Universitat Oberta de Catalunya

uoc.edu

ANALÍTICA DE CLIENTS

REPTE 5. SENTIMENT ANALYSIS I TOPIC MODELING

AUTOR: JOAN MIQUEL ALFONSO GARCIA
PROFESSOR: POL CAPDEVILA SALINAS

Grau de ciències de dades aplicades
Primer semestre de 2024

PAC 5: Sentiment Analysis i Topic Modeling

Presentació

Aquesta PAC presenta un cas d'ús per aplicar data mining amb diferentes tècniques de preprocessament per passar de dades desestructurades a una estructura òptima. Es treballarà amb un dataset que conté tuits extrets des de l'API de Twitter de diferents usuaris. A partir d'aquest dataset, es treballaran diferentes tècniques de preprocessament de text per a l'anàlisi d'opinions basat en tuits.

Competències

En aquesta PAC es treballen les següents competències:

1. Identificar, comprendre i reconèixer noves oportunitats de millora en qualsevol mena d'organització que poden ser resoltes de manera eficient i efectiva mitjançant la ciència de les dades.
2. Utilitzar de forma combinada els fonaments matemàtics, estadístics i de programació per desenvolupar solucions a problemes en l'àmbit de la ciència de les dades.
3. Resumir, interpretar, presentar i contrastar de manera crítica els resultats obtinguts utilitzant les eines d'anàlisis i visualització més adequades.

Objectius

Els objectius d'aquesta PAC són:

- Conèixer i aprendre les tècniques de Text Mining per passar de dades desestructurades a una estructura òptima.
- Conèixer i saber utilitzar l'anàlisi de sentiments.
- Conèixer i saber utilitzar algorismes de Topic Modelling
- En un cas d'ús determinat saber prendre decisions que aportin valor al negoci en funció de l'estudi de les dades.
- Entendre les implicacions ètiques de la governança de les dades.

Descripció de la PAC

Aquesta activitat ens permetrà posar en pràctica els coneixements i procediments treballats en aquest repte.
A partir d'un dataset, demanem que resolgueu una sèrie d'exercicis on s'hauran d'aplicar els procediments que hem anat treballant.

Recursos

Els recursos d'aprenentatge relacionats amb aquesta PAC es poden trobar a l'aula de l'assignatura.

Criteris d'avaluació

- La PAC ha de resoldre de manera individual.
- És necessari justificar totes les respostes de les preguntes de la PAC.
- La nota de la PAC5 serà part de la nota d'avaluació contínua de l'assignatura amb un pes del 20% respecte al total. Per a més informació sobre el model d'avaluació de l'assignatura podeu consultar el pla docent.

Format i data de lliurament

És necessari lliurar un document PDF amb les respostes de tots els exercicis i un arxiu .py o .ipynb amb el codi realitzat per cada pregunta que aplicació.
És necessari realitzar el document PDF amb la solució de la PAC amb un processador de textos ja que no s'acceptaran solucions a mà i escanejades.
Aquests arxius han de lliurar-se en l'espai de Lliurament de l'aula abans de les 23.59 del dia 24/06/2024. No s'acceptaran lliuraments fora de termini.

Repte

Des de l'empresa de Twitter realitzen preprocessament del llenguatge natural (NLP) per a la creació de models que preprocesen i comprenen el llenguatge natural, entenen l'agrupació semàntica de les paraules, la conversió de text en veu, la traducció del llenguatge i molt més. S'han posat en contacte amb nosaltres perquè volen fer una anàlisi de sentiment per interpretar i classificar les emocions (positives, negatives i neutres) dels seus usuaris contra diverses aerolínies d'USA i ens han demanat ajuda.
La seva idea és poder identificar el sentiment del públic cap a determinades paraules o temes i poder filtrar discursos negatius no desitjats en les seves plataformes.
Ens han proporcionat un dataset parcial que conté informació dels tuits. Respon a les següents preguntes raonadament.

Activitat 1: Anàlisi de les dades (25%)

Per a aquesta activitat es recomana utilitzar el paquet NLTK (Natural Language Toolkit) de Python.

a) Explora el dataset, quantes mostres té? Quines variables tenen valors nuls? Realitza els procediments adequats per tenir un dataset amb informació constructiva.

El dataset té 14640 mostres i 15 variables.

Cadascuna de les variables és:

tweet_id	numèric
airline_sentiment	categòrica
airline_sentiment_confidence	numèric
negativereason	categòrica
negativereason_confidence	numèric
airline	categòrica
airline_sentiment_gold	categòrica
name	categòrica
negativereason_gold	categòrica
retweet_count	numèrica
text	categòrica

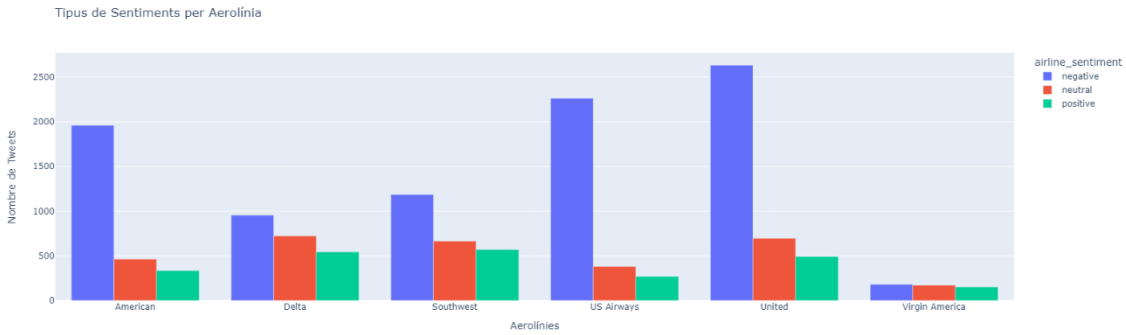
tweet_coord	categòrica
tweet_created	categòrica → l'haurem de posar com datetime
tweet_location	categòrica
user_timezone	categòrica

Pel que fa als nulls en cada variable tenim:

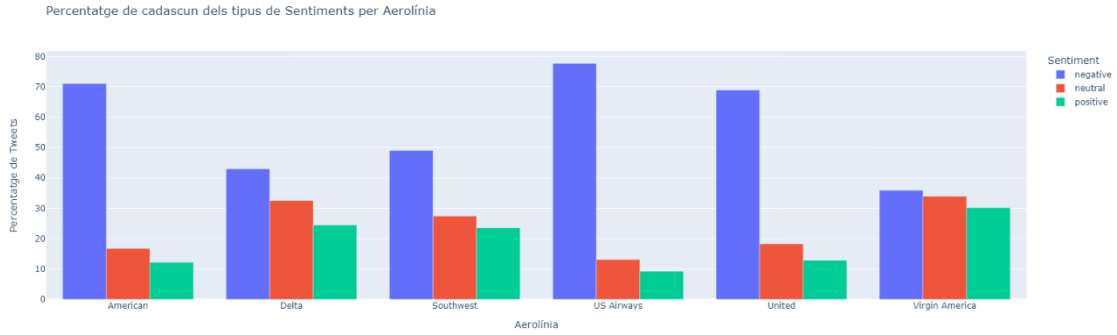
tweet_id	0
airline_sentiment	0
airline_sentiment_confidence	0
negativereason	5462
negativereason_confidence	4118
airline	0
airline_sentiment_gold	14600
name	0
negativereason_gold	14608
retweet_count	0
text	0
tweet_coord	13621
tweet_created	0
tweet_location	4733
user_timezone	4820

Com en aquest cas solament emprarem posteriorment les variables tweet_id, airline_sentiment, negativereason i text. Que no tenen cap valor nul (negativereason quan l'emprem és en els valors que són airline_sentiment negatiu, i en aquest cas no hi ha cap valor nul). No és necessari fer cap tractament dels valors nuls. Tampoc és necessari normalitzar dades, ja que airline_sentiment i negativereason seran les variables objectius i text, que és la variable explicativa la transformarem posteriorment.

b) Divideix els tuits en tres grups: positius, negatius i neutres, realitza un gràfic de barres de sentiments dels usuaris per cada aerolínia. Quines aerolínies tenen comentaris més negatius? I comentaris més positius? Observant els gràfics, què dedueixes dels tuits negatius?



L'aerolínia que més comentaris negatius té és United amb 2663. L'aerolínia amb més comentaris positius és Southwest amb 570.



Respecte a percentatge del total de tweets de cada aerolínia, veiem que el percentatge major de tweets negatius és per a US Airways, mentre que el percentatge major de tweets positius és per a Virgin America.

Així, podem deduir d'aquesta distribució, que els viatgers no solen escriure tweets d'experiències positives, sinó que tenen tendència a escriure tweets quan tenen una experiència negativa. Per tant, fins a l'aerolínia que té major percentatge de tweets positius i neutres, té més percentatge de tweets positius.

c) Per a tenir una idea de les paraules més freqüents en els tuits negatius i positius, representa un núvol de paraules amb el paquet wordcloud de Python. Quines són les paraules més freqüents en tuits amb sentiments negatius? I en els positius?

Wordcloud sentiments negatius:

Paraules més freqüents en tweets de sentiment negatiu:

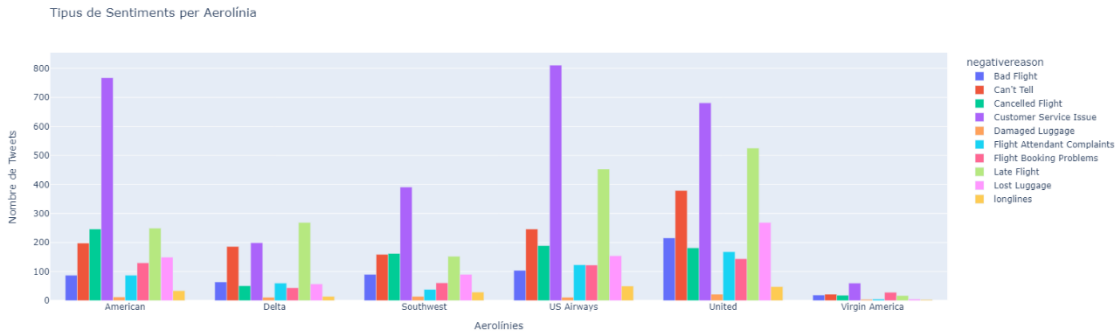
'flight': 2782,
'get': 981,
'cancelled': 912,
'service': 674,
'hours': 598,
'customer': 592,
'hold': 586,
'time': 533,
'2': 520,
"i'm": 497,
'still': 474,
'plane': 473

Wordcloud sentiments positius:

Paraules més freqüents en tweets de sentiment positiu:

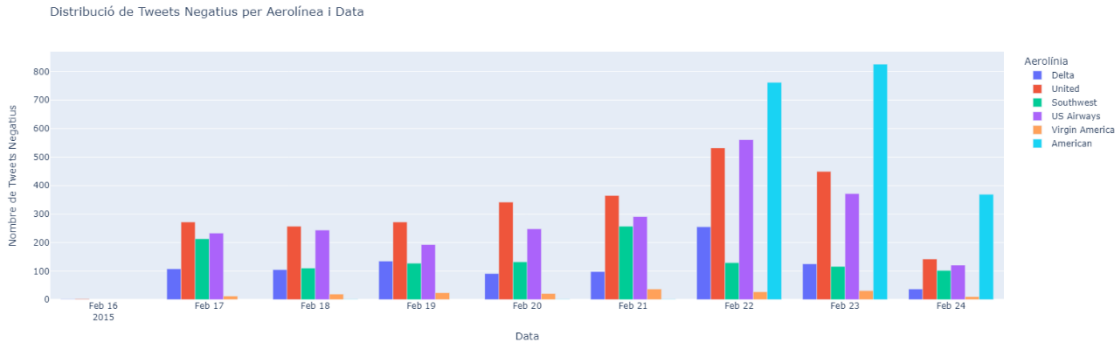
'thanks': 478,
'thank': 453,
'flight': 349,
'great': 217,
'you!': 134,
'service': 133,
'love': 124,
':)': 119,
'thanks!': 118,

d) Extreu conclusions sobre les raons dels sentiments negatius dels tuits dels clients representant la distribució de la variable 'negativereason' per cada aerolínia. Quina és la raó que genera més sentiments negatius? Destacaries alguna aerolínia? Justifica la teva resposta.



La raó que genera més comentaris negatius és una incidència amb el servei al client. Entre les aerolínies, destaca l'aerolínia Delta, la qual és l'única aerolínia que no té com a raó principal les incidències amb el servei al client, encara que la resta de raons estan paregudes a la resta de les aerolínies, per la qual cosa, ens pot dir que el servei d'atenció al client d'aquesta aerolínia possiblement és molt millor que el de les seues competidores.

e) Existeix alguna relació entre els sentiments negatius dels usuaris en les aerolínies i la data? El dataset té registres des de 2015-feb-17 fins a 2015-feb-24, justifica la teva resposta mitjançant la visualització en un gràfic de barres de la distribució de tuits negatius en cada aerolínia per cada data. Quines conclusions obtens?



Veiem com la majoria dels tweets negatius es distribueixen el 22 i el 23 de febrer. És curiós com els tweets negatius de United i US Airways, els dies anteriors tenen una quantitat considerable de tweets negatius (després creix, aquest el 22 de febrer), però American Airlines no té cap comentari negatiu fins al 22 de febrer, i a partir d'aquest dia és l'aerolínia amb més tweets negatius en diferència.

Buscant informació es veu que hi va haver una onada de fred al febrer (on en algunes ciutats com Chicago es van assolir temperatures que no s'havien assolit des de 1934). A més el 14-15 de febrer hi va haver una tempesta de neu. A més he trobat que hi va haver

més de 1000 cancel·lacions i 5000 retards (açò es trasllada amb problemes amb els serveis als consumidors, ja que són els encarregats de gestionar aquesta problemàtica), el diumenge 22 de febrer i el dilluns 23 de febrer. Si bé, no podem trobar cap incidència que afectara particularment a American Airlines, per la qual cosa no podem explicar perquè no hi ha cap tweet negatiu els dies anteriors al 22 de febrer, i perquè tots aquests tweets negatius es concentren el 22, 23 i 24 de febrer per a aquesta aerolínia.

https://en.wikipedia.org/wiki/February_2015_North_American_cold_wave
https://en.wikipedia.org/wiki/February_14%E2%80%9315,_2015_North_American_blizzard
https://eu.usatoday.com/story/todayinthesky/2015/02/21/cancellations-near-1000-from-latest-air-travel-weather-woes/23797113/

Activitat 2: Preprocesamiento de text (30%)

La part més important del NLP és el preprocessament de text per passar d'un llenguatge lliure a un format òptim en màquina per al seu posterior processament.
Atès que hi ha molts paràmetres innecessaris en el conjunt de dades, ens quedarem amb les columnes 'airline_sentiment', 'tweet_id' i el text del tuit.

Realitza els següents apartats sobre la columna 'text' i emmagatzema els resultats en una columna nova, 'cleaned_text':

f) La primera implementació que farem es basarà a aplicar **tokenització**, és a dir, per a cada seqüència de caràcters, ens permet dividir en peces discretes anomenades tokens (1 paraula = 1 token). Normalment, aquest procés implica l'eliminació de certs caràcters, com els signes de puntuació.
Es recomana utilitzar la biblioteca NLTK que té un *tokenitzador* incorporat.

En aquest apartat també he realitzat l'eliminació de signes de puntuació i convertir totes les paraules a minúscula, encara que corresponguera a l'apartat h).

	tweet_id	airline_sentiment	text	cleaned_text
1	570301130888122368	positive	@VirginAmerica plus you've added commercials to the experience... tacky.	[virginamerica, plus, you, ve, added, commercials, to, the, experience, tacky]
3	570301031407624196	negative	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse	[virginamerica, it, s, really, aggressive, to, blast, obnoxious, entertainment, in, your, guests, faces, amp, they, have, little, recourse]
4	570300817074462722	negative	@VirginAmerica and it's a really big bad thing about it	[virginamerica, and, it, s, a, really, big, bad, thing, about, it]
5	570300767074181121	negative	@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing.\nit's really the only bad thing about flying VA	[virginamerica, seriously, would, pay, a, flight, for, seats, that, didn, t, have, this, playing, it, s, really, the, only, bad, thing, about, flying, va]
6	570300616901320704	positive	@VirginAmerica yes, nearly every time I fly VX this "ear worm" won't go away :)	[virginamerica, yes, nearly, every, time, i, fly, vx, this, ", ear, worm, ", won, ', t, go, away]
...	...	...	...	...
14633	569587705937600512	negative	@AmericanAir my flight was Cancelled Flightled, leaving tomorrow morning. Auto rebooked for a Tuesday night flight but need to arrive Monday.	[americanair, my, flight, was, cancelled, flightled, leaving, tomorrow, morning, auto, rebooked, for, a, tuesday, night, flight, but, need, to, arrive, monday]
14634	569587691626622976	negative	@AmericanAir right on cue with the delays👌	[americanair, right, on, cue, with, the, delays👌]
14635	569587686496825344	positive	@AmericanAir thank you we got on a different flight to Chicago.	[americanair, thank, you, we, got, on, a, different, flight, to, chicago]
14636	569587371693355008	negative	@AmericanAir leaving over 20 minutes Late Flight. No warnings or communication until we were 15 minutes Late Flight. That's called shitty customer svc	[americanair, leaving, over, minutes, late, flight, no, warnings, or, communication, until, we, were, minutes, late, flight, that, s, called, shitty, customer, svc]
14638	569587188687634433	negative	@AmericanAir you have my money, you change my flight, and don't answer your phones! Any other suggestions so I can make my commitment??	[americanair, you, have, my, money, you, change, my, flight, and, don, t, answer, your, phones, any, other, suggestions, so, i, can, make, my, commitment]

11541 rows × 4 columns

g) Eliminem les **Stop Words**, que bàsicament són preposicions o adverbis que no ajuden a determinar la qualitat semàntica del tuit. Per a això, es recomana utilitzar el mòdul de corpus i importar la funció de stopwords.

	tweet_id	airline_sentiment	text	cleaned_text
1	570301130888122368	positive	@VirginAmerica plus you've added commercials to the experience... tacky.	[virginamerica, plus, added, commercials, experience, tacky]
3	570301031407624196	negative	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse	[virginamerica, really, aggressive, blast, obnoxious, entertainment, guests, faces, amp, little, recourse]
4	570300817074462722	negative	@VirginAmerica and it's a really big bad thing about it	[virginamerica, really, big, bad, thing]
5	570300767074181121	negative	@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing.\nit's really the only bad thing about flying VA	[virginamerica, seriously, would, pay, flight, seats, playing, really, bad, thing, flying]
6	570300616901320704	positive	@VirginAmerica yes, nearly every time I fly VX this "ear worm" won't go away :)	[virginamerica, yes, nearly, every, time, fly, ear, worm, away]
...	...	...	...	...
14633	569587705937600512	negative	@AmericanAir my flight was Cancelled Flightled, leaving tomorrow morning. Auto rebooked for a Tuesday night flight but need to arrive Monday.	[americanair, flight, cancelled, flightled, leaving, tomorrow, morning, auto, rebooked, tuesday, night, flight, need, arrive, monday]
14634	569587691626622976	negative	@AmericanAir right on cue with the delays👌	[americanair, right, cue, delays👌]
14635	569587686496825344	positive	@AmericanAir thank you we got on a different flight to Chicago.	[americanair, thank, got, different, flight, chicago]
14636	569587371693355008	negative	@AmericanAir leaving over 20 minutes Late Flight. No warnings or communication until we were 15 minutes Late Flight. That's called shitty customer svc	[americanair, leaving, minutes, late, flight, warnings, communication, minutes, late, flight, called, shitty, customer, svc]
14638	569587188687634433	negative	@AmericanAir you have my money, you change my flight, and don't answer your phones! Any other suggestions so I can make my commitment??	[americanair, money, change, flight, answer, phones, suggestions, make, commitment]

11541 rows × 4 columns

h) A continuació, heu de convertir totes les paraules en minúscules i eliminar els **signes de puntuació** que no aporten significat. Es recomana utilitzar el paquet re de Python.

A efectes gramaticals, els documents utilitzen diferents formes d'una paraula (look, looks, looking, looking) que en moltes situacions tenen qualitats semàntiques molt similars. El **stemming** redueix les paraules a la seva paraula stem o arrel, de 'jumping, jumped' seria 'jump'. La **lematització** és l'eliminació de lletres amb prefix o sufix d'una paraula, el resultat pot ser o no una paraula del corpus lingüístic, de 'am,are a 'be'.

i) Realitza una de les dues tècniques anteriors sobre les dades treballades. Per a fer això, es recomana utilitzar el mòdul de stem de NLTK que ofereix diversos algorismes. Justifica l'elecció de l'algorisme seleccionat.

Com a eina de suport us adjuntem un blog per manejar dades NLP i preprocessament de dades.

	tweet_id	airline_sentiment	text	cleaned_text
1	570301130888122368	positive	@VirginAmerica plus you've added commercials to the experience... tacky.	[virginamerica, plus, added, commercial, experience, tacky]
3	570301031407624196	negative	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse	[virginamerica, really, aggressive, blast, obnoxious, entertainment, guest, face, amp, little, recourse]
4	570300817074462722	negative	@VirginAmerica and it's a really big bad thing about it	[virginamerica, really, big, bad, thing]
5	570300767074181121	negative	@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing.\nit's really the only bad thing about flying VA	[virginamerica, seriously, would, pay, flight, seat, playing, really, bad, thing, flying]
6	570300616901320704	positive	@VirginAmerica yes, nearly every time I fly VX this "ear worm" won't go away :)	[virginamerica, yes, nearly, every, time, fly, ear, worm, away]
...	...	...	...	...
14633	569587705937600512	negative	@AmericanAir my flight was Cancelled Flightled, leaving tomorrow morning. Auto rebooked for a Tuesday night flight but need to arrive Monday.	[americanair, flight, cancelled, flightled, leaving, tomorrow, morning, auto, rebooked, tuesday, night, flight, need, arrive, monday]
14634	569587691626622976	negative	@AmericanAir right on cue with the delays🔥	[americanair, right, cue, delays🔥]
14635	569587686496825344	positive	@AmericanAir thank you we got on a different flight to Chicago.	[americanair, thank, got, different, flight, chicago]
14636	569587371693355008	negative	@AmericanAir leaving over 20 minutes Late Flight. No warnings or communication until we were 15 minutes Late Flight. That's called shitty customer svc	[americanair, leaving, minute, late, flight, warning, communication, minute, late, flight, called, shitty, customer, svc]
14638	569587188687634433	negative	@AmericanAir you have my money, you change my flight, and don't answer your phones! Any other suggestions so I can make my commitment??	[americanair, money, change, flight, answer, phone, suggestion, make, commitment]

11541 rows × 4 columns

En base als resultats obtinguts, identifiques alguna paraula que no aporti significat al sentiment del tuit a primera vista? Què creus que pot estar passant amb les ortografies incorrectes?

Sí, tenim paraules que són genèriques i no podem dir que aporten cap significat als sentiments, com poden ser “really”, “flight”, “flying”, “thing”...

En les errades ortogràfiques, en primer lloc, el stop words, no la detecta com a una stop word, doncs no l’eliminarà, ja que aquesta funció actua detectant i eliminant les paraules que té programades com stop words.

Respecte a la lematització, és possible que no puga detectar correctament la seua forma base correcta i no realitzar correctament la lematització d’aquesta paraula. Així, les errades ortogràfiques poden afectar negativament el nostre model.

j) Una altre metodologia interessant per aquest tipus de cas d’ús són els anomenats Transformers de text, per exemple tenim els models BERT o GPT. Explica la principal diferència entre ells.

Els models BERT (Biderectional Encoder Representations from Transformers) i GPT (Generative Pre-trained Transformer) són dos tipus de models de transformació per al processament del llenguatge natural. Aquest model tenen les següents diferències.

Respecte a l’arquitectura, BERT utilitza una arquitectura Transformer bidireccional, és a dir, processa el text d’entrada en ambdues direccions simultàniament. Mentre que GPT empra una arquitectura Transformer

unidireccional, processant el text d'esquerra a dreta, permetent predir la paraula següent en una seqüència.

Pel que fa a com es va entrenar cada model, BERT va ser entrenat mitjançant una tasca de model de llenguatge emmascarat (MLM), aquesta tècnica s'emmascaren algunes paraules d'una oració i el model prediu les paraules segons el context. Mentre que GPT fou entrenat amb un model de llenguatge causal (CLM) on el model prediu la paraula següent en una seqüència.

BERT és molt bo per respondre tasques de classificació, però no pot generar text. La qual cosa no passa amb GPT, que és adequat per a la generació de text, traducció...

https://vitalflux.com/bert-vs-gpt-differences-real-life-examples/

k) A partir d'un BERT, 'bert-base-cased' (https://huggingface.co/bert-base-cased), computa els vectors representatius del text dels tweets i guarda'ls a una nova columna anomenada "embeddings_text".

	tweet_id	airline_sentiment	text	cleaned_text	embeddings_text
1	570301130888122368	positive	@VirginAmerica plus you've added commercials to the experience... tacky.	[virginamerica, plus, added, commercial, experience, tacky]	[0.34247345, 0.22057413, 0.008949173, -0.3577943, -0.27422562, -0.08858224, 0.22067057, 0.0928336, 0.09489015, -1.0199862, -0.1529686, 0.21117334, -0.28810266, -0.29660738, -0.4194882, -0.02842718, 0.12961699, 0.1166038, -0.07249626, -0.5568561, 0.22686996, -0.28318593, 0.52959985, -0.17645791, 0.35819685, 0.018861694, 0.22552766, -0.011022514, -0.12486034, 0.23631243, 0.12823568, 0.44937393, -0.05863446, 0.20718014, -0.16310391, -0.041884314, -0.070925914, -0.29174942, -0.028660534, 0.015902378, -0.44798213, 0.2681426, 0.7953395, -0.066303894, 0.21035163, -0.2701232, 0.03737183, 0.08218401, -0.1773099, -0.029167602, 0.049675606, -0.4884795, 0.08896434, -0.06098332, 0.12510744, 0.4850182, -0.41119772, 0.12975779, -0.5801559, 0.23747961, -0.07376934, 0.0873681, 0.26947057, 0.29070324, -0.36722124, -0.14208436, 0.027412184, 0.11861879, -0.3248775, 0.2924903, -0.13809787, 0.15445977, 0.5266443, 0.8911838, 0.34949285, -0.2768948, 0.55006516, 0.005939119, 0.1753455, 0.09649554, 0.03220263, 0.17608324, -0.11541479, -0.20296584, -0.23194173, -0.015657783, 0.15126485, -0.01664991, 0.111244574, 0.43098176, 0.41382527, -0.12565134, -0.3699866, 0.27756983, 0.043457013, 0.0651534, -0.25218514, 0.07145675, 5.9267406, 0.16856298, ...]
3	570301031407624196	negative	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse	[virginamerica, really, aggressive, blast, obnoxious, entertainment, guest, face, amp, little, recourse]	[0.48871082, 0.2622475, 0.06119926, -0.31489217, -0.08650649, -0.11688634, 0.15016045, 0.11267535, 0.022885753, -1.1291866, -0.18148486, 0.1966306, -0.26415908, 0.0393627, -0.7071261, 0.12903662, 0.18057232, 0.09339162, -0.1426701, -0.03624054, 0.072604425, -0.19664378, 0.44522297, -0.29489917, 0.23105955, 0.12809914, 0.27934968, 0.29582134, -0.185061, 0.24890839, 0.32102168, 0.40487805, -0.1429067, 0.15904151, -0.3869651, 0.021616692, -0.2773302, -0.16128547, -0.10159528, -0.12664665, -0.2780022, 0.33990094, 0.47406015, 0.008521909, 0.2625427, -0.5495919, 0.023400156, -0.11849601, -0.19705603, 0.11786363, 0.14734362, 0.025323933, 0.02911701, 0.30724624, -0.054124616, 0.24350512, -0.3142185, 0.14119533, -0.6852971, 0.176424, -0.052242737, 0.091496214, 0.417991, -0.15708973, -0.3415683, 0.0013449641, -0.2298149, 0.22359551, -0.13981323, 0.090473354, -0.083367296, 0.06497366, 0.25313488, 0.7239459, 0.43130913, -0.26689968, 0.6841533, 0.053413015, 0.16033706, 0.20776244, -0.114443064, 0.31158444, -0.25697246, -0.29937327, -0.4344422, -0.046752647, 0.09240279, -0.16981435, -0.03806989, 0.21487516, 0.65549505, -0.10835452, -0.6154646, 0.020074446, -0.15604578, 0.4578292, -0.010118667, 0.073750414, 5.1386957, -0.18991332, ...]

Finalment, dividirem les dades en conjunt d'entrenament, test i validació per realitzar l'activitat 3 i poder classificar tuits en positius i negatius. Per a fer això es considera com a variable explicativa 'cleaned_text' pel primer cas d'ús, 'embeddings_text' pel segon cas d'ús i com a variable objectiu pels dos casos d'ús 'airline_sentiment'.

l) Donant-te suport en la funció train_test_split de scikit-learn, divideix les dades d'entrenament 70%, validació 10% i test 20%; un fitxer per a cada cas d'ús. Per tant, tindrem dos fitxers de train anomenats X_train_cas1, X_train_cas2…

Es recomana incloure random_state=126 perquè els resultats siguin reproduïbles.

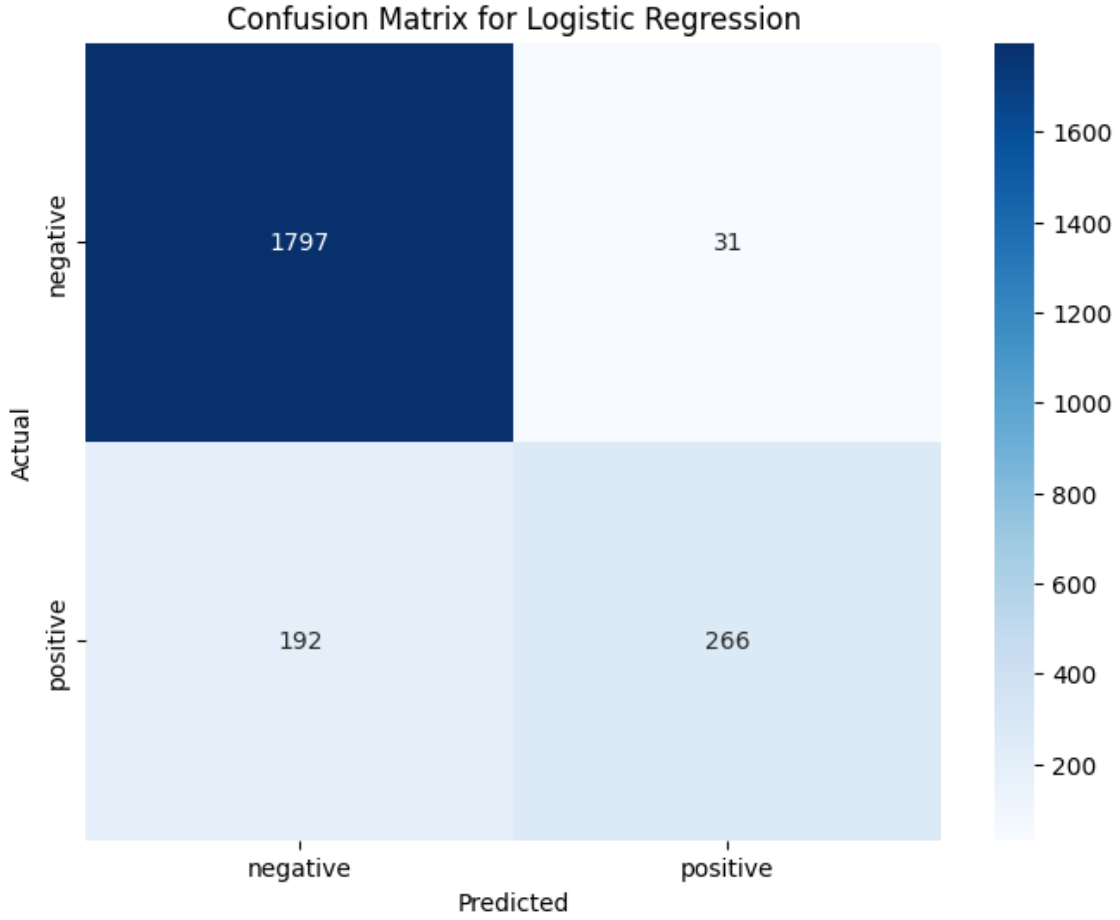
m) La freqüència de termes o TF-IDF sol utilitzar-se per produir pesos associats a paraules que poden ser útils en cerques de recuperació d'informació. Utilitza la implementació de scikit-learn de TfidfVectorizer per convertir la col·lecció de tuits sense processar en una matriu de funcions **TF-IDF** sobre les mostres del primer cas d'ús, és a dir, sobre X_train_cas1,X_val_cas1,X_test_cas1.

Activitat 3: Topic Modelling (35%)

Ja tenim les dades preparades per a poder entrenar els models predictius per classificar els tuits en positius i negatius.

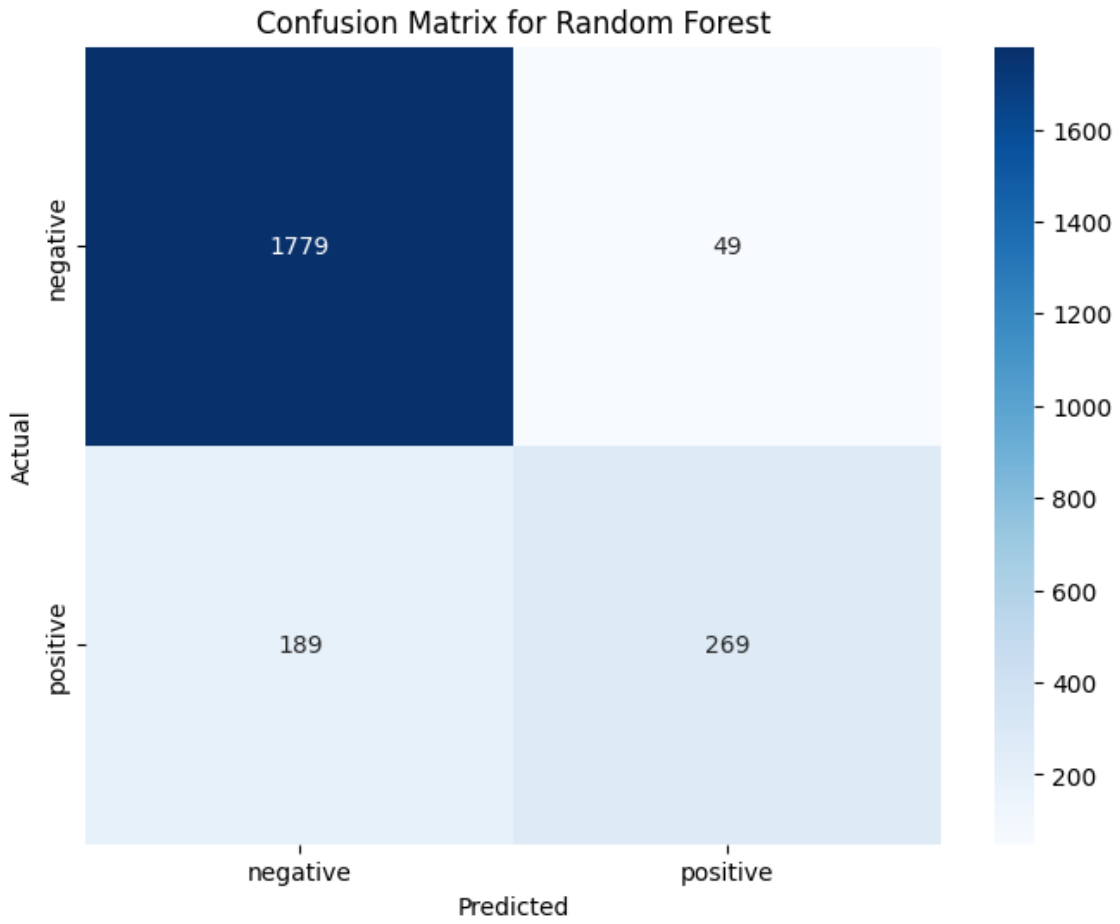
Amb els fitxers del primer cas d'ús, realitza:

a) Entrena un Logistic Regression, un Random Forest, un LinearSVC i un Gradient Boosting Classifier de la llibreria de sklearn amb els hiperparàmetres en default excepte el random state, que usarem el 126. Construeix la matriu de confusió per cada model amb el dataset de proves i contesta les següents preguntes:



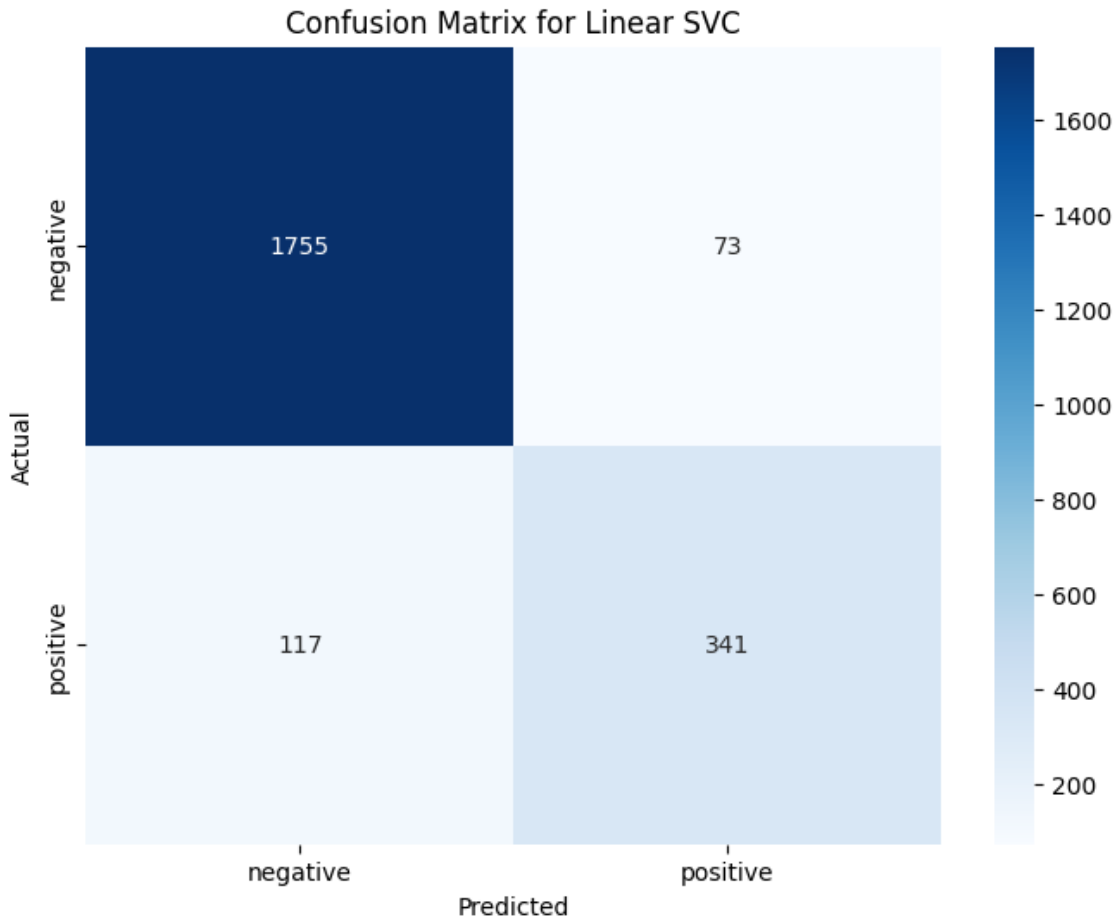
Classification Report for Logistic Regression:

	precision	recall	f1-score	support
negative	0.90	0.98	0.94	1828
positive	0.90	0.58	0.70	458
accuracy			0.90	2286
macro avg	0.90	0.78	0.82	2286
weighted avg	0.90	0.90	0.89	2286



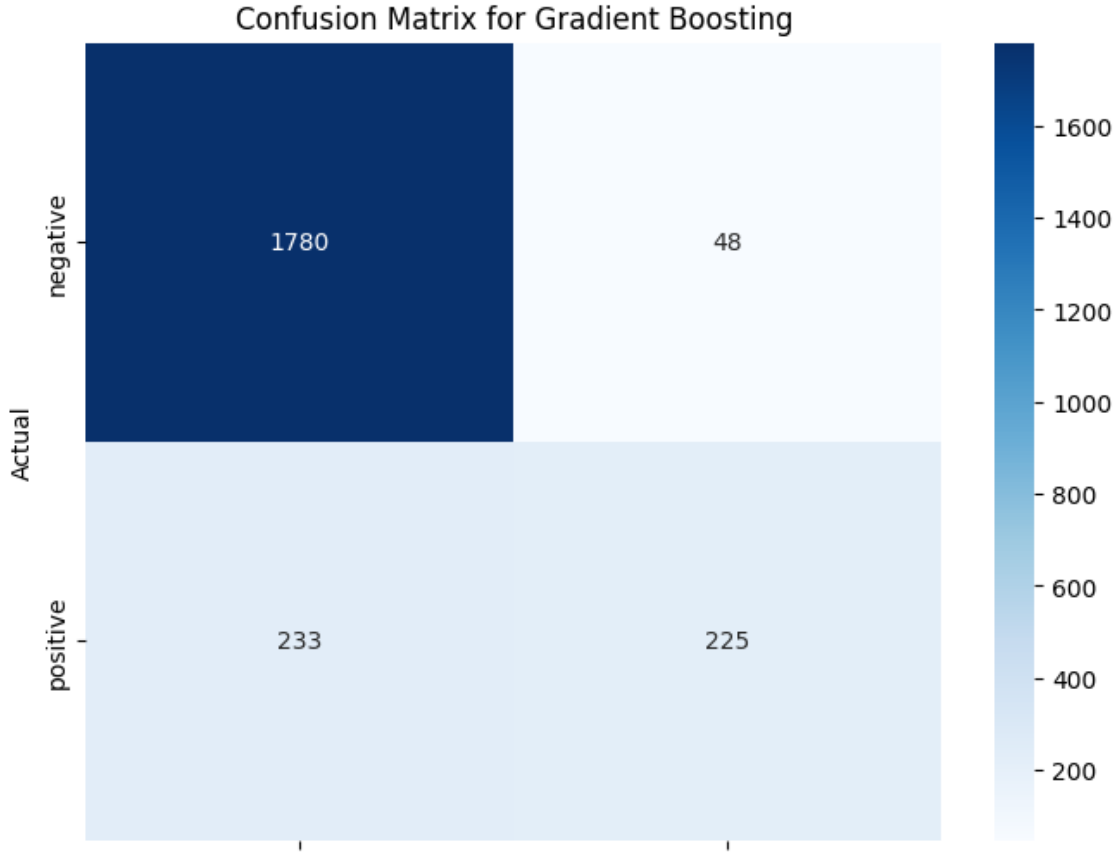
Classification Report for Random Forest:

	precision	recall	f1-score	support
negative	0.90	0.97	0.94	1828
positive	0.85	0.59	0.69	458
accuracy			0.90	2286
macro avg	0.87	0.78	0.82	2286
weighted avg	0.89	0.90	0.89	2286



Classification Report for Linear SVC:

	precision	recall	f1-score	support
negative	0.94	0.96	0.95	1828
positive	0.82	0.74	0.78	458
accuracy			0.92	2286
macro avg	0.88	0.85	0.87	2286
weighted avg	0.91	0.92	0.92	2286



	precision	recall	f1-score	support
negative	0.88	0.97	0.93	1828
positive	0.82	0.49	0.62	458
accuracy			0.88	2286
macro avg	0.85	0.73	0.77	2286
weighted avg	0.87	0.88	0.86	2286

i) Quin model té millor accuracy?

El model amb millor accuracy és Linear SVC amb un 0.92, seguit de Logistic Regression i Random Forest amb un 0.9 i Gradient Boosting amb 0.88.

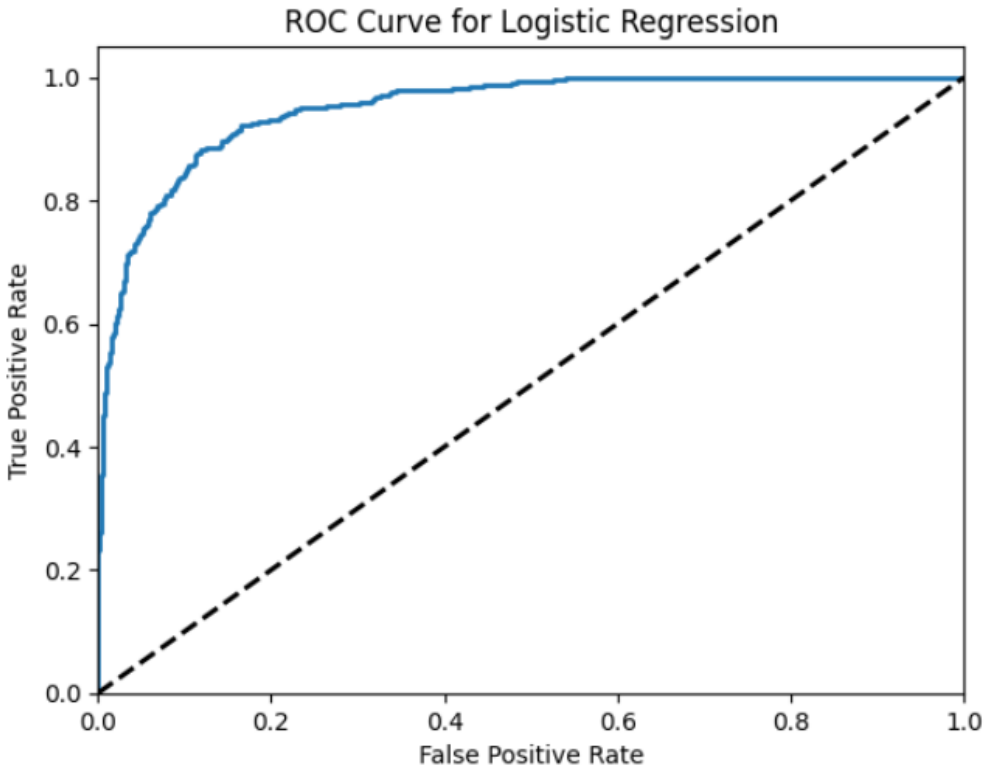
ii) Quin té millor recall?

El millor recall de mitjana ponderada ho té també Linear SVC amb un 0.92, seguit de Logistic Regression i Random Forest amb un 0.9 i Gradient Boosting amb 0.88.

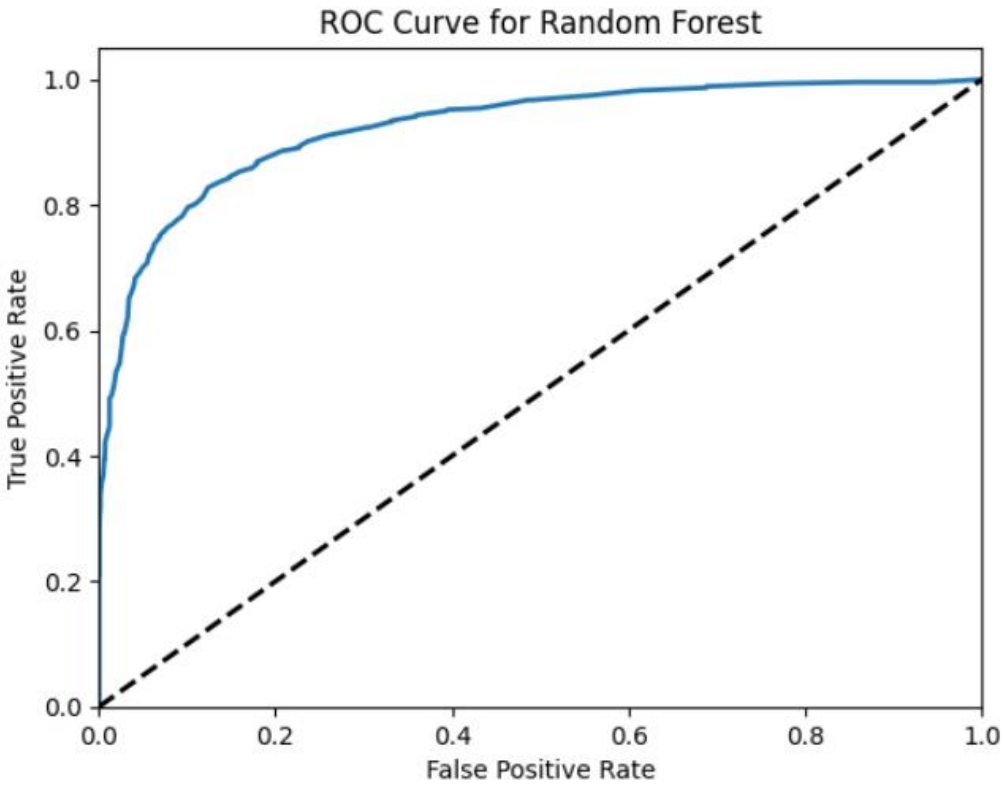
Respecte al millor Recall respecte a les prediccions positives, Linear SVC és el que millor puntuació té amb 0.74, el qual és una puntuació decent, ja que de cada 100 prediccions positives, 74 són correctes. La resta de models tenen menys de 0.6, la qual cosa ens diu que classifica massa tweets negatius com positius.

iii) Implementa la Corba ROC per cada model. Explica amb les teves paraules quina informació ens proporciona aquesta mètrica.

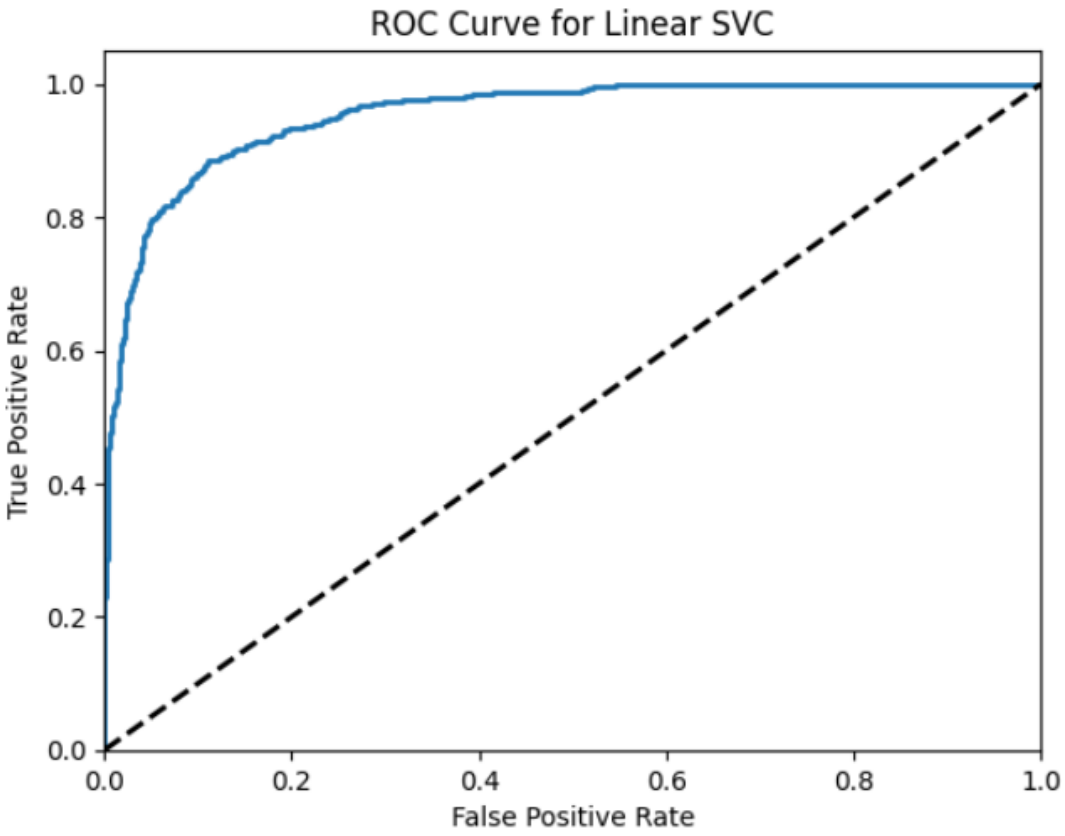
```
ROC AUC Score for Logistic Regression: 0.9498091311285869
```



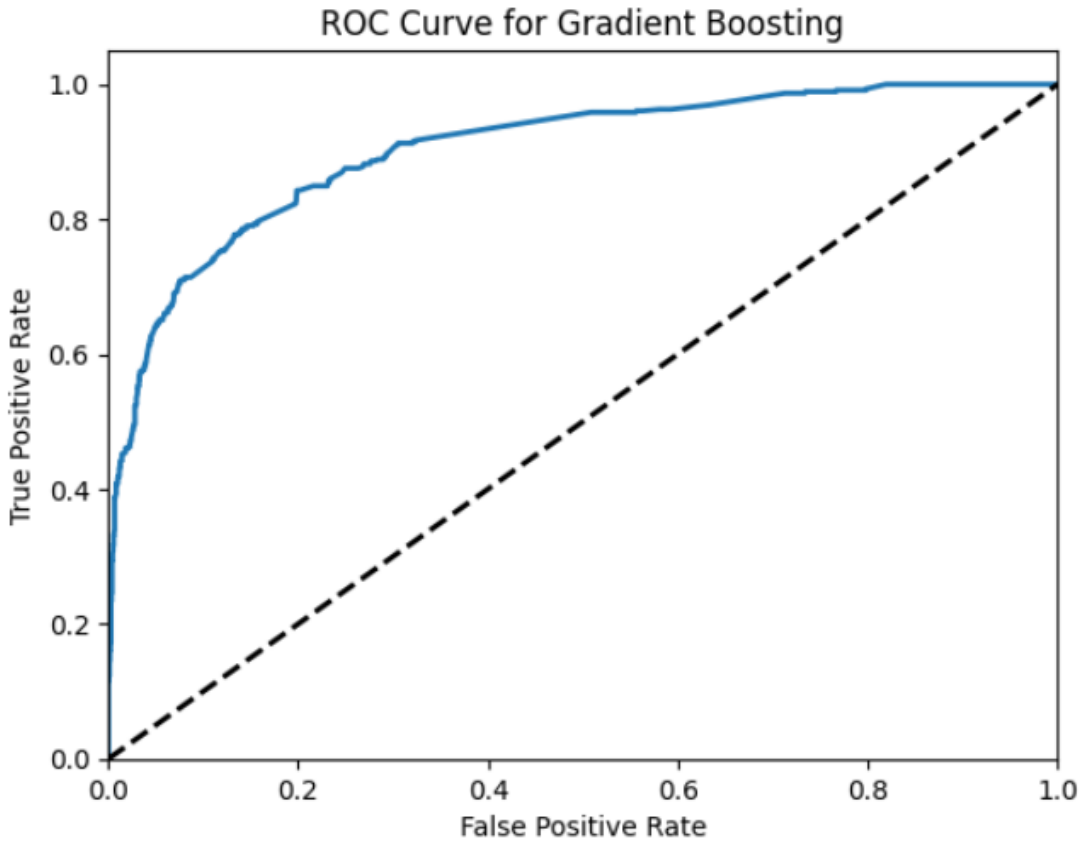
```
ROC AUC Score for Random Forest: 0.9242377189378232
```



```
ROC AUC Score for Linear SVC: 0.9539000315327797
```



```
ROC AUC Score for Gradient Boosting: 0.9031017983239851
```



La corba ROC és una representació gràfica sobre el rendiment d'un model, així la gràfica tenim a l'eix X, la taxa de Fals Positius, i a l'eix Y, la taxa de positius vertaders. A meitat tenim una línia que ens indica un AUC de 0.5. Açò significa que per damunt d'aquesta línia el model té capacitat de discriminar entre les classes positives i negatives (i que no és a l'atzar), com més proper és a 1, millor és el model. Per sota d'aquesta línia, ens indica que el model està invertint les classes i que és pitjor que fer un model que classificarà a l'atzar.

La millor corba ROC és lògicament la del model que millor Accuracy i Recall té, que és el model amb Linear SVC. Així, amb una taxa de Fals positius menor a 0.1, el model ja identifica correctament més del 80% dels positius vertaders.

iv) Tria un model i implementa un Random Search (entre 2 i 5 paràmetres per cada array del grid) per trobar una combinació millor respecte al

default dels hiperparàmetres. Realitza la matriu de confusió. Quin % de millora s'ha guanyat d'accuracy i recall si ho comparem amb la versió anterior?

El millor model té aquests hiperparàmetres: 'tol': 0.1, 'max_iter': 1000, 'C': 1.

	precision	recall	f1-score	support
negative	0.94	0.96	0.95	1828
positive	0.83	0.74	0.78	458
accuracy			0.92	2286
macro avg	0.88	0.85	0.87	2286
weighted avg	0.92	0.92	0.92	2286

En aquest cas no hem guanyat cap millora en l'accuracy ni el recall, solament hem millorat un 0.01, la precisió. Per tant el percentatge de millora en l'accuracy i el recall ha sigut d'un 0%. Per la qual cosa el model per defecte ja estava prou prop de la millor configuració.

b) En les dades tenim una feature que ha estat computada amb l'objectiu de saber el perquè del sentiment negatiu.
   i) Partint del dataset de training, crea un de nou amb només els tuits negatius i amb el dataset de test realitza el mateix per després poder comprovar el bé que funciona el model.
   ii) Realitza un petit estudi dels motius negatius. Estan balancejades les classes? En cas negatiu, comenta (no ho implementis) que faries per balancejar el dataset.

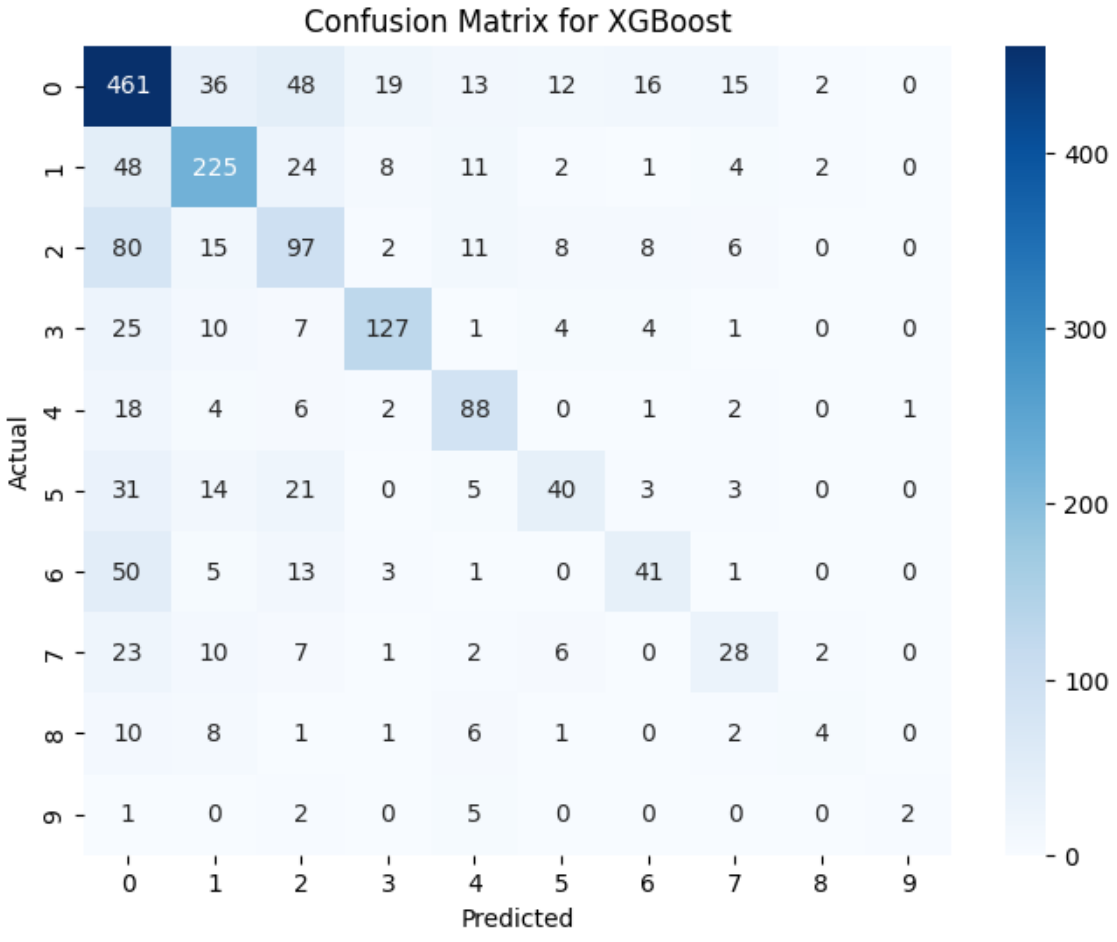
Veiem com el dataset no està balancejat, ja que hi ha raons amb més de 800, 1000 i 2000 casos, mentre que altres tenen menys de 150 casos. El problema d'emprar datasets desbalancejats és que el model pot ignorar les classes minoritàries, ja que s'obtenen millors resultats (millor Accuracy) si no prediem cap classe minoritària.

La solució a aquest problema és balancejar les dades, per fer-ho podem fer un sobremostreig de les classes minoritàries, o un submostreig de les classes majoritàries.

iii) Entrena un XGBoost (amb el valor 126 per reproduir resultats) que intenti predir el motiu d'un tuit negatiu. Implementa també un Random Search (entre 2 i 5 paràmetres per cada array del grid) per personalitzar els hiperparàmetres del model.

La millor combinació de paràmetres és: 'subsample': 0.7, 'n_estimators': 500, 'max_depth': 3, 'learning_rate': 0.05.

iv) Quin encert té el model en funció de cada motiu, és a dir, quin percentatge d'encert té respecte a cadascun? (ex: el model encerta un 60% del tipus “Customer Service Issue”...)



	precision	recall	f1-score	support
Bad Flight	0.55	0.34	0.42	117
Can't Tell	0.43	0.43	0.43	227
Cancelled Flight	0.78	0.71	0.74	179
Customer Service Issue	0.62	0.74	0.67	622
Damaged Luggage	0.67	0.20	0.31	10
Flight Attendant Complaints	0.45	0.35	0.40	79
Flight Booking Problems	0.55	0.36	0.44	114
Late Flight	0.69	0.69	0.69	325
Lost Luggage	0.62	0.72	0.66	122
longlines	0.40	0.12	0.19	33
accuracy			0.61	1828
macro avg	0.57	0.47	0.49	1828
weighted avg	0.60	0.61	0.60	1828

Veiem com l'accuracy general és del 61%, si bé sembla un poc baixa, com tenim 10 categories diferents, no està malament (si ho férem a l'atzar encertaríem al voltant d'un 10%).

Respecte a la precisió per categories ordenades per freqüència:

Customer Service Issue: 62%
Late Flight: 69%
Can't Tell: 43%
Cancelled Flight: 78%
Lost Luggage: 62%
Bad Flight: 55%
Flight Booking Problems: 55%
Flight Attendant Complaints: 45%
Longline: 40%
Damaged Luggage: 67%

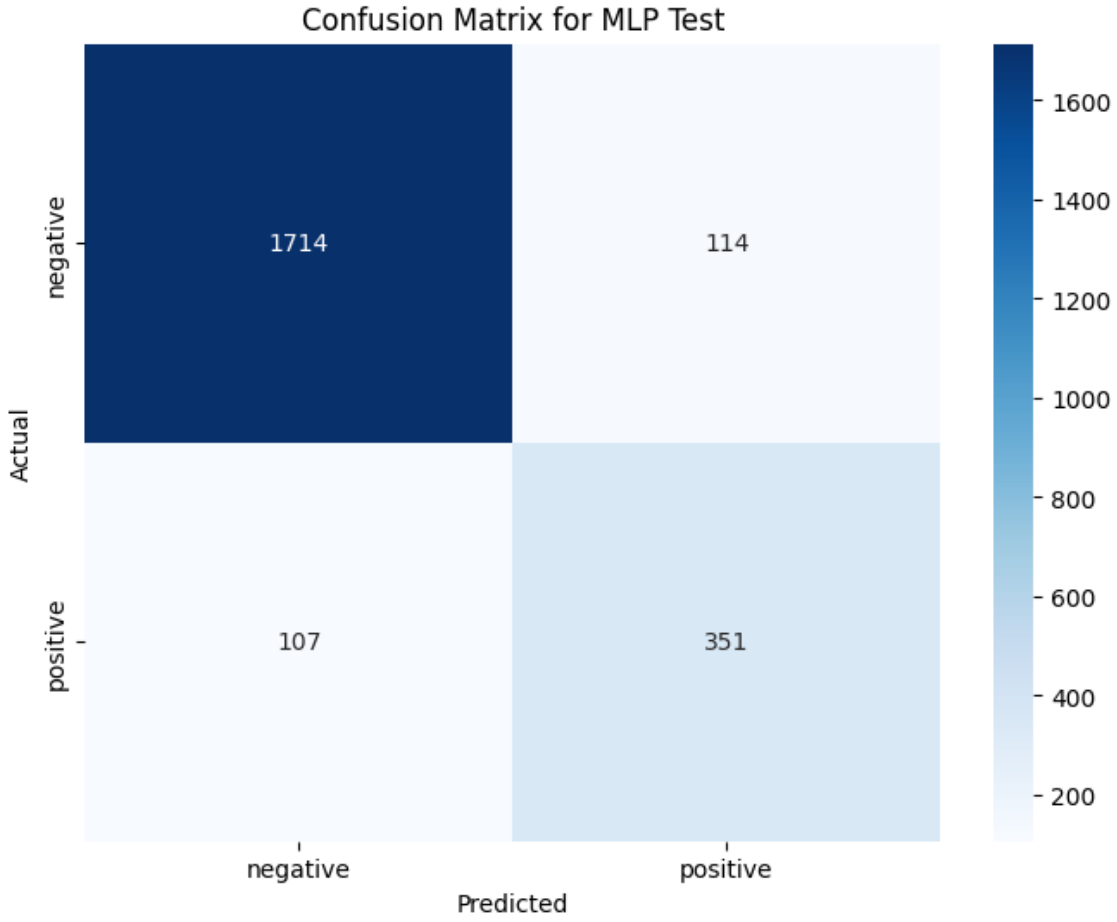
Com veiem de forma general les categories més freqüents solen tenir millors resultats (exceptuant Can't Tell, el qual és una categoria particular) que les categories menys freqüents.

Amb els fitxers del segon cas d'ús, realitza:

c) Amb la llibreria scikit-learn, entrena una xarxa neuronal MLP que ens permeti classificar els tweets en positius i negatius amb un layer de 128 (potser, en funció del hardware has de reduir la dimensió o el conjunt de dades d'entrenament). Les èpoques i els altres hiperparàmetres necessaris els pots revisar i escollir llegint la documentació de la llibreria. Quin encert té el model? Computa les matrius de confusió sobre el dataset de train i el de test.

Classification Report for MLP Train:

	precision	recall	f1-score	support
negative	1.00	1.00	1.00	6423
positive	1.00	1.00	1.00	1655
accuracy			1.00	8078
macro avg	1.00	1.00	1.00	8078
weighted avg	1.00	1.00	1.00	8078



```
Classification Report for MLP Test:

              precision    recall  f1-score   support

    negative       0.94      0.94      0.94      1828
    positive       0.75      0.77      0.76       458

    accuracy                           0.90      2286
   macro avg       0.85      0.85      0.85      2286
weighted avg       0.90      0.90      0.90      2286
```

Veiem com en el conjunt d'entrenament té una precisió de quasi el 100%, la qual cosa ens podria dir que el model està sobreentrenat. Però veiem el resultat del conjunt de test té una precisió del 90%, la qual cosa és un molt bon resultat i podem dir que el model està una mica sobreentrenat, però de totes formes, el model funciona prou bé.

d) A partir dels models entrenats en les activitats 3a i 3b, on tenim un model que ens serveix per analitzar el sentiment dels tuits i un altre que ens ajuda a saber el perquè d'un tuit negatiu, crea una funció que a partir d'un tuit, ens digui si és positiu o negatiu i en cas de negatiu que ens intenti comentar el perquè.
Es recomana passar com a paràmetres els models a la funció o crear variables globals.

```
tweet_classifactor("The Customer Service person was awful, he didn't help us")
```

```
('negative', 'Customer Service Issue')
```

```
[193] tweet_classifactor("The flight with @Delta was amazing, thanks")
```

```
'positive'
```

```
[194] tweet_classifactor("We arrived very late to our destination")
```

```
('negative', 'Late Flight')
```

[191]

	text	prediction	airline_sentiment	negativereason
3099	@united Another unfortunate case of bad luck, usually maintenance issue. They are now swapping planes to ONT, will get in Late Flight. :(	(negative, Late Flight)	negative	Late Flight
8958	@JetBlue A flight delay due to pilots oversleeping is apparently an uncontrollable irregularity that is not eligible for delay compensation.	(negative, Late Flight)	negative	Late Flight
13237	@AmericanAir delayed flight six hours. Missed international connection= extra night in a hotel and still have to pay for a 15lb surfboardbag	(negative, Late Flight)	negative	Late Flight
5813	@SouthwestAir is seriously THE WORST. I don't remember the last time I or someone I knew had a flight that wasn't delayed / Cancelled Flightled / etcl	(negative, Cancelled Flight)	negative	Cancelled Flight
3715	@united Will have to try standby in Denver tonight or will have to Cancelled Flight father son trip till next year. Thx for trying.	positive	positive	NaN
12691	@AmericanAir "sorry you were disappointed" #outoftouchwithreality #people have kids and jobs	(negative, Customer Service Issue)	negative	Cancelled Flight
835	@united Thanks for the reminder. It's been a fun ride. http://t.co/pPVA4Rch9f	positive	positive	NaN
2531	@united will the kudos to capt. Herman be relayed to his chief pilot or should I be emailing someone?	positive	positive	NaN
13925	@AmericanAir The delay is nothing but the personnel being so combative up to the point of saying "what's the hury, the plane is not leaving	(negative, Late Flight)	negative	Late Flight
3465	@united - watched the entire #UNCvsDUKE game on the Tarmac before Cancelled Fighting my flight because crew timed out, right before my 4hr flight....	(negative, Cancelled Flight)	negative	Cancelled Flight
7922	@JetBlue thank you so much for your effort	positive	positive	NaN
9191	@USAirways is the worst. I have to pay $200 just to NOT take my first flight in a round trip! Otherwise they'll Cancelled Flight my whole flight.	(negative, Can't Tell)	negative	Can't Tell
1305	@united never ever again will I be Flight Booking Problems a flight with United or any affiliate if there is a chance to get on a United flight	(negative, Flight Booking Problems)	negative	Can't Tell
3966	@united tried to check it 40 mins before a flight rather than 45 mins and we are stonewalled by your employees...	(negative, Flight Attendant Complaints)	negative	Flight Attendant Complaints
8276	@JetBlue @TSA JetBlue never disappoints !	positive	positive	NaN
10042	@USAirways quit being cheap and hire some call center employees. It shouldn't take half a day to call you	(negative, Customer Service Issue)	negative	Customer Service Issue

Veiem com la nostra funció funciona correctament, i sembla prou precís, sobretot respecte a la classificació de positiu i negatiu.

Activitat 4: Data Governance (10%)

a) Què és el govern de les dades? Quins són els seus pilars? Quin és el seu objectiu?

El govern de les dades són les polítiques, procediments i normes que s'implementen per a garantir que les dades d'una organització siguen precisos, és a dir, que menegen de forma correcta quan s'ingressen, emmagatzemen, menegen, accedeixen i eliminen.

Els pilars als quals es basa el govern de dades és:

- Disponibilitat, que és la capacitat de garantir que les dades siguen accessibles en el moment i lloc necessaris per als usuaris autoritzats.
- Usabilitat, que implica que les dades puguen ser fàcilment comprensibles i utilitzables pels usuaris finals.
- Consistència, tracta de la uniformitat de les dades a través de l'empresa.
- Integritat de les dades, és mantenir la precisió i coherència de les dades al llarg del seu cicle de vida.
- Seguretat de les dades, és la protecció de les dades contra accessos no autoritzats, robatoris o pèrdues.
- Compliment de normatives, els diferents països han desenvolupat lleis i regulacions sobre l'ús de dades, les quals l'empresa ha de complir.

L'objectiu principal del govern de les dades és garantir que les dades s'utilitzen correctament, tant a l'àmbit ètic com regulador. Aquest govern de les dades tracta tant d'evitar l'entrada d'errades en els sistemes, com bloquejar el possible ús indegut de dades personals sobre clients i altres informacions confidencials.

https://www.sap.com/latinamerica/products/technology-platform/master-data-governance/what-is-data-governance.html
https://en.wikipedia.org/wiki/Data_governance
https://www.iebschool.com/blog/data-governance-big-data/

b) Tria un cas d'ús del recurs *Data management and usi: casi studies of technologies and governance* i explica com han abordat la governança de les dades. Explica les raons/motius pels quals han decidit actuar de tal manera. Hauries pres tu alguna decisió diferent? Per què?

He triat el cas de govern de les dades en el cas de localització personal, on ens exposa el cas d'aplicacions que utilitzen i recopilen dades de localització que després són emprats per a diferents fins com: la planificació urbana, serveis personalitzats o anàlisi de comportament.

En aquest cas, s'ha abordat la governança de dades, pel fet d'assegurar la privacitat i la seguretat dels usuaris, tal com indica la normativa del GPDR. Aquesta normativa considera les dades de localització com a dades personals, ja que tenen una alta re-identificabilitat, com poden ser les coordenades, adreces IP i adreces MAC/IMEI.

Per altra banda, també es centren problemes sobre la transparència i el consentiment per l'ús de les dades dels usuaris a aquestes aplicacions. Així, moltes vegades, encara que els usuaris aproven un consentiment per a què l'aplicació utilitze les seues dades, no queda clar alguns aspectes, com quan s'estan recopilant les dades o que es capturen dades de localització sense que els usuaris ho sàpiguen.

Aleshores, les raons per les quals s'han pres decisions sobre el govern de dades de localització es prenen, en primer lloc, per complir la normativa referent a la normativa europea GPDR. També es prenen per protegir la privacitat i evitar que es puguen revelar informació delicada sobre comportaments dels usuaris, i per

tant que els consumidors no es senten “espiats” i siguen emprant i mantenint la confiança en les aplicacions que empre dades de localització.

c) Avui dia és necessari que les persones confiïn en l'ús de la Intel·ligència Artificial i la governança de les dades ens ajuda, però no és suficient. Aquí, per exemple, entra l’Explainability (tal hi com vam veure en el repte 3), la Robustesa o el “Fairness” (la idea és que el model de IA sigui just i imparcial). Explica els motius pels quals es necessita que els models siguin robustos i imparcials amb exemples específics per cada concepte. En el cas de robustesa, explica un exemple d'un possible atac “maliciós” a un model.

   En primer lloc, necessitem que els models siguen robustos per a què aquest puguen mantenir el seu rendiment, encara que hi haja condicions adverses com poden ser dades, sorolloses, incompletes o manipulades. Així és important que els models siguen robustos per a proporcionar seguretat als models (que puguen resistir atacs maliciosos mantenint la seua integritat) i fiabilitat als models (que encara que hi haja variacions en les dades d’entrada els models continuen sent consistents i precisos).

   En segon lloc, tenim la imparcialitat, el qual tracta que els models no discriminen a diferents grups de persones pels seus trets identitaris, com poden ser la raça, l’edat, el gènere... Els motius, perquè els models siguen imparcials, són morals (que tothom tinga un tracte just i equitatiu) i legals (La majoria dels països democràtics tenen lleis contra la discriminació que les empreses han de complir).

   Un exemple d’atac maliciós que pot afectar a un model com el que hem creat seria un atac de dades enverinades. Així, l’atacant inclou en el dataset d’entrenament tweets modificats de forma maliciosa, per a manipular a la classificació del model. Així els atacants identifiquen quines són les paraules claus que fan que el model puga classificar a un tweet com a positiu o negatiu i les barregen per a confondre al model.

   Per exemple, amb dades normals tindríem com a tweet positiu “El vol va ser puntual i el servei excel·lent”, i com a tweet negatiu “El vol va tenir un retard de 3 hores, molt mal servei”. En un model maliciós inclouríem tweets etiquetats com a positius com “El vol va ser puntual, però l’aterratge va ser perillós” o negatius com “El vol va tenir un retard, però el personal era amable”.

Açò fa que el model aprén associacions incorrectes, com classificar com a positiu qualsevol tweet que continga la paraula “retard”.

La forma de prevenir aquests tipus d’atacs, seria emprar dades verificades per humans o utilitzar tècniques que detecten anomalies i incoherències en els models.

https://www.cobalt.io/blog/data-poisoning-attacks-a-new-attack-vector-within-ai
https://en.wikipedia.org/wiki/Fairness_(machine_learning)
https://oa.upm.es/75829/1/TFM_CARLOS_SANCHEZ_VELAZQUEZ.pdf

BIBLIOGRAFIA

https://en.wikipedia.org/wiki/February_2015_North_American_cold_wave
https://en.wikipedia.org/wiki/February_14%E2%80%9315,_2015_North_American_blizzard
https://eu.usatoday.com/story/todayinthesky/2015/02/21/cancellations-near-1000-from-latest-air-travel-weather-woes/23797113/
https://www.sap.com/latinamerica/products/technology-platform/master-data-governance/what-is-data-governance.html
https://en.wikipedia.org/wiki/Data_governance
https://www.iebschool.com/blog/data-governance-big-data/
https://www.cobalt.io/blog/data-poisoning-attacks-a-new-attack-vector-within-ai
https://en.wikipedia.org/wiki/Fairness_(machine_learning)
https://oa.upm.es/75829/1/TFM_CARLOS_SANCHEZ_VELAZQUEZ.pdf
https://vitalflux.com/bert-vs-gpt-differences-real-life-examples/