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.

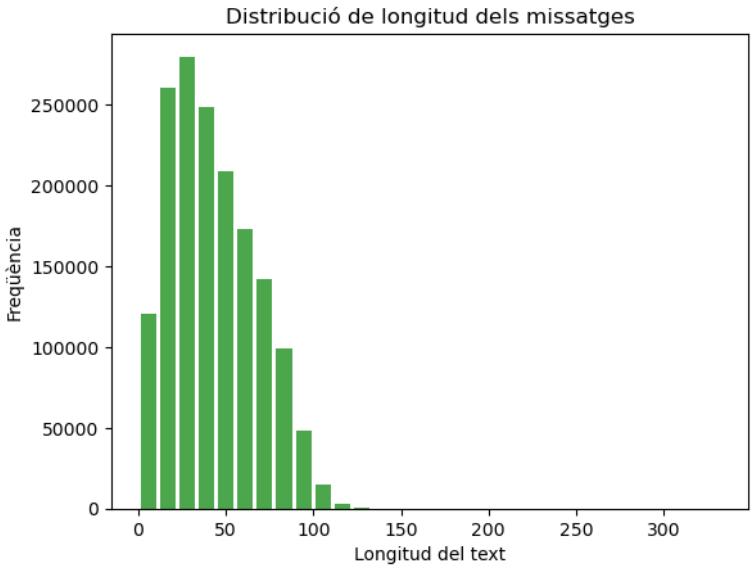
En el nostre cas el codi ens ha donat el següent recompte:

```
Twitter data carregada correctament.

La cantitat de etiquetes amb valor 0 es: 799999
La cantitat de etiquetes amb valor 4 es: 800000

No Hi ha altres valors que no son 0 o 4 a la columna'target'.
```

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:

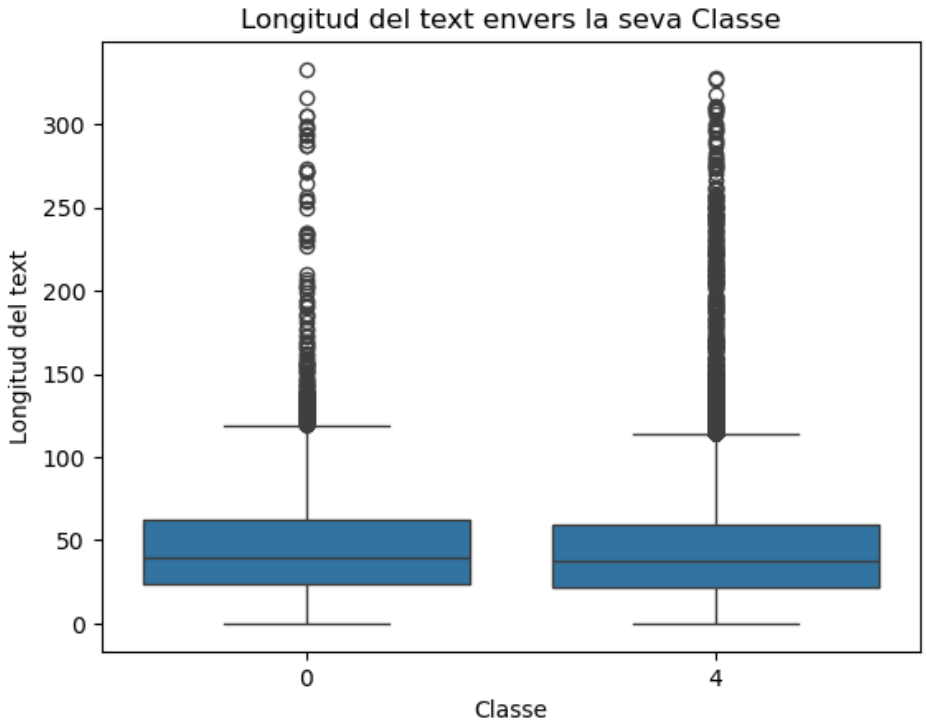


De la gràfica es pot observar que la majoria dels missatges tenen una longitud compresa entre els 0 i 100 caràcters, amb un pic molt destacat al voltant dels 50 caràcters. A mesura que augmenta la longitud dels missatges, la seva freqüència disminueix de manera consistent. Encara que hi ha un petit pic entorn dels 120 caràcters, aquest descens continua posteriorment.

Addicionalment, s'han calculat diverses estadístiques descriptives. Per exemple, el nombre total de caràcters en tot el conjunt de dades i la longitud mitjana dels missatges, que és de 74 caràcters. Aquesta dada concorda amb la gràfica, ja que la mitjana es troba en el rang de longituds més freqüent. També s'han identificat altres estadístiques interessants, com la desviació estàndard, el missatge més curt, el missatge més llarg (374 caràcters), així com els valors dels quartils i la mediana.

```
Estadístiques de la longitud del text:
count    1.599999e+06
mean     7.409009e+01
std      3.644114e+01
min      6.000000e+00
25%      4.400000e+01
50%      6.900000e+01
75%      1.040000e+02
max      3.740000e+02
Name: len_text, dtype: float64
```

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.



Les distribucions de longitud són pràcticament idèntiques en ambdues classes, amb la major part dels textos situats entre els 40 i 100 caràcters, i una mediana molt similar, al voltant dels 50-70 caràcters. Aquestes dades evidencien que la longitud del text no sembla ser un factor discriminant entre les dues classes.

A més, hi ha alguns valors atípics (outliers) que corresponen a textos més llargs, alguns fins i tot superant els 300 caràcters. No obstant això, aquests casos són poc freqüents i no modifiquen significativament el patró general. Això reforça la idea que la longitud del text aporta poca informació per diferenciar les dues classes

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:

	total	Negative	Positive	Positive %
good	91328	29207	62121	68.019665
day	87058	39858	47200	54.216729
get	82044	45542	36502	44.490761
like	78567	41048	37519	47.754146
go	73932	45582	28350	38.346048
today	68210	38115	30095	44.121097
work	65000	45455	19545	30.069231
love	64686	16990	47696	73.734657
going	64617	33685	30932	47.869756
got	61296	33289	28007	45.691399
lol	59269	23135	36134	60.966104
time	57945	27516	30429	52.513590
back	56934	33076	23858	41.904662
u	55858	23937	31921	57.146693
one	53932	27257	26675	49.460432
im	52239	31256	20983	40.167308
know	52043	26310	25733	49.445651
really	50026	31497	18529	37.038740
dont	18226	12764	5462	29.968177
cant	17677	12424	5253	29.716581

A partir de la taula, es pot observar que algunes paraules estan fortament associades amb un sentiment específic. Per exemple, paraules com 'love' (73.73%), 'good' (68.01%) i 'lol' (60.96%) mostren una clara associació amb sentiments positius, la qual cosa reflecteix que s'utilitzen sovint en contextos positius. D'altra banda, paraules com 'dont' (29.97%) i 'cant' (29.72%) es troben majoritàriament en missatges amb connotacions negatives.

Algunes paraules, com 'today' (44.10%) i 'time' (52.51%), tenen una distribució més equilibrada entre sentiments positius i negatius, suggerint que el seu significat pot dependre del context en què es fan servir. Aquest patró subratlla la importància del context per interpretar correctament l'emoció associada a certes paraules.

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 de la paraula “day”, i fent servir expressions elegres com “happy mothers day” o “hope great day”

En els tweets negatius, els bígramas més comuns inclouen frases amb la paraula “miss”, així com expressions que es poden identificar fàcilment com a desmotivades o sense sentiment d’alegria, com podria ser “dont feel good”.

```
Dades de Twitter carregades correctament.
Bigrames més comuns en tweets positius:
good morning: 7990
im going: 4867
good luck: 3806
good night: 3667
looking forward: 3369
happy birthday: 3107
good day: 2894
great day: 2790
getting ready: 2766
mother day: 2725

Bigrames més comuns en tweets negatius:
feel like: 6502
dont know: 6224
im going: 5895
im sorry: 5588
dont want: 5124
look like: 4358
im gonna: 3921
dont think: 3376
im sad: 2898
dont like: 2731
```

En els tweets positius, els trígrames més comuns reflecteixen celebracions i agraïments. Frases com "happy mother day" i "100 follower day" indiquen moments especials i fites importants per als usuaris. Altres combinacions, com "using add train" i "add train pay", semblen estar relacionades amb promocions o automatismes en els missatges. També hi ha referències a premis i esdeveniments, com "mtv movie award", que destaquen un ambient festiu i positiu.

En canvi, en els tweets negatius, els trígrames més comuns expressen pèrdua, tristesa i frustració. Frases com "im gonna miss", "hope feel better" i "dont feel good" reflecteixen emocions personals intenses. També s'observen expressions de suport, com "help good home" i "lost help good", que suggereixen la cerca de consol o ajuda en moments difícils. En conjunt, aquests trígrames mostren una clara diferència en el to i les emocions expressades respecte als tweets positius.

```
Trigrames més comuns en tweets positius:
happy mother day: 1964
100 follower day: 1486
follower day using: 1484
day using add: 1484
using add train: 1484
add train pay: 1484
train pay vip: 1484
mtv movie award: 1024
new moon trailer: 489
hope great day: 363

Trigrames més comuns en tweets negatius:
im gonna miss: 929
hope feel better: 851
im going miss: 751
dont feel good: 750
feel like im: 641
im sorry hear: 567
lost help good: 550
help good home: 550
getting ready work: 504
dont want work: 504
```