

TREBALL FINAL DE GRAU



ESCOLA
POLITÈCNICA SUPERIOR
UNIVERSITAT DE LLEIDA
INSPIRING THE FUTURE

Estudiant: Víctor Fernández Florensa

Titulació: Grau en Enginyeria Informàtica

Títol de Treball Final de Grau: **Processament de Llenguatge Natural aplicat a ressenyes de videojocs.**

Director/a: **Nil Torrent Bureu i Jordi Planes Cid**

Presentació

Mes: Juny

Any: 2024

Índex

Acrònims	6
Glossari	6
1 Agraïments	7
2 Introducció	8
2.1 Motivació	8
2.2 Objectius	8
2.3 Metodologia	8
2.4 Programes i llibreries utilitzades	10
3 Conceptes teòrics	12
3.1 Introducció al Processament de Llenguatge Natural	12
3.1.1 Definició	12
3.1.2 Objectius	12
3.1.3 Com funciona el NLP	12
3.1.3.a Preprocessament de les dades	13
3.1.3.b Desenvolupament d'algorithmes	15
3.1.4 Àrees de treball	15
3.1.5 Reptes del llenguatge	17
3.1.6 Importància en la indústria i aplicacions	18
3.2 Algorismes per al Processament del Llenguatge Natural	19
3.2.1 Mètodes bàsics	19
3.2.1.a Bag-of-words	19
3.2.1.b TF-IDF	19
3.2.1.c Word embeddings	20
3.2.1.d Word2Vec	23
3.2.2 Mètodes avançats	25
3.2.2.a Xarxes Neuronals	25
3.2.2.b Transformadors	27
3.2.2.c BERT	29
3.2.2.d RoBERTa	29
3.2.3 Reptes tècnics del NLP	30
3.3 Anàlisi de sentiments	31
3.3.1 Definició	31
3.3.2 Tipus	31
3.3.3 Beneficis	31
3.3.4 Aplicacions	32
4 Context del problema	33
4.1 The Last of Us	33
4.1.1 Història i llançament del videojoc	33
4.1.2 Sobre Naughty Dog i context de desenvolupament	33
4.1.3 Relevància en la indústria dels videojocs	34
4.1.4 Referències i inspiració d'altres treballs en la narrativa	35
4.1.5 Recepció del mercat	37
4.1.6 The Last of Us: Left Behind	37
4.1.6.a Història i llançament	37
4.1.6.b Recepció del mercat	38
4.2 The Last of Us: Part II	39
4.2.1 Història i llançament del videojoc	39
4.2.2 Sobre Naughty Dog i context de desenvolupament	39
4.2.3 Rellevància en la indústria dels videojocs	40
4.2.4 Referències i inspiració d'altres treballs en la narrativa	41
4.2.5 Recepció del mercat	42

5 Estudi preliminar del conjunt de dades	43
5.1 Informació sobre el dataset	43
5.2 Hipòtesis inicials	44
5.3 Preprocessament i visualització de les dades	46
5.3.1 The Last of Us (2013) i The Last of Us Remaster (2014)	46
5.3.1.a Ressenyes dels crítics	46
5.3.1.b Ressenyes dels usuaris	49
5.3.2 The Last of Us: Part 2 (2020)	54
5.3.2.a Ressenyes dels crítics	54
5.3.2.b Ressenyes dels usuaris	55
6 Anàlisi de sentiments	61
6.1 Etiquetatge multilingüe amb xlm-RoBERTa-base	61
6.2 Etiquetatge amb diversos models específics	64
6.3 Model de freqüència de paraules	68
7 Predicció de la puntuació dels usuaris a partir dels comentaris	71
7.1 Entrenament del model de regressió	71
7.2 Descripció de les mètriques d'avaluació utilitzades	71
7.3 Visualització de l'ajust del model	73
7.4 Resultats obtinguts	74
8 Conclusions	76
9 Possibles ampliacions de l'estudi	78
Annex	79

Índex de figures

1	Diagrama de processos que mostra la relació entre les diferents fases de CRISP-DM. Autor: Kenneth Jensen. Font: Wikipedia	9
2	Diagrama del cicle de vida de les dades i les eines utilitzades	10
3	Processament de dades amb transformers	11
4	El NLP és una branca de la Intel·ligència Artificial que pot utilitzar tècniques de Machine Learning (ML) i Deep Learning (DL) per millorar el model. Autor: Balakrishnan Sathiyakugan. Font: [1]	12
5	Il·lustració de com el processament del llenguatge natural converteix text no estructurat en dades estructurades, les quals poden ser analitzades posteriorment per algoritmes de ML en l'àmbit de la medicina. Autor: Avishek Choudhury. Font: [2]	13
6	Exemple de tokenització d'una frase. Autor: Maleesha De Silva. Font: [3]	14
7	Exemple de la tècnica de stopword removal. Autor: Maleesha De Silva. Font: [3]	14
8	Exemple de preprocessament de dades i correlació de patrons. Autor: Mohd Sanad Zaki-rizvi. Font [4]	14
9	Exemples de xarxes neuronals. Font: Neural Network course, University of Toronto. Font: [5]	15
10	El NLU és un subconjunt de NLP. Autor: MacCartney 2014, slide 8 Font: [6]	16
11	Camps del NLP segons la tasca que desenvolupen. Autor: Fabio Chiusano. Font: [7]	17
12	Exemple com la mateixa frase en un context o en un altre canvia completament el significat. Autor: John Teleska. Font: [8]	17
13	Exemple que mostra la tècnica de la bossa de paraules. Per cada text es calcula la freqüència d'aparició de cada paraula i es crea una matriu. Autor: Alexis Perrier. Font: [9]	19
14	Exemple que mostra la tècnica de bossa de paraules amb la tècnica de vectorització. Autor: Alexis Perrier. Font: [9]	19
15	Quan s'aplica TF-IDF a un conjunt de documents, s'obté una representació vectorial de cada document, on cada dimensió del vector correspon a una paraula, i el valor de cada dimensió és el seu pes TF-IDF. Autor: Turing. Font: [10]	20
16	Les paraules similars tenen representació vectorial similar. Autor: Fabio Chiusano. Font: [11]	21
17	Diagrama en forma d'arbre que mostra els diferents mètodes i la seva classificació en dos grans grups: les de context independent i les de context dependent. Autor: Fabio Chiusano. Font: [11]	22
18	Exemple de com la mateixa paraula adopta la mateixa representació en contextos diferents. Autor: Fabio Chiusano. Font: [11]	22
19	Exemple de com la mateixa paraula adopta diferents representacions en contextos diferents. Autor: Fabio Chiusano. Font: [11]	23
20	En aquest exemple, s'intenta predir la paraula "listen", partint d'una mida de finestra de 5 paraules circumdants, que capturen el context que envolta la paraula. Autor: Turing. Font: [10]	23
21	Diagrama de les dues variants de Word2Vec. Autor: "Efficient Estimation of Word Representations in Vector Space", 2013. Font: [12]	24
22	Diagrama d'una Xarxa Neuronal. Autor: -. Font: [13]	25
23	Diagrama d'una neurona artificial. Autor: Neural Network course, University of Toronto. Font: [5]	25
24	Exemple d'una xarxa neuronal. Es pot observar una capa d'entrada amb dues neurones, una capa oculta amb dues neurones més i una capa de sortida amb una neurona. (ex: problema de classificació binària) Autor: Neural Network course, University of Toronto. Font: [5]	26
25	Fòrmules de la propagació cap endavant (esquerra) i retropropagació (dreta). Autor: Frederick kratzert's blog. Font: [14]	26
26	Exemples de funcions d'activació. Autor: -. Font: [5]	27
27	Els Transformadors, de vegades anomenats models fonamentals, són la tecnologia subjacent de moltes aplicacions recents. Autor: Rick Merritt. Font: [15]	28
28	Evolució de l'Aprenentatge Automàtic. Autor: Rick Merritt. Font: [15]	28
29	RoBERTa està entrenat en un conjunt de dades més gran i té més paràmetres que BERT, la qual cosa el fa més potent i flexible. Autor: Vyacheslav Efimov. Font: [16]	29

30	Exemple de classificació de tres comentaris segons el seu sentiment. Autor: MonkeyLearn. Font: [17]	31
31	Portada The Last of Us PS3. Imatge promocional. Autors: Naughty Dog	33
32	Concept art dels protagonistes. Authors: Naughty Dog. Font: The Last of Us Wiki [18] . .	35
33	De desconeuguts a família: el viatge de Joel i Ellie. Autors: Naughty Dog. Font: Metratge del joc.	35
34	Entorns i caracterització del videojoc - 1. Concept Art. Autors: Naughty Dog. Font: [19]	36
35	Entorns i caracterització del videojoc - 2. Concept Art. Autors: Naughty Dog. Font: [19]	36
36	Portada contingut ampliable The Last of Us: Left Behind. Imatge promocional. Autors: Naughty Dog.	37
37	Portada The Last of Us: Part II. Imatge promocional. Autors: Naughty Dog.	39
38	Les cinemàtiques del videojoc s'apropen tènicament al fotorealisme. Autors: Naughty Dog. Font: Metratge del joc	40
39	El joc mostra diversos punts de vista del conflicte. Autors: Naughty Dog. Font: Metratge del joc.	41
40	Reconstrucció de la ciutat de Seattle en un entorn postapocalíptic. Autors: Naughty Dog. Font: Metratge del joc	42
41	Gràfic de barres de ressenyes especialitzades per plataforma	46
42	Gràfic de línies de les ressenyes especialitzades en el temps PS3	46
43	Gràfic de línies de les ressenyes especialitzades acumulades en el temps PS3	47
44	Gràfic de línies de les ressenyes especialitzades en el temps PS4 Remaster	47
45	Gràfic de línies de les ressenyes especialitzades acumulades en el temps PS4 Remaster . .	48
46	Gràfic de capsa de les puntuacions especialitzades rebudes d'ambdós videojocs	48
47	Gràfic de barres de ressenyes d'usuaris per plataforma	49
48	Gràfic de línies de les ressenyes d'usuaris en el temps PS3	49
49	Gràfic de línies de les ressenyes d'usuaris acumulades en el temps PS3	50
50	Gràfic de línies de les ressenyes d'usuaris en el temps PS4	50
51	Gràfic de línies de les ressenyes d'usuaris en el temps PS4	51
52	Gràfic de barres agrupades de la puntuació dels usuaris normalitzada per plataforma . .	51
53	Gràfic de barres de la puntuació dels usuaris agrupada per idioma i normalitzada en el seu grup	52
54	Gràfic de dispersió de <i>views/votes</i> de les ressenyes d'usuaris, acolorit per <i>score</i>	52
55	Regressió lineal per als dos subgrups en l'escala <i>vots/visualitzacions</i>	53
56	Gràfic de línies de les ressenyes especialitzades en el temps PS4	54
57	Gràfic de línies de les ressenyes especialitzades acumulades en el temps PS4	54
58	Gràfic de capsa de les puntuacions especialitzades rebudes de PS4	55
59	Gràfic de línies de les ressenyes d'usuaris en el temps PS4	55
60	Gràfic de línies de les ressenyes d'usuaris acumulades en el temps PS4	56
61	Gràfic de barres de la puntuació dels usuaris	56
62	Gràfic de línies de les ressenyes d'usuaris acumulades en el temps PS4	57
63	Gràfic de línies de les ressenyes d'usuaris acumulades en el temps PS4	57
64	Gràfic de la representació dels idiomes en les ressenyes.	58
65	Gràfic de barres de la puntuació dels usuaris agrupada per idioma i normalitzada segons el grup	58
66	Gràfic de dispersió de <i>views/votes</i> de les ressenyes d'usuaris, acolorit per <i>score</i>	59
67	Regressió lineal per als dos subgrups en l'escala <i>vots/visualitzacions</i>	60
68	Polaritat de les ressenyes segons l'anàlisi de sentiments.	62
69	Puntuació dels comentaris classificats com positius.	62
70	Puntuació dels comentaris classificats com negatius.	63
71	Puntuació dels comentaris classificats com neutres.	63
72	Polaritat de les ressenyes segons l'anàlisi de sentiments amb models específics.	65
73	Gràfic de línies del sentiment de les ressenyes al llarg del temps	65
74	Gràfic de línies (ampliat) del sentiment de les ressenyes al llarg del temps	66
75	Puntuació dels comentaris classificats com positius.	66
76	Puntuació dels comentaris classificats com negatius.	67
77	Puntuació dels comentaris classificats com neutres.	67
78	Gràfic de freqüència de les paraules més repetides de les ressenyes.	69

79	Mosaic de les paraules més freqüents, segmentat segons la nota que els jugadors van donar al videojoc. (Word Cloud)	70
80	Gràfic de valors observats vs predictius. Regressió Lineal.	73
81	Diagrama de violí de valors observats vs predictius. Regressió Lineal.	73
82	Gràfic de Residus Estandarditzats vs Valors Predictius. Regressió Lineal.	74
83	Per entendre en millor els càlculs de la propagació endavant d'una Xarxa Neuronal es va fer aquest exercici durant el període de documentació.	79

Acrònims

- CRISP-DM** Cross Industry Standard Process for Data Mining. 3, 8, 9
- NLP** Natural Language Processing. 3, 12, 15–18, 21, 28–31, 76
- ML** Machine Learning. 3, 11–13, 15, 18, 19, 31
- DL** Deep Learning. 3, 11, 12, 15, 18, 21, 25
- NLU** Natural Language Understanding. 3, 15, 16
- TF-IDF** Term Frequency-Inverse Document Frequency. 3, 19, 20
- GOTY** Game Of The Year. 8
- TFG** Treball de Final de Grau. 9
- IA** Intel·ligència Artificial. 12
- NLG** Natural Language Generation. 15
- CBOW** Continuous Bag-of-Words. 23
- MMLU** Massive Multitask Language Understanding. 64
- MAE** Mean Absolute Error. 71
- MSE** Mean Squared Error. 71
- RMSE** Root Mean Square Error. 71

Glossari

Intel·ligència Artificial La Intel·ligència Artificial és un camp de la informàtica que se centra en crear algorismes i models estadístics que puguin realitzar tasques històricament relacionades amb la intel·ligència humana, com l'aprenentatge, el raonament i la percepció. 3, 12

Machine Learning El Machine Learning o aprenentatge automàtic és una disciplina del camp de la intel·ligència artificial que, a través d'algorismes, dota els ordinadors de la capacitat d'identificar patrons en dades massives i elaborar prediccions (anàlisi predictiu). 3, 12

Deep Learning El deep learning o aprenentatge profund és una subdisciplina del Machine Learning que utilitza xarxes neuronals profundes, és a dir, xarxes amb una o més capes ocultes, per analitzar grans quantitats de dades i extreure característiques complexes de manera automàtica. 3, 12

1 Agraïments

En primer lloc, m'agradaria donar les gràcies al Nil Torrent per co-tutoritzar i guiar aquest treball de fi de grau juntament amb el Jordi Planes. La seva dedicació docent i estima per l'ensenyament han sigut de gran inspiració en la meva formació i una de les raons per voler aprofundir més en el camp de la ciència de dades i la intel·ligència artificial. En l'àmbit personal, no tinc més que bones paraules per la seva predisposició a ajudar-me i pel tracte que he rebut, que ha estat essencial per a la realització del treball.

També m'agradaria donar les gràcies a l'Escola Politècnica Superior de Lleida per oferir una educació pública, de qualitat i accessible per tots els que ens volem dedicar al sector de la informàtica de forma professional.

En tercer lloc, vull agrair el suport de la meva família, a qui els ho dec tot. Sempre m'han dit que la millor herència que uns pares poden donar és una bona educació. El seu ajut al llarg de la carrera, tant en els bons com en els mals moments, ha estat un pilar imprescindible en qui soc i on he arribat.

Per acabar, el meu agraïment als amics que he fet durant el grau: el Pau Llobet, el David Castro, la Chaymaa Dkouk, l'Elena Barrachina, l'Aleix Drudis, l'Eric Mesa, el Catalin Doja, el Víctor Mateu, el Zihan Chen i el David Carreras. Amb tots ells he compartit infinitat d'hores a la universitat, ens hem fet costat mútuament i és el que m'emporto d'aquesta experiència. Espero que la nostra amistat perduri durant molt de temps.

2 Introducció

2.1 Motivació

La idea d'aquest treball sorgeix a principis d'any quan vaig contactar amb el Nil Torrent, professor adjunt de l'assignatura de Programació Avançada en d'Inteligència Artificial, per proposar-li fer un TFG que fos una extensió de l'assignatura que impartia. Quan vaig obtenir la confirmació de què me'l podia portar, vaig quedar amb Jordi Planes, professor agregat del departament d'Enginyeria Informàtica i Disseny Digital, i el Nil per presentar-los diverses opcions que pensava que eren viables per investigar.

Un dels consells que em van donar va ser que escollís un tema que realment m'apassionés, perquè li hauria de dedicar moltes hores. Després d'uns dies de reflexió, ho vaig tenir clar: volia fer un treball que d'alguna forma estigués relacionat amb els videojocs.

Vaig pensar que la millor forma de fer-ho era amb un treball de Processament de Llenguatge Natural que analitzés les ressenyes de la meva saga de videojocs preferida, "The Last of Us". A més, vaig creure que el conjunt de dades a analitzar podia ser interessant a causa de la gran quantitat de missatges polaritzants entorn de la preqüela, The Last of Us: Part II, que va obtenir el reconeixement de Game Of The Year (GOTY) 2020 amb una nota mitjana dels usuaris d'un 5'8 a Metacritic, que contrastava molt amb el 9'3 de l'opinió crítica de medis especialitzats. A més, ja existia un estudi previ a la plataforma de Kaggle sobre el mateix tema que podia consultar per veure que s'havia fet i com es podia millorar. Un cop definits els objectius del projecte, vaig presentar la proposta als dos tutors i va ser acceptada.

2.2 Objectius

La tasca principal d'aquest projecte és fer un estudi aplicant diversos models de Processament de Llenguatge Natural per analitzar un conjunt de dades reals de les ressenyes de diversos videojocs de la mateixa franquícia.

En la part teòrica, es farà una introducció al processament de llenguatge natural i els algorismes o tècniques de processament més coneguts. També s'explicarà concretament l'anàlisi de sentiments i els models que millor s'adapten per resoldre el problema. A més, es farà una introducció de la saga de videojocs "The Last of Us" i la seva preqüela per contextualitzar el problema, fent un resum de la informació dels videojocs per una millor comprensió de les dades.

En la part pràctica, es farà una anàlisi de dades dels dos videojocs de la saga, tant de les opinions crítiques com d'usuaris. A continuació es farà una anàlisi de sentiments amb diversos models preentrenats i es compararan els resultats obtinguts. També s'utilitzaran dos models de traducció i moltes de les tècniques de preprocessament de dades per fer un recompte de la freqüència de les paraules i mostrar-les segons la seva polaritat.

Per últim, s'entrenarà un model per predir la nota que un usuari pot donar al videojoc a partir del seu comentari, a partir de les puntuacions de l'anàlisi de sentiments.

2.3 Metodologia

Per l'elaboració d'aquest treball s'ha seguit de forma adaptada la metodologia Cross Industry Standard Process for Data Mining (CRISP-DM), vista en l'assignatura de Programació Avançada en Intel·ligència Artificial. Aquesta metodologia és molt popular en les empreses, ja que proporciona una aproximació a la metodologia tradicional de gestió de projectes en l'empresa en el context de la ciència de dades. CRISP-DM defineix un projecte com una seqüència de fases:

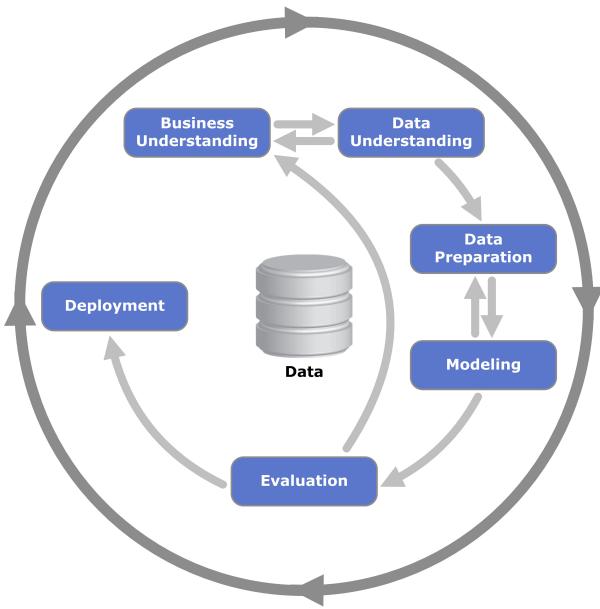


Figura 1: Diagrama de processos que mostra la relació entre les diferents fases de CRISP-DM. Autor: Kenneth Jensen. Font: Wikipedia

En la primera fase, segons aquesta metodologia, s'explora i es clarifiquen els objectius del projecte de ciència de dades per assegurar que estiguin alineats amb els objectius del negoci. En aquest cas, el primer pas ha consistit a **definir els objectius** del Treball de Final de Grau (TFG) i parlar amb el tutor sobre el plantejament, l'estat actual del problema i planificació de les tasques a realitzar.

Una vegada està clar el que el client demana, o en aquest cas, els objectius a assolir del TFG, es passa a la fase de **comprendió de les dades**. Això consisteix a dur a terme una anàlisi exploratòria inicial amb la finalitat d'obtenir una visió general de què es pot aconseguir amb les dades i detectar possibles problemes o mancances. Veure: *Estudi preliminar del conjunt de dades*. [5] Aquesta fase complementa la feina de la fase anterior, duent a terme una anàlisi guiada pel coneixement del problema adquirit. Veure: *Context del problema*. [4]

La tercera fase consisteix en la **preparació de les dades**. L'objectiu final d'aquesta fase és obtenir les dades finals sobre les quals s'aplicaran els models. Això implica la neteja de les dades i una exploració visual per detectar correlacions entre les variables, la distribució de les dades, valors atípics, etc. Veure *Preprocessament i visualització de les dades*. [5.3]

En la quarta fase, s'apliquen tècniques d'anàlisi de dades per **construir o emprar models preentrenats** que permetin assolir els objectius del projecte. Veure *Anàlisi de sentiments* [6] i *Predictió de la puntuació del videojoc a partir dels comentaris* [7]

A continuació, hi ha una **fase d'avaluació** en la qual es comprova el grau d'acompliment dels objectius respecte als resultats del model. També s'ha de revisar tot el procés que s'ha seguit fins a arribar a aquest punt i establir les accions pertinents, ja sigui validar el projecte i passar a producció, establir noves línies d'investigació o repetir fases anteriors per buscar millores.

Per últim, segons CRISP-DM, es duen a terme les accions necessàries per a la implementació en l'entorn de producció. Es desenvolupen plans per integrar els resultats en els processos de negoci existents i es realitza una fase de manteniment o es torna a iterar en la metodologia. En aquest cas, es desenvoluparà unes diapositives amb els **resultats obtinguts i un resum de la feina feta** per la defensa al tribunal del TFG. [20] [21]

2.4 Programes i llibreries utilitzades

Per a l'obtenció de dades s'utilitzarà **Kaggle**, que és la plataforma que reuneix la comunitat de Data Science més gran del món. En ella els usuaris poden trobar i publicar datasets, explorar i construir models de dades, i participar en competicions de ciència de dades per resoldre problemes reals. Segons la informació de la pàgina web del dataset “The Last of Us Reviews”, les dades de les ressenyes recopilades provenen d'una pàgina web anomenada **Metacritic**. Aquesta pàgina aglutina una gran quantitat de ressenyes de pel·lícules, programes de televisió, àlbums de música i videojocs.

L'entorn escollit per analitzar i desenvolupar tot el projecte és **Google Colab** és un servei de Google que permet escriure i executar codi Python des del navegador.

Per guardar i compartir el codi s'ha escollit **Google Drive** per la seva integració amb Google Colab i la fàcil utilització del servei.

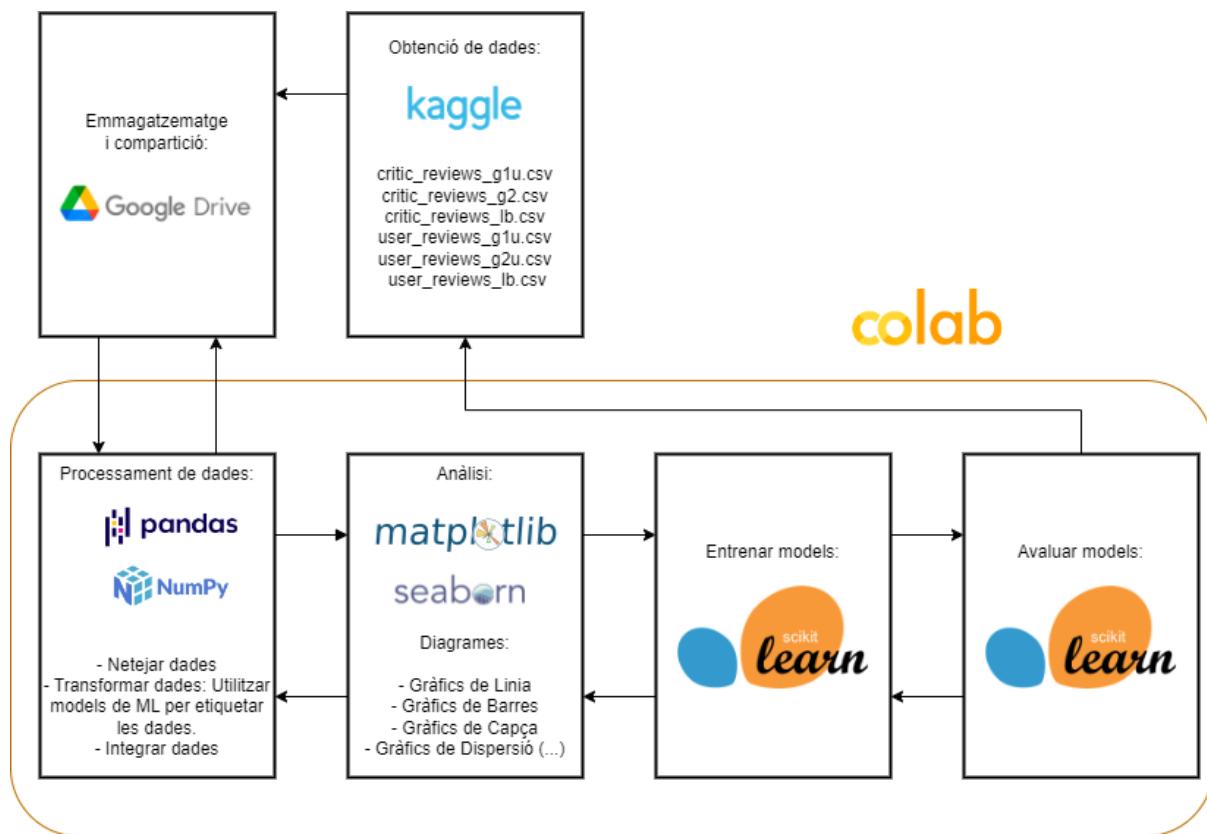


Figura 2: Diagrama del cicle de vida de les dades i les eines utilitzades

També s'han importat una sèrie de llibreries per facilitar el preprocessament i l'anàlisi dels conjunts de dades:

- **Pandas** és una llibreria de Python utilitzada per netejar, transformar i integrar dades. Facilita la manipulació de dades en formats tabulars (DataFrames).
- **Numpy** és una llibreria de Python que proporciona suport per a arrays multidimensionals i una varietat de funcions matemàtiques per a operacions numèriques.
- **Matplotlib** és una llibreria de Python utilitzada per crear diferents tipus de gràfics.
- **Seaborn** és una llibreria de Python basada en Matplotlib que facilita la creació de gràfics estadístics i la visualització de dades complexes.
- **Scikit-learn** és una llibreria de Python que proporciona eines per l'entrenament i avaluació de models.

models de ML, incloent-hi algorismes de regressió, classificació, clustering, i reducció de dimensialitat.

A més, per dur a terme l'anàlisi de sentiments i la traducció de les ressenyes es va importar la llibreria de ***transformers*** per utilitzar diversos models de la pàgina web **Hugging Face**. Hugging Face és una plataforma i comunitat de ciència de dades i ML que ajuda els usuaris a construir, desplegar i entrenar models d'aprenentatge automàtic. Per importar els models es va fer ús dels *pipelines* que permeten importar els models al projecte d'una forma molt senzilla.

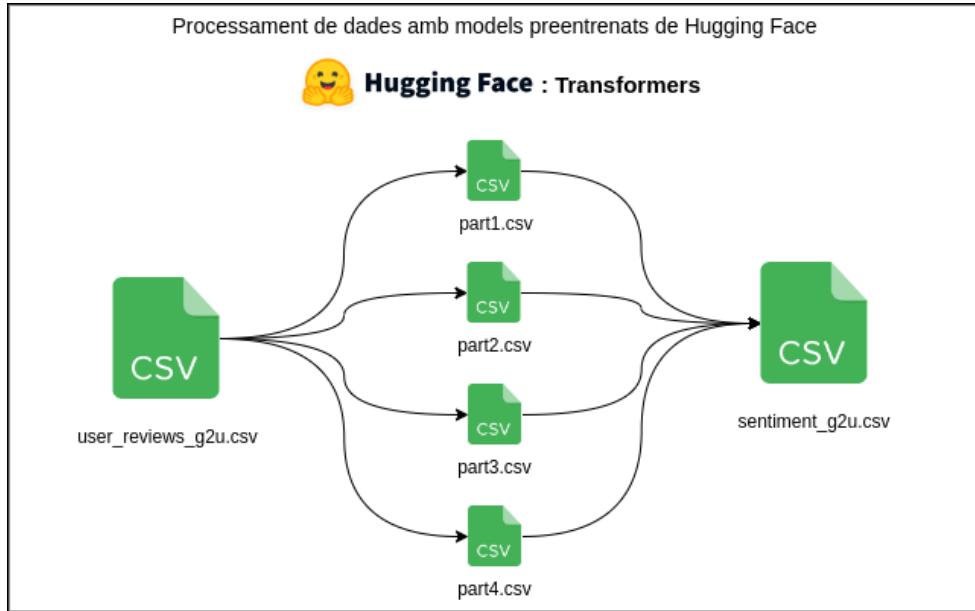


Figura 3: Processament de dades amb transformers

El següent esquema il·lustra com les dades són dividides, processades i integrades de nou en un mateix conjunt de dades. Aquesta estratègia es va utilitzar perquè el processament del conjunt de dades de forma seqüencial requeria molt temps i es va decidir paral·lelitzar el procés amb múltiples màquines virtuals de Google Colab per reduir temps de càlcul.

Altres llibreries que també es van utilitzar:

- **SciPy** és una llibreria de Python utilitzada per a computació científica i tècnica.
- **Tqdm** és una llibreria de Python utilitzada per visualitzar barres de progrés de bucles, iteracions i altres processos llargs en programes Python.
- **PyTorch** és una llibreria que proporciona eines per a la creació i entrenament de models de DL, així com per a la manipulació de tensors.

3 Conceptes teòrics

3.1 Introducció al Processament de Llenguatge Natural

3.1.1 Definició

El Processament de Llenguatge Natural (*PLN*; o *NLP* del seu nom en anglès, *Natural Language Processing*) és un camp interdisciplinari de les ciències de la computació i de la lingüística que, mitjançant models estadístics i d'Inteligència Artificial (IA), estudia les interaccions entre els computadors i el llenguatge humà. [1]

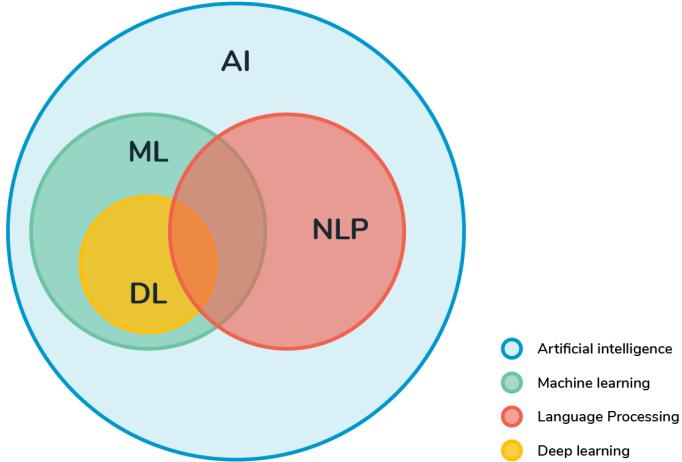


Figura 4: El NLP és una branca de la Intel·ligència Artificial que pot utilitzar tècniques de Machine Learning (ML) i Deep Learning (DL) per millorar el model. Autor: Balakrishnan Sathiyakugan. Font: [1]

3.1.2 Objectius

En informàtica, les llengües que utilitzen els humans per comunicar-se s'anomenen “llengües naturals”. Alguns exemples en són l'anglès, el francès i el català. Els primers ordinadors van ser dissenyats per resoldre equacions i processar números. No estaven destinats a entendre les llengües naturals. Els ordinadors tenen els seus propis llenguatges de programació (C, Java, Python) i protocols de comunicació (TCP/IP, HTTP, MQTT).

La forma tradicional d'interaccionar amb els ordinadors i donar-los instruccions és mitjançant un teclat i un ratolí. Llavors, per què no parlar amb l'ordinador i deixar-lo respondre en una llengua natural? Aquest és un dels objectius del NLP: donar a un ordinador l'habilitat “d'entendre” i generar llenguatge humà. [22]

3.1.3 Com funciona el NLP

El NLP fa servir moltes tècniques diferents per permetre que els ordinadors entenguin el llenguatge natural com ho fan els humans. Normalment, consta de dues fases principals: el preprocessament de dades i el desenvolupament d'algorismes. [23]

1. **El preprocessament de dades** implica preparar i netejar les dades de text perquè les màquines puguin analitzar-les. Després d'això, s'obté unes dades les quals són més adients per entrenar un model d'IA.
2. Un cop les dades han estat preprocessades, **s'utilitza un algorisme per processar-les**. Es pot utilitzar qualsevol algorisme d'IA per aquest propòsit. Des de ML, DL o altres models probabilístics.

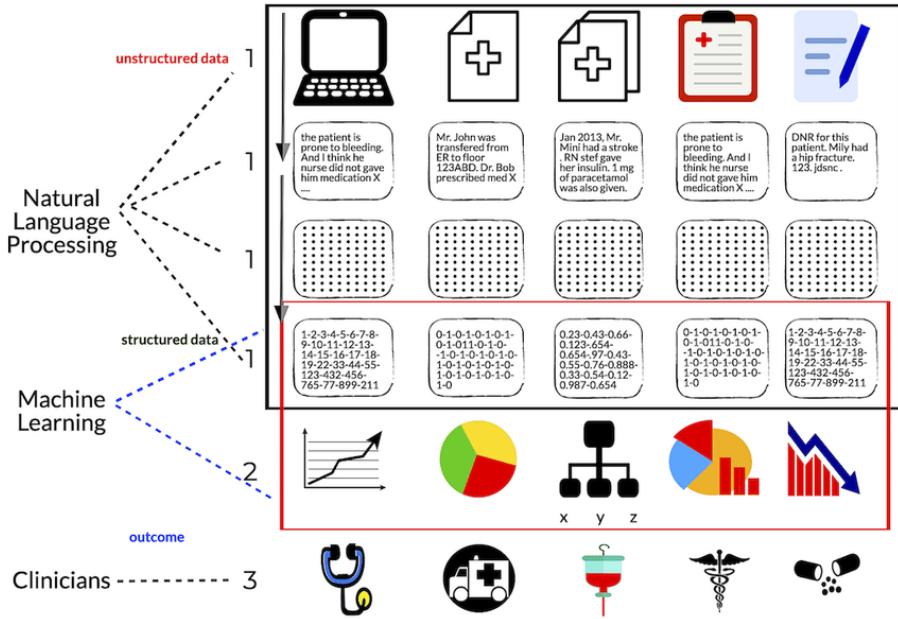


Figura 5: Il·lustració de com el processament del llenguatge natural converteix text no estructurat en dades estructurades, les quals poden ser analitzades posteriorment per algoritmes de ML en l'àmbit de la medicina. Autor: Avishek Choudhury. Font: [2]

3.1.3.a Preprocessament de les dades

El preprocessament posa les dades en una forma manipulable i destaca les característiques del text amb les quals un algorisme pot treballar. Més formalment, tal com s'ha vist en l'assignatura de Processadors de llenguatge [24], el tractament de dades es pot fer a diversos nivells:

1. **L'anàlisi lèxica:** Aquest nivell implica tots els passos inicials per netejar i normalitzar el text. Es llegeix la seqüència de caràcters de l'entrada i se separa en lexemes o tokens els components identificats.
2. **L'anàlisi sintàctica:** Aquest nivell se centra en l'anàlisi de l'estrucció gramatical del text, incloent-hi la identificació de les parts del discurs (substantius, verbs, adjectius, etc.), l'anàlisi de la sintaxi de les oracions i la construcció de l'estrucció jeràrquica del text.
3. **L'anàlisi semàntica:** Se centra principalment en la identificació de significats literals de les paraules i les relacions semàntiques entre elles.
4. **L'anàlisi pragmàtica:** Considera el significat del text en un context més ampli, incloent-hi factors com el propòsit comunicatiu, les intencions de l'autor, les implicacions socials i les inferències pragmàtiques que es poden fer del text.

En el preprocessament s'utilitzen diverses tècniques per transformar les dades. Algunes de les quals són:

- **Neteja de text:** En aquesta fase es transforma tot el text a minúscules, s'elimina els caràcters que no són paraules o espais en blanc i qualsevol dígit numèric present.
- **Tokenització:** Com s'ha avançat a la part de l'anàlisi lèxica, consisteix a separar els paràgrafs i frases en unitats més petites.

'I see a cup of coffee' → 'I', 'see', 'a', 'cup', 'of', 'coffee'

Figura 6: Exemple de tokenització d'una frase. Autor: Maleesha De Silva. Font: [3]

- **Transformació de contraccions:** Les contraccions són formes abreviades de paraules o frases, sovint formades combinant dues paraules, i són habituals en el llenguatge de cada dia.
- **Stopword removal**, o en català, “eliminació de paraules buides”, consisteix a eliminar les paraules que no aporten un valor significatiu al text. (articles, preposicions, conjuncions)

[This', 'is', 'an', 'example', 'for', 'stop', 'word', 'removal'] → [This', 'example', 'stop', 'word', 'removal']

Figura 7: Exemple de la tècnica de stopword removal. Autor: Maleesha De Silva. Font: [3]

- **Correcció ortogràfica:** consisteix a reconèixer paraules mal escrites i transformar-les a la seva forma correcta.
- **Eliminar paraules curtes o estranyes:** Les paraules poc freqüents, que es produueixen rarament, podrien no contribuir significativament a la comprensió global del text i podrien introduir soroll en l'anàlisi.
- **Stemming\Lematitació:** La primera fa referència al procés de transformar les paraules a la seva forma base o arrel, eliminant sufíxos i prefixos. (automatitzar, automàtic, automatització → automat). Mentre que la segona agrupa totes les formes flexionades d'una paraula sota una única forma base. (gat, gata, gats, gates → gat)
- **Named Entity Recognition (NER):** Consisteix a etiquetar les paraules segons categories previament definides, com ara persones, organitzacions, llocs, dates, quantitats, etc.
- **Part Of Speech (POS) tagging:** Depenent del problema a resoldre pot ser interessant assignar una etiqueta gramatical a cada paraula en un text, indicant la seva funció gramatical, com a substantiu, verb, adjetiu, etc.

* No totes aquestes tècniques s'utilitzaran en tots els casos, normalment dependrà del tipus de dades a tractar i del problema que es vulgui resoldre.

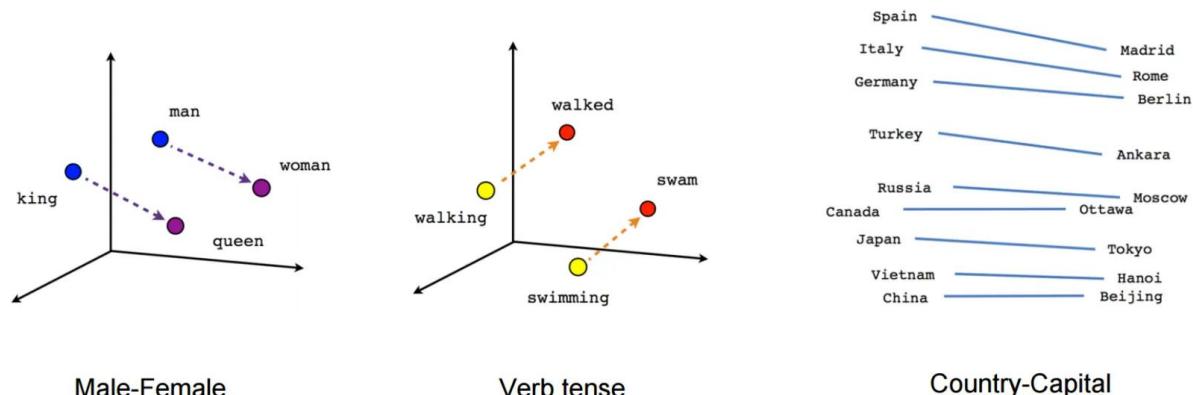


Figura 8: Exemple de preprocessament de dades i correlació de patrons. Autor: Mohd Sanad Zakirizvi. Font [4]

Després de fer el preprocessament, adaptarem l'entrada de manera que el nostre model pugui entendre les dades fàcilment i connectar les correlacions. [25] [22]

3.1.3.b Desenvolupament d'algorismes

Un cop s'ha fet el preprocessament, es pot procedir a examinar la informació estructurada. Aquesta és la fase d'anàlisi, en què s'estreuen diferents característiques per dur a terme la tasca prevista. Encara que hi ha un gran ventall de tècniques que es poden utilitzar, la gran majoria es poden classificar en tres tipus:

- **Sistemes basats en regles lingüístiques.** El NLP clàssic, arrelat en la lingüística i en els enfocaments simbòlics dels anys 50, es basava en regles elaborades manualment per a la sintaxi i la gramàtica per a derivar l'estructura del text i el seu significat. Per contra, les regles que es necessitaven eren nombroses i tenien problemes amb el llenguatge col·loquial. Aquests mètodes poden ser consumir molt temps i no generalitzen tan bé amb dades no vistes.
- **Sistemes basats en ML.** En la dècada dels 80, el gir cap a enfocaments estadístics va impulsar models de ML, DL els quals perfeccionen les seves pròpies regles mitjançant processament i aprenentatge reiterat. Aquests mètodes generalitzen millor amb les dades no vistes, però requereixen una gran quantitat de dades d'entrenament etiquetades i poden ser computacionalment costosos.

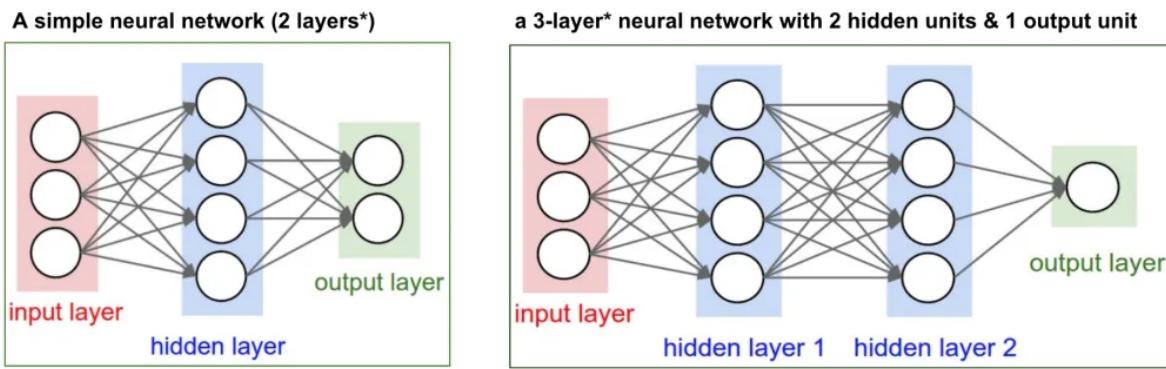


Figura 9: Exemples de xarxes neuronals. Font: Neural Network course, University of Toronto. Font: [5]

- **Sistemes híbrids.** Molts models actuals combinen el ML amb mètodes basats en regles peraprofitar les fortaleses de tots dos sistemes.

[26] [23] [6]

3.1.4 Àrees de treball

Les principals àrees de treball del NLP s'engloben en dos grans grups:

- **Natural Language Understanding (NLU):** Consisteix a convertir text en llenguatge natural a representacions estructurades o codificades. L'objectiu és resoldre ambigüïtats, obtenir context i entendre el significat de què s'està dient.
- **Natural Language Generation (NLG):** Consisteix a generar text de manera coherent i correcta a partir d'informació o dades estructurades prèvies.

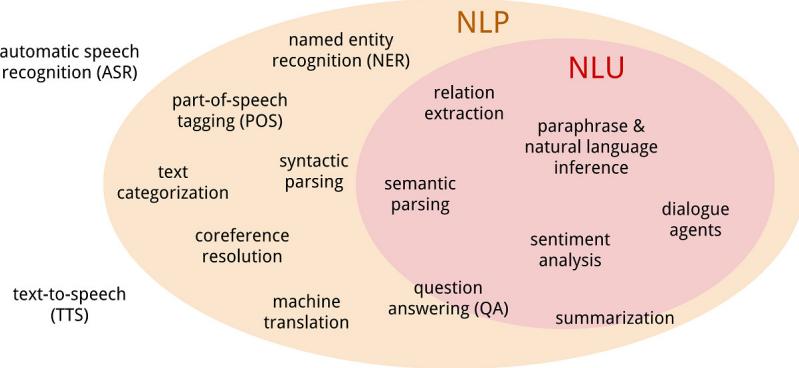


Figura 10: El NLU és un subconjunt de NLP. Autor: MacCartney 2014, slide 8 Font: [6]

A continuació, s'expliquen alguns dels problemes que es poden resoldre:

- **Resposta a preguntes:** Donada una pregunta un motor de NLP que té una basta amplitud de coneixement és capaç de proporcionar una resposta. L'aplicació del llenguatge per investigar dades no només millora el nivell d'accessibilitat, sinó que també redueix les barrires de l'anàlisi dins de les organitzacions, més enllà de la comunitat d'analistes i desenvolupadors de programari.
- **Traducció automàtica:** En el passat, els serveis de traducció no eren massa bons, ja que molts idiomes no permeten una traducció literal, però últimament han avançat molt. Gràcies al NLP, els traductors en línia poden traduir idiomes amb major precisió i oferir resultats gramaticalment correctes. Això resulta de gran utilitat quan intentem comunicar-nos amb algú en un altre idioma. Per descomptat, la traducció automàtica de textos no pot substituir els humans en textos complexos i on la fiabilitat sigui crucial (facció, textos legals, etc.). No obstant això, s'ha estès en els continguts d'internet i les xarxes socials. Sobretot, perquè és gratuïta, instantània i, en la majoria dels casos, acceptable per als nostres interessos.
- **Resum de text:** El NLP pot ser encarregat de resumir un document o un llibre sencer. Pot proporcionar un resum equilibrat d'una història publicada en diferents llocs web amb punts de vista diferents.
- **Anàlisi de sentiments:** A partir de ressenyes de productes o missatges en xarxes socials, la tasca consisteix a determinar si el sentiment és positiu, neutral o negatiu. Això és útil per als departaments d'atenció al client, enginyeria i màrqueting.
- **Classificació de text:** El NLP pot classificar les notícies per temes o detectar correu brossa. Aquests tòpics ens informen de manera general del contingut i serveixen per a una tasca més complexa: establir relacions entre conceptes.
- **Text a veu:** Donat un text, transforma les paraules i produeix una representació parlada. El text a veu es pot utilitzar per ajudar les persones amb discapacitat visual.
- **Veu a text:** Donat un fragment de so d'una persona o persones parlant, reconeix el discurs i determina la representació textual.
- **De text a vídeo:** Donada una descripció d'un vídeo, genera un vídeo que coincideixi amb la descripció.

Existeix una gran quantitat de problemes, i aquí només se n'esmenten alguns. La gran majoria d'aquests problemes necessiten una resolució en temps real. [27] [22] [28] [7]



Figura 11: Camps del NLP segons la tasca que desenvolupen. Autor: Fabio Chiusano. Font: [7]

3.1.5 Reptes del llenguatge

Hi ha nombrosos reptes en el processament del llenguatge natural, alguns dels quals són:

- **La precisió del llenguatge.** Tradicionalment, els ordinadors han requerit que els humans els parlin en un llenguatge de programació precís, no ambigu i altament estructurat, o mitjançant un nombre limitat de comandes clarament definides. No obstant això, el llenguatge humà no sempre és precís; sovint és ambigu i l'estrucció lingüística pot dependre de moltes variables complexes, incloent-hi l'argot, els dialectes regionals i el context social.

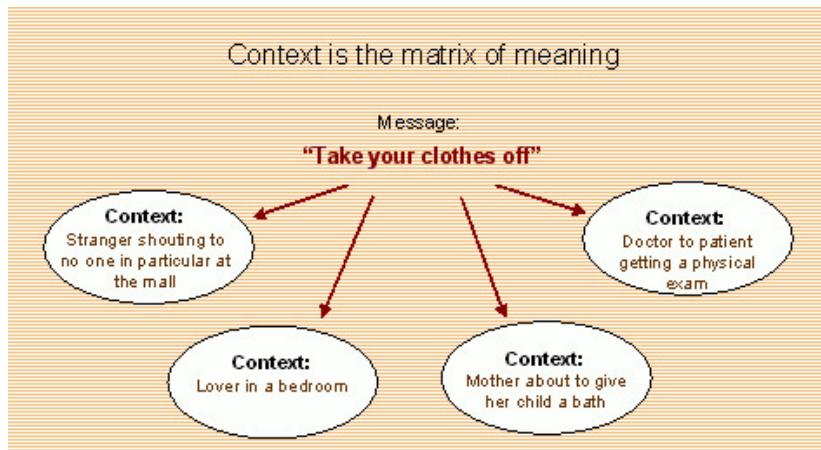


Figura 12: Exemple com la mateixa frase en un context o en un altre canvia completament el significat. Autor: John Teleska. Font: [8]

- **La tonalitat de veu i inflexió.** El NLP encara no ha estat perfeccionat i té dificultats en l'ús

abstracte del llenguatge, com per exemple el sarcasme. Aquestes situacions normalment requereixen comprendre les paraules utilitzades i el seu context dins d'una conversa. A més, una frase pot canviar de significat segons quina paraula o sillaba es posi èmfasi. Aquest i molts altres detalls es poden perdre amb els models actuals i són molt importants en el reconeixement de la parla.

- **L'evolució del llenguatge.** El processament del llenguatge natural també es veu compromès pel fet que el llenguatge, i la manera com la gent l'utilitza, canvia contínuament. Les regles computacionals estrictes que funcionen ara poden quedar-se obsoletes a mesura que les característiques del llenguatge del món real canvien amb el temps.
- **Els esbiaixos.** Els sistemes de NLP poden reflectir en els seus processos els esbiaixos de les dades amb els que han sigut entrenats. Això suposa un problema molt gran, ja que una persona o un collectiu pot ser discriminat.

3.1.6 Importància en la indústria i aplicacions

Avui dia, les empreses utilitzen grans quantitats de dades no estructurades, la gran majoria en format de text, i necessiten una manera d'analitzar-les eficientment. Molta de la informació creada en línia i emmagatzemada en bases de dades és llenguatge humà natural, i fins fa poc, les empreses no podien analitzar aquestes dades de manera efectiva. Ara, amb les millores en els mètodes de DL i ML, els algorismes poden ampliar l'abast i la comprensió de les dades i proporcionar coneixements molt valuosos. [23]

Quant a les aplicacions del NLP son nombroses, destaca la creixent popularització dels chatbots multimodals, com Chatgpt (OpenAI) o Gemini (Google), Llama 3 (Meta), també sistemes automatitzats per processar currículums, filtres de spam als correus, detecció de plagis, anàlisi i categorització d'històrics mèdics, agilització de recerca acadèmica, audiollibres per a persones amb discapacitat visual, atenció al client automàtica, anàlisi de ressenyes de clients, etc.

3.2 Algorismes per al Processament del Llenguatge Natural

3.2.1 Mètodes bàsics

3.2.1.a Bag-of-words

En català, bossa de paraules, és un model de llenguatge estadístic molt senzill utilitzat per analitzar textos basant-se en el recompte de paraules.

	about	bird	heard	is	the	word	you
About the bird , the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

Figura 13: Exemple que mostra la tècnica de la bossa de paraules. Per cada text es calcula la freqüència d'aparició de cada paraula i es crea una matriu. Autor: Alexis Perrier. Font: [9]

Aquesta tècnica es pot implementar com un diccionari en Python, amb cada clau corresponent a una paraula i cada valor establert com el nombre de vegades que aquesta paraula apareix en el text. De la matriu generada serà necessari reduir la mida per evitar realitzar càculs sobre matrius gegants. La idea és eliminar tants tokens com sigui possible sense descartar informació rellevant aplicant les tècniques mencionades en l'apartat de *Preprocessament de dades*. Sec. 3.1.3.a. Amb la matriu reduïda, es podria calcular directament el sentiment del text utilitzant un diccionari que assigni un valor de sentiment a cada paraula i fent la mitjana d'aquests valors. [9]

Si volguéssim entrenar un model de ML amb aquestes dades, hauríem de fer un pas addicional nomenat vectorització. La **vectorització** és el procés de convertir dades en forma de text o d'altra naturalesa en vectors o arrays numèrics. [29]

- **bird** → **[5, 1, 1]**
- **the** → **[2, 1, 2]**
- **word** → **[0, 0, 1]**
- ...

Figura 14: Exemple que mostra la tècnica de bossa de paraules amb la tècnica de vectorització. Autor: Alexis Perrier. Font: [9]

També caldria assignar a cada categoria o *feature* de la taula un índex numèric arbitrari, ja que molts d'aquests algorismes requereixen només dades numèriques com a entrada.

Limitacions: Aquest model ignora el context al descartar el significat de les paraules i centrar-se només en la freqüència d'aparició. Això pot ser un problema important, perquè l'organització de les paraules en una frase pot canviar completament el significat de la frase i el model no pot tenir això en compte.

3.2.1.b TF-IDF

La tècnica de Term Frequency-Inverse Document Frequency (TF-IDF) és una tècnica per mesurar la rellevància d'una paraula en un document dins d'una col·lecció de documents. Això es fa multiplicant dues mètriques:

- Freqüència de terme (TF): quantes vegades apareix una paraula en un document.

$$\text{TF}(t, d) = \frac{\text{Nombre de vegades que el terme } t \text{ apareix en el document } d}{\text{Nombre total de termes en el document } d}$$

- Freqüència de document inversa (IDF): Mesura la importància d'una paraula en tot el conjunt de documents. La idea és que les paraules comunes en molts documents tenen menys valor distintiu que les paraules que apareixen en pocs documents.

$$\text{IDF}(t, D) = \log \left(\frac{\text{Nombre total de documents}}{\text{Nombre de documents que contenen el terme } t} \right)$$

TF-IDF va ser dissenyat inicialment per a la cerca de documents, on s'executa una consulta i el sistema ha de trobar els documents més rellevants. Suposem que la consulta és el text “El lleopard”. El sistema donaria a cada document una puntuació més alta proporcionalment a les freqüències de les paraules de la consulta trobades en el document, donant més pes a paraules rares com “lleopard” respecte a paraules comunes com “El”. [10]

	ablaze	accident	car	caught	fire	jam	kind	sadly	set	swear	true	up	world
0	0.00	0.00	0.00	0.0	0.0	0.00	0.67	0.53	0.00	0.00	0.53	0.0	0.00
1	0.47	0.00	0.00	0.0	0.0	0.47	0.00	0.00	0.47	0.37	0.00	0.0	0.47
2	0.00	0.59	0.47	0.0	0.0	0.00	0.00	0.00	0.00	0.47	0.47	0.0	0.00
3	0.00	0.00	0.64	0.4	0.4	0.00	0.00	0.32	0.00	0.00	0.00	0.4	0.00

Figura 15: Quan s'aplica TF-IDF a un conjunt de documents, s'obté una representació vectorial de cada document, on cada dimensió del vector correspon a una paraula, i el valor de cada dimensió és el seu pes TF-IDF. Autor: Turing. Font: [10]

Limitacions: El valor de TF-IDF està influït per la longitud del document. En general, els documents més llargs poden tenir valors de TF més alts simplement perquè hi ha més paraules. Això pot causar problemes si volem comparar la importància relativa de les paraules en documents de longituds diferents. Tampoc té en compte el context de les paraules dins del document ni les relacions semàntiques entre elles. [7]

3.2.1.c Word embeddings

En català, incrustació de mots, és una tècnica que permet representar paraules com a vectors, a vegades amb desenes o centenars de dimensions, facilitant així la identificació de relacions semàntiques i similituds entre les paraules. Això és especialment útil en l'anàlisi de sentiments, ja que permet identificar matisos subtils de l'expressió emocional dels textos.

Per obtenir la representació vectorial d'una paraula es pot utilitzar un model preentrenat amb una gran quantitat de textos, com per exemple, **spaCy**.

```
nlp = spacy.load('en')
print(nlp('peace').vector)
```

Output: [5.2907305, -4.20267, 1.6989858, -1.422668, -1.500128, ...]

Cada dimensió del vector capture un aspecte diferent del significat de la paraula. L'objectiu d'utilitzar els “embeddings” es basa en la premissa que les paraules amb un significat similar tenen una representació vectorial similar.

Llavors, podem mesurar la distància entre les paraules amb diversos algorismes per determinar la similitud semàntica entre elles. Els més comuns soLEN ser:

- Distància de Manhattan
- Distància Euclidiana
- Distància del Cosinus

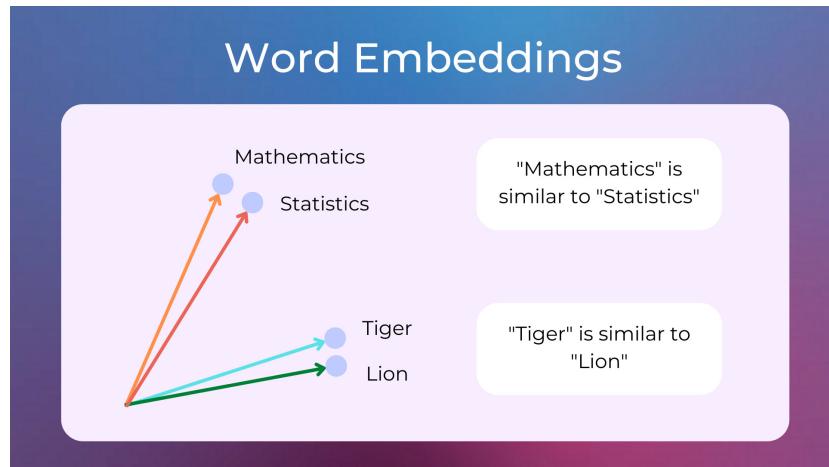


Figura 16: Les paraules similars tenen representació vectorial similar. Autor: Fabio Chiusano. Font: [11]

En la pràctica, generalment s'utilitza la distància del cosinus, perquè genera valors molt més petits en mesurar l'angle entre dos vectors. Els resultats poden estar entre -1 i 1.

- 1 indica que els vectors són completament alineats, és a dir, apunten exactament en la mateixa direcció i, per tant, les paraules són semànticament molt similars.
- 0 indica que els vectors són ortogonals, és a dir, no tenen cap similitud semàntica.
- -1 indica que els vectors són completament oposats, és a dir, apunten en direccions oposades i, per tant, les paraules tenen significats contraris.

Aquesta forma de representar les dades és un dels avenços clau per millorar el rendiment dels mètodes de DL en problemes de NLP, com l'anàlisi de sentiments. [30]

Limitacions: Les “embeddings” de paraules depenen en gran mesura de la qualitat i la mida del *corpus* (conjunt de dades d’input). Si el *corpus* és esbiaixat o insuficient, les representacions resultants també seran esbiaixades o poc precises. A més a més, si una paraula no es troba en el vocabulari utilitzat per entrenar els “embeddings”, llavors no té representació vectorial.

En l’última dècada han aparegut diversos mètodes per crear aquestes incrustacions de les paraules o “embeddings” [11]. En la següent imatge es mostren alguns d’ells.

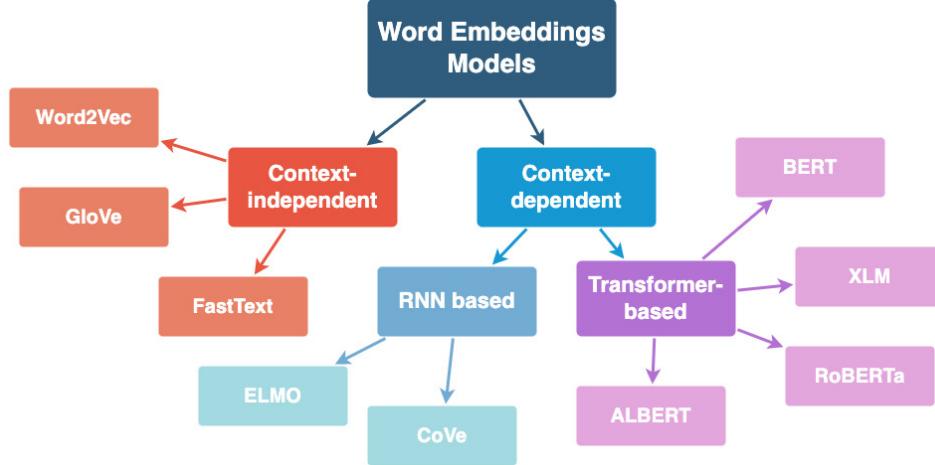


Figura 17: Diagrama en forma d'arbre que mostra els diferents mètodes i la seva classificació en dos grans grups: les de context independent i les de context dependent. Autor: Fabio Chiusano. Font: [11]

- Les “embeddings” de **context independent**: són representacions que no tenen en compte el context de la paraula en el text. Això vol dir que les paraules homònimes com ”serra”(que pot ser l'eina o la muntanya) no seran desambiguades segons el context i sempre tindran una única representació que, segons com, capture els dos significats.

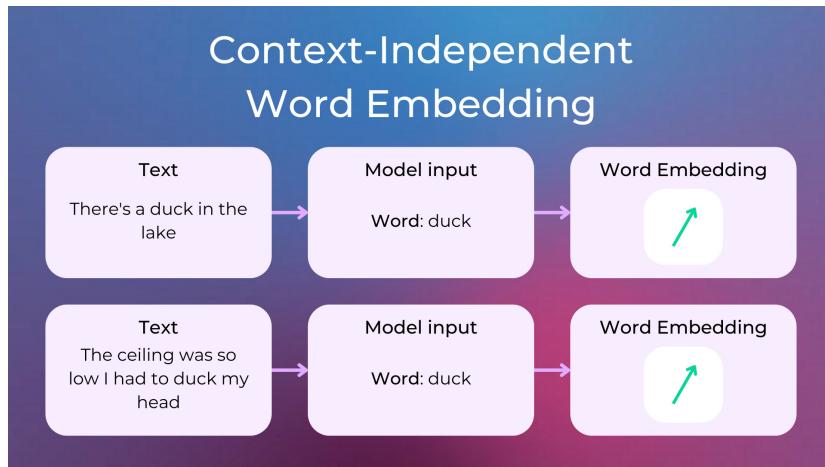


Figura 18: Exemple de com la mateixa paraula adopta la mateixa representació en contexts diferents. Autor: Fabio Chiusano. Font: [11]

- Les “embeddings” de **context dependent** aprenen diferents representacions de la mateixa paraula basada en el seu context.

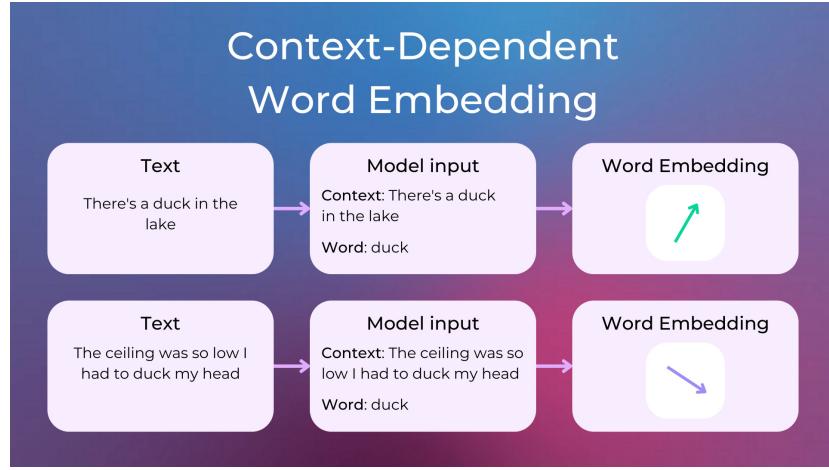


Figura 19: Exemple de com la mateixa paraula adopta diferents representacions en contextos diferents.
Autor: Fabio Chiusano. Font: [11]

3.2.1.d Word2Vec

Word2Vec és un algorisme que utilitza tècniques estadístiques per crear representacions numèriques (veccors) de paraules de manera autònoma i eficient, basant-se en l'anàlisi de grans col·leccions de text.

Va ser desenvolupat per Tomas Mikolov a Google l'any 2013 com una resposta per fer més eficient l'entrenament de les “embeddings” basades en xarxes neuronals, i des de llavors s'ha convertit en l'estàndard de facto per al desenvolupament d’“embeddings” preentrenats.

Word2Vec engloba dos variants per la generació de “embeddings” de les paraules:

- *Continuous Bag-of-Words (CBOW)*
- *Continuous Skip-Gram Model*

El primer mètode, CBOW, s'entrena amb un *corpus* molt gran de paraules. Durant l'entrenament, el model intenta predir cada paraula basant-se en el context de les paraules circumdants. A mesura que el model es va entrenant amb moltes frases i contextos diferents, aprèn les representacions vectorials (embeddings) de les paraules. Aquestes embeddings capturen les relacions semàntiques i sintàctiques entre les paraules. Per tant, si una paraula és polisèmica (té més d'un significat), el model aprendrà una representació que serà una mena de mitjana dels diferents significats dels contextos del corpus de l'entrenament.



Figura 20: En aquest exemple, s'intenta predir la paraula “listen”, partint d'una mida de finestra de 5 paraules circumdants, que capturen el context que envolta la paraula. Autor: Turing. Font: [10]

El segon, Continuous Skip-Gram, funciona just a l'invers, intenta predir les paraules circumdants donada només la paraula actual. En ambdós casos, l'objectiu final de Word2Vec és aprendre els pesos de la capa oculta de la xarxa neuronal. Aquests pesos s'utilitzaran com a “embeddings” de les paraules.

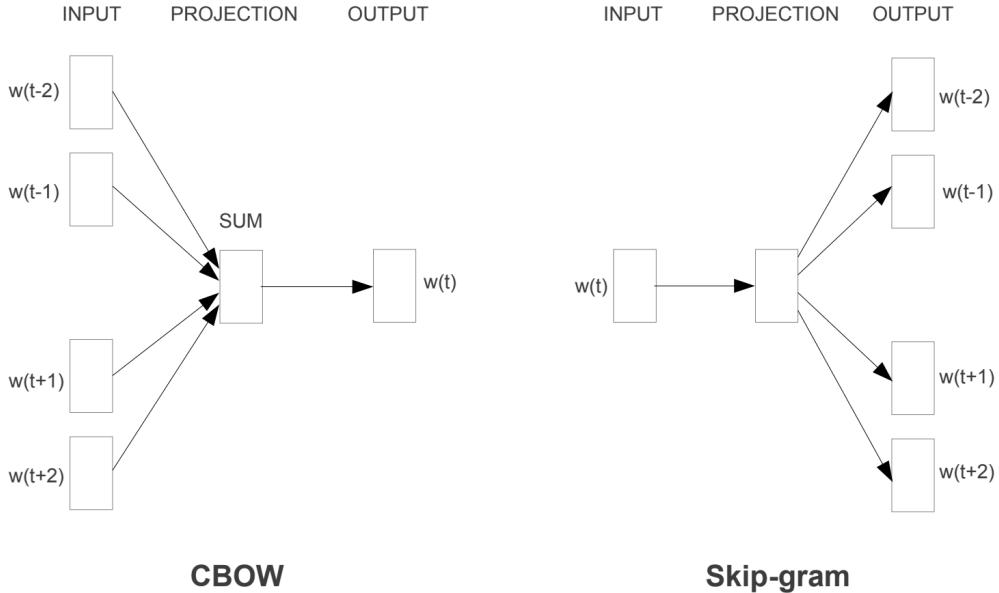


Figura 21: Diagrama de les dues variants de Word2Vec. Autor: "Efficient Estimation of Word Representations in Vector Space", 2013. Font: [12]

El benefici clau d'aquest enfocament Word2Vec és que les “embeddings” de paraules es poden aprendre de manera més eficient, permetent generar vectors amb més dimensions a partir de *corpus* de text molt més extensos (milions de paraules).

Limitacions: Word2Vec genera una única representació vectorial per a cada paraula, independentment del seu context. Això vol dir que no pot diferenciar entre paraules polisèmiques. A més, a causa de la limitació de la mida de finestra, no pot modelar relacions de llarg abast amb un context més gran. A més a més, requereix una gran quantitat de dades d’entrenament. [31] [10] [12]

3.2.2 Mètodes avançats

3.2.2.a Xarxes Neuronals

Una xarxa neuronal és un model computacional inspirat en la manera en què funciona el cervell humà. Aquest tipus de model és utilitzat principalment en el camp del DL per reconèixer patrons i fer prediccions.

Les xarxes neuronals estan formades per “neurones artificials”, també conegudes com a nodes, que estan organitzades en capes. Hi ha tres tipus principals de capes en una xarxa neuronal:

- **Capa d'entrada:** Aquesta és la primera capa de la xarxa i rep les dades d'entrada. Cada neurona en aquesta capa representa una característica del conjunt de dades d'entrada.
- **Capes ocultes:** Aquestes són les capes intermèdies entre la capa d'entrada i la capa de sortida. Les neurones en aquestes capes processen les dades rebudes de la capa anterior aplicant pesos i funcions d'activació per detectar característiques complexes i patrons.
- **Capa de sortida:** Aquesta és l'última capa de la xarxa, on cada neurona representa una possible sortida o resultat de la xarxa. Les dades processades a través de les capes ocultes arriben a la capa de sortida per proporcionar el resultat final.

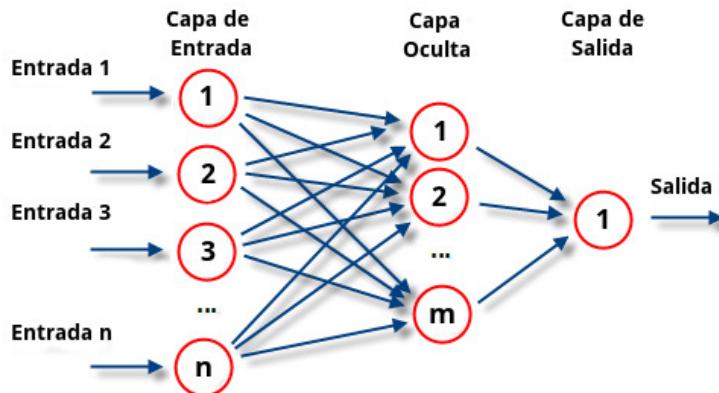


Figura 22: Diagrama d'una Xarxa Neuronal. Autor: -. Font: [13]

Cada capa consta d'un cert nombre de neurones, o nodes. Les connexions entre neurones estan representades per pesos que determinen la força de la connexió. La xarxa neuronal aprèn aquests pesos durant un procés d'entrenament per optimitzar el seu rendiment en una tasca específica.

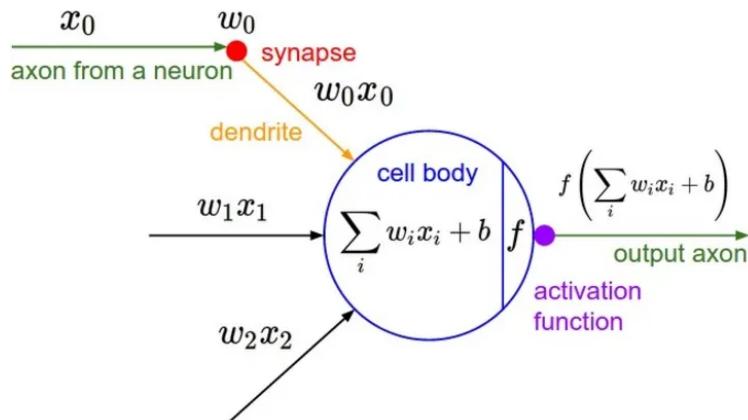


Figura 23: Diagrama d'una neurona artificial. Autor: Neural Network course, University of Toronto. Font: [5]

Illustrative example of a neural network

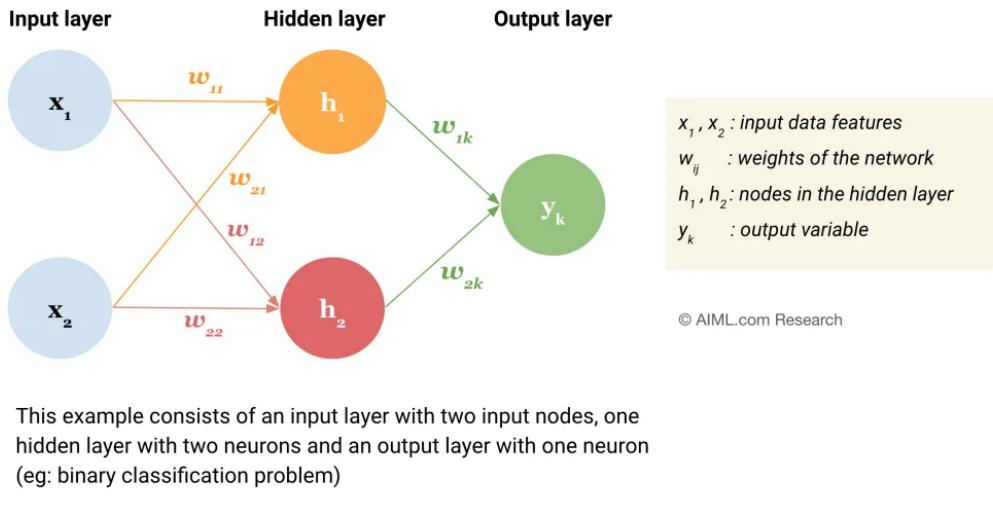


Figura 24: Exemple d'una xarxa neuronal. Es pot observar una capa d'entrada amb dues neurones, una capa oculta amb dues neurones més i una capa de sortida amb una neurona. (ex: problema de classificació binària) Autor: Neural Network course, University of Toronto. Font: [5]

Procés d'entrenament d'una Xarxa Neuronal:

L'entrenament d'una xarxa neuronal consta de dues fases interconnectades: la propagació cap endavant i la retropropagació.

- En la primera fase, coneguda com a **propagació cap endavant**, les dades d'entrada són processades a través de les diverses capes de neurones de la xarxa. Cada capa realitza càlculs específics utilitzant els pesos i els biaixos assignats, i genera prediccions o estimacions de sortida.
- Una vegada s'han obtingut les prediccions, comença la segona fase: **la retropropagació**. La retropropagació, o propagació enrere, ajusta els pesos per minimitzar l'error en les prediccions de la xarxa. Conceptualment, la retropropagació comença per l'últim paràmetre i avança enrere per estimar tots els altres paràmetres.

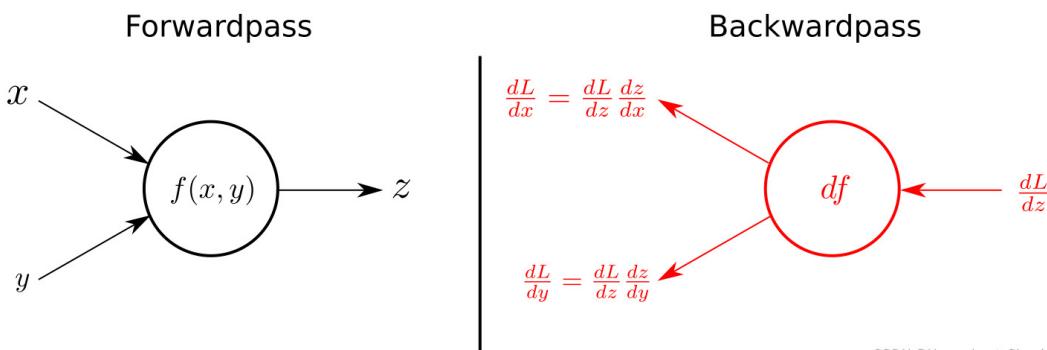


Figura 25: Fórmules de la propagació cap endavant (esquerra) i retropropagació (dreta). Autor: Frederick kratzert's blog. Font: [14]

Així, mentre la propagació cap endavant proporciona les prediccions de la xarxa, la retropropagació utilitza aquestes prediccions per calcular l'error i ajustar els paràmetres de la xarxa. Aquest procés permet a la xarxa aprendre dels seus propis errors durant el procés d'entrenament, optimitzant els paràmetres per millorar el rendiment global de la xarxa.

Més formalment, els passos generals d'entrenament són els següents:

1. En l'etapa de propagació cap endavant, les dades flueixen a través de la xarxa per obtenir les sortides.
2. La funció de pèrdua es fa servir per calcular l'error total.
3. A continuació, s'utilitza l'algoritme de retropropagació per calcular el gradient de la funció de pèrdua respecte a cada pes i biaix.
4. Finalment, s'utilitza el Descens del Gradient per actualitzar els pesos i els biaixos en cada capa.
5. Es repeteixen els passos anteriors per minimitzar l'error total de la xarxa neuronal.

En l'Apèndix s'adjunta un exercici de propagació cap endavant on es veu en més detall els càlculs de cada neurona i les funcions predites.

Hiperparàmetres d'un model de Xarxa Neuronal:

A continuació es presenta alguns hiperparàmetres que es poden ajustar per millorar el rendiment d'una xarxa neuronal:

- **Nombre de capes ocultes**, també conegut com la profunditat de la xarxa, permeten a la xarxa aprendre nivells de característiques més abstractes i complexes, però alhora també poden portar a sobreajustament.
- **Nombre de neurones per capa**: impacta la capacitat de la xarxa per captar patrons complexos. Capes més grans poden captar característiques més intrincades, però d'igual manera, poden comportar sobreajustament.
- **La funció d'activació**: introduceix no-linealitat al model, permetent captar relacions complexes en les dades.

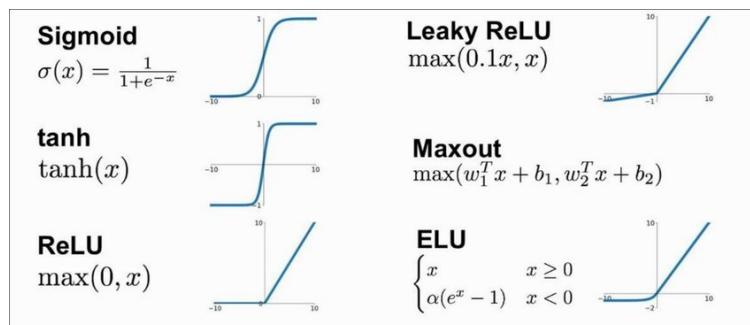


Figura 26: Exemples de funcions d'activació. Autor: -. Font: [5]

- **La funció de pèrdua**: quantifica la diferència entre els valors predictius i els valors reals. L'entrenament de la xarxa neuronal té com a objectiu minimitzar la funció de pèrdua. Per exemple: Error Quadràtic Mitjà (MSE), Pèrdua Huber per a regressió, Entropia creuada per a classificació.
- **Les èpoques**: són el nombre de vegades que tot el conjunt de dades d'entrenament es passa a través de la xarxa durant l'entrenament. Per exemple: 1, 10, 50, 100.
- **El nombre d'iteracions**: és el nombre de vegades que el model realitza propagacions cap endavant i cap enrere, així com actualitzacions dels seus pesos i biaixos.

Les xarxes neuronals són extremadament versàtils i s'utilitzen en una àmplia varietat d'aplicacions, com el reconeixement de veu, la visió per computador, la traducció automàtica, la detecció de frauds, l'**anàlisi de sentiments**, etc. [5] [13] [32] [33] [14]

3.2.2.b Transformadors

Els Transformadors, en anglès *Transformers*, són un tipus d'arquitectura de xarxes neuronals profundes utilitzada en el camp del Processament del Llenguatge Natural. Van ser introduïts per primer cop en

l'article “Attention is All You Need” de Vaswani, el 2017. Aquesta arquitectura va suposar una revolució en el camp del NLP, ja que va oferir millors significatives en la manera com les màquines processaven i generaven el llenguatge.

Els Transformadors van resoldre algunes de les limitacions dels models anteriors, com les Xarxes Neuronal Recurrents (RNN) i els Long Short-Term Memory networks (LSTM). Fins llavors era complicat paral·lelitzar els algorismes per fer-los més eficients, ja que aquests models processaven les dades de manera seqüencial, cosa que impedia aprofitar plenament les capacitats de parallelització dels processadors moderns. A més, les RNN tenien dificultats per gestionar les dependències a llarg termini en les seqüències de text. Això feia que, a mesura que la cadena d’informació es feia més llarga, la probabilitat que la informació es perdés augmentés.

En canvi, els Transformadors utilitzen un mecanisme conegut com a “**atenció**” per gestionar aquestes dependències. El mecanisme d’atenció permet que cada paraula d’una seqüència de text pugui considerar directament altres paraules de la mateixa seqüència, independentment de la seva distància en el text. Això no només millora la capacitat per capturar relacions a llarg termini, sinó que també permet un processament paral·lel molt més eficient. [34] [15]

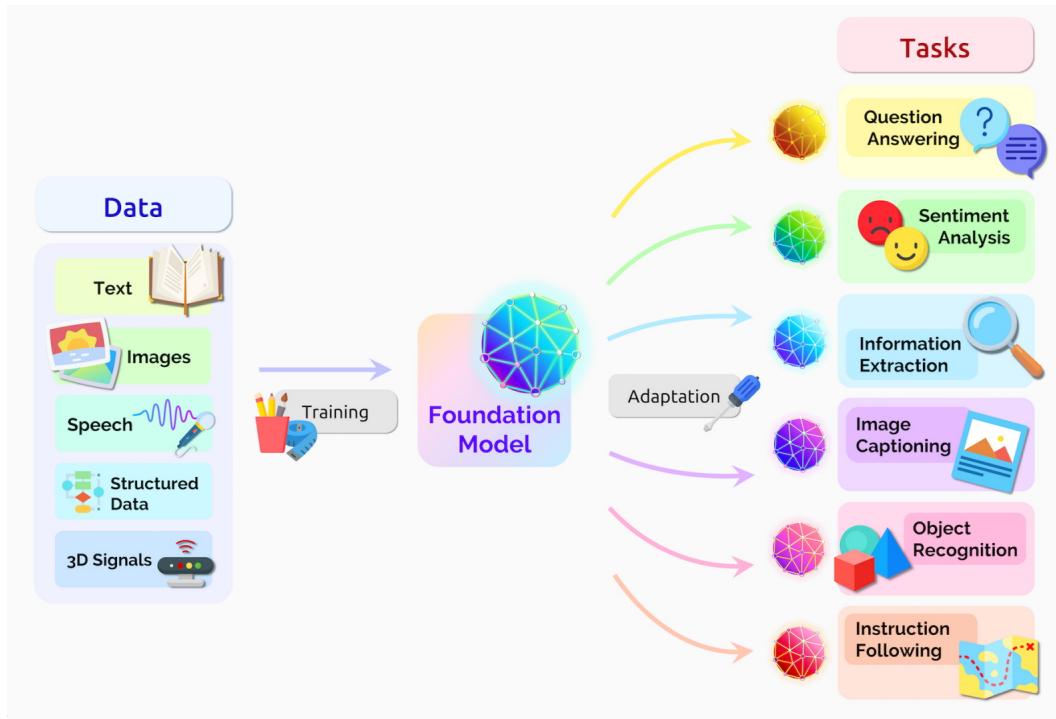


Figura 27: Els Transformadors, de vegades anomenats models fonamentals, són la tecnologia subjacent de moltes aplicacions recents. Autor: Rick Merritt. Font: [15]

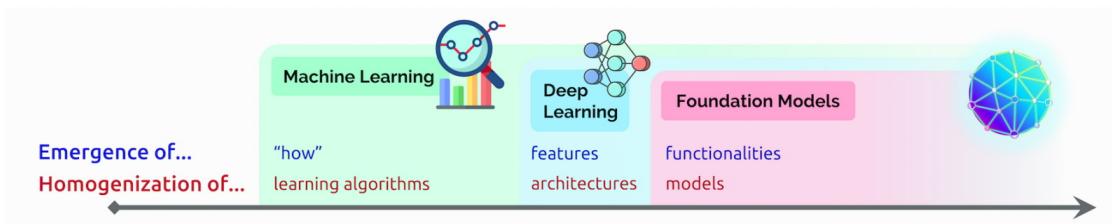


Figura 28: Evolució de l’Aprendentatge Automàtic. Autor: Rick Merritt. Font: [15]

3.2.2.c BERT

BERT és l'acrònim de “*Bidirectional Encoder Representations from Transformers*”. És un mètode avançat desenvolupat per Google per al processament de llenguatge natural que va ser publicat el 2018.

BERT utilitzà un enfocament bidireccional per comprendre el context de les paraules en una frase. Això vol dir que, a diferència dels models unidireccionals anteriors, que llegeixen el text de manera seqüencial (d'esquerra a dreta o de dreta a esquerra), BERT té en compte tant el context anterior com el posterior a una paraula per a una comprensió més rica i precisa.

Adoptant aquest enfocament, els models BERT preentrenats es poden ajustar finament amb només una capa de sortida addicional per crear models d'avantguarda per a una àmplia gamma de tasques, com ara la resposta a preguntes i la inferència de llenguatge, sense necessitar modificacions substancials al model subjacent.

Una enquesta de literatura del 2020 va concloure que en poc més d'un any, BERT s'ha convertit en una línia de base omnipresent en experiments de NLP, comptant més de 150 publicacions de recerca que analitzen i milloren el model.

Les seves implicacions inclouen la millora dels motors de cerca, fent-los més precisos en entendre les intencions dels usuaris, i l'enriquiment d'aplicacions de traducció, sistemes de resposta a preguntes i assistents虚拟s, augmentant la seva precisió i naturalitat. BERT també gestiona millor l'ambigüïtat del llenguatge i representa un gran avanç en la recerca d'IA, establint nous estàndards en el processament del llenguatge natural. [35]

3.2.2.d RoBERTa

RoBERTa (Robustly Optimised BERT Approach) és una variant del popular model BERT, que ha aconseguit resultats d'avantguarda en diverses tasques de NLP.

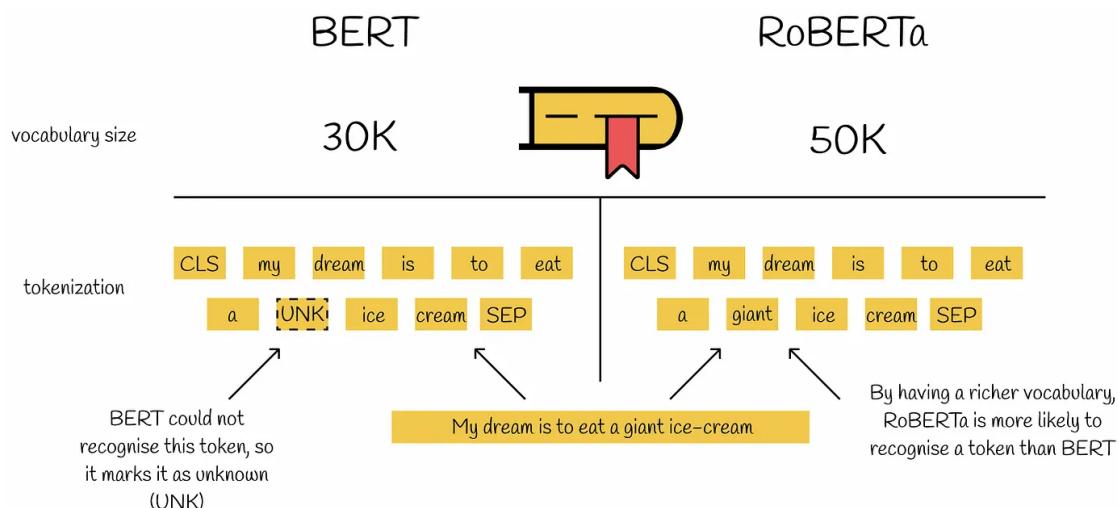


Figura 29: RoBERTa està entrenat en un conjunt de dades més gran i té més paràmetres que BERT, la qual cosa el fa més potent i flexible. Autor: Vyacheslav Efimov. Font: [16]

Com il·lustra la imatge, l'avantatge de RoBERTa respecte a BERT és que ha estat entrenat en un corpus de dades molt més gran. Això inclou tota la Wikipedia en anglès, i molts altres datasets com BookCorpus, Common Crawl i OpenWebText. També en l'entrenament s'han suprimit o actualitzat altres fases d'entrenament que utilitzava BERT i que el fan que de forma global un millor model.

3.2.3 Reptes tècnics del NLP

- **Obtenció de dades de qualitat.** Els sistemes de NLP necessiten grans quantitats de dades d'alta qualitat per entrenar-se. Aconseguir aquestes dades pot ser costós i discriminatòria les dades que no aporten valor requereix molt de temps. A més, la majoria de les dades disponibles a internet estan en anglès, fet que fa que els models entrenats en aquest idioma tinguin un rendiment superior comparat amb els models en altres llengües.
- **Privacitat i seguretat de les dades.** L'ús de grans quantitats de dades per entrenar models de PLN pot plantejar problemes de privacitat i seguretat. Garantir que les dades utilitzades no contenen informació sensible o personal és crucial però difícil de gestionar.
- **Escalabilitat del model.** Entrenar models grans de PLN requereix recursos computacionals massius, incloent-hi processadors gràfics (GPUs) i temps de càlcul considerable. Això pot limitar l'accés a aquests models a grans organitzacions amb recursos suficients.
- **Generalització a nous dominis.** Els models de PLN sovint tenen dificultats per generalitzar a nous dominis o contextos diferents dels quals han estat entrenats.
- **Processament de textos llargs.** El processament de textos llargs planteja reptes únics, com ara:
 - **Limitacions de longitud d'entrada:** La majoria dels models de PLN estan dissenyats per manejar seqüències d'entrada d'una longitud fixa, típicament 512 tokens. Els textos més llargs han de ser dividits en fragments més petits per ser processats adequadament, cosa que pot fer que es perdi context i es redueixi l'eficàcia del model en capturar les relacions a llarg termini dins del text.
 - **Complexitat del model:** Alguns enfocaments per manejar textos llargs impliquen l'ús de models més complexos o tècniques avançades com atenció jeràrquica o models amb mecanismes d'atenció a llarg termini. Aquests mètodes poden ser difícils d'implementar i entrenar.
 - **Cost computacional:** El processament de textos llargs requereix molts més recursos computacionals, el que pot resultar un obstacle per a moltes aplicacions.

El NLP juga un paper vital en la tecnologia i en la manera com els humans interactuen amb ella. Tot i tenir els seus reptes, s'espera que el NLP esdevingui més precís amb models més sofisticats, més accessible i més rellevant en nombroses indústries. El que està clar és que seguirà sent una part molt important tant de la indústria com de la vida quotidiana. [1]

3.3 Anàlisi de sentiments

3.3.1 Definició

L'anàlisi de sentiments és una tècnica de processament del llenguatge natural (NLP) utilitzada per determinar si les dades són positives, negatives o neutres.

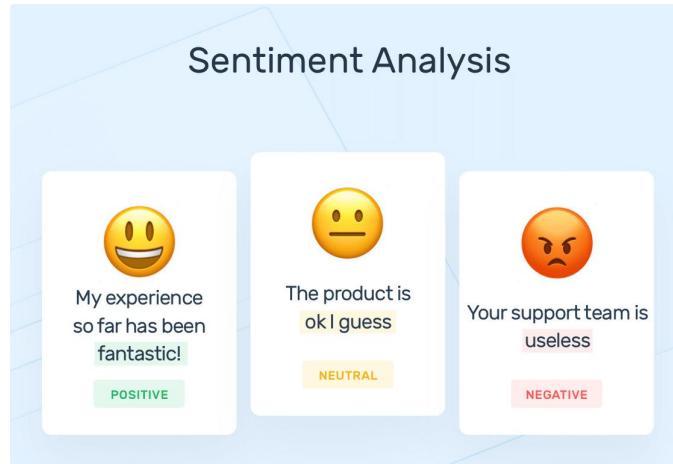


Figura 30: Exemple de classificació de tres comentaris segons el seu sentiment. Autor: MonkeyLearn. Font: [17]

3.3.2 Tipus

Alguns dels tipus més utilitzats d'anàlisis de sentiments són:

- **Anàlisi de sentiments gradual:** S'afegeixen diverses categories per incloure diferents nivells entre el positiu i el negatiu. Un exemple podria ser: molt positiu, positiu, neutral, negatiu o molt negatiu.
- **Detecció d'emocions:** La detecció d'emocions en l'anàlisi de sentiments permet anar més enllà de la polaritat per detectar emocions, com ara felicitat, frustració, ira i tristesa. Hi ha dues aproximacions a aquest problema, el primer és mitjançant una llista de paraules amb les emocions que representa cadascuna i el segon és utilitzant models complexos de ML.
- **Anàlisi de sentiments basat en aspectes:** se centra a identificar i analitzar els sentiments associats amb aspectes específics o característiques d'un producte, servei o situació. La diferència clau entre l'anàlisi de sentiments tradicional i l'anàlisi de sentiments basat en aspectes rau en el nivell de granularitat. Mentre que el sentiment analysis tradicional s'enfoca en el sentiment general d'un text, l'anàlisi de sentiments basat en aspectes desglossa aquest sentiment en aspectes específics i evalua els sentiments associats amb cadascun d'ells de manera individual.
- **Anàlisi de sentiments multilingüe:** és una tècnica que permet analitzar sentiments en textos que es troben en diversos idiomes. Aquesta tècnica és molt útil en contextos globals, com ara les xarxes socials, fòrums en línia, ressenyes de productes i altres tipus de dades textuales recopilades de diferents parts del món. [17]

3.3.3 Beneficis

Els beneficis generals de l'anàlisi de sentiments inclouen:

- **Classificació de dades a gran escala:** L'anàlisi de sentiments permet a les empreses gestionar grans volums de dades no estructurades de manera eficient i rendible.
- **Anàlisi en temps real:** L'anàlisi de sentiments és capaç d'identificar problemes crítics en temps real. Per exemple, pot detectar l'escalada d'una crisi de relacions públiques a les xarxes socials o la imminent desafecció d'un client insatisfet. Els models d'anàlisi de sentiments permeten identificar aquestes situacions immediatament, facilitant així la presa de mesures ràpides i efectives.

- **Criteris consistents:** L'ús d'un sistema centralitzat d'anàlisi de sentiments permet a les empreses aplicar criteris uniformes a totes les seves dades, la qual cosa contribueix a millorar la precisió i a obtenir coneixements més fiables i coherents.

3.3.4 Aplicacions

L'anàlisi de sentiments té múltiples aplicacions que poden beneficiar significativament les empreses en diversos àmbits. Una de les principals aplicacions és la **monitorització de xarxes socials**, on les empreses poden supervisar el sentiment general dels consumidors cap a la seva marca en plataformes com Twitter, Facebook i Instagram.

A més, l'anàlisi de sentiments s'utilitza en el **suport al client** per processar i interpretar les interaccions dels clients amb els serveis de suport. Aquesta eina ajuda a identificar ràpidament els problemes més comuns i els clients insatisfets, permetent a l'equip de suport prioritzar i resoldre les qüestions més crítiques amb major eficiència.

Una altra aplicació important és l'**escucha dels empleats**, que se centra a analitzar les enquestes de clima laboral, els comentaris i les discussions internes per entendre millor el sentiment dels empleats envers l'empresa. Mitjançant aquesta informació, les organitzacions poden abordar problemes interns, millorar l'ambient de treball i prendre mesures proactives per augmentar la motivació i la productivitat del personal.

L'anàlisi de sentiments també s'aplica a l'estudi de les ressenyes i els **comentaris sobre productes**. Aquesta informació és crucial per entendre les percepcions dels consumidors, identificar àrees de millora i desenvolupar noves característiques o productes que responden millor a les necessitats dels clients.

4 Context del problema

4.1 The Last of Us

4.1.1 Història i llançament del videojoc

The Last of Us, abreviat TLOU, és un videojoc d'acció i terror desenvolupat per Naughty Dog i publicat per Sony Computer Entertainment en exclusiva per a PlayStation 3. Es va llançar a la venda el 14 de juny de 2013.



Figura 31: Portada The Last of Us PS3. Imatge promocional. Autors: Naughty Dog

TLOU es desenvolupa en un món postapocalíptic devastat per una pandèmia que ha col·lapsat la societat. El joc està ambientat principalment als Estats Units, on ciutats i pobles han estat abandonats o destruïts, deixant els supervivents lluitant per la seva vida en un entorn desolat i perillós. L'escassetat de recursos, la presència de faccions hostils i la naturalesa que reclama el que és seu, caracteritzen l'atmosfera del videojoc.

La història segueix a Joel, un home marcat per la tragèdia, i Ellie, una adolescent valenta i decidida, mentre viatgen per un món infestat de fongs que converteixen els humans en criatures hostils. El seu objectiu inicial és lliurar a Ellie, qui sembla ser immune al virus, a un grup de científics que busquen trobar una cura. La relació entre els dos protagonistes evoluciona al llarg del joc, passant de ser desconeguts a formar un lligam paternal que impulsa gran part de la narrativa.

4.1.2 Sobre Naughty Dog i context de desenvolupament

Quant a l'estudi de desenvolupament de videojocs Naughty Dog, podem parlar de grans títols originals independents. Destaca la saga de Crash Bandicoot, que es va convertir en un top vendes de PlayStation 1 i la saga Jak and Daxter, per PlayStation 2. Però, sens dubte, el joc que marca un precedent dins de l'estudi i que després serviria com a base per construir TLOU és la saga Uncharted, per PlayStation 3.

La primera entrega d'aquest videojoc, 'Uncharted: Drake's Fortune', es va llançar el 2007 com un joc d'aventures a l'estil d'"Indiana Jones", amb un gran èmfasi en la narrativa i en les noves capacitats tècniques de la consola. Molts medis especialitzats, com Kotaku e IGN el nomenen joc de l'any.

Aquesta fórmula de joc lineal, d'aventures, amb una narrativa cinematogràfica, uns personatges icònics, una molt bona banda sonora i, en general, un nivell de qualitat excepcional, es va mantenint durant els següents anys, amb "Uncharted 2: Among Thieves" (2009) i "Uncharted 3: Drake's Deception" (2011).

Segons un article publicat a la revista especialitzada IGN [36], per primera vegada en la història de la companyia, l'estudi de Naughty Dog es divideix en dos equips, un desenvolupa Uncharted 3, mentre que l'altre comença a treballar en The Last of Us paral·lelament.

Naughty Dog amb tots aquests títols aconsegueix fer-se un renom dins la indústria dels videojocs i reuneix un gran nombre de fans que esperen amb expectació la sortida del nou títol de l'empresa americana.

Els desenvolupadors de TLOU, en el documental de “Grounded: The Making of The Last of Us” [37] expliquen que busquen una experiència diametralment oposada a Uncharted amb una jugabilitat realista, pausada i silenciosa que sorprengui l'audiència.

A més, el context en què es desenvolupa TLOU coincideix amb una creixent popularitat dels jocs de zombis. Això es pot veure amb altres companyies que en aquell moment exploten el gènere, com “Resident Evil” (Capcom), “Call of Duty: Black Ops” (Treyarch), “Dead Rising 2” (Capcom), entre altres, que aconsegueixen molt bones vendes.

4.1.3 Relevància en la indústria dels videojocs

The Last of Us va ser un joc revolucionari que va assumir grans riscos en el moment i el context del seu llançament. En l'article d'Edge Online [38], Bruce Straley, director del joc, parla del risc que va implicar incloure una nena a la portada d'un joc de més divuit anys i, al mateix temps, involucrar nens dins de la narrativa que es trobaven realment en perill o podien morir. Hi havia molt pocs exemples de grans produccions que haguessin fet això, però, segons el cineasta, encara que fos controvertit i dur, era part del món i de la història que intentaven narrar i era molt important mostrar-ho.

A més, també parla de la polarització que provoquen les decisions finals del guió d'aquest joc. Straley afirma que, abans de llançar el joc, van fer entrevistes per sondejar l'opinió de la gent i hi havia persones que deien: “M'encanta el joc, m'encanta la mecànica, m'encanta el combat, però heu d'arreglar el final”. No obstant això, el director de TLOU afirma que prefereix generar passió en els jugadors en lloc de fer un final feliç i rebre reaccions indiferents. The Last of Us deixa un final obert, lliure a interpretació. “Encara que no existís una segona part, la història ha concluït”, afirma Druckmann.

El cineasta, a un article de The Verge, [19] també diu que el seu objectiu era fer un joc on hi hagués una protagonista que fos molt memorable i no estigués hipersexualitzada. “Era una oportunitat per canviar la indústria”.



Figura 32: Concept art dels protagonistes. Authors: Naughty Dog. Font: The Last of Us Wiki [18]

TLOU va portar el gènere d'acció i aventures a noves cotes de qualitat amb el seu enfocament en la narrativa i el realisme emocional. Va introduir mecaniques de joc innovadores i un disseny de nivell que fomentava l'exploració i la furtivitat, la qual cosa va influir en altres jocs posteriors.

El joc va ser elogiat per la seva representació de personatges femenins, en particular Ellie, que és una de les poques protagonistes femenines en la indústria dels videojocs que és forta, complexa i multifacètica.

TLOU es va convertir en un fenomen cultural, generant discussions sobre temes com la moralitat, la supervivència i el lligam entre pares i fills. La seva influència es va estendre més enllà del món dels videojocs, amb adaptacions a altres mitjans com còmics i una sèrie de televisió en desenvolupament. [39]



Figura 33: De desconegeuts a família: el viatge de Joel i Ellie. Autors: Naughty Dog. Font: Metratge del joc.

4.1.4 Referències i inspiració d'altres treballs en la narrativa

Neil Druckmann, director creatiu del videojoc, explica a The Verge [19] que, en un inici, un dels seus grans referents per desenvolupar la història va ser la pel·lícula de *Night of the Living Dead* de Jorge A.

Romero, director consagrat per ser el creador del gènere de pel·lícules de zombis modernes.

En l'article, Druckmann reconeix que la història de TLOU es va cuinar a foc lent, durant deu anys i va ser a costa de prova i error fins que, finalment, l'estudi de Naughty Dog va donar amb la tecla.

En els vídeos promocionals del videojoc [40], Bruce Straley, director de joc, i Neil Druckmann, parlen de les referències i nomenen còmics, novel·les i pel·lícules que els van inspirar per desenvolupar el videojoc. Destaquen: *Children of men*, *The road*, *28 Days Later*, *Walking Dead*, *The Last Man*, entre altres.



Figura 34: Entorns i caracterització del videojoc - 1. Concept Art. Autors: Naughty Dog. Font: [19]

Quant als infectats, a un altre vídeo [41] expliquen que van ser inspirats per un segment del documental de natura de la BBC *Planet Earth* (2006), que presentava el fong *Cordyceps*. Tot i que aquest fong existeix a la vida real, només és capaç de prendre el control d'alguns insectes, però a TLOU exploten la idea de l'evolució del fong als éssers humans per donar sentit al videojoc.

En un altre article d'Edge Online [38], Bruce Straley, director del joc, afirma que el departament d'art es va basar en fotografies de Robert Polidori's de l'Huracà Katrina com a referència de les àrees inundades del videojoc.



Figura 35: Entorns i caracterització del videojoc - 2. Concept Art. Autors: Naughty Dog. Font: [19]

També en un article publicat a MMGN [42] el dissenyador de TLOU, Ricky Cambier, va citar els videojocs *Ico* i *Resident Evil 4* com a principals influències en el disseny del videojoc.

Com es pot observar, Naughty Dog pren referències de la vida real, de pel·lícules, sèries, còmics i jocs del gènere. Molta gent coincideix que Naughty Dog no ha inventat res nou amb TLOU, però reconeixen el gran mèrit d'agrupar totes les peces del trencaclosques i fer un gran joc amb carisma i una identitat pròpia que el fa diferent dels altres.

4.1.5 Recepció del mercat

En el seu llançament The Last of Us rep "aclamacions universals", segons el portal de ressenyes, Metacritic i es converteix en el cinquè joc millor valorat per a PlayStation 3. Els revisors elogien el desenvolupament dels personatges, la història i el missatge subjacent, el disseny visual i sonor. IGN va qualificar The Last of Us de "obra mestra" i "el millor exclusiu de PlayStation 3" i Andy Kelly, un reconegut periodista de videojocs, va declarar que era "el millor moment de Naughty Dog".

Quant a l'opinió pública, la majoria dels comentaris i crítics fan referència a aquests temes:

- És una opinió generalitzada que molta gent que ha jugat TLOU creu que és el *Game of The Year*.
- Entre tanta saturació de jocs de matar sense cap preocupació, és una renovació i un canvi d'aires la nova jugabilitat, amb la gestió de recursos i els enfrontaments molt tensos.
- La història és bastant simplista i el gènere zombis tampoc és original.
- El viatge dels personatges es fa una mica llarg i repetitiu en algunes parts del joc.
- La construcció dels personatges durant la història és molt bona.
- El final no deixa indiferent a ningú.
- Memorable banda sonora.

Per entendre el context del llançament també és molt útil la ressenya de TLOU de DayoScript a YT (2013).[43]

4.1.6 The Last of Us: Left Behind

4.1.6.a Història i llançament

The Last of Us: Left Behind va ser llançat a escala mundial el 14 de febrer de 2014 per a la PlayStation 3. Aquest contingut descarregable es va publicar per complementar la història principal i narra dues noves històries: els esdeveniments que van tenir lloc tres setmanes abans dels fets de The Last of Us, i els esdeveniments que es desenvolupen durant els capítols de tardor i hivern de la història principal.



Figura 36: Portada contingut ampliable The Last of Us: Left Behind. Imatge promocional. Autors: Naughty Dog.

En la primera història, tal com s'explica al vídeo de YouTube del canal oficial de PlayStation [44], Naughty Dog explora una forma de jugar i de redefinir que pot mostrar un videojoc. La trama se centra en conèixer millor la vida de la Ellie i explorar una faceta seva molt diferent de la que té amb el Joel. Això també es reflecteix en la jugabilitat, en contraposició al combat, per permetre als jugadors identificar-s'hi més. És tot molt més minimalistà i se simplifica molt les mecaniques del joc. “La relació d’Ellie amb Riley és molt interessant, perquè permet veure qui ha influït a la Ellie en ser la persona que és” explica Jacob Minkoff, dissenyador principal de Naughty Dog.

En la segona història, Ellie cerca subministraments mèdics per guarir Joel en un centre comercial abandonat de Colorado, mentre enfronta diversos enemics. Les dues històries s’alternen en el temps i es complementen per generar contrastos.

4.1.6.b Recepció del mercat

The Last of Us: Left Behind va ser el contingut descarregable (DLC) més ben valorat de tota la plataforma PlayStation 3. [45]. L’ampliació va afegir una nova perspectiva a la història principal del joc, centrant-se en el personatge d’Ellie i la seva relació amb la seva amiga Riley. Els crítics van elogiar la narrativa, el desenvolupament del personatge i la jugabilitat del DLC.

Quant a l’opinió pública, la majoria dels comentaris i crítics fan referència a aquests temes:

- És una opinió generalitzada que al públic li va agradar aquest contingut addicional.
- Es manté la consistència en la qualitat de la jugabilitat i s’afegeixen noves mecaniques que enriqueixen l’experiència.
- Els jugadors van apreciar l’oportunitat de submarir-se més en la història d’Ellie i d’explorar el seu passat.
- Left Behind és una de les primeres grans superproduccions que inclou al collectiu LGTB de manera significativa en la trama.

4.2 The Last of Us: Part II

4.2.1 Història i llançament del videojoc

The Last of Us Part II, també abreviat TLOU2, és la continuació de l'acamat joc d'acció i aventura The Last of Us (2013). Desenvolupat per Naughty Dog i publicat per Sony Interactive Entertainment en exclusiva per a PlayStation 4, es va llançar a la venda el 19 de juny de 2020.



Figura 37: Portada The Last of Us: Part II. Imatge promocional. Autors: Naughty Dog.

Ambientat cinc anys després de The Last of Us, el joc se centra en dos personatges jugables en un Estats Units postapocalíptic les vides dels quals s'entrellacen: Ellie, que busca venjança després de patir una tragèdia, i Abby, una soldada que es veu immersa en un conflicte entre la seva milícia i un culte religiós.

4.2.2 Sobre Naughty Dog i context de desenvolupament

El desenvolupament de The Last of Us Part II va començar el 2014, poc després del llançament de The Last of Us Remastered. Neil Druckmann, després de dirigir "Uncharted 4", va tornar com a director creatiu, coescrivint la història amb Halley Gross.

Segons documents de l'editor Sony Interactive Entertainment, el desenvolupament de 70 mesos va arribar al seu punt màxim amb 200 empleats a temps complet i va costar al voltant de 220 milions de dòlars, convertint-lo en un dels videojocs més cars de desenvolupar de la història. [46]

Naughty Dog va empènyer les capacitats tècniques de la PlayStation 4 al màxim en crear la Part II, afegint més enemics i entorns més grans que en els jocs anteriors. Druckmann va assenyalar que qualsevol disminució en el detall arruïnaria el sentit d'autenticitat, cosa que requeriria una optimització constant de la tecnologia.



Figura 38: Les cinemàtiques del videojoc s'apropen tècnicament al fotorealisme. Autors: Naughty Dog. Font: Metratge del joc

També es va augmentar les opcions d'accessibilitat introduïdes en Uncharted 4 per assegurar que tots els jugadors poguessin completar la història, i els desenvolupadors van assistir a conferències i van treballar amb advocats.

En l'article de GQ Magazine [47] es narra molts dels moments claus del desenvolupament. En ell s'explica com Naughty Dog també es va enfocar a un repte sense precedents: haver de fer front a la pandèmia de la Covid-19. Es va imposar un retard indefinit en la data de llançament que, segons Druckmann, va suposar un cop devastador per a l'equip. Tots els 350 desenvolupadors havien dedicat una part significativa de les seves vides a aquest projecte, havent-hi guardat els secrets als seus amics i familiars, i ara havien de completar el desenvolupament des de casa. A més, hi havia milions de dòlars en joc. El director també destaca la coincidència de llançar un joc que tracta sobre un virus apocalíptic durant la pandèmia més letal en un segle.

A més, l'estudi va intentar mantenir el mateix nivell d'expectació que el primer joc. Es van revelar estratègicament tràilers de la seqüela on semblava que la història seria una altra i el misteri envoltant la història d'Ellie i Joel va alimentar la curiositat i les discussions entre els fans. [48]

El 27 d'abril, més d'una hora de videojoc i diversos moments clau de la història es van filtrar. La informació incompleta, l'homofòbia i la transfòbia es van estendre al voltant de la discussió de la Part II. El període següent va ser un dels pitjors en la història de Naughty Dog. Druckmann reconeix que va rebre una tempesta de missatges antisemites en les setmanes següents.

Després del retard indefinit a causa de la pandèmia, The Last of Us Part II finalment es va llançar al mercat, tres mesos després, el 19 de juny de 2020.

4.2.3 Rellevància en la indústria dels videojocs

Neil Druckmann explica en GQ Magazine [47] que The Last of Us Part II és el joc AAA més divers mai fet abans. Ellie és una protagonista obertament lesbiana, però també hi ha personatges trans i de minories ètniques en rols principals, portant els límits de la representació més enllà de qualsevol precedent. “En un mitjà sovint tan anti progressista com aquest, el joc sembla ser un gran gest de rebel·lia”, explica el cineasta. Fins i tot una escena de sexe ‘elegant’, una raresa en els videojocs.

Aquest enfocament en la diversitat i l'autenticitat va establir un precedent dins de la història dels videojocs i la representació de personatges. [49]

El joc també presenta decisions morals i ètiques que confronten al mateix jugador i fan plantejar-li preguntes sobre si són correctes les accions preses pels personatges.

A més, es presenten diferents faccions amb ideologies i objectius diferents que reflecteixen els conflictes polítics i socials que poden esdevenir-se en una situació de crisi. El control de la informació i manipulació del poble són molt notables en funció d'en quin bàndol es troba el jugador i permeten veure diferents punts de vista de la història i empatitzar amb els personatges, establint paralel·lismes entre ficció i realitat.



Figura 39: El joc mostra diversos punts de vista del conflicte. Autors: Naughty Dog. Font: Metratge del joc.

TLOU: Part II va causar un gran impacte cultural. Va generar una gran quantitat de discussions i debats dins de la comunitat de jugadors i més enllà. La seva narrativa i temes provocatius, així com les seves decisions de disseny innovadores, el van convertir en un tema de conversa durant molt temps.

The Last of Us: Part II va ser l'exclusiu més venut de PS4.

4.2.4 Referències i inspiració d'altres treballs en la narrativa

Els temes de venjança i retribució en la narrativa van ser inspirats per les experiències de Druckmann en créixer a Israel, on la violència era un tema freqüent. Altres temes que s'inclouen són el tribalisme, el trauma i la cerca de la justícia. [50]

El cineasta també explica que la seqüela està clarament inspirada, en “El Padrí” de Francis Ford Coppola. [47]

L'amor i relacions humanes continuen sent temes molt importants en aquesta segona part. Halley Gross va dirigir moltes escenes romàntiques segons s'explica a GQ Magazine i va aportar un punt de vista diferent de la visió de Drukmann [46].

Pel que fa al món The Last of Us, expandeix el què ja era, però mantenint la coherència amb la primera part. En una part del joc es pot trobar una referència a *City Of Thieves* novel·la de David Benioff. [51]

Els artistes de Naughty Dog van viatjar a Seattle per analitzar l'arquitectura, la vegetació, els materials, la topografia, l'il·luminació i capturar textures fotorealistes. [52]



Figura 40: Reconstrucció de la ciutat de Seattle en un entorn postapocalíptic. Autors: Naughty Dog. Font: Metratge del joc.

4.2.5 Recepció del mercat

La Part II va rebre elogis per la seva jugabilitat, disseny de so, banda sonora, interpretacions, personatges i fidelitat visual, tot i que la seva narrativa i temes van dividir als medis especialitzats i als jugadors.

Aquest últim factor va ser decisiu en el seu llançament i va suposar un bombardeig de ressenyes negatives a Metacritic, amb una puntuació de 5,7 basada en més de 150.000 comentaris d'usuaris.

Algunes de les crítiques que se li van fer:

1. La inclusió de personatges LGBTQ+ en rols principals i l'exploració de relacions homosexuals en el joc van ser objecte de crítiques per part d'alguns sectors conservadors.
2. Les decisions controvertides preses pels personatges principals, així com el desenvolupament de la trama en certs punts del joc, van generar debats entre els jugadors i crítiques sobre la coherència de la història i la motivació dels personatges.
3. La necessitat de defensar-se de gossos entrenats i la violència i duresa de les escenes. Aquest fet va ser molt controvertit i considerat problemàtic per alguns jugadors i defensors dels drets animals.
4. La quantitat i la intensitat de la violència gràfica en el joc van ser objecte de discussió, amb alguns considerant que creuava límits ètics i altres elogiant el seu realisme i la immersió en el món de ficció.
5. La duració del joc. Algunes persones també es van queixar respecte a la duració de la Part II i qüestionaven si realment era necessari que fos tan llarga. DayoScript va titular l'anàlisi del videojoc com una “ambició desmesurada”.

Segons paraules d'aquest analista: “The Last of Us: Part II no és una obra complaent, no existeix per fer sentir bé als fans sinó per confrontar-los al que van jugar el 2013. L'obra està per desafiar el que el públic esperava que fos. Criticar aquestes idees pot conduir a una reacció molt agressiva, que és el que ha passat, però és això, el que ho fa precisament més necessari.” [53] [54]

5 Estudi preliminar del conjunt de dades

5.1 Informació sobre el dataset

Nom del projecte: The Last of Us Reviews

Enllaç al projecte: <https://www.kaggle.com/datasets/lazaro97/the-last-of-us-reviews>

Quantitat de dades no repetides: 36021 valors

1. critic_reviews_g1u.csv: 168 valors
2. critic_reviews_g2.csv: 130 valors
3. critic_reviews_lb.csv: 69 valors
4. user_reviews_g1u.csv: 4873 valors
5. user_reviews_g2u.csv: 30655 valors
6. user_reviews_lb.csv: 185 valors

Mida del conjunt de dades: 32.9 MB

Variable objectiu: Puntuació (score)

Variables del sistema: Id, Review, Review Type, Views, Votes, Date, Language, Score, Platform, Split

Inspecció de les variables:

1. Id:
 - Descripció: Nom d'usuari del jugador.
 - Tipologia: Variable independent, categòrica nominal, inherent del dataset.
 - Implicació: Cada Id és únic i ens permet diferenciar als jugadors.
2. Review
 - Descripció: La ressenya de l'usuari.
 - Tipologia: Variable textual, categòrica nominal, inherent del dataset.
 - Implicació: Ens dona la informació útil de l'opinió del client.
3. Review Type
 - Descripció: La llargada de la ressenya. ‘Expanded’ (més extensa, detallada) o ‘Normal’ (més concisa, breu).
 - Tipologia: Variable dependent de ‘Review’, categòrica nominal, inherent del dataset.
 - Implicació: Pot servir per classificar les ressenyes.
4. Views
 - Descripció: Quantitat de persones que han vist la ressenya.
 - Tipologia: Variable independent, numèrica discreta, inherent del dataset.
 - Implicació: Com més views més probable és que la gent voti el comentari.
5. Votes
 - Descripció: Quantitat de vots que els jugadors han donat a la ressenya.
 - Tipologia: Variable dependent de ‘Views’ i ‘Review’, numèrica discreta, inherent del dataset.
 - Implicació: Si una ressenya té molts vots vol dir que més gent està d'acord amb el comentari i s'hauria de tenir més en compte.
6. Date
 - Descripció: Quan la ressenya va ser publicada.

- Tipologia: Variable independent, temporal, numèrica discreta, inherent del dataset.
- Implicació: Necessaria per valorar el context del comentari en el temps.

7. Language

- Descripció: Idioma utilitzat en la ressenya.
- Tipologia: Variable dependent de ‘Review’, categòrica nominal, inherent del dataset.
- Implicació: L’idioma ens pot donar una idea del perfil d’usuari que ha jugat el joc.

8. Score

- Descripció: Puntuació assignada per l’usuari.
- Tipologia: Variable independent, numèrica discreta, inherent del dataset.
- Implicació: Gran indicatiu de si li ha agradat o no el producte al jugador.

9. Platform

- Descripció: En quina plataforma el joc va ser jugat.
- Tipologia: Variable independent, categòrica nominal, inherent del dataset.
- Implicació: La plataforma té una importància moderada, ja que en el remaster de PS4 es van afegir algunes millors del joc original de PS3.

10. Split

- Descripció: Distribució recomanada per fer un test/train del model.
- Tipologia: Variable independent, categòrica nominal, ampliada respecte al dataset original.
- Implicació: Pot ser útil per fer un primer estudi del dataset.

Potencials desavantatges:

- Quant a TLOU i Left Behind: El dataset a analitzar pot tenir una sobrerepresentació de comentaris positius.
- Quant a Left Behind: potser no seran suficients les ressenyes del dataset.
- Alguns datasets presenten Id’s nuls.
- En alguns datasets el score va de 0 a 10 mentre que altres de 0 a 100.
- No tots els datasets tenen el mateix nombre de variables. Els que són de ressenyes de medis especialitzats només tenen: Id, Review, Date, Score.
- Podria donar-se el cas que un jugador hagi fet més d’una ressenya o una per cada joc, per tant, el Id no és únic per cada ressenya.
- Existeix una sobrerepresentació de comentaris anglesos.
- A les ressenyes de TLOU: Left Behind també hi ha alguna ressenya amb el camp ‘Review’ a null.

5.2 Hipòtesis inicials

Relacionades amb The Last of Us

- La percepció del videojoc ha millorat amb el temps, reflectint una major apreciació per la seva narrativa i jugabilitat.
- La percepció de la jugabilitat en “The Last of Us” afecta significativament la qualificació general dels comentaris, fins i tot en un joc on la narrativa és el focus principal.
- Els temes representatius en els comentaris negatius inclouen problemes amb les mecàniques del joc, errors i dificultats tècniques.
- La sortida de The Last of Us: Part II va suposar un increment en la quantitat de jugadors en la primera part, a causa de la continuïtat directa de la història.

Relacionades amb TLOU: Left Behind

- La naturalesa del producte com a ampliació del primer joc influeix en les ressenyes. Hi haurà més comentaris positius, ja que qui compra el DLC ja li va agradar el joc original.
- La inclusió de personatges LGTB afecta la percepció dels jugadors i la puntuació del videojoc.
- La jugabilitat alternativa centrada en els personatges i no tant en el gènere d'acció i terror rep més elogis que crítiques.
- La brevetat del contingut addicional pot influir en la percepció dels jugadors sobre la qualitat del DLC.

Relacionades amb TLOU: Part II

- La mort d'un personatge principal a l'inici de la història repercutex negativament en les ressenyes.
- Amb el pas del temps, l'opinió pública de The Last of Us: Part II ha millorat positivament.
- L'idioma dels comentaris pot influir significativament en les tendències de mercat i en les opinions expressades sobre el joc.
- Els usuaris tendeixen a valorar més la narrativa d'amunt del criteri tècnic.
- Les visualitzacions i vots dels comentaris de Metacrític estan estretament relacionats amb l'opinió general del joc.

5.3 Preprocessament i visualització de les dades

5.3.1 The Last of Us (2013) i The Last of Us Remaster (2014)

5.3.1.a Ressenyes dels crítics

Aquest apartat fa referència a la visualització de les dades del dataset *critic_reviews_g1u.csv* que és el més complet i una extensió del dataset original *critic_reviews_g1.csv*.

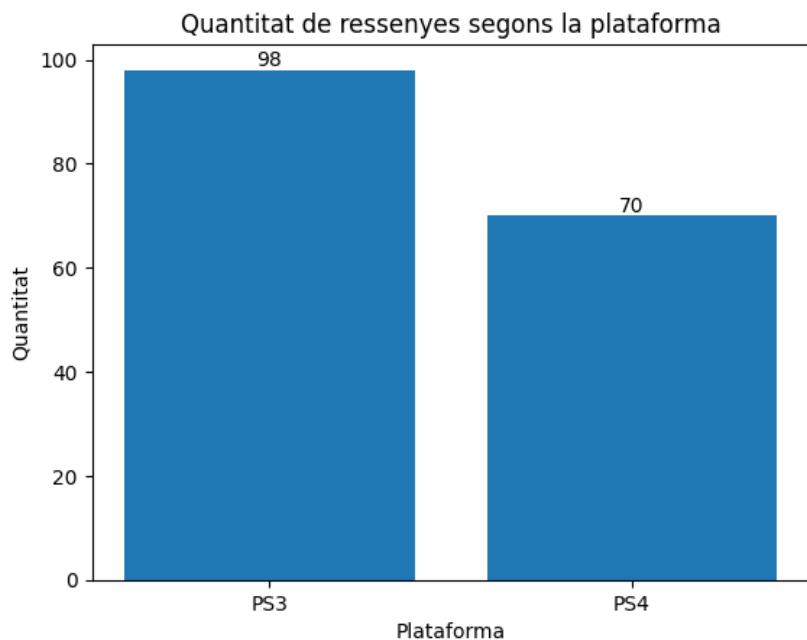


Figura 41: Gràfic de barres de ressenyes especialitzades per plataforma

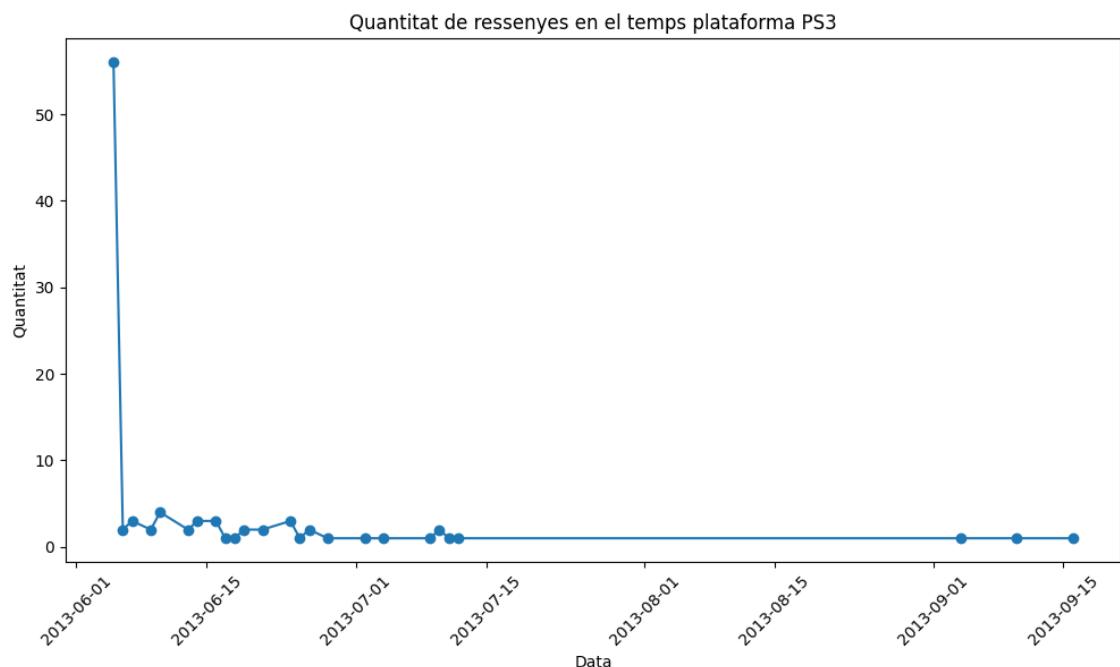


Figura 42: Gràfic de línies de les ressenyes especialitzades en el temps PS3

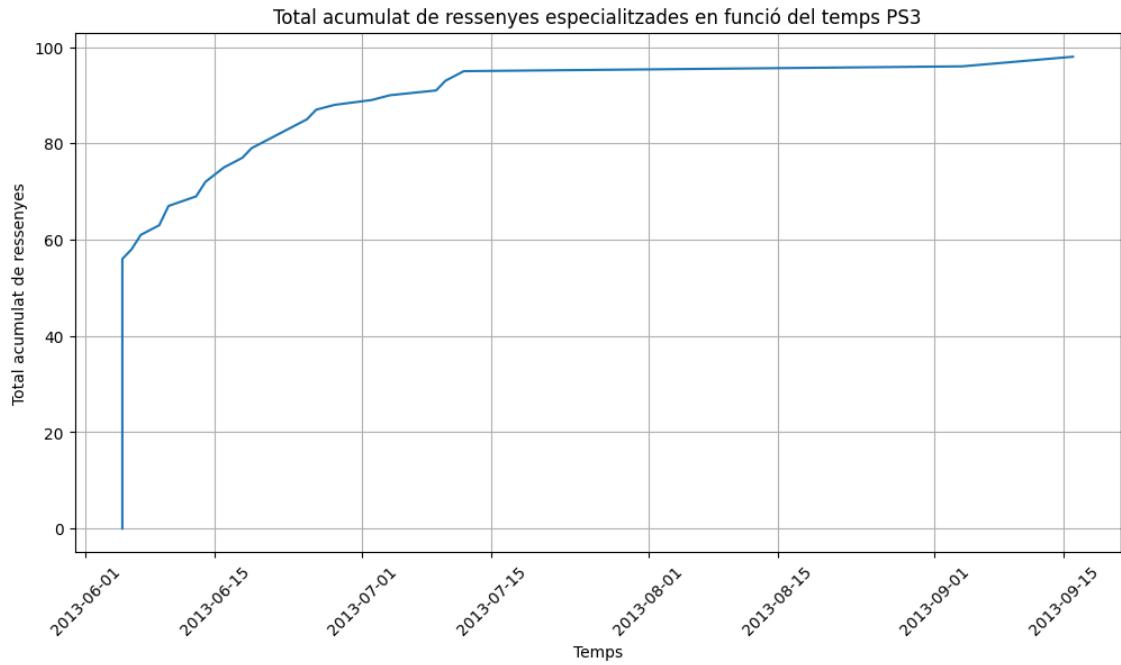


Figura 43: Gràfic de línies de les ressenyes especialitzades acumulades en el temps PS3

Es pot observar com la gran majoria de ressenyes, com és habitual en aquest tipus de productes, es fan el primer mes des de la sortida del videojoc al mercat. El primer dia hi ha un pic de ressenyes també pel fet que els medis especialitzats generalment són proveïts amb el videojoc setmanes abans i ja tenen preparada la ressenya pel dia de sortida.

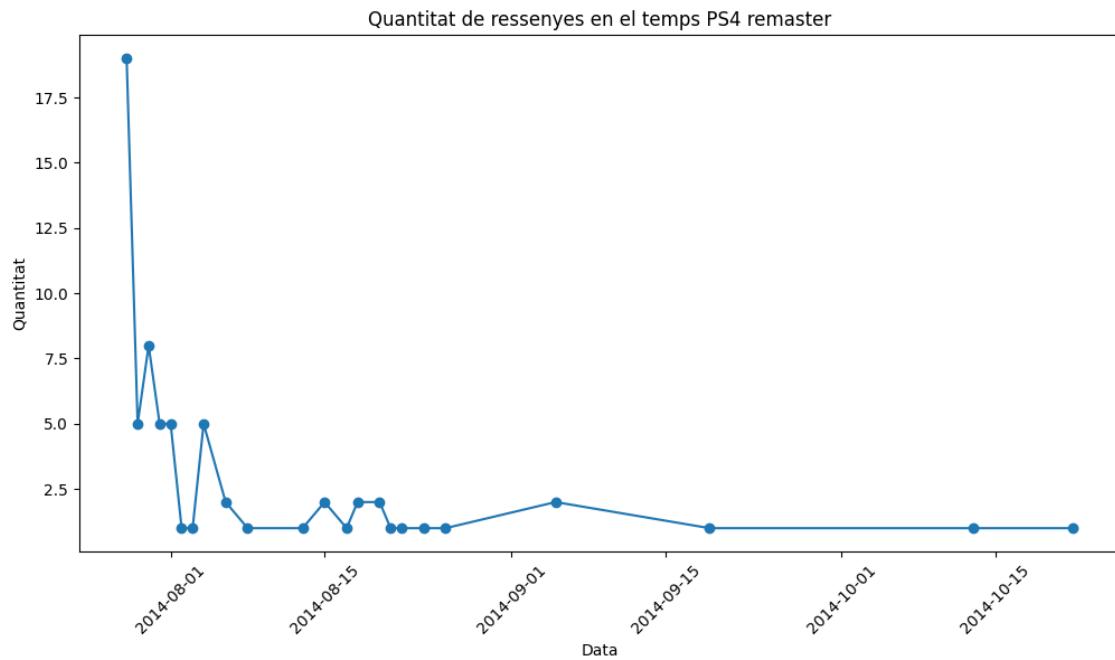


Figura 44: Gràfic de línies de les ressenyes especialitzades en el temps PS4 Remaster

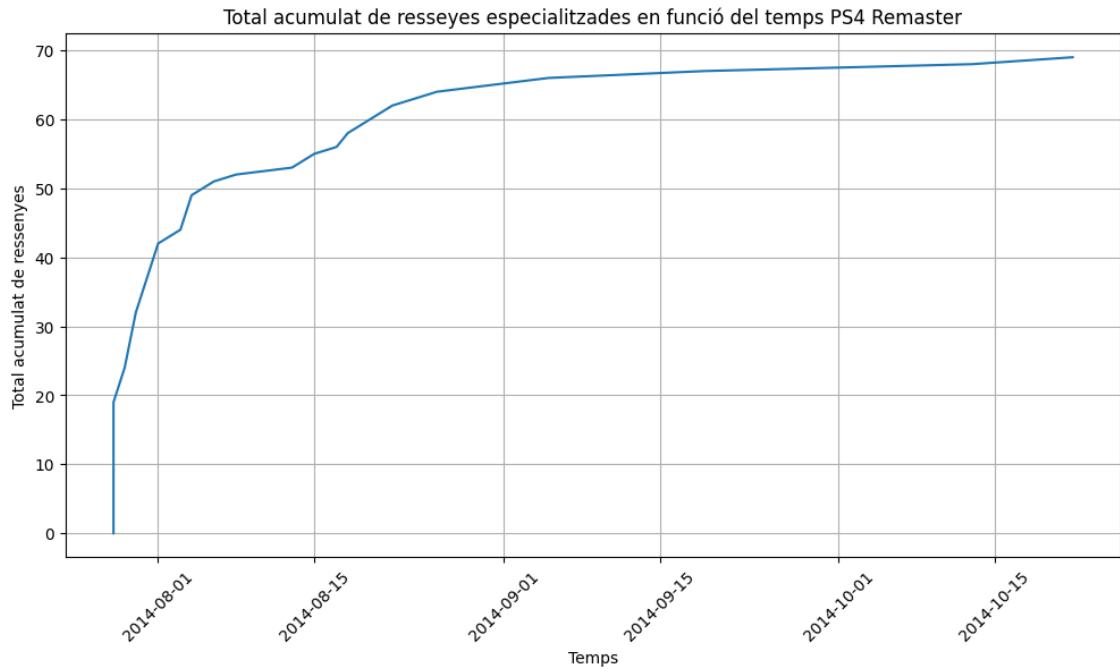


Figura 45: Gràfic de línies de les ressenyes especialitzades acumulades en el temps PS4 Remaster

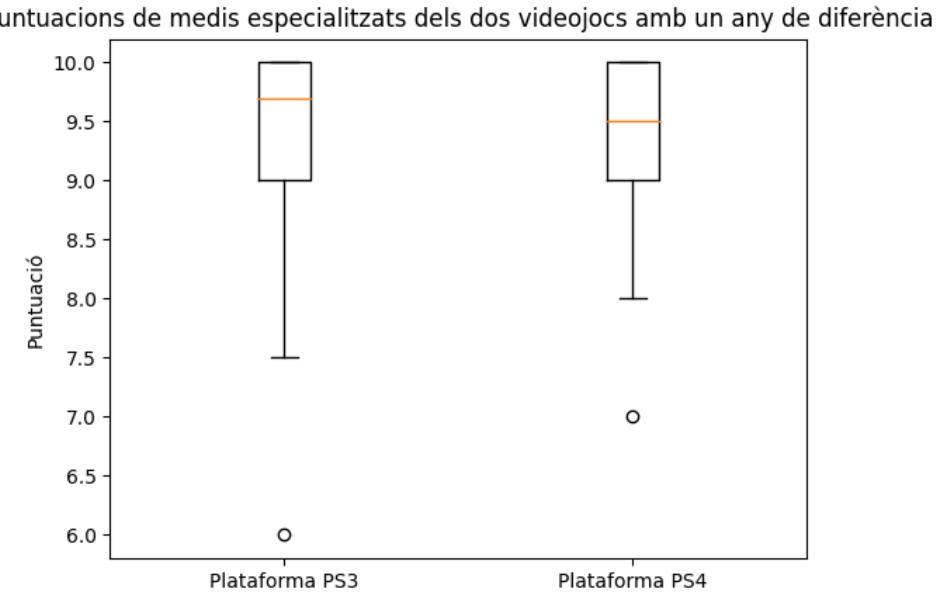


Figura 46: Gràfic de capsas de les puntuacions especialitzades rebudes d'ambdós videojocs

La forma de representar les qualificacions en el cas dels medis especialitzats és diferent de la dels usuaris, ja que en un cas es permeten valors decimals i en l'altre enters, respectivament.

La variable score relacionada amb la plataforma és interessant analitzar-la per veure si el salt generacional entre el joc original (PS3) i el remaster (PS4) ha influït significativament en les puntuacions.

Es pot observar que el joc remasteritzat, llançat un any més tard, obté puntuacions similars al joc original. Tot i això, en aquesta segona edició s'incorpora el contingut addicional 'Left Behind', que abans era de pagament, així com millores en el rendiment i els aspectes visuals.

5.3.1.b Ressenyes dels usuaris

Aquest apartat fa referència a la visualització de les dades del dataset *user_reviews_g1u.csv* que és el més complet i una extensió del dataset original *user_reviews_g1.csv*.

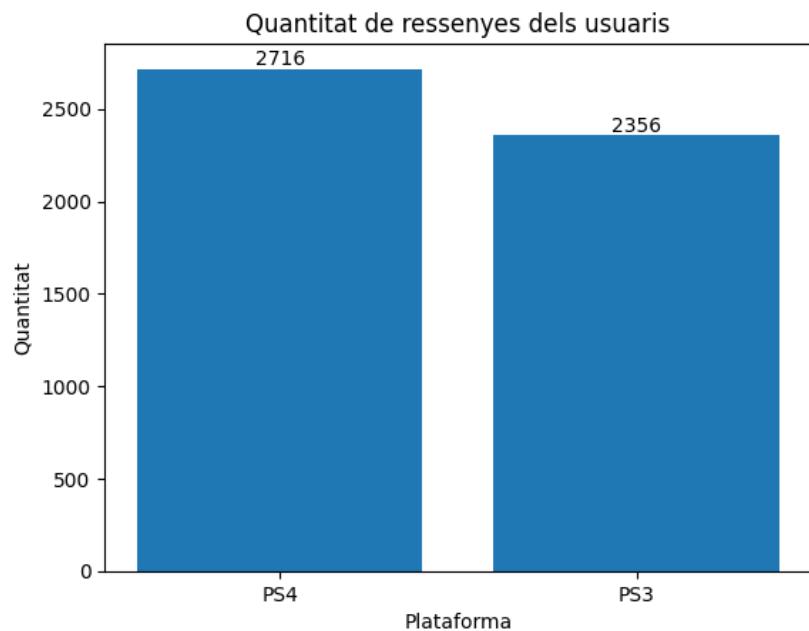


Figura 47: Gràfic de barres de ressenyes d'usuaris per plataforma

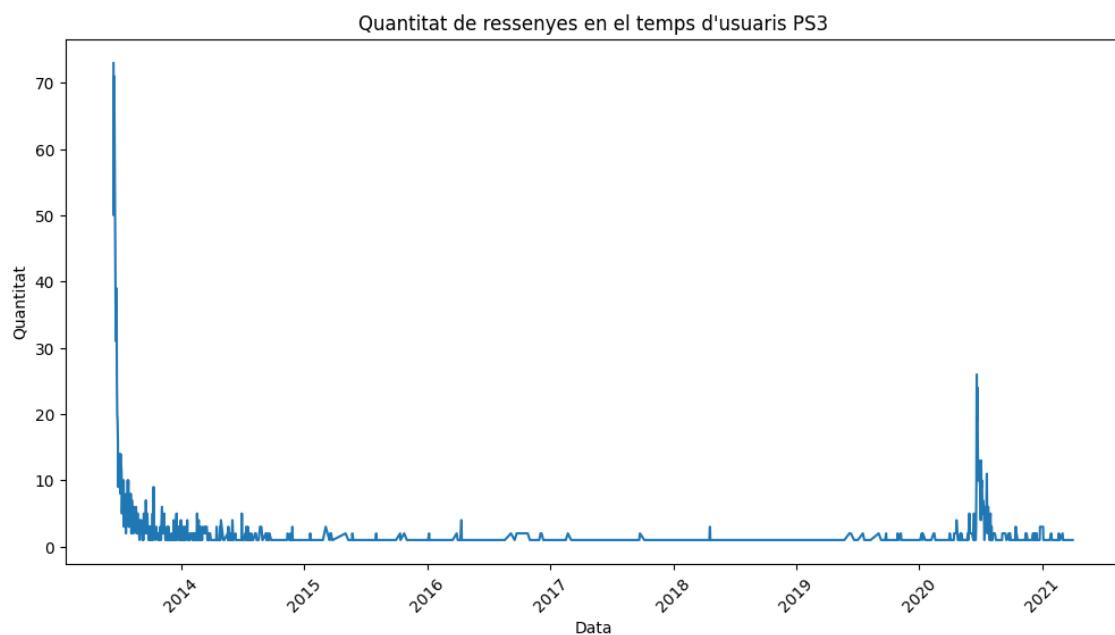


Figura 48: Gràfic de línies de les ressenyes d'usuaris en el temps PS3

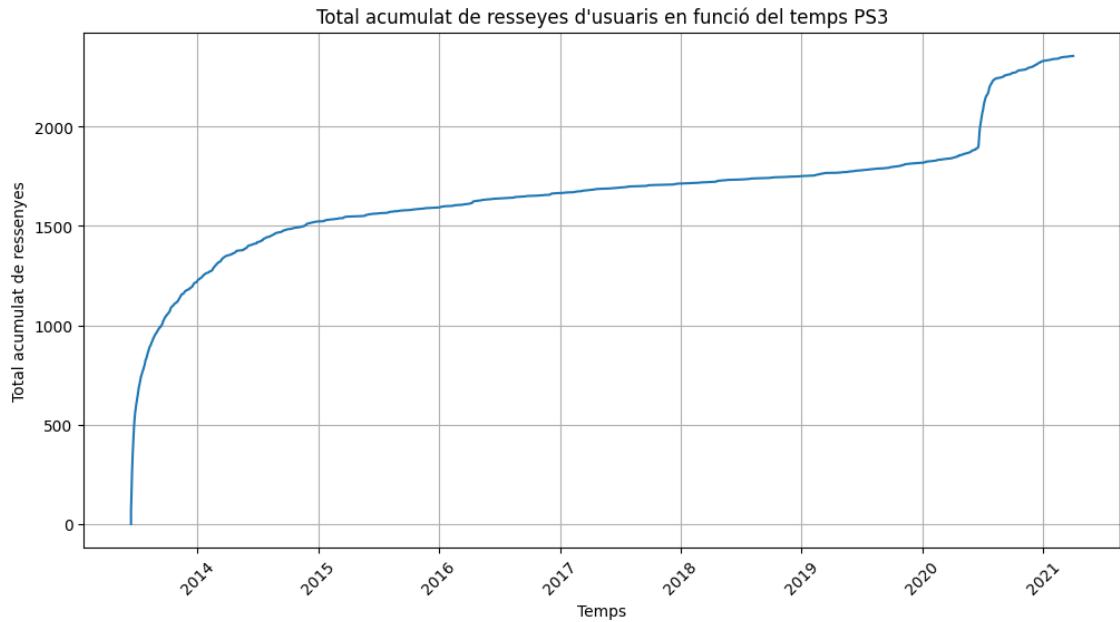


Figura 49: Gràfic de línies de les ressenyes d'usuaris acumulades en el temps PS3

En ambdós gràfics es pot observar un augment de les ressenyes a meitats del 2020 que coincideix amb el llançament del The Last of Us: Part II, el 19 de juny d'aquell any, que va incentivar als usuaris a tornar a jugar també la primera part. El mateix es pot observar en els gràfics referents a la versió de PS4.

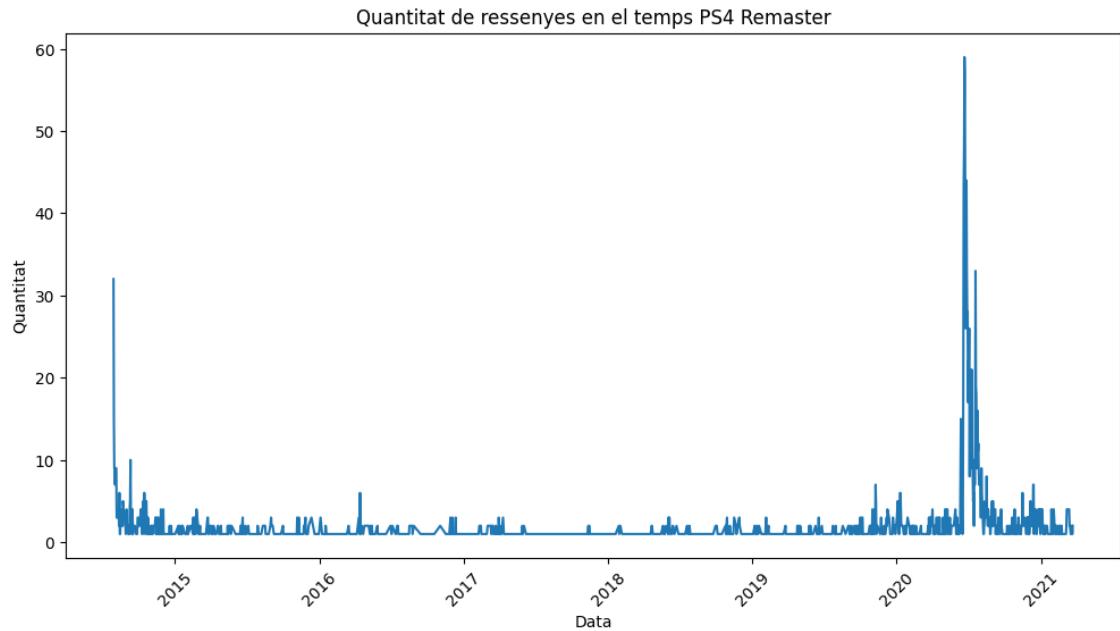


Figura 50: Gràfic de línies de les ressenyes d'usuaris en el temps PS4

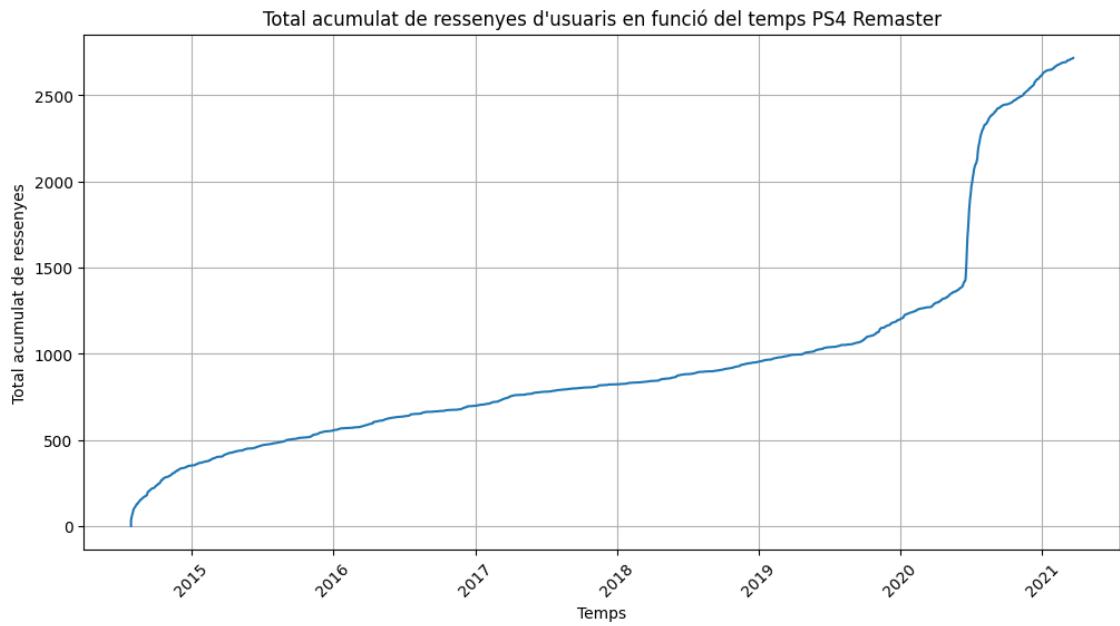


Figura 51: Gràfic de línies de les ressenyes d'usuaris en el temps PS4

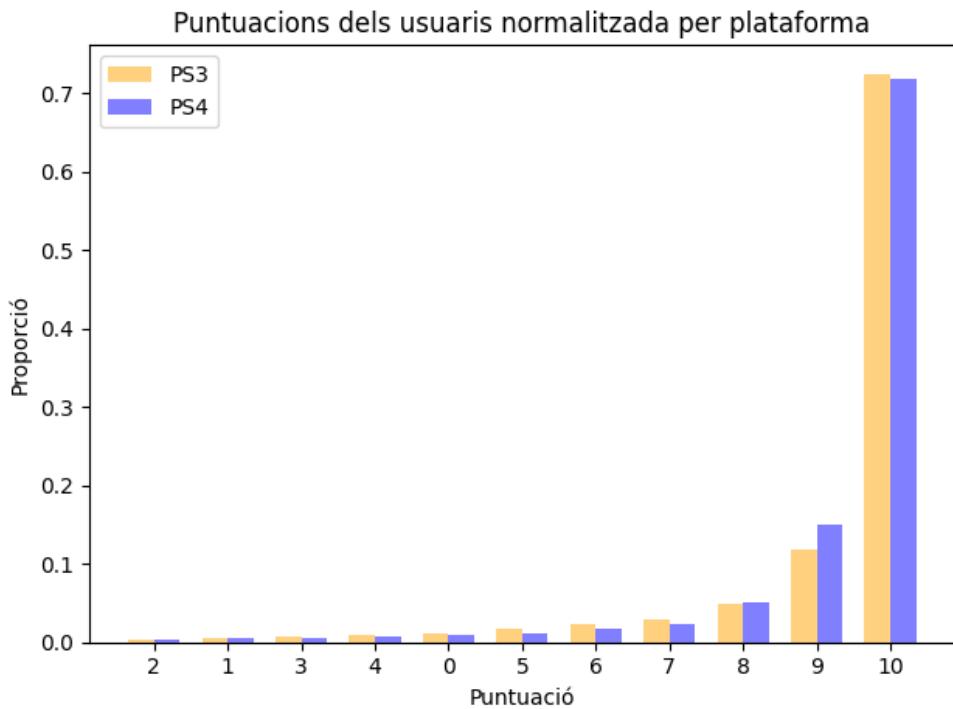


Figura 52: Gràfic de barres agrupades de la puntuació dels usuaris normalitzada per plataforma

No es nota una diferència significativa entre les puntuacions dels dos videojocs amb un any de separació. Sembla que l'opinió es manté en el temps.

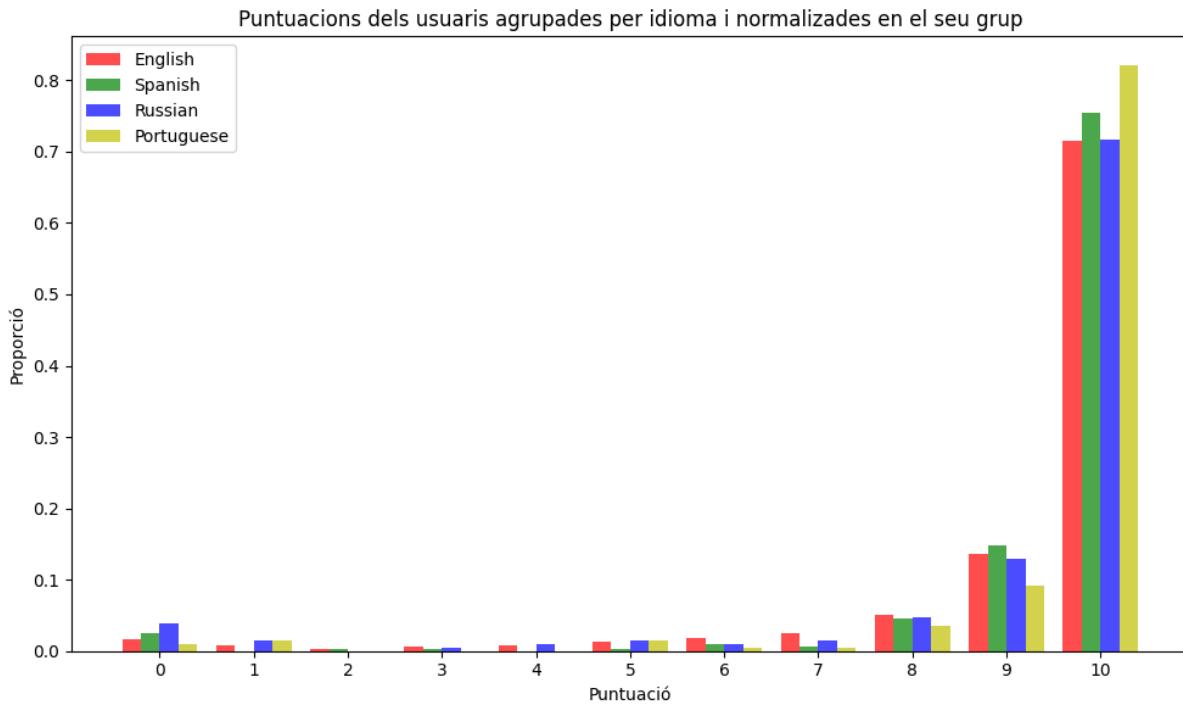


Figura 53: Gràfic de barres de la puntuació dels usuaris agrupada per idioma i normalitzada en el seu grup

Quant a l'idioma de les ressenyes, s'analitza els que tenen més representació (almenys 200 comentaris) i es mira si existeix alguna correlació d'aquesta variable amb la puntuació del videojoc. Com es pot observar, s'obtenen resultats bastant similars. Sembla que ni el doblatge ni altres factors externs alteren l'estadística global.

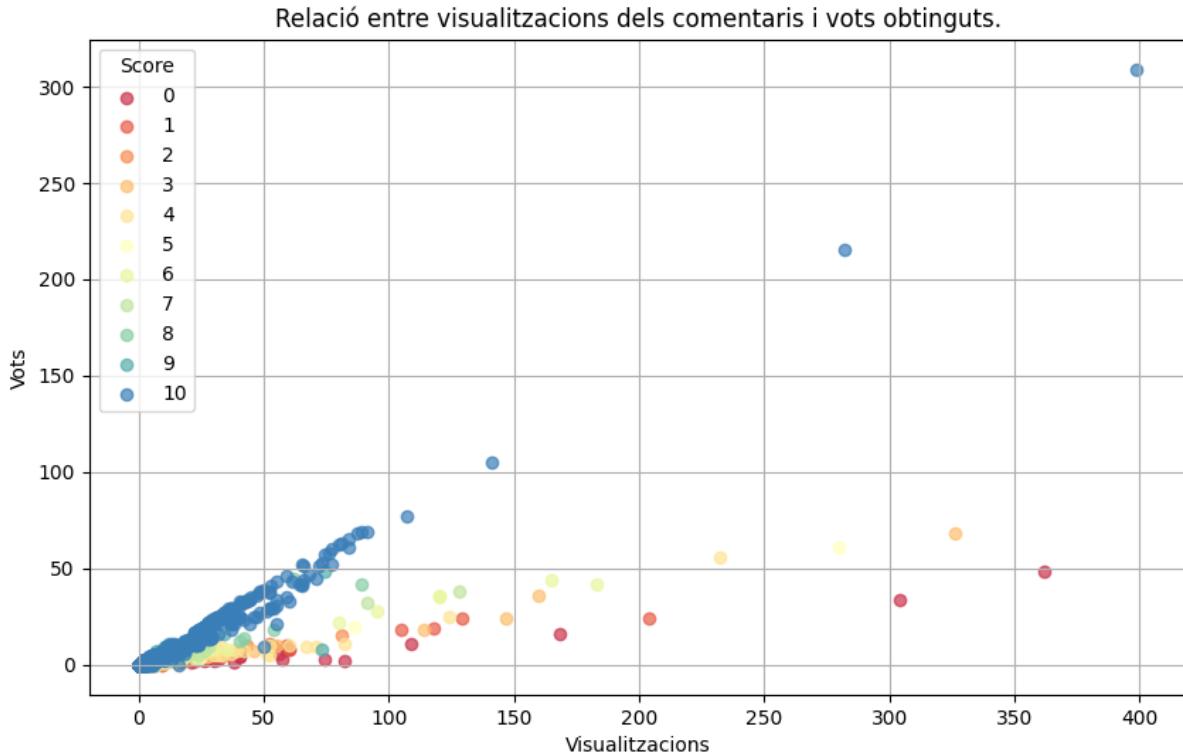


Figura 54: Gràfic de dispersió de *views/votes* de les ressenyes d'usuaris, acolorit per *score*

Si s'analitzen els vots dels comentaris dels usuaris, es pot observar com els comentaris que més suport reben són els que valoren amb notes més altes al videojoc.

Segons DayoScript, en la indústria dels videojocs, una superproducció d'aquest estil generalment parteix d'una nota base que és un 8 en el cas de les ressenyes crítiques. És a dir, un joc es considera "recomanable" quan almenys té aquesta puntuació de base.

En aquest cas, el que podem avaluar són els comentaris dels usuaris, tot i que podem seguir el mateix criteri. Les ressenyes es divideixen en dos grups: "recomanable" ($\text{score} \geq 8$) i mediocre ($\text{score} < 8$).

Si es fa la mitjana entre els vots de cada grup s'obté:

- 1 vot per cada 1614 visualitzacions en les ressenyes amb valoracions 'recomanables'
- 1 vot per cada 4915 visualitzacions en les ressenyes amb valoracions 'mediocres'.

En altres paraules, un comentari que atorga una bona nota al videojoc és tres vegades més votat que un que el puntuà per avall d'aquest llindar.

En aplicar una regressió lineal als dos grups, es pot observar gràficament aquesta tendència.

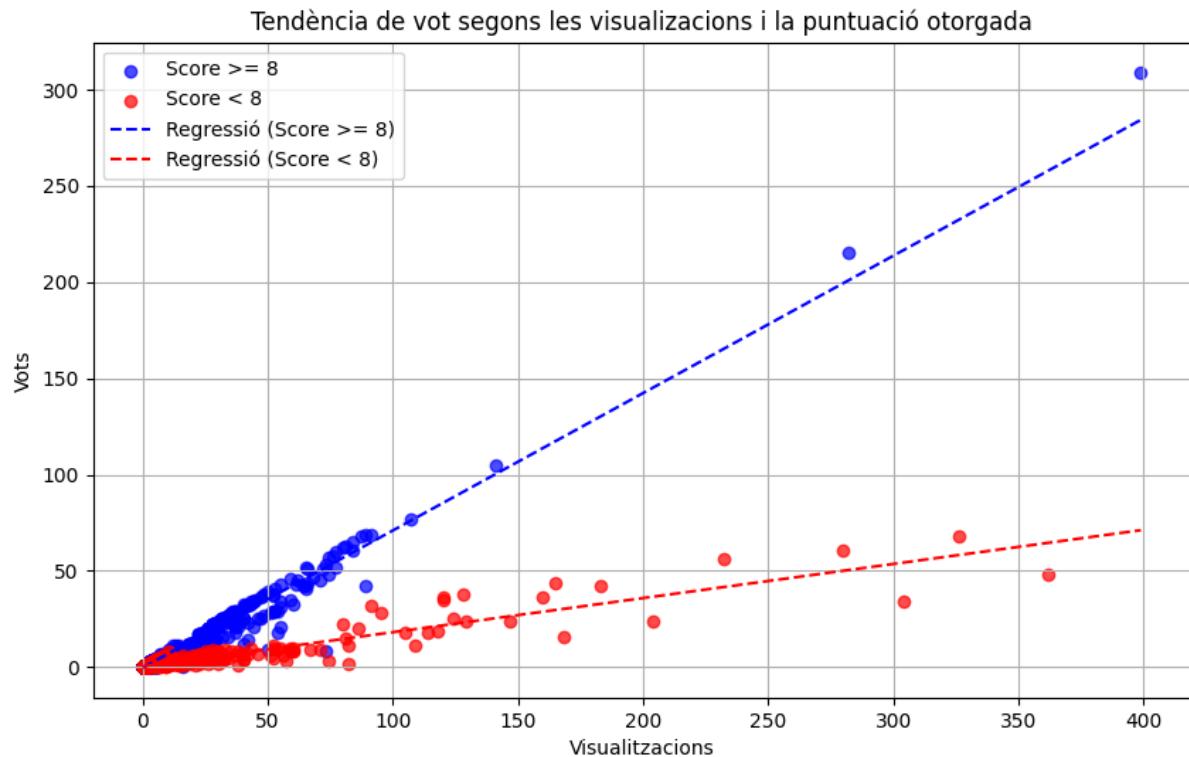


Figura 55: Regressió lineal per als dos subgrups en l'escala vots/visualitzacions

5.3.2 The Last of Us: Part 2 (2020)

5.3.2.a Ressenyes dels crítics

En aquest cas, dataset conté 121 ressenyes de medis especialitzats per la plataforma de PS4.

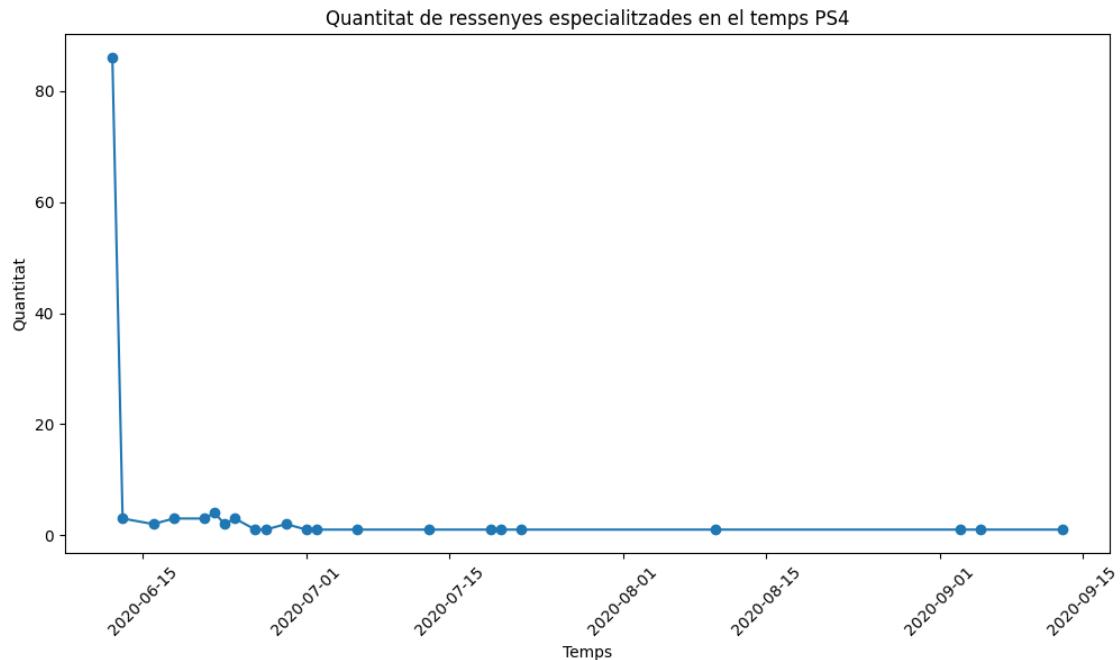


Figura 56: Gràfic de línies de les ressenyes especialitzades en el temps PS4

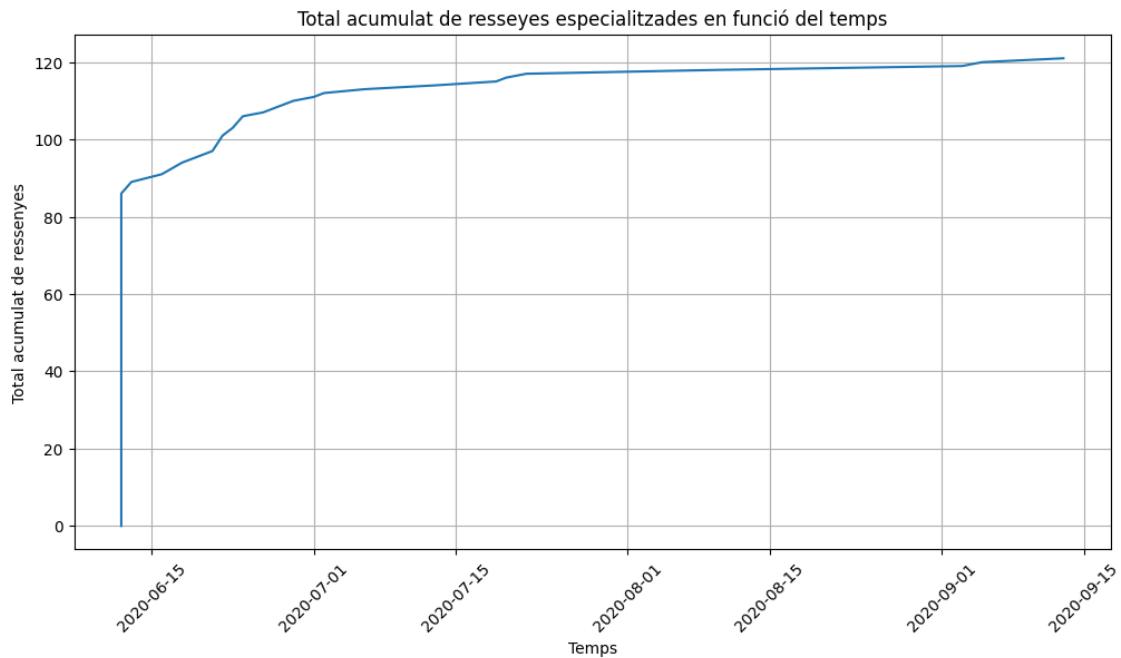


Figura 57: Gràfic de línies de les ressenyes especialitzades acumulades en el temps PS4

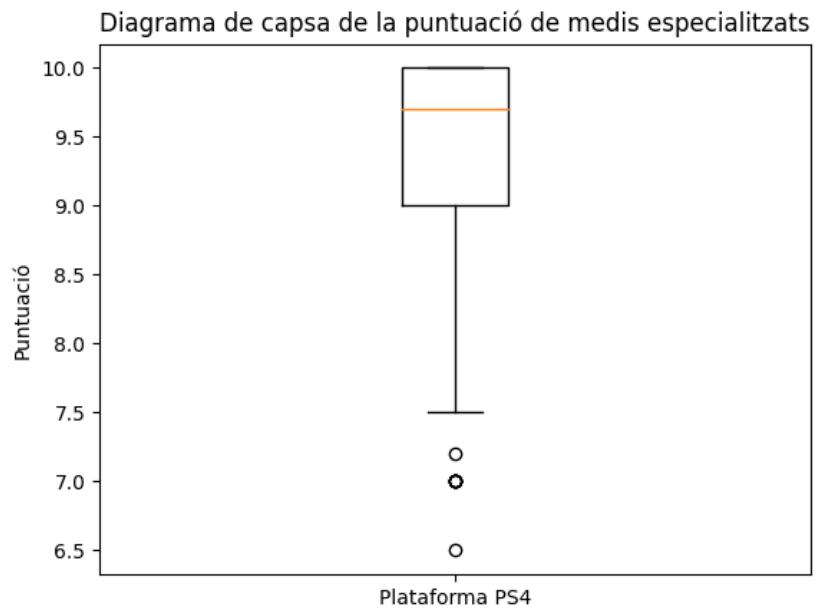


Figura 58: Gràfic de capsas de les puntuacions especialitzades rebudes de PS4

5.3.2.b Ressenyes dels usuaris

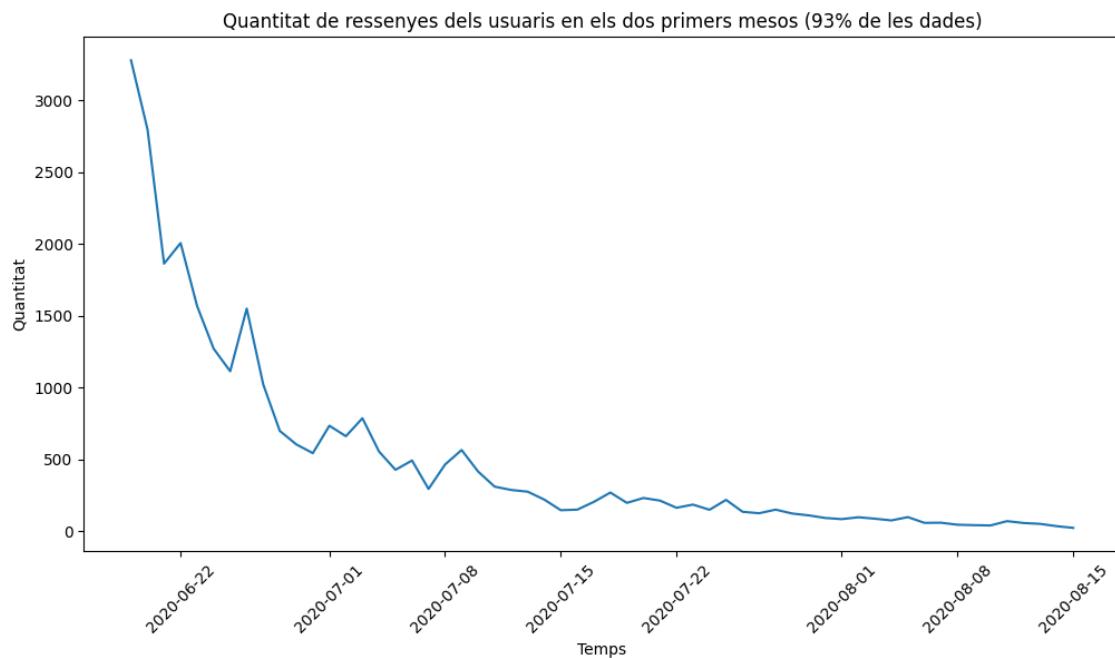


Figura 59: Gràfic de línies de les ressenyes d'usuaris en el temps PS4

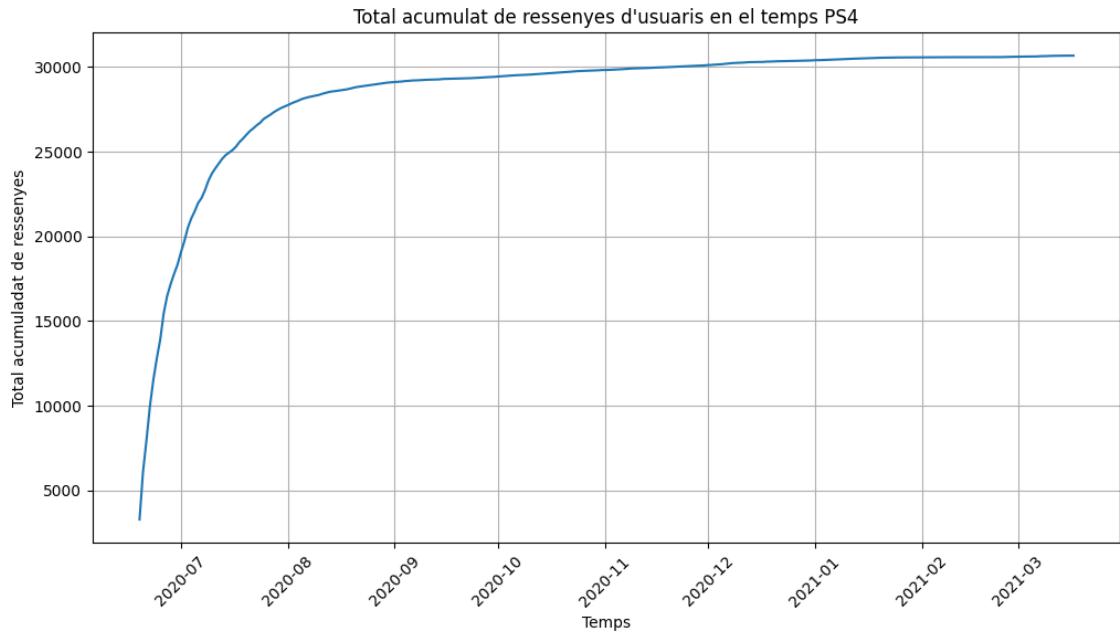


Figura 60: Gràfic de línies de les ressenyes d'usuaris acumulades en el temps PS4

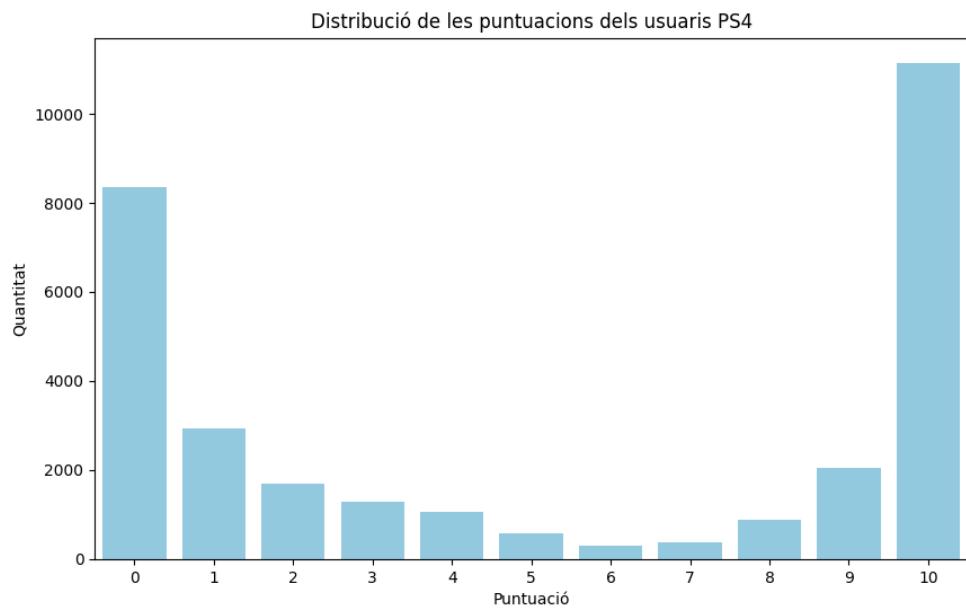


Figura 61: Gràfic de barres de la puntuació dels usuaris

A diferència de les puntuacions dels medis especialitzats on hi ha unes puntuacions que tendeixen a l'excel·lent, en el cas dels usuaris es pot observar una clara polarització en les puntuacions on els valors extrems 0 i 10 són els més freqüents mentre que els intermedis són gairebé inexistentes. Això demostra que els fans no puntuen segons la qualitat tècnica del videojoc, que és impecable, sinó per les emocions que els fa sentir el videojoc en jugar-lo.

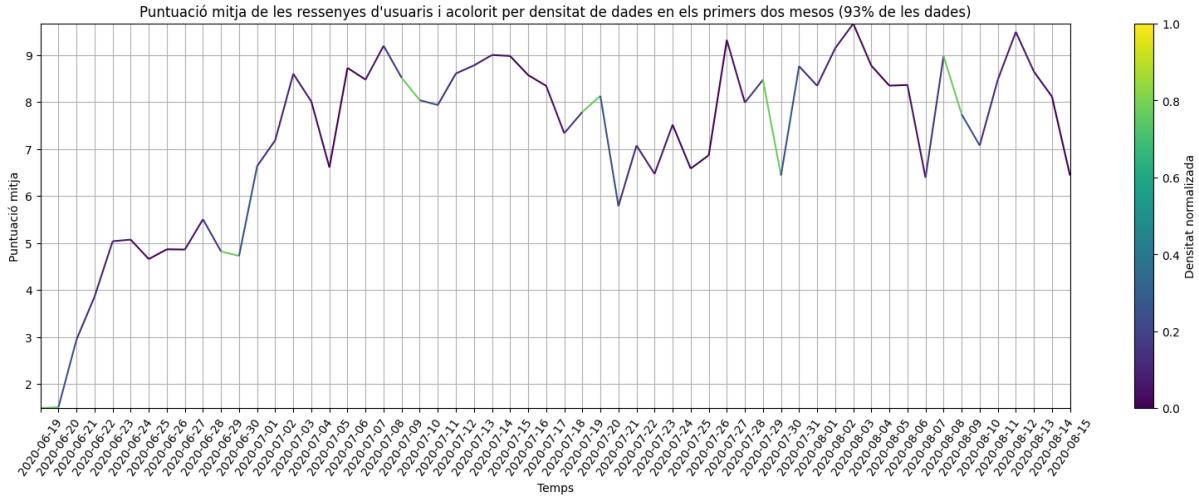


Figura 62: Gràfic de línies de les ressenyes d'usuari acumulades en el temps PS4

En el següent gràfic es mostren els dos primers mesos de llançament del videojoc i és interessant analitzar les oscil·lacions de la mitjana de les puntuacions dels usuaris per dia. Les àrees verdes indiquen que en aquests dies es van rebre un gran nombre de ressenyes. Un dels factors que podria influir en què moltes persones comentessin en massa el mateix dia podria deures un desplegament enrederit en alguna regió geogràfica en específic o campanyes de desprestigi dirigides per part *d'influencers* molt coneeguts que influeixin en l'opinió dels seus seguidors.

També es pot observar que els primers 15 dies del llançament el joc acumula moltes puntuacions negatives. Segons l'estudi teòric previ, s'atribueix a les decisions narratives que succeeixen en la primera hora de joc. Molts fans que van jugar la primera part queden decebuts i deslegitimen aquesta segona part com si no fos una digna successora.

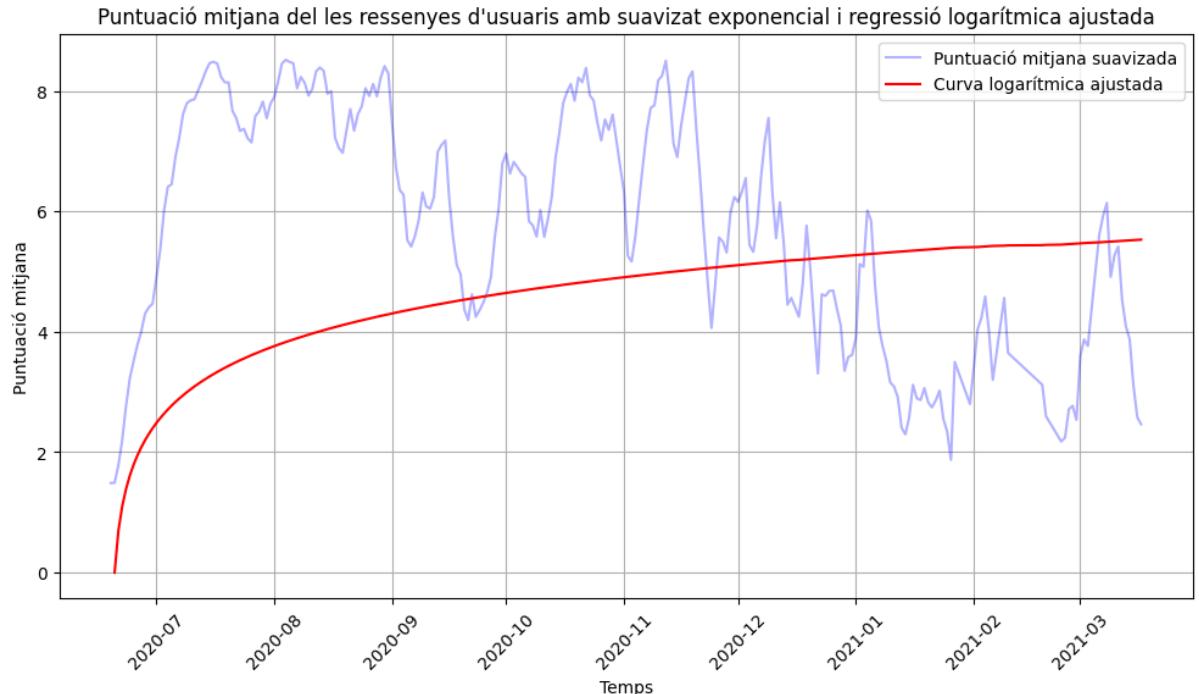


Figura 63: Gràfic de línies de les ressenyes d'usuari acumulades en el temps PS4

Si es veu tot l'historic amb el 10% de les dades restants, s'observa com les puntuacions tendeixen a la

baixa. La mitja de tot el conjunt del dataset és de 5.2. No obstant això, si s'analitza la font fiable més actualitzada que és Metacrític, actualment la mitja està en 5.8, el que indica que els últims mesos ha rebut puntuacions més positives per part dels usuaris.

Idioma de les ressenyes: TLOU Part II

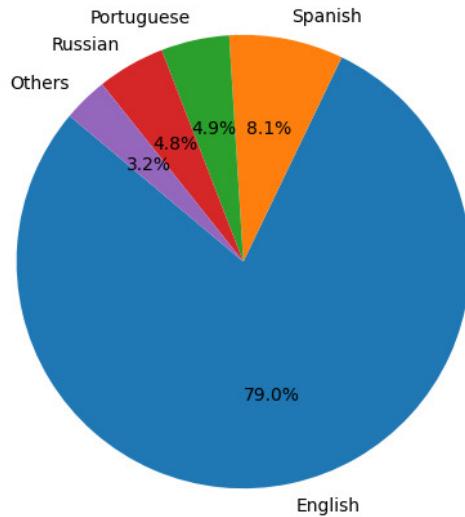


Figura 64: Gràfic de la representació dels idiomes en les ressenyes.

Si es mostra les ressenyes segons l'idioma s'observa que els idiomes més representatius són l'anglès, l'espanyol, el portuguès i el rus amb gairebé el 97% de les dades.

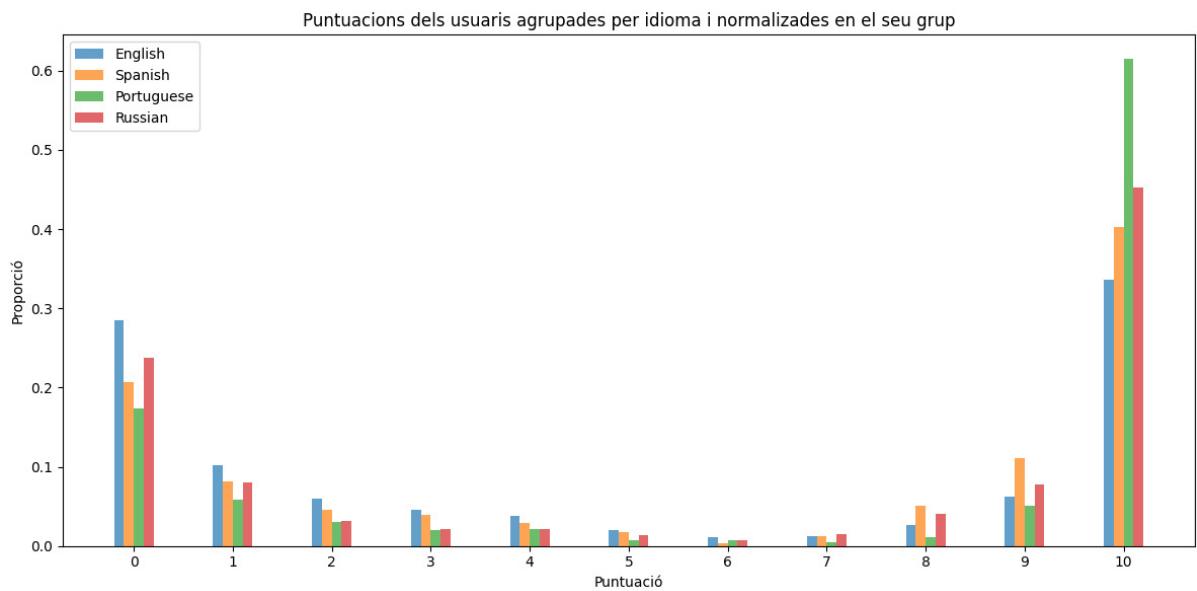


Figura 65: Gràfic de barres de la puntuació dels usuaris agrupada per idioma i normalitzada segons el grup

Per l'elaboració d'aquest gràfic s'escullen els idiomes amb més de 1000 ressenyes. Es pot veure com els comentaris anglesos són els que pitjor puntuen el videojoc, mentre que els portuguesos els que millor. Les causes d'aquest esbiaix poden ser moltes, des de la cultura i les sensibilitats dels usuaris, com la traducció i el doblatge, les diferències geogràfiques, entre altres.

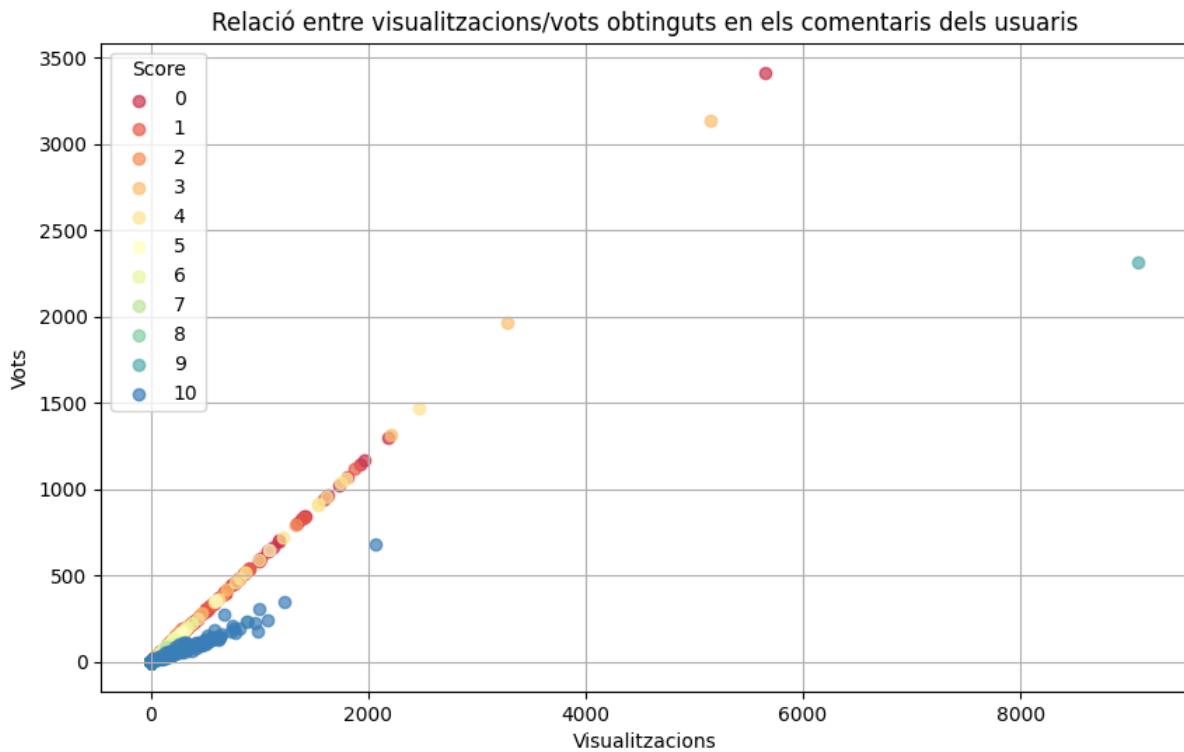


Figura 66: Gràfic de dispersió de *views/votes* de les ressenyes d'usuari, acolorit per *score*

En aquest cas, si s'analitzen els vots dels comentaris dels usuari, es pot observar com els comentaris que més suport reben són els que valoren amb notes més baixes al videojoc.

La forma en la qual se segmenten les dades segueix la mateixa metodologia explicada en les ressenyes d'usuari del primer videojoc.

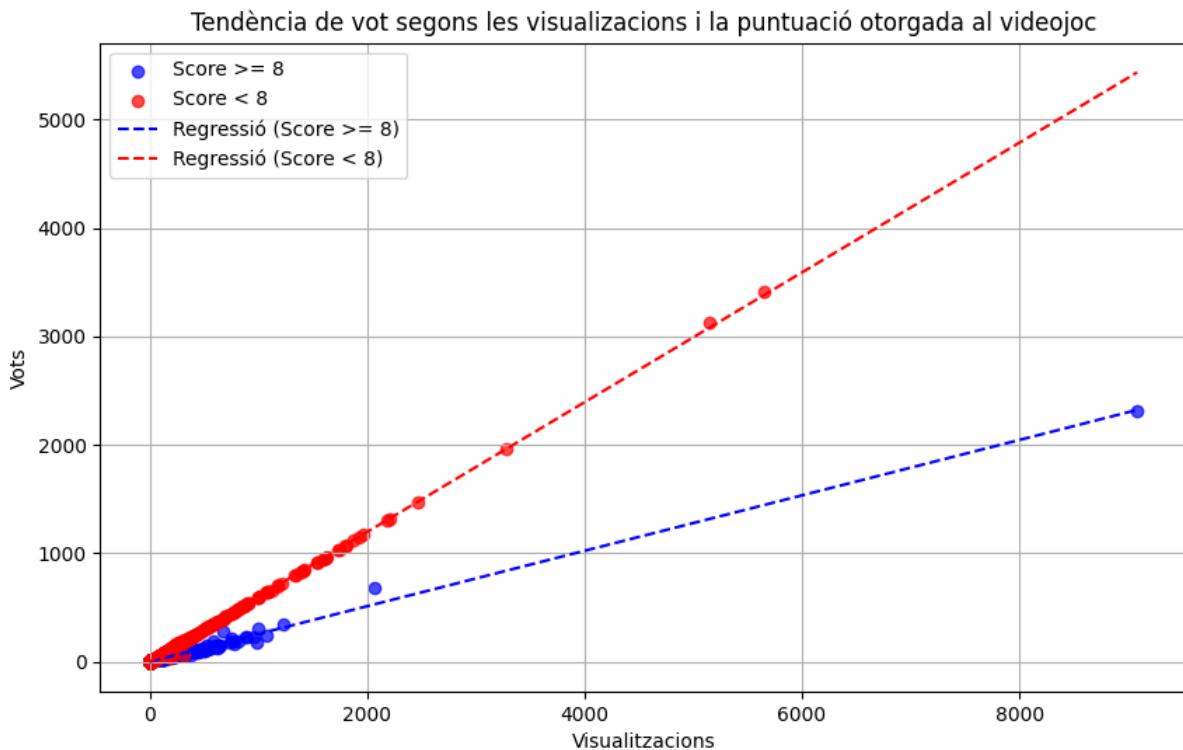


Figura 67: Regressió lineal per als dos subgrups en l'escala vots/visualitzacions

Segons la informació que proveeix la plataforma de Metacrític, es pot observar com els usuaris puntuuen més aquelles ressenyes que acaben atorgant al videojoc puntuacions per davall del 8.

Si es fa la mitjana entre els vots de cada grup s'obté:

- 1 vot per cada 2406 visualitzacions en les ressenyes amb valoracions ‘recomanables’
- 1 vot per cada 1688 visualitzacions en les ressenyes amb valoracions ‘mediocres’

És a dir, els comentaris amb puntuacions ‘mediocres’ al videojoc són 1.4 vegades més recolzats que els que donen puntuacions ‘recomanables’ dins d’aquest sector.

6 Anàlisi de sentiments

6.1 Etiquetatge multilingüe amb xlm-RoBERTa-base

Per fer la classificació de sentiments s'ha optat per utilitzar el dataset del videojoc “The Last of Us: Part II”, ja que segons s'ha vist a la visualització de dades és el dataset que més diversitat de comentaris presenta.

Com que el conjunt de dades no és prou gran per entrenar un model d'anàlisi de sentiments, s'ha acordat amb el tutor que per a l'estudi s'utilitzarà un model preentrenat de la plataforma <https://huggingface.co/>.

Per analitzar les ressenyes, en la primera iteració s'ha triat el model “**cardiffnlp/twitter-xlm-roberta-base-sentiment**” [55], tal com es va investigar a la part de teoria.

Aquest és un model multilingüe entrenat amb 198 milions de tuits i ajustat per a l'anàlisi de sentiments. Aquest model és especialment útil perquè cobreix gairebé la totalitat dels idiomes del dataset, incloent-hi els més representatius del conjunt de dades que són l'anglès, l'espanyol, el portuguès i el rus (97% de les dades).

A continuació, s'ha fet un programa per processar les ressenyes i guardar en el conjunt de dades les mètriques de l'anàlisi de sentiments.

El primer problema que s'ha hagut de solucionar és que la majoria de models tenen un límit de 512 tokens de buffer. Això és problemàtic perquè alguns comentaris superen els 3000 caràcters i són massa llargs. En aquesta primera iteració, s'ha segmentat les ressenyes que superen aquest límit en fragments afegint una capa de programació per d'amunt del model. S'ha fet l'anàlisi de sentiments de cada fragment i després s'ha calculat la mitjana ponderada en funció de la llargada dels fragments. Aquesta solució no és perfecta, però és la millor que s'ha trobat de moment.

Quant a la forma de computar les ressenyes, com que n'hi ha vora unes 30000, s'ha hagut de dividir-les en diversos arxius de mides “computables”. Concretament en 10 parts de 3000 ressenyes cada una. Després, mitjançant les diverses sessions que ofereix Google Colab, s'han paral·lelitzat en 6 màquines virtuals els arxius que, cada un, ha trigat prop de 30 mins en fer l'anàlisi de sentiments. Després s'ha tornat a ajuntar tots els documents en un únic conjunt de dades.

En aquesta primera iteració es pren la decisió de classificar les ressenyes amb una fiabilitat de menys del 55% de precisió com a neutrals. És a dir, per exemple, donades les següents puntuacions: (Positiu: 0.32, Neutral: 0.00, Negatiu: 0.11) com que cap supera el 0.55, s'opta per l'opció més conservadora que és etiquetar-les com a neutrals.

Polaritat de les ressenyes segons l'anàlisi de sentiments

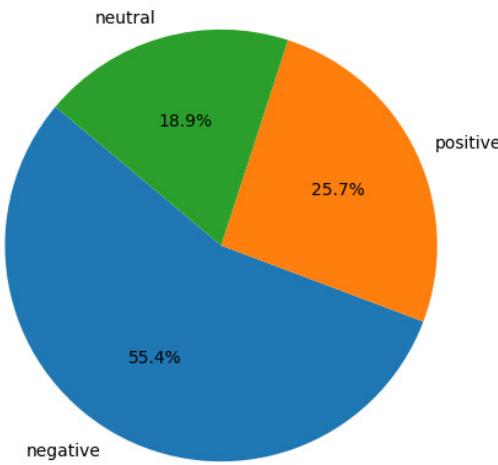


Figura 68: Polaritat de les ressenyes segons l'anàlisi de sentiments.

Una forma per comparar si la classificació s'està fent correctament és mirar si les ressenyes tenen relació amb la puntuació donada al videojoc. S'assumeix, per tant, que si la puntuació és alta implica que la ressenya serà positiva i si la puntuació és baixa implica que la ressenya serà negativa.

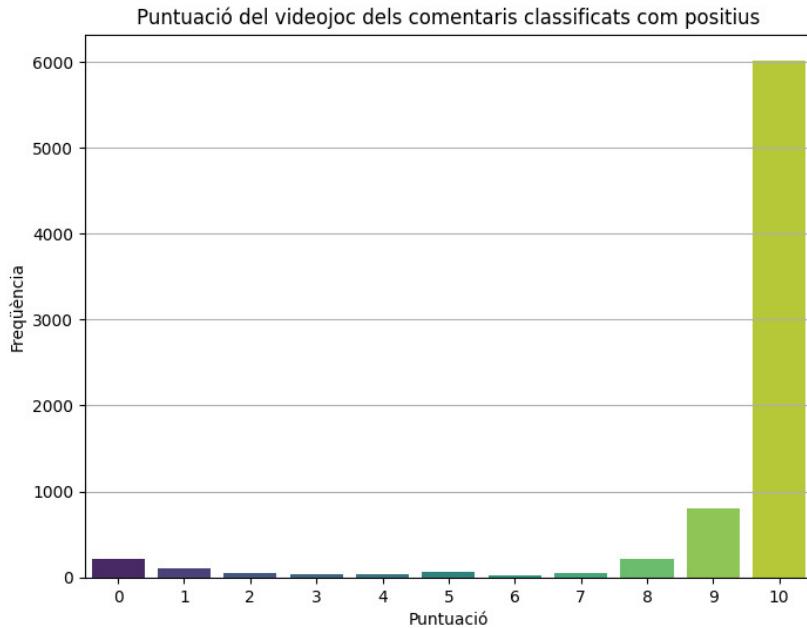


Figura 69: Puntuació dels comentaris classificats com positius.

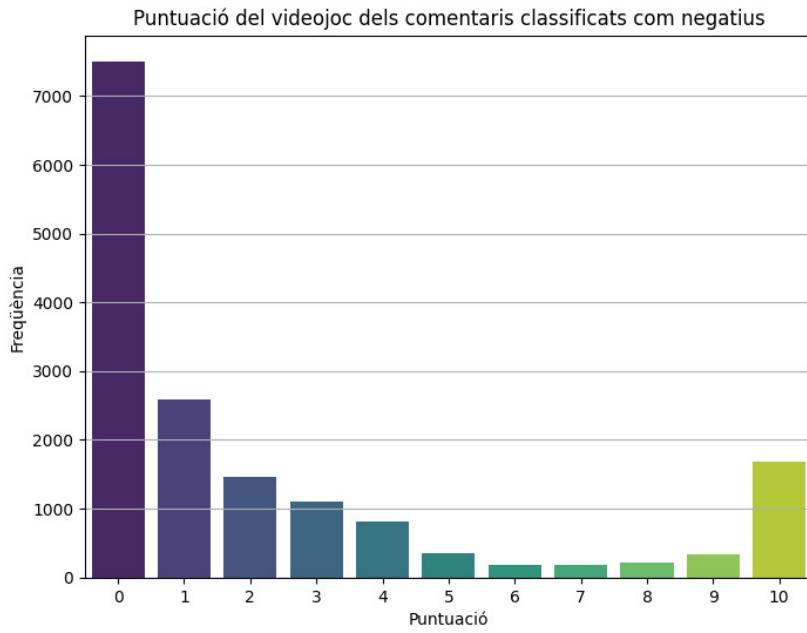


Figura 70: Puntuació dels comentaris classificats com negatius.

Quan una ressenya es classifica com a negativa sembla que hi ha més marge d'error. Una hipòtesi que es baralla és que moltes ressenyes tant positives com negatives tracten sobre la història i els esdeveniments que hi succeeixen. En moltes s'hi inclouen paraules com “assassinat”, “matar”, “mort” i “traïció” que poden influir en la sobrerepresentació dels comentaris negatius en l'anàlisi de sentiments.

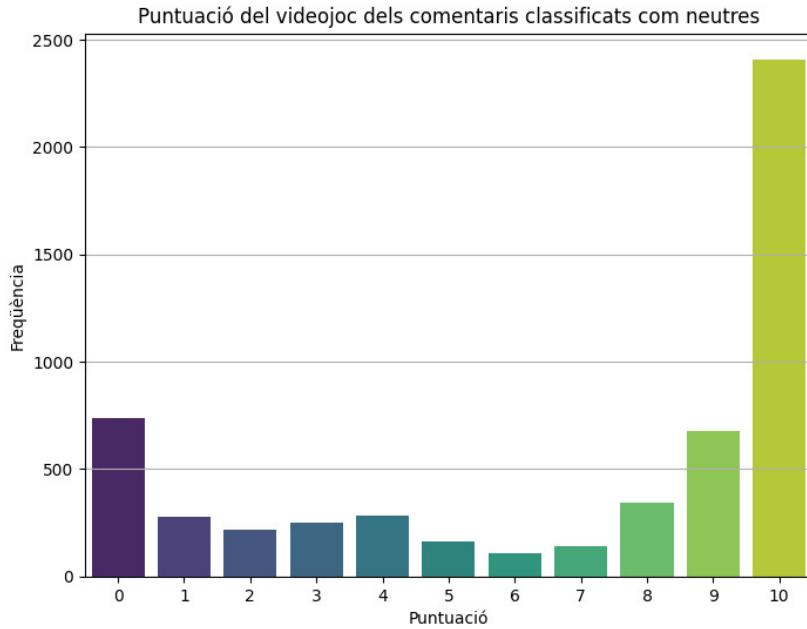


Figura 71: Puntuació dels comentaris classificats com neutres.

En aquest últim grup s'ajunten totes aquelles ressenyes que, o bé la seva naturalesa no és polaritzant, o l'anàlisi de sentiment no ha pogut treure resultats prou concloents. Això pot ser degut a ressenyes mal escrites, ambigües, massa llargues, etc. També s'observa que hi ha un gran percentatge de comentaris que han puntuat favorablement el videojoc i estan classificats com neutres.

En la segona iteració de l'etiquetatge s'intentarà afinar més en la classificació dels comentaris positius i veure si amb altres models s'obtenen millors resultats.

6.2 Etiquetatge amb diversos models específics

En aquesta segona iteració s'ha revisat quines són les formes de processar textos de gran longitud. Les opcions que hi ha són:

- Processar la seqüència per parts i afegir una capa d'amunt del model per ajuntar les parts.
- Truncar l'input directament amb el tokenitzador.
- Utilitzar un model de processament de textos llargs com *Longformer* o *BigBird*.
- Convertir un model ja existent perquè accepti inputs més llargs.

En l'estudi publicat a les Actes de la Conferència de 2022 sobre “Mètodes Empírics en el Processament de Llenguatge Natural” a Abu Dhabi [56], uns investigadors van demostrar que **RoBERTa-large** pot ser utilitzat amb tècniques de PEFT (Parameter-Efficient Fine-Tuning) per analitzar el sentiment de textos llargs, incloent-hi ressenyes de pel·lícules i ressenyes de productes. Aquest model presentat va superar altres models de referència, incloent-hi BERT i RoBERTa-base.

Tanmateix, encara que en aquest paper s'exposa com el model presentat millorar els resultats d'anàlisi de sentiments en textos llargs, en l'estat actual de l'art, RoBERTa-Large només està entrenat en anglès, per tant, no es pot utilitzar en contextos multilingües. A més, no hi ha un model capaç d'acceptar més de 512 tokens d'entrada. Les tècniques PEFT que s'utilitzen per processar els comentaris són: des de fragmentar els prompts per analitzar-los i després tornar-los a concatenar, fins a modificar els prompts perquè capturin només característiques específiques, categories o emocions que puguin permetre al model inferir millor els sentiments.

Augmentar la finestra de tokens no és una tasca senzilla, ja que depèn de l'arquitectura amb la qual s'ha entrenat el model. En altres paraules, la mida de finestra per processar tokens és proporcional a la quantitat de recursos (memòria) assignada i la llargada del corpus del text amb què s'ha entrenat el model. Per aquest motiu BERT i RoBERTa tenen una mida fixada de 512 tokens. [57]

A Hugging Face no existeix cap model especialitzat en l'anàlisi de sentiments que accepti corpus de text llargs. Transformar un model ja existent, tampoc sembla una solució realista pel temps que es disposa. Per tant, la decisió presa en aquesta segona iteració és la de truncar “de forma nativa” les ressenyes amb el tokenitzador i utilitzar models *fine-tunned* ja preentrenats que puguin estar més optimitzats pels diferents idiomes de les ressenyes. A continuació es nomena i es dona una breu explicació dels models escollits:

- Ressenyes en anglès: Per l'anàlisi de textos només en anglès s'ha trobat un model molt similar a l'utilitzat en la primera iteració, però que només està especialitzat en aquest idioma. “**siebert/sentiment-roberta-large-english**” [58]
- Ressenyes en espanyol: En aquest cas s'opta per un analitzador de sentiments basat en BERT i *fine-tunejat* amb un gran corpus de dades que conté 11.500 tuits en espanyol recollits de diverses regions, tant positius com negatius. “**VerificadoProfesional/SaBERT-Spanish-Sentiment-Analysis**”. [59]
- Ressenyes en rus: No hi ha massa a elegir, però el model millor valorat és “**seara/rubert-tiny2-russian-sentiment**” [60], que està basat en BERT i *fine-tunejat* amb diversos datasets Russos sembla la millor opció. També s'ha mirat la possibilitat de fer l'anàlisi de sentiments amb Flan-T5-xl que és un Massive Multitask Language Understanding (MMLU) com GPT-3, però consumia massa memòria ram i no s'ha pogut executar.
- Ressenyes en portuguès i la resta d'idiomes: S'ha decidit utilitzar el mateix model que en la primera iteració “**cardiffnlp/twitter-xlm-roberta-base-sentiment**”. [55]

Per processar les ressenyes, de forma similar a la primera iteració s'ha dividit les ressenyes en diversos arxius computables per diverses màquines virtuals a Google Colab.

Polaritat de les ressenyes segons l'anàlisi de sentiments

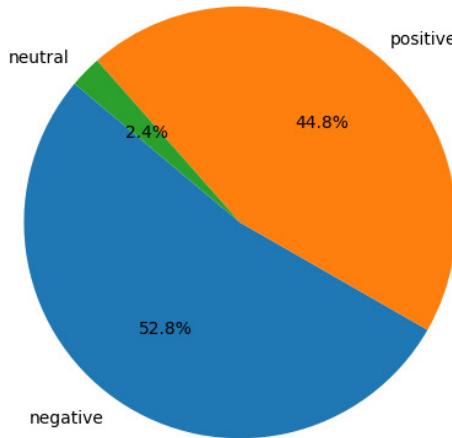


Figura 72: Polaritat de les ressenyes segons l'anàlisi de sentiments amb models específics.

En aquesta segona versió s'ha aconseguit millorar la detecció de comentaris positius significativament. També s'ha de dir, que s'ha utilitzat alguns models que fan classificació binària i, per tant, els comentaris neutrals tenen menys representació.

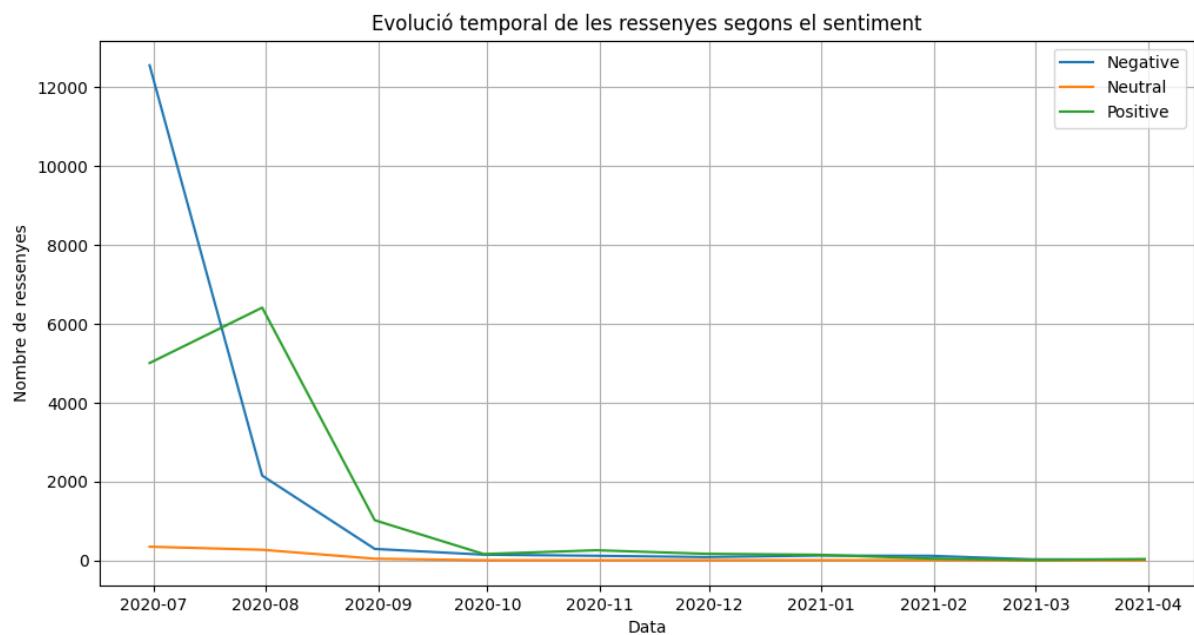


Figura 73: Gràfic de línies del sentiment de les ressenyes al llarg del temps

És curiós com la majoria dels comentaris negatius són en les primeres 24 hores, però el joc dura entre 25 i 30 hores. O no han jugat el joc completament o només han jugat fins a la part controvertida de la narrativa, que succeeix dues hores després de l'inici de la història.

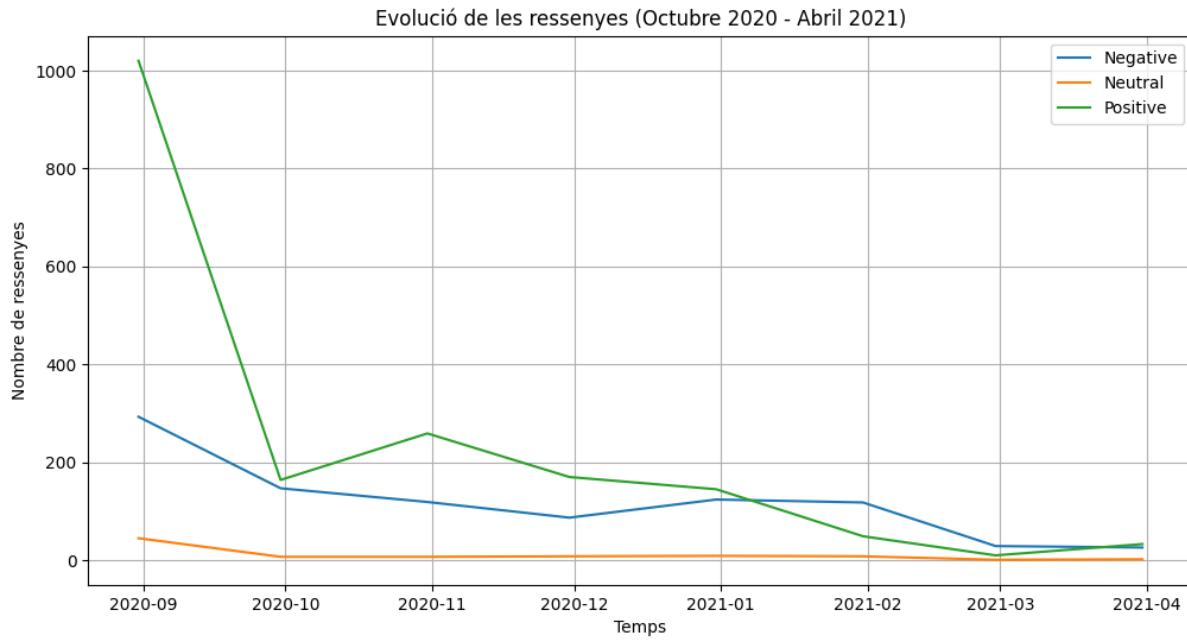


Figura 74: Gràfic de línies (ampliat) del sentiment de les ressenyes al llarg del temps

En ampliar el gràfic per als últims mesos, s'observa que la tendència és més positiva que negativa. Això indica que la percepció del videojoc ha millorat.

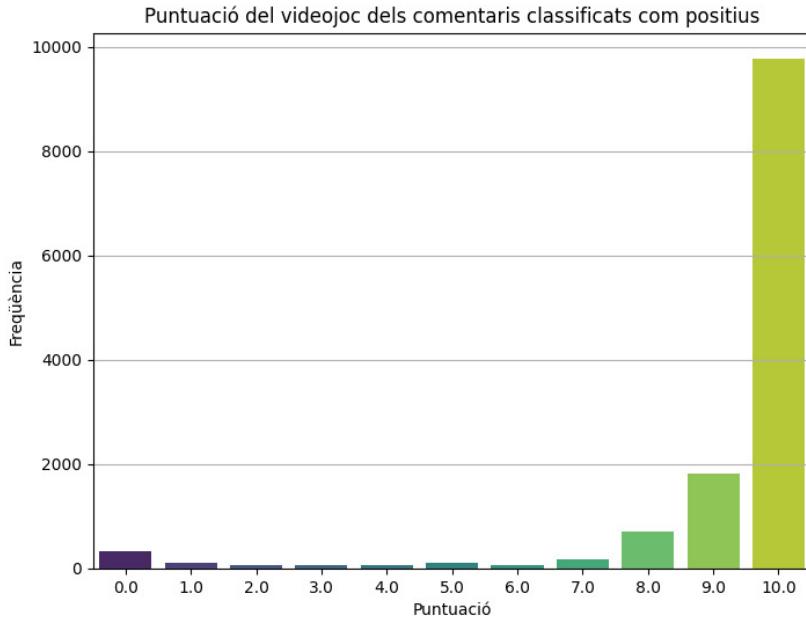


Figura 75: Puntuació dels comentaris classificats com positius.

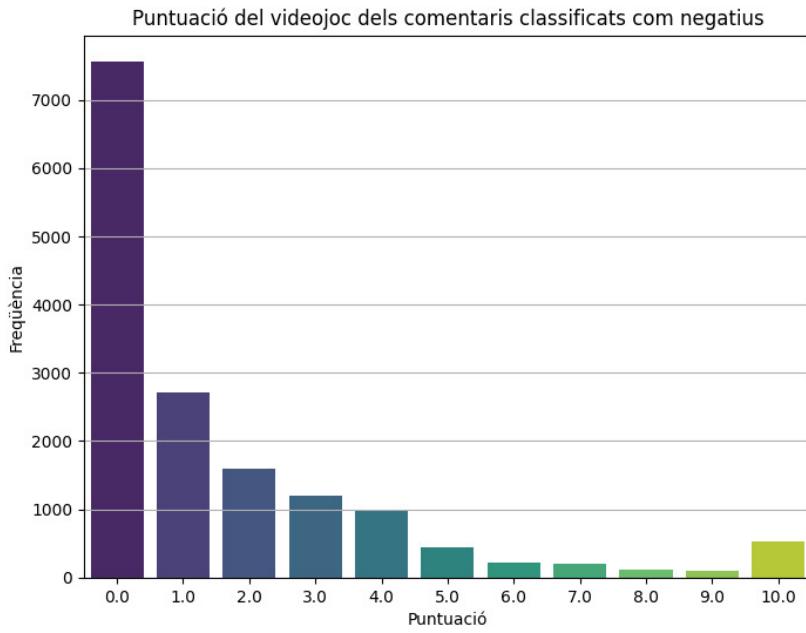


Figura 76: Puntuació dels comentaris classificats com negatius.

La proporció de comentaris negatius amb bones puntuacions disminueix respecte al primer model. Si es mira alguns dels comentaris negatius amb puntuació de 10 sembla que molts d'ells critiquen als jugadors que puntuuen amb mala nota el videojoc. Per demostrar aquesta hipòtesi es podria fer una classificació d'aquest grup de comentaris segons quins parlen del videojoc i quins critiquen els usuaris que no els ha agradat el videojoc. Si aquesta hipòtesi fos certa, voldria dir que no sempre una puntuació alta implica una ressenya positiva.

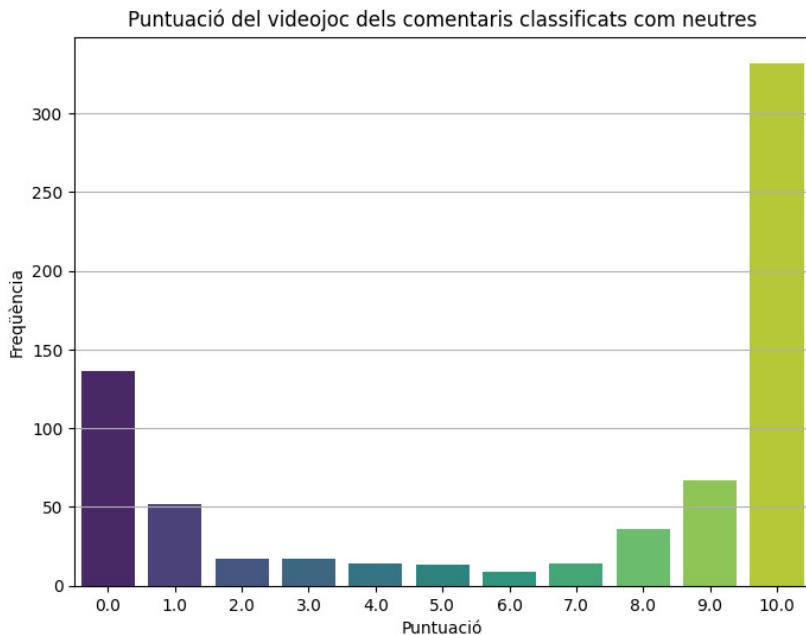


Figura 77: Puntuació dels comentaris classificats com neutres.

Els comentaris neutres encara que tenen també una distribució similar a la del primer model, s'ha aconseguit reduir la seva representació d'un 18.9% a un 2.4% del total de les dades.

Una altra dada interessant és que les ressenyes amb classificació **positiva** tenen una longitud mitjana de **433 caràcters**, mentre que les **negatives** tenen una longitud mitjana de **217 caràcters** i les **neutres**, en canvi, una mitjana de **831 caràcters**. Això posa en evidència que quan es tracta de comentaris negatius, els comentaris són més curts i directes. Mentre que els positius són més llargs i específics. Per últim, els neutres són amb diferència els més llargs i, per tant, els més difícils d'avaluar i treure un sentiment predominant del text.

6.3 Model de freqüència de paraules

Per mantenir la continuïtat de l'estudi, s'utilitza el conjunt de dades de The Last of Us: Part II. Per a la comptabilització de la freqüència de les paraules, és crucial que totes estiguin en el mateix idioma. Amb aquest objectiu, s'han explorat dues opcions per traduir totes les ressenyes a l'anglès.

- **SeamlessM4Tv2** [61] en la seva versió SeamlessM4Tv2ForTextToText és el model de traducció més recent de Meta. Després de provar el model, s'ha observat que el temps de computació per a cada ressenya és excessivament llarg. S'ha decidit llavors canviar a una opció més lleugera per millorar l'eficiència del procés. Igualment, es deixa un exemple de com s'ha processat amb aquest model algunes ressenyes als arxius de la pràctica.
- **Helsinki-NLP/opus-mt-mul-en** [62] és un projecte de traducció automàtica que forma part de l'*Open Parallel Corpus* (OPUS). Aquest model està especialitzat en la traducció de múltiples idiomes a l'anglès. Ha resultat ser lleugerament més ràpid que SeamlessM4Tv2 encara que sembla que la traducció és una feina computacionalment més costosa que l'etiquetatge de sentiments.

D'igual manera que amb l'etiquetatge de sentiments, s'ha dividit el dataset en parts petites per poder computar-les de forma paralela en múltiples màquines virtuals.

Després d'ajuntar totes les parts amb molta cura, s'ha procedit a fer el preprocessament de les dades. S'ha utilitzat la gran majoria de tècniques explicades a la part de teoria [63] [64] [65]:

1. Convertir a minúscules el text.
2. Convertir els números a paraules.
3. Eliminar signes de puntuació i símbols.
4. Eliminar espais en blanc redundants, tabulats, etc.
5. Eliminar *Stopwords*.
6. Lemmatitzar totes les paraules a la seva arrel.
7. Taggejar totes les paraules segons la seva categoria gramatical per a després només escollir els noms i adjetius.
8. Eliminar paraules massa petites o massa llargues.
9. Excloure paraules que no aportaven valor de forma específica.
10. Eliminar totes les paraules que no aportessin sentiment, mitjançant el model Vader.

Després amb l'eina de CountVectorizer s'ha comptabilitzat la freqüència de les paraules resultants i s'ha obtingut els següents resultats:

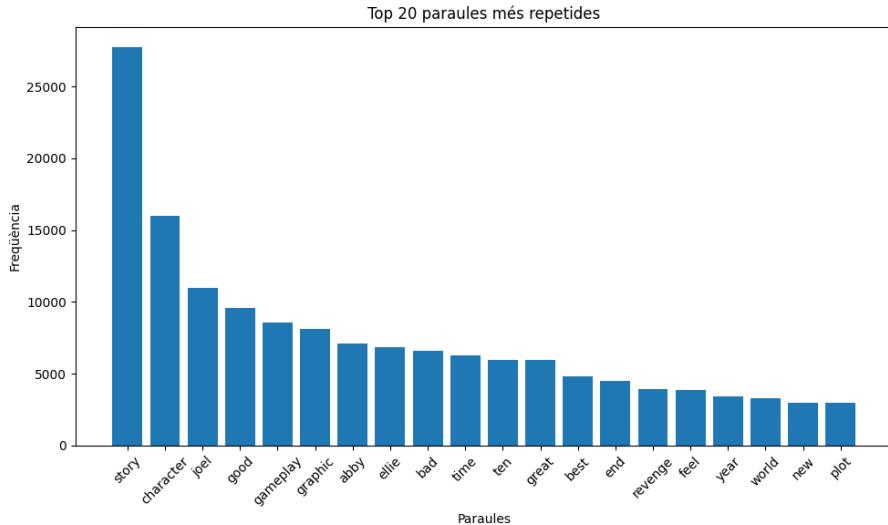


Figura 78: Gràfic de freqüència de les paraules més repetides de les ressenyes.

Per poder crear aquest gràfic s'han exclòs algunes paraules que no tenien massa rellevància, com ara: “game”, “play”, “first”, “last”, “part”, “one”, “two”, “naughty”, “dog”. També s'ha intentat filtrar les paraules amb molt de pes que apareixen en totes les puntuacions, com ara “good”, “great”, ja que no té gaire sentit que un comentari amb puntuació de 0 contingui aquestes paraules; probablement es referien al primer videojoc. També s'exclou “bad”, ja que és una paraula molt freqüent i és l'antònim de “good”.

Com es pot observar en el gràfic, la narrativa sembla tenir una importància destacada per a les opinions dels usuaris, ja que moltes de les paraules més utilitzades fan referència a: “story”, “character”, “joel”, “ellie”, “abby”, “plot”. També es fa ús de la paraula “time”, que pot referir-se a la durada del videojoc, una de les principals crítiques rebudes, així com al desenllaç del mateix “end”. D'altra banda, hi ha paraules que aborden l'aspecte tècnic com “gameplay”, “graphic”, “ten”.

Per altra banda, per fer els mosaics amb les paraules més utilitzades, s'ha fet ús del model Vader [66] per discriminari aquelles paraules que no representen un sentiment i alhora acolorir amb un gradient de verd a roig les paraules segons la seva polaritat. Per altra banda, les paraules negatives referents a la narrativa també apareixen a gairebé tots llocs: “revenge”, “kill”, “death”, “hate” el que sustenta la hipòtesi que podria ser un dels motius pels quals alguns comentaris amb bona puntuació estiguin classificats com negatius.



Figura 79: Mosaic de les paraules més freqüents, segmentat segons la nota que els jugadors van donar al videojoc. (Word Cloud)

7 Predicció de la puntuació dels usuaris a partir dels comentaris

7.1 Entrenament del model de regressió

Per l'entrenament del model de regressió s'han utilitzat els valors resultats de l'etiquetatge de l'anàlisi de sentiments. Com a variables independents s'utilitzen els valors 'positive', 'negative' i 'neutral' mentre que com a variable dependent (objectiu) s'utilitza la variable 'score'.

7.2 Descripció de les mètriques d'avaluació utilitzades

- **L'error mitjà absolut (Mean Absolute Error (MAE))** és una mètrica d'avaluació utilitzada per mesurar la precisió d'un model de predicció, especialment en problemes de regressió. L'objectiu del MAE és quantificar la precisió de les prediccions fets per un model en relació amb els valors reals. Es calcula fent la mitjana de les diferències absolutes entre els valors predictius i els valors reals. [67]

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

En no elevar els errors al quadrat, MAE no penalitza tant els errors grans, fet que el fa poc sensible als valors anòmals. Per tant, no és una mètrica recomanable per a models en què cal prestar especial atenció a aquests errors. El més desitjable és que l'error sigui proper a 0. El MAE és fàcil d'interpretar perquè utilitza les mateixes unitats que les dades originals, la qual cosa permet entendre immediatament la magnitud de l'error.

- **L'error mitjà quadràtic (Mean Squared Error (MSE))** és una altra mètrica d'avaluació de models de regresió. Es calcula fent la mitjana dels quadrats de les diferències entre els valors predictius i els valors reals.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

En elevar al quadrat els valors, magnifica els errors grans, per la qual cosa s'ha d'utilitzar amb cura quan es té valors anòmals (outliers) en el conjunt de dades. Pot prendre valors entre 0 i infinit. Com més a prop de zero estigui la mètrica, millor.

- **L'arrel quadrada de la mitjana de l'error quadrat (Root Mean Square Error (RMSE))** és una altra mètrica d'avaluació comuna en problemes de regressió. Es calcula fent l'arrel quadrada del MSE.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

L'avantatge d'aquesta mètrica és que presenta l'error en les mateixes unitats que la variable objectiu, fet que la fa més fàcil d'entendre.

- **R^2 , també conegut com a coeficient de determinació**, és una mesura estadística que mesura el percentatge de la variació en la variable dependent (la que estem intentant predir) en funció de les variables independents (les que utilitzem per fer la predicció). Per exemple, un valor de $R^2 = 0.50$ indica que el 50% de la variabilitat de la variable dependent és explicada per les variables independents utilitzades en el model de regressió. [68]

En un gràfic de regressió, R^2 ens indica com de prop estan els punts de dades reals de la línia de regressió ajustada pel model. Valors propers a 1 indiquen que el model s'ajusta bé a les dades, és a dir, les variables independents estan molt relacionades amb la variable dependent. En contrast, un valor proper a 0 indica que el model no ajusta bé les dades i que les variables independents tenen

una relació més feble amb la variable dependent.

És important tenir en compte que aquest coeficient per si sol no pot determinar per si sol si els punts de les dades o les prediccions estan esbiaixades.

El coeficient de determinació es calcula com la proporció de la variació total que és explicada pel model:

$$R^2 = \frac{SS_{\text{reg}}}{SS_{\text{tot}}}$$

Variació explicada (o regressió):

$$SS_{\text{reg}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Variació total:

$$SS_{\text{tot}} = \sum_{i=1}^n (y_i - \bar{y})^2$$

On:

- y_i són els valors observats de la variable dependent.
- \bar{y} és la mitjana dels valors observats de la variable dependent.
- \hat{y}_i són els valors predictius del model.
- n és el nombre total de mostres.

- **Coeficients de la Regressió.** Un dels mètodes més senzills per entendre la influència relativa de les variables independents és examinar els coeficients del model de regressió.

Els coeficients de regressió obtinguts de la regressió lineal són:

- negative: -8.63
- positive: -0.84
- neutral: -2.98

Mentre que el valor de *intercept* és 10.1.

La interpretació que es fa d'aquestes dades és que quan totes les variables (negative, positive i neutral) tenen un valor de 0, "l'interceptació" de la regressió pren el valor de 10.1, que correspon a la puntuació màxima possible del videojoc. Això significa que sense cap influència del sentiment de les ressenyes, el model prediu la puntuació màxima del joc.

Quan la variable "positive" és molt alta i les altres dues són baixes, l'impacte sobre "l'interceptació" és mínim, ja que la resta que se li fa a 10.1 és petita. Per tant, en aquest cas, el model prediu una puntuació molt alta per al videojoc, ja que la influència positiva és dominant.

Per contra, si la variable "negative" és molt alta mentre que les altres dues són baixes, la resta que se li fa a 10.1 és significativament gran (aproximadament 8). Això provoca que el model predigui una puntuació molt baixa per al videojoc.

7.3 Visualització de l'ajust del model

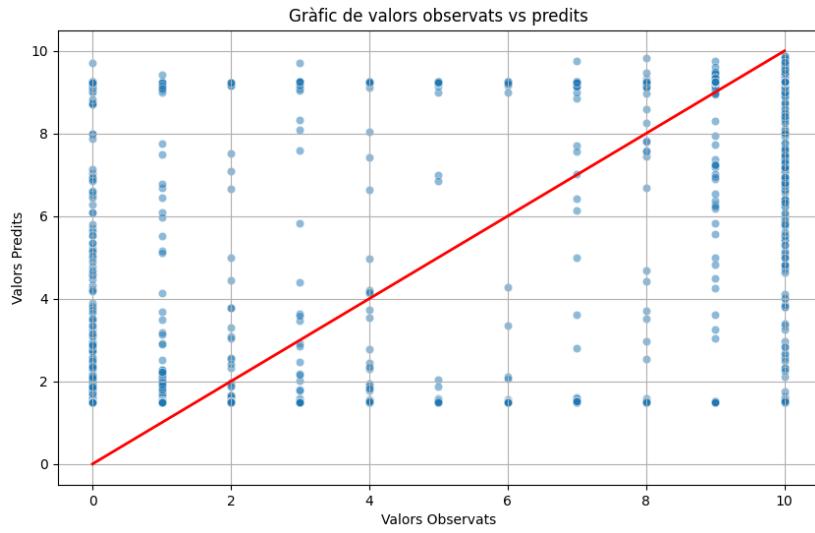


Figura 80: Gràfic de valors observats vs predictis. Regressió Lineal.

D'aquest gràfic observem un patró asimètric on les dades es concentren més en els extrems i menys en el centre. També hi ha bastant outliers quan la puntuació és un 10 o és un 0. Això depèn de la naturalesa de les dades i de com el model d'anàlisi de sentiments ha predit la polaritat dels comentaris.

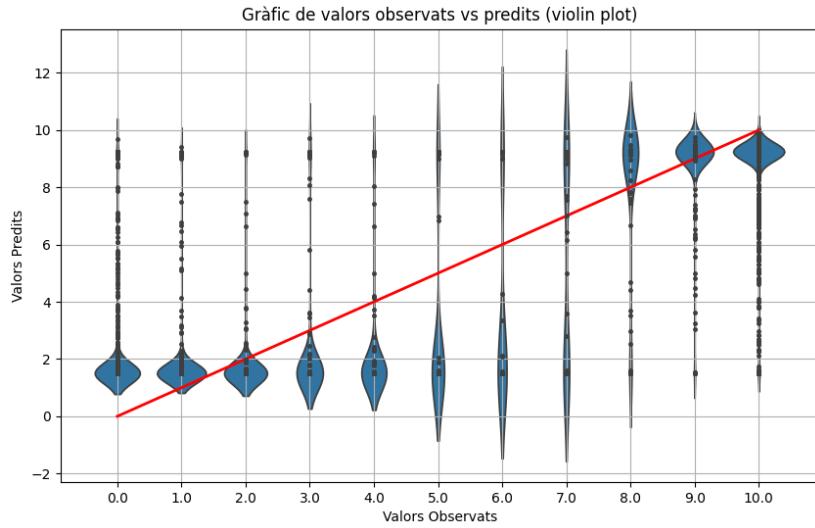


Figura 81: Diagrama de violí de valors observats vs predictis. Regressió Lineal.

Per veure la densitat de les dades s'utilitza un diagrama de violí.

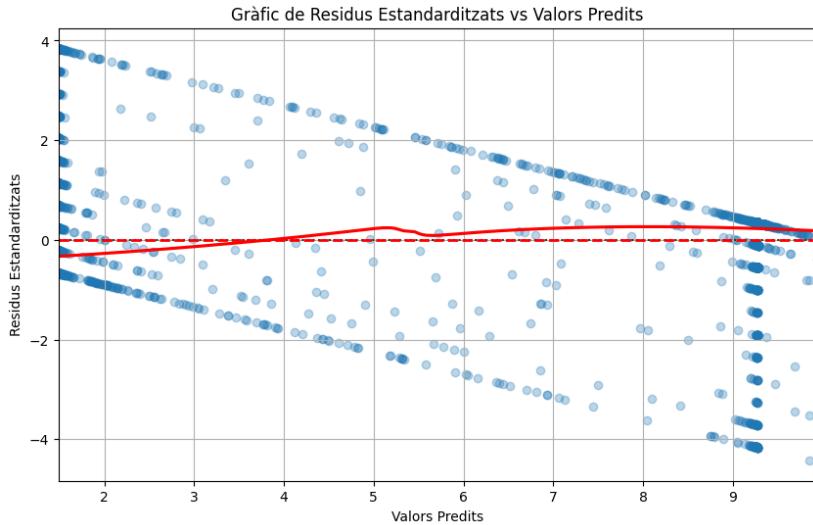


Figura 82: Gràfic de Residus Estandarditzats vs Valors Predicts. Regressió Lineal.

Els valors residuals són els valors observats menys els valors predictos. Els residus mostren un patró de dispersió que sembla disminuir a mesura que els valors predictos augmenten.

7.4 Resultats obtinguts

Per calcular els resultats, ja que no es té un dataset massa extens s'ha utilitzat la validació creuada (Cross Validation) amb $k=5$ folds per tenir uns resultats els més fiables possibles.

Aquestes són els resultats de la regressió lineal:

1. Mitjana Mean Absolute Error (MAE): 1.499
2. Mitjana Mean Squared Error (MSE): 4.994
3. Mitjana Root Mean Squared Error (RMSE): 2.234
4. Mitjana R-squared (R^2): 0.742

Anàlisi de cada mètrica:

1. Una MAE d'1.499 significa que, de mitjana, les prediccions del model estan aproximadament a 1.5 punts dels valors reals de les puntuacions dels usuaris.
2. Una MSE de 4.994 significa que, de mitjana, els errors quadràtics de les prediccions són aproximadament de 5 punts.
3. Una RMSE de 2.234 indica que les prediccions del model tenen un error d'aproximadament 2.23 en les mateixes unitats que les puntuacions de les ressenyes dels usuaris.
4. Una R^2 indica que el model pot explicar aproximadament el 74.2% de la variància en la puntuació de les ressenyes dels usuaris.

En els resultats obtinguts poden ser bons o dolents en funció de l'estàndard de precisió que es vulgui assolir. Considero que una desviació MAE d'1.5 punts respecte als valors reals està força bé. A més, que la R^2 sigui de 74.2% vol dir que el model s'adequa a la major part de les dades.

També s'ha fet un model amb una predició logarítmica de grau quatre que ha obtingut resultats lleugerament millors a la regressió lineal. Els resultats obtinguts de la validació creuada són els següents:

- Mitjana Mean Absolute Error (MAE): 1.460
- Mitjana Mean Squared Error (MSE): 4.794
- Mitjana Root Mean Squared Error (RMSE): 2.189

- Mitjana R-squared (R^2): 0.752

Tanmateix, per una altra distribució de dades més complexa o si es volgués filar prim, aquest model no seria adequat i es requeriria un model que considerés altres variables per fer una predicció encertada. Per exemple, enfocament híbrid entre una anàlisi de sentiments i un model de classificació.

8 Conclusions

En el present treball de “Processament de llenguatge natural aplicat a ressenyes de videojocs” s’ha fet un estudi teòric sobre els fonaments del NLP i s’ha vist algunes de les tècniques més utilitzades d’avantguarda per dur a terme l’anàlisi de textos. Tota aquesta investigació ha culminat en l’anàlisi de diversos grups de dades relacionats amb les ressenyes dels videojocs de la saga “The Last of Us”, on s’ha posat en pràctica els coneixements apresos.

En el document, s’ha realitzat una extensa labor de contextualització del problema, sustentada en una gran quantitat d’articles i entrevistes en diferents mitjans, per entendre millor les percepcions dels jugadors. Aquesta contextualització ha estat de gran valor a l’hora d’analitzar les dades.

En la part pràctica, s’ha seguit la metodologia CRISP-DM, que estableix les diferents fases necessàries per desenvolupar un projecte de ciència de dades de manera estructurada i eficient. S’han plantejat hipòtesis sobre els diversos datasets analitzats i algunes de les conclusions a les que s’ha arribat són:

- Quant al primer videojoc, The Last of Us, no s’han trobat evidències que el pas del temps hagi millorat la percepció dels usuaris sobre el videojoc. Això es pot observar en les diverses gràfiques de puntuació referents a les versions de PS3 i PS4, amb més d’un any de diferència de llançament, que no reflecteixen un canvi notable.
- El que sí que s’ha trobat és una correlació directa entre un increment en la quantitat de ressenyes i notorietat de la primera part amb la sortida del segon videojoc, a causa de la continuïtat directa de la història.
- Els usuaris tendeixen a valorar més la narrativa per damunt del criteri tècnic del videojoc. Aquesta hipòtesi queda corroborada per totes les mencions als personatges principals i al guió, per damunt d’altres aspectes, en l’anàlisi de la freqüència de paraules.
- L’idioma dels comentaris pot influir significativament en les tendències de mercat i en les ressenyes sobre el videojoc. Aquesta hipòtesi no s’ha pogut corroborar, ja que, tot i que s’ha observat que els comentaris en portuguès en els dos videojocs valoren molt millor el videojoc en comparació amb altres idiomes, no s’ha trobat un motiu específic evident. Farien falta més dades i un estudi més profund d’aquest fenomen.
- Les visualitzacions i vots dels comentaris d’usuaris a Metacritic estan estretament relacionats amb l’opinió general del joc. Aquesta hipòtesi és certa i es pot veure en ambdós videojocs, ja que hi ha una relació directa entre aquestes dues variables i les puntuacions de les ressenyes positives o negatives.
- Amb el pas del temps, l’opinió pública de The Last of Us: Part II ha millorat positivament. Segons l’anàlisi de sentiments feta amb les dades històriques dels usuaris, es pot veure com hi ha una tendència a l’alça dels comentaris positius sobre el videojoc.

En aquest estudi, s’ha implementat una àmplia gamma de models per a l’anàlisi de textos, com ara: twitter-xlm-roberta-base-sentiment, sentiment-roberta-large-english, SaBERT-Spanish-Sentiment-Analysis, rubert-tiny2-russian-sentiment i el model Vader per fer l’anàlisi de sentiments.

També s’han utilitzat dos models de traducció: el model de Meta, SeamlessM4T-v2, en la seva versió de text a text; i el model multilingüe de l’Open Parallel Corpus, opus-mt-mul-en, per a la traducció de les ressenyes. A més, s’ha dut a terme un extens preprocessament de dades aplicant les tècniques vistes en la part teòrica. Finalment, s’ha elaborat un WordCloud per visualitzar de manera gràfica les paraules clau i més rellevants de les ressenyes analitzades.

A més a més, també s’ha desenvolupat un model de predicció de la puntuació dels comentaris dels usuaris gràcies als resultats de l’anàlisi de sentiments i s’ha documentat totes les mètriques d’avaluació utilitzades.

Per conoure, aquest treball m’ha fet adonar de la complexitat i la diversitat de tècniques disponibles per a l’anàlisi de llenguatge natural i m’ha fet indagar en els models transformadors que són la base

de molts dels avenços actuals. També m'ha inspirat a seguir investigant en aquest camp que encara té molts reptes i oportunitats per descobrir i explorar. Aquesta experiència ha contribuït, encara més si cap, al meu desig de dedicar-me a la ciència de dades i la intel·ligència artificial. Penso que és un camp amb un potencial enorme per contribuir al progrés tecnològic i a la comprensió del món a través de la informació, l'anàlisi i la predicción de dades.

9 Possibles ampliacions de l'estudi

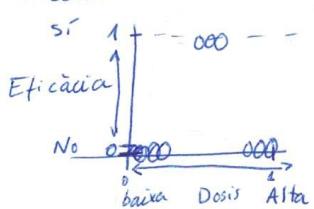
En el present treball, s'han abordat diverses hipòtesis i s'han obtingut resultats significatius. No obstant això, existeixen diverses vies potencials per ampliar aquest estudi en el futur:

- **Exploració d'hipòtesis obertes:** Durant l'anàlisi s'han pogut corroborar algunes hipòtesis, però d'altres, per falta de temps, no s'han pogut demostrar amb suficient evidència. Un exemple destacat és la presència de comentaris amb sentiment positiu associats a puntuacions baixes, especialment puntuacions de 0. Aquesta observació, indicada en l'apartat 6.2, com d'altres exposades en l'estudi preliminar 5.2, podrien ser objecte d'un estudi més profund de cara a una ampliació.
- **Extracció de característiques del text:** Un altre enfocament prometedor seria realitzar una extracció de característiques del text per identificar els temes principals que determinen una puntuació positiva o negativa. Mitjançant tècniques de processament de llenguatge natural, es podrien analitzar les ressenyes per trobar patrons comuns i paraules clau que influeixen significativament en les valoracions dels usuaris. Aquesta anàlisi temàtica podria proporcionar una comprensió més profunda i una ampliació de les dades amb les quals poder treballar.
- **Desenvolupament d'un model de predicció de puntuació més complex:** Utilitzant les noves dades obtingudes de l'extracció de característiques, es podria definir un model de predicció de la puntuació dels usuaris més sofisticat. Això permetria avaluar si es pot obtenir una millora significativa en la precisió de les prediccions comparada amb el model actual. A més, aquest model podria integrar altres variables com la longitud de la ressenya, la freqüència de certes paraules clau i el sentiment general del text.
- **Web Scraping:** Una altra ampliació important seria l'ús de tècniques de Web Scraping per obtenir conjunts de dades més actualitzats. Això permetria veure si les observacions realitzades en aquest estudi es mantenen vigents o si han canviat amb el temps. Es podrien recollir noves ressenyes tant dels jocs ja analitzats com d'altres remakes o remasters de versions per a PS5. Aquesta actualització de dades podria aportar noves pistes sobre les tendències en les valoracions dels usuaris.

Annex

Problema: "Es desenvolupa un medicament per tractar una enfermetat i es suministra a 3 grups de persones amb 3 dosis diferents: Grup 1: Dosis alta, Grup 2: Dosis mitja, Grup 3: Dosis baixa."

Dades:



$$f(x) = \log(1 + e^x) = y - \text{card}$$

$$\text{Dosis} \cdot \text{pes} + \text{biaix} = x - \text{card}$$

$$0 \cdot (-34'4) + 2'14 = 2'14$$

$$f(2'14) = \log_e(1 + e^{2'14}) = 2'25$$

$$0'1 \cdot (-34'4) + 2'14 = -1'3$$

$$f(-1'3) = \log_e(1 + e^{-1'3}) = 0'24$$

$$0'05 \cdot (-34'4) + 2'14 = 0'42$$

$$f(0'42) = \log_e(1 + e^{0'42}) = 0'92$$

$$0'3 \cdot (-34'4) + 2'14 = -8'18$$

$$f(-8'18) = \log_e(1 + e^{-8'18}) = 0'0003$$

$$0 \cdot (-2'52) + 1'29 = 1'29$$

$$f(1'29) = \log_e(1 + e^{1'29}) = 1'53$$

$$0'1 \cdot (-2'52) + 1'29 = 1'04$$

$$f(1'04) = \log_e(1 + e^{1'04}) = 1'34$$

$$0'3 \cdot (-2'52) + 1'29 = 0'53$$

$$f(0'53) = \log_e(1 + e^{0'53}) = 1$$

$$0'6 \cdot (-2'52) + 1'29 = 0'22$$

$$f(0'22) = \log_e(1 + e^{0'22}) = 0'8$$

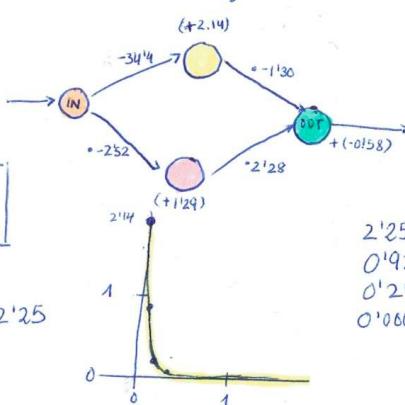
$$1 \cdot (-2'52) + 1'29 = -1'23$$

$$f(-1'23) = \log_e(1 + e^{-1'23}) = 0'25$$

$$0'05 \cdot (-2'52) + 1'29 = 1'16$$

$$f(1'16) = \log_e(1 + e^{1'16}) = 1'44$$

Suposem que ja s'ha entrenat el model i els pesos i biaixos ja estan calculats. Suposem que utilitzem la f.a. Soft Pln.

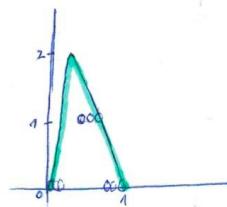


$$2'25 \cdot (-1'30) = -2'92$$

$$0'92 \cdot (-1'30) = -1'2$$

$$0'24 \cdot (-1'30) = -0'31$$

$$0'0003 \cdot (-1'30) = -0'0004$$



Total:

$$\begin{aligned} \text{INPUT } 0 &\Rightarrow -2'92 + 3'5 - 0'58 = 0 \\ \text{INPUT } 0'05 &\Rightarrow -1'2 + 3'3 - 0'58 = 1'55 \\ \text{INPUT } 0'1 &\Rightarrow -0'31 + 3 - 0'58 = 2'11 \\ \text{INPUT } 0'3 &\Rightarrow 0'0004 + 2'28 - 0'58 = 1'7 \\ \text{INPUT } 0'6 &\Rightarrow 0 + 1'8 - 0'58 = 1'22 \\ \text{INPUT } 1 &\Rightarrow 0 + 0'57 - 0'58 \approx 0 \end{aligned}$$

$$\begin{aligned} 1'44 + 2'28 &= 3'7 \\ 1'53 + 2'28 &= 3'8 \\ 1'34 + 2'28 &= 3 \\ 1 + 2'28 &= 2'28 \\ 0'18 + 2'28 &= 1'8 \\ 0'25 + 2'28 &= 0'57 \end{aligned}$$

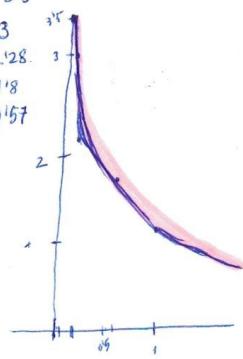
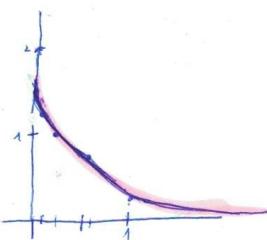


Figura 83: Per entendre en millor els càlculs de la propagació endavant d'una Xarxa Neuronal es va fer aquest exercici durant el període de documentació.

Planificació setmanal TFG		
29-05 Maig	Teoria: Fonaments del llenguatge natural i metodologia	-
06-12 Maig	Teoria: Fonaments del llenguatge natural	Pràctica: Model anàlisi de sentiments
13-19 Maig	Teoria: Anàlisi de sentiments.	Pràctica: Model anàlisi de sentiments
20-26 Maig	Teoria: Continuar anàlisi de sentiments.	Pràctica: Model de predicció de la puntuació
27-02 Juny	Teoria: Mètriques d'avaluació regresió.	Pràctica: Model de predicció de la puntuació
03-09 Juny	Teoria: Documentació anàlisi de sentiments (pràctica)	Pràctica: Freqüency Word model
10-16 Juny	Teoria: Documentació model de regresió (pràctica)	Pràctica: Freqüency Word model
17-23 Juny	Teoria: Documentació model de Word Freqüency model (pràctica)	-
24-28 Juny	Revisar contingut	

Taula 1: Planificació setmanal TFG

Referències

- [1] Balakrishnan Sathiyakugan. Learn Natural Language Processing from scratch, Jul 24, 2018. MEDIUM. [Consultat el 6 de maig de 2024]. Disponible a: <https://blog.goodaudience.com/learn-natural-language-processing-from-scratch-7893314725ff>. Citada en seccions: (document), 3.1.1, 4 i 3.2.3.
- [2] Avishek Choudhury. Schematic illustration of how natural language processing converts unstructured text to machine-readable structured data, which can then be analyzed by machine-learning algorithms., Jul 2020. ResearchGate. [Consultat el 26 de juny de 2024]. Disponible a: https://www.researchgate.net/figure/Schematic-illustration-of-how-natural-language-processing-converts-unstructured-text-to_fig1_343194021. Citada en seccions: (document) i 5.
- [3] Maleesha De Silva. Preprocessing Steps for Natural Language Processing (NLP): A Beginner's Guide!, Abril 30, 2023. Medium. [Consultat el 27 de juny de 2024]. Disponible a: <https://medium.com/@maleeshadesilva21/preprocessing-steps-for-natural-language-processing-nlp-a-beginners-guide-d6d9bf7689c9>. Citada en seccions: (document), 6 i 7.
- [4] Mohd Sanad Zakirizvi. BERT: A Comprehensive Guide to the Groundbreaking NLP Framework, 19 Gen, 2024. analiticavidhya. [Consultat el 28 de juny de 2024]. Disponible a: <https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/>. Citada en seccions: (document) i 8.
- [5] AIML. Explain the basic architecture of a Neural Network, model training and key hyperparameters, Març 8, 2024. AIML.com. [Consultat el 30 de maig de 2024]. Disponible a: <https://aiml.com/what-is-the-basic-architecture-of-an-artificial-neural-network-ann/>. Citada en seccions: (document), 9, 23, 24, 26 i 3.2.2.a.
- [6] Bill MacCartney. Understanding Natural Language Understanding, Jul 16, 2014. The Stanford NLP Group. [Consultat el 11 de maig de 2024]. Disponible a: <https://nlp.stanford.edu/~wcmac/papers/20140716-UNLU.pdf>. Citada en seccions: (document), 3.1.3.b i 10.
- [7] Fabio Chiusano. The Natural Language Processing Community, -. NPPlanet. [Consultat el 18 de maig de 2024]. Disponible a: <https://www.nlplanet.org/>. Citada en seccions: (document), 3.1.4, 11 i 3.2.1.b.
- [8] John Teleska. Context is the matrix of meaning, Dec 17, 2010. Blogspot [Consultat el 26 de juny de 2024]. Disponible a: <https://johnteleska.blogspot.com/2010/12/context-is-matrix-of-meaning.html>. Citada en seccions: (document) i 12.
- [9] AlexisPerrier. Introduction to Natural Language Processing, Dec 15, 2022. OPENCLASSROOMS. [Consultat el 18 de maig de 2024]. Disponible a: <https://openclassrooms.com/en/courses/6532301-introduction-to-natural-language-processing/8081284-apply-a-simple-bag-of-words-approach>. Citada en seccions: (document), 13, 3.2.1.a i 14.
- [10] Turing. A Guide on Word Embeddings in NLP, Gen 23, 2024. Commerce.ai. [Consultat el 18 de maig de 2024]. Disponible a: <https://www.turing.com/kb/guide-on-word-embeddings-in-nlp>. Citada en seccions: (document), 3.2.1.b, 15, 20 i 3.2.1.d.
- [11] Fabio Chiusano. Representing Texts as Vectors: Word Embeddings, -. NPPlanet. [Consultat el 19 de maig de 2024]. Disponible a: <https://www.nlplanet.org/course-practical-nlp/01-intro-to-nlp/11-text-as-vectors-embeddings>. Citada en seccions: (document), 16, 3.2.1.c, 17, 18 i 19.
- [12] Jason Brownlee. What Are Word Embeddings for Text?, Agost 7, 2019. Machine Learning Mastery. [Consultat el 19 de maig de 2024]. Disponible a: <https://machinelearningmastery.com/what-are-word-embeddings/>. Citada en seccions: (document), 21 i 3.2.1.d.
- [13] Atria. Qué son las redes neuronales y sus funciones, Mar 7, 2024. Atria. [Consultat el 25 de maig de 2024]. Disponible a: <https://atriainnovation.com/blog/>

- que-son-las-redes-neuronales-y-sus-funciones/. Citada en seccions: (document), 22 i 3.2.2.a.
- [14] ajitjaokar. An elegant way to represent forward propagation and back propagation in a neural network, Jul 27, 2019. Data Science Central. [Consultat el 29 de maig de 2024]. Disponible a: <https://www.datasciencecentral.com/an-elegant-way-to-represent-forward-propagation-and-back/>. Citada en seccions: (document), 25 i 3.2.2.a.
- [15] Rick Merritt. What Is a Transformer Model?, Març 25, 2022. blogs.nvidia.com. [Consultat el 31 de maig de 2024]. Disponible a: <https://blogs.nvidia.com/blog/what-is-a-transformer-model/>. Citada en seccions: (document), 3.2.2.b, 27 i 28.
- [16] Vyacheslav Efimov. Large Language Models: RoBERTa — A Robustly Optimized BERT Approach, Sept 25, 2023. Medium. [Consultat el 26 de juny de 2024]. Disponible a: <https://towardsdatascience.com/roberta-1ef07226c8d8>. Citada en seccions: (document) i 29.
- [17] MonkeyLearn. Sentiment Analysis: A Definitive Guide, -. MonkeyLearn. [Consultat el 14 de maig de 2024]. Disponible a: <https://monkeylearn.com/sentiment-analysis/>. Citada en seccions: (document), 30 i 3.3.2.
- [18] The Last of Us, 2024. The Last of Us Wiki [Consultat el 26 de juny de 2024]. Disponible a: https://thelastofus.fandom.com/wiki/The_Last_of_Us. Citada en seccions: (document) i 32.
- [19] Andrew Webster. The power of failure: Making 'The Last Of Us', Sep 19, 2013. The Verge. [Consultat el 4 de març de 2024]. Disponible a: <https://www.theverge.com/2013/9/19/4744008/making-the-last-of-us-ps3>. Citada en seccions: (document), 4.1.3, 4.1.4, 34 i 35.
- [20] Pablo Haya. La metodología CRISP-DM en ciencia de datos, Nov 29, 2021. Instituto de Ingeniería del conocimiento. [Consultat el 12 de maig de 2024]. Disponible a: <https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/>. Citada en secció: 2.3.
- [21] Juan Francisco Vallalta. CRISP-DM: una metodología para minería de datos en salud, Nov 4, 2019. Health Data Miner. [Consultat el 12 de maig de 2024]. Disponible a: <https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/>. Citada en secció: 2.3.
- [22] Arvindpdmn. Natural Language Processing, Feb 15, 2022. DEVOOPEDIA. [Consultat el 6 de maig de 2024]. Disponible a: <https://devopedia.org/natural-language-processing>. Citada en seccions: 3.1.2, 3.1.3.a i 3.1.4.
- [23] Alexander S. Gillis. natural language processing (NLP), Feb 2024. [Consultat el 7 de maig de 2024]. Disponible a: <https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP>. Citada en seccions: 3.1.3, 3.1.3.b i 3.1.6.
- [24] Teresa Alsinet. Notes de classe: Fases d'anàlisi. Processadors de Llenguatge. Departament d'Informàtica i Enginyeria Industrial. Lleida: Escola Politècnica Superior. 2024. Citada en secció: 3.1.3.a.
- [25] Awaldeep Singh. What is named entity recognition?, Dec 30, 2023. MEDIUM. [Consultat el 9 de maig de 2024]. Disponible a: <https://medium.com/@awaldeep/understanding-the-essentials-nlp-text-preprocessing-steps-b5d1fd58c11a>. Citada en secció: 3.1.3.a.
- [26] IBM. What is named entity recognition?, -. [Consultat el 9 de maig de 2024]. Disponible a: <https://www.ibm.com/topics/named-entity-recognition>. Citada en secció: 3.1.3.b.
- [27] Diego Santos. Procesamiento de lenguaje natural: qué es, ejemplos y herramientas, Oct 25, 2023. Hubspot. [Consultat el 7 de maig de 2024]. Disponible a: <https://blog.hubspot.es/marketing/procesamiento-de-lenguaje-natural>. Citada en secció: 3.1.4.
- [28] IBM. What is natural language processing (NLP)?, -. [Consultat el 7 de maig de 2024]. Disponible a: <https://www.ibm.com/topics/natural-language-processing>. Citada en secció: 3.1.4.

- [29] CodeAcademy. Bag of Words, -. CodeAcademy. [Consultat el 18 de maig de 2024]. Disponible a: <https://www.codecademy.com/learn/dscp-natural-language-processing/modules/dscp-bag-of-words/cheatsheet>. Citada en secció: 3.2.1.a.
- [30] CodeAcademy. Word Embeddings, -. CodeAcademy. [Consultat el 17 de maig de 2024]. Disponible a: <https://www.codecademy.com/learn/dscp-natural-language-processing/modules/dscp-word-embeddings/cheatsheet>. Citada en secció: 3.2.1.c.
- [31] Frederik Bussler. Amazon Product Review Analysis: The Ultimate Guide (2021), Maig 19, 2021. Commerce.ai. [Consultat el 18 de maig de 2024]. Disponible a: <https://www.commerce.ai/blog/amazon-product-review-analysis-the-ultimate-guide>. Citada en secció: 3.2.1.d.
- [32] StatQuest with Josh Starmer. Redes neuronales pt. 1: Dentro de la Caja Negra, Agost 31, 2020. YouTube. [Consultat el 28 de maig de 2024]. Disponible a: <https://www.youtube.com/watch?v=Cq0fi41LfDw>. Citada en secció: 3.2.2.a.
- [33] StatQuest with Josh Starmer. Redes neuronales (Parte 2): Ideas principales de propagación hacia atrás, Oct 19, 2020. YouTube. [Consultat el 28 de maig de 2024]. Disponible a: <https://www.youtube.com/watch?v=IN2XmBhILt4&t=30s>. Citada en secció: 3.2.2.a.
- [34] Giuliano Giacaglia. How Transformers Work, Mar 11, 2019. Medium. [Consultat el 22 de maig de 2024]. Disponible a: <https://towardsdatascience.com/transformers-141e32e69591>. Citada en secció: 3.2.2.b.
- [35] Manish Shivanandhan. BERT Explained – The Key to Advanced Language Models, Mar 4, 2024. Free Code Camp. [Consultat el 25 de maig de 2024]. Disponible a: <https://www.freecodecamp.org/news/bert-explained-the-key-to-advanced-language-models/>. Citada en secció: 3.2.2.c.
- [36] Colin Moriarty. Naughty Dog Officially Split Into Two Teams, Dec 12, 2011. IGN. [Consultat el 4 de març de 2024]. Disponible a: <https://www.ign.com/articles/2011/12/12/naughty-dog-officially-split-into-two-teams>. Citada en secció: 4.1.2.
- [37] Playstation. Grounded: The Making of The Last of Us, Feb 28, 2014. Youtube. [Consultat el 19 de març de 2024]. Disponible a: <https://www.youtube.com/watch?v=yH5MgEbB0ps>. Citada en secció: 4.1.2.
- [38] Edge Staff. The Last Of Us: the definitive postmortem – spoilers be damned, Jun 18, 2013. Edge Online. [Consultat el 4 de març de 2024]. Disponible a: <https://web.archive.org/web/20130621235831/http://www.edge-online.com/features/the-last-of-us-the-definitive-postmortem-spoilers-be-damned/>. Citada en seccions: 4.1.3 i 4.1.4.
- [39] Filmaffinity. The Last of Us (Serie de TV), Gen 16, 2023. Filmaffinity. [Consultat el 1 de juny de 2024]. Disponible a: <https://www.filmaffinity.com/es/film205905.html>. Citada en secció: 4.1.3.
- [40] Playstation. The Last of Us - Director's Video Blog - Gamescom 2012 Presentation, Jul 3, 2013. Youtube. [Consultat el 5 de març de 2024]. Disponible a: https://www.youtube.com/watch?v=BJFDAubW_XE&t=19s. Citada en secció: 4.1.4.
- [41] Playstation. The Inspirations for The Last of Us, Abr 5, 2012. Youtube. [Consultat el 5 de març de 2024]. Disponible a: <https://www.youtube.com/watch?v=xZkCBHmeeMg>. Citada en secció: 4.1.4.
- [42] Gaetano Prestia. The Last Of Us inspired by Ico, RE4, Jun 10, 2013. MMGM. [Consultat el 4 de març de 2024]. Disponible a: <https://web.archive.org/web/20130610034155/http://ps3.mmgn.com/News/the-last-of-us-inspired-by-ico-re4>. Citada en secció: 4.1.4.
- [43] DayoScript. The Last of Us [Anàlisis] - Post Script, Jul 3, 2013. Youtube. [Consultat el 4 de març de 2024]. Disponible a: https://www.youtube.com/watch?v=BJFDAubW_XE. Citada en secció: 4.1.5.
- [44] Playstation. From Dreams - The Making of The Last of Us: Left Behind, Feb 28, 2014. Youtube. [Consultat el 4 de març de 2024]. Disponible a: https://www.youtube.com/watch?v=v7WEeNH_C2I. Citada en secció: 4.1.6.a.

- [45] Eric Monacelli. Left Behind is now the highest rated PS3 DLC ever, Març 14, 2014. PlayStation Blog. [Consultat el 4 de març de 2024]. Disponible a: <https://blog.playstation.com/archive/2014/03/14/last-us-passes-6-million-sales/>. Citada en secció: 4.1.6.b.
- [46] Austin Wood. PlayStation Accidentally Reveals More Than 200 Millions of Development Costs for Horizon Forbidden West and The Last of Us Part 2, Juny 28, 2023. GamesRadar+. [Consultat el 1 d'abril de 2024]. Disponible a: <https://acortar.link/JyQgIU>. Citada en seccions: 4.2.2 i 4.2.4.
- [47] Sam White. The Last of Us Part II: how Naughty Dog made a classic amidst catastrophe, Juny 9, 2020. GQ Magazine. [Consultat el 31 de març de 2024]. Disponible a: <https://www.gq-magazine.co.uk/culture/article/the-last-of-us-part-ii-neil-druckmann-interview>. Citada en seccions: 4.2.2, 4.2.3 i 4.2.4.
- [48] Matthew Roberts. How Naughty Dog marketed The Last Of Us: Part II, Jul 29, 2020. Approval Studio. [Consultat el 5 d'abril de 2024]. Disponible a: <https://approval.studio/blog/how-naughty-dog-marketed-the-last-of-us-part-ii/>. Citada en secció: 4.2.2.
- [49] Ilaria Mangiardi. Reframing Love: Why The Last of Us Part 2 Still Has Major Impact, Dec 30, 2022. The Gamer. [Consultat el 2 d'abril de 2024]. Disponible a: <https://www.thegamer.com/reframing-love-relationships-the-last-of-us-part-2/>. Citada en secció: 4.2.3.
- [50] Emanuel Maiberg. The Not So Hidden Israeli Politics of 'The Last of Us Part II', Jul 15, 2020. VICE. [Consultat el 5 d'abril de 2024]. Disponible a: <https://www.vice.com/en/article/bv8da4/the-not-so-hidden-israeli-politics-of-the-last-of-us-part-ii>. Citada en secció: 4.2.4.
- [51] Melissa Sarnowski. The Last Of Us Part 2's Most Meta Easter Egg Connects To Game Of Thrones, Jul 17, 2022. ScreenRant. [Consultat el 5 d'abril de 2024]. Disponible a: <https://screenrant.com/last-us-part-2-game-thrones-easter-egg/>. Citada en secció: 4.2.4.
- [52] Joe Juba. The Last Of Us Part II Interview – Adding Depth, Staying Grounded, And The Cost Of Revenge, Juny 01, 2020. GameInformer. [Consultat el 5 d'abril de 2024]. Disponible a: <https://www.gameinformer.com/preview/2020/06/01/the-last-of-us-part-ii-interview-adding-depth-staying-grounded-and-the-cost-of>. Citada en secció: 4.2.4.
- [53] Dayoscript. La ambició desmedida de The Last of Us 2 [Análisis] - Post Script, Jul 22, 2020. Youtube. [Consultat el 2 d'abril de 2024]. Disponible a: <https://www.youtube.com/watch?v=Uk28T6vB33Y>. Citada en secció: 4.2.5.
- [54] Rob Zacny. 'The Last of Us Part II' Is a Grim and Bloody Spectacle, but a Poor Sequel, Juny 12, 2020. VICE. [Consultat el 2 d'abril de 2024]. Disponible a: <https://www.vice.com/en/article/wxqnxy/last-of-us-part-2-review>. Citada en secció: 4.2.5.
- [55] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France, June 2022. European Language Resources Association. Citada en seccions: 6.1 i 6.2.
- [56] Anònim. The Application of RoBERTa-large (PEFT) in Long Text Processing, Maig 27, 2024. Seduca.ai. [Consultat el 31 de maig de 2024]. Disponible a: <https://blog.seduca.ai/id/10457/>. Citada en secció: 6.2.
- [57] Noe Casas. Do transformers (e.g. BERT) have an unlimited input size?, Març 31, 2023. StackExchange. [Consultat el 05 de juny de 2024]. Disponible a: <https://datascience.stackexchange.com/questions/120601/do-transformers-e-g-bert-have-an-unlimited-input-size>. Citada en secció: 6.2.
- [58] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023. Citada en secció: 6.2.
- [59] Azul Fuentes, Dante Reinaudo, Lucía Pardo, and Roberto Iskandarani. Spanish Pre-Trained BERT Model and Evaluation Data, 2024. HuggingFace. [Consultat el 26 de juny de 2024]. Disponible a: <https://github.com/cjhutto/vaderSentiment>. Citada en secció: 6.2.

- [60] Rubert-tiny2 a model fine-tuned for sentiment classification, Agost 25, 2023. (Codi font). HuggingFace. [Consultat el 26 de juny de 2024]. Disponible a: <https://huggingface.co/seara/rubert-tiny2-russian-sentiment>. Citada en secció: 6.2.
- [61] Seamless Comunication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. Seamless: Multilingual expressive and streaming speech translation, 2023. ArXiv. [Consultat el 26 de juny de 2024]. Disponible a: <https://huggingface.co/facebook/seamless-m4t-v2-large>. Citada en secció: 6.3.
- [62] Jörg Tiedemann. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT In Proceedings of the Fifth Conference on Machine Translation, pages 1174–1182. Online, Nov 2020. Association for Computational Linguistics [Consultat el 26 de juny de 2024]. Disponible a: <https://www.aclweb.org/anthology/2020.wmt-1.139>. Citada en secció: 6.3.
- [63] Awaldeep Singh. Understanding the Essentials: NLP Text Preprocessing Steps!, Dec 30, 2023. Medium. [Consultat el 26 de juny de 2024]. Disponible a: <https://medium.com/@awaldeep/understanding-the-essentials-nlp-text-preprocessing-steps-b5d1fd58c11a>. Citada en secció: 6.3.
- [64] Cristhian Boujon. How to list the most common words from text corpus using Scikit-Learn?, Març 18, 2018. Medium. [Consultat el 26 de juny de 2024]. Disponible a: <https://medium.com/@cristhianboujon/how-to-list-the-most-common-words-from-text-corpus-using-scikit-learn-dad4d0cab41d>. Citada en secció: 6.3.
- [65] Byte Insights. Exploring Fascinating Insights with Word Frequency Analysis, Jun 20, 2023. Medium. [Consultat el 26 de juny de 2024]. Disponible a: <https://medium.com/@ByteInsights/exploring-fascinating-insights-with-word-frequency-analysis-5da2113df864>. Citada en secció: 6.3.
- [66] C. J. Hutto. Vader-sentiment-analysis, (2014). [Codi font]. GitHub. [Consultat el 26 de juny]. Disponible a : <https://github.com/cjhutto/vaderSentiment>. Citada en secció: 6.3.
- [67] Datos.gob.es. ¿Cómo sé si mi modelo de predicción es realmente bueno?, Gen 26, 2021. datos.gob.es. [Consultat el 23 de juny de 2024]. Disponible a: <https://datos.gob.es/es/blog/como-se-si-mi-modelo-de-prediccion-es-realmente-bueno>. Citada en secció: 7.2.
- [68] Minitab Blog Editor. Análisis de Regresión: ¿Cómo Puedo Interpretar el R cuadrado y Evaluar la Bondad de Ajuste?, Abr 4, 2019. Minitab. [Consultat el 23 de juny de 2024]. Disponible a: <https://blog.minitab.com/es/analisis-de-regresion-como-puedo-interpretar-el-r-cuadrado-y-evaluar-la-bondad-de-ajuste>. Citada en secció: 7.2.