# Anàlisi del Dataset

## Comprovació Dataset Balancejat

Comprovar si un dataset és balancejat és important a l’hora de començar un projecte ja que si no ho és, a l’hora d’entrenar un model de Machine learning el model pot esbiaixar les predicciona cap a la classe majoritària i produir per tant un error més elevat.

Per tal de comprovar si el dataset és balancejat el que hem fet ha sigut crear un codi que compti quan “tweets” tenen com a “target” un 0 i quins tenen un 4, el target ens indica si un “tweet” és positiu o negatiu (0 per els negatius, 4 per els positius), si el nombre de posts positius i negatius és semblant, llavors es pot considerar que el dataset està balancejat.

En el nostre cas el codi ens ha donat el següent recompte:

Texto

Descripción generada automáticamente

Ón es pot veure que pràcticament hi ha els mateixos valors positius com negatius, per tant podem considerar que el nostre dataset està balancejat.

## Distribució de longitud de missatges

Una altre cosa que ens ha semblat que és important saber del dataset, és fer un recompte de les longituds que tenen tots els ‘tweets’ per tal de veure quines són les tendències. Per fer això, simplement hem creat un codi que compti quants missatges de cada longitud hi ha, i el resultat ha sigut el següent:

## Gráfico, Histograma Descripción generada automáticamente

De la gràfica es pot observar que la tendència dels missatges oscil·la majoritàriament entre els valors compresos entre els 30 fins a una mica menys de 100 caràcters i veiem com a mesura que va pujant el nombre de caràcters, el seu nombre de missatges també descendeix, menys aproximadament sobre els 120, on hi torna a haver un pic, però que després torna a seguir amb la tendència.

Addicionalment també hem calculat les estadístiques, on hem pogut identificar dades com el nombre totol de caràcters que hi ha en tot el dataset, la mitjana per missatge (74) la qual te força sentit fixant-nos en la gràfica anterior ja que es el valor que queda enmig entre el 50 i el 100, que és la longitud que tenen la majoria dels missatges. També ho ha altres estadístiques interessants com la desviació (std), el missatge amb menys caràcters i el missatge amb el que més (374), el primer i segon quartil, juntament amb la mediana.

Texto

Descripción generada automáticamente

Per últim hem generat dos diagrames de caixa per a cada classe (0 és negatiu i 4 positiu) per tal de veure si la longitud del text tendeix cap a alguna de les dues.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Les distribucions de longitud són molt similars en les dues classes, amb la majoria de textos entre 40 i 100 caràcters i una mediana al voltant dels 70 caràcters. Això indica que la longitud del text no és un factor discriminant entre textos positius i negatius.

Tot i que hi ha outliers (textos més llargs, superiors als 300 caràcters), aquests són pocs i no afecten significativament la distribució general. Aquest patró suggereix que la longitud del text aporta poca informació per diferenciar sentiments.

## Freqüència de paraules

Un altre patró que pot ser útil es mirar quines són les paraules que més s’han fet servir en els tweets del dataset, això en podria donar alguna pista sobre les tendències de certes paraules cap a comentaris negatius o positius.

El primer que hem fet ha sigut comptabilitzar i mostrar les 20 paraules que més s’han fet servir:

Texto

Descripción generada automáticamente

Com hem vist que la majoria de paraules són connectors, els quals no aporten un gran pes al significat del missatge hem decidit ampliar i trobar les 100 paraules que amb més freqüència:

**i: 750749, to: 564517, the: 520037, a: 377505, my: 314039, and: 298369, you: 270068, is: 236006, it: 230842, for: 215692, in: 214201, of: 183427, im: 177478, on: 167078, me: 160074, so: 150709, have: 144505, that: 141325, but: 127457, just: 125053, with: 114866, be: 112521, at: 111280, its: 108027, not: 106355, was: 104073, this: 92983, now: 90357, good: 89398, up: 88449, day: 82362, get: 81484, all: 80707, out: 80516, like: 77748, are: 77696, no: 73515, go: 72911, dont: 66923, do: 65961, your: 64996, today: 64601, too: 64204, going: 64087, love: 63463, work: 62843, cant: 62597, got: 60751, time: 56107, from: 55996, back: 55974, lol: 55146, u: 52775, what: 52204, one: 52185, will: 51829, know: 51193, we: 50902, about: 50207, really: 49550, am: 49344, had: 47678, can: 46952, see: 45833, some: 45328, well: 44412, if: 42996, still: 42800, want: 42070, new: 41996, night: 41064, how: 41049, think: 41043, amp: 40630, thanks: 39340, home: 39197, as: 38975, when: 38753, there: 38227, oh: 37986, 2: 37942, more: 37451, much: 36532, off: 36137, miss: 36101, here: 35730, they: 35650, need: 35524, last: 35237, an: 35024, then: 33513, been: 33426, morning: 33149, hope: 33059, has: 32904, great: 32833, tomorrow: 32694, ill: 31638, twitter: 31252, or: 31060**

D’aquestes paraules hem fet un filtratge nosaltres i ens hem quedat amb aquelles paraules que per si soles ja poden tenir cert significat i podria dir-nos si tenen una tendència més positiva o negativa. Per fer això el que hem fet ha sigut de cada paraules de les que hem escollit comptabilitzar quants cops sortia en un missatge positius i quants en un de negatiu, i els resultats han estat els següents:

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Es pot veure que hi ha paraules molt associades amb un sentiment específic. Per exemple, "not" és una paraula fortament vinculada a sentiments negatius, surt 108,689 vegades en total, de les quals un 69.05% es troben en missatges negatius. Altres paraules com "no", "cant" i "miss" també mostren una clara relació amb sentiments negatius. En canvi, paraules com "love" (73.73%), "good" (68.01%) i "lol" (60.96%) acostumen a ser relacionats amb sentiments positius. Una de les paraules més destacades és "thanks", amb un 85.61% d'associació amb sentiments positius, la qual cosa té basant sentit tenint en compte que és una paraula que sempre s’acostuma a fer servir en contextos positius.

D'altra banda, algunes paraules, com "out" (46.67%) i "time" (52.51%), mostren una distribució més equilibrada entre sentiments negatius i positius, donant a entendre que el seu significat pot variar segons el context.

## Bigrames i Trigrames

Relacionat amb les paraules i sobre si el seu significat és positiu o negatiu, hem pensat que també seria interessant buscar quines són les combinacions de 2 i 3 paraules que tendeixen més a missatges positius i negatius.

Per fer això hem comptabilitzat aquells grups de 2 i 3 paraules que més es repetien sense tenir en compte les ‘stopwords’, es a dir, els connectors en anglés, ja que com hem esmentat abans, aquests apareixen molt i no aporten un significat fort al missatge.

Pel que fa als bígramas , els tweets positius mostren un ús significatiu d’enllaços com "http twitpic" i "bit ly", acompanyats de frases com "good morning" o "good night", que transmeten sentiments positius per iniciar o concloure el dia. També es destaquen expressions com "just got", que poden estar associades a esdeveniments feliços o situacions informals.

En els tweets negatius, els bígramas més comuns inclouen frases com "feel like" i "don know", suggerint confusió o incomoditat. Tot i que enllaços com "http twitpic" i "bit ly" continuen presents, són menys prominents, i altres frases com "don want" o "looks like" reflecteixen frustració o decepció.

Texto

Descripción generada automáticamente

En els tweets positius, els trígramas més comuns estan dominats per enllaços, com "http twitpic com", "http bit ly" i "http tinyurl com", indicant que molts usuaris comparteixen contingut extern amb un sentiment positiu. També es destaquen frases com "100 followers day" o "followers day using", que reflecteixen un interès per l'augment de seguidors i el creixement a les xarxes socials. Alguns altres trígramas, com "add train pay", semblen missatges promocionals o automatitzats.

En canvi, en els tweets negatius, tot i que els enllaços continuen sent comuns, com "http twitpic com" i "http bit ly", la seva presència és menys prominent que en els positius. Frases com "just got home" i "hope feel better" indiquen preocupacions personals o desitjos de millora. Altres trígramas, com "lost help good", reflecteixen sentiments de pèrdua o necessitat d'ajuda.

Texto

Descripción generada automáticamente