre associades de la mateixa manera, de manera que si en tot el corpus d'entrenament una paraula només apareix un cop, o bé n'apareix més d'un però amb associacions diferents, l'etiquetador serà incapaç de veure-hi una pauta i etiquetarà la paraula com a desconeguda. Per evitar aquest problema, se solen combinar els etiquetadors amb una tècnica coneguda com a *backoff*, de tal manera que si l'etiquetador per trigrames (és a dir, fent servir els dos mots immediatament adjacents) no pot establir-ne la funció gramatical, aleshores ho intentarà l'etiquetador per bigrames (és a dir, fent servir els mots immediatament adjacents), i si tampoc no n'és capaç aleshores ho farà l'etiquetador de unigrames, que acabarà etiquetant el mot amb l'etiqueta més freqüent que es dóna en aquest mot en tot el corpus d'entrenament. Com es pot deduir fàcilment, els corpus grans proporcionen un millor entrenament a l'algorisme classificador, per tal com cada mot té més possibilitats d'aparició en contextos diferents. D'altra banda, el text amb els mots per classificar cal passar-lo a l'algorisme degudament tokenitzat. El següent fragment de codi combina els tres tipus d'etiquetador que acabem d'esmentar i entrena l'algorisme amb el corpus *brown* d'NLTK:

[14] Si més no, aquest és el cas dels algorismes basats en l'aprenentatge automàtic, perquè les tècniques de *POS tagging* "es poden dividir en tres grans grups d'acord amb el tipus de coneixement que fan servir (...): coneixement lingüístic, informació estadísitca o tècniques d'aprenentatge automàtic. També es poden trobar sistemes híbrids i també hi ha certs sistemes difícils de classificar" (*íbid.*: 9).

```
from nltk.corpus import brown
tagged_sents=brown.tagged_sents()
unigram_tagger=nltk.UnigramTagger(tagged_sents)
bigram_tagger=nltk.BigramTagger(tagged_sents,
       backoff=unigram_tagger)
trigram_tagger=nltk.TrigramTagger(tagged_sents,
       backoff=bigram_tagger)
tokens=nltk.tokenize.word_tokenize(tweet)
analisi = trigram_tagger.tag(tokens)
print(analisi)
```

Amb aquest etiquetador es poden classificar paraules de frases escrites en anglès, perquè el corpus d'entrenament només aplega textos anglesos. Si en volguéssim crear un per a analitzar paraules de frase escrites en altres llengües, com el català, només hauríem de canviar el corpus d'entrenament. Per descomptat, hauria d'estar prèviament etiquetat.

2.2.3 Càlcul de distribucíó de freqüències

Aquesta tècnica pertany a l'àrea del processament de llenguatge natural anomenada 'anàlisi textual i processament de corpus", i si bé la seva aplicació a millorar la precisió de l'algorisme classificador no és directa, tanmateix el càlcul de distribució de freqüències ens pot ajudar a conèixer millor el nostre corpus de treball, cosa fonamental a l'hora d'implementar un codi *ad hoc* que resolgui els problemes que cada corpus porta implícits. Primer de tot cal distingir entre els conceptes de *token* i *ocurrència*. El primer es refereix a un mot únic i diferenciat d'un altre, mentre que el segon es refereix als cops que apareixen al corpus. Per exemple, la llista ['my', 'taylor', 'is', 'rich', 'and', 'my', 'parents', 'are', 'poor'] conté 9 elements, dels quals 8 són tokens diferents, perquè del mot 'my' hi ha dues ocurrències, això és que apareix dos cops. Així doncs, el càlcul de distribució de freqüències ens diu (a) quants tokens apareixen en un corpus textual, (b) la seva freqüència absoluta o quantes ocurrències hi ha de cada token, (c) la seva freqüència relativa o proporció d'ocurrències envers el total d'ocurrències comptant-hi tots els diferents tokens presents en el corpus i (d) la posició que ocupa en la classificació de mots amb més freqüència.

Arribats a aquest punt, convé d'explicar breument en què consisteix la *Llei de Zipf*. Aquesta descriu la distribució relativa de mots en un corpus textual, la qual és molt similar independentment del tipus de corpus (que pot ser un joc de dades textuals com els nostres, o bé un llibre): "la Llei de Zipf afirma que donat

un corpus, la freqüència d'una paraula és inversament proporcional a la seva posició a la taula de freqüències" (Oliver, 2010: 39). Això vol dir que en un corpus textual qualsevol, molt pocs tokens acaparen la majoria d'ocurrències, mentre que una gran quantitat de tokens hi tenen molt poques o una sola ocurrència en tot el corpus.

3.Instal·lació de les eines necessàries

Per a la consecució del present estudi hem triat el llenguatge de programació *Python*, atès que disposa de potents llibreries que faciliten tant (a) la preparació de corpus textuals i la creació de xarxes neuronals per a l'anàlisi del sentiment en què se centra la primera fase del treball, com (b) el processament del llenguatge natural en què se centra la segona fase del treball.

Amb la intenció de poder treballar en un entorn local, hem instal·lat la darrera versió (3.9.4) de l'intèrpret de *Python* en un ordinador, juntament amb l'IDE *Sublime Text 3* per tal d'editar-ne el codi en un entorn còmode. Els enllaços d'aquests dos programes són els següents:

<https://www.python.org/>

<https://www.python.org/downloads/>

<https://www.sublimetext.com/>

Amb tot, com que també ens ha calgut la potència de processament d'un entorn remot, en diverses ocasions també hem fet servir l'eina *Google Colab*, per tal com permet de codificar en *Python* i qualsevol llibreria que poguéssim necessitar ja hi és instal·lada. L'enllaç d'aquesa plataforma és el següent: https://colab.research.google.com/notebooks/intro.ipynb

3.1 Per a la creació de xarxes neuronals centrades en l'anàlisi del sentiment

En aquest cas, les llibreries necessàries cobreixen els sis passos obligats a l'hora d'efectuar una anàlisi del sentiment amb xarxes neuronals, que són: (a) la importació, lectura i manipulació del(s) joc(s) de dades mitjançant la llibreria *pandas*[15], (b) el pre-processament d'aquests jocs de dades mitjançant la llibreria *nltk*[16], (c) la vectorització de les dades mitjançant les llibreries *gensim*[17] i *scikit-learn*[18], (d) la separació de les dades en sengles subjocs d'entrenament (*train*) i comprovació (*test*), (e) la creació i

[15] Per a la documentació de la llibreria *Pandas* de Python: https://pandas.pydata.org/docs/
[16] Per a la documentació de la llibreria *nltk* de Python: https://www.nltk.org/
[17] Per a la documentació de la llibreria *gensim* de Python: https://radimrehurek.com/gensim/
[18] Per a la documentació de la llibreria *scikit-learn* de Python: https://scikit-learn.org/stable/

entrenament de la xarxa neuronal i (f) la generació de mètriques d'avaluació, aquestes tres darreres tasques mitjançant la llibreria *scikit-learn*. Tal com ha estat dit abans, l'entorn de *Google Colab* ja porta instal·lades totes les llibreries que el programador de Python pot necessitar, però si es treballa en entorn local aleshores sí que cal instal·lar les anteriors llibreries. A través del sistema de gestió de paquets *pip* de què disposa *Python*, és extremadament senzill d'efectuar les instal·lacions a través de la línia d'ordres:

```
python –m pip install -- upgrade pip
pip install pandas
pip install nltk
pip install scikit-learn
pip install -- upgrade setuptools
pip install gensim
```

3.2 Per al processament del llenguatge natural

Les eines que ens calen per efectuar les operacions de processament del llenguatge natural previstes per a la segona fase del treball estan gairebé totes contingudes a la llibreria *nltk* (sigles de "natural language tool kit"), que abans ja hem instal·lat per tal com també era necessària per a les tasques de preprocessament del joc de dades textual. Tanmateix, la instal·lació de tots els recursos disponibles en aquesta llibreria demana l'execució d'un parell d'ordres més en una consola de *python*:

```
>> import nltk
>> nltk.data
```

En el transcurs de les operacions previstes a la fase 2, d'aquest enorme repositori de recursos farem servir el corpus textual *brown*[19] del qual acabem de parlar a l'anterior apartat 2.2.1, en aquest cas per localitzar-hi els mots estranys o mal formats.

D'altra banda, per a les operacions de correcció automàtica ens caldrà instal·lar el que en l'argot de la PLN s'anomenen *spelling correctors*. Per a Python n'hi ha diversos de disponibles. Després d'haver-ne comprovat l'eficàcia un per un, hem optat per *autocorrect*. La instal·lació s'efectua amb la següent comanda:

[19] Hem escollit aquest corpus precisament perquè és el mateix amb què s'entrena l'algorisme d'etiquetatge morfosintàctic que farem servir més endavant.

```
pip install autocorrect
```

Per a les operacions relacionades amb la distribució de freqüències necessitarem un mètode per comptar els elements únics d'una llista d'elements repetits. Tot i que la implenentació d'aquesta funcionalitat es pot fàcilment a mà, amb la importació de la classe *Counter* de la llibreria *collections* ens estalviarem algunes línies de codi:

```
from collections import Counter
```

Finalment, la modificació que resol el problema trobat en el mètode per a escalar els valors en el vector de mots (vegeu l'apartat 5.2) requereix també la classe *csr_matrix* de la llibreria *Scipy*:

```
from scipy.sparse import csr_matrix
```

4.Evolució del projecte: fase 1

4.1 Cerca i tria de les dades

Les dades necessàries per a un estudi enfocat en l'anàlisi del sentiment són de tipus textual, preferiblement extretes d'una xarxa social en què els usuaris manifesten les seves opinions sobre un determinat tema. Els dos únics camps indispensables amb què aquestes dades han d'estar disposades són (a) el text en format de cadena que expressi una postura al voltant d'allò que s'està debatent i (b) la classificació d'aquestes postures sintetitzant l'opinió o sentiment de llur emissor, en un format que tant pot ser numèric com textual, però sempre de tipus categòric, és a dir havent restringit les possiblitats de classificació a unes opcions limitades i preestablertes (per exemple: *positiu* o *negatiu*, o bé *a favor*, *en contra* i *indiferent*, etc).

D'entre tots els recursos de dades públiques per a l'anàlisi estadística i *machine learning* disponibles, hem observat que *Kaggle* és la font que allotja més jocs de dades per a l'anàlisi del sentiment, això és que satisfan els requisits suara apuntats. Centrant-nos doncs en aquest lloc, hem escollit fins a tres jocs de dades: a banda d'atènyer els requisits que els fan aptes per a l'anàlisi del sentiment, val a dir que comparteixen el tret que la classificació de les observacions és binària i el llenguatge emprat en les opinions és l'anglès; alhora, mentenen certes diferències pel que fa a la mida, la paritat en l'etiquetatge (és a dir, si hi ha el mateix nombre d'observacions amb una i altra classificació), la temàtica i, com més endavant es revelarà fonamental, la distribució dels mots de determinades categories gramaticals segons l'etiquetatge de les observacions. Les respectives url's tant del repositori com dels tres jocs de dades són les següents:

<https://www.kaggle.com/datasets>
<https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech>
<https://www.kaggle.com/kazanova/sentiment140>
<https://www.kaggle.com/gargmanas/sentimental-analysis-for-tweets>

4.2 Anàlisi preliminar de les dades

El *dataset_1* recull opinions manifestades pels usuaris de la popular xarxa social *Twitter*. El tema al voltant del qual giren aquestes opinions és el de les qüestions racials i de gènere, i tal com s'explica a la descripció del repositori, "label '1' denotes the tweet is racist/sexist and label '0' denotes the tweet is not racist/sexist". El text de les observacions ha estat anonimitzat, de manera que els noms originals dels usuaris, ja

fossin autors de la piulada o bé referències a d'altres usuaris, apareixen genèricament com a @user. El que no ha estat eliminat ni tractat, sinó que s'ha deixat tal qual constava a les piulades originals, són les *tags* o etiquetes encapçalades pel signe # amb què els mateixos usuaris de *Twitter* i altres xarxes socials solen classificar les seves pròpies opinions. El joc de dades consta de tres camps, que són l'*id* o identificador numèric de l'observació, la *label* o categoria en què ha estat classificada l'observació i el *tweet* o opinió textual expressada per l'usuari. La mida és aproximadament 32K amb un biaix pronunciat del 92.89% d'observacions pertanyents a la categoria 1 contra el 7.01% restant d'observacions pertanyents a la categoria 2.

El *dataset_2* també aplega piulades de *Twitter*, però en aquest cas no se n'ha restringit la temàtica. Tal com s'explica a la descripció del repositori, "The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment". El text de les observacions no ha estat anonimitzat, però se n'ha extret el nom de l'autor per poblar un camp propi alhora que s'hi deixaven les referències a d'altres usuaris. A diferència del *dataset_1*, les opinions no contenen cap etiqueta encapçada pel símbol #. El joc de dades consta de sis camps, que són el *target* o categoria en què ha estat classificada l'observació, *l'id* o identificador numèric de l'observació, la *date* o data en què es va emetre la piulada, un *flag* que indica si la piulada és resposta a una altra piulada, l'*user* o nom d'usuari que va fer la piulada i el *text* o opinió textual expressada per l'usuari. L'original a *Kaggle* conté fins a 1.600.000 observacions, això és una mida de 1600K, amb paritat estricta d'etiquetatge. Aquí l'hem reduit a 32K respectant paritat.

Finalment, el *dataset_3* reuneix piulades de *Twitter* seleccionades amb la finalitat, segons es fa saber a la descripció del repositori, de predir "if a person is depressed from their use of words on social media", de tal manera que cada observació ha estat etiquetada en funció de si aquesta havia estat generada per algú diagnosticat amb depressió, on "0 stands for NO and 1 stands for YES". El text de les observacions ha estat anonimitzat, però s'hi han conservat els *nicknames* que referencien les interaccions amb altres usuaris. Les piulades no contenen cap etiqueta encapçalada pel símbol #. El joc de dades consta de tres camps, que són *Index* o un identificador numèric que no és seqüencial i probablement es refereix al número de piulada extreta d'un conjunt molt més gran, *message to examine* o opinió textual expressada per l'usuari i la *label (depression result)* o categoria en què ha estat classificada l'observació. El nombre total de registres és de 10314, dels quals 8000 observacions pertanyen a la categoria 1 (això

és un 77.56% del total) mentre que les 2314 restants pertanyen a la categoria 2 (això és un 22.43% del total), de manera que no hi ha paritat en l'etiquetatge. Més endavant, un cop haguem obtingut els resultats de la fase 1, entrarem al detall de la distribució de mots classificats per categories gramaticals en cadascun dels jocs de dades.

4.3 Preprocessament bàsic de les dades

Per començar, hem reduït la dimensionalitat del *dataset_2* a 32K, que és la mateixa mida que té el *dataset_1*, bo i respectant la paritat original en la distribució de les observacions segons categories. Tot seguit de suprimir per innecessaris els camps diferents del text i la categoria, hem sotmès els tres datasets a les següents tècniques de preprocessament més amunt descrites: (a) conversió de les cadenes textuals a lletra minúscula, (b) eliminació dels caràcters no alfanumèrics, (c) eliminació dels espais sobrers, (d) eliminació del tercer caràcter igual consecutiu, (e) eliminació de les *stopwords*, (f) tokenització i (g) lematització. A més a més, com que tots tres reculls contenen tant url's com noms d'usuari, hem eliminat les primeres i hem anonimitzat els segons. En una de les múltiples proves que hem efectuat per tal de comparar els resultats del model entrenat amb el *dataset_1* amb els resultats del model entrenat amb el *dataset_2*, que no conté tags, hem esborrat els tags del primer.

4.4 Vectorització de les dades lingüístiques

Alternativament, hem fet servir les tres tècniques de vectorització abans descrites amb la intenció de comparar resultats: *Bag-of-Words*, *Tf-idf* i *Word2Vec*. Val a dir que d'aquesta darrera tècnica hem implementat concretament l'anomenada *Doc2Vec*, que n'és una evolució directa.

4.5 Creació de la xarxa neuronal per a l'anàlisi del sentiment

Mitjançant la llibreria *scikit-learn* de *Python*, la creacio d'una xarxa neuronal esdevé una tasca extremadament senzilla: la funció *MLPClassifier* (MLP són les sigles de *Multi Layer Perceptron* o "perceptró multi-capa") sense parametritzar crea un model de xarxa neuronal que per defecte empra l'algorisme d'activació *RELU*, preveu una quantitat de 100 neurones per a les capes ocultes i un nombre màxim de 200 iteracions després de les quals el model atura l'entrenament encara que no s'hagi arribat a la convergència.

4.6 Entrenament del model amb els jocs de dades

Tal com és preceptiu, hem separat els jocs de dades preprocessats en dos subjocs d'entrenament i comprovació en una proporció del 80% contra el 20% restant. En principi s'ha procedit de la mateixa manera tant per al *dataset_1* i el *dataset_3*, en què no hi ha paritat d'observacions pel que fa a les categories, com per al *dataset_2*, on sí hi ha paritat. Tanmateix, aquest excessiu biaix del primer corpus (de 31962 observacions, un 92.98% pertanyen a la categoria 1 per un 7.01% restant que pertanyen a la categoria 2) distorsionava la veritable capacitat de predicció del model, per tal com l'aposta única per una determinada categoria constant té moltes probabilitats de resultar reeixida. Per aquest motiu s'ha procedit a efectuar un segon entrenament del *dataset_1* basat en l'equiparació de les categories, això és seleccionant del subconjunt d'entrenament un altre subconjunt amb la mateixa quantitat d'observacions per a totes dues possibles categories. Semblantment s'ha fet per al *dataset_3*, que també presenta biaix, si bé no tant pronunciat.

4.7 Descripció de les proves efectuades i avaluació de resultats

Cada joc de dades, convenientment preprocessat i tokenitzat (vegeu la secció *i* de l'anterior apartat 2.1), així com vectoritzat mitjançant cadascun dels mètodes de vectorització triats (vegeu la secció *ii* de l'anterior apartat 2.1), ha estat sotmès a tres entrenaments consecutius segons el model escollit de perceptró multicapa o MLP (vegeu l'anterior subapartat 2.1.1) i s'ha calculat la mitjana de les precisions obtingudes, de manera que ens puguem fer una idea del marge de precisió en les diferents metodologies d'entrenament. L'algorisme de classificació entrenat amb *dataset_1* i *dataset_3* sense paritat en la categoria de les observacions (vegeu *Taula 1* i *Taula 3*) ha tingut com a resultat un grau de precisió molt per sobre de la precisió obtinguda pel mateix algoritme entrenat amb el *dataset_2* (vegeu *Taula 2*), el qual originalment ja venia amb paritat en la categoria de les observacions. Aquests resultats són els que hem pres com a referència per ajustar l'entrenament dels altres dos jocs de dades.

Pel que fa al *dataset*_1, el mateix algorisme entrenat a partir d'un subconjunt d'entrenament paritari (vegeu *Taula 4*) ha obtingut uns resultats més propers i equiparables als del *dataset_2*. Tanmateix, podem observar com encara se situen al voltant d'uns 14 punts per sobre quant a precisió. Aleshores, i descartada ja la diferència pel que fa a la paritat de les categories en l'entrenament, ens hem aventurat a pensar

que aquesta es devia a una diferent qualitat interna de les dades lingüístiques contingudes en els jocs de dades. Efectivament, al *dataset_1* juntament amb el text de la piulada s'hi inclouen etiquetes d'usuari (els famosos *tags*, encapçalats pel signe # de què hem parlat més amunt), que en certa manera ja estan classificant el text. Per descomptat, no és que l'algorisme sàpiga el significat d'aquests adjectius, però sí que es pot "adonar" durant l'entrenament que sempre que apareix una determinada etiqueta (com ara #sad o #happy), l'observació que la conté està classificada en una categoria, i viceversa. Per tant, aquests mots de variabilitat i repertori més restringits que la resta de text lliure, actuarien com una mena d'atributs categoritzadors a dins de la cadena mateixa de text, independentment del camp extern de classificació. Per provar tal hipòtesi, hem procedit a esborrar les etiquetes en el *dataset_1* i tornat a efectuar l'entrenament. En aquest cas, la precisió obtinguda per l'algoritme (vegeu *Taula 5*) encara ha baixat una mica més, bo i situant-se en un rang més proper al del nostre referent (només a 6-7 punts per sobre, és a dir la meitat que en la prova anterior).

Pel que fa al *dataset_3*, l'entrenament amb un subconjunt paritari no ha suposat una reducció en la precisió de l'algorisme equiparable a la que ha patit el *dataset_1* per mitjà d'aquest mateix procediment (vegeu *Taula 6*) i els valors obtinguts encara són sospitosament alts –i en el cas del sistema de vectorització *Bag-of-Words*, fins i tot s'ha produït un lleuger augment. Com ha estat dit abans, a les piulades del *dataset_3* no hi figuren tags que puguem eliminar per tal d'equiparar precisions amb el *dataset_2*. Per tant, deu ser un altre tret inherent a les dades del *dataset_3* no observable a simple vista el que influeix en l'altíssima precisió de l'algorisme, per sobre dels resultats amb els altres jocs de dades. Probablement, aquest també és el motiu causant que la precisió de l'algorisme amb el *dataset_1* encara se situï clarament per sobre de la precisió amb el *dataset_2*, malgrat l'evident reducció després d'haver forçat el paritatge del primer en la classificació d'observacions i haver-ne eliminat tots els tags. Més endavant, l'anàlisi en profunditat de tots tres jocs de dades pel que fa a la distribució interna dels mots desvelarà quines són aquestes diferències ocultes entre els diferents reculls.

Quant a resultats segons mètodes de vectorització, en totes les proves *Tf-idf* és el que ha obtingut millors marques[20], seguit de *Bag-of-Words*, tret del cas amb el *dataset*_3, on el mètode capdavanter ha estat *Bag-of-Words*. En darrer lloc sempre ha quedat *Word2Vec*.

Quant a l'estructura de xarxa neuronal, hem efectuat diverses proves bo i alterant diversos paràmetres de la funció creadora, en concret el paràmetre *activation*, que permet d'especificar la funció d'activació d'entre les possibles opcions *identity*, *logistic*, *tanh* i *relu*, el paràmetre *hidden_layer_sizetuple*, que permet d'especificar el nombre de neurones a les capes ocultes i l'atribut *n_layers*, que permet d'especificar la quantitat de capes ocultes. No hem pogut apreciar una variació significativa en els resultats d'una configuració a una altra, de manera que per a les proves de la fase 2 amb *POS tagging* i pesos incorporats al vector de mots, hem deixat la configuració que venia per defecte (vegeu l'anterior apartat 4.5, on se'n descriuen els valors). Cal dir també que l'entrenament del *dataset_1* paritari vectoritzat amb el sistema *vec2word* no ha assolit la convergència amb les iteracions previstes per defecte, de manera que hem fet servir el paràmetre *max_iter* per augmentar-lo a 500 iteracions, topall que ha estat suficient per completar l'entrenament.

[20] La superioritat del vectoritzador *Tf-idf* –per damunt de *Word2vec*– en precisió obtinguda a l'hora d'entrenar algorismes de classificació amb dades textuals de dimensions entre 10K i 50K (com són els nostres datasets de 32K) ha quedat palesada per **Chakraborty, Elhence & Arora** (2019) i per **Irawaty, Andreswari & Pramesti** (2020).

5.Evolució del projecte: fase 2

5.1 Eliminació del soroll informatiu i reconeixement del llenguatge informal

Per saber de manera eficient i fiable quins mots no estan escrits segons les convencions i regles del llenguatge formal (que són els que poden provocar un etiquetatge morfosintàctic deficient), farem servir el següent fragment de codi per crear una llista que inclogui tots els mots del nostre corpus que no tenen correspondència amb el corpus *brown* de la llibreria *nltk*. Aquesta relació apareixerà en format de diccionari on la clau és cada classe de mot i el valor el nombre d'ocurrències:

```
unknownWords=[]
occurrences={}
brownWords = [word.lower() for word in nltk.corpus.brown.words()]
for row in tweet_dataset['tokens']:
	for token in row:
		if(token not in brownWords):
			unknownWords.append(token)
sorted_unkwnownWords =
	dict(sorted(Counter(unknownWords).items(),key=lambda item:
		item[1],reverse=True))
print(sorted_unkwnownWords)
```

No la copiarem aquí sencera perquè la llista és molt llarga, però hi apareixen formes com l'esmentada *bihday*, la qual és una evident errada tipogràfica que el corrector automàtic pot resoldre[21], mots propis de l'argot en xarxes socials com ara *retweet*, interjeccions com ara *ouch*, substantius informals com ara *dude* i abreviatures com ara *u* i *ur* dels pronoms *you* i *your* respectivament. De cara doncs a assegurar un bon etiquetatge morfosintàctic, pot ser interessant de reconstruir aquestes dues darreres formes abreujades, cosa que faríem cridant un mètode que recorrés el text de cada piulada bo i comparant els mots d'aquesta piulada amb les claus d'un diccionari propi de correccions: en cas que coincideixin, es procediria a substituir el mot de la piulada que consta com a clau de diccionari pel valor d'aquesta clau, això és la correcció que volem:

```
def correction(tweet):
	dictionary={'u':'you','ur':'your'
	}
	tweetwords=''
	for word in tweet.split():
		if(word in dictionary):
			word=dictionary[word]
		#elif word not in brownWords:
		#	word=spell(word)
		tweetwords+=(word+' ')
	tweet = tweetwords
	return tweet
```

[21] "(...) although it is difficult to detect the intent of the user, some content seems likely to have been accidentally mis-spelled (...) Automatic spelling correction may improve performance in these cases" (Derczinski, 2013: 201).

A banda del diccionari propi, elaborat a partir del coneixement directe que tenim del corpus, també podem indicar al mètode que corregeixi automàticament tota paraula no coincident amb el corpus *brown* de referència. Tanmateix, aquesta opció no sempre és recomanable atès l'advertiment que ja hem fet abans en el sentit que els correctors automàtics sovint "poden malinterpretar la intenció del parlant". Per això les línies de codi corresponents estan comentades.

5.2 Etiquetatge morfosintàctic

Tot i que a l'apartat teòric sobre l'etiquetatge morfosintàctic s'ha explicat com crear un etiquetador propi, l'etiquetador per defecte d'*nltk* fa una anàlisi força més ràpida i acurada, perquè compta amb models prèviament entrenats a partir de corpus encara més grans que el corpus *brown*. És per això que l'hem fet servir per al nostre treball. També ens hem adonat que l'etiquetatge guanya precisió si en el text que volem analitzar (és a dir, el joc de dades per fer la classificació del sentiment) es mantenen les *stop-words*. Recordem que a la primera fase les havíem eliminades tot seguint el procediment estàndard en preprocessament de dades textuals. Té sentit que amb les *stop-words* l'anàlisi gramatical per context obtingui millors resultats, perquè aquest tipus de mots són molt recurrents i a més a més exerceixen importants funcions d'estructuració de la frase, és a dir del context. El mateix fenomen de guany en la precisió a l'hora d'etiquetar gramaticalment s'esdevé quan no s'aplica la lematització, que és una altra de les tècniques de preprocessament textual explicades a l'anterior apartat 2.1.i. En aquest, cas la supressió de les flexions lèxiques dificulta l'etiquetador de distingir entre diferents tipolodies de mots a través de la forma externa, de manera que ha de fiar la desambiguació a l'anàlisi del context. D'altra banda, ens interessa conèixer la funció gramatical de les paraules per poder destacar aquelles amb més càrrega semàntica, de tal manera que guanyin més de pes en el procés d'aprenentatge per a la classificació del sentiment[22]. Al contrari que les esmentades *stop-words*, els mots amb més càrrega de significat pel que fa a les valoracions personals són, sens cap mena de dubte, els adjectius: amb ells manifestem opinions, gustos i preferències. Tot seguit vénen els substantius, atès que les entitats que anomenen porten implícites una sèrie

[22] "We notice that not all the basic atoms or words of a natural language sentence can be mapped to the corresponding parts of an image. For instance, a determiner (...) or preposition (...) cannot directly correspond to any explicit region of an image. On the contrary, the image regions including objects, their attributes and activities that usually make sense to caption generation, often correspond to nouns, verbs and adjectives of a natural language sentence, respectively" (He, 2017: 1-2).

de qualitats. Els verbs descriuen accions, i de vegades també serveixen per manifestar l'opinió. Pel que fa als adverbis, només els anomenats adverbis de modalitat poden expressar alguna mena de valoració subjectiva, de manera que caldria distingir-los de la resta (temporals i local), cosa que un etiquetador morfosintàctic no pot fer. Quant a la resta de tipologies, pronoms, preposicions i conjuncions formen part de les *stop-words*, és a dir que exerceixen importants funcions d'estructuració de la frase però posseeixen molt poca càrrega semàntica, i les interjeccions solen ser molt poc freqüents (tot i que no tant en textos informals). Així doncs, ens hem centrat a reconèixer i aplegar els adjectius i els substantius. El següent fragment de codi recorre tots els tokens del corpus d'entrenament, en fa l'etiquetatge morfosintàctic i desa adjectius i substantius en llistes separades:

```
JJ_list = []
NN_list = []
for items in train.tokens:
	analysis = nltk.pos_tag(items)
	for i in analysis:
		word=i[0]
		tag=i[1]
		myTuple=(word, tag)
		if(tag=='JJ'):
			JJ_list.append(word)
		elif(tag=='NN' or tag=='NNS'):
			NN_list.append(word)
```

Un cop identificats els dos tipus de mots que ens interessaven, els hem adjudicat un major pes en la vectorització del corpus d'entrenament. Per fer-ho, primer hem cercat la posició que ocupa cada element de la llista d'adjectius i substantius en el vector de *tfidf*, i aleshores hem multiplicat aquesta posició per un escalar. Aquest fragment de codi és el corresponent a l'increment de pes en els substantius (per als adjectius, només canvia la llista que recorre):

```
for word in NN_list:
	try:
		position = vectoriser.vocabulary_[word]
		X_train[:, position] *=4.0
	except:
		pass
```

Arribats a aquest punt, s'ha de fer esment del problema que ens hem trobat en aplicar aquesta solució al *dataset_2*, que és ostensiblement més gran que no el *dataset_1* o el *dataset_3*. A causa de la major freqüència dels mots, alguns dels valors en el vector de *tfidf* eren tan grans que a l'hora d'escalar-los se sobrepassava la capacitat d'emmagatzematge del dataframe. Com que aquest només admet codificació en *float32*

i *float64*, hem hagut de modificar l'anterior funció, de tal manera que desisteixi de fer l'escalatge si el valor obtingut sobrepassés el topall. La modificació és la següent (en aquest cas, fent servir la llista dels adjectius):

```
max_num = (np.finfo(np.float64).max/escalar
for word in JJ_list:
	try:
		position = vectoriser.vocabulary_[word]
		llista = csr_matrix(X_train[:, position]).data
		is_possible=True
		for i in llista:
			if(i>(max_num):
				is_possible=False
		if(is_possible):
			X_train[:, position] *=escalar
	except:
		pass
```

Ara bé, tot i que aquest mètode és capaç de solventar el problema, afecta plenament la nostra estratègia d'adjudicar un pes més gran a determinades categories gramaticals presents al joc de dades, atès que si d'entrada un mot ja té un valor de vectorització molt alt, el codi impedirà que se li assigni encara més pes en funció de la seva categoria gramatical, i aleshores la seva influència semàntica no hi estarà representada. Una altra forma d'evitar el problema de capacitat però que no restringeix el creixement vectorial de cap mot en particular és escalar el vector a la baixa abans de l'escalatge dels mots significatius, de manera que aquest també es pot aplicar de forma proporcional. El següent fragment de codi ho exemplifica:

```
X_train=X_train.multiply(0.01)

for word in JJ_list:
	try:
		position = vectoriser.vocabulary_[word]
		X_train[:, position] *=1.04
	except:
		pass

for word in NN_list:
	try:
		position = vectoriser.vocabulary_[word]
		X_train[:, position] *=1.05
	except:
		pass
```

També es pot assegurar la robustesa del mètode combinant les dues anteriors solucions.

D'altra banda, hem pogut comprovar com paraules mal tokenitzades[23] o bé l'escriptura de les quals difereix de l'estàndard, causen errors en l'etiquetatge gramatical. Això succeeix perquè els etiquetadors morfosintàctics s'entrenen amb corpus textuals prèviament etiquetats on els mots solen estar escrits seguint les pautes de la llengua formal, i si aquestes formes no coincideixen amb la forma dels mots que se li passen per classificar, aleshores no els reconeix. Per tal com en el nostre estudi volem provar com la millora en la precisió de l'algorisme classificador depèn en part del pes que se li assigni a determinats mots, i per això cal que l'etiquetador morfosintàctic els reconegui, aleshores ens hem d'assegurar que a l'etiquetador li arriba un text com més net i estandarditzat millor. El següent exemple mostra una piulada del corpus amb un error de classificació per culpa d'una tokenització deficient:

```
[('i', 'NN'), ("'m", 'VBP'), ('sure', 'JJ'), ('they', 'PRP'), ('are',
'VBP'), ('just', 'RB'), ('as', 'IN'), ('happy', 'JJ')]
```

Com es pot veure, etiqueta el primer mot com a nom (*NN*), quan es tracta del pronom personal de primera persona del singular. Això és pot resoldre mitjançant el mètode corrector que hem vist a l'apartat anterior, si a les claus del nostre diccionari de substitucions hi anem afegint totes aquelles formes que hem observat que l'etiquetador classifica malament, per exemple:

```
dictionary = {      'u': 'you',
                    'ur':'your',
                    'i':'I',
                    "i'm": 'I am',
                    "i've":'i have'}, etcètera.
```

Havent convertit el mot en majúscules i desfeta la contracció amb el verb, l'etiquetador ja reconeix que es tracta d'un pronom (*PR*):

```
[('I', 'PRP'), ('am', 'VBP'), ('sure', 'JJ'), ('they', 'PRP'), ('are',
'VBP'), ('just', 'RB'), ('as', 'IN'), ('happy', 'JJ')]
```

[23] "Occasionally the tokeniser or original author makes tokenisation errors (...) These are hard to correct. Specific subtypes of error, such as the joined words in the example, could be checked for a forcibly fixed, though this requires distinguishing intentional from unintentional word usage" (Derczinski, 2013: 201).

5.3 Distribució de freqüències

Per tal d'efectuar un estudi sobre la distribució de freqüències prou ampli però alhora precís que ens permeti de constatar les diferències ocultes entre els diversos jocs de dades, l'hem aplicat tant a la versió íntegra com d'entrenament de tots tres datasets, en combinació amb l'etiquetatge gramatical dels mots restringit als substantius i als adjectius, que són els dos tipus de mots que ens interessa d'observar. Per tal de calcular la distribució de freqüències, primer de tot ens cal obtenir la llista de tots els mots que apareixen en un corpus, i en el nostre cas les subllistes corresponents a adjectius i substantius per separat, cosa que obté el següent mètode:

```
def aux(dataset):

        all_words=[]
        JJ_words=[]
        NN_words=[]

        no_mind_words=['user']
        # List of all words + JJ / NN in dataset
        for row in dataset:
                analysis = nltk.pos_tag(row)
                for i in analysis:
                        word=i[0]
                        tag=i[1]
                        if(word not in no_mind_words):
                                if(tag=='JJ' or tag=='JJR' or tag=='JJS'):
                                        JJ_words.append(word)
                                elif(tag=='NN' or tag=='NNS'):
                                        NN_words.append(word)

                for token in row:
                        all_words.append(token)

        return all_words, JJ_words, NN_words
```

Per tal que aquest mètode ens generi les llistes d'ocurrències aparegudes segregant-les per etiqueta de les piulades, només cal passar-li com a argument un dataset amb el valor del camp *etiqueta* condicionat. Aquestes llistes no només ens permetran de calcular la distribució de freqüències, sinó també de recomptar tokens i ocurrències de mots en general, adjectius i substantius en tot el corpus, o bé en una i altra categoria.

Aquest fragment de codi imprimeix els primers 100 tokens més sovintejats del corpus, així com el nombre absolut de vegades que apareixen i el seu percentatge d'aparició:

```
wordsFrequency_all=FreqDist(words_list)
for mc in wordsFrequency_all.most_common(100):
      word=mc[0]
      abs_freq=mc[1]
      rel_freq=wordsFrequency_all.freq(word)
      string=str(abs_freq)+"\t"+str(rel_freq)+"\t"+word
      print(string)
```

Tenint en compte la classificació de les observacions en dues categories, anem a esbrinar si hi ha algun biaix en la distribució d'aquests mots que pugui explicar les diferències de precisió observades anteriorment entre els jocs de dades, i que en part es mantenien tot i haver estat equiparats tant en paritat d'etiquetes com en presència de mots clau.

Abans però cal tenir en compte aquestes dues prevencions: (a) com ja ha estat dit més amunt, es pot esdevenir que l'etiquetador morfosintàctic no hagi classificat correctament un mot –les confusions entre substantius i adjectius solen ser freqüents, per tal com la col·locació els pot fer ambivalents– i (b) només hem considerat els 100 primers mots més comuns, i atenent el que ja hem explicat de la *Llei de Zipf*, per sota d'aquesta classificació hi deu haver una enorme quantitat de mots ja no tan freqüents, diversos dels quals poden tenir igualment presència exclusiva en una sola classificació: aquests ens han passat per alt, però entenem que la seva baixa freqüència no els fa significatius tot i contribuir a un cert biaix en el corpus. A les taules, hem destacat els mots esbiaixats amb color blau.

Pel que fa al *dataset_1*, a la versió íntegra podem observar que diversos adjectius i substantius només apareixen en observacions classificades com a positives. Quant als adjectius (vegeu *Taula 17*), aquests són: *ready* amb 254 ocurrències de les 343571 de tots els mots i 28929 de tots els adjectius trobats al corpus sencer (vegeu *Taula 16*), *polar* (216), *amazing* (145), *useful* (87), *empty* (82), *sunday* (68), *bull* (52), *easy* (52), *orlando* (50), *vast* (50), *saturday* (49), *birthday* (46) i *busy* (44). Quant als substantius (vegeu *Taula 18*), aquests són: *birthday* amb 519 ocurrències de les 343571 de tots els mots i 28929 de tots els substantius trobats al corpus sencer (vegeu *Taula 16*), *tomorrow* (437), *bull* (277), *climb* (195), *friday* (182), *happiness* (135), *baby* (131), *orlando* (119), *environment* (117), *simulator* (112), *simulation* (110) i *moment* (104). Si ens limitem a la versió d'entrenament, els adjectius que només apareixen en observacions classificades com a positives (vegeu *Taula* 20) són *polar* amb 21 ocurrències de les 41450 de tots els

mots i 3622 de tots els substantius trobats al corpus d'entrenament (vegeu *Taula 19*), *ready* (14), *useful* (8) i *amazing* (8), mentre que els substantius són *birthday* amb 36 ocurrències de les 41450 de tots els mots i 10024 de tots els substantius trobats al corpus d'entrenament (vegeu *Taula 19*) i *tomorrow* (28). No s'ha observat cap adjectiu o substantiu que aparegui exclusivament en observacions classificades com a negatives.

Pel que fa al *dataset_2*, tant a la versió íntegra com a la versió d'entrenament, no s'ha observat cap adjectiu o substantiu completament esbiaixat (vegeu taules 22 a 27).
Finalment, pel que fa al *dataset_3*, a la versió íntegra podem observar que diversos adjectius i substantius només apareixen en observacions classificades o bé com a positives o bé com a negatives. Quant als adjectius (vegeu *Taula 29*), *quot* amb 89 ocurrències de les 161837 de tots els mots i 13840 de tots els adjectius trobats al corpus sencer (vegeu *Taula 28*), *amp* (51) i *cute* (40) només apareixen en observacions classificades com a positives, mentre que *emoji* (63), *bipolar* (31), *intellectual* (31), *depression* (23) i *chronic* (23) només apareixen en observacions classificades com a negatives. Quant als substantius (vegeu *Taula 30*), *quot* amb 253 ocurrències de les 161837 de tots els mots i 41334 de tots els substantius trobats al corpus sencer (vegeu *Taula 28*), *amp* (123), *yay* (58) i *it* (47) només apareixen en observacions classificades com a positives, mentre que *depression* (2307), *anxiety* (384), *emoji* (151) i *mentalhealth* (56) només apareixen en observacions classificades com a negatives. Si ens limitem a la versió d'entrenament, els adjectius *quot* amb 20 ocurrències de les 67737 de tots els mots i les 5742 de tots els adjectius trobats al corpus d'entrenament (vegeu *Taula 31*), *amp* (17) i *depression* (14) només apareixen en observacions classificades com a positives, mentre que *mental* (80), *emoji* (50), *emotional* (33), *bipolar* (26), *intellectual* (25) i *depression* (14) només apareixen en observacions classificades com a negatives (vegeu *Taula 32*), mentre que els substantius *depression* amb 1821 ocurrències de les 67737 de tots els mots i les 18071 de tots els substantius trobats al corpus d'entrenament (vegeu *Taula 31*), *anxiety* (306), *emoji* (121), *mentalhealth* (38), *illness* (34), *disorder* (26), *tears* (22) i *symptoms* (21) només apareixen en observacions classificades com a negatives (vegeu *Taula 33*). En aquest darrer cas, no s'ha observat que cap substantiu aparegui exclusivament en observacions positives.

Per tant i a tall de resum, podem dir que la diferència més evident entre jocs de dades pel que fa al biaix d'adjectius i substantius és que (a) el *dataset_2* no conté adjectius ni substantius que apareguin exclusivament en observacions en funció de la seva etiqueta,

(b) en canvi, els altres dos datasets sí que en tenen, i (c) de forma més accentuada el *dataset_3*, perquè si bé aquest compta amb menys quantitat de mots exclusius per etiqueta que el *dataset_1*, alguns dels seus mots esbiaixats tenen una gran presència en el corpus, principalment el substantiu més freqüent de tots (*depression* amb 2307 ocurrències, que se situa a una gran distància del segon amb 579 ocurrències), seguit pel que ocupa la sisena posició (*anxiety* amb 384 ocurrències), fenomen que encara s'agreuja en el corpus d'entrenament per tal com en aquest cas els dos primers mots amb més ocurrències presenten biaix.

D'altra banda, i posats a detallar encara més l'abast d'aquests biaixos, podríem entrar en l'anàlisi de les *collocations* o *concurrències*, que és la freqüència amb què un mot n'acompanya un altre bo i alterant-ne el significat. Per exemple, els adjectius *happy* i *sad*, que tenen una evident càrrega simbòlica, no presenten biaix exclusiu en cap dels corpus. Tanmateix, és lògic de relacionar el primer amb piulades que manifestin sentiments positius i el segon amb piulades que manifestin sentiments negatius. Però què passa si alguns d'aquests adjectius estan modificats per una partícula negativa? Aleshores el sentit primigeni d'aquests adjectius es podria invertir. Podem implementar una senzilla funció que explori aquestes concurrències:

```
def neg_collocations(dataset, word):
negation_list=[     'no',
                    'not',
                    'never',
                    'nor',
                    'neither',
                    'nowhere]
counter=0

for row in dataset:
       if word in row:
              for elem in row:
                     if elem in negation_list:
                            analysis = nltk.pos_tag(row)
                            print(analysis)
                            counter+=1
print(counter)
```

Aquest mètode demana com a paràmetres el corpus a explorar i un mot de concurrència, cerca en el corpus aquest mot i si a la mateixa piulada hi apareix una de les partícules de negació que consten a la llista, aleshores en fa l'etiquetatge morfosintàctic que treu per pantalla juntament amb el recompte de totes aquelles observacions on hi concorren el mot cercat amb una partícula de negació.

5.4 Descripció de les proves efectuades i avaluació de resultats

Per tal com el *dataset_1* és un corpus que se situa a mig camí entre els altres dos quant a precisió de l'algorisme classificador, aquest ha estat la base de les nostres exploracions en l'assignació de pes a adjectius i substantius. De primer ho hem provat en adjectius i substantius per separat, i a partir d'aquests resultats parcials hem decidit la millor combinació en l'assignació de pes a adjectius i substantius alhora. Tenint en compte que (a) un dels trets propis del *dataset_1* és que el corpus original comptava amb tags i (b) a la fase anterior ja havíem pogut comprovat com aquests empenyien la precisió cap amunt, hem efectuat les proves tant amb els tags originals del corpus com sense. El valor seguit del signe ~ és la mitjana que s'havia obtingut en la classificació feta a la fase 1, és a dir sense l'assignació de pes per tipus de paraula. Seguint la mateixa pauta efectuada fins ara, de cada 3 medicions se n'ha extret la mitjana, i aquest cop també se n'ha calculat la diferència en la precisió respecte de les mitjanes anteriors. Quan alguna classificació ha obtingut un resultat per sota de la mitjana anterior, aquestes s'han representat amb color vermell al dessota. Els resultats discrepants de la mitjana de totes les observacions d'un escalar, bé per més baixos o bé per més alts dels esperats, també s'han consignat amb color groc els primers i amb color verd els segons. Els dos millors escalars pel que fa a resultats estan encerclats amb color verd i blau respectivament.

La *Taula 9* comprèn l'assignació d'un pes complementari a adjectius sols i els corresponents resultats de l'algorisme classificador. Hem anat augmentant l'escalar a mesura que la precisió millorava (de *2 a *4), fins que ha tocat sostre (en *5) i finalment l'increment es tornava més petit (en *6). Aquest comportament és semblant en les dues versions del corpus, amb tags i sense tags.

La *Taula 10* comprèn l'assignació d'un pes complementari a substantius sols i els corresponents resultats de l'algorisme classificador. Contràriament al que ha passat amb els adjectius, el comportament ha estat diferent depenent de la versió del corpus, amb tags i sense tags. En el primer cas, si bé la millor classificació torna a ser per a l'escalar *5, es produeix una dramàtica baixada a l'escalar *4 a causa d'uns valors sorprenentment baixos que no semblen esporàdics. A diferència també dels adjectius, l'escalar *6 fa disminuir abruptament la precisió de l'algorisme. En el segon cas, l'increment de la millora ha estat gradual fins a tocar sostre a l'escalar *4, que obté la diferència positiva més gran de totes les proves. A partir d'aquí, el decrement en la millora ha estat gradual.

La *Taula 11* comprèn l'assignació combinada de pes tant a adjectius com a substantius, per a totes dues versions del corpus amb tags i sense tags. En el primer cas, el mateix escalar *5 a adjectius i substantius (que ja havia estat capdavanter per separat) ha obtingut una notable millora en la precisió de l'algorisme de fins a 7 punts. En el segon cas, hem provat diverses combinacions fins a arribar a la conclusió que les millors són la que escala *4 els adjectius i escala *3 els substantius amb un increment d'entre 7 i 9 punts, i la que escala *5 els adjectius i escala *4 els substantius, amb un increment que en el millor dels casos s'enfila fins als 9 punts, si bé cal fer avinent que han sovintejat alguns valors inesperadament baixos.

Finalment, la *Taula 12* recull les mateixes proves que la taula anterior, però amb l'afegit que aquest cop hem sotmès les dades textuals al mètode de depuració descrit a l'anterior apartat 8.1. Podem observar com la millora de resultats ha estat molt notable, probablement perquè l'etiquetador morfosintàctic ha guanyat en precisió i per tant l'assignació de pesos en funció de si el mot és adjectiu o subtantiu té un efecte més notori. En aquest cas, la millor precisió per al corpus sense tags s'ha aconseguit no només amb la ja coneguda combinació d'escalar els adjectius *3 i els substantius *4, sinó també amb la combinació, fins ara més aviat modesta en resultats, d'escalar *3 tant adjectius com substantius. Com a nota curiosa, cal fer avinent que la precisió per al corpus sense tags no ha experimentat cap canvi.

Amb aquests resultats provats en el *dataset_1*, hem procedit a aplicar les anteriors tècniques als altres dos jocs de dades. Pel que fa al *dataset_2*, l'assignació de pes complementari a adjectius i substantius ha provocat un retrocés en la precisió de l'algorisme classificador (vegeu *Taula 13*). Pel que fa al *dataset_3*, la minva de precisió encara ha estat pitjor (vegeu *Taula 14*). Aleshores hem volgut comprovar fins a quin punt l'assignació de pes complementari a adjectius i substantius només és eficaç en corpus textuals com el *dataset_1*, això és tematitzats (a diferència del *dataset_2*, que no ho està) i sense un biaix semàntic tan exagerat com el del *dataset_3*. Per fer-ho, hem aplicat al *dataset_3* una purga dels substantius més freqüents que només apareixen en observacions classificades d'una determinada manera, bo i afegint a l'esmentat mètode de depuració una llista complementària de mots que cal eliminar del corpus, en aquest cas *depression*, *anxiety*, *emoji*, *intellectual* i *bipolar*. En aquest cas l'increment de pes a adjectius i substantius sí que ha millorat resultats (tot i que moderadament) en la precisió de l'algorisme classificador (vegeu *Taula 15*).

6 Conclusions i futures línies d'investigació

Les característiques lingüístiques d'un corpus textual afecten la capacitat dels algorismes d'aprenentatge computacional a l'hora de classificar i predir el sentiment. Les tècniques pròpies del processament del llenguatge natural per mitjans computacionals són indicades per conèixer aquestes característiques, que altrament es mantenen ocultes a un examen superficial, així com per fer-ne un tractament adequat en cas necessari.

La distinció d'aquelles classes gramaticals amb més càrrega semàntica (fonamentalment adjectius i substantius) per tal d'assignar-los un pes superior en la vectorització dels mots, s'ha revelat com una tècnica que pot millorar notablement els resultats de la classificació en l'anàlisi del sentiment, sempre i quan el corpus (a) aplegui opinions al voltant d'una temàtica restringida, com s'ha vist en el cas del *dataset_1* i el *dataset_3* en front del *dataset_2*, i (b) els mots d'aparició més freqüent no presentin un biaix exclusiu cap a una o altra categoria, com s'ha vist en el cas del *dataset_1* en front del *dataset_3*. Aquesta darrera situació s'ha pogut revertir parcialment amb la supressió d'aquests mots completament esbiaixats, fins al punt que l'algorisme classificador ha millorat lleugerament resultats respecte de l'entrenament sense dita supressió. En certa manera, ha estat com fer una passa enrera per poder-ne fer dues cap endavant. Un examen més detallat del corpus possiblement afinaria sobre el tractament a aplicar-hi. Pel que fa al *dataset_2*, la dispersió total quant a la temàtica ha provocat que en aquest corpus no hi apareguin ni adjectius ni substantius exclusivament esbiaixats, de manera que no s'hi ha pogut aplicar una estratègia semblant a l'aplicada al *dataset_3*. Tanmateix, una anàlisi de les freqüències i els biaxos parcials (això és, un adjectiu o substantiu amb una presència majoritària, tot i que no exclusiva, a les piulades d'una determinada categoria) potser podria indicar la fórmula per revertir aquesta situació de bloqueig. Aquest tipus d'exploració, juntament amb la cerca d'un mètode per automatitzar la detecció i tractament de freqüències i biaixos lèxics aplicada a millorar la classificació en l'anàlisi del sentiment, s'obren com a dues possibles futures línies d'investigació en l'espai comú que s'obre entre l'aprenentatge computacional i el processament del llenguatge natural.

Annex A: Relació de lliurables

- Arxius “nn_dataset_1”, “nn_dataset_2” i “nn_dataset_3” en format *.py* que contenen els programes en Python on es duen a terme tots els passos anteriorment descrits per a l’anàlisi del sentiment en un corpus textual amb refinament per PLN, en les variants per al *dataset_1*, el *dataset_2* i el *dataset_3* respectivament.
- Jocs de dades “dataset_1”, “dataset_2” i “dataset_3” en format *.csv*.

Annex B: taules de resultats

	dataset_1 (31962 observacions) **Categoria_1:** 29720 / 31962 = 92.98% **Categoria_2:** 2242 / 31962 = 7.01%					
	BagOfWords		**TfIdf**		**doc2Vec**	
Matriu de confusió	121	67	123	65	70	118
	123	65	122	66	73	115
	121	67	124	64	97	91
	69	2940	19	2990	31	2978
	70	2939	17	2992	28	2981
	58	2951	19	2990	32	2977
Precisió	95.75 %		97.37 %		95.34 %	
	95.78 %		97.40 %		95.53 %	
	96.09 %		97.40 %		96.15 %	
Mitjana	**95.87 %**		**97.39 %**		**95.67 %**	

Taula 1

	dataset_2 (31960 observacions) **Categoria_1:** 15980 / 31962 = 50% **Categoria_2:** 15980 / 31962 = 50%					
	BagOfWords		**TfIdf**		**doc2Vec**	
Matriu de confusió	1104	480	1150	434	1002	582
	1161	423	1098	486	1094	490
	1158	426	1172	412	1181	466
	409	1203	402	1210	490	1104
	465	1147	343	1269	568	1044
	431	1181	415	1197	466	1210
Precisió	72.18 %		73.84 %		65.89 %	
	72.22 %		74.06 %		66.90 %	
	73.19 %		74.12 %		72.84 %	
Mitjana	**72.53 %**		**74.01 %**		**68.54 %**	

Taula 2

	dataset_3 (10314 observacions) **Categoria_1:** 8000 / 10314 = 77.56% **Categoria_2:** 2314 / 10314 = 22.43%					
	BagOfWords		**TfIdf**		**doc2Vec**	
Matriu de confusió	468	22	432	58	414	76
	466	24	431	59	409	81
	468	22	433	57	407	83
	0	1573	18	1555	36	1537
	0	1573	19	1554	36	1537
	0	1573	20	1553	32	1541
Precisió	98.93 %		96.32 %		94.57 %	
	98.84 %		96.22 %		94.33 %	
	98.93 %		96.27 %		94.42 %	
Mitjana	**98.90 %**		**96.27 %**		**94.44 %**	

Taula3

#tags	dataset_1 (31962 observacions) – train paritari (3688 observacions) **Categoria_1:** 1834 / 3688 = 50% **Categoria_2:** 1834 / 3688 = 50%							
	BagOfWords		**TfIdf**		**doc2Vec***			
Matriu de confusió	344	64	355	53	317	91	327	1376
	345	63	359	49	309	99	300	1262
	344	64	359	49	315	93	294	1111
	899	5086	817	5168	1312	4672	81	4609
	842	5143	802	5183	1275	4710	108	4723
	812	5713	793	5192	1254	4731	114	4874
Precisió	84.94 %		86.39 %		78.05 %		77.21 %	
	85.84 %		86.69 %		78.51 %		78.57 %	
	86.30 %		86.83 %		78.93 %		80.84 %	
Mitjana	**85.69 %**		**86.64 %**		**78.50 %**		**78.87 %**	

Taula 4

no #tags	dataset_1 (31962 observacions) – train paritari (3688 observacions) **Categoria_1:** 1834 / 3688 = 50% **Categoria_2:** 1834 / 3688 = 50%							
	BagOfWords		**TfIdf**		**doc2Vec***			
Matriu de confusió	316	92	313	95	269	112	278	130
	310	98	319	89	265	143	275	133
	313	95	325	83	277	131	281	127
	1262	4723	1079	4906	1630	4355	1663	4322
	1173	4812	1091	4894	1446	4539	1576	4409
	1152	4833	1076	4909	1519	4466	1599	4386
Precisió	78.82 %		81.64 %		72.75 %		71.95 %	
	80.12 %		81.54 %		74.19 %		73.00 %	
	80.49 %		81.87 %		75.14 %		73.27 %	
Mijana	**79.81 %**		**81.68 %**		**74.03 %**		**72.74 %**	

Taula 5

(*) La primera columna mostra els resultats de la classificació havent entrenat l'algorisme MLP tal com s'ha fet per als altres vectoritzadors, això és deixant els paràmetres per defecte en què el nombre d'iteracions està establert a 200. En aquest cas, l'entrenament no va arribar a la convergència. La segona columna en canvi mostra els resultats de la classificació havent entrenat l'algorisme MLP amb el paràmetre que regula el nombre d'iteracions establert a 500. En aquest cas, l'entrenament va poder acabar amb convergència. Amb tot, els resultats d'un i altre intent tampoc difereixen gaire.

	dataset_3 (10314 observacions) – train paritari (3648 observacions) **Categoria_1:** 1824 / 3648 = 50% **Categoria_2:** 1824 / 3648 = 50%					
	BagOfWords		**TfIdf**		**doc2Vec**	
Matriu de confusió	471 476 469	19 23 21	461 457 463	29 33 27	449 445 439	41 45 51
	19 24 27	1554 1549 1546	141 127 135	1432 1446 1438	160 154 155	1413 1419 1418
Precisió	98.16 % 97.72 % 97.67 %		91.76 % 92.24 % 92.15 %		90.26 % 90.35 % 90.01 %	
Mitjana	**97.85 %**		**92.05 %**		**90.21 %**	

Taula 6

no exlusive labelled words	dataset_3 (10314 observacions) **Categoria_1:** 8000 / 10314 = 77.56% **Categoria_2:** 2314 / 10314 = 22.43%					
	BagOfWords		**TfIdf**		**doc2Vec**	
Matriu de confusió	374 373 373	116 117 117	383 380 385	107 110 105	290 295 290	200 195 200
	114 101 107	1459 1472 1466	62 63 67	1511 1510 1506	58 57 61	1515 1516 1512
Precisió	88.85 % 89.43 % 89.14 %		91.81 % 91.61 % 91.66 %		87.49 % 87.78 % 87.35 %	
Mitjana	**89.14 %**		**91.69 %**		**87.54 %**	

Taula 7

no exlusive labelled words	dataset_3 (10314 observacions) – train paritari (3648 observacions) **Categoria_1:** 1824 / 3648 = 50% **Categoria_2:** 1824 / 3648 = 50%					
	BagOfWords		**TfIdf**		**doc2Vec**	
Matriu de confusió	438 441 439	52 49 51	437 440 435	53 50 55	373 392 374	117 98 116
	263 258 243	1319 1315 1330	238 245 258	1335 1328 1315	273 310 271	1300 1263 1302
Precisió	84.73 % 85.12 % 85.75 %		85.89 % 85.70 % 84.83 %		81.09 % 80.22 % 81.24 %	
Mitjana	**85.20 %**		**85.47 %**		**80.85 %**	

Taula 8

dataset_1 amb adjectius destacats x POS tagging									
amb #tags ~ 86.64									
*2		*3		*4		*5		*6	
85.64	88.09	89.45	90.80	89.86	88.14	92.27	90.92	89.86	90.36
88.47	89.83	91.80	88.80	90.78	90.83	92.46	91.34	92.28	89.12
89.64	89.11	89.39	91.00	91.06	92.19	91.41	91.06	91.27	92.24
87.91	**89.01**	**90.21**	**90.20**	**90.57**	**90.39**	**92.04**	**91.11**	**91.14**	**90.57**
(+1.28)	**(+2.37)**	**(+3.58)**	**(+3.56)**	**(+3.93)**	**(+3.75)**	**(+5.41)**	**(+4.47)**	**(+4.50)**	**(+3.94)**
sense #tags ~ 81.68									
*2		*3		*4		*5		*6	
81.08	83.18	84.63	83.32	82.91	86.12	85.62	85.35	85.84	84.68
84.57	80.32	84.45	86.23	87.01	85.92	86.31	88.61	84.09	86.00
82.18	81.40	86.01	83.96	85.39	85.82	84.51	83.09	84.18	83.95
82.61	**81.63**	**85.03**	**84.50**	**85.10**	**85.95**	**85.48**	**85.68**	**84.70**	**84.87**
(+0.94)	**(-0.04)**	**(+3.35)**	**(+2.83)**	**(+3.43)**	**(+4.28)**	**(+3.81)**	**(+4.01)**	**(+3.03)**	**(+3.20)**
						61.286			

Taula 9

dataset_1 amb substantius destacats x POS tagging									
amb #tags ~ 86.64									
*2		*3		*4		*5		*6	
89.91	91.20	87.39	93.32	78.08	79.14	94.16	94.58	83.37	80.77
88.48	87.45	91.75	89.14	94.00	92.85	90.55	89.42	88.69	92.64
89.41	84.45	92.72	93.75	94.47	86.42	94.38	93.55	85.65	82.04
89.26	**87.70**	**90.56**	**92.07**	**88.85**	**86.14**	**93.03**	**92.52**	**85.91**	**85.156**
(+2.63)	**(+1.06)**	**(+3.92)**	**(+5.43)**	**(+2.2)**	**(-0.49)**	**(+6.39)**	**(+5.88)**	**(-0.73)**	**(-1.48)**
		83.247						**46.832** **72.220** **94.353**	
sense #tags ~ 81.68									
*2		*3		*4		*5		*6	
86.46	84.06	89.26	83.67	89.48	89.67	86.46	85.35	82.01	87.12
88.92	85.71	84.76	86.53	89.84	88.08	85.78	81.85	85.28	82.77
79.99	86.34	86.06	83.98	87.64	85.65	87.97	88.81	85.71	88.67
85.12	**85.37**	**86.70**	**84.72**	**88.99**	**87.80**	**86.74**	**85.34**	**84.33**	**86.19**
(+3.45)	**(+3.69)**	**(+5.02)**	**(+3.05)**	**(+7.13)**	**(+6.12)**	**(+5.06)**	**(+3.66)**	**(+2.48)**	**(+4.51)**
				82.871		**77.819** **93.383**		**63.945** **94.488**	

Taula 10

dataset_1 amb adjectius i substantius destacats x POS tagging									
amb #tags ~ 86.64									
								*5*5	
								94.44 93.94 93.16	93.30 93.86 93.13
								93.85 **(+7.21)**	**93.43** **(+6.79)**
sense #tags ~ 81.68									
*3*3		*4*3		*4*4		*4*5		*5*4	
81.29 89.48 92.55	93.43 84.14 86.05	85.64 88.83 90.81	91.05 92.67 89.63	92.60 90.95 86.06	93.85 88.84 90.03	87.12 89.41 90.06	88.19 87.79 86.66	90.94 90.97 90.80	85.84 86.07 92.75
87.778 **(+6.09)**	**87.87** **(+6.19)**	**88.430** **(+6.74)**	**91.12** **(+9.44)**	**87.87** **(+6.19)**	**90.91** **(+9.23)**	**88.86** **(+7.18)**	**87.55** **(+5.86)**	**90.906** **(+9.23)**	**88.226** **(+6.55)**
				79.759		**78.179** **80.432**		**70.908** **76.865**	

Taula 11

dataset_1 amb adjectius i substantius destacats x POS tagging + correcció x spelling									
amb #tags ~ 86.64									
								*5*5	
								91.16 92.35 95.43	94.27 94.90 92.88
								92.98 **(+6.34)**	**94.02** **(+7.38)**
								88.38	
sense #tags ~ 81.68									
*3*3		*4*3		*4*4		*4*5		*5*4	
93.40 94.59 92.91	93.26 93.38 94.49	94.59 94.29 94.62	93.20 93.35 92.98	94.46 92.04 91.77	93.88 92.90 86.44	94.09 94.32 93.38	93.32 92.52 92.82	90.00 94.23 94.81	93.72 93.98 89.64
93.63 **(+11.95)**	**93.71** **(+12.03)**	**94.50** **(+12.82)**	**93.18** **(+11.50)**	**92.76** **(+11.08)**	**90.84** **(+9.16)**	**93.93** **(+12.25)**	**92.89** **(+11.21)**	**93.01** **(+11.30)**	**93.45** **(+10.76)**
		88.61						**84.03**	

Taula 12

dataset_2 amb adjectius i substantius destacats x POS tagging				
~ 74.01				
***3*3**	***4*3**	***4*4**	***4*5**	***5*4**
71.59 71.03 71.67	68.80 66.86 71.63	71.71 69.26 64.41	70.90 69.71 71.24	65.38 69.91 70.27
71.43 **(-2.58)**	**69.10** **(-4.91)**	**68.46** **(-5.55)**	**70.61** **(-3.39)**	**68.52** **(-5.49)**

Taula 13

dataset_3 original amb adjectius i substantius destacats x POS tagging				
~ 85.47				
***3*3**	***4*3**	***4*4**	***4*5**	***5*4**
76.25 76.30 76.34	76.30 76.30 76.25	76.30 76.78 76.25	76.25 76.25 76.25	77.12 76.29 76.30
76.27 **(-9.20)**	**76.28** **(-9.19)**	**76.44** **(-9.03)**	**76.25** **(-9.22)**	**76.57** **(-8.90)**

Taula 14

dataset_3 amb els principals substantius i adjectius d'etiquetatge exclusiu eliminats i la resta destacats x POS tagging									
~ 85.47									
***3*3**		***4*3**		***4*4**		***4*5**		***5*4**	
86.72 87.11 88.03	87.93 88.27 86.33	89.14 87.74 90.01	89.63 87.11 86.82	88.17 86.18 86.04	88.37 87.69 85.80	86.09 89.09 89.58	86.72 84.25 87.88	88.56 86.14 89.43	87.83 87.83 88.66
86.55 **(+1.08)**	**87.51** **(+2.04)**	**88.96** **(+3.49)**	**87.85** **(+2.38)**	**86.80** **(+1.33)**	**87.29** **(+1.82)**	**88.25** **(+2.78)**	**86.28** **(+0.81)**	**88.04** **(+2.57)**	**88.10** **(+2.63)**
80.76 **83.62**				**80.29** **84.68**		**82.26** **83.57**		**84.25**	

Taula 15

dataset_1						
labels	paraules		adjectius		substantius	
	tipus	ocurrències	tipus	ocurrències	tipus	ocurrències
all	22118	343571	4614	28929	14541	85077
positive	20897	316513	4323	26472	13750	78529
negative	4556	27058	849	2456	2491	6548

Taula 16

Freqüència i distribució dels adjectius al dataset_1 sencer								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
1576	0.0544800884	happy	1565	0.0591190692	happy	125	0.0508957654	white
930	0.0321487831	new	866	0.0327138108	new	90	0.0366449511	black
793	0.0274128871	good	761	0.0287473556	good	86	0.0350162866	amp
489	0.0169040376	great	461	0.0174146267	great	64	0.0260586319	new
452	0.015625	best	451	0.0170368691	thankful	35	0.0142508143	racist
451	0.0155904314	thankful	444	0.0167724388	best	34	0.0138436482	right
435	0.0150373340	more	416	0.0157147174	more	32	0.0130293159	latest
345	0.0119261615	first	331	0.0125037775	first	32	0.0130293159	good
334	0.0115459070	amp	325	0.0122771229	last	28	0.0114006514	great
333	0.0115113384	last	290	0.0109549712	sad	25	0.0101791530	same
297	0.0102668694	sad	281	0.0106149894	next	24	0.0097719869	own
292	0.0100940265	next	273	0.0103127833	much	23	0.0093648208	many
290	0.0100248893	much	266	0.0100483529	many	22	0.0089576547	other
289	0.0099903207	many	261	0.0098594741	little	21	0.0085504885	daily
266	0.0091952433	little	260	0.0098216983	old	21	0.0085504885	real
266	0.0091952433	old	254	0.0095950438	ready	19	0.0077361563	more
254	0.0087804203	ready	248	0.0093683892	amp	17	0.0069218241	free
216	0.0074668141	polar	216	0.0081595648	polar	17	0.0069218241	much
213	0.0073631084	beautiful	212	0.0080084617	beautiful	14	0.0057003257	first
196	0.0067754424	angry	192	0.0072529465	angry	14	0.0057003257	thought
188	0.0064988938	free	177	0.0066863100	big	13	0.0052931596	american
187	0.0064643252	black	171	0.0064596554	free	12	0.0048859934	most
184	0.0063606194	big	153	0.0057796917	other	12	0.0048859934	true
182	0.0062914823	white	152	0.0057419159	bad	12	0.0048859934	better
175	0.0060495022	other	145	0.0054774856	amazing	11	0.0044788273	anti
174	0.0060149336	right	142	0.0053641583	true	11	0.0044788273	porn
160	0.0055309734	bad	140	0.0052886068	nice	11	0.0044788273	next
159	0.0054964048	real	140	0.0052886068	right	11	0.0044788273	happy
154	0.0053235619	true	138	0.0052130553	real	11	0.0044788273	mad
152	0.0052544247	own	128	0.0048352976	own	11	0.0044788273	racial
145	0.0050124446	amazing	127	0.0047975219	strong	11	0.0044788273	such
144	0.0049778761	nice	119	0.0044953158	full	10	0.0040716612	islamic
137	0.0047358960	same	118	0.0044575400	few	9	0.0036644951	african
132	0.0045630530	strong	118	0.0044575400	such	9	0.0036644951	uk
129	0.0044593473	such	112	0.0042308854	same	9	0.0036644951	worst
125	0.0043210730	few	107	0.0040420066	better	8	0.0032573289	best
124	0.0042865044	full	106	0.0040042308	awesome	8	0.0032573289	semitic
119	0.0041136615	better	103	0.0038909035	top	8	0.0032573289	different
109	0.0037679756	awesome	98	0.0037020247	hard	8	0.0032573289	hard
106	0.0036642699	hard	97	0.0036642490	black	8	0.0032573289	non
105	0.0036297013	top	95	0.0035886974	long	8	0.0032573289	last
102	0.0035259955	only	95	0.0035886974	only	8	0.0032573289	bad
97	0.0033531526	long	93	0.0035131459	friday	8	0.0032573289	wrong
95	0.0032840154	most	89	0.0033620429	cold	7	0.0028501628	liberal
93	0.0032148783	cold	87	0.0032864913	wonderful	7	0.0028501628	asian
93	0.0032148783	friday	87	0.0032864913	useful	7	0.0028501628	only
89	0.0030766039	live	86	0.0032487156	excited	7	0.0028501628	few
89	0.0030766039	wonderful	85	0.0032109398	live	7	0.0028501628	t
87	0.0030074668	excited	83	0.0031353883	young	7	0.0028501628	stupid

87	0.0030074668	young	83	0.0031353883	most	7	0.0028501628	sad
87	0.0030074668	dead	83	0.0031353883	dead	7	0.0028501628	big
87	0.0030074668	useful	82	0.0030976125	empty	7	0.0028501628	brown
85	0.0029383296	high	81	0.0030598368	high	7	0.0028501628	s
83	0.0028691924	wrong	75	0.0028331822	important	7	0.0028501628	denial
82	0.0028346238	empty	75	0.0028331822	wrong	7	0.0028501628	dear
80	0.0027654867	important	75	0.0028331822	perfect	7	0.0028501628	dangerous
79	0.0027309181	latest	73	0.0027576307	open	7	0.0028501628	sick
77	0.0026617809	perfect	72	0.0027198549	able	7	0.0028501628	fat
77	0.0026617809	open	69	0.0026065276	special	6	0.0024429967	old
73	0.0025235066	able	68	0.0025687518	sunday	6	0.0024429967	co
73	0.0025235066	special	65	0.0024554245	sure	6	0.0024429967	small
68	0.0023506637	sunday	65	0.0024554245	favorite	6	0.0024429967	palestinian
68	0.0023506637	sure	64	0.0024176488	lovely	6	0.0024429967	nude
67	0.0023160951	daily	63	0.0023798730	whole	6	0.0024429967	fuck
66	0.0022815265	favorite	61	0.0023043215	super	6	0.0024429967	enough
66	0.0022815265	whole	57	0.0021532184	safe	6	0.0024429967	total
65	0.0022469579	lovely	57	0.0021532184	single	6	0.0024429967	ag
63	0.0021778207	super	57	0.0021532184	white	6	0.0024429967	jewish
63	0.0021778207	single	55	0.0020776669	less	6	0.0024429967	wish
60	0.0020741150	safe	55	0.0020776669	proud	6	0.0024429967	blicqer
60	0.0020741150	proud	54	0.0020398912	positive	6	0.0024429967	single
59	0.0020395464	less	53	0.0020021154	funny	6	0.0024429967	systemic
57	0.0019704092	positive	52	0.0019643396	bull	6	0.0024429967	hot
56	0.0019358407	funny	52	0.0019643396	easy	5	0.0020358306	poor
55	0.0019012721	least	52	0.0019643396	final	5	0.0020358306	evil
55	0.0019012721	crazy	51	0.0019265639	least	5	0.0020358306	proud
53	0.0018321349	worst	51	0.0019265639	crazy	5	0.0020358306	dumb
53	0.0018321349	final	50	0.0018887881	orlando	5	0.0020358306	sexist
52	0.0017975663	different	50	0.0018887881	gt	5	0.0020358306	nationalist
52	0.0017975663	bull	50	0.0018887881	vast	5	0.0020358306	excellent
52	0.0017975663	easy	49	0.0018510123	saturday	5	0.0020358306	pa
51	0.0017629977	gt	48	0.0018132366	late	5	0.0020358306	little
51	0.0017629977	small	47	0.0017754608	latest	5	0.0020358306	common
51	0.0017629977	late	46	0.0017376851	birthday	5	0.0020358306	acceptable
50	0.0017284292	orlando	46	0.0017376851	daily	5	0.0020358306	obama
50	0.0017284292	vast	46	0.0017376851	early	5	0.0020358306	strong
49	0.0016938606	social	45	0.0016999093	small	5	0.0020358306	full
49	0.0016938606	saturday	44	0.0016621335	different	5	0.0020358306	important
49	0.0016938606	hot	44	0.0016621335	due	5	0.0020358306	social
49	0.0016938606	early	44	0.0016621335	social	5	0.0020358306	equal
48	0.0016592920	sick	44	0.0016621335	worst	5	0.0020358306	israeli
48	0.0016592920	american	44	0.0016621335	red	5	0.0020358306	human
47	0.0016247234	due	44	0.0016621335	busy	5	0.0020358306	trump
46	0.0015901548	birthday	44	0.0016621335	local	5	0.0020358306	main
45	0.0015555862	green	43	0.0016243578	available	4	0.0016286644	nice
45	0.0015555862	red	43	0.0016243578	hot	4	0.0016286644	green
45	0.0015555862	local	43	0.0016243578	fantastic	4	0.0016286644	young
44	0.0015210176	poor	42	0.0015865820	sweet	4	0.0016286644	hateful
44	0.0015210176	fantastic	42	0.0015865820	simple	4	0.0016286644	familiar
44	0.0015210176	busy	42	0.0015865820	cool	4	0.0016286644	mexican

Taula 17

Freqüència i distribució dels substantius al dataset_1 sencer								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
2080	0.0244484408	day	2072	0.0263851570	day	143	0.0218387293	amp
1101	0.0129412179	time	1079	0.0137401469	time	84	0.0128283445	people
1016	0.0119421230	today	1002	0.0127596174	today	73	0.0111484422	libtard
846	0.0099439331	amp	740	0.0094232703	people	67	0.0102321319	women
824	0.0096853438	people	729	0.0092831947	life	65	0.0099266951	trump
735	0.0086392326	life	703	0.0089521068	amp	64	0.0097739767	racist
519	0.0061003561	birthday	519	0.0066090234	birthday	62	0.0094685400	racism
510	0.0059945696	way	492	0.0062652013	way	41	0.0062614538	woman
476	0.0055949316	days	470	0.0059850501	days	39	0.0059560171	men
460	0.0054068667	father	456	0.0058067720	father	37	0.0056505803	comments
437	0.0051365233	tomorrow	437	0.0055648231	tomorrow	35	0.0053451435	sex
424	0.0049837206	week	420	0.0053483426	week	34	0.0051924251	man
398	0.0046781151	world	389	0.0049535840	morning	33	0.0050397067	race
391	0.0045958367	morning	372	0.0047371034	world	31	0.0047342700	obama
340	0.0039963797	night	338	0.0043041424	night	31	0.0047342700	video
335	0.0039376094	weekend	335	0.0042659399	weekend	30	0.0045815516	thanks
331	0.0038905932	work	318	0.0040494594	work	29	0.0044288332	year
329	0.0038670851	year	300	0.0038202447	year	28	0.0042761148	america
302	0.0035497255	things	295	0.0037565740	things	27	0.0041233964	act
293	0.0034439390	love	288	0.0036674349	love	26	0.0039706780	world
277	0.0032558740	bull	277	0.0035273593	bull	26	0.0039706780	condemns
275	0.0032323659	everyone	264	0.0033618153	everyone	25	0.0038179596	nothing
268	0.0031500875	thanks	253	0.0032217397	family	25	0.0038179596	guy
263	0.0030913172	family	240	0.0030561958	home	25	0.0038179596	girl
249	0.0029267604	man	238	0.0030307275	thanks	24	0.0036652412	president
246	0.0028914982	home	237	0.0030179933	friends	24	0.0036652412	suppoers
241	0.0028327279	friends	230	0.0029288543	city	23	0.0035125229	word
235	0.0027622036	thank	227	0.0028906518	thank	22	0.0033598045	pay
235	0.0027622036	city	226	0.0028779177	bear	22	0.0033598045	time
228	0.0026799252	fathers	226	0.0028779177	fathers	21	0.0032070861	girls
227	0.0026681711	bear	215	0.0027378420	man	20	0.0030543677	police
220	0.0025858927	someone	211	0.0026869054	gt	20	0.0030543677	factory
217	0.0025506306	game	208	0.0026487030	game	20	0.0030543677	s
217	0.0025506306	gt	207	0.0026359688	tonight	19	0.0029016493	w
208	0.0024448440	tonight	206	0.0026232347	someone	19	0.0029016493	sea
200	0.0023508116	nothing	195	0.0024831590	climb	19	0.0029016493	shepherd
196	0.0023037953	news	190	0.0024194883	summer	18	0.0027489309	way
195	0.0022920413	climb	184	0.0023430834	music	18	0.0027489309	country
191	0.0022450250	music	182	0.0023176151	friday	17	0.0025962125	hate
191	0.0022450250	years	180	0.0022921468	news	17	0.0025962125	person
191	0.0022450250	summer	176	0.0022412102	years	17	0.0025962125	history
182	0.0021392385	friday	175	0.0022284761	nothing	16	0.0024434941	school
181	0.0021274845	girl	168	0.0021393370	something	16	0.0024434941	book
178	0.0020922223	trump	158	0.0020119955	school	16	0.0024434941	news
174	0.0020452061	school	158	0.0020119955	thing	16	0.0024434941	media
173	0.0020334520	thing	156	0.0019865272	girl	15	0.0022907758	thing
171	0.0020099439	something	151	0.0019228565	friend	15	0.0022907758	fuck
155	0.0018218790	dad	151	0.0019228565	dad	15	0.0022907758	years
154	0.0018101249	friend	150	0.0019101223	weeks	15	0.0022907758	paladino
152	0.0017866168	weeks	149	0.0018973882	place	14	0.0021380574	won
152	0.0017866168	place	147	0.0018719199	sunday	14	0.0021380574	someone
152	0.0017866168	video	140	0.0017827808	need	14	0.0021380574	today
148	0.0017396005	sunday	138	0.0017573125	show	14	0.0021380574	job
148	0.0017396005	person	136	0.0017318442	kids	13	0.0019853390	chair
146	0.0017160924	girls	135	0.0017191101	hea	13	0.0019853390	remarks
145	0.0017043384	need	135	0.0017191101	happiness	13	0.0019853390	whites
144	0.0016925843	show	132	0.0016809076	team	13	0.0019853390	please
144	0.0016925843	team	131	0.0016681735	food	13	0.0019853390	videos
143	0.0016808303	kids	131	0.0016681735	baby	13	0.0019853390	work
143	0.0016808303	women	131	0.0016681735	lot	13	0.0019853390	ignorance
139	0.0016338140	everything	131	0.0016681735	person	13	0.0019853390	kind
138	0.0016220600	lot	129	0.0016427052	everything	13	0.0019853390	oil
137	0.0016103059	hea	128	0.0016299710	tweets	13	0.0019853390	blacks

137	0.0016103059	pay	125	0.0015917686	girls	12	0.0018326206	comment
135	0.0015867978	happiness	125	0.0015917686	god	12	0.0018326206	americans
133	0.0015632897	food	124	0.0015790344	buffalo	12	0.0018326206	jews
131	0.0015397816	baby	121	0.0015408320	video	12	0.0018326206	muslims
131	0.0015397816	tweets	121	0.0015408320	house	12	0.0018326206	team
129	0.0015162734	god	120	0.0015280979	fun	12	0.0018326206	others
128	0.0015045194	sex	119	0.0015153637	orlando	12	0.0018326206	polarisation
127	0.0014927653	house	119	0.0015153637	gorilla	11	0.0016799022	business
127	0.0014927653	buffalo	117	0.0014898954	environment	11	0.0016799022	color
125	0.0014692572	one	116	0.0014771613	one	11	0.0016799022	children
125	0.0014692572	job	115	0.0014644271	pay	11	0.0016799022	christmas
123	0.0014457491	fun	115	0.0014644271	face	11	0.0016799022	course
123	0.0014457491	face	113	0.0014389588	trump	11	0.0016799022	towards
121	0.0014222410	book	113	0.0014389588	end	11	0.0016799022	look
121	0.0014222410	gorilla	113	0.0014389588	guys	11	0.0016799022	everyone
119	0.0013987329	orlando	113	0.0014389588	month	11	0.0016799022	message
118	0.0013869788	end	112	0.0014262247	simulator	11	0.0016799022	feel
118	0.0013869788	guys	111	0.0014134905	hours	10	0.0015271838	attack
117	0.0013752247	environment	111	0.0014134905	job	10	0.0015271838	family
117	0.0013752247	look	110	0.0014007564	simulation	10	0.0015271838	aicle
113	0.0013282085	hours	107	0.0013625539	season	10	0.0015271838	power
113	0.0013282085	month	106	0.0013498198	look	10	0.0015271838	ppl
112	0.0013164545	simulator	105	0.0013370856	book	10	0.0015271838	problem
110	0.0012929463	simulation	104	0.0013243515	conference	10	0.0015271838	ass
108	0.0012694382	season	104	0.0013243515	moment	10	0.0015271838	post
107	0.0012576842	times	102	0.0012988832	get	10	0.0015271838	everything
107	0.0012576842	get	102	0.0012988832	head	10	0.0015271838	campaign
107	0.0012576842	country	101	0.0012861490	song	9	0.0013744654	game
107	0.0012576842	guy	99	0.0012606807	times	9	0.0013744654	vote
107	0.0012576842	children	99	0.0012606807	hour	9	0.0013744654	fans
106	0.0012459301	s	99	0.0012606807	twitter	9	0.0013744654	reflections
105	0.0012341760	conference	98	0.0012479466	body	9	0.0013744654	hey
105	0.0012341760	others	97	0.0012352124	d	9	0.0013744654	disease
105	0.0012341760	please	96	0.0012224783	money	9	0.0013744654	matter
105	0.0012341760	twitter	96	0.0012224783	children	9	0.0013744654	words
104	0.0012224220	head	96	0.0012224783	lol	9	0.0013744654	porn
104	0.0012224220	moment	96	0.0012224783	poetry	9	0.0013744654	one

Taula 18

dataset_1.train						
labels	paraules		adjectius		substantius	
	tipus	ocurrències	tipus	ocurrències	tipus	ocurrències
all	6595	41450	1216	3622	3865	10024
positive	4081	19399	694	1625	2365	4712
negative	4072	22051	762	1997	2188	5312

Taula 19

Freqüència i distribució dels adjectius al joc d'entrenament del dataset_1								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
117	0.0321252059	happy	108	0.0656534954	happy	107	0.0535803705	white
114	0.0313014827	white	58	0.0352583586	new	70	0.0350525788	black
103	0.0282811641	new	52	0.0316109422	good	68	0.0340510766	amp
87	0.0238879736	amp	30	0.0182370820	thankful	45	0.0225338007	new
78	0.0214168039	good	30	0.0182370820	great	29	0.0145217826	racist
77	0.0211422295	black	28	0.0170212765	best	28	0.0140210315	right
54	0.0148270181	great	26	0.0158054711	first	26	0.0130195292	good
41	0.0112575507	more	25	0.0151975683	last	24	0.0120180270	great
37	0.0101592531	first	24	0.0145896656	more	22	0.0110165247	latest
36	0.0098846787	best	21	0.0127659574	polar	20	0.0100150225	own
34	0.0093355299	right	19	0.0115501519	amp	18	0.0090135202	same
33	0.0090609555	last	18	0.0109422492	old	17	0.0085127691	many
31	0.0085118066	many	17	0.0103343465	next	17	0.0085127691	more
30	0.0082372322	thankful	16	0.0097264437	angry	16	0.0080120180	other
29	0.0079626578	racist	16	0.0097264437	little	14	0.0070105157	free
26	0.0071389346	much	15	0.0091185410	sad	14	0.0070105157	daily
26	0.0071389346	next	15	0.0091185410	big	13	0.0065097646	much
25	0.0068643602	own	14	0.0085106382	ready	13	0.0065097646	real
24	0.0065897858	old	14	0.0085106382	many	12	0.0060090135	american
24	0.0065897858	other	13	0.0079027355	much	11	0.0055082623	thought
24	0.0065897858	latest	11	0.0066869300	strong	11	0.0055082623	first
23	0.0063152114	free	11	0.0066869300	few	9	0.0045067601	anti
22	0.0060406370	same	10	0.0060790273	beautiful	9	0.0045067601	better
21	0.0057660626	little	10	0.0060790273	excited	9	0.0045067601	next
21	0.0057660626	big	9	0.0054711246	free	9	0.0045067601	happy
21	0.0057660626	polar	8	0.0048632218	better	9	0.0045067601	true
20	0.0054914881	angry	8	0.0048632218	useful	9	0.0045067601	porn
19	0.0052169137	sad	8	0.0048632218	cold	8	0.0040060090	best
18	0.0049423393	real	8	0.0048632218	other	8	0.0040060090	islamic
18	0.0049423393	daily	8	0.0048632218	true	8	0.0040060090	uk
17	0.0046677649	better	8	0.0048632218	amazing	8	0.0040060090	such
17	0.0046677649	true	7	0.0042553191	nice	8	0.0040060090	bad
15	0.0041186161	strong	7	0.0042553191	black	8	0.0040060090	most
15	0.0041186161	few	7	0.0042553191	top	8	0.0040060090	racial
14	0.0038440417	american	7	0.0042553191	live	8	0.0040060090	african
14	0.0038440417	ready	7	0.0042553191	open	8	0.0040060090	last
14	0.0038440417	bad	7	0.0042553191	huge	7	0.0035052578	different
14	0.0038440417	most	7	0.0042553191	white	7	0.0035052578	non
14	0.0038440417	such	6	0.0036474164	bad	7	0.0035052578	worst
12	0.0032948929	hard	6	0.0036474164	most	7	0.0035052578	wrong
11	0.0030203185	wrong	6	0.0036474164	right	7	0.0035052578	fat
11	0.0030203185	excited	6	0.0036474164	hard	7	0.0035052578	mad
11	0.0030203185	thought	6	0.0036474164	such	6	0.0030045067	dear
11	0.0030203185	open	6	0.0036474164	friday	6	0.0030045067	old
11	0.0030203185	nice	6	0.0036474164	birthday	6	0.0030045067	semitic
11	0.0030203185	beautiful	6	0.0036474164	wonderful	6	0.0030045067	stupid
11	0.0030203185	live	5	0.0030395136	late	6	0.0030045067	fuck
11	0.0030203185	cold	5	0.0030395136	sunday	6	0.0030045067	jewish
10	0.0027457440	different	5	0.0030395136	vast	6	0.0030045067	big
10	0.0027457440	uk	5	0.0030395136	sure	6	0.0030045067	systemic
10	0.0027457440	only	5	0.0030395136	empty	6	0.0030045067	asian

10	0.0027457440	mad	5	0.0030395136	famous	6	0.0030045067	denial
9	0.0024711696	anti	5	0.0030395136	awesome	6	0.0030045067	single
9	0.0024711696	t	5	0.0030395136	perfect	6	0.0030045067	nude
9	0.0024711696	stupid	5	0.0030395136	real	6	0.0030045067	hard
9	0.0024711696	islamic	5	0.0030395136	only	6	0.0030045067	dangerous
9	0.0024711696	single	5	0.0030395136	young	6	0.0030045067	brown
9	0.0024711696	nude	5	0.0030395136	lucky	5	0.0025037556	t
9	0.0024711696	young	5	0.0030395136	own	5	0.0025037556	palestinian
9	0.0024711696	porn	4	0.0024316109	wrong	5	0.0025037556	wish
8	0.0021965952	late	4	0.0024316109	t	5	0.0025037556	blicqer
8	0.0021965952	wonderful	4	0.0024316109	same	5	0.0025037556	ag
8	0.0021965952	top	4	0.0024316109	previous	5	0.0025037556	only
8	0.0021965952	sure	4	0.0024316109	early	5	0.0025037556	co
8	0.0021965952	useful	4	0.0024316109	sexy	5	0.0025037556	human
8	0.0021965952	racial	4	0.0024316109	long	5	0.0025037556	obama
8	0.0021965952	african	4	0.0024316109	bull	5	0.0025037556	pa
8	0.0021965952	amazing	4	0.0024316109	gorgeous	5	0.0025037556	social
7	0.0019220208	dear	4	0.0024316109	grr	5	0.0025037556	evil
7	0.0019220208	important	4	0.0024316109	pathetic	5	0.0025037556	little
7	0.0019220208	non	4	0.0024316109	dead	5	0.0025037556	dumb
7	0.0019220208	early	4	0.0024316109	daily	5	0.0025037556	sexist
7	0.0019220208	wish	4	0.0024316109	dont	4	0.0020030045	important
7	0.0019220208	worst	4	0.0024316109	safe	4	0.0020030045	illegal
7	0.0019220208	sick	4	0.0024316109	cute	4	0.0020030045	male
7	0.0019220208	dumb	4	0.0024316109	follow	4	0.0020030045	present
7	0.0019220208	awesome	4	0.0024316109	lovely	4	0.0020030045	sad
7	0.0019220208	ok	4	0.0024316109	orlando	4	0.0020030045	enough
7	0.0019220208	special	3	0.0018237082	third	4	0.0020030045	teen
7	0.0019220208	dead	3	0.0018237082	clean	4	0.0020030045	least
7	0.0019220208	fat	3	0.0018237082	hot	4	0.0020030045	open
7	0.0019220208	huge	3	0.0018237082	different	4	0.0020030045	small
7	0.0019220208	safe	3	0.0018237082	important	4	0.0020030045	michelle
7	0.0019220208	deep	3	0.0018237082	nohern	4	0.0020030045	total
6	0.0016474464	semitic	3	0.0018237082	tonight	4	0.0020030045	alive
6	0.0016474464	hot	3	0.0018237082	able	4	0.0020030045	trash
6	0.0016474464	fuck	3	0.0018237082	sorry	4	0.0020030045	sick
6	0.0016474464	jewish	3	0.0018237082	fabulous	4	0.0020030045	hateful
6	0.0016474464	least	3	0.0018237082	nude	4	0.0020030045	hopeful
6	0.0016474464	systemic	3	0.0018237082	quick	4	0.0020030045	poor
6	0.0016474464	sexy	3	0.0018237082	w	4	0.0020030045	nice
6	0.0016474464	small	3	0.0018237082	entire	4	0.0020030045	mexican
6	0.0016474464	alive	3	0.0018237082	queen	4	0.0020030045	less
6	0.0016474464	asian	3	0.0018237082	ok	4	0.0020030045	civil
6	0.0016474464	w	3	0.0018237082	single	4	0.0020030045	young
6	0.0016474464	denial	3	0.0018237082	longer	4	0.0020030045	guilty
6	0.0016474464	poor	3	0.0018237082	national	4	0.0020030045	blatant
6	0.0016474464	less	3	0.0018237082	brilliant	4	0.0020030045	shit
6	0.0016474464	excellent	3	0.0018237082	stupid	4	0.0020030045	common
6	0.0016474464	friday	3	0.0018237082	mad	4	0.0020030045	special

Taula 20

Freqüència i distribució dels substantius al joc d'entrenament del dataset_1								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
156	0.0154900208	amp	131	0.0275267913	day	113	0.0212725903	amp
137	0.0136034157	day	74	0.0155494851	time	68	0.0128012048	people
119	0.0118161056	people	63	0.0132380752	today	65	0.0122364457	libtard
92	0.0091351405	time	51	0.0107165370	people	56	0.0105421686	women
73	0.0072485353	today	43	0.0090355116	amp	50	0.0094126506	racism
65	0.0064541753	libtard	36	0.0075646144	birthday	48	0.0090361445	racist
63	0.0062555853	women	33	0.0069342298	life	47	0.0088478915	trump
58	0.0057591103	trump	32	0.0067241017	way	34	0.0064006024	woman
53	0.0052626352	racist	29	0.0060937171	days	32	0.0060240963	men
51	0.0050640452	racism	28	0.0058835889	tomorrow	30	0.0056475903	man
45	0.0044682752	way	27	0.0056734608	things	28	0.0052710843	comments
43	0.0042696852	man	26	0.0054633326	week	27	0.0050828313	sex
42	0.0041703902	year	26	0.0054633326	morning	26	0.0048945783	race
39	0.0038725052	life	24	0.0050430762	work	24	0.0045180722	video
38	0.0037732102	woman	23	0.0048329480	father	23	0.0043298192	obama
38	0.0037732102	world	22	0.0046228199	night	23	0.0043298192	world
36	0.0035746201	birthday	21	0.0044126917	year	21	0.0039533132	america
35	0.0034753251	video	20	0.0042025635	family	21	0.0039533132	condemns
34	0.0033760301	things	20	0.0042025635	home	21	0.0039533132	act
34	0.0033760301	men	18	0.0037823072	game	21	0.0039533132	girl
34	0.0033760301	days	16	0.0033620508	poetry	21	0.0039533132	suppoers
33	0.0032767351	girl	16	0.0033620508	weekend	21	0.0039533132	year
33	0.0032767351	sex	15	0.0031519226	friday	20	0.0037650602	girls
32	0.0031774401	work	15	0.0031519226	world	20	0.0037650602	thanks
32	0.0031774401	thanks	15	0.0031519226	everyone	20	0.0037650602	pay
30	0.0029788501	family	14	0.0029417944	gt	19	0.0035768072	nothing
29	0.0028795551	week	13	0.0027316663	man	19	0.0035768072	guy
28	0.0027802601	comments	13	0.0027316663	thank	18	0.0033885542	time
28	0.0027802601	morning	12	0.0025215381	city	17	0.0032003012	sea
28	0.0027802601	tomorrow	12	0.0025215381	friends	17	0.0032003012	shepherd
27	0.0026809651	girls	12	0.0025215381	love	17	0.0032003012	country
27	0.0026809651	race	12	0.0025215381	weeks	17	0.0032003012	word
27	0.0026809651	nothing	12	0.0025215381	girl	16	0.0030120481	president
26	0.0025816701	father	12	0.0025215381	thanks	16	0.0030120481	factory
25	0.0024823751	obama	11	0.0023114099	trump	16	0.0030120481	police
25	0.0024823751	everyone	11	0.0023114099	fathers	14	0.0026355421	someone
25	0.0024823751	home	11	0.0023114099	something	14	0.0026355421	w
24	0.0023830801	night	11	0.0023114099	music	13	0.0024472891	person
24	0.0023830801	person	11	0.0023114099	dad	13	0.0024472891	s
24	0.0023830801	act	11	0.0023114099	person	13	0.0024472891	history
24	0.0023830801	guy	11	0.0023114099	bull	13	0.0024472891	news
24	0.0023830801	pay	11	0.0023114099	video	13	0.0024472891	fuck
23	0.0022837851	suppoers	10	0.0021012817	head	13	0.0024472891	way
23	0.0022837851	game	10	0.0021012817	bear	13	0.0024472891	thing
23	0.0022837851	news	10	0.0021012817	news	13	0.0024472891	school
22	0.0021844901	someone	10	0.0021012817	sunday	12	0.0022590361	years
22	0.0021844901	thing	10	0.0021012817	hours	12	0.0022590361	job
21	0.0020851951	america	10	0.0021012817	orlando	12	0.0022590361	media
21	0.0020851951	condemns	10	0.0021012817	kids	12	0.0022590361	kind
21	0.0020851951	s	9	0.0018911536	dance	12	0.0022590361	please
20	0.0019859001	hate	9	0.0018911536	tonight	12	0.0022590361	whites
20	0.0019859001	gt	9	0.0018911536	thing	11	0.0020707831	message
20	0.0019859001	country	9	0.0018911536	team	11	0.0020707831	hate
20	0.0019859001	word	9	0.0018911536	attack	11	0.0020707831	paladino
19	0.0018866051	kind	9	0.0018911536	hate	11	0.0020707831	remarks
19	0.0018866051	job	9	0.0018911536	side	11	0.0020707831	jews
19	0.0018866051	team	9	0.0018911536	lot	11	0.0020707831	course
19	0.0018866051	sea	8	0.0016810254	anyone	11	0.0020707831	chair
19	0.0018866051	w	8	0.0016810254	followers	11	0.0020707831	towards
18	0.0017873100	music	8	0.0016810254	s	11	0.0020707831	won
18	0.0017873100	school	8	0.0016810254	d	11	0.0020707831	book

18	0.0017873100	media	8	0.0016810254	someone	11	0.0020707831	comment
18	0.0017873100	police	8	0.0016810254	summer	11	0.0020707831	blacks
18	0.0017873100	attack	8	0.0016810254	money	10	0.0018825301	problem
17	0.0016880150	years	8	0.0016810254	nothing	10	0.0018825301	muslims
17	0.0016880150	thank	8	0.0016810254	house	10	0.0018825301	team
17	0.0016880150	shepherd	8	0.0016810254	show	10	0.0018825301	children
16	0.0015887200	anyone	8	0.0016810254	place	10	0.0018825301	oil
16	0.0015887200	president	8	0.0016810254	happiness	10	0.0018825301	today
16	0.0015887200	poetry	7	0.0014708972	girls	10	0.0018825301	family
16	0.0015887200	kids	7	0.0014708972	tags	10	0.0018825301	feel
16	0.0015887200	friends	7	0.0014708972	kind	10	0.0018825301	color
16	0.0015887200	factory	7	0.0014708972	baby	10	0.0018825301	everyone
16	0.0015887200	children	7	0.0014708972	wow	10	0.0018825301	business
16	0.0015887200	love	7	0.0014708972	hea	10	0.0018825301	look
16	0.0015887200	campaign	7	0.0014708972	post	10	0.0018825301	ppl
16	0.0015887200	weekend	7	0.0014708972	get	10	0.0018825301	christmas
16	0.0015887200	book	7	0.0014708972	gift	9	0.0016942771	polarisation
15	0.0014894250	friday	7	0.0014708972	right	9	0.0016942771	power
15	0.0014894250	history	7	0.0014708972	job	9	0.0016942771	campaign
15	0.0014894250	please	7	0.0014708972	friend	9	0.0016942771	fans
15	0.0014894250	course	7	0.0014708972	car	9	0.0016942771	attack
15	0.0014894250	look	7	0.0014708972	women	9	0.0016942771	matter
14	0.0013901300	city	7	0.0014708972	damn	9	0.0016942771	others
14	0.0013901300	something	7	0.0014708972	campaign	9	0.0016942771	videos
14	0.0013901300	dad	7	0.0014708972	p	8	0.0015060240	lady
14	0.0013901300	post	6	0.0012607690	tips	8	0.0015060240	vote
14	0.0013901300	fans	6	0.0012607690	hope	8	0.0015060240	porn
14	0.0013901300	weeks	6	0.0012607690	name	8	0.0015060240	anyone
14	0.0013901300	house	6	0.0012607690	change	8	0.0015060240	work
14	0.0013901300	ppl	6	0.0012607690	tweets	8	0.0015060240	list
14	0.0013901300	fuck	6	0.0012607690	end	8	0.0015060240	words
13	0.0012908350	message	6	0.0012607690	climb	8	0.0015060240	americans
13	0.0012908350	fathers	6	0.0012607690	places	8	0.0015060240	times
13	0.0012908350	lot	6	0.0012607690	xx	8	0.0015060240	aicle
13	0.0012908350	words	6	0.0012607690	congrats	8	0.0015060240	industry
13	0.0012908350	power	6	0.0012607690	sex	8	0.0015060240	hey
13	0.0012908350	oil	6	0.0012607690	photo	8	0.0015060240	rights
13	0.0012908350	wow	6	0.0012607690	point	8	0.0015060240	one
13	0.0012908350	business	6	0.0012607690	children	8	0.0015060240	everything

Taula 21

dataset_2						
labels	paraules		adjectius		substantius	
	tipus	ocurrències	tipus	ocurrències	tipus	ocurrències
all	26386	440008	5796	35893	17344	100578
positive	17908	214252	3653	18308	11670	51255
negative	16744	225756	3599	17585	10672	49323

Taula 22

Freqüència i distribució dels adjectius al dataset_2 sencer								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
1950	0.0543281419	good	1392	0.0760323355	good	558	0.0317315894	good
879	0.0244894547	new	574	0.0313524142	new	468	0.0266135911	last
703	0.0195859916	last	504	0.0275289490	great	417	0.0237133920	bad
671	0.0186944529	url	474	0.0258903211	url	415	0.0235996588	sad
669	0.0186387317	great	340	0.0185711164	happy	329	0.0187091270	much
549	0.0152954615	bad	340	0.0185711164	nice	305	0.0173443275	new
548	0.0152676009	much	235	0.0128359187	last	277	0.0157520614	sick
491	0.0136795475	more	227	0.0123989512	more	264	0.0150127949	more
481	0.0134009416	nice	219	0.0119619838	much	197	0.0112027295	url
457	0.0127322876	happy	202	0.0110334280	quot	178	0.0101222632	sorry
444	0.0123701000	sad	196	0.0107057024	best	174	0.0098947967	little
345	0.0096119020	little	193	0.0105418396	awesome	171	0.0097241967	poor
313	0.0087203633	quot	180	0.0098317675	first	165	0.0093829968	great
303	0.0084417574	first	171	0.0093401791	little	143	0.0081319306	cold
300	0.0083581756	sick	161	0.0087939698	ready	141	0.0080181973	nice
298	0.0083024545	next	159	0.0086847279	next	139	0.0079044640	next
286	0.0079681274	best	148	0.0080838977	sure	126	0.0071651976	long
285	0.0079402669	ready	139	0.0075923093	better	124	0.0070514643	ready
254	0.0070765887	sure	132	0.0072099628	bad	123	0.0069945976	first
246	0.0068537040	awesome	130	0.0071007209	big	120	0.0068239977	tired
241	0.0067144011	old	126	0.0068822372	old	117	0.0066533977	happy
230	0.0064079346	better	126	0.0068822372	cool	115	0.0065396644	old
219	0.0061014682	big	114	0.0062267861	beautiful	113	0.0064259311	early
218	0.0060736076	long	111	0.0060629233	other	111	0.0063121978	many
215	0.0059900259	other	100	0.0054620930	same	111	0.0063121978	quot
210	0.0058507229	same	97	0.0052982302	glad	110	0.0062553312	same
204	0.0056835594	sorry	96	0.0052436093	few	108	0.0061415979	stupid
199	0.0055442565	many	93	0.0050797465	welcome	107	0.0060847313	hard
196	0.0054606747	cold	92	0.0050251256	long	106	0.0060278646	sure
190	0.0052935112	poor	91	0.0049705046	free	104	0.0059141313	other
185	0.0051542083	few	88	0.0048066419	many	96	0.0054591981	wish
184	0.0051263477	beautiful	88	0.0048066419	ok	91	0.0051748649	better
174	0.0048477418	cool	81	0.0044242953	funny	91	0.0051748649	late
171	0.0047641601	hard	79	0.0043150535	right	90	0.0051179982	best
161	0.0044855542	tired	78	0.0042604325	least	89	0.0050611316	few
161	0.0044855542	early	78	0.0042604325	lol	89	0.0050611316	big
159	0.0044298331	ok	78	0.0042604325	amp	79	0.0044924651	hot
152	0.0042348090	least	78	0.0042604325	cute	76	0.0043218652	able
152	0.0042348090	late	74	0.0040419488	sweet	74	0.0042081319	least
151	0.0042069484	hot	73	0.0039873279	excited	73	0.0041512652	wrong
139	0.0038726214	right	72	0.0039327070	hot	71	0.0040375319	ok
130	0.0036218761	amp	70	0.0038234651	amazing	70	0.0039806653	beautiful
128	0.0035661549	lol	66	0.0036049814	such	67	0.0038100653	busy
125	0.0034825732	free	66	0.0036049814	own	62	0.0035257321	only
124	0.0034547126	stupid	65	0.0035503604	real	61	0.0034688655	ugh
121	0.0033711308	wish	64	0.0034957395	hard	60	0.0034119988	right
120	0.0033432702	glad	61	0.0033318767	wonderful	58	0.0032982655	ta
120	0.0033432702	real	61	0.0033318767	late	57	0.0032413989	soo
119	0.0033154096	busy	60	0.0032772558	true	56	0.0031845322	hungry
116	0.0032318279	ta	58	0.0031680139	ta	55	0.0031276656	dead

114	0.0031761067	able	55	0.0030041511	whole	55	0.0031276656	real
113	0.0031482461	such	53	0.0028949093	cold	54	0.0030707989	horrible
109	0.0030368038	own	52	0.0028402883	busy	53	0.0030139323	awesome
106	0.0029532220	funny	48	0.0026218046	early	52	0.0029570656	amp
104	0.0028975009	true	48	0.0026218046	full	50	0.0028433323	lol
104	0.0028975009	wrong	46	0.0025125628	fine	50	0.0028433323	outside
103	0.0028696403	soo	46	0.0025125628	soo	50	0.0028433323	worst
102	0.0028417797	whole	45	0.0024579418	sunny	49	0.0027864657	sleep
102	0.0028417797	sweet	43	0.0023487000	lovely	49	0.0027864657	full
100	0.0027860585	welcome	42	0.0022940790	red	48	0.0027295990	cool
99	0.0027581979	only	42	0.0022940790	interesting	47	0.0026727324	whole
98	0.0027303373	cute	41	0.0022394581	tired	47	0.0026727324	such
97	0.0027024768	full	41	0.0022394581	special	46	0.0026158657	jealous
94	0.0026188950	amazing	40	0.0021848372	pretty	45	0.0025589991	red
91	0.0025353132	excited	40	0.0021848372	fun	45	0.0025589991	enough
87	0.0024238709	red	38	0.0020755953	easy	45	0.0025589991	bored
84	0.0023402891	sunny	38	0.0020755953	able	44	0.0025021324	true
79	0.0022009862	sleep	38	0.0020755953	green	43	0.0024452658	na
76	0.0021174045	crazy	37	0.0020209744	only	43	0.0024452658	own
76	0.0021174045	wonderful	37	0.0020209744	different	42	0.0023883992	saturday
75	0.0020895439	hungry	36	0.0019663535	crazy	42	0.0023883992	super
75	0.0020895439	fun	35	0.0019117325	lt	41	0.0023315325	worse
74	0.0020616833	dead	35	0.0019117325	fantastic	40	0.0022746659	crazy
74	0.0020616833	super	34	0.0018571116	short	40	0.0022746659	due
72	0.0020059621	ugh	34	0.0018571116	lucky	39	0.0022177992	oh
71	0.0019781015	outside	34	0.0018571116	favorite	39	0.0022177992	boring
70	0.0019502409	oh	33	0.0018024907	live	39	0.0022177992	snow
69	0.0019223804	most	33	0.0018024907	perfect	39	0.0022177992	sunny
69	0.0019223804	pretty	33	0.0018024907	most	38	0.0021609326	terrible
68	0.0018945198	fine	33	0.0018024907	twitter	38	0.0021609326	slow
68	0.0018945198	enough	32	0.0017478697	super	36	0.0020471993	ive
67	0.0018666592	saturday	32	0.0017478697	extra	36	0.0020471993	most
65	0.0018109380	easy	32	0.0017478697	high	36	0.0020471993	tonight
64	0.0017830774	jealous	31	0.0016932488	wrong	35	0.0019903326	upset
64	0.0017830774	short	31	0.0016932488	oh	35	0.0019903326	fun
61	0.0016994957	interesting	31	0.0016932488	gorgeous	34	0.0019334660	free
60	0.0016716351	green	30	0.0016386279	sleep	32	0.0018197327	uni
60	0.0016716351	lovely	30	0.0016386279	black	32	0.0018197327	sigh
59	0.0016437745	ive	30	0.0016386279	online	32	0.0018197327	female
59	0.0016437745	na	29	0.0015840069	sad	32	0.0018197327	okay
58	0.0016159139	favorite	28	0.0015293860	hilarious	32	0.0018197327	ill
58	0.0016159139	horrible	28	0.0015293860	second	31	0.0017628660	english
57	0.0015880533	bored	28	0.0015293860	less	31	0.0017628660	white
56	0.0015601927	tonight	28	0.0015293860	delicious	30	0.0017059994	mad
56	0.0015601927	lt	27	0.0014747651	productive	30	0.0017059994	small
56	0.0015601927	special	26	0.0014201441	sorry	30	0.0017059994	huge
56	0.0015601927	high	26	0.0014201441	friday	30	0.0017059994	short
55	0.0015323322	black	25	0.0013655232	wish	29	0.0016491327	pretty
55	0.0015323322	due	25	0.0013655232	saturday	28	0.0015922661	sweet
55	0.0015323322	different	25	0.0013655232	quick	28	0.0015922661	ouch

Taula 23

Freqüència i distribució dels substantius al dataset_2 sencer								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
1805	0.0179462705	day	936	0.0182616330	day	869	0.0176185552	day
1539	0.0153015570	today	675	0.0131694468	today	864	0.0175171826	today
1189	0.0118216707	time	667	0.0130133645	thanks	714	0.0144760051	work
1028	0.0102209230	work	599	0.0116866647	time	590	0.0119619650	time
926	0.0092067847	night	574	0.0111989074	morning	461	0.0093465523	night
925	0.0091968422	morning	532	0.0103794751	quot	397	0.0080489832	home
829	0.0082423591	quot	465	0.0090722856	night	351	0.0071163554	morning
797	0.0079241981	thanks	464	0.0090527753	twitter	297	0.0060215315	quot
749	0.0074469565	twitter	433	0.0084479562	url	289	0.0058593354	tomorrow
662	0.0065819562	home	386	0.0075309725	lol	285	0.0057782373	twitter
648	0.0064427608	lol	314	0.0061262315	work	276	0.0055957666	school
641	0.0063731631	url	265	0.0051702272	tomorrow	262	0.0053119234	lol
554	0.0055081628	tomorrow	265	0.0051702272	home	240	0.0048658840	way
493	0.0049016683	way	264	0.0051507170	amp	230	0.0046631389	week
466	0.0046332199	people	261	0.0050921861	haha	220	0.0044603937	people
443	0.0044045417	amp	253	0.0049361037	way	208	0.0042170995	url
407	0.0040466105	days	246	0.0047995317	people	205	0.0041562759	days
402	0.0039968979	week	236	0.0046044288	everyone	190	0.0038521582	hours
401	0.0039869553	school	219	0.0042727538	fun	182	0.0036899620	sorry
366	0.0036389667	haha	207	0.0040386303	thank	179	0.0036291385	amp
346	0.0034401161	everyone	202	0.0039410789	days	177	0.0035885895	house
340	0.0033804609	fun	200	0.0039020583	hey	171	0.0034669424	bed
339	0.0033705184	tonight	172	0.0033557701	week	171	0.0034669424	phone
338	0.0033605758	something	170	0.0033167495	tonight	169	0.0034263933	tonight
331	0.0032909781	thing	169	0.0032972392	something	169	0.0034263933	something
312	0.0031020700	life	168	0.0032777290	life	165	0.0033452952	thing
287	0.0028535067	yeah	166	0.0032387084	thing	152	0.0030817265	man
286	0.0028435641	house	159	0.0031021363	yeah	151	0.0030614520	someone
285	0.0028336216	man	151	0.0029460540	yay	150	0.0030411775	friends
283	0.0028137366	bed	138	0.0026924202	love	144	0.0029195304	life
282	0.0028037940	hours	138	0.0026924202	yes	138	0.0027978833	weather
280	0.0027839090	hey	133	0.0025948687	man	136	0.0027573343	weekend
264	0.0026248284	friends	128	0.0024973173	things	133	0.0026965107	nothing
258	0.0025651732	weekend	125	0.0024387864	school	131	0.0026559617	show
248	0.0024657479	phone	124	0.0024192761	world	130	0.0026356872	thanks
246	0.0024458629	things	123	0.0023997658	music	129	0.0026154126	class
246	0.0024458629	show	122	0.0023802555	weekend	128	0.0025951381	year
237	0.0023563801	someone	121	0.0023607452	movie	128	0.0025951381	yeah
236	0.0023464375	weather	120	0.0023412350	hope	122	0.0024734910	yesterday
234	0.0023265525	sorry	119	0.0023217247	stuff	121	0.0024532165	fun
233	0.0023166099	thank	116	0.0022631938	friend	118	0.0023923930	game
222	0.0022072421	hope	115	0.0022436835	show	118	0.0023923930	things
221	0.0021972996	year	114	0.0022241732	friends	117	0.0023721184	one
210	0.0020879317	nothing	113	0.0022046629	hi	117	0.0023721184	aww
209	0.0020779892	love	112	0.0021851526	bed	116	0.0023518439	car
206	0.0020481616	one	109	0.0021266217	birthday	110	0.0022301968	everyone
205	0.0020382190	yesterday	109	0.0021266217	house	105	0.0021288242	haha
205	0.0020382190	stuff	107	0.0020876012	sun	102	0.0020680007	hope
200	0.0019885064	class	106	0.0020680909	lot	101	0.0020477262	anyone
199	0.0019785639	yay	103	0.0020095600	please	100	0.0020274516	sucks
199	0.0019785639	game	103	0.0020095600	coffee	99	0.0020071771	damn
199	0.0019785639	sun	99	0.0019315188	song	95	0.0019260791	rain
198	0.0019686213	world	98	0.0019120085	weather	95	0.0019260791	anything
196	0.0019487363	friend	93	0.0018144571	year	95	0.0019260791	headache
193	0.0019189087	movie	92	0.0017949468	hours	94	0.0019058045	bit
192	0.0019089661	music	92	0.0017949468	job	94	0.0019058045	baby
185	0.0018393684	please	91	0.0017754365	blog	92	0.0018652555	sun
183	0.0018194833	bit	89	0.0017364159	x	91	0.0018449810	head
183	0.0018194833	lot	89	0.0017364159	party	86	0.0017436084	boo
182	0.0018095408	yes	89	0.0017364159	one	86	0.0017436084	stuff

181	0.0017995983	aww	89	0.0017364159	bit	85	0.0017233339	hour
178	0.0017697707	coffee	88	0.0017169056	wow	83	0.0016827849	dog
173	0.0017200580	job	88	0.0017169056	luck	82	0.0016625103	please
164	0.0016305752	birthday	86	0.0016778850	girl	81	0.0016422358	job
164	0.0016305752	car	86	0.0016778850	someone	80	0.0016219613	hey
164	0.0016305752	baby	86	0.0016778850	tweet	80	0.0016219613	friend
163	0.0016206327	anyone	84	0.0016388645	hello	80	0.0016219613	sleep
162	0.0016106902	girl	83	0.0016193542	yesterday	79	0.0016016868	money
161	0.0016007476	anything	81	0.0015803336	game	77	0.0015611378	spring
160	0.0015908051	party	80	0.0015608233	video	77	0.0015611378	family
152	0.0015112648	x	80	0.0015608233	followers	77	0.0015611378	lot
151	0.0015013223	luck	79	0.0015413130	lunch	77	0.0015611378	pain
150	0.0014913798	wow	78	0.0015218027	book	77	0.0015611378	homework
149	0.0014814372	room	78	0.0015218027	xx	77	0.0015611378	room
146	0.0014516096	everything	77	0.0015022924	glad	76	0.0015408632	girl
145	0.0014416671	hour	77	0.0015022924	phone	76	0.0015408632	throat
145	0.0014416671	food	77	0.0015022924	nothing	75	0.0015205887	company
145	0.0014416671	hair	77	0.0015022924	tweets	75	0.0015205887	weeks
143	0.0014217820	god	77	0.0015022924	name	75	0.0015205887	coffee
142	0.0014118395	times	76	0.0014827821	idea	74	0.0015003142	world
140	0.0013919545	family	76	0.0014827821	god	74	0.0015003142	everything
140	0.0013919545	video	74	0.0014437615	times	74	0.0015003142	computer
140	0.0013919545	followers	74	0.0014437615	hair	72	0.0014597652	food
139	0.0013820119	song	73	0.0014242512	food	72	0.0014597652	movie
138	0.0013720694	xx	72	0.0014047410	room	71	0.0014394907	party
137	0.0013621269	hi	72	0.0014047410	everything	71	0.0014394907	love
136	0.0013521843	sleep	71	0.0013852307	class	71	0.0014394907	hair
135	0.0013422418	head	70	0.0013657204	baby	71	0.0014394907	crap
135	0.0013422418	lunch	69	0.0013462101	pic	70	0.0014192161	mom
134	0.0013322993	tweets	68	0.0013266998	dinner	69	0.0013989416	music
133	0.0013223567	oh	68	0.0013266998	hahaha	68	0.0013786671	times
132	0.0013124142	idea	68	0.0013266998	guys	68	0.0013786671	gon
131	0.0013024717	spring	67	0.0013071895	place	67	0.0013583926	oh
131	0.0013024717	kids	67	0.0013071895	kids	67	0.0013583926	end
130	0.0012925291	guys	66	0.0012876792	congrats	67	0.0013583926	omg
128	0.0012726441	gon	66	0.0012876792	course	67	0.0013583926	god
128	0.0012726441	tweet	66	0.0012876792	anything	64	0.0012975690	break
127	0.0012627015	name	66	0.0012876792	oh	64	0.0012975690	kids
126	0.0012527590	mom	64	0.0012486586	aww	63	0.0012772945	luck
124	0.0012328739	money	64	0.0012486586	ha	63	0.0012772945	x

Taula 24

dataset_2.train						
labels	paraules		adjectius		substantius	
	tipus	ocurrències	tipus	ocurrències	tipus	ocurrències
all	7281	50104	1309	4074	4474	11609
positive	4666	24369	784	2088	2754	5969
negative	4520	25735	800	1986	2649	5721

Taula 25

Freqüència i distribució dels adjectius al joc d'entrenament del dataset_2								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
216	0.0546282245	good	147	0.0745814307	good	69	0.0347957639	good
92	0.0232675771	last	52	0.0263825469	new	60	0.0302571860	last
83	0.0209914011	new	49	0.0248604769	url	48	0.0242057488	bad
74	0.0187152250	url	43	0.0218163368	great	46	0.0231971759	sad
63	0.0159332321	great	37	0.0187721968	happy	32	0.0161371659	more
62	0.0156803237	more	32	0.0162354134	last	31	0.0156328794	sick
59	0.0149215983	bad	30	0.0152207001	more	31	0.0156328794	new
54	0.0136570561	much	30	0.0152207001	best	29	0.0146243066	much
51	0.0128983308	nice	30	0.0152207001	nice	25	0.0126071608	url
48	0.0121396054	happy	25	0.0126839167	much	22	0.0110943015	little
48	0.0121396054	sad	24	0.0121765601	quot	22	0.0110943015	poor
42	0.0106221547	best	23	0.0116692034	first	21	0.0105900151	nice
39	0.0098634294	little	22	0.0111618467	awesome	20	0.0100857286	great
38	0.0096105209	first	21	0.0106544901	next	16	0.0080685829	cold
38	0.0096105209	quot	19	0.0096397767	big	16	0.0080685829	long
34	0.0085988872	sick	17	0.0086250634	welcome	15	0.0075642965	first
32	0.0080930703	next	17	0.0086250634	little	14	0.0070600100	old
32	0.0080930703	awesome	16	0.0081177067	other	14	0.0070600100	quot
28	0.0070814365	other	15	0.0076103500	ok	13	0.0065557236	hard
27	0.0068285280	big	14	0.0071029934	hot	13	0.0065557236	late
27	0.0068285280	old	13	0.0065956367	old	12	0.0060514372	best
23	0.0058168942	long	12	0.0060882800	few	12	0.0060514372	same
23	0.0058168942	cold	11	0.0055809233	bad	12	0.0060514372	tired
23	0.0058168942	hot	11	0.0055809233	sure	12	0.0060514372	other
23	0.0058168942	poor	11	0.0055809233	ready	12	0.0060514372	sorry
22	0.0055639858	ok	11	0.0055809233	amp	11	0.0055471507	next
21	0.0053110773	few	11	0.0055809233	beautiful	11	0.0055471507	better
21	0.0053110773	sure	10	0.0050735667	right	11	0.0055471507	happy
20	0.0050581689	ready	9	0.0045662100	real	11	0.0055471507	wish
20	0.0050581689	same	9	0.0045662100	cool	10	0.0050428643	sleep
20	0.0050581689	hard	9	0.0045662100	own	10	0.0050428643	red
19	0.0048052604	late	8	0.0040588533	same	10	0.0050428643	horrible
19	0.0048052604	right	8	0.0040588533	wonderful	10	0.0050428643	fun
18	0.0045523520	welcome	8	0.0040588533	many	10	0.0050428643	awesome
18	0.0045523520	beautiful	8	0.0040588533	sweet	10	0.0050428643	many
18	0.0045523520	many	8	0.0040588533	funny	10	0.0050428643	sure
17	0.0042994436	better	7	0.0035514967	ta	9	0.0045385779	ready
17	0.0042994436	amp	7	0.0035514967	glad	9	0.0045385779	few
16	0.0040465351	fun	7	0.0035514967	long	9	0.0045385779	right
16	0.0040465351	sorry	7	0.0035514967	twitter	9	0.0045385779	hot
15	0.0037936267	red	7	0.0035514967	cold	8	0.0040342914	big
15	0.0037936267	real	7	0.0035514967	excited	8	0.0040342914	early
15	0.0037936267	tired	7	0.0035514967	free	8	0.0040342914	stupid
15	0.0037936267	wish	7	0.0035514967	hard	7	0.0035300050	most
14	0.0035407182	early	6	0.0030441400	late	7	0.0035300050	true
14	0.0035407182	cool	6	0.0030441400	fine	7	0.0035300050	na
14	0.0035407182	wonderful	6	0.0030441400	sunny	7	0.0035300050	lol
13	0.0032878098	sleep	6	0.0030441400	hilarious	7	0.0035300050	beautiful
13	0.0032878098	funny	6	0.0030441400	lucky	7	0.0035300050	ok

13	0.0032878098	own	6	0.0030441400	fun	7	0.0035300050	only
13	0.0032878098	least	6	0.0030441400	better	7	0.0035300050	soo
12	0.0030349013	true	6	0.0030441400	early	7	0.0035300050	least
12	0.0030349013	lol	6	0.0030441400	least	7	0.0035300050	worst
11	0.0027819929	glad	6	0.0030441400	second	7	0.0035300050	enough
11	0.0027819929	sunny	5	0.0025367833	interesting	6	0.0030257186	real
11	0.0027819929	most	5	0.0025367833	lol	6	0.0030257186	able
11	0.0027819929	horrible	5	0.0025367833	delicious	6	0.0030257186	boring
11	0.0027819929	soo	5	0.0025367833	gorgeous	6	0.0030257186	wonderful
10	0.0025290844	super	5	0.0025367833	different	6	0.0030257186	jealous
10	0.0025290844	ta	5	0.0025367833	lt	6	0.0030257186	saturday
10	0.0025290844	twitter	5	0.0025367833	cute	6	0.0030257186	amp
10	0.0025290844	sweet	5	0.0025367833	amazing	6	0.0030257186	bored
10	0.0025290844	free	5	0.0025367833	red	6	0.0030257186	ugh
10	0.0025290844	na	5	0.0025367833	true	5	0.0025214321	super
10	0.0025290844	jealous	5	0.0025367833	super	5	0.0025214321	due
10	0.0025290844	only	5	0.0025367833	pretty	5	0.0025214321	whole
9	0.0022761760	fine	5	0.0025367833	busy	5	0.0025214321	funny
9	0.0022761760	busy	4	0.0020294266	videos	5	0.0025214321	serious
9	0.0022761760	different	4	0.0020294266	fabulous	5	0.0025214321	cool
9	0.0022761760	such	4	0.0020294266	crazy	5	0.0025214321	such
9	0.0022761760	excited	4	0.0020294266	strange	5	0.0025214321	rainy
9	0.0022761760	uni	4	0.0020294266	soo	5	0.0025214321	wrong
8	0.0020232675	lt	4	0.0020294266	most	5	0.0025214321	dear
8	0.0020232675	lucky	4	0.0020294266	massive	5	0.0025214321	uni
8	0.0020232675	second	4	0.0020294266	wish	5	0.0025214321	expensive
8	0.0020232675	bored	4	0.0020294266	such	5	0.0025214321	past
8	0.0020232675	ugh	4	0.0020294266	o	5	0.0025214321	open
8	0.0020232675	stupid	4	0.0020294266	online	5	0.0025214321	hungry
8	0.0020232675	special	4	0.0020294266	exciting	5	0.0025214321	sunny
8	0.0020232675	worst	4	0.0020294266	uni	4	0.0020171457	busy
8	0.0020232675	pretty	4	0.0020294266	jealous	4	0.0020171457	different
7	0.0017703591	full	4	0.0020294266	high	4	0.0020171457	safe
7	0.0017703591	lovely	4	0.0020294266	fellow	4	0.0020171457	short
7	0.0017703591	able	4	0.0020294266	special	4	0.0020171457	worse
7	0.0017703591	interesting	4	0.0020294266	re	4	0.0020171457	lovely
7	0.0017703591	whole	4	0.0020294266	sorry	4	0.0020171457	bright
7	0.0017703591	safe	4	0.0020294266	live	4	0.0020171457	usual
7	0.0017703591	english	4	0.0020294266	quick	4	0.0020171457	puppy
7	0.0017703591	oh	4	0.0020294266	ish	4	0.0020171457	oh
7	0.0017703591	amazing	4	0.0020294266	smart	4	0.0020171457	sleepy
7	0.0017703591	saturday	3	0.0015220700	responsible	4	0.0020171457	full
7	0.0017703591	xx	3	0.0015220700	full	4	0.0020171457	own
7	0.0017703591	high	3	0.0015220700	vicky	4	0.0020171457	xx
7	0.0017703591	open	3	0.0015220700	lovely	4	0.0020171457	glad
7	0.0017703591	enough	3	0.0015220700	complete	4	0.0020171457	low
6	0.0015174506	due	3	0.0015220700	curious	4	0.0020171457	painful
6	0.0015174506	smart	3	0.0015220700	earlier	4	0.0020171457	white
6	0.0015174506	hungry	3	0.0015220700	odd	4	0.0020171457	female
6	0.0015174506	short	3	0.0015220700	english	4	0.0020171457	english
6	0.0015174506	serious	3	0.0015220700	fantastic	4	0.0020171457	fat

Taula 26

Freqüència i distribució dels substantius al joc d'entrenament del dataset_2								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
203	0.0176399026	today	116	0.0198868506	day	117	0.0206167400	today
202	0.0175530066	day	86	0.0147436996	today	86	0.0151541850	day
136	0.0118178658	time	77	0.0132007543	morning	83	0.0146255506	work
121	0.0105144247	morning	75	0.0128578775	thanks	61	0.0107488986	time
115	0.0099930483	work	75	0.0128578775	time	44	0.0077533039	home
107	0.0092978797	night	73	0.0125150008	night	44	0.0077533039	morning
88	0.0076468543	twitter	61	0.0104577404	quot	38	0.0066960352	twitter
84	0.0072992700	quot	53	0.0090862334	url	34	0.0059911894	night
83	0.0072123740	thanks	50	0.0085719183	lol	33	0.0058149779	way
82	0.0071254779	lol	50	0.0085719183	twitter	32	0.0056387665	lol
82	0.0071254779	url	35	0.0060003428	fun	30	0.0052863436	week
77	0.0066909975	home	34	0.0058289045	amp	29	0.0051101321	url
63	0.0054744525	way	33	0.0056574661	home	26	0.0045814977	people
57	0.0049530761	amp	32	0.0054860277	work	25	0.0044052863	tonight
56	0.0048661800	week	32	0.0054860277	tomorrow	25	0.0044052863	school
56	0.0048661800	tomorrow	30	0.0051431510	way	24	0.0042290748	tomorrow
52	0.0045185957	people	27	0.0046288359	hey	23	0.0040528634	thing
47	0.0040841153	fun	26	0.0044573975	week	23	0.0040528634	amp
47	0.0040841153	school	26	0.0044573975	people	23	0.0040528634	quot
46	0.0039972193	tonight	22	0.0037716440	haha	21	0.0037004405	house
40	0.0034758428	life	22	0.0037716440	school	21	0.0037004405	weather
39	0.0033889468	thing	21	0.0036002057	tonight	21	0.0037004405	life
37	0.0032151546	days	21	0.0036002057	days	20	0.0035242290	hours
34	0.0029544664	haha	20	0.0034287673	love	20	0.0035242290	sorry
34	0.0029544664	house	20	0.0034287673	yay	18	0.0031718061	yesterday
32	0.0027806743	someone	20	0.0034287673	show	17	0.0029955947	bed
32	0.0027806743	hours	19	0.0032573289	yes	16	0.0028193832	days
31	0.0026937782	hey	19	0.0032573289	life	16	0.0028193832	someone
31	0.0026937782	weather	18	0.0030858906	everyone	16	0.0028193832	phone
31	0.0026937782	show	18	0.0030858906	hope	15	0.0026431718	game
29	0.0025199860	music	18	0.0030858906	stuff	15	0.0026431718	man
29	0.0025199860	love	18	0.0030858906	friend	15	0.0026431718	aww
29	0.0025199860	something	17	0.0029144522	music	15	0.0026431718	nothing
28	0.0024330900	man	16	0.0027430138	someone	15	0.0026431718	one
28	0.0024330900	everyone	16	0.0027430138	something	15	0.0026431718	rain
28	0.0024330900	weekend	16	0.0027430138	thing	15	0.0026431718	class
28	0.0024330900	friend	16	0.0027430138	thank	15	0.0026431718	friends
28	0.0024330900	one	15	0.0025715755	song	14	0.0024669603	yeah
27	0.0023461939	yeah	15	0.0025715755	world	14	0.0024669603	hour
27	0.0023461939	phone	14	0.0024001371	tweets	14	0.0024669603	weekend
27	0.0023461939	hope	14	0.0024001371	weekend	13	0.0022907488	weeks
26	0.0022592978	yesterday	14	0.0024001371	lot	13	0.0022907488	everything
26	0.0022592978	stuff	14	0.0024001371	hi	13	0.0022907488	name
26	0.0022592978	friends	13	0.0022286987	man	13	0.0022907488	year
25	0.0021724018	yay	13	0.0022286987	house	13	0.0022907488	something
25	0.0021724018	bed	13	0.0022286987	lunch	13	0.0022907488	car
25	0.0021724018	nothing	13	0.0022286987	party	13	0.0022907488	money
24	0.0020855057	yes	13	0.0022286987	one	12	0.0021145374	music
24	0.0020855057	party	13	0.0022286987	yeah	12	0.0021145374	please
23	0.0019986096	lot	13	0.0022286987	hair	12	0.0021145374	fun
23	0.0019986096	things	12	0.0020572604	followers	12	0.0021145374	baby
23	0.0019986096	year	12	0.0020572604	wow	12	0.0021145374	anything
23	0.0019986096	song	12	0.0020572604	food	12	0.0021145374	things
23	0.0019986096	world	12	0.0020572604	idea	12	0.0021145374	haha
22	0.0019117135	anything	12	0.0020572604	hours	11	0.0019383259	b
22	0.0019117135	name	12	0.0020572604	girl	11	0.0019383259	party
22	0.0019117135	class	12	0.0020572604	luck	11	0.0019383259	kids
22	0.0019117135	food	11	0.0018858220	phone	11	0.0019383259	x
22	0.0019117135	girl	11	0.0018858220	things	11	0.0019383259	pain
22	0.0019117135	sorry	11	0.0018858220	hmm	11	0.0019383259	show
21	0.0018248175	game	11	0.0018858220	friends	10	0.0017621145	coffee

21	0.0018248175	tweets	11	0.0018858220	movie	10	0.0017621145	throat
21	0.0018248175	aww	10	0.0017143836	blog	10	0.0017621145	friend
20	0.0017379214	coffee	10	0.0017143836	sun	10	0.0017621145	cold
20	0.0017379214	followers	10	0.0017143836	look	10	0.0017621145	sleep
20	0.0017379214	hair	10	0.0017143836	anything	10	0.0017621145	boo
20	0.0017379214	everything	10	0.0017143836	weather	10	0.0017621145	girl
20	0.0017379214	summer	10	0.0017143836	year	10	0.0017621145	sucks
20	0.0017379214	x	10	0.0017143836	d	10	0.0017621145	everyone
19	0.0016510253	please	10	0.0017143836	summer	10	0.0017621145	office
19	0.0016510253	sun	10	0.0017143836	birthday	10	0.0017621145	summer
19	0.0016510253	job	10	0.0017143836	nothing	10	0.0017621145	food
19	0.0016510253	idea	10	0.0017143836	coffee	10	0.0017621145	anyone
19	0.0016510253	rain	10	0.0017143836	job	9	0.0015859030	lot
19	0.0016510253	money	10	0.0017143836	guys	9	0.0015859030	computer
18	0.0015641293	weeks	9	0.0015429453	family	9	0.0015859030	damn
18	0.0015641293	hour	9	0.0015429453	myspace	9	0.0015859030	brother
18	0.0015641293	luck	9	0.0015429453	dinner	9	0.0015859030	job
18	0.0015641293	baby	9	0.0015429453	name	9	0.0015859030	t
18	0.0015641293	video	9	0.0015429453	video	9	0.0015859030	love
18	0.0015641293	wow	9	0.0015429453	cheers	9	0.0015859030	hope
18	0.0015641293	kids	9	0.0015429453	hahaha	9	0.0015859030	site
18	0.0015641293	thank	9	0.0015429453	afternoon	9	0.0015859030	spring
17	0.0014772332	sleep	9	0.0015429453	guy	9	0.0015859030	m
17	0.0014772332	b	9	0.0015429453	x	9	0.0015859030	years
17	0.0014772332	lunch	8	0.0013715069	company	9	0.0015859030	shit
17	0.0014772332	years	8	0.0013715069	picture	9	0.0015859030	video
16	0.0013903371	pain	8	0.0013715069	pic	9	0.0015859030	sun
16	0.0013903371	hi	8	0.0013715069	bed	8	0.0014096916	thanks
16	0.0013903371	spring	8	0.0013715069	follow	8	0.0014096916	saturday
16	0.0013903371	god	8	0.0013715069	thats	8	0.0014096916	bit
15	0.0013034410	family	8	0.0013715069	congrats	8	0.0014096916	god
15	0.0013034410	car	8	0.0013715069	yesterday	8	0.0014096916	page
15	0.0013034410	picture	8	0.0013715069	years	8	0.0014096916	cat
15	0.0013034410	dinner	8	0.0013715069	trip	8	0.0014096916	room
15	0.0013034410	room	8	0.0013715069	tea	8	0.0014096916	song
15	0.0013034410	hahaha	8	0.0013715069	god	8	0.0014096916	stuff
15	0.0013034410	movie	8	0.0013715069	com	8	0.0014096916	followers
15	0.0013034410	anyone	8	0.0013715069	iphone	8	0.0014096916	world
14	0.0012165450	company	8	0.0013715069	test	7	0.0012334801	dog

Taula 27

dataset_3						
labels	paraules		adjectius		substantius	
	tipus	ocurrències	tipus	ocurrències	tipus	ocurrències
all	15526	161837	3187	13840	9888	41334
positive	11490	105839	2255	9028	7439	25763
negative	7366	55998	1476	4812	4111	15571

Taula 28

Freqüència i distribució dels adjectius al dataset_3 sencer								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
708	0.0511560693	good	622	0.0688967656	good	313	0.0650457190	url
565	0.0408236994	url	267	0.0295746566	new	104	0.0216126350	mental
322	0.0232658959	great	252	0.0279131590	url	86	0.0178719866	good
316	0.0228323699	new	249	0.0275808595	great	84	0.0174563591	more
254	0.0183526011	happy	215	0.0238147984	happy	73	0.0151704073	great
223	0.0161127167	more	147	0.0162826761	last	68	0.0141313383	much
195	0.0140895953	much	139	0.0153965440	more	63	0.0130922693	emoji
173	0.0125	last	130	0.0143996455	nice	57	0.0118453865	real
141	0.0101878612	nice	127	0.0140673460	much	53	0.0110141313	other
140	0.0101156069	best	116	0.0128489144	awesome	49	0.0101828761	new
122	0.0088150289	awesome	104	0.0115197164	best	44	0.0091438071	bad
108	0.0078034682	bad	89	0.0098582188	quot	43	0.0089359933	emotional
105	0.0075867052	better	87	0.0096366858	ready	41	0.0085203657	hard
105	0.0075867052	mental	77	0.0085290208	better	41	0.0085203657	many
102	0.0073699421	real	75	0.0083074878	first	39	0.0081047381	happy
102	0.0073699421	sure	75	0.0083074878	next	36	0.0074812967	best
96	0.0069364161	first	71	0.0078644217	sure	36	0.0074812967	heavy
94	0.0067919075	ready	69	0.0076428887	cool	35	0.0072734829	sad
93	0.0067196531	next	64	0.0070890562	big	31	0.0064422277	bipolar
89	0.0064306358	quot	64	0.0070890562	bad	31	0.0064422277	red
88	0.0063583815	other	63	0.0069782897	beautiful	31	0.0064422277	sure
85	0.0061416184	little	63	0.0069782897	little	31	0.0064422277	intellectual
81	0.0058526011	many	55	0.0060921577	few	28	0.0058187863	better
81	0.0058526011	big	51	0.0056490917	free	28	0.0058187863	severe
75	0.0054190751	cool	51	0.0056490917	amp	27	0.0056109725	own
71	0.0051300578	beautiful	50	0.0055383252	glad	26	0.0054031587	last
70	0.0050578034	few	49	0.0054275587	long	25	0.0051953449	such
70	0.0050578034	long	49	0.0054275587	old	25	0.0051953449	social
68	0.0049132947	same	49	0.0054275587	amazing	24	0.0049875311	same
67	0.0048410404	free	47	0.0052060256	excited	23	0.0047797173	depression
67	0.0048410404	hard	47	0.0052060256	right	23	0.0047797173	chronic
66	0.0047687861	old	45	0.0049844926	real	22	0.0045719035	worst
65	0.0046965317	right	45	0.0049844926	least	22	0.0045719035	only
63	0.0045520231	such	45	0.0049844926	funny	22	0.0045719035	serious
63	0.0045520231	emoji	44	0.0048737261	early	22	0.0045719035	little
60	0.0043352601	least	44	0.0048737261	same	22	0.0045719035	true
57	0.0041184971	amazing	44	0.0048737261	lol	21	0.0043640897	long
56	0.0040462427	glad	43	0.0047629596	hot	21	0.0043640897	first
52	0.0037572254	ok	42	0.0046521931	sweet	20	0.0041562759	pic
52	0.0037572254	funny	42	0.0046521931	ok	20	0.0041562759	physical
52	0.0037572254	own	40	0.0044306601	many	18	0.0037406483	easy
51	0.0036849710	early	40	0.0044306601	cute	18	0.0037406483	right
51	0.0036849710	amp	38	0.0042091271	such	18	0.0037406483	next
50	0.0036127167	excited	35	0.0038768276	other	17	0.0035328345	big
50	0.0036127167	true	34	0.0037660611	busy	17	0.0035328345	depressed
49	0.0035404624	hot	30	0.0033229951	fine	17	0.0035328345	low
47	0.0033959537	lol	30	0.0033229951	favorite	17	0.0035328345	old
47	0.0033959537	red	30	0.0033229951	whole	17	0.0035328345	suicidal
45	0.0032514450	sad	28	0.0031014621	fun	17	0.0035328345	deep
44	0.0031791907	emotional	28	0.0031014621	true	17	0.0035328345	past

43	0.0031069364	sweet	28	0.0031014621	welcome	17	0.0035328345	high
42	0.0030346820	whole	27	0.0029906956	late	17	0.0035328345	light
42	0.0030346820	heavy	27	0.0029906956	saturday	16	0.0033250207	clinical
40	0.0028901734	cute	26	0.0028799291	hard	16	0.0033250207	free
38	0.0027456647	social	25	0.0027691626	own	16	0.0033250207	positive
37	0.0026734104	busy	23	0.0025476295	wonderful	15	0.0031172069	few
37	0.0026734104	easy	23	0.0025476295	wish	15	0.0031172069	able
35	0.0025289017	only	23	0.0025476295	full	15	0.0031172069	major
33	0.0023843930	fine	22	0.0024368630	live	15	0.0031172069	least
33	0.0023843930	full	22	0.0024368630	crazy	15	0.0031172069	cold
32	0.0023121387	cold	20	0.0022153300	pretty	14	0.0029093931	medical
31	0.0022398843	favorite	19	0.0021045635	friday	14	0.0029093931	worse
31	0.0022398843	fun	19	0.0021045635	easy	13	0.0027015793	weary
31	0.0022398843	crazy	19	0.0021045635	special	13	0.0027015793	daily
31	0.0022398843	bipolar	18	0.0019937970	sunny	13	0.0027015793	panic
31	0.0022398843	intellectual	18	0.0019937970	safe	12	0.0024937655	whole
30	0.0021676300	positive	18	0.0019937970	oh	12	0.0024937655	strong
30	0.0021676300	low	18	0.0019937970	white	12	0.0024937655	human
30	0.0021676300	live	18	0.0019937970	soo	12	0.0024937655	due
29	0.0020953757	able	17	0.0018830305	cold	11	0.0022859517	wrong
29	0.0020953757	late	17	0.0018830305	lovely	11	0.0022859517	nice
29	0.0020953757	wish	17	0.0018830305	enough	11	0.0022859517	normal
29	0.0020953757	welcome	17	0.0018830305	follow	11	0.0022859517	sleep
29	0.0020953757	severe	16	0.0017722640	wrong	11	0.0022859517	black
28	0.0020231213	saturday	16	0.0017722640	lucky	11	0.0022859517	healthy
28	0.0020231213	high	16	0.0017722640	super	11	0.0022859517	important
27	0.0019508670	wrong	16	0.0017722640	most	11	0.0022859517	huge
26	0.0018786127	black	16	0.0017722640	green	11	0.0022859517	cannabis
26	0.0018786127	serious	16	0.0017722640	red	10	0.0020781379	active
25	0.0018063583	white	16	0.0017722640	perfect	10	0.0020781379	different
25	0.0018063583	worst	15	0.0016614975	fantastic	10	0.0020781379	full
24	0.0017341040	most	15	0.0016614975	haha	10	0.0020781379	ok
24	0.0017341040	wonderful	15	0.0016614975	black	10	0.0020781379	seasonal
23	0.0016618497	perfect	15	0.0016614975	lt	10	0.0020781379	difficult
23	0.0016618497	depression	15	0.0016614975	poor	10	0.0020781379	actual
23	0.0016618497	chronic	15	0.0016614975	ta	10	0.0020781379	terrible
22	0.0015895953	super	15	0.0016614975	interesting	10	0.0020781379	pensive
22	0.0015895953	pretty	15	0.0016614975	sunday	10	0.0020781379	open
22	0.0015895953	deep	15	0.0016614975	proud	10	0.0020781379	young
22	0.0015895953	second	15	0.0016614975	tired	10	0.0020781379	suicide
22	0.0015895953	worse	14	0.0015507310	tonight	9	0.0018703241	crazy
21	0.0015173410	safe	14	0.0015507310	able	9	0.0018703241	single
21	0.0015173410	single	14	0.0015507310	positive	9	0.0018703241	everyday
21	0.0015173410	huge	14	0.0015507310	jealous	9	0.0018703241	second
21	0.0015173410	enough	14	0.0015507310	blue	9	0.0018703241	twitter
21	0.0015173410	sorry	14	0.0015507310	sorry	9	0.0018703241	higher
21	0.0015173410	past	14	0.0015507310	followfriday	8	0.0016625103	fake
21	0.0015173410	proud	13	0.0014399645	worth	8	0.0016625103	folded
21	0.0015173410	open	13	0.0014399645	only	8	0.0016625103	most
21	0.0015173410	major	13	0.0014399645	low	8	0.0016625103	sick

Taula 29

Freqüència i distribució dels substantius al dataset_3 sencer								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
2307	0.0558136159	depression	491	0.0190583394	day	2307	0.1481600411	depression
579	0.0140078385	day	336	0.0130419593	thanks	384	0.0246612292	anxiety
550	0.0133062369	url	323	0.0125373597	time	325	0.0208721340	url
467	0.0112982048	twitter	320	0.0124209137	today	294	0.0188812536	twitter
418	0.0101127401	time	259	0.0100531770	lol	283	0.0181748121	com
384	0.0092901727	anxiety	253	0.0098202849	quot	225	0.0144499389	pic
380	0.0091934001	today	240	0.0093156852	night	212	0.0136150536	face
355	0.0085885711	thanks	225	0.0087334549	url	167	0.0107250658	people
309	0.0074756858	com	196	0.0076078096	morning	151	0.0096975146	emoji
276	0.0066773116	lol	179	0.0069479486	haha	130	0.0083488536	life
260	0.0062902211	night	173	0.0067150564	twitter	95	0.0061010853	time
260	0.0062902211	pic	148	0.0057446725	hey	91	0.0058441975	health
253	0.0061208690	quot	147	0.0057058572	tomorrow	88	0.0056515316	day
247	0.0059757100	people	138	0.0053565190	everyone	82	0.0052661999	someone
228	0.0055160400	face	135	0.0052400729	work	71	0.0045597585	heart
218	0.0052741084	life	131	0.0050848115	thank	61	0.0039175390	things
203	0.0049112111	morning	130	0.0050459961	home	60	0.0038533170	today
180	0.0043547684	haha	128	0.0049683654	way	56	0.0035964292	mentalhealth
179	0.0043305753	way	123	0.0047742887	amp	51	0.0032753195	way
167	0.0040402574	everyone	116	0.0045025812	fun	50	0.0032110975	stress
162	0.0039192916	work	94	0.0036486434	weekend	48	0.0030826536	kids
156	0.0037741326	tomorrow	89	0.0034545666	friends	47	0.0030184316	thing
152	0.0036773600	hey	88	0.0034157512	life	47	0.0030184316	years
151	0.0036531668	emoji	87	0.0033769359	days	46	0.0029542097	risk
145	0.0035080079	home	85	0.0032993052	week	45	0.0028899877	mom
140	0.0033870421	thank	83	0.0032216744	yeah	45	0.0028899877	post
129	0.0031209174	fun	83	0.0032216744	movie	44	0.0028257658	person
123	0.0029757584	someone	80	0.0031052284	people	43	0.0027615438	lot
123	0.0029757584	amp	79	0.0030664130	school	41	0.0026330999	illness
121	0.0029273721	friends	79	0.0030664130	something	40	0.0025688780	something
119	0.0028789858	something	78	0.0030275977	love	40	0.0025688780	suicide
115	0.0027822131	days	72	0.0027947055	song	39	0.0025046560	world
113	0.0027338268	thing	71	0.0027558902	man	35	0.0022477682	thoughts
113	0.0027338268	things	68	0.0026394441	world	35	0.0022477682	development
107	0.0025886679	world	68	0.0026394441	tonight	34	0.0021835463	exercise
105	0.0025402816	school	67	0.0026006288	friend	34	0.0021835463	children
104	0.0025160884	week	66	0.0025618134	thing	33	0.0021193243	study
99	0.0023951226	weekend	65	0.0025229980	birthday	33	0.0021193243	brain
95	0.0022983500	love	64	0.0024841827	sun	32	0.0020551024	friends
95	0.0022983500	health	64	0.0024841827	hope	32	0.0020551024	help
94	0.0022741568	man	63	0.0024453673	house	32	0.0020551024	disorder
91	0.0022015773	lot	62	0.0024065520	bed	31	0.0019908804	therapy
91	0.0022015773	mom	60	0.0023289213	nothing	31	0.0019908804	symptoms
90	0.0021773842	nothing	59	0.0022901059	show	31	0.0019908804	tone
88	0.0021289979	yeah	59	0.0022901059	summer	30	0.0019266585	nothing
88	0.0021289979	friend	58	0.0022512906	please	30	0.0019266585	tears
88	0.0021289979	heart	58	0.0022512906	yay	30	0.0019266585	everything
85	0.0020564184	movie	57	0.0022124752	hi	29	0.0018624365	joy
83	0.0020080321	year	56	0.0021736598	yes	29	0.0018624365	everyone
80	0.0019354526	bed	56	0.0021736598	music	29	0.0018624365	year
79	0.0019112594	music	55	0.0021348445	job	28	0.0017982146	days
78	0.0018870663	song	54	0.0020960291	year	28	0.0017982146	anything
76	0.0018386800	please	52	0.0020183984	things	28	0.0017982146	shit
74	0.0017902937	sun	49	0.0019019524	guys	27	0.0017339926	work
74	0.0017902937	tonight	49	0.0019019524	party	26	0.0016697707	school
74	0.0017902937	house	49	0.0019019524	tweet	26	0.0016697707	god
73	0.0017661005	hi	49	0.0019019524	dinner	26	0.0016697707	pain
73	0.0017661005	hope	49	0.0019019524	followers	26	0.0016697707	eyes
72	0.0017419073	years	48	0.0018631370	lot	26	0.0016697707	issues
72	0.0017419073	person	48	0.0018631370	bit	24	0.0015413268	death

71	0.0017177142	job	47	0.0018243217	lt	24	0.0015413268	problems
71	0.0017177142	kids	46	0.0017855063	hours	23	0.0014771048	food
68	0.0016451347	birthday	46	0.0017855063	mom	23	0.0014771048	anyone
68	0.0016451347	post	45	0.0017466909	xx	23	0.0014771048	music
68	0.0016451347	summer	45	0.0017466909	ha	23	0.0014771048	cause
65	0.0015725552	show	44	0.0017078756	gt	23	0.0014771048	reason
63	0.0015241689	hours	43	0.0016690602	aww	23	0.0014771048	man
62	0.0014999758	food	43	0.0016690602	baby	23	0.0014771048	hand
62	0.0014999758	family	43	0.0016690602	family	22	0.0014128829	addiction
61	0.0014757826	anyone	43	0.0016690602	game	22	0.0014128829	self
59	0.0014273963	yes	42	0.0016302449	luck	22	0.0014128829	others
58	0.0014032031	tweet	41	0.0015914295	someone	22	0.0014128829	skin
58	0.0014032031	god	41	0.0015914295	hair	21	0.0013486609	mind
58	0.0014032031	yay	41	0.0015914295	oh	21	0.0013486609	joke
57	0.0013790100	times	40	0.0015526142	coffee	21	0.0013486609	friend
56	0.0013548168	bit	40	0.0015526142	girl	21	0.0013486609	mood
56	0.0013548168	mentalhealth	40	0.0015526142	x	21	0.0013486609	teens
55	0.0013306236	anything	39	0.0015137988	times	21	0.0013486609	ptsd
55	0.0013306236	guys	39	0.0015137988	weather	21	0.0013486609	drug
54	0.0013064305	game	39	0.0015137988	food	21	0.0013486609	video
54	0.0013064305	stress	39	0.0015137988	yesterday	20	0.0012844390	state
53	0.0012822373	baby	38	0.0014749835	stuff	20	0.0012844390	war
53	0.0012822373	party	38	0.0014749835	anyone	20	0.0012844390	night
52	0.0012580442	stuff	38	0.0014749835	phone	19	0.0012202170	sucks
52	0.0012580442	girl	37	0.0014361681	ya	19	0.0012202170	week
52	0.0012580442	followers	36	0.0013973527	hahaha	19	0.0012202170	ass
51	0.0012338510	everything	35	0.0013585374	pic	19	0.0012202170	point
50	0.0012096579	dinner	35	0.0013585374	while	19	0.0012202170	thanks
50	0.0012096579	video	35	0.0013585374	wow	19	0.0012202170	fact
49	0.0011854647	hair	35	0.0013585374	idea	19	0.0012202170	support
48	0.0011612715	xx	34	0.0013197220	cool	19	0.0012202170	family
47	0.0011370784	weather	34	0.0013197220	thats	19	0.0012202170	puff
47	0.0011370784	book	34	0.0013197220	congrats	18	0.0011559951	part
47	0.0011370784	lt	34	0.0013197220	glad	18	0.0011559951	times
47	0.0011370784	yesterday	33	0.0012809067	trip	18	0.0011559951	bed
47	0.0011370784	luck	32	0.0012420913	room	18	0.0011559951	please
47	0.0011370784	risk	32	0.0012420913	movies	18	0.0011559951	medium
46	0.0011128852	room	32	0.0012420913	breakfast	18	0.0011559951	book
46	0.0011128852	one	32	0.0012420913	god	17	0.0010917731	cure
45	0.0010886921	coffee	32	0.0012420913	album	17	0.0010917731	sign

Taula 30

dataset_3.train						
labels	paraules		adjectius		substantius	
	tipus	ocurrències	tipus	ocurrències	tipus	ocurrències
all	8907	67737	1758	5742	5279	18071
positive	4363	23563	788	1982	2592	5757
negative	6405	44174	1267	3760	3524	12314

Taula 31

Freqüència i distribució dels adjectius al joc d'entrenament del dataset_3								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
308	0.0525149190	url	149	0.0707838479	good	242	0.0643617021	url
211	0.0359761295	good	66	0.0313539192	url	80	0.0212765957	mental
115	0.0196078431	great	64	0.0304038004	new	68	0.0180851063	more
101	0.0172208013	more	59	0.0280285035	great	62	0.0164893617	good
99	0.0168797953	new	44	0.0209026128	happy	56	0.0148936170	great
80	0.0136402387	mental	36	0.0171021377	last	52	0.0138297872	much
78	0.0132992327	happy	33	0.0156769596	more	50	0.0132978723	emoji
69	0.0117647058	much	26	0.0123515439	nice	46	0.0122340425	real
58	0.0098891730	real	24	0.0114014251	awesome	43	0.0114361702	other
58	0.0098891730	other	24	0.0114014251	best	35	0.0093085106	new
57	0.0097186700	last	22	0.0104513064	sure	34	0.0090425531	happy
54	0.0092071611	best	21	0.0099762470	ready	33	0.0087765957	many
50	0.0085251491	emoji	20	0.0095011876	amazing	33	0.0087765957	emotional
50	0.0085251491	bad	20	0.0095011876	quot	33	0.0087765957	bad
49	0.0083546462	sure	19	0.0090261282	first	31	0.0082446808	hard
43	0.0073316283	many	18	0.0085510688	big	31	0.0082446808	heavy
40	0.0068201193	hard	17	0.0080760095	amp	30	0.0079787234	best
40	0.0068201193	better	17	0.0080760095	much	27	0.0071808510	sad
36	0.0061381074	nice	17	0.0080760095	bad	27	0.0071808510	sure
33	0.0056265984	emotional	16	0.0076009501	long	27	0.0071808510	red
32	0.0054560954	big	16	0.0076009501	better	26	0.0069148936	bipolar
32	0.0054560954	first	15	0.0071258907	other	25	0.0066489361	intellectual
32	0.0054560954	heavy	15	0.0071258907	glad	24	0.0063829787	better
31	0.0052855924	long	15	0.0071258907	next	22	0.0058510638	same
31	0.0052855924	red	15	0.0071258907	few	21	0.0055851063	social
30	0.0051150895	sad	14	0.0066508313	cool	21	0.0055851063	last
30	0.0051150895	same	13	0.0061757719	least	20	0.0053191489	such
30	0.0051150895	such	12	0.0057007125	real	19	0.0050531914	true
29	0.0049445865	awesome	12	0.0057007125	busy	19	0.0050531914	severe
29	0.0049445865	next	12	0.0057007125	excited	18	0.0047872340	only
27	0.0046035805	few	11	0.0052256532	ok	17	0.0045212765	own
27	0.0046035805	right	11	0.0052256532	cute	17	0.0045212765	chronic
26	0.0044330775	bipolar	10	0.0047505938	many	17	0.0045212765	serious
26	0.0044330775	ready	10	0.0047505938	right	17	0.0045212765	right
25	0.0042625745	true	10	0.0047505938	beautiful	16	0.0042553191	worst
25	0.0042625745	own	10	0.0047505938	late	15	0.0039893617	long
25	0.0042625745	little	10	0.0047505938	hot	15	0.0039893617	little
25	0.0042625745	social	10	0.0047505938	free	15	0.0039893617	cold
25	0.0042625745	intellectual	10	0.0047505938	little	14	0.0037234042	easy
25	0.0042625745	amazing	10	0.0047505938	such	14	0.0037234042	physical
24	0.0040920716	free	10	0.0047505938	whole	14	0.0037234042	depression
24	0.0040920716	cold	9	0.0042755344	hard	14	0.0037234042	suicidal
23	0.0039215686	old	9	0.0042755344	old	14	0.0037234042	old
22	0.0037510656	only	9	0.0042755344	cold	14	0.0037234042	pic
22	0.0037510656	least	8	0.0038004750	own	14	0.0037234042	big
21	0.0035805626	glad	8	0.0038004750	same	14	0.0037234042	next
21	0.0035805626	whole	8	0.0038004750	sweet	14	0.0037234042	free
20	0.0034100596	easy	8	0.0038004750	crazy	13	0.0034574468	low
20	0.0034100596	quot	7	0.0033254156	sunny	13	0.0034574468	first
19	0.0032395566	ok	7	0.0033254156	welcome	13	0.0034574468	past

19	0.0032395566	severe	7	0.0033254156	blue	13	0.0034574468	depressed
18	0.0030690537	cool	6	0.0028503562	easy	12	0.0031914893	few
17	0.0028985507	beautiful	6	0.0028503562	lovely	12	0.0031914893	high
17	0.0028985507	amp	6	0.0028503562	fantastic	12	0.0031914893	clinical
17	0.0028985507	chronic	6	0.0028503562	fun	12	0.0031914893	panic
17	0.0028985507	serious	6	0.0028503562	early	12	0.0031914893	deep
16	0.0027280477	low	6	0.0028503562	full	12	0.0031914893	major
16	0.0027280477	hot	6	0.0028503562	true	11	0.0029255319	daily
16	0.0027280477	able	6	0.0028503562	favorite	11	0.0029255319	medical
16	0.0027280477	worst	6	0.0028503562	safe	11	0.0029255319	cannabis
15	0.0025575447	excited	6	0.0028503562	sorry	11	0.0029255319	whole
14	0.0023870417	physical	5	0.0023752969	wish	11	0.0029255319	able
14	0.0023870417	depression	5	0.0023752969	lol	10	0.0026595744	strong
14	0.0023870417	busy	5	0.0023752969	sunday	10	0.0026595744	nice
14	0.0023870417	suicidal	5	0.0023752969	white	10	0.0026595744	human
14	0.0023870417	pic	5	0.0023752969	perfect	10	0.0026595744	normal
14	0.0023870417	full	5	0.0023752969	tonight	10	0.0026595744	wrong
14	0.0023870417	past	5	0.0023752969	fine	10	0.0026595744	healthy
14	0.0023870417	wrong	5	0.0023752969	worth	9	0.0023936170	seasonal
14	0.0023870417	crazy	5	0.0023752969	friday	9	0.0023936170	open
13	0.0022165387	high	5	0.0023752969	ta	9	0.0023936170	light
13	0.0022165387	deep	5	0.0023752969	nite	9	0.0023936170	sleep
13	0.0022165387	depressed	5	0.0023752969	soo	9	0.0023936170	active
12	0.0020460358	late	5	0.0023752969	wonderful	9	0.0023936170	least
12	0.0020460358	open	5	0.0023752969	saturday	9	0.0023936170	different
12	0.0020460358	clinical	5	0.0023752969	live	9	0.0023936170	worse
12	0.0020460358	sick	5	0.0023752969	able	9	0.0023936170	single
12	0.0020460358	panic	4	0.0019002375	follow	8	0.0021276595	fucking
12	0.0020460358	black	4	0.0019002375	dead	8	0.0021276595	sick
12	0.0020460358	normal	4	0.0019002375	alive	8	0.0021276595	full
12	0.0020460358	sleep	4	0.0019002375	most	8	0.0021276595	black
12	0.0020460358	early	4	0.0019002375	lucky	8	0.0021276595	twitter
12	0.0020460358	live	4	0.0019002375	social	8	0.0021276595	ok
12	0.0020460358	major	4	0.0019002375	goodnight	8	0.0021276595	terrible
11	0.0018755328	daily	4	0.0019002375	wrong	7	0.0018617021	pensive
11	0.0018755328	strong	4	0.0019002375	only	7	0.0018617021	beautiful
11	0.0018755328	white	4	0.0019002375	san	7	0.0018617021	positive
11	0.0018755328	perfect	4	0.0019002375	sick	7	0.0018617021	actual
11	0.0018755328	medical	4	0.0019002375	ive	7	0.0018617021	manic
11	0.0018755328	cannabis	4	0.0019002375	super	7	0.0018617021	excellent
11	0.0018755328	single	4	0.0019002375	black	7	0.0018617021	okay
11	0.0018755328	sorry	4	0.0019002375	red	7	0.0018617021	dark
11	0.0018755328	cute	4	0.0019002375	uni	7	0.0018617021	negative
11	0.0018755328	healthy	4	0.0019002375	weird	7	0.0018617021	live
10	0.0017050298	wish	4	0.0019002375	afraid	7	0.0018617021	stupid
10	0.0017050298	positive	4	0.0019002375	fave	7	0.0018617021	second
10	0.0017050298	most	4	0.0019002375	funny	7	0.0018617021	everyday
10	0.0017050298	light	3	0.0014251781	shit	7	0.0018617021	higher
10	0.0017050298	human	3	0.0014251781	low	7	0.0018617021	young
10	0.0017050298	blue	3	0.0014251781	strange	7	0.0018617021	entire

Taula 32

Freqüència i distribució dels substantius al joc d'entrenament del dataset_3								
all_tweets			positive_tweets			negative_tweets		
fAbs	fRel	paraula	fAbs	fRel	paraula	fAbs	fRel	paraula
1821	0.0993128272	depression	115	0.0190966456	day	1821	0.1478804612	depression
316	0.0172338568	url	88	0.0146130853	thanks	306	0.0248497644	anxiety
306	0.0166884816	anxiety	79	0.0131185652	time	261	0.0211953873	url
263	0.0143433682	twitter	68	0.0112919295	today	229	0.0185967191	twitter
233	0.0127072425	com	60	0.0099634672	quot	223	0.0181094688	com
187	0.0101985165	pic	56	0.0092992361	lol	176	0.0142926750	pic
180	0.0098167539	day	55	0.0091331783	url	170	0.0138054247	face
173	0.0094349912	face	54	0.0089671205	night	141	0.0114503816	people
157	0.0085623909	people	52	0.0086350049	hey	121	0.0098262140	emoji
154	0.0083987783	time	41	0.0068083693	haha	103	0.0083644632	life
126	0.0068717277	life	39	0.0064762537	morning	75	0.0060906285	time
121	0.0065990401	emoji	34	0.0056459647	thank	73	0.0059282117	health
120	0.0065445026	today	34	0.0056459647	twitter	71	0.0057657950	someone
103	0.0056173647	thanks	32	0.0053138492	amp	65	0.0052785447	day
81	0.0044175392	someone	31	0.0051477914	home	56	0.0045476693	heart
77	0.0041993891	way	30	0.0049817336	way	52	0.0042228357	today
73	0.0039812390	health	30	0.0049817336	work	49	0.0039792106	things
68	0.0037085514	lol	29	0.0048156758	everyone	47	0.0038167938	way
68	0.0037085514	night	28	0.0046496180	tomorrow	41	0.0033295436	stress
66	0.0035994764	things	26	0.0043175024	fun	41	0.0033295436	years
63	0.0034358638	heart	23	0.0038193291	life	40	0.0032483352	thing
60	0.0032722513	quot	23	0.0038193291	love	38	0.0030859184	mentalhealth
59	0.0032177137	thing	22	0.0036532713	friends	38	0.0030859184	risk
55	0.0029995636	hey	21	0.0034872135	show	36	0.0029235017	lot
54	0.0029450261	work	21	0.0034872135	weekend	36	0.0029235017	mom
53	0.0028904886	everyone	20	0.0033211557	school	35	0.0028422933	kids
52	0.0028359511	lot	20	0.0033211557	movie	35	0.0028422933	suicide
50	0.0027268760	years	20	0.0033211557	sun	34	0.0027610849	person
50	0.0027268760	friends	19	0.0031550979	house	34	0.0027610849	illness
44	0.0023996509	stress	19	0.0031550979	luck	33	0.0026798765	post
44	0.0023996509	home	19	0.0031550979	thing	31	0.0025174598	world
44	0.0023996509	morning	19	0.0031550979	yeah	30	0.0024362514	brain
43	0.0023451134	world	18	0.0029890401	hope	29	0.0023550430	everything
43	0.0023451134	mom	18	0.0029890401	bed	29	0.0023550430	something
43	0.0023451134	something	17	0.0028229823	things	28	0.0022738346	friends
42	0.0022905759	haha	17	0.0028229823	man	28	0.0022738346	children
42	0.0022905759	thank	17	0.0028229823	bit	27	0.0021926262	development
41	0.0022360383	kids	16	0.0026569246	week	26	0.0021114178	therapy
40	0.0021815008	person	16	0.0026569246	people	26	0.0021114178	disorder
38	0.0020724258	mentalhealth	16	0.0026569246	lot	24	0.0019490011	tone
38	0.0020724258	post	15	0.0024908668	dinner	24	0.0019490011	work
38	0.0020724258	risk	15	0.0024908668	tonight	24	0.0019490011	joy
37	0.0020178883	love	14	0.0023248090	something	24	0.0019490011	shit
37	0.0020178883	school	14	0.0023248090	song	24	0.0019490011	thoughts
36	0.0019633507	days	13	0.0021587512	days	24	0.0019490011	everyone
36	0.0019633507	tomorrow	13	0.0021587512	gt	24	0.0019490011	problems
35	0.0019088132	suicide	13	0.0021587512	guys	23	0.0018677927	year
35	0.0019088132	nothing	12	0.0019926934	yay	23	0.0018677927	exercise
34	0.0018542757	fun	12	0.0019926934	friend	23	0.0018677927	help
34	0.0018542757	everything	12	0.0019926934	nothing	23	0.0018677927	days
34	0.0018542757	illness	12	0.0019926934	stuff	23	0.0018677927	nothing
33	0.0017997382	year	12	0.0019926934	baby	22	0.0017865843	study
32	0.0017452006	man	12	0.0019926934	world	22	0.0017865843	anything
32	0.0017452006	amp	12	0.0019926934	hair	22	0.0017865843	tears
31	0.0016906631	brain	11	0.0018266356	girl	21	0.0017053759	symptoms
30	0.0016361256	music	11	0.0018266356	birthday	21	0.0017053759	death
30	0.0016361256	week	11	0.0018266356	yes	20	0.0016241676	pain
30	0.0016361256	bed	11	0.0018266356	music	20	0.0016241676	god
29	0.0015815881	friend	11	0.0018266356	hours	20	0.0016241676	eyes

29	0.0015815881	house	11	0.0018266356	aww	19	0.0015429592	addiction
29	0.0015815881	sun	11	0.0018266356	o	19	0.0015429592	music
29	0.0015815881	children	11	0.0018266356	pic	19	0.0015429592	issues
27	0.0014725130	therapy	11	0.0018266356	breakfast	19	0.0015429592	cause
27	0.0014725130	hours	11	0.0018266356	tweet	18	0.0014617508	ptsd
27	0.0014725130	development	11	0.0018266356	job	18	0.0014617508	teens
27	0.0014725130	shit	11	0.0018266356	times	18	0.0014617508	self
26	0.0014179755	times	10	0.0016605778	someone	18	0.0014617508	state
26	0.0014179755	disorder	10	0.0016605778	movies	18	0.0014617508	puff
25	0.0013634380	anything	10	0.0016605778	please	18	0.0014617508	anyone
25	0.0013634380	joy	10	0.0016605778	hehe	17	0.0013805424	friend
25	0.0013634380	thoughts	10	0.0016605778	com	17	0.0013805424	others
25	0.0013634380	show	10	0.0016605778	year	17	0.0013805424	school
25	0.0013634380	bit	10	0.0016605778	ya	17	0.0013805424	cure
25	0.0013634380	help	9	0.0014945200	glad	17	0.0013805424	joke
25	0.0013634380	problems	9	0.0014945200	x	17	0.0013805424	reason
24	0.0013089005	tone	9	0.0014945200	followers	17	0.0013805424	drug
24	0.0013089005	hope	9	0.0014945200	room	16	0.0012993340	hours
24	0.0013089005	stuff	9	0.0014945200	june	16	0.0012993340	war
24	0.0013089005	pain	9	0.0014945200	weather	16	0.0012993340	book
24	0.0013089005	book	9	0.0014945200	ice	16	0.0012993340	mind
24	0.0013089005	weekend	9	0.0014945200	years	16	0.0012993340	ass
24	0.0013089005	god	9	0.0014945200	cream	16	0.0012993340	food
23	0.0012543630	anyone	9	0.0014945200	college	16	0.0012993340	skin
23	0.0012543630	study	9	0.0014945200	girls	15	0.0012181257	mood
23	0.0012543630	exercise	8	0.0013284623	thats	15	0.0012181257	times
23	0.0012543630	yeah	8	0.0013284623	ok	15	0.0012181257	fact
22	0.0011998254	family	8	0.0013284623	band	15	0.0012181257	thanks
22	0.0011998254	hi	8	0.0013284623	summer	15	0.0012181257	man
22	0.0011998254	tears	8	0.0013284623	book	15	0.0012181257	video
22	0.0011998254	movie	8	0.0013284623	t	15	0.0012181257	patients
22	0.0011998254	please	8	0.0013284623	guy	15	0.0012181257	feeling
22	0.0011998254	cause	8	0.0013284623	family	14	0.0011369173	medium
21	0.0011452879	job	8	0.0013284623	hmm	14	0.0011369173	family
21	0.0011452879	reason	8	0.0013284623	oh	14	0.0011369173	hi
21	0.0011452879	food	8	0.0013284623	check	14	0.0011369173	love
21	0.0011452879	video	8	0.0013284623	hi	14	0.0011369173	night
21	0.0011452879	symptoms	8	0.0013284623	myspace	14	0.0011369173	support
21	0.0011452879	place	8	0.0013284623	mcflyforgermany	14	0.0011369173	treatment
21	0.0011452879	issues	8	0.0013284623	place	14	0.0011369173	cannabis
21	0.0011452879	baby	8	0.0013284623	afternoon	14	0.0011369173	one

Taula 33

Bibliografia[24]

Cliche, M., "Twiter Sentiment Analysis with CNNs and LSTMs", *Proceedings of the 11th International Workshop on Semantinc Evaluations* (2017), pages 573-580 <https://www.aclweb.org/anthology/S17-2094.pdf>

Bird, S., **Klein**, E. & **Loper,** E. (2009): *Natural language processing with Python.* O'Reilly.

Chakraborty, R., **Elhence**, A. & **Arora**, K. (2019): "Sparse Victory – A Large Scale Systematic Comparison of Count-Based and Prediction-Based Vectorizers for Text Classification", A: *Proceedings of Recent Advances in Natural Language Processing*, 2019, Varna, pp. 188-197 <https://www.aclweb.org/anthology/R19-1022.pdf>

Charniak, E. (ed) (1993): *Statistical language learning.* Cambridge: The MIT Press.

Das, Oaindrila & **Balanbataray**, Rakesh (2014): "Sentiment analysis of movie reviews using POS tags and term frequencies", A: *International Journal of Computer Applications*, vol.96, núm.25 (juny del 2014).

Davydova, Olga (2018): "Test preprocessing in Python: steps, tools, and examples", A: *DataMonsters* (data de publicació: 18 d'octubre de 2018): <https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908>

Derczinski, L. *et alii* (2013): "Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data", *Preceedings or Recent Advances in Natural Language Processing*, pp.198-206 <https://www.aclweb.org/anthology/R13-1026.pdf>

Dubey, Prauveen (2018): "An introduction to Bag of Words and how to code it in Python for NLP", A: *freeCodeCamp* (data de publicació: 18 de desembre de 2018): <https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/>

Gamallo Otero, Pablo & **González López,** Isaac, "A grammatical formalism based on patterns of Part of Speech tags", *Universidad de Santiago de Compostela* (2011)

Giménez, J & **Màrquez,** L. (2004). "SVMTool: a general POS tagger generator based on Support Vector Machines", *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisboa.

Hastie, T., **Tibshirani**, R. & **Friedman**, J. (2001): *The elements of Statistical Learning*. New York, Springer-Verlag.

[24] Totes les fonts en línia han estat consultades per darrer cop el dia 8 de juny de 2021.

Irawaty, I., **Andreswari**, R. & **Pramesti**, D. (2020): "Vectorizer Comparison for Sentiment Analysis on Social Media Youtube: A Case Study," A: *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, pp. 69-74 <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9274650&isnumber=9274561

Khong, WH., **Soon**, LK., **Goh**, HN. & **Haw**, SC. (2018): "Leveraging Part-of-Speech Tagging for Sentiment Analysis in Short Texts and Regular Texts". A: Ichise R., Lecue F., Kawamura T., Zhao D., Muggleton S., Kozaki K. (eds): *Semantic Technology*. JIST 2018. Lecture Notes in Computer Science, vol 11341. Springer, Cham. <https://doi.org/10.1007/978-3-030-04284-4_13>

Langley, P. (1996): *Elements of Machine Learning*. New York, Morgan Kaufmann Pubilshers.

Màrquez, L. (2001): "Aprendizaje automático y procesamiento del lenguage natural". Ponència al curs: *La ingeniería lingüística en la sociedad de la información*. Fundación Duques de Soria

Mitkov, R. (2003, ed): *The Oxforx handbook of computational linguistics.* Oxford University Press.

Moré, J., **Climent,** S., **Oliver,** A**.** & **Taulé,** M**.** (2005): "Análisis de los fenómenos lingüísticos de los mensajes de correo electrónico en catalán desde la perspectiva de la traducción automática", *Procesamiento del lenguage natural*, (35): pàgs.45-50.

Nakawaga, T., **Kudoh,** T. & **Matsumoto**, Y. (2001): "Unknown Word Guessing and Part-of-Speech tagging using Suport Vector Machines", *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, pàg.325-331.

Palmer, D. (2000): *Handbook of natural language processing*, cap.2: "Tokenisation and sentence segmentation", (pàgs.11-36). New York: Marcel Dekker, Inc.

O'Connor, B. *et alii* (2013): "Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters", *NAACL* <http://www.cs.cmu.edu/~ark/TweetNLP/owoputi+etal.naacl13.pdf>

Oliver, Antoni (2010): "Anàlisi textual i processament de corpus" i "Etiquetatge morfosintàctic", a: Antoni Oliver (coord): *Lingüística computacional*. Barcelona, FUOC.

Singh, Ankit (2018): "Breakthrough Research Papers and Models for Sentiment Analysis", A: *ParalledDots* (data de publicació: 19 de juliol de 2018) <https://blog.paralleldots.com/data-science/breakthrough-research-papers-and-models-for-sentiment-analysis/>

Thede, S. (1998): "Predicting Part-of-Speech information about unknown words using statistical methods". A: *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, COLING-ACL"*, pàgs.1505-1507.

Tripathi, Mayank (2018): "How to process textual data using TF-IDF in Python, A: *freeCodeCamp* (data de publicació: 6 de juny de 2018): <https://www.freecodecamp.org/news/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3/>

Vinodhini, G. & **Chandrasekaran**, R.M. (2016): "A comparative performance evaluation of neural network based approach for sentiment classification of online reviews", A: *Journal of King Saud University – Computer and Information Sciences*, volume 28, Issue 1, January 2016, pp.2-12 <https://www.sciencedirect.com/science/article/pii/S1319157815001020>

He, X. *et alii* (2017): "Image caption generation with Part-of-Speech guidance", *Pattern Recognition Letters* <https://www.researchgate.net/publication/320494335_Image_Caption_Generation_with_Part_of_Speech_Guidance>