Tendència Geopolítica en Temps Real

Predicció electoral a través de les preocupacions socials extretes de la premsa, mitjançant tècniques de *BDA in streaming, NLP,* i *ML*.

Maria Roser Santacreu Gou

Analytics for Data Streaming

Àrea 1: *Data Analytics in Industrial and Business Environments*

Tutor/a de TF

Rafael Luque Ocaña

Professor/a responsable de l'assignatura

Susana Acedo Nadal

Juny 2025

A la Universitat Oberta de Catalunya, als meus tutors, professors i col·laboradors pel seu suport, dedicació i guia al llarg d'aquest camí. La seva aportació ha estat fonamental per a la realització d'aquest treball final de màster.

Al meu home per motivar-me constantment, i ajudar-me a fer realitat els meus somnis. A la mare per la creativitat, força i resiliència, que em va transmetre, i al pare per la seva persistència i aquelles meravelloses tardes de cafè on intercanviàvem opinions diverses de política intentant arreglar el món.

Fitxa del Treball Final

Títol del treball:	Tendència Geopolítica en Temps Real
Nom de l'autor/a:	Maria Roser Santacreu Gou
Nom del Tutor/a de TF:	Rafael Luque Ocaña
Nom del/de la PRA:	Susana Acedo Nadal
Data de lliurament:	06/2025
Titulació o programa:	TFM MUCD
Àrea del Treball Final:	Àrea 1: *Data Analytics in Industrial and Business Environments*
Idioma del treball:	Català
Paraules clau	Processament de llenguatge natural (*NLP*) Anàlisis de dades massives en temps real (*BDA in Streaming*) Aprenentatge automàtic (*ML*)

Resum del Treball

Avui dia amb el *Big Data* i les noves tecnologies la següent dita ja no hauria de tenir sentit:

> "L'home és l'únic animal que ensopega dues vegades amb la mateixa pedra"[1]

Si s'interpreta com que no es pot recordar tot el que s'aprèn ni la història que precedeix, en futures situacions similars es caurà en els mateixos errors.

Per tant, si globalment s'analitza l'evolució de les notícies de diferents mitjans de comunicació i països, es pot preveure el futur i no ensopegar-hi de nou?

En aquest sentit, aquest treball pren com a fonament l'extracció de les preocupacions de la població de diferents països, utilitzant com a fonts de dades les notícies emeses per diversos mitjans de comunicació, juntament amb els resultats electorals d'alguns països. On mitjançant el *NLP*, l'anàlisi en temps real i el *ML*, es pronosticarà els resultats electorals.

Donat que en molts casos no tindrem els resultats electorals i prou dades antigues,

s'agruparà els països segons les seves preocupacions en períodes de temps concrets, i s'hi aplicarà el model entrenat amb països similars. Aquest pas inclou la dificultat de tractar amb diferents llengües amb les quals rebrem les notícies, per tant, es traduiran a l'anglès durant la ingesta.

Addicionalment, agrupar els països per preocupacions similars dona l'oportunitat de visualitzar informació extra a l'usuari final, el qual podrà observar i aprofundir en les diferències de temes concrets entre països, i reflexionar sobre com ho van solucionar realitzant investigacions addicionals. Per exemple: si un país no té absència escolar i el nostre sí, aquest fet dona lloc a una investigació addicional extra de com es va resoldre.

Si bé la idea principal és pronosticar els resultats electorals, la visualització de sèries temporals de les principals preocupacions dels diferents països, en distintes granularitats i períodes, dona una trajectòria de l'evolució dels països que pot ser molt rellevant, i de la qual se'n pot aprendre i avançar-se al futur en situacions inicials similars.

Abstract

Nowadays, with Big Data and new technologies, the following saying should no longer make sense:

> *"Man is the only animal that trips twice over the same stone"* [1]

If this is interpreted as the humanity cannot remember everything learned or the history that precedes us, then in the same situations same mistakes will be repeated.

Therefore, if the evolution of news from different media and countries are analyzed, can one predict the future and avoid trips again?

In this point, this work is based on the population's concerns extraction of different countries, using the news reports from various media sources, in conjunction with the old election results of some countries, and by applying through NLP, real-time analysis, and ML, the election results will be predicted.

Since election results will not be available for many countries and also there will not be enough old news, countries will be grouped according on their similarity concerns in a specific time periods, and the model trained in similar country will be applied in these similar countries to predict next elections results. This step includes the difficulty of dealing with different languages in which we will receive the news; so all news will be translated into English during ingestion.

Additionally, grouping countries by similar concerns provides to the finally user the opportunity to view additional information, allowing them to observe and delve deeper into the differences in specific topics across countries and reflect on how they fixed them doing an additional research. For example, if one country doesn't have school absences and ours does, this fact leads to further investigation into how it was resolved.

While the primary goal is to predict election results, visualizing time series of the main concerns of different countries, at different granularities of periods, provides a trajectory of the countries' evolution. And this can be very relevant, offering valuable insights that enable learning and anticipating the future in similar early-stage situations.

Índex

Llista de Figures

Llista de Taules

1. Introducció

A continuació es descriu el context del treball final, la rellevància que pot tenir a la societat, la motivació que porta el desenvolupament d'aquest projecte, així com els objectius principals i secundaris.

També es defineix l'impacte que pot tenir en la sostenibilitat ètica-social, en la diversitat de l'àmbit de la competència de compromís ètic i global (CCEG), i els objectius de desenvolupament sostenibles (ODS).

Tot seguit s'explica la metodologia de treball que se seguirà durant el desenvolupament del projecte, es defineix l'arquitectura d'alt nivell, eines i flux principal. Finalment, es determinen les tasques, les seves dependències, la planificació temporal, i els riscos que poden afectar tant a la planificació, com a la qualitat del resultat del model de predicció.

1.1 Context i justificació del Treball

Si s'observa el present i la trajectòria de la geopolítica, sembla que s'ha oblidat el passat i no es tem les conseqüències d'alguns actes. Sense tenir en compte l'egoisme d'alguns dirigents, en països amb democràcia en primera instància, es el poble el que escull a les persones que governen. Per tant, en aquesta línia es pretén facilitar una trajectòria de la història a tothom, de la mateixa forma que els progenitors transfereix la seva saviesa als seus descendents per a que tingui un millor futur. Així doncs, aquest treball pretén transmetre coneixement i saviesa a la població, per conscienciar del futur que es pot tenir basant-se en les decisions que s'estan prenent.

Què li preocupa a la societat perquè prengui la decisió de votar un partit polític o altre? Quins són aquests termes (preocupacions) que fan decantar la balança a posicions polítiques de dreta i esquerra?

En certa manera el que vol el poble és millorar la seva forma de vida, però si el coneixement i la saviesa únicament els tenen les persones que estan al capdavant; a cas no tenen avantatja sobre el poble?, a cas la informació no és poder com bé sempre s'ha sentit dir, doncs compartim-la!

En tot cas, si bé s'acostuma a gravar amb foc els successos traumàtics, que passa amb els que no ho són? O que passa amb les persones que no han tingut aquest trauma? Així i tot, en ocasions similars podran evitar-ho? I si hi ha transferència de coneixement? Per exemple:

> *"De segur que donada la desgràcia de les inundacions de València, moltes persones que vulguin comprar una casa demanaran, abans de prendre la*

decisió, un informe de riscos d'inundacions. Però, per què es va construir en terrenys inundables? I per què es va permetre?"

Tanmateix, que passa a mesura que el temps passa? Anem oblidant perquè no ho podem recordar tot, tenim d'altres preocupacions..., i les noves generacions no ho recordaran.

De manera que, tenir la facilitat de consultar la informació mitjançant una línia temporal que mostra les preocupacions de diferents països, amb una visualització en temps real, dinàmica, interactiva, que permeti filtrar per països, temes, amb diferents granularitats de temps, i que també mostri la predicció dels resultats de les següents eleccions polítiques partint de les preocupacions actuals, pot ser molt rellevant per a la població i la presa de decisions.

1.2 Explicació de la motivació personal

En record a les converses de persones que es preocupen pels seus essers estimats, que sempre explicaven i expliquen la seva vida i experiències, per a que les tinguis presents i no caiguis en els errors que van caure, perquè volen que a tu et vagi millor. Històries personals que t'expliquen repetidament, malgrat que a vegades es fan pesades, saps que és per a que et puguis preparar per afrontar contratemps futurs, poder evitar-los o inclús canviar-los.

Aquesta transferència de saviesa que passa de persones a persones és la que em motiva, i que avui dia, amb la ingent quantitat de dades que tenim i generem, les quals podem processar, tractar, filtrar, de les que obtenim informació, coneixement i finalment saviesa, em pregunto: podríem preveure els següents passos en l'evolució d'un país? Es podria millorar la presa de decisions basant-nos en models científics de transferència d'experiència? Aquesta saviesa a l'abast de tots podria conscienciar la població per tal de millorar el nostre futur i no caure en errors passats?

En resum, podem amb el *Big Data* i el *Machin learning* millorar el món?

Si bé la motivació personal abraça un repte fora del meu abast, en soc optimista i crec que sí que és possible. Tot i que un altre repte és el de les persones que en última instància prenen les decisions, les quals en lloc de mirar per un bé comú, facin cas omís de les prediccions del model final per motius egoistes.

1.3 Objectius del Treball

Com a objectiu principal s'estableix que segons les notícies dels mitjans de comunicació conjuntament amb els resultats electorals d'anys anteriors, es crearà un model que pronostiqui els següents resultats electorals, resumint amb dreta o esquerra (tal com ho entenem a qui a Espanya).

Com probablement no es disposarà de tota la informació dels resultats electorals dels diferents països dels quals rebrem les notícies i de les notícies antigues, s'intentarà extrapolar els resultats. Per tant, com a objectiu secundari es farà una classificació de països mitjançant les similituds dels continguts dels titulars de les notícies, les quals es creu que reflecteixen les preocupacions de la població, i donat que hi intervenen dades de diferents països, per tal de cercar les similituds s'haurà de traduir a un idioma comú que serà l'anglès.

Finalment, un cop agrupats els països per similituds, s'utilitzarà el model de pronòstic electoral en aquells clústers de països on es tingui algun país amb resultats electorals i suficients notícies antigues, les quals s'hagin pogut fer servir per a l'entrenament del model.

Tant els pronòstics electorals, com els resultats de les similituds de les preocupacions dels països, es visualitzaran en temps real amb un decalatge de 24 hores per qüestions tècniques, en un *dashboard* on l'usuari final podrà filtrar i agrupar per països, temes i data.

Sens dubte, els models de pronòstic i agrupacions, s'hauran d'anar reentrant amb la nova informació cada cert temps, perquè el pronòstic electoral tingui un bon percentatge d'encert.

1.4 Impacte en sostenibilitat, ètic-social i de diversitat

Donat que l'objectiu principal és esbrinar les preocupacions dels diferents països i basant-se en aquestes pronosticar la posició política dels resultats electorals, i el secundari és classificar els països segons les similituds de les seves preocupacions, queda pales que s'ha d'establir durant totes les fases del projecte (disseny, desenvolupament i conclusions), la competència de compromís ètic i global (CCEG), ja que afecta a les tres dimensions que el CCEG inclou: sostenibilitat, comportament ètic i responsabilitat social, i diversitat i drets humans. A continuació es reflexionarà sobre cada dimensió establint els impactes positius i negatius, i en aquest últim cas com es poden minimitzar, on s'inclouran els objectius i metes de desenvolupament sostenible (ODS) per al 2030, tal com es defineixen a les Nacions Unides [2].

Dimensió de sostenibilitat

En un escenari ideal, i independentment del pronòstic electoral, si els usuaris revisen el passat i la seva trajectòria (preocupacions), i observen que els resultats no són del seu gust, en situacions similar podran prendre la decisió de canviar-lo avanç que es reprodueixi. Per exemple, si ens fixem en la contaminació de l'aigua del Mar Menor, i que ha portat a aquesta situació, en situacions i ubicacions similars les persones que prenen decisions poden preveure i evitar aquesta catàstrofe.

Per tant, depenent del tema que es pugui cercar a les línies temporals de preocupacions dels diferents països, que sortiran a la visualització del projecte, i de l'ús que en facin les

persones que decideixen, poden millorar l'impacte mediambiental en aquest exemple. Així mateix, depenent del tema a investigar, més una intervenció positiva de l'usuari per a decidir millorar el futur, intervindrien tots els ODS definits per aquesta dimensió:

- ODS 7 - Energia assequible i neta
- ODS 9 - Indústria, innovació i infraestructura
- ODS 11 - Ciutats i comunitats sostenibles
- ODS 12 - Consum i producció responsables
- ODS 13 - Acció pel clima
- ODS 14 - Vida sota l'aigua
- ODS 15 - La vida a la terra

En conclusió, l'impacte positiu en la dimensió sostenibilitat ha de ser dut a terme per l'usuari que extreu la saviesa del projecte.

Pel que fa a l'impacte negatiu, tret del mal ús de la informació que en pugui fer l'usuari final, per exemple prioritzant motius personals en detriment del benefici global, únicament s'observa els recursos energètics i tecnològics necessaris per a l'elaboració i funcionament del projecte. Tot i que el desenvolupament d'aquest projecte es durà a terme amb una maquina local, tret de la visualització (*dashboard* que podrà ser consultat des d'una URL pública), la idea final seria el ús de serveis *cloud* escalables per poder gestionar el màxim volum de notícies de tots els països del món, i així tenir una visió mundial de la trajectòria de la humanitat.

Així mateix, en aquest context l'empremta ecològica augmentarà, tot i que les pròpies eines que s'utilitzen per al desenvolupament d'aquest projecte *Big Data*, més la configuració adequada, la gestió eficient del tractament de les dades i les millores en els centres de dades per a reduir el consum d'energia, anomenats *Green Data Centers*, en minimitzen l'impacte negatiu.

Dimensió de comportament ètic i de responsabilitat social

Si bé cada país evoluciona al seu ritme, aquest ritme es pot accelerar si tenim mostres de com han evolucionat països més avançats. Es clar que sempre dependrà de les persones que dirigeixen un país, una institució o empresa, prendre com a mostra la trajectòria fructuosa d'altres i aplicar-ne els passos.

Però que passa quan el poble també té accés a la mateixa informació per a poder evolucionar i millorar el seu futur? De segur que les persones dirigents hauran d'acabar fent un canvi a favor del poble, ja que al final el poble els supera en número, i sens dubte que ningú vol pobresa ni fam, i tothom vol com a mínim un treball digne, pau i justícia, per tant, la informació en mans de tots contribueix a un bé comú. Amb tot això en aquesta dimensió també hi intervenen tots els ODS definits:

- ODS 1 - Sense pobresa
- ODS 2 - Zero fam
- ODS 6 - Aigua neta i sanejament

- ODS 8 - Treball digne i creixement econòmic
- ODS 16 - Pau, justícia i institucions fortes

Amb relació a l'impacte negatiu, igual que en la dimensió anterior, tenim el mal ús de la informació que en pugui fer l'usuari final, per tant, per minimitzar-lo la visualització final de projecte serà pública per a que tot poble hi tingui accés, i no solament les persones que prenen decisions en l'espai institucional o governamental.

Dimensió diversitat de gènere i drets humans

Tot i que la idea principal del projecte és extreure les preocupacions de les notícies rebudes de diferents mitjans de comunicació de diferents països, aquestes notícies poden incloure desigualtat de gènere, diversitat o incompliment dels drets humans, degut probablement a la cultura en si, de cada país, i a la divulgació de dades personal, com per exemple noms propis, o ideologies subjacents del mateix mitjà de comunicació. A més, la comparació en sí de preocupacions entre països pot influir en l'autoestima de viure en un país subdesenvolupat i potenciar la discriminació. Per tant, en aquesta última dimensió de CCEG també hi intervenen els ODS definits:

- ODS 5 - Igualtat de gènere
- ODS 10 - Reducció de les desigualtats

Així mateix, per tal de minimitzar aquest impacte negatiu inherent a la mateixa solució i que aporta la mateixa font de dades, durant el procés de desenvolupament i transformació de les dades s'eliminaran els noms propis, referències a institucions, així com la lematització de les paraules. Tot i que aquestes solucions probablement no serà suficient per exhaurir el problema, s'implementarà una revisió constant per tal de continuar minimitzant aquest impacte establint noves mesures per intentar no ferir la sensibilitat de cap col·lectiu.

Pel que fa als impactes positius, com bé s'ha dit que: la informació és poder, dependrà en si mateix de la utilitat i sentit que li vulgui donar l'usuari final. Així i tot, tenir una visió àmplia del que passa a altres països, sempre pot enriquir el coneixement i consegüentment millorar en futur.

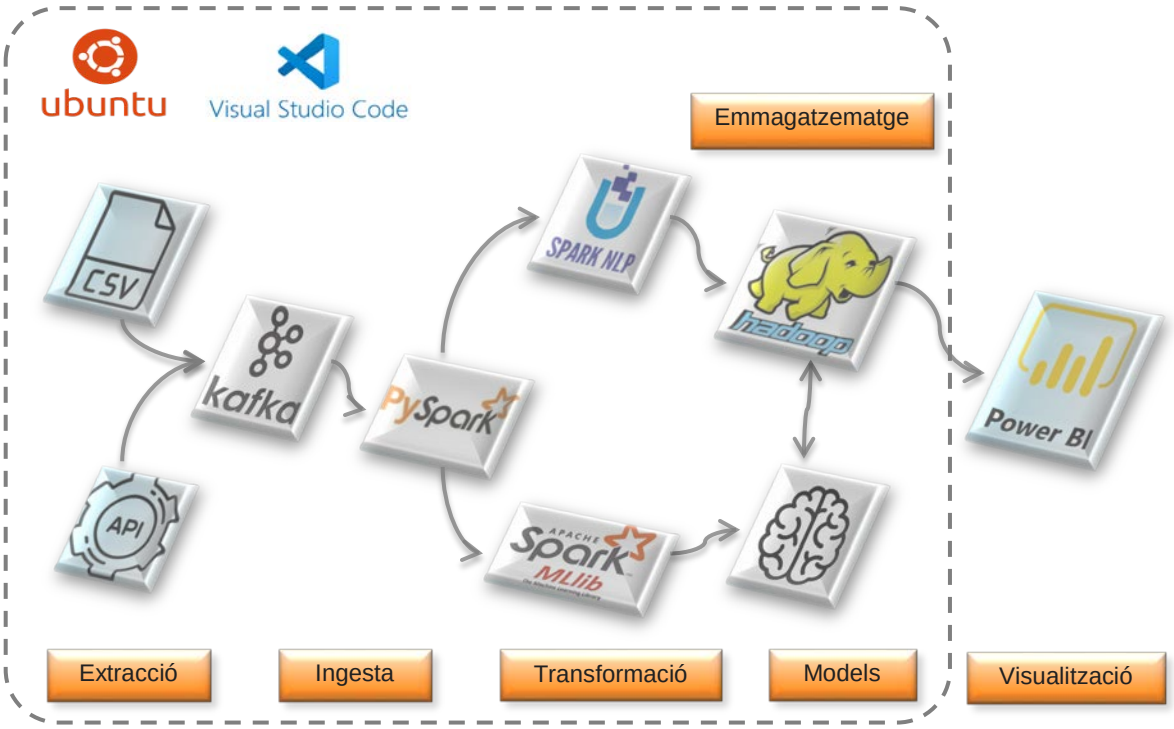
1.5 Enfocament i mètode seguit

Pel que fa a l'estratègia per al desenvolupament del projecte, d'entre els diferents mètodes que existeixen com: la metodologia en cascada, àgil, PRINCE2, PMBOK®, entre d'altres, s'utilitzarà una metodologia àgil per a la creació del nou producte, ja que l'objectiu és clar però la solució no del tot. Per tant, les etapes del cicle de vida de les dades: captura, emmagatzematge, processament, anàlisis, visualització i publicació, seran cíclicament millorades per tal de consolidar un producte bàsic i funcional tan aviat com es pugui, el qual s'anirà millorant fins a aconseguir el producte final desitjat.

Tot i que l'objectiu final del producte és clar, una visualització en temps real de l'evolució de les preocupacions, països assimilats, i la predicció de resultats electorals, s'ha seleccionat aquest tipus de desenvolupament perquè els mètodes àgils utilitzen controls empírics, els quals abracen la variabilitat donada la indeterminació i complexitats que poden sorgir.

En aquest sentit, a continuació es mostra un primer diagrama d'alt nivell per tenir una visió global, amb el qual es pretén informar de l'arquitectura del projecte, eines que s'usaran i el flux de les dades. S'ha de tenir present que, com el desenvolupament es farà en local, tret de *Power BI*, únicament es tindrà un broker de Kafka, per tant, no hi haurà rèpliques de la informació.

Figura 1: Arquitectura d'alt nivell, eines i flux principal



Font: Elaboració pròpia

1.6 Planificació del Treball

La següent taula descriu els recursos necessaris per realitzar el projecte:

Taula 1: Recursos necessaris

Nom	Versió
Windows 11 Pro i7 RAM 16 GB	24H2
Visual Studio Code	1.97.2

Ubuntu	24.04.1 LTS (GNU/Linux 5.15.167.4-microsoft-standard-WSL2 x86_64)
Java OpenJDK	64-Bit Server VM Corretto-11.0.26.4.1
Python	3.12.3
Apache Kafka	3.9.0
Apache Hadoop	3.4.1

Font: Elaboració pròpia

Les tasques a realitzar es descriuen a la següent taula, així com la data inicial, final i la duració.

Taula 2: Descripció de les tasques

Tasques		Data Inici	Data Fi	Duració
PAC 1		**28/02/25**	**08/03/25**	**9**
0. Preparació de l'entorn		**28/02/25**	**08/03/25**	**9**
0.1	**Instal·lació *Visual Studio Code***	28/02/25	28/02/25	1
0.2	**Instal·lació *Ubuntu* via *WSL***	01/03/25	01/03/25	1
0.3	**Instal·lació *Java***	02/03/25	02/03/25	1
0.4	**Instal·lació *Apache Kafka*** Creació del clúster amb un únic broker.	03/03/25	05/03/25	3
0.5	**Instal·lació *Apache Hadoop***	06/03/25	08/03/25	3
PAC 2		**09/03/25**	**29/03/25**	**21**
1. Extracció de dades		**09/03/25**	**21/03/25**	**13**
1.1	**CSV resultats electorals** Cercar els històrics dels resultats electorals dels països dels quals es rebran les notícies des de l'*API*, ideal seria mínim de tres països.	09/03/25	10/03/25	2
1.2	***News* CSV inicial** Cerca de fitxer inicial amb milers de notícies de diferents països amb estructura similar a les que es rebran de l'*API*, per tal de tenir un *dataset* base prou gran per a entrenar els models de pronòstic i agrupació.	09/03/25	10/03/25	2
1.3	**Descàrrega via API cada 24 hores.** Desenvolupar el codi per descarregar diàriament les notícies des de l'*API*.	11/03/25	15/03/25	5

Tasques		Data Inici	Data Fi	Duració
1.4	**Filtratge inicial** Netejar les dades de camps innecessaris per al projecte.	16/03/25	17/03/25	2
1.5	**Preprocessament inicial** Traduir a l'anglès les notícies provinents de països no angloparlants.	18/03/25	21/03/25	4
2. Ingesta de dades *Kafka*		**22/03/25**	**29/03/25**	**8**
2.1	**Creació de tòpics** Definició i creació dels tòpics.	22/03/25	23/03/25	2
2.2	**Productor** Creació del productor i desenvolupament de la ingesta al tòpic.	24/03/25	25/03/25	2
2.3	**Consumidor** Creació del consumidor, subscripció als tòpics, model de classificació preentrenat i guardar resultats a local, i les *news* a *HDFS*.	26/03/25	29/03/25	4
PAC 3		**31/03/25**	**04/05/25**	**35**
3. Processament en temps real *PySpark Spark NLP*		31/03/25	11/04/25	12
3.1	**Transformació amb *PySpark Pipeline*** Neteja de les dades (titulars i descripció de les notícies), expandir paraules, eliminar puntuacions, espais extra i *stop words*, *tokenizer* i eliminar segons quines categories de paraules com determinants, noms propis..., lematitzar.	31/03/25	03/04/25	4
3.2	**Extracció de característiques** Crear noves característiques, com fer recompte de paraules (preocupacions) per dia, setmana, mes i any, per cada país.	04/04/25	07/04/25	4
3.3	**Emmagatzematge *Hadoop HDFS*** Crear un esquema i emmagatzemar les noves característiques.	08/04/25	11/04/25	4
4. Creació dels models *PySpark ML*, *Pipeline*		12/04/25	20/04/25	9
4.1	**Model de predicció** Vectoritzar característiques, preparar data set entrenament, prova i test i crear el model de predicció.	12/04/25	15/04/25	4

Tasques		Data Inici	Data Fi	Duració
4.2	**Model d'agrupació** Crear el model d'agrupació de països segons similitud de característiques (preocupacions vectoritzades), en el mateix període de temps.	16/04/25	19/04/25	4
4.3	**Emmagatzematge *Hadoop HDFS* Resultats** Emmagatzemar resultats de les prediccions electorals de països assimilats amb el model d'agrupació, i també els resultats de les agrupacions de països per tal de mostrar-ho a la visualització.	20/04/25	20/04/25	1
5. Visualització *Power BI real time*		21/04/25	04/05/25	14
5.1	**Connexió *fonts de dades.*** Crear la connexió de les dades local.	21/04/25	22/04/25	2
5.2	**Mostrar predicció** De les pròximes eleccions dels països assimilats al de l'entrenament del model de classificació de països per similituds de preocupacions.	23/04/25	24/04/25	2
5.3	**Mapa dels països** Assimilats colors dels països segons predicció electoral pot mostrar les 5 principals preocupacions dins de cada país amb un pop-up o llegenda).	25/04/25	26/04/25	2
5.4	**Línia temporal de les preocupacions** Amb zooms, anuals, mensuals, setmanals, diaris.	27/04/25	29/04/25	3
5.5	**Afegir filtres per països.**	30/04/25	30/04/25	1
5.6	**Afegir filtres per termes de preocupacions.**	01/05/25	01/05/25	1
5.7	**Assimilar els colors** Dels mateixos termes (preocupacions) entre països (així es pot observar les similituds entre països).	02/05/25	03/05/25	2
5.8	**Definir autoactualització cada 24 hores**	04/05/25	04/05/25	1
PAC 4.0 - Redacció de la memòria (Preliminar)		**05/05/25**	**18/05/25**	**14**

PAC 4.1 - Redacció de la memòria (Final)	**19/05/25**	**25/05/25**	**7**
PAC 4.2 - Presentació audiovisual del treball	**26/05/25**	**03/06/25**	**9**
PAC 5.1 - Lliurament de la documentació al tribunal	**04/06/25**	**06/06/25**	**3**
PAC 5.2 - Defensa pública del treball	**07/06/25**	**27/06/25**	**21**

Font: Elaboració pròpia

Mitjançant el següent diagrama de Gantt es mostra la planificació temporal de cada tasca. On les fites parcials de cadascuna de les PAC esdevenen l'ampliació incremental dels subapartats de les 5 fases presentades del projecte. D'aquesta forma s'assegura el correcte flux de dades entre fases i eines, el que minimitza en estats avançats del projecte haver de canviar d'eines per incompatibilitats. Tanmateix, a continuació també es defineixen les tasques generals a realitzar dins de cada prova d'avaluació continuada.

PAC 01 - Entorn i definició i planificació del treball.

0. Preparació de l'entorn.

PAC 02 - Extracció i Ingesta, i l'estudi de l'art.

1. Extracció de dades.
2. Ingesta de dades *Kafka*.

PAC 03 - Processament, models i visualització.

3. Processament en temps real *PySpark, Spark NLP*.
4. Creació dels models *Spark ML Pipeline*.
5. Visualització *Power BI real time*.

PAC 4.0 - Redacció de la memòria (Preliminar).

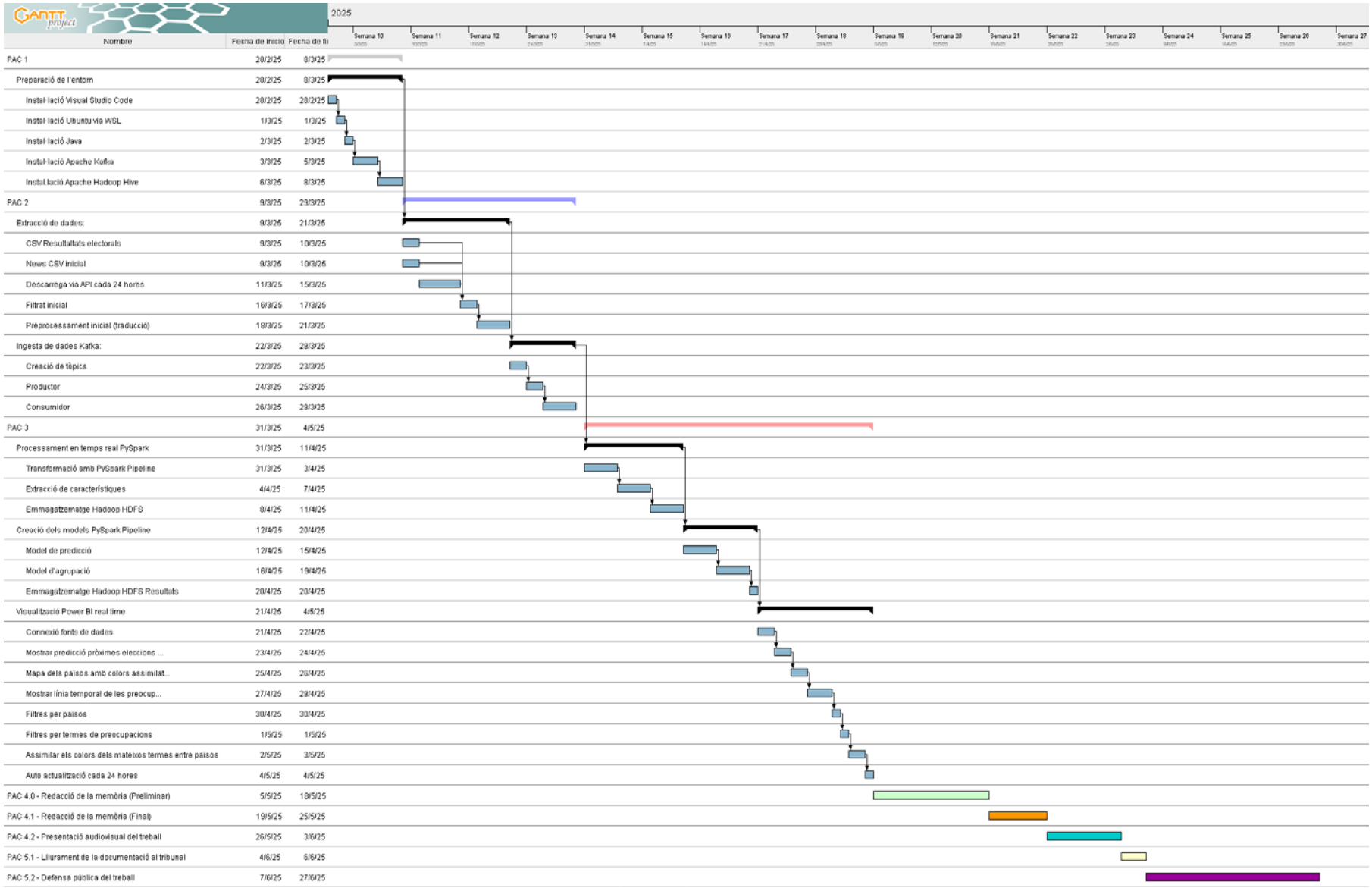
PAC 4.1 - Redacció de la memòria (Final).

PAC 4.2 - Presentació audiovisual del treball.

PAC 5.1 - Lliurament de la documentació al tribunal.

PAC 5.2 - Defensa pública del treball.

Figura 2: Planificació temporal - Diagrama Gantt



Font: Elaboració pròpia

A continuació es detallen els possibles riscos que es poden presentar i afectar el seguiment de la planificació establerta.

- Incompatibilitats entre eines com per exemple que *Power BI* web no funcioni correctament amb la cerca de dades a la instal·lació d'*Apache Hive* local. En tal cas es provaria de fer la visualització amb *Tableau,* tot i que s'hauria de revisar que *Tableau* actualitza automàticament els gràfics cada dia amb les noves dades processades. Si finalment la visualització web no és possible, s'implementaria amb la instal·lació d'escriptori de *Power BI* o *Tableau*.

- Amb relació al model predictiu dels resultats electorals hi ha el risc que l'avaluació no doni resultats prou significatius. En aquest cas es provaria de modificar el conjunt de dades, incloent o excloent més categories de paraules (preocupacions i d'altre contingut). Així i tot, si finalment els resultats no són rellevants, és a dir, el pronòstic és aleatori en aquells països on tenim les dades dels resultats electorals, donaria lloc a pensar que, amb els titulars de les notícies i una descripció no és suficient per predir els resultats de les pròximes eleccions. Per tant, una segona investigació seria afegir noves característiques com per exemple els partits que han governat als països limítrofs, no obstant això, aquest últim repunt quedaria fora de l'abast del projecte actual, ja que amb el *timing* actual no podem recuperar tota la història electoral de tots els països, en conseqüència quedaria com una millora en futures versions.

- Pel que fa a les fonts de dades s'ha de tenir present que provenen de dues *API*, de les quals s'extrauran les notícies actuals amb un usuari amb perfil gratuït, per tant, es té límit de descàrrega diària, i a més no se sap si prioritzen en enviar certes notícies i de certs mitjans de comunicació, en conseqüència podrien estar esbiaixades. En tot cas, això no endarrereix la planificació però si la qualitat final dels models, i per això la solució seria més complicada, tot i que es podria cercar diferents mitjans de comunicació de cada país i fer un *web scraping* diari dels titulars, però amb el *timing* que es té no és factible, a més possiblement en moltes webs no es permet i seria il·legal. També es podria fer un *upgrade* del perfil de les *API* i descarregar més notícies diàries de cada país, així hi hauria més probabilitat d'ampliar les categories i la procedència de mitjans de comunicació, però la quota mensual queda fora d'abast. Per tant, es confiarà en el fet que les *API* retornin notícies aleatòries.

- Un altre punt complicat pel que fa a l'extracció de preocupacions dels països, es que a les mateixes notícies de cada país no es parli únicament dels esdeveniments del mateix país, sinó també d'altres països. Aquest fet fa pensar que no són preocupacions en si del país que notifica la notícia, sinó del que se'n parla. En tot cas es podria descartar les notícies que parlen d'altres països, però per no descartar mostres i característiques, a priori es consideraran igualment, ja que també pot ser un punt rellevant o preocupant per la població del mateix país saber que està passant als països veïns.

- Pel que fa al model de similitud entre països per les preocupacions, com a risc afegit es té que les traduccions automàtiques de les notícies no siguin precises, i encara que la traducció es fa de la notícia completa abans de ser processada, per a que les paraules no perdin el significat contextual, una mala traducció pot afectar els resultats del model de predicció de posició política, i no assimilar correctament els països per les preocupacions reals. Com a alternativa es podria fer ús d'un altre traductor automàtic gratuït i comparar els resultats de les traduccions amb un model de similitud de significat de frases, o fer servir un de pagament, però aquesta última solució augmentaria el cost del projecte, i per aquest motiu queda descartada.

1.7 Breu sumari de productes obtinguts

Els productes obtinguts en aquest treball seran en última instància una visualització que mostrarà els resultats del processament de la ingesta massiva de dades, notícies provinents de diversos països, les quals fluiran per un sistema que les netejarà i transformarà, extraient les característiques apropiades per a entrenar el model de predicció de la posició política, el qual farà prediccions diàries simulant eleccions presidencials per cada país. A més l'obtenció de les preocupacions i termes rellevants de la població de diversos països, per tal de fer un estudi i seguiment d'aquestes preocupacions entre els diferents països seleccionant diverses dates.

1.8 Breu descripció dels altres capítols de la memòria

En el capítol dos es realitza una recerca de la bibliografia existent amb relació a la predicció dels resultats electorals, on s'explicarà una breu descripció dels enfocaments d'acord amb els conjunts de dades que fan servir, la metodologia que segueixen i els resultats que obtenen, els quals s'acompanyaran d'una petita discussió, a manera de comparativa amb el treball que es pretén dur a terme en aquest document.

En el tercer capítol s'explica els materials i mètodes utilitzats en el projecte, mostrant per entrar en context i a mode introductori el diagrama d'activitat. Es descriu les parts més rellevants del codi: l'extracció dels conjunts de dades, la neteja, transformació i l'extracció de característiques amb *NLP*, el flux de dades, l'emmagatzematge, com es genera el conjunt de dades per a l'entrenament del model de predicció i l'avaluació, com es crea, optimitza i entrena el model de predicció, i com s'aplica el model de classificació de notícies. Després s'il·lustra el model de dades i la visualització en *Power BI*. Per finalitzar, es descriuen les limitacions trobades durant el desenvolupament i com s'han afrontat.

El capítol quatre mostra els resultats obtinguts del desenvolupament del treball descrit al capítol tres utilitzant la metodologia definida.

Al capítol cinc s'hi trobaran les conclusions del treball així com les reflexions sobre l'assoliment dels objectius, el compliment de la planificació i metodologia, i els impactes previstos i no previstos de les diferents dimensions de sostenibilitat, ètic-social i de diversitat. Per finalitzar, es proposen millores i noves línies d'investigació.

2. Estat de l'art

En aquest capítol es realitza una recerca existent de la bibliografia recent amb relació a la a l'objectiu principal a resoldre, la predicció dels resultats electorals, generalitzant el resultat com la posició del partit guanyador (de dretes o esquerres), sobre la base de les preocupacions del poble extretes dels titulars de les notícies del país.

Per tant, es realitza una revisió de com els investigadors han abordat problemàtiques o objectius similars, a manera de tenir una visió global dels treballs i recerques, i per tal de definir les novetats que s'aporten en aquest àmbit.

A continuació es descriuen diversos treballs que aborden la predicció electoral des de diferents perspectives:

- En el treball (López, Cazorla i Martin, 2024)[3] publicat a la "*Revista de Investigaciones Políticas y Sociológicas*" anomenat "*Medición psicofisiológica de las emociones políticas. Un análisis de sus antecedentes y propuesta metodológica*.", es va avaluar que els models tradicionals de predicció electoral estan perdent precisió a causa de la informació incorrecta derivada de variables sociodemogràfiques aportades per enquestes, i al canvi emocional resultat de l'espectacularització de la política actual.

  A causa d'aquest decaïment sorgeix la necessitat de cercar noves idees per tal de millorar la precisió dels models de predicció electoral, i es proposa un nou model que combina autoinformes i respostes fisiològiques, com ara la psicologia i la neurociència per avaluar les emocions dels votants a un estímul directe, el que condueix al reconeixement de les emocions per augmentar la precisió de la predicció electoral.

  A grans trets els tres models clàssics, anteriors als anys setanta, que es comenten a l'article i que expliquen el comportament polític, els quals definirien el vot entre republicà o demòcrata, es basen en:

  Estereotips per a la classificació dels votants (empresari o treballador, església o estat, classe social...)

  1. La transferència d'afinitats partidistes induïdes pels familiars.
  2. La racionalitat, benefici o el perjudici de les polítiques dels candidats.
  3. Per tant, cap dels models clàssics inclou directament les emocions dels votants en el comportament electoral.

  S'explica que el model més important, en el pla explicatiu emocional, és el de la intel·ligència afectiva (AIT), desenvolupat per George Marcus, Russell Neuman i Michael Mackuen a partir de la dècada de los 2000, el qual intenta explicar com les emocions afecten el

comportament polític. Teoria basada en la neurociència i la psicologia, on el comportament polític esdevé avaluacions afectives resumides principalment en dos axiomes:

1. Emocions d'entusiasme i d'aversió, implica que les persones confien en els seus hàbits i tradicions.
2. La por evita que les persones confiïn en els seus hàbits partidistes de pensament i acció, i això implica atenció sobre l'amenaça i aprenentatge.

En aquest sentint l'objectiu principal de l'article prenent passar de la teoria, descrita en el paràgraf anterior, al mesurament de les emocions mitjançant tècniques de registre autor referencials i psicofisiològiques, i així definir les possibles aplicacions en l'àmbit polític.

- **Models autoreferencials**: es tracta d'una enquesta per mesura les emocions sentides i la seva intensitat a respostes fisiològiques provocades per estímuls en el moment del mesurament.

- **Models psicofisiològics**: mesura les emocions centrant-se en la relació entre les respostes fisiològiques de l'organisme i l'estat emocional de la persona, les quals es poden dividir en tres grans grups:

    1. **El sistema nerviós autònom**: el qual es divideix en el sistema parasimpàtic (relaxació) i simpàtic (activació). Per al mesurament s'utilitzen sensors de conductància de la pell, es mesura el sistema cardiovascular, la pressió arterial, els quals s'ha demostrat que poden estar vinculats a estats emocionals, estandarditzant els resultats a emocions vinculades a estímuls concrets amb una alta precisió.

    2. **El sistema nerviós central:** compost pel cervell i la medul·la espinal, i les emocions es mesuren amb electroencefalografia i les imatges de tomografia, tot i que aquesta última tècnica es va descartar per la gran complexitat d'interpretar les imatges. Pel que fa a l'electroencefalografia es comenta que la precisió d'identificar les emocions va ser del 99% amb algoritmes de *Deep Learning* en un altre treball (Suzuki et al., 2021)[4] desenvolupat al *Instituto de Tecnología de Shibaura*.

    3. **Els reflexes**: on es mesuren amb electromiogrames la resposta elèctrica dels músculs del coll, esquena i el parpelleig. On es determina que l'expressió facial és una forma molt efectiva de precisar les emocions.

Es comenta que com ús potencial es pot aplicar la teoria de la intel·ligència afectiva ajudada pel mesurament de les emocions, combinant mesuraments autoreferencials i psicofisiologies, on s'explica que diverses investigacions afirmen les diferents activacions segons la ideologia.

Com a conclusió, s'especifica que donada la multitud d'estudis que corroboren la determinació automàtica de les emocions mitjançant models psicofisiològics, com el de

mesurament del sistema nerviós autònom, i sumant els qüestionaris autoreferencials, dona lloc a l'exploració de solucions amb millor precisió de predicció del vot, fet que combinat amb equacions estructurals o models bayesians, aportaria components explicatius relacionats amb les emocions, que es podrien usar en el màrqueting polític, electoral o consultoria política.

Comentaris:

Si bé la idea d'incorporar l'avaluació de les emocions al model de predicció electoral és rellevant i afegeix una alta precisió als resultats de les prediccions, s'observa com a limitació la metodologia de detectar les emocions. Doncs seria necessari realitzar qüestionaris i mesuraments psicofisiològics per determinar les emocions i posterior predicció del vot, fet que inclou temps i equips cars.

Per tant, com a diferència significativa amb el treball que es pretén realitzar en aquest document tenim que:

- Tot i que no es considera el mesurament de les emocions directes, s'incorporen al model de forma indirecta, mitjançant les preocupacions extretes de les notícies emeses. Ja que les emocions són reaccions a un determinat estímul i aquest pot ser a les preocupacions de la societat i com aquesta evoluciona.
- Pel que fa a la immediatesa tenim que el plantejament del model del treball d'aquest document, no requereix d'estudis de pacients, perquè tot recau en la ingesta en temps real de les notícies. Tot i que, si bé és cert que es requereix un volum substancial de dades, notícies més l'històric de resultats electorals, per l'entrenament del model, i que s'haurà d'anar ajustant amb el temps, no intervenen directament les persones sinó l'evolució del seu vot en vers a les preocupacions, i en comparació és molt més ràpid.

- En el treball (Donnin et al., 2024)[5] publicat a la "*Harvard Data Science Review"* anomenat "*Election Night Forecasting With DDHQ: A Real-Time Predictive Framework*", es va desenvolupar un model de predicció en temps reals dels resultats de les eleccions primàries i adaptable a les generals, amb una metodologia que combina informes de les votacions en directe, dades geoespacials i informació demogràfica, per estimar el candidat guanyador i la distribució de vots. El qual combinant diverses tècniques estadístiques, el model proporciona un marc robust i precís de la representació dels resultats en temps real.

  El model de Primàries en Viu, permet el recompte de vots en temps real, també ofereix una interpretació sobre l'estat de les eleccions basat en els vots ja comptabilitzats, i estima amb precisió els resultats electorals finals, i que amb la metodologia del creuament de dades geoespacials i demogràfiques, pretén augmentar la confiança en el procés electoral, donades les acusacions de frau electoral succeïdes recentment.

  La ingesta de dades ve definida per:

1. Durant la nit electoral, amb el recompte complet i parcial en directe els analistes busquen predir els resultats electorals.
2. Les condicions especials de recompte de cada comptat, diferències entre els patrons de votació i el report d'informes. On, quan el comptat ha reportat quasi el recompte complet, s'hi incorporen:
    - Dades geogràfiques: regió d'origen dels candidats on solen tenir millors resultats combinats amb la distància on resideixen, on la força estimada disminueix a mesura que augmenta la distància entre el comptat d'origen i la residència actual del candidat, i que es combinen amb enquestes sobre els candidats.
    - Dades demogràfiques: les que inclouen factors crítics com la raça, els ingressos, l'educació, el partidisme, la densitat de la població o la taxa de títols universitaris. La qual cosa permet una anàlisi més completa dels patrons de votació, per exemple els votants tendeixen a votar més candidats amb similituds demogràfiques, afirmacions que es basen en nombroses anàlisis realitzades per l'equip del model de Primàries en Viu.

    Tot plegat proporciona una predicció més confiable i precisa dels resultats finals. Encara que les dades incompletes dels recomptes de comptats s'exclouen, ja que poden donar lloc a prediccions falses.

Per tant, amb el conjunt de dades complet de com a mínim tres comptats s'activa el model demogràfic, el qual fa servir una regressió d'equacions de família binomial per a predir la proporció de vots de cada candidat. Regressió entrenada amb dades de comptats complets o quasi completes, segons es van incorporant al model. El model d'agregació (prediccions amb dades geogràfiques, i prediccions amb dades demogràfiques) va variant la ponderació segons els totals de comptats informats.

Finalment, determinada la mitja i la desviació estàndard estimada pel percentatge de vots d'un candidat en cada comptat, el model realitza 10.000 simulacions basades en simulacres de participacions, realistes, per comtat.

Comentaris:

Si bé és un model molt precís, en el qual intervenen diferents fonts de dades de diferents dimensionalitats (votacions en directe, dades geoespacials i informació demogràfica) el model és per fer servir durant les votacions, i té la limitació que la ingesta de dades ha de ser parcialment completa o quasi completa, pel que fa al recompte de vots per comptat, i percentatge de recompte de vots que varia segons cada comptat, per tal que la predicció sigui precisa.

Addicionalment, tot i que el model pretén també retornar la confiança de la població en el procés electoral explicant amb claredat el complex procés electoral que tenen. Aquest no preveu una tendència a mitjà o llarg termini, tal com es pretén en el treball que es proposa en aquest document, el qual també es pot usar com a predicció del que pot passar si no es

varia la trajectòria de les preocupacions actuals, és a dir, pot mostrar una visió general als governs i als seus analistes polítics, de com ho estan fent i cap a on van si se segueix pel mateix camí, i això és tenir informació per antelació, que per tant, pot donar l'oportunitat de rectificar la trajectòria, tenint en compte que a més proximitat de les eleccions menys probabilitat hi ha de canviar la trajectòria, tret de successos catastròfics.

Per finalitzar, la idea de classificar el poble amb dades demogràfiques, mitjançant quotes de raça, de nivell econòmic, de partidisme, d'educació, de taxes de títols ..., esdevé classista i podria ferir la sensibilitat d'alguns col·lectius. Per aquest motiu, el que es pretén en el treball d'aquest document és tractar a les persones per igual, tot i que les mateixes notícies poden inferir aquestes mateixes classificacions entre països, segons les preocupacions que anunciïn les notícies, però tal com s'ha comentat amb anterioritat cada país va al seu ritme.

- En el treball (Denicia, Ballinas, Minquiz i Medina, 2025)[6] publicat a la "*Revista Científica De Sistemas E Informática*, *5*(1), e763" anomenat "*Análisis de sentimientos en la red social X para la evaluación del posicionamiento de candidatos en elecciones políticas*", es va desenvolupar un anàlisi de sentiment de publicacions de la xarxa social de X, sobre quatre candidats a les votacions primàries per a la presidència del partit de MORENA de Mèxic en el 2024. Si bé els resultats van mostrar que els candidats millor posicionats a les votacions van ser aquells els quals indicaven major quantitat de publicacions, a la xarxa social X, amb sentiments de polaritat positiva, el guanyador final no va coincidir amb el predit. Per tant, l'estudi conclou que es requereix afegir altres variables per realitzar una predicció més precisa.

  La metodologia per a l'extracció de dades es va realitzar amb eines de mineria de dades, en concret amb *RapidMiner Studio versió 9.7*, la qual permet extracció de publicacions establint criteris de cerca mitjançant el connector de *searchTwitter*. Per al processament de text i l'anàlisi de sentiment es va utilitzar l'extensió de *MeaningCourd* de *RapidMiner*.

  Les etapes del procés d'investigació van ser: la recol·lecció de missatges de X durant el transcurs de la campanya, el preprocessament, construcció de conjunts, classificació de sentiment en 5 categories, i l'anàlisi de resultats, els que incloïen primerament la percepció de cada aspirant, i en segon lloc la classificació de sentiment, els resultats dels quals es van comparar amb els resultats reals.

  Per a la percepció dels candidats es va construir un *Word Cloud* per cada un d'ells, d'acord amb la freqüència de les paraules en les publicacions dels internautes.

  Per a la classificació de sentiment es van establir cinc conjunts en els quals es realitza una comparativa per a cada candidat:

  1. Tots els *tweets* sense duplicats.
  2. Tots els *retweets*.
  3. Únicament els *tweets* escrits i compartits des d'un dispositiu mòbil.
  4. Tots els *tweets* compartits en un episodi polític important.

5. Tots els tweets publicats en un episodi polític important des d'un dispositiu mòbil.

Finalment, es conclou que considerar únicament la polaritat de sentiment no n'és prou per predir els resultats de les eleccions, degut parcialment al biaix que hi pot haver envers la percepció dels usuaris a la publicació d'un fet polític de rellevància, ja que es pot tornar més negatiu.

Addicionalment, s'indica que durant l'anàlisi textual es van trobar moltes publicacions repetides, fet que va donà lloc a pensar que hi va haver un ús de bots tendenciosos, que van esbiaixar la polaritat sobre un tema.

Comentaris:

Tot i que els resultats no van predir amb precisió els resultats electorals, l'anàlisi obra l'oportunitat a aprofundir en l'extracció de sentiment combinant-ho amb altres variables. Per tant, si bé el procés de predicció és durant la campanya electoral fins al dia de les votacions, aquest model es podria incorporar durant la campanya de les presidencials per extraure el sentiment vers els partits d'esquerra o dreta i ampliar així la robustesa del treball que s'explica en aquest document. Tot i que aquesta incorporació de classificació de sentiments inclouria afegir els candidats de tots els països que intervenen en el treball, que ara mateix esdevenen els Estats Units d'Amèrica, els països pertanyents a la Unió Europea i el Regne Unit, per tant, això implicaria que cada període electoral s'hauria d'actualitzar els candidats, i requeriria un manteniment del model final més complex.

Addicionalment, no tots els votants tenen compte a les xarxes socials, i en aquesta situació s'estaria excloent a bona part d'ells i els seus sentiments. En conseqüència, s'hauria de fer un estudi per saber la ponderació en vers a la població que té compte de la xarxa social X respecte del total, amb la que s'hauria d'incorporar als resultats del treball d'aquest document.

- En el treball (Topîrceanu, 2025)[7] publicat a la "*Matemáticas*, *13*(4), 604" anomenat "*Macro-Scale Temporal Attenuation for Electoral Forecasting: A Retrospective Study on Recent Elections*", es va desenvolupar un model d'atenuació temporal (TA) a macroescala, que integra l'opinió a microescala i les teories de difusió d'epidèmies temporals, per tal de millorar la precisió dels pronòstics electorals fent ús d'enquestes preelectorals. On es va descobrir que el moment de les enquestes influeix en la fluctuació de les opinions, a mesura que s'aproximen les dates de les eleccions. El model es va provar amb diferents variants, fent ús d'un conjunt de dades, enquestes públiques confiables, de 10 eleccions celebrades entre el 2020 i el 2024 de tot el món, i els resultats obtinguts van concloure que el model d'atenuació temporal va superar significativament els pronòstics d'altres models estadístics, el que suggereix que a mesura que les dades d'enquestes electorals mundials són més accessibles, l'error del pronòstic disminueix.

  La ingesta de dades esdevé principalment les enquestes realitzades en diferents dies durant el període electoral, on cada enquesta dona els percentatges de suport a cada candidat en

un sistema de múltiples candidats, més opinions diàries que emplenaran els dies que no s'hagin realitzat enquestes, és a dir amb vectors de sondejos d'opinions continus (diaris), d'aquesta forma es podran visualitzar les fluctuacions i períodes d'estabilitat.

De la ingesta principal a microescala, enquestes i opinions diàries, se cerca el comportament a macroescala, fent ús de les teories de difusió epidèmiques adaptades, per tal d'observar l'evolució de la dinàmica de l'opinió després de cada injecció d'opinió, en els individus en el context d'opinió electoral.

El model es prova i s'avalua amb diferents mètriques, i fa rellevància en el fet que la precisió del pronòstic electoral és superior en 6 de cada 10 conjunts de dades electorals. Addicionalment, és demostrar que els resultats suggereixen que la consciència temporal juga un paper molt important en la predicció electoral del que es reconeixia anteriorment. En aquest sentit, es comenta que la capacitat per modelar la dinàmica temporal i l'evolució de l'opinió es pot estendre a la predicció de resultats en altres dominis.

Comentaris:

Si bé és un model amb resultats notables de pronòstics electorals, que no fa ús de dades demogràfiques ni econòmiques, ni de cap context polític, sí que requereix enquestes públiques confiables, fet que depèn de la publicació i, per tant, de l'accessibilitat que li vulgui donar cada centre d'estadístiques dels diferents països. Tanmateix, la combinació d'enquestes confiables en el temps del procés electoral combinat amb les teories de difusió d'epidèmies, resulta un enfocament innovador que com s'ha demostrat obté resultats significativament superiors a altres models.

Si es compara amb l'enfocament del treball d'aquest document, trobem similituds en la demostració de la tendència evolutiva de la predicció de candidats, en el cas que ens ocupa, de dretes o esquerres. Encara que el treball descrit se centra en el període preelectoral i que la metodologia és diferent, aquest s'assimila a l'evolució de les preocupacions que es volen extreure de les notícies en un període determinat. Per tant, si s'observa amb perspectiva, poden estar interrelacionades amb les opinions de les enquestes dels diferents candidats els quals proposen certes solucions, a problemes o preocupacions actuals, en els seus programes electorals.

- Finalment, s'estudia el llibre (Francisco i Fernández, 2023)[8] publicat per "*Universidad de Alicante. Obets Ciencia Abierta. Alicante: Limencop*." i anomenat "*Métodos y Modelos para la Predicción Electoral: Una Guía Práctica*", on es proposa enfocaments teòrics i pràctics amb una visió general i actualitzada dels principals models i metodologies utilitzats, en la ciència social computacional, per a la predicció electoral.

  S'hi introdueixen els conceptes bàsics amb relació a les teories i visions de l'estudi de la democràcia, incloent-hi la democràcia liberal, participativa i la deliberativa. Així com els diferents tipus de sistemes electorals: majoritaris, proporcionals i mixtes. També es descriu els diferents tipus de partits polítics, com ara els de masses, de quadres i atracció.

Així mateix, es detalla generalment el comportament electoral i els factors determinants que poden influir en el vot, com ara les identitats socials, les avaluacions econòmiques o les actituds polítiques.

Pel que fa a les fonts de dades utilitzades per a predir els resultats electorals, es fa menció de les enquestes, models estadístics, mercats d'apostes, mitjans de comunicació i anàlisis de sentiment, i xarxes socials, on s'expliquen els avantatges, complicacions de recol·lecció, i limitacions en vers a biaixos, errors de mostreig, influència de factors externs, dades falses, respostes ambigües..., tot i que per abordar-les poden utilitzar-se tècniques d'ajust, ponderació, imputació, validació creuada, combinació de diferents fonts, entre d'altres.

També s'explica la rellevància de la preparació de les dades en la implementació de models predictius, com la neteja, la transformació, selecció atributs rellevants, la normalització o estandarditzación de variables..., on la representativitat i la qualitat tenen un impacte directe en el rendiment i l'avaluació dels models. Així com diverses eines i diferents llenguatges de programació amb les que du a terme la preparació de les dades, com ara *Python, R, Excel, SQL, KNIME, Talend, MATLAB, Orange...*

Es fan aclariments en relació amb els models estadístics i economètrics, on aquests últims esdevenen extensions dels primers en l'anàlisi de dades econòmiques i polítiques, que tracten problemes específics com per exemple l'autocorrelació. Els mètodes i tècniques habituals per estimar-los i avaluar-los solen ser: regressions lineal, logística i anàlisis de sèries temporals.

En la secció dels mètodes d'aprenentatge automàtic es descriuen els models d'aprenentatge automàtic supervisat, no supervisat i de reforç, amb alguns exemples d'algoritmes com la regressió lineal o les màquines de suport vectorial (SVM) o xarxes de neurones artificials (ANN) per als mètodes supervisats, k-means per als no supervisats, i l'algoritme *Q-learning* i l'aprenentatge per reforç profund per a l'aprenentatge per reforç.

S'esmenta un punt important respecte a la problemàtica de la privacitat i seguretat de la informació personal, en el que s'ha de garantir el compliment de les regulacions de privacitat de dades, com el Reglament General de Protecció de Dades (GDPR).

Amb relació al pas del temps, es comenta la bona praxi del manteniment i monitoratge dels models d'aprenentatge automàtic, a manera de garantir la precisió i l'eficiència d'aquests. El que implica l'avaluació del model en temps real a mesura que s'hi ingesten dades noves. Així mateix, esmenta la biblioteca "MLflow" com a opció popular per a la gestió del cicle de vida dels models d'aprenentatge automàtic.

Respecte a la comunicació de les prediccions electorals, es puntualitza que ha de ser efectiva, transparent, comprensible, en la que s'ha d'informar la incertesa i el marge d'error, per tal que els interessats puguin interpretar els resultats amb una prospectiva adequada i comprendre la incertesa de les estimacions a manera de prendre decisions ben informades. Pel que fa a les visualitzacions dels pronòstics es poden fer servir mapes de densitat o mapes coropletes[10], on els colors graduats mostrarien l'interval de confiança dels resultats i permeten visualitzar clarament on els pronòstics són més precisos.

Comentaris:

Pel que fa als factors determinants que es descriuen en vers al condicionament del vot (identitats socials, avaluacions econòmiques o actituds polítiques), es pot pensar que agrupen les preocupacions que es pretenen extreure de les notícies, en el treball d'aquest document, tret la de les identitats socials, ja que és més classista i assimilaria el vot als candidats que representen la mateixa classe social, per exemple.

A causa dels canvis i tendències en el comportament electoral, es referència la capacitat dels polítics electes que han de tenir per a adaptar-se a les demandes recents i canviants dels ciutadans. Demandes que poden quedar reflectides en les notícies emeses pels mitjans de comunicació, tal com es pretén esbrinar en el treball d'aquest document; si els resultats de predicció del model resulten concloents vers les preocupacions que s'extrauran de les notícies.

S'ha trobat especialment rellevant la descripció de l'enfocament psicològic en el comportament electoral, on s'explica que els votants poden desenvolupar lleialtats partidistes, basades en les seves experiències polítiques i socials. En aquest sentit, aquest tipus de votants probablement no tendeixen a canviar el seu vot segons les preocupacions extretes de les notícies. Per tant, això implica que la predicció dels resultats electorals, en el treball d'aquest document, pot veure's afectat en el cas que hi hagi un ampli percentatge de la població que sigui lleial, i que, en conseqüència, no els afectin les preocupacions actuals o bé en facin cas omís.

Amb relació als mitjans de comunicació es comenta la seva influència en vers a les actituds i participació dels votants, així mateix, es reflexiona sobre si les preocupacions que es volen extraure de les notícies són genuïnes i independents?, o venen definides per influir? Si bé és cert que és un aspecte a tenir en compte, les fonts de dades que es pretenen fer servir, extrauen les notícies de diversos mitjans de comunicació, en la majoria dels països i, per tant, es consideraran independent i no manipulades per cap posició política. En tot cas es pot preveure en una segona versió del projecte verificar la procedència equitativa de les notícies dels diversos mitjans de comunicació. El que implicaria emmagatzemar també els mitjans de comunicació de les notícies a manera d'avaluar la varietat de procedència i evitar biaixos. En aquest sentit, tenim les notícies dels últims cinquanta anys dels Estats Units d'Amèrica, les quals procedeixen únicamentdel diari *New York Times*, així que es podria considerar un biaix, tot i que tenim entre 3.000 i 8.000 notícies mensuals, però en pròximes versions s'hauria d'incorporar més varietat de mitjans de comunicació pel que fa als Estats Units d'Amèrica.

Pel que fa als mètodes d'aprenentatge automàtic que es descriuen, destaquem el supervisat per als casos on es tinguin prou dades per a l'entrenament, com per exemple per als Estats Units, on es tenen notícies de bastants anys enrere amb els seus presidents i amb la seva posició política. En el cas on no tinguem prou dades de notícies per realitzar un entrenament de predicció, com malauradament en els casos dels països de la Unió Europea, s'utilitzaran models d'aprenentatge no supervisat per classificar-los segons la similitud de preocupacions

en períodes concrets amb les dades extretes de les notícies dels Estats Units, com per exemple els models d'agrupació o classificació de textos.

A propòsit del compliment de la protecció personal de les dades, en vers al reglament GDPR establert per a tots els ciutadans de la Unió Europea[9], tot i que a priori es recullen dades de diferents mitjans de comunicació les quals poden incloure dades personals, durant el processament d'aquestes es pretenen anonimitzar, eliminant noms propis, tret dels presidents de cada país, que en tot cas si fos necessari també s'eliminarien.

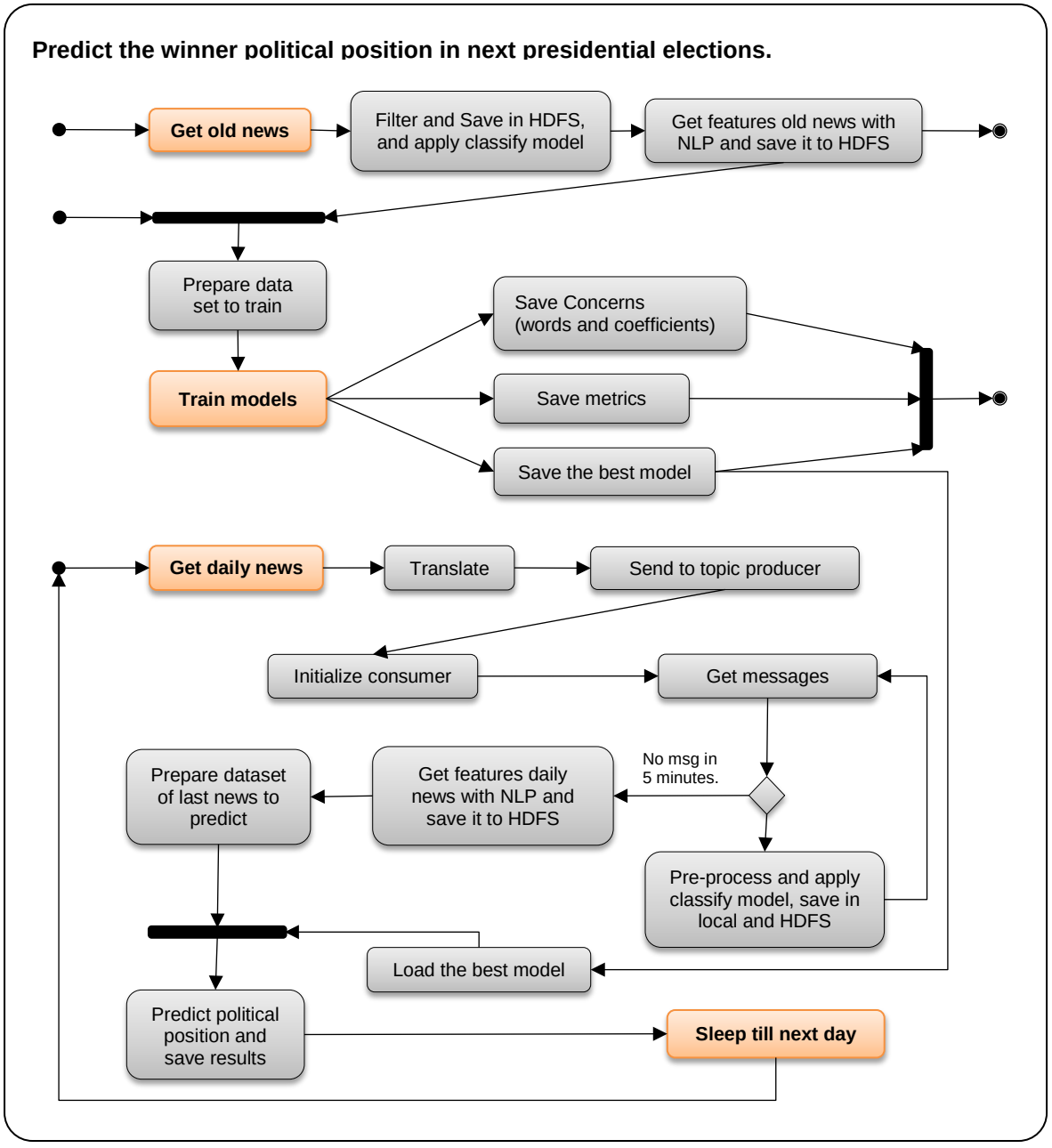
Tot i que es comenta que l'anàlisi de sentiments es considera un valor afegit com a eina de predicció en l'aprenentatge automàtic, per comprendre les opinions i emocions dels usuaris, productes o serveis, que en especial va quedar demostrat en l'estudi de Tumasjan et al. (2010) (citat per Francisco i Fernández, 2023), la correlació entre: el volum de *tweets* i la polaritat de sentiments, amb els resultats de les eleccions. Així mateix, no es creu que sigui efectiu aplicar-lo al text de les notícies, ja que no són escrites pel poble en general, sinó per professionals de la comunicació, que 'suposadament' exposen els successos sense polaritat. En tot cas, tal com s'ha comentat amb anterioritat, si finalment els resultats del treball d'aquest document no són rellevants, es consideraria afegir aquest nou estudi com a complement de predicció, això sí, enfocat amb l'extracció de polaritat de sentiments únicament del govern actual, amb relació als comentaris i opinions de la xarxa social X, per tal de combinar-ho com a *assembly models* al de predicció d'aquest treball, així es milloraria el rendiment predictiu i el faria més robust.

3. Materials i mètodes

En aquest capítol, es descriuran els aspectes més rellevants del disseny i desenvolupament del treball, les decisions preses, i els criteris utilitzats per donar resposta als objectius principals.

Per a la descripció general del producte obtingut a continuació es mostra el diagrama d'activitat, facilitant així la comprensió del flux de control entre les activitats involucrades a alt nivell de l'execució del sistema.

Figura 3: Diagrama d'activitats



Font: Elaboració pròpia

El codi, l'executable de la visualització, exemples de *logs* i resultats de fitxers com a exemple generats per l'última execució a dia 5 de juny es pot trobar a l'URL de *GitHub*: https://github.com/sangoumr/TFM

Un cop descarregat el repositori, la documentació del codi es pot trobar a l'accés directe "Documentation - index.html", la qual s'ha generat amb l'aplicació DoxyGen.

La visualització està publicada amb data del 5 de juny del 2025 i com es comentarà a la secció de limitacions, no es pot actualitzar automàticament, però si es pot explorar i interactuar fins aquesta data:
https://app.powerbi.com/view?r=eyJrIjoiY2MxN2FhYWItMTZhNS00NjgxLThjOWMtYTM3YzQ4ZWU3MWQzIiwidCI6ImMyM2M0ZThiLTRlMGMtNDY3MC1iMmFlLTZhZTA2MThkZDBjNyIsImMiOjl9

3.1 Configuració de l'entorn

La configuració de l'entorn per al desenvolupament es realitza en local, i el manual d'instal·lació s'inclou en l'Annex 1.

Un cop l'entorn està configurat i iniciat, només cal engegar els serveis de *Kafka* i *HDFS,* tal com s'indica al manual.

Tot seguit s'ha d'obtenir les *key* de les diferents *API*, donar-se d'alta a cada una, i intercanviar-les en el fitxer de constants.py.

```
## Key for downloading current news from the United States.
# https://newsapi.org
NEWS_API_KEY =  "********************"
## Key for downloading current news for the member countries of the European Union,
# https://newsdata.io
NEWS_DATA_IO_KEY = "********************"
## Key for downloading old US news from the New York Times (downloaded since the
1970s),
# https://api.nytimes.com
NEWS_NY_TIME_KEY = "********************"
```

Ara ja es pot iniciar la descàrrega de les dades antigues executant el fitxer:

- get_old_news_us.py

Un cop finalitzat el procés de descàrrega ja es pot iniciar l'entrenament del model, executant el fitxer:

- train_model.py

Finalment, un cop tenim el model entrenat, podem iniciar la descàrrega diària de les notícies, on es tracten i s'injecten al tòpic del productor, es consumeixen, processen,

classifiquen i es fa la predicció, procés que no finalitza i que automàticament cada 24h es repeteix tal com es mostra a la figura 3 del diagrama d'activitats, executant el fitxer:

- main.py

En aquest punt la visualització en *Power BI* ja està llesta per capturar les dades diàriament i mostrar-les. Prèvia actualització de la ruta dels conjunts de dades des de *Power Query* accedint a l'editor avançat, per exemple per al fitxer concersn_summary.csv:

```
Origen =
Csv.Document(File.Contents("\\wsl.localhost\Ubuntu\home\roser\TFM\data\cleaned\concerns_summary.csv")
```

- Geopolitical Trend in Real Time.pbix

3.2 Els conjunts de dades

Els conjunts de dades principals, dels quals extraurem les característiques provenen de diferents interfícies de programació d'aplicacions (*API*), de les quals es descarregaran diàriament les notícies dels països per a ser processades, a més de l'obtenció del conjunt de dades inicial, amb milers de notícies per a ser utilitzat com a conjunt d'entrenament en els models d'aprenentatge automàtic supervisat. Supervisat perquè se li inclourà la classe, és a dir, amb base a la data de publicació de la notícia s'hi incorporarà la posició política del president del país del qual procedeix la notícia en aquella data.

3.2.1 Conjunt de dades inicial

Prové del "*The New York Times Developer Network*" de la secció d'arxiu, de la qual s'han descarregat mensualment les notícies publicades des del 1970 fina ara.

Figura 4: API NYT

- The New York Times

```
https://api.nytimes.com/svc/archive/v1/2024/4.json?api-key=yourkey
```

Font: Elaboració pròpia

Tot i haver filtrat els atributs que retorna la consulta, únicament s'ha seleccionat el títol, la descripció i la data de publicació, genera un *dataset* de 1,5 Gb, on per cada mes s'han obtingut entre 3000 i 8000 notícies.

Encara que les notícies venen únicament d'un únic mitjà de comunicació, indici d'un possible biaix, en cas d'estar polititzat, en processos posteriors es tractarà de minimitzar-lo eliminant persones, organitzacions, localitzacions, mitjançant el model entrenat "*English Berttest BertForTokenClassification from RtwC*".

Aquest conjunt inicial, juntament amb els presidents dels Estats Units d'Amèrica, composts pel seu nom, nomenament i la seva posició política, serviran per a entrenar el model supervisat de regressió logística, fent que la variable objectiu o depenent sigui la posició política del guanyador o guanyadora de les eleccions presidencials, on la posició política se simplificarà a binomial (esquerra o dreta).

Per tant, aquest conjunt de dades i el model, donaran resposta a l'objectiu principal, predicció de la posició política i extracció de preocupacions, on es determinarà:

- L'existència o l'absència de relació entre les característiques de les notícies (variables independents) i la posició política de les eleccions (variable depenent).
- I el grau de magnitud d'aquesta relació, la que s'avaluarà amb diferents mesures.

3.2.2 Llistat de presidents i posició política (l'etiqueta a predir)

Amb relació a les etiquetes (la variable a predir), s'han descarregat manualment des de diferents webs oficials i no oficials, els presidents, primers ministres i caps de governs dels diferents països, amb la data de nomenament i la posició del partit en el qual pertanyien quan van ser escollits.

- **Presidents dels EUA**: https://www.whitehousehistory.org/the-presidents-timeline
- **Presidents d'Espanya**: https://www.lamoncloa.gob.es/presidente/presidentes/Paginas/index.aspx
- **Presidents de França**: https://es.wikipedia.org/wiki/Presidente_de_Francia

Aquest últim conjunt de dades ha estat un repte obtenir-lo, ja que cada país té el seu propi sistema electoral, en el que el poble no sempre vota directament el president, primer ministre o cap de govern, sinó als ministres i aquests elegeixen el president, primer ministre o cap de govern.

En aquestes llistes de presidents s'han exclòs els presidents interins i els mandataris en període de dictadura, ja que el poble no els ha votat. Per tant, s'ha decidit que aquests períodes buits entren directament dins de l'últim president electe, encara que hagi mort, dimitit o hi hagi hagut un cop d'estat.

Per facilitar l'actualització manual, la llista de presidents dels diferents països s'emmagatzemen en diferents fitxers amb la mateixa estructura CSV, per exemple per a Espanya queda:

Fitxer: presidents-es.csv

```
Nom,Nomenament,Partit
```

```
Adolfo Suárez González,03-07-1976,dreta
Leopoldo Calvo-Sotelo Bustelo,26-02-1981,dreta
Felipe González Márquez,02-12-1982,esquerra
Felipe González Márquez,24-07-1986,esquerra
Felipe González Márquez,06-12-1989,esquerra
Felipe González Márquez,10-07-1993,esquerra
José María Aznar López,05-05-1996,dreta
José María Aznar López,27-04-2000,dreta
José Luis Rodríguez Zapatero,17-04-2004,esquerra
José Luis Rodríguez Zapatero,12-04-2008,esquerra
Mariano Rajoy Brey,21-12-2011,dreta
Mariano Rajoy Brey,31-10-2016,dreta
Pedro Sánchez Pérez-Castejón,02-06-2018,esquerra
Pedro Sánchez Pérez-Castejón,08-01-2020,esquerra
Pedro Sánchez Pérez-Castejón,17-11-2023,esquerra
```

Tot i que per a ser utilitzats en el procés de generació de les etiquetes predictives i els períodes electorals, en generar el conjunt de dades d'entrenament del model, i per a fer la predicció de les notícies actuals, els fitxers dels diferents països es combinen en un diàriament, just abans de generar els *datasets*, així si hi ha una actualització es té en compte. El fitxer combinat queda com:

Fitxer: presidents_combined_final.csv

```
iso_code,nom,nomenament,posicio_politica,label
at,Bruno Kreisky,21-04-1970,esquerra,0
at,Fred Sinowatz,24-05-1983,esquerra,0
de,Angela Merkel,17-12-2013,dreta,1
de,Angela Merkel,14-10-2018,dreta,1
de,Olaf Scholz,08-12-2021,esquerra,0
...
dk,Hilmar Baunsgaard,02-02-1968,esquerra,0
dk,Jens Otto Krag,11-10-1971,esquerra,0
dk,Anker Jørgensen,05-10-1972,esquerra,0
es,Pedro Sánchez Pérez-Castejón,08-01-2020,esquerra,0
es,Pedro Sánchez Pérez-Castejón,17-11-2023,esquerra,0
fi,Urho Kekkonen,16-03-1968,dreta,1
fi,Urho Kekkonen,16-03-1978,dreta,1
...
```

Amb relació a la posició política, si bé amb els presidents dels Estats Units s'ha trobat que tots són republicans o demòcrates, amb els dels països de la Unió Europea i el Regne Unit, s'han trobat posicions de centre i independents. En aquests últims casos s'han unificat, per tant, l'etiqueta resultant de la combinació dels fitxers serà [centre,2], tot i que el model únicament prediu 0 o 1 (esquerra i dreta respectivament). Pel que fa a les posicions extremes s'han integrat a esquerra o dreta, segons l'extrem.

3.2.3 Conjunt de dades diari i ingesta *Kafka*

Donat que l'ús de la versió gratuïta de les *API* per a la descàrrega diària de les últimes notícies dels diferents països té limitacions, es reparteix la càrrega entre dues *API*, una per a descarregar les últimes notícies dels Estats Units, i l'altra per a descarregar les últimes notícies dels diferents països de la Unió Europea i el Regne Unit.

- Les últimes notícies dels Estats Units, provenen de diferents mitjans de comunicacions, com ara:

Figura 5: News API

Top headlines

Mitjans de comunicació EEUU:

- BBC News
- SciTechDaily
- DW (English)
- Mavs Moneyball
- Associated Press
- ...

```
https://newsapi.org/v2/top-headlines?country=us&apiKey=yourkey&pageSize=10
```

Font: Elaboració pròpia

- Les últimes notícies dels països de la Unió Europea, provenen també de diferents mitjans de comunicacions per a cada país, per exemple per a França:

Figura 6: NewsDATA.IO API

Mitjans de comunicació per França :

- 20 minutes
- Adents
- Crash
- Expatica
- French Morning
- Le Progrès
- …

```
https://newsdata.io/api/1/latest?apikey=yourkey&country=fr&size=10
```

Font: Elaboració pròpia

Encara que, les descàrregues diàries de les últimes notícies dels diferents països provenen de diferents mitjans de comunicació, igualment s'aplicarà el model de reconeixement d'entitats nombrades i s'eliminaran, ja que el model entrenat no les ha considerat en haver-les eliminat en el conjunt d'entrenament i, per tant, podria tenir un impacte negatiu en la predicció. A més les entitats d'un país poden no ser d'altres, i considerant que les notícies descarregades de països no angloparlants es tradueixen a l'anglès, és possible que el model *Bert* no les reconegui del tot, en tot cas podrien considerar-se soroll per al model i per aquest motiu s'eliminen.

El sistema s'inicia posant en marxa el productor de Kafka, tot seguit llegeix el fitxer que incorpora els països dels quals es vol extreure notícies, i utilitza el codi ISO del país i el codi ISO de la llengua que s'hi parla, per descarregar les notícies de les *API*. A continuació es mostra un exemple del fitxer de països.

Fitxer: paisos-ue.csv

```
"Nom","Name","Codi ISO","Codi ISO Llengua"
```

```
"Alemanya","Germany","DE","de"
"Àustria","Austria","AT","de"
"Bèlgica","Belgium","BE","nl,fr,de"
...
```

D'aquesta forma si en qualsevol moment es vol incorporar un país, només cal afegir-lo en aquest fitxer.

Així que, per cada línia d'aquest fitxer amb el codi de país i llenguatge, el sistema fa un *request* a les *API* amb límit de notícies configurable, amb versió gratuïta unes 10, on s'estableix l'idioma amb el qual es volen rebre les notícies de cada país, ja que altrament pot ser diferent i complica la traducció.

Un cop descarregades les notícies es tradueixen una a una amb *Google Translate*, limitant prèviament el total de caràcters a 2500, perquè a vegades si la descripció de la notícia l'excedia retornava error per *time-out* o excés, establint també l'idioma d'origen i el de destí del text de la notícia. Encara que *Google Translate* dona l'opció de reconèixer automàticament l'idioma origen, aquesta configuració ha donat problemes, ja que algun cop el castellà l'ha interpretat com a portuguès en frases curtes. Per aquest motiu tant en els *requests* de les *API* com en la traducció, s'estableix l'idioma origen en el qual es vol descarregar la notícia, i per a la traducció l'idioma origen en el que està la notícia.

```
https://clients5.google.com/translate_a/t?client=dict-chrome-
ex&sl=es&tl=en&q=texte
```

Així que, per cada bloc de notícies rebudes de cada país i traduïdes s'envien al tòpic definit del productor establert. A continuació es mostra un exemple dels missatges enviats de notícies de la República Txeca, amb negreta s'observa la traducció:

```
## Data send to Kafka successfully:
[['cz','Svatba, pohřeb nebo doktor. Nově dostanou zaměstnanci volno i zpáteční
cestu z nemocnice', 'Změna po devatenácti letech. Novela nařízení o překážkách v
práci má zaměstnancům konečně ulevit. Zpřehlední situace kolem volna při svatbě,
pohřbu i doprovodu rodiny k lékaři. Zohlední stejnopohlavní páry a do pravidel
přidá i návrat z nemocnice. Pracovní volno se tak přiblíží realitě běžného
života.', 'en', 'Wedding, funeral or doctor. New employees will get off and return
trip from the hospital', 'Change after nineteen years. The amendment to the
obstacles to work at work is finally supposed to relieve employees. It will make
the situation around free at the wedding, funeral and family accompaniment to the
doctor. It takes into account the same -sex pairs and adds a return from the
hospital to the rules. Working leave will thus approach the reality of everyday
life.', '2025-04-20 19:00:00'],...]
```

A tall de control, en finalitzar la descàrrega de notícies, traducció i enviament dels missatges al tòpic, es mostra un resum on es pot visualitzar el total de notícies tractades per país, i observa que els països angloparlants no han estat traduïts, i que en aquesta ocasió les consultes a les *API* han retornat les notícies demanades, excepte en el cas dels Estats Units, 18 en lloc de 20, configuració establerta al fitxer de constants.py, tal com es mostra a continuació:

```
## Number of news to download every day for US
NEWS_US = "20"
## Number of news to download every day for UE countries
NEWS_UE_COUNTRIES = "10"
```

```
+--------+-----------------------+------------------------+
| country|    total_news_download|   total_news_translated|
+--------+-----------------------+------------------------+
|      at|                     10|                      10 |
|      be|                     10|                      10 |
|      bg|                     10|                      10 |
|      cy|                     10|                      10 |
|      cz|                     10|                      10 |
|      de|                     10|                      10 |
|      dk|                     10|                      10 |
|      ee|                     10|                      10 |
|      es|                     10|                      10 |
|      fi|                     10|                      10 |
|      fr|                     10|                      10 |
|      gb|                     10|                       0 |
|      gr|                     10|                      10 |
|      hr|                     10|                      10 |
|      hu|                     10|                      10 |
|      ie|                     10|                      10 |
|      it|                     10|                      10 |
|      lt|                     10|                      10 |
|      lu|                     10|                      10 |
|      lv|                     10|                      10 |
|      mt|                     10|                      10 |
|      nl|                     10|                      10 |
|      pl|                     10|                      10 |
|      pt|                     10|                      10 |
|      ro|                     10|                      10 |
|      se|                     10|                      10 |
|      si|                     10|                      10 |
|      sk|                     10|                      10 |
|      us|                     18|                       0 |
+--------+-----------------------+------------------------+
```

Un cop descarregades, traduïdes i enviades totes les notícies dels països definits, s'inicia el consumidor de Kafka[11] subscrit al tòpic definit al productor, que les va recollint en bloc, les filtra per país i les va guardant en format parquet, per ubicació de país, al sistema de fitxers distribuïts de *Hadoop* amb el següent esquema:

```
# Define schema for the incoming data.
 schema_news = StructType([
     StructField("iso_code", StringType(), True),
     StructField("title", StringType(), True),
     StructField("description", StringType(), True),
     StructField("lang_tranlation", StringType(), True),
     StructField("title_translated", StringType(), False),
     StructField("description_translated", StringType(), True),
     StructField("pubDate", TimestampType(), False)
 ])
```

A continuació es mostra un bloc rebut del tòpic que s'emmagatzema a *HDFS*:

```
+--------+--------------------+-----------------+----------+------------------+-------------------+-------------------+
|iso_code|               title|      description|lang_trnsl|  title_translated|description_transla|            pubDate|
+--------+--------------------+-----------------+----------+------------------+-------------------+-------------------+
|      ee|VIDEO ) Padel  k...|4. Aprilliltoi...|       en|Video) Padel is r...|  On April 4, a...|2025-04-20 14:00:00|
|      ee|GALERII ) Täna  ...|Lossis möödus ...|       en|Gallery) Today wa...|  In the castle...|2025-04-20 13:30:00|
|      ee|ÕL TV "MERILY  ...|Uus nädal, uued...|       en|OIL TV "Merily Ti..|  New week, new ...|2025-04-20 13:05:00|
|      ee|FC Kuressaare tu...|FC Kuressaare t...|       en|FC Kuressaare ack..|  FC Kuressaare ...|2025-04-20 13:01:00|
|      ee|Saaremaa naiskon...|Lõppenud nädal... |       en|The Saaremaa wome..|  The Volleyball...|2025-04-20 12:31:00|
|      ee|AMETLIK: Tammeka...|Tartu Tammeka j...|       en|Official: Tammeka..|  Tartu Tammeka ...|2025-04-20 12:29:00|
|      ee|FOTOD | Prints A...|Neil lihavõtete...|       en|Photos Prince And..|  Prince Andrew ...|2025-04-20 12:25:39|
|      ee|VIDEO | Evelin V...|Evelin Võigemas...|       en|Video | Evelin Võ..|  Evelin Võigema...|2025-04-20 12:00:00|
|      ee|Leisi saab juurd...|Leisi aleviku k...|       en|Leisi can get 1.7..|  The continuati...|2025-04-20 11:56:00|
```

```
|      ee|TERVIS ) "Ma ei ...|Ingridi* teekon...|      en|Health) "I don't ...|  Ingrid* The j...|2025-04-20 11:09:00|
+--------+--------------------+-----------------+--------+--------------------+------------------+--------------------+
## Append to ee conuntry parquet: 10 news.
```

A més, per cada bloc rebut s'extreu la classe de cada notícia i s'emmagatzema localment, com a resum de preocupacions diàries per a mostrar-ho a la visualització. Per obtenir la classe es crea una canalització amb les etapes:

1. **DocmentAssembler**: Prepara les dades en format processable per *Spark NLP*.
2. **Tokenizer**: Retorna el document en *tokens*.
3. ***DistilBertForSequenceClassification***: Finalment s'aplica el model entrenat *DistilBERT Sequence Classification* [12], que prèviament durant la preparació de l'entorn s'ha descarregat i ara es carrega, ja que carregar-lo cada cop que s'usa requereix més recursos. Amb aquest model s'identifica de quin tipus és cada notícia (*Business, Sci/Tech, Sports, World*). El model té una avaluació molt bona, doncs té un F1-Score de 0,90 per cada classe. A continuació es mostra un exemple de la classificació, i que com a resum de preocupacions generals diàries de cada país es guardarà en el fitxer que es mostra tot seguit.

```
+--------+----------+-------------------------------------------------------------+----------+
|iso_code|pubDate   |text_news                                                    |result    |
+--------+----------+-------------------------------------------------------------+----------+
|us      |2025-03-17|Rory McIlroy takes drama out of playoff to win THE PLAYERS...|[Sports]  |
|us      |2025-03-17|Fast-fashion staple Forever 21 files for bankruptcy again... |[Business]|
|us      |2025-03-17|Rory McIlroy turns frustration into triumph to conquer TH... |[Sports]  |
|us      |2025-03-17|Serbia protests: How much trouble is Aleksandar Vui in? -... |[World]   |
+--------+----------+-------------------------------------------------------------+----------+
```

Fitxer: concerns_summary.csv

```
iso_code,pubDate,summary_concern
at,2025-03-18,World
at,2025-03-18,World
at,2025-03-18,Business
us,2025-04-22,Sports
us,2025-04-22,Business
lt,2025-04-23,Sci/Tech
```

Passat un temps, si no hi ha més missatges el consumidor de *Kafka* finalitza, i s'inicia el procés de neteja i transformació amb *NLP*, el qual s'explica en l'apartat següent.

```
   last_msg_time = time.time()
   while True:
        messages = consumer.poll(timeout_ms=10000)
        current_time = time.time()
        elapsed_time = current_time - last_msg_time

        if messages is None:
            continue
        if not messages:
            print("No messages received.")
            if elapsed_time >= MAX_TIME_WAITING:
                 print("No messages received for 5 minutes. Exiting...")
               consumer.close()
                 print("Closed consumer. Final data written to HDFS.")
               break
            continue
                 ……
...
No messages received.
No messages received for 5 minutes. Exiting...
Closed consumer. Final data written to HDFS.
## Start NLP process at: 2025-04-21 11:10:46.368733
```

3.3 Neteja de les dades (*Spark NLP Pipeline*)

Per a la neteja preliminar de les dades es defineix una canalització amb una seqüència d'etapes per al processament de paraules[13], les quals són:

1. **DocmentAssembler**: Prepara les dades en format processable per *Spark NLP*.
2. **Tokenizer**: Retorna el document en *tokens*.
3. **LemmatizerModel**: De cada *token* retorna la paraula base. Es fa servir un model ja preentrenat en anglès, anomenat *"lemma_antbnc"*
4. **Normalizer**: Normalitza els *tokens* per eliminar tots els caràcters bruts del text seguint un patró d'expressió regular, eliminar caràcters especials, puntuació i les paraules inferiors a 3 caràcters.
5. **StopWordsCleaner**: Elimina les paraules buides aplicant el model entrenat en anglès, anomenat *"stopwords_en"*.
6. ***BertForTokenClassification***: Finalment s'aplica el model entrenat *English berttest BertForTokenClassification from RtwC*[14], el qual durant la preparació de l'entorn s'ha descarregat i ara es carrega, ja que carregar-lo cada cop que s'usa requereix més recursos. Aquest model té una avaluació de la mesura F1-Score del 0,94, per tant, el reconeixement d'entitats nombrades (NER) és molt bo, tot i que per a les traduccions de les notícies d'altres idiomes és probable que no les identifiqui amb aquesta bonança. En tot cas, s'inicia la identificació de les entitats anomenades. (Tot i que, sovint s'hauran d'actualitzar amb versions recents per identificar millor les entitats.)

Un cop instanciat el *Pipeline* i les seves etapes s'entrena el model i es transforma el conjunt de dades:

```
# Instance Pipeline setting stages.
pipeline_norm = Pipeline(stages=[document_assembler,
                                 tokenizer,
                                 lemmatizer,
                                 normalizer,
                                 stop_words,
                                 token_classifier])
model = pipeline_norm.fit(df_news_f)
result = model.transform(df_news_f)
```

A continuació es mostra un exemple del resultat de la canalització per algunes etapes d'una notícia, on es podrà observar les transformacions:

Taula 3: Pipeline Processament de Llenguatge Natural

STAGE	Transformation
News_Translated	Walls, pits and cobbled streets that seem detained over time. This is one of Girona's most beautiful medieval peoples. More information: this is the Catalan people who are in danger of extinction: the smallest in Catalonia and very few inhabitants
Token	[Walls, ,, pits, and, cobbled, streets, that, seem, detained, over, time, ., This, is, one, of, Girona's, most, beautiful, medieval, peoples, ., More, information, :, this, is, the, Catalan, people, who, are, in, danger, of, extinction, :, the, smallest, in, Catalonia, and, very, few, inhabitants]
Lemma	[Walls, ,, pit, and, cobble, street, that, seem, detain, over, time, ., This, be, one, of, Girona's, most, beautiful, medieval, people, ., More, information, :, this, be, the, Catalan, people, who, be, in, danger, of, extinction, :, the, small, in, Catalonia, and, very, few, inhabitant]
Normalized	[Walls, pit, and, cobble, street, that, seem, detain, over, time, This, one, Gironas, most, beautiful, medieval, people, More, information, this, the, Catalan, people, who, danger, extinction, the, small, Catalonia, and, very, few, inhabitant]
Stop_Words	[Walls, pit, cobble, street, detain, time, Gironas, beautiful, medieval, people, information, Catalan, people, danger, extinction, small, Catalonia, inhabitant]
NER	[O, O, O, O, O, O, B-LOC, O, O, O, O, B-MISC, O, O, O, O, B-LOC, O]

Font: Elaboració pròpia

Un Cop finalitzada la canalització se seleccionen i combinen els *tokens* resultants de l'etapa de *Stop Words* amb la classificació *NER*, i s'eliminen organitzacions, localitzacions i persones, opció configurable al fitxer de constants.py:

```
NER_EXCLUDE = ".*-(PER|ORG|LOC)".
```

Com es pot observar s'ha decidit no eliminar les entitats MISC, ja que es pensa que poden ser rellevants per a la predicció, tot i que com s'ha comentat poden ser diferents en cada país i, per tant, en predir la posició política siguin irrellevants sinó han estat considerades a l'entrenament. En tot cas, durant les proves realitzades sí que s'ha vist alguna entitat MISC internacional que té especial rellevància avui dia com UCRANIAN.

A continuació, els *tokens* es converteixen a minúscules, ja que si es feia abans del NER no reconeixia correctament algunes entitats, i tot seguit s'exclouen referències temporals, webs i dominis, opció configurable al fitxer de constants.py:

```
## Exclude temporal references and web domains
EXCLUDE_WORDS = (
    "^(www|http|https|telnet|mailto|ftps|"
    "monday|tuesday|wednesday|thursday|friday|saturday|sunday|"
    "january|february|march|april|may|june|july|august|september|"
    "october|november|december|"
    "year|month|week|day|today|yesterday|tomorrow|"
    "now|soon|later|before|after|early|late|"
    "morning|afternoon|evening|night|midnight|noon|"
    "recently|previously|currently|already|ago|"
    "second|seconds|minute|minutes|hour|hours|"
    "weekend|holiday|season|spring|summer|autumn|fall|winter|"
    "decade|decades|century|centuries|millennium|millennia|"
    "always|never|sometimes|once|still|eventually|immediately)"
)
```

Finalment, es combinen els *tokens* finals per país i per dia de publicació i es guarden com a parquet i com a CSV local, on diàriament s'aniran acumulant.

```
### 9: Save into HDFS as parquet:  /TFM/cleaned/news
### 10: Save into local CSV:  /home/roser/TFM/data/cleaned/news_clean.csv
```

(El fitxer local s'utilitzarà en la visualització de *Power BI* per a mostrar el *Word Cloud*, ja que altrament com a *parquet* al sistema distribuït de fitxer de *Hadoop* la connexió que es tenia no era gaire estable, i a més es complica a l'haver de recuperar múltiples fitxers.)

A continuació es pot visualitzar una mostra de com queda la combinació de les notícies processades per al dia 20 d'abril del 2025 per a Espanya emmagatzemades al CSV local:

Fitxer: news_clean.csv

```
iso_code,pubDate,words_tex
es,2025-04-20,news novelty sinful brothel bad political practice corruption
walls pit cobble street detain time beautiful medieval people information
catalan people danger extinction small inhabitant positive unpunctual
university professor scientific disseminator investigate predictive model dream
```

```
perform televised live event combine science comedy home clear small bathroom
gain style order functional elegant shelf information launch beautiful dish
handmade cost euro streaming content platform continue expand series series
film program production occupy miniseries consume case talk sacrifice painters
facet painter ahead criticize low price healthy delicious ingredient perfect
include recipe information innovative seafood triumph recommend doctor load
omega celebrate historical moment future world country subject deep change
geopolitical socioeconomic situation world full uncertainty challenge people
reaffirm place draw future unbeatable time write suffer addition find eye
tribulation endless err insufferable bachelor geography historian vocation dive
encyclopedia search romans escape bike friend leave begin work monastery strike
capital emblematic space act host guide displace extensive historical cultural
knowledge
```

3.4 Extracció de característiques

En aquest punt ja es tenen les dades netes i preparades per a generar el conjunt d'entrenament i test per a entrenar el model, i per a la predicció de les posicions polítiques segons les notícies diàries que es reben.

Per a generar el conjunt de dades per a l'entrenament del model, cal incorporar l'etiqueta per a cada agrupació de notícies diàries, per tant, es necessitarà la llista de períodes electorals dels presidents dels Estats Units, dels quals s'extraurà la data inici i final del mandat de cada president i la posició política, a continuació es mostra un exemple, tot i que com a punt rellevant, la data de nomenament dels presidents dels Estats Units és aproximadament 76 dies més tard del dia de les eleccions, per aquest motiu a la data de nomenament se li ha restat aquests dies. Així a l'hora de seleccionar les notícies publicades en cada període electoral aquestes no influiran en la predicció, ja que durant aquests 76 dies ja se sap, o bé hi ha una alta certesa de qui és el guanyador, per aquest motiu s'exclouen durant el procés d'entrenament i test del model. Aquesta mesura no hauria sigut necessària si durant la recollida del conjunt de dades dels presidents dels Estats Units, s'hagués anotat el dia de les eleccions en lloc del nomenament. En tot cas és important excloure-les, ja que com s'explicarà en breu, les notícies pròximes al dia de les eleccions prenen més rellevància en ponderar-les.

iso_code	nom	nomenament	posicio_politica	label	before_nomenament	days_before_apointment
us	Richard Nixon	1972-11-05	dreta	1	1968-11-05	1461
us	Gerald Ford	1974-05-25	dreta	1	1972-11-05	566
us	Jimmy Carter	1976-11-05	esquerra	0	1974-05-25	895
us	Ronald Reagan	1980-11-05	dreta	1	1976-11-05	1461
us	Ronald Reagan	1984-11-05	dreta	1	1980-11-05	1461

Com es pot observar també s'han calculat els dies de cada període en el qual ha governat el president electe. Aquesta nova característica s'utilitzarà per a ponderar les notícies diàries, així segons si les notícies emeses estan més pròximes al dia de les eleccions el model els hi donarà major importància, i a la inversa, com més dies faci de les notícies, menys importants seran per al model, encara que si és una notícia que es manté en el temps la ponderació i el mateix model (aquest punt del model s'explicarà en la secció de la creació dels models) li donarà rellevància. Aquest matís de ponderar les notícies segons la data de publicació s'ha incorporat pensant en la justificació del context

i l'abstracte, on es comentava que a mesura que el temps passa es van oblidant segons quins successos s'han viscut, vist o sentit, i per aquest motiu s'han ponderat. A continuació es mostra un exemple del conjunt de dades de notícies antigues ponderat, remarcant les notícies agrupades per dia amb més i menys pes:

iso_code	pubDate	words_text	day_to_app	weight	word_count	before_app	next_app	label
us	1971-01-01	theodore winner l...	674	**0.001481...**	179	1968-11-05	1972-11-05	1
us	1971-01-02	jan announce pres...	673	0.001483...	146	1968-11-05	1972-11-05	1
us	1971-01-03	award grant total...	672	0.001485...	109	1968-11-05	1972-11-05	1
us	1971-01-04	election represen...	671	0.001488...	153	1968-11-05	1972-11-05	1
us	1971-01-05	success forthcomi...	670	0.001490...	126	1968-11-05	1972-11-05	1
us	1971-01-06	hong kong jan mun...	669	0.001492...	115	1968-11-05	1972-11-05	1
us	1971-01-07	mask robber arm p...	668	0.001494...	148	1968-11-05	1972-11-05	1
us	1971-01-08	mrs widow chairma...	667	0.001497...	103	1968-11-05	1972-11-05	1
...								
...								
...								
us	1972-10-26	ceasefirea ceasef...	10	0.090909...	199	1968-11-05	1972-11-05	1
us	1972-10-27	report gain reven...	9	0.1	130	1968-11-05	1972-11-05	1
us	1972-10-28	letters editor da...	8	0.111111...	173	1968-11-05	1972-11-05	1
us	1972-10-29	obvious theater o...	7	0.125	241	1968-11-05	1972-11-05	1
us	1972-10-30	riverside oct fol...	6	0.142857...	137	1968-11-05	1972-11-05	1
us	1972-10-31	undersigned econo...	5	0.166666...	120	1968-11-05	1972-11-05	1
us	1972-11-01	receive million c...	4	0.2	118	1968-11-05	1972-11-05	1
us	1972-11-02	style choreograph...	3	0.25	145	1968-11-05	1972-11-05	1
us	1972-11-03	puissance strengt...	2	0.333333...	112	1968-11-05	1972-11-05	1
us	1972-11-04	bank hold company...	1	**0.5**	77	1968-11-05	1972-11-05	1

En canvi, per a generar el conjunt de dades per a la predicció de les notícies actuals, es calcula des de les últimes eleccions/nomenament de cada país fins avui dia, per calcular els dies transcorreguts i poder ponderar les notícies diàries de cada país. Per exemple, per al dia 20 d'abril tenim que:

```
## Show days from last presidential elections to today of each country:
```

iso_code	nom	nomenament	posicio_politica	label	today	**days_last_apointment**
at	Christian Stocker	2025-03-03	dreta	1	2025-04-20	48
be	Bart De Wever	2025-02-03	dreta	1	2025-04-20	76
bg	Rumen Radev	2017-01-22	esquerra	0	2025-04-20	3010
cy	Níkos Christodoul...	2023-02-28	centre	2	2025-04-20	782
cz	Petr Pavel	2023-03-09	dreta	1	2025-04-20	773
de	Olaf Scholz	2021-12-08	esquerra	0	2025-04-20	1229
dk	Mette Frederiksen	2022-11-01	esquerra	0	2025-04-20	901
...						
pt	Marcelo Rebelo de...	2021-03-09	dreta	1	2025-04-20	1503
ro	Ilie Bolojan	2025-02-12	dreta	1	2025-04-20	67
se	Ulf Kristersson	2022-10-18	dreta	1	2025-04-20	915
si	Nataša Pirc Musar	2022-12-23	centre	2	2025-04-20	849
sk	Peter Pellegrini	2024-06-15	esquerra	0	2025-04-20	309
us	Donald Trump	2025-01-20	dreta	1	2025-04-20	90

Tal com es pot observar hi ha països amb presidents de posició central amb etiqueta a 2, tal com s'ha explicat en el punt 3.1, la posició independent i centre s'han unificat per a simplificar el model, en tot cas es podria considerar com a millora en futures actualitzacions fer una classificació amb múltiples classes, ja que les posicions polítiques extremes també s'han agrupat als seus extrems.

Un cop creat el *data frame* anterior ja es pot preparar el *dataset* per a predir per país en les pròximes eleccions la posició política, a continuació es mostra un exemple per al país d'Àustria, remarcant el dia amb més pes, ja que és el més pròxim al dia actual, i mostrant també la part del codi per calcular-lo:

```
# Filter between dates range.
df_filtrat_data = df_last_news_filtered.filter((F.col("pubDate") > F.lit(date_start))
& (F.col("pubDate") <= F.lit(date_end)))

# Calculate days to next appointment.
df_filtrat_data = df_filtrat_data.withColumn("day_last_app",
F.datediff(F.lit(date_end), F.col("pubDate")))

# Calculate weight, weight up if the publication date is close to the next
appointment, in this case today. Calculate based on the number of days since the last
presidential election, calculated in the previous step.
df_ponderat = df_filtrat_data.withColumn("weight", 1 / (1 +
F.col("day_last_app").cast("double")))
```

```
## Preparing weighted samples [at] from last appointment 2025-03-03 to 2025-04-20: label: 1
+--------+----------+--------------------+------------+---------+----------+----------+----------+-----+
|iso_code|   pubDate|          words_text|day_last_app|   weight|word_count|before_app|  next_app|label|
+--------+----------+--------------------+------------+---------+----------+----------+----------+-----+
|      at|2025-03-23|germans drive end...|          28|0.0344825|       101|2025-03-03|2025-04-20|    1|
|      at|2025-04-07|wide comparison s...|          13|0.0714285|       334|2025-03-03|2025-04-20|    1|
|      at|2025-03-26|incension questio...|          25|0.0384615|       266|2025-03-03|2025-04-20|    1|
|      at|2025-03-25|rosa pink paysafe...|          26|0.0370370|       209|2025-03-03|2025-04-20|    1|
|      at|2025-03-28|dpa afx petrol pe...|          23|0.0416666|       214|2025-03-03|2025-04-20|    1|
+--------+----------+--------------------+------------+---------+----------+----------+----------+-----+
```

Tot i que actualment no es fa servir, s'ha decidit calcular les paraules diàries de les notícies agrupades per país, la columna *"word_count"* en fa referència. Si bé s'ha pogut observar que per cada país el valor acostuma a ser similar, sí que s'ha vist a vegades el que sembla *outliers*, malgrat això, no es tracten, ja que les notícies venen de diferents mitjans de comunicació, i cadascun deu tenir la seva forma d'explicar, o bé en segons quins temes n'aprofundeixen més. A continuació s'observa les diversitats entre països i en el mateix país:

```
## Preparing weighted samples [us] from last appointment 2025-01-20 to 2025-04-20: label: 1
+--------+----------+--------------------+------------+---------+----------+----------+----------+-----+
|iso_code|   pubDate|          words_text|day_last_app|   weight|word_count|before_app|  next_app|label|
+--------+----------+--------------------+------------+---------+----------+----------+----------+-----+
|      us|2025-03-20|respect spirit ru...|          31|  0.03125|        99|2025-01-20|2025-04-20|    1|
|      us|2025-03-24|prime minister pl...|          27|0.0357142|        76|2025-01-20|2025-04-20|    1|
|      us|2025-03-21|controversial fil...|          30|0.0322580|        87|2025-01-20|2025-04-20|    1|
|      us|2025-04-01|statement attack ...|          19|     0.05|        56|2025-01-20|2025-04-20|    1|
|      us|2025-03-30|add security cabi...|          21|0.0454545|        82|2025-01-20|2025-04-20|    1|
+--------+----------+--------------------+------------+---------+----------+----------+----------+-----+


## Preparing weighted samples [it] from last appointment 2022-02-03 to 2025-04-20: label: 0
+--------+----------+--------------------+------------+---------+----------+----------+----------+-----+
|iso_code|   pubDate|          words_text|day_last_app|   weight|word_count|before_app|  next_app|label|
+--------+----------+--------------------+------------+---------+----------+----------+----------+-----+
|      it|2025-04-07|world story small...|          13|0.0714285|       220|2022-02-03|2025-04-20|    0|
|      it|2025-04-08|background murder...|          12|0.0769230|       137|2022-02-03|2025-04-20|    0|
|      it|2025-03-28|router excellent ...|          23|0.0416666|       137|2022-02-03|2025-04-20|    0|
|      it|2025-03-27|reward fage contr...|          24|     0.04|       181|2022-02-03|2025-04-20|    0|
|      it|2025-03-18|collection signat...|          33|0.0294117|      1527|2022-02-03|2025-04-20|    0|
+--------+----------+--------------------+------------+---------+----------+----------+----------+-----+
```

En aquest punt els conjunts de dades ja estan preparats per a entrenar el model, amb les dades de les notícies antigues, i per a predir les posicions de les notícies recents. A continuació s'explica com s'ha creat el model de predicció de les posicions polítiques.

3.5 Creació dels models *PySpark ML, Pipeline*

Amb el conjunt de dades etiquetat i amb els pesos ja es pot procedir a l'entrenament del model. En primer lloc, s'entrenarà un model de Regressió Logística[15], establint com a model base el model *Naive Bayes*, classificador probabilista simple basat en teorema de *Bayes*, el qual suposa independència entre característiques a diferència de la Regressió logística.

Per ambdós models s'utilitzarà una graella i validació creuada per cercar els millors paràmetres, avaluant la mètrica *AUC* (*Area Under de Curve*), tot i que també es contempla F1-Score (per les notícies diàries, i per als períodes electorals de mandats complets). Per tant, s'emmagatzema el model amb *AUC* superior, que com més pròxim a 1 sigui millor discriminarà, i aquest model serà el que s'utilitzarà per a fer les prediccions diàries.

Addicionalment, s'emmagatzema també les paraules del vocabulari amb els coeficients que s'extrauen del model, obtenint així les preocupacions i el seu pes, objectiu que també s'havia definit.

De les paraules agrupades de les notícies diàries de cada país amb els seus pesos, es crearà el conjunt d'entrenament i el de test, que com es veurà estan esbiaixades en un 25% aproximadament a favor de la classe 1 (posició política dreta) fet que pot influir en els resultats finals.

```
## Label count in news_train::
+-----+-----+
|label|count|
+-----+-----+
|    1| 9172|
|    0| 6619|
+-----+-----+
```

```
## Label count in news_test:
+-----+-----+
|label|count|
+-----+-----+
|    1| 2253|
|    0| 1575|
+-----+-----+
```

A continuació es crea una canalització amb la següent seqüència d'etapes, en aquest cas per al model de Regressió Logística:

1. **Tokenizer**: Retorna el text en *tokens*.
2. **CountVectorizer**[16]: Converteix la llista de *tokens* en un vector numèric on cada número representa la freqüència de la paraula del vocabulari.
3. **IDF**[16]**:** Calcular la freqüència inversa per donar més rellevància a les paraules importants, reduint els pesos dels termes que apareixem amb més freqüències.
4. **LogisticRegresion()**: Crea el model per a la classificació que utilitzarà les característiques extretes, indicant-li la columna de pes 'weight'.

```
 # Tokenizer text to get: num features, prepare count Vectorised, and
 # converts hashed symbols to TF-IDF.
tokenizer = Tokenizer(inputCol="words_text", outputCol="words")

# Get count Vectorizet to get vocaulary.
```

```
cv = CountVectorizer(inputCol="words", outputCol="countV")

# Convert hashed symbols to TF-IDF.
tf_idf = IDF(inputCol='countV', outputCol='features')

# Create a logistic regression object and add everything to a pipeline.
logistic = LogisticRegression(weightCol="weight",
                              featuresCol="features",
                              labelCol="label",
                              predictionCol="prediction")
pipeline = Pipeline(stages=[tokenizer, cv, tf_idf, logistic])
```

Per al model de *Naive Bayes* es crea una nova canalització canviat l'última etapa:

```
# Create model Naive Bayes.
nb = NaiveBayes(weightCol="weight",
                featuresCol="features",
                labelCol="label",
                predictionCol="prediction")

pipeline_nb = Pipeline(stages=[tokenizer, cv, tf_idf, nb])
```

L'optimització dels hiperparàmetres amb validació creuada se centra en la mida del vocabulari, que s'ha extret de l'etapa de "CountVectorizer", i la regularització per a evitar el *overfitting* en el model de *Logistic Regression*. Per al model de Naive *Bayes* es focalitza en el suavitzat per evitar el problema de la probabilitat zero, i el tipus de model "Multinomial" o "Gaussian", el primer per a dades contínues i el segon per a discretes, a més de la mida del vocabulari.

```
# Add grid for countVectorizer parameters and logistic regression parameters, and
build.
params = params.addGrid(cv.vocabSize, [1000, 10000,  num_features]) \
            .addGrid(logistic.regParam, [0.001, 0.01, 0.1]).build()

print('Number of models to be tested: ', len(params))
evaluator = BinaryClassificationEvaluator()
# Create cross-validator for logistic regression.
crossval = CrossValidator(estimator=pipeline,
                          estimatorParamMaps=params,
                          evaluator=evaluator,
                          numFolds=5)

# Run cross-validation, and choose the best set of parameters.
cv_model_l = crossval.fit(news_train)

# Construction of the parameter grid.
param_grid_nb = ParamGridBuilder() \
        .addGrid(cv.vocabSize, [1000, 10000, num_features]) \
        .addGrid(nb.smoothing, [0.5, 1.0, 1.5]) \
        .addGrid(nb.modelType, ["multinomial", "gaussian"]) \
        .build()

print('Number of models to be tested: ', len(param_grid_nb))
# CrossValidator for Naive Bayes.
crossval_nb = CrossValidator(estimator=pipeline_nb,
                        estimatorParamMaps=param_grid_nb,
                        evaluator=evaluator,
                        numFolds=5)
```

```
    # Run cross-validation, selecting the best parameters and model.
    cv_model_nb = crossval_nb.fit(news_train)
```

L'avaluació dels diferents models, tant per notícies diàries, com agrupant aquestes per períodes de mandats electorals, mostren el millor AUC en el Model de Regressió Logística optimitzat amb la cerca d'hiperparàmetres en la validació creuada, tal com es pot veure a continuació la comparativa:

```
## Naive Bayes                  AUC: 0.4848 F1-Score: 0.5197 F1-Score-period: 0.4500
## Naive Bayes (CV)             AUC: 0.5617 F1-Score: 0.5733 F1-Score-period: 0.8667
## Logistic Regression          AUC: 0.7081 F1-Score: 0.6585 F1-Score-period: 0.9321
## Logistic Regression (CV)     AUC: 0.7563 F1-Score: 0.6915 F1-Score-period: 0.9321

## Report metrics of prediction by day:
              precision    recall  f1-score   support

           0       0.65      0.57      0.61      1617
           1       0.72      0.79      0.75      2289

    accuracy                           0.70      3906
   macro avg       0.69      0.68      0.68      3906
weighted avg       0.69      0.70      0.69      3906


## Report metrics of prediction grouped by election period
              precision    recall  f1-score   support

           0       1.00      0.83      0.91         6
           1       0.90      1.00      0.95         9

    accuracy                           0.93        15
   macro avg       0.95      0.92      0.93        15
weighted avg       0.94      0.93      0.93        15
```

Tot i que s'ha vist que el F1-Score per període és alt per al model de *Naive Bayes* optimitzat, i per al model de Regressió Logística base és igual al de la validació creuada, la resta de mesures són inferiors, i aquest fet pot ser a causa de les dades seleccionades del conjunt de test, i a la independència de les característiques en el model de *Naive Bayes*, i amb relació al model base de Regressió logística tot i que les mesures AUC són similars, van de 5 punts, a més la validació creuada és més robusta.

Pel que fa al resum de mètriques de predicció de notícies diàries, s'observa que el model classifica millor la classe 1 (dreta), amb una diferència de 14 punts a la mètrica f1-score, fet que pot ser causat pel biaix de les classes que es tenen al conjunt de dades, així el model aprèn millor a classificar la classe majoritària. Tot i que, si s'observa l'agrupació dels resultats per període electoral, mètrica f1-score, informa que la classe millor classificada és també la 1 (dreta), però en aquest cas amb una diferència de 4 punts, aquest fet pot ser arran dels pocs períodes electorals que es tenen al conjunt de dades, que esdevindrien per al conjunt de test 15, és a dir les notícies diàries seleccionades aleatòriament per al conjunt de test pertanyen a 15 períodes de mandats electorals, d'un total de 17 presidències des dels 1970.

Es podria pensar, i perquè no s'han agrupat les notícies diàries per període electoral, doncs perquè únicament s'hagués obtingut un conjunt de dades amb 17 mostres, i no

s'hagués pogut incorporar l'oblit en el temps, tal com s'ha implementat amb el pes que se li ha donat a les notícies més o menys llunyanes respecte al seu període electoral, en l'apartat anterior al preparar el conjunt de dades.

Per tant, serà el model que es guardarà i s'utilitzarà per a la predicció diària de les notícies de la resta de països seleccionats. Addicionalment, es guarden les mètriques i les paraules del vocabulari amb els seus coeficients, per mostrar-ho a la visualització. A continuació es mostren les 10 més positives i negatives del model:

```
## More positive words of model:
        word    weigh
        teem 0.165099
   ukrainian 0.142513
   threepart 0.138753
      sandal 0.136789
     omicron 0.130181
    pictures 0.129711
    glossary 0.128541
     healing 0.127708
moderndance 0.126893
```

```
## More negative words of model:
        word     weigh
noncallable -0.197734
    noncash -0.192939
   dealbook -0.190634
  cincluded -0.183377
supervisors -0.170288
        ply -0.166269
      serbs -0.165417
    patriot -0.158624
       wins -0.157935
```

Per finalitzar, es mostra la predicció del conjunt de test agrupada per període electoral:

```
#Sample of prediction grouped by election period and
counting label predicted:
+--------+----------+----------+-----+----------+-----+
|iso_code|before_app|  next_app|label|prediction|count|
+--------+----------+----------+-----+----------+-----+
|      us|1968-11-05|1972-11-05|    1|       0.0|   25|
|      us|1968-11-05|1972-11-05|    1|       1.0|  182|
|      us|1972-11-05|1974-05-25|    1|       1.0|   70|
|      us|1972-11-05|1974-05-25|    1|       0.0|   28|
|      us|1974-05-25|1976-11-05|    0|       0.0|   53|
|      us|1974-05-25|1976-11-05|    0|       1.0|  108|
|      us|1976-11-05|1980-11-05|    1|       1.0|  183|
|      us|1976-11-05|1980-11-05|    1|       0.0|   70|
...
|      us|2004-11-05|2008-11-05|    0|       0.0|  183|
|      us|2004-11-05|2008-11-05|    0|       1.0|  112|
|      us|2008-11-05|2012-11-05|    0|       0.0|  175|
|      us|2008-11-05|2012-11-05|    0|       1.0|   98|
|      us|2012-11-05|2016-11-05|    1|       0.0|  132|
|      us|2012-11-05|2016-11-05|    1|       1.0|  149|
|      us|2016-11-05|2020-11-05|    0|       0.0|  187|
|      us|2016-11-05|2020-11-05|    0|       1.0|  122|
|      us|2020-11-05|2024-11-05|    1|       1.0|  201|
|      us|2020-11-05|2024-11-05|    1|       0.0|   83|
+--------+----------+----------+-----+----------+-----+
```

Amb el model entrenat i emmagatzemat, diàriament amb la descàrrega de les notícies dels diferents països, després de netejar-les i ser processades es fa la predicció per a cada país de les notícies emeses des de l'últim nomenament fins al dia actual, obtenint el recompte diari de classificacions i finalment la predicció, la que es combina amb el fitxer de presidents CSV i que s'utilitzarà en la visualització:

Fitxer: predictions.csv

```
iso_code,nom,nomenament,posicio_politica,label
at,Bruno Kreisky,21-04-1970,esquerra,0
```

```
at,Fred Sinowatz,24-05-1983,esquerra,0
at,Franz Vranitzky,16-06-1986,esquerra,0
at,Viktor Klima,28-01-1997,esquerra,0
...
bg,Petar Stoyanov,22-01-1997,dreta,1
bg,Georgi Parvanov,22-01-2002,esquerra,0
bg,Rosen Plevneliev,22-01-2012,dreta,1
bg,Rumen Radev,22-01-2017,esquerra,0
cy,Makarios III,26-02-1968,esquerra,0
cy,Makarios III,18-02-1973,esquerra,0
cy,Makarios III,07-12-1974,esquerra,0
...
us,Barack Obama,20-01-2013,esquerra,0
us,Donald Trump,20-01-2017,dreta,1
us,Joe Biden,20-01-2021,esquerra,0
us,Donald Trump,20-01-2025,dreta,1
at,PREDICTION,2025-04-22,dreta,1
be,PREDICTION,2025-04-22,dreta,1
bg,PREDICTION,2025-04-22,dreta,1
...
nl,PREDICTION,2025-04-22,esquerra,0
pt,PREDICTION,2025-04-22,dreta,1
ro,PREDICTION,2025-04-22,dreta,1
se,PREDICTION,2025-04-22,dreta,1
si,PREDICTION,2025-04-22,dreta,1
sk,PREDICTION,2025-04-22,dreta,1
us,PREDICTION,2025-04-22,esquerra,0
```

A tall de curiositat, com es pot observar en la predicció del dia 22 d'abril, per als Estats Units prediu que, si hi hagués eleccions presidencials guanyaria l'esquerra.

Referent al model d'agrupació de països segons similitud de característiques en el mateix període de temps, finalment no s'ha implementat perquè solament es tenen notícies antigues dels Estats Units, i no s'ha pogut construir cap altre model amb notícies antigues d'altres països, per tant, el model de predicció entrenat amb les notícies dels Estats Units serà el que s'utilitza per a predir la resta de països.

3.6 Visualització *Power BI real time (window 24h)*

Un cop desenvolupat el diagrama d'activitats, el qual recull l'extracció de dades, la neteja, l'obtenció de característiques adequades per a l'entrenament del model i la predicció en *real time*, amb una finestra de 24 hores, ja es tenen els jocs de dades finals per crear la visualització i donar resposta als pronòstics de les següents eleccions de cada país tractat, així com la projecció de l'evolució de les preocupacions i a quins països afecten.

A tall d'introducció, i encara que s'han anat esmentant en els punts anteriors, a continuació es descriuen els fitxers que s'utilitzaran en el model de la visualització, agrupats per dimensions, i el perquè d'algunes decisions preses, on tots es trobaran en local a la ruta: \\wsl.localhost\Ubuntu\home\roser\TFM\data\cleaned"

Països

- **paisos-ue.csv**: Conté el detall dels països que es tractaran de la Unió Europea, codi ISO del país, nom, i codis ISO dels llenguatges que s'hi parlen.
- **pais-us.csv**: Conté el detall del país dels Estats Units, codi ISO del país, nom, i codi ISO dels llenguatges que s'hi parlen.

Notícies

- **concerns_summary.csv**: Conté la classificació de les notícies diàries de cada país. Les classes són [*Business, World, Sci/Tech, Sports*]. (S'actualitza automàticament cada dia.)
- **news_clean.csv**: Conté les paraules netes i agrupades per dia de les notícies extretes de cada país diàriament. (S'actualitza automàticament cada dia.)
- **old_news_clean.csv**: Conté les paraules netes i agrupades per dia de les notícies antigues dels Estats Units des del 1970. (Es crea un únic cop en descarregar i processar les notícies antigues.)

Model predictiu

- **predictions.csv**: A més de les prediccions de la posició política del dia actual de cada país, conté tots els presidents dels diversos països des del 1970. S'ha de tenir en compte que per a una correcta visualització de l'evolució les dates s'han expandit, així a mesura que els dies passen es continua visualitzant al mapa quina posició política governava a cada país, fins al dia actual de la predicció. (Aquest fitxer es crea automàticament cada dia.)
- **probability_predictions.csv**: Per tal de donar informació més precisa es mostra la probabilitat de les prediccions, les quals s'han extret del recompte de prediccions diàries de les notícies de cada país, des del nomenament del president de les últimes eleccions, donant així una referència detallada de per quant guanya una posició de l'altra. En aquest apartat tot i que el model de predicció únicament prediu 0 o 1 (esquerra o dreta), en fer el recompte si hi ha empat la predicció mostrarà un 2 el qual s'indicarà amb un color distint al mapa de que és un empat, i que, a diferència de les dades antigues dels presidents passats, els quals eren independents o de centre que es van reagrupar a centre amb la classe 2, aquest també es mostrarà en un color diferent dels empats, per tal de no confondre a l'usuari. Així que en total al mapa es mostraran 4 colors, per a dreta el blau, esquerra el vermell, centre/independent amb taronja, i empat amb groc, on empat únicament es mostrarà el dia actual de la predicció si es dona el cas. (Aquest fitxer es crea automàticament cada dia.)
- **concerns_words.csv**: Conté el vocabulari i els coeficients del model entrenat. Així es pot visualitzar les paraules més positives/negatives i estudiar-les. (Es crea únicament en finalitzar l'entrenament del model.)

- **metrics.csv**: Inclou el nom del model que ha extret millor resultats, així com les mètriques d'avaluació que s'han considerat i els seus valors. (Es crea únicament en finalitzar l'entrenament del model.)

3.6.1 Model de la visualització

Amb relació al model de la visualització, com es veurà a continuació a la figura 7, s'han combinat les taules de països per simplificar les relacions, i la visualització dels noms dels països en els objectes que ho requereixin, com per al mapa, taules il·lustratives i interactives.

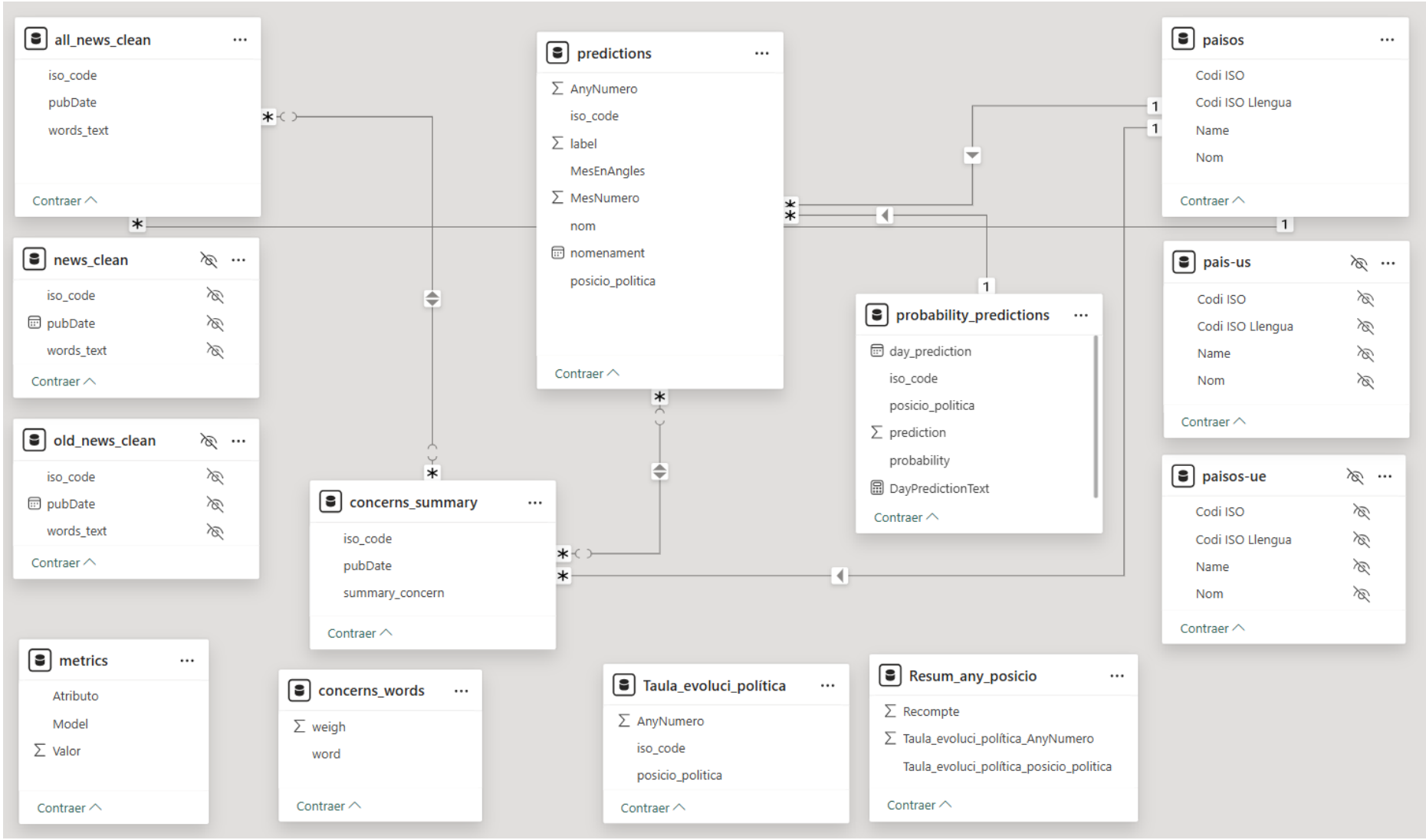
A més també s'han combinat les notícies actuals amb les antigues, aquestes últimes són les que es van fer servir per a l'entrenament del model, així mateix, es pot seguir l'evolució de la trajectòria fàcilment seleccionant les dates. Aquesta unió quedarà reflectida al núvol de paraules on, mentre hi hagi diversos colors indicarà que hi participen diversos països. Ara bé, com les notícies actuals parteixen de mitjans de març, tot el període anterior únicament mostrarà les paraules de les notícies antigues que s'originen de les notícies dels Estats Units, amb un únic color.

Pel que fa als a les taules que representen els fitxers de concerns_words.csv i metrics.csv, no estan relacionades amb les altres perquè esdevenen una informació més tècnica del model entrenat que s'utilitza per a la predicció. Tot i que s'han volgut mostrar pels usuaris més experts, tenint així un detall més profund dels paràmetres que resulten de l'entrenament model.

A partir de la taula prediccions s'ha generat dues taules més: "Taula_evolució_política" i "Resum_any_posició", amb la finalitat de donar resposta a la tendència evolutiva de la posició política amb un gràfic de línies, el qual mostra per a cada any el recompte de posicions polítiques dels països intervinents. Aquestes taules no s'han relacionat amb cap altra, arran de què la gràfica és informativa i no s'ha volgut que interaccioni amb la resta. En tot cas, a partir de les divergències que s'observin a les dades, en captura l'any, aquest pot fer-se servir per explorar la resta de la visualització en el filtre temporal.

La següent figura mostra la representació gràfica del model de dades per a la visualització amb les relacions que hi intervenen, a tall de sincronitzar la informació seleccionada mitjançant els filtres.

Figura 7: Model de la visualització



Font: Elaboració pròpia

Pel que fa a les transformacions que han sofert les dades utilitzant les funcions DAX de *Power Query* han estat:

- Posar les paraules en format tipus títol, per tal de facilitar la lectura principalment en el *Word Cloud*.

- Traduir les etiquetes de la posició política a l'anglès, ja que la visualització mostra la informació principal en anglès:

```
each if [posicio_politica] = "esquerra" then "Left"
     else if [posicio_politica] = "dreta" then "Right"
     else if [posicio_politica] = "centre" then "Center"
     else if [posicio_politica] = "empat" then "Tie"
     else [posicio_politica],
Replacer.ReplaceValue,
{"posicio_politica"}
```

- S'han arrodonit els valors de les mesures de l'avaluació del model a 3 decimals, i transposat la taula, a més d'intercanviar el punt per la coma:

```
#"Canviar punt per coma" = Table.TransformColumns(
   #"Texto recortado",{
      {"AUC", each Text.Replace(Text.From(_, "en-US"), ".", ","), type text},
      {"F1-Score_day", each Text.Replace(Text.From(_, "en-US"), ".", ","),
type text},
      {"F1-Score_period", each Text.Replace(Text.From(_, "en-US"), ".", ","),
type text}}
),
#"Convertit a número" = Table.TransformColumnTypes(
    #"Canviar punt per coma",{
        {"AUC", type number},
        {"F1-Score_day", type number},
        {"F1-Score_period", type number}}
),
#"Round" = Table.TransformColumns(
    #"Convertit a número",{
        {"AUC", each Number.Round(_, 3), type number},
        {"F1-Score_day", each Number.Round(_, 3), type number},
        {"F1-Score_period", each Number.Round(_, 3), type number}}
),
#"Transposar" = Table.UnpivotOtherColumns(#"Round", {"Model"}, "Atributo",
"Valor")
```

- S'ha realitzat el mateix procediment de canviar el punt per la coma, i arrodonit a tres decimals els coeficients dels paràmetres del model, a més de posar en tipus títol el vocabulari del model per facilitar-ne la lectura.

- La taula "Resum_any_posició" s'ha creat en dos passos a partir de la taula "predictions", agrupant per codi de país, any i posició política, excloent les prediccions perquè no surtin els resultats duplicats en l'any actual i el recompte no sigui real:

```
Taula_evoluci_política = GROUPBY(
   FILTER(predictions,predictions[nom]<>"PREDICTION"),
    predictions[iso_code],
    predictions[AnyNumero],
    predictions[posicio_politica])
```

  I a partir d'aquesta taula es genera el recompte que s'utilitzarà per al gràfic de línies que s'explicarà en el següent apartat:

```
Resum_any_posicio = GROUPBY(
    'Taula_evoluci_política',
    'Taula_evoluci_política'[AnyNumero],
    'Taula_evoluci_política'[posicio_politica],
    "Recompte", COUNTX(CURRENTGROUP(),
'Taula_evoluci_política'[posicio_politica]))
```

- S'ha creat una jerarquia per a la data de nomenament per facilitar el filtre individual d'any, mes i dia com a botons, ja que no s'ha trobat la forma de mostrar-ho com una barra de progrés amb una única selecció de data. Tot i que es podia haver fet amb selecció de data inici i fi, aquest rang de dates podria incorporar dos nomenaments en el mateix període d'un país, i en aquest cas, si la posició política hagués sigut diferent no hi hauria coherència en quin color mostrar el país. En tot cas aquesta situació quedaria com un millora en futures versions en les quals donar-hi una pensada.

3.6.2 La visualització

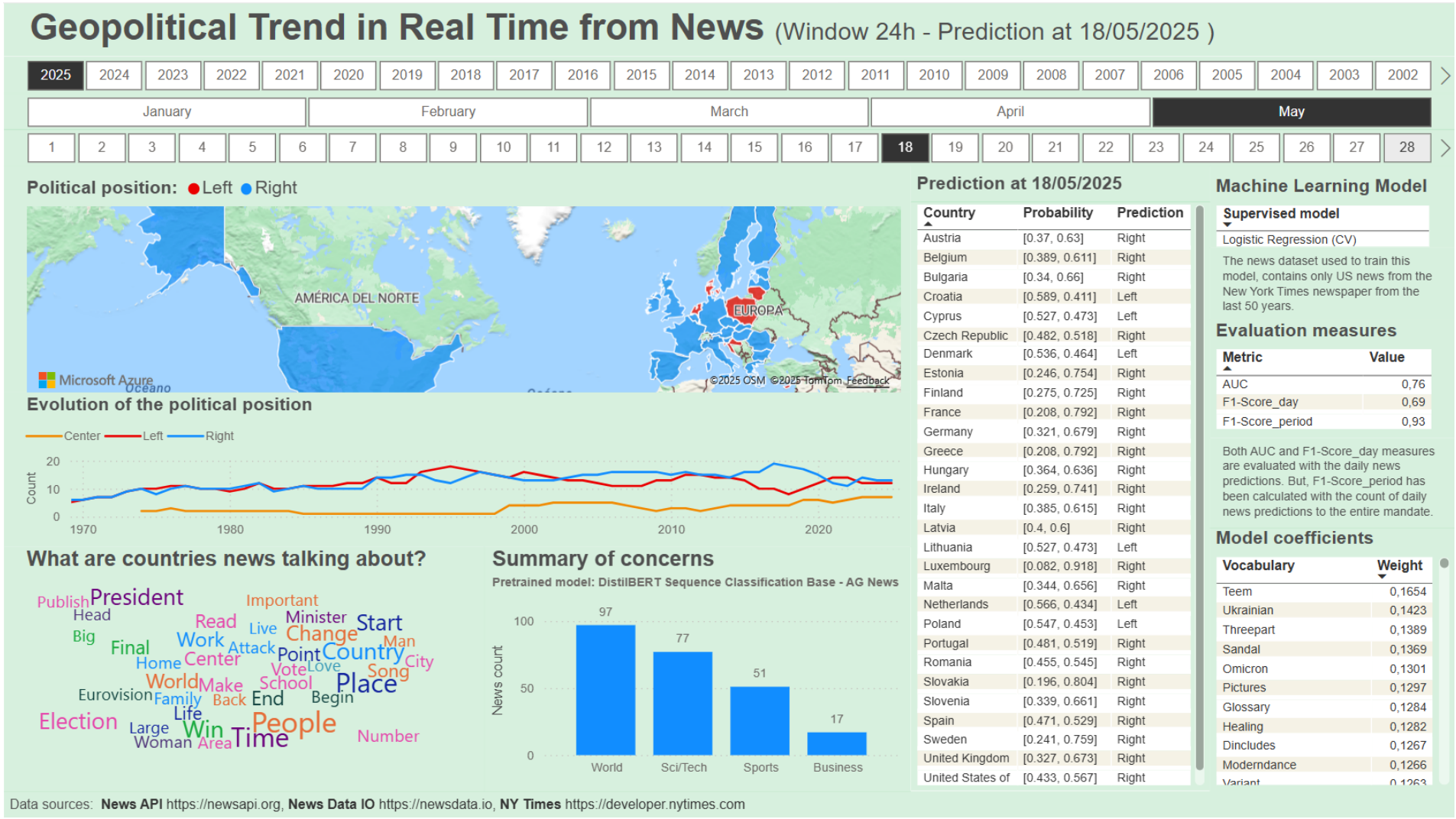
A continuació es mostra la visualització final realitzada amb *Power BI*[17], en la que s'explicaran els diferents filtres, gràfics, taules i les seves interaccions, i que per a la disposició dels diferents elements de la visualització s'ha seguit els criteris bàsics d'ordre natural:

- **En primera posició a dalt i a l'esquerra**: el títol amb una grandària del tipus de lletra rellevant, que resumeix de què va la visualització, seguit d'un subtítol el qual fa més entenedor els detalls de la finestra d'actualització i el dia de la predicció.
- **D'esquerra a dreta**: els filtres principals temporals on seleccionar dates per mostrar l'evolució de les relacions entre posició política i paraules de les notícies.
- **De dalt a baix**: el cos principal que mostra els gràfics amb filtres secundaris i la taula de prediccions, amb el mapa coroplètic com a gràfic principal amb el qual s'obté una focalització involuntària directa al país interessat, seguit d'un gràfic de línies el que resumeix l'evolució de la posició política dels últims 50 anys, a continuació el núvol de paraules acolorit on els usuaris curiosos es fixin en les paraules més comentades de les notícies, i un diagrama de barres vertical que mostra el recompte de tipus de notícies.
- **Als racons**: fonts, llegendes i informació més tècnica per als usuaris experimentats.

En relació al *data-ink Ratio*, proporció de tinta dedicada a la representació de dades, en aquesta visualització seria baix, per culpa del color de fons verd innecessari, però que s'ha volgut posar perquè el mapa quedi integrat, i que a més en ser un color clar de tipus pastel no li treu rellevància a les dades principals, les quals es mostren amb colors més variats i forts.

Amb relació al color intercalat dels registres les taules també s'ha escollit similar als colors del propi mapa, en aquest cas a ubicacions amb poca vegetació, el qual a més facilita la lectura horitzontal entre les diferents columnes.

Figura 8: Visualització Power BI



Font: Elaboració pròpia

Filtres temporals

S'han extret les dates de la taula de prediccions i es mostren en format botó: any, mes i dia. Tot i les limitacions per mostrar-ho com a barra de progrés amb una única data (vegeu punt 3.6.1), permet accedir directament a la data sense haver de desplaçar-se.

La interacció amb aquest filtre afectarà el mapa coroplètic canviant de color els països en vers a la posició política del govern en aquella data, tret del dia actual de la predicció, que mostrarà el color de l'etiqueta predita. També afectarà el núvol de paraules mostrant les paraules més freqüents del dia seleccionat, i filtrarà el recompte del resum de la classificació de notícies en el gràfic de barres del dia seleccionat.

Mapa coroplètic

Aquest mapa mostra els països acolorits segons la posició política del dia seleccionat en el filtre temporal. El mapa permet fer *pan* i *zoom*, a més de seleccionar un país concret, on aquesta selecció interaccionarà directament a:

- **Taula de prediccions**: mostrant únicament el registre del país seleccionat.
- ***Word Cloud***: mostrarà les paraules més freqüents de les notícies del país seleccionat.
- **Gràfic de barres**: filtrarà mostrant únicament el recompte de les notícies classificades per al país seleccionat.

Gràfic de línies

Aquest gràfic pretén mostrar l'evolució de la posició política a manera de visualitzar tendències, observar-hi períodes rellevants, i extraure l'any del punt d'inflexió, amb la intenció d'utilitzar-lo en els filtres temporals per explorar la resta de dades que es mostren a la visualització, donant resposta a la pregunta: que va passar en aquelles dates?, o bé, de què es parlava a llavors i quines eren les preocupacions generals?, en aquell període o data concreta interessada.

Word Cloud

El núvol de paraules mostra les paraules de les notícies de la data seleccionada dels països, on la mida de cada paraula indica la freqüència d'aparició, i els diferents colors denoten la pertinença a diversos països.

En seleccionar una paraula, aquest interacciona filtrant els països en els quals també si ha trobat a les notícies diàries. Per tant, tant el mapa coroplètic com la taula de prediccions mostraran únicament els països filtrats on, per aquest dia concret la paraula seleccionada hi era present a les notícies. De la mateixa forma en tornar a clicar la mateixa paraula el *Word Cloud* torna a l'estat normal, mostrant les paraules freqüents de tots els països que tinguin notícies aquell dia.

Si bé és cert que a mitjans de març és quan s'incorpora la resta de notícies dels països seleccionats, hi ha uns dies que les notícies antigues descarregades i processades se superposen amb les diàries per al país dels Estats Units, per tant, per aquests dies encara que se seleccioni únicament els Estats Units, es veuran dos grups de colors, aquest comportament no estava previst i s'ha observat durant l'exploració final de la visualització. Així mateix, aquest comportament incoherent quedaria com a millora per a solucionar-ho i evitar aquesta superposició en noves versions del producte final.

Resum de preocupacions

El gràfic de barres del resum de preocupacions, extretes amb el model preentrenat de classificació de notícies *DistilBert Sequence Classification Base – AG News*, mostra el recompte de les classes per a tots els països del dia seleccionat. A l'interactuar seleccionat una classe, es filtrant igualment els països en la taula de prediccions, i també en el mapa coroplètic, quedant únicament aquells on hi ha notícies de la classe seleccionada per al dia concret.

No hi ha interacció amb el núvol de paraules, ja que resultaria incoherent, donat que les preocupacions s'han extret per a cada notícia de cada dia de cada país, i en el núvol de paraules es mostren les paraules resultants de la unió de les notícies de cada país per a cada dia. Per tant, en seleccionar la classe "ciència/tecnologia", no es pot saber quines paraules intervenen en aquesta classe, ja que tal com s'ha comentat intervé la descripció sencera de la notícia per a ser classificada.

Si bé el títol d'aquest gràfic inclou la paraula preocupacions, tal com s'hi ha fet referència des de l'inici d'aquest document, que preocupa a les persones perquè decideixin votar per a una posició política o altra, també es podria haver posat com a títol, resum d'interessos, si es visualitza des de la perspectiva que els mitjans de comunicacions informen del que pot interessar a la població. Tot i que preocupacions denota a priori problemes, el to que se li ha volgut donar, es mirat des de la perspectiva que s'informa a la població del que li pot preocupar com a interès especial. Encara que si es mira des de la perspectiva de les funcions dels polítics, en resulta que aquests s'han de preocupar dels problemes dels ciutadans, per tant, preocupacions, problemes i interessos quedarien entrellaçat.

Taula de prediccions

La taula de les prediccions del dia actual, o més ben dit de l'últim dia predit si encara no s'ha obtingut les últimes dades generades, en tot cas el dia predit sempre es mostrarà al títol de la taula de prediccions i al títol de la visualització. Com a interacció, filtra igual que el mapa coroplètic, focalitzant-hi únicament el país seleccionat, mostrant només les paraules més freqüents també del país seleccionat al *Word Cloud*, i resumint el recompte de preocupacions igualment.

Tot i que inicialment no estava previst mostrar la probabilitat, es va considerar especialment rellevant, ja que indica per quant es decanta per una posició política o per l'altra, proporcionant així més informació a l'usuari:

"Es prediu una victòria de... a però per poc!"

Taules d'informació tècnica del model predictiu

Les taules d'informació del model predictiu són només informatives, i no tenen cap mena d'interacció amb la resta. Tot i que per als usuaris més experts donen confiança i credibilitat extra als resultats de les prediccions mostrades a la visualització. Seria interessant que en futures versions es mostrés també els valors dels paràmetres optimitzats del model.

Addicionalment, encara que el vocabulari amb el seu coeficient no interaccionen amb les paraules del *Word Cloud*, es pot visualitzar ordenant la taula, les que tenen un pes més positiu i més negatiu, a més de cercar-les alfabèticament. En tot cas, quedaria com una millora en noves versions poder filtrar la resta de gràfics pel vocabulari del model, així com es fa seleccionant una paraula del *Word Cloud,* i també crear un mapa de calor mostrant les paraules mes i menys rellevants d'aquell dia per aquell país.

3.7 Limitacions

La limitació més destacada ha estat els recursos de memòria de la instal·lació de l'entorn en un orinador de taula, ja que no ha permès realitzar el tractament de dades amb *Apache HIVE* i s'ha optat per l'ús directe de HDFS. Aquest canvia ha comportat que la connexió de *Power BI* a les dades finals per a la visualització hagi canviat.

Així i tot, el volum de notícies antigues descarregades per a l'entrenament del model ha estat tal, 1,5 Gb aproximadament, que en generar l'extracció de característiques ha sigut inviable el seu tractament. Per aquest motiu, finalment s'ha seleccionat dels últims 55 anys únicament 10 notícies diàries, que resulten unes 300 mensuals. Tenint en compte que es van descarregar mensualment entre 3000 i 8000 notícies mensuals, s'ha descartat una quantitat immensa de mostres per limitacions de l'entorn. Tot i això, sembla que l'avaluació del model ha donat bastant bons resultats, fet que s'haurà d'anar contrastant a mesura que se celebren eleccions als països Europeus contemplats.

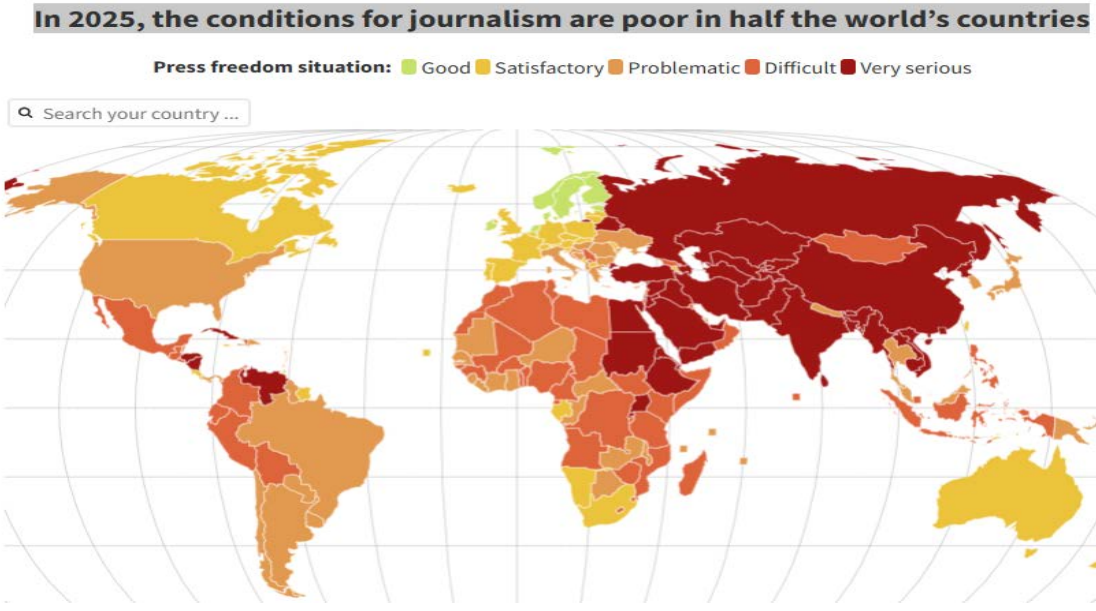
En aquest sentit, tot i que es podia haver optat per un entorn *Cloud* preconfigurat i escalable com AWS, aquesta opció en local ha donat l'oportunitat de resoldre diferents dificultats aprenent aspectes de sistemes no relacionats directament amb el treball d'aquest document, tot i que quedaria pendent com a millora incloure tota la configuració i dependències de l'entorn així com el sistema, en un contenidor com *Docker compose*, donant lloc a la facilitat de reproducció. Cal comentar també que aquestes limitacions de recursos han condicionat el desenvolupament del codi, fent que sigui seqüencial en lloc de paral·lel, com seria normal amb un productor i un consumidor.

Amb relació a les versions gratuïtes de les *API* no han permès la descàrrega de notícies antigues per països, per tant, no s'ha pogut validar la bondat del model entrenat únicament amb notícies dels Estats Units d'Amèrica, utilitzant les notícies antigues d'altres països. Si bé s'ha trobat notícies antigues d'Europa, o bé no estaven classificades per país o no tenien la data de publicació, de manera que no es podien relacionar amb el període electoral correcte per etiquetar-les amb la posició política adequada. Tot i que aquest fet (model entrenat únicament amb notícies dels Estats Units), s'ha indicat a la visualització per tal que l'usuari pugui interpretar correctament les prediccions.

Addicionalment, l'ús d'*API* gratuïtes per a descarregar les últimes notícies de cada país, amb un decalatge de 24 hores, no ha resultat real ben bé del tot. Doncs a vegades en descarregar les últimes notícies d'un país algunes estaven publicades amb data d'ahir, d'avanç d'ahir, i d'altres amb data d'avui. Per tant, pot haver donat lloc a notícies descarregades no processades adequadament. Aquesta situació podria solucionar-se amb una versió de pagament, al poder escollir les notícies d'un dia concret, en tenir accés a l'històric. O bé si no s'hagués tingut limitacions de memòria, s'hauria pogut realitzar el processament en paral·lel i no de forma seqüencial tal com s'ha fet.

En tot cas amb una versió de pagament també es podria haver descarregat més notícies per cada país, tenint així més varietat de dades provinents de més diversos mitjans de comunicació. Ja que, si es revisa l'article de: Bocandé, Anne (2025). "RSF World Press Freedom Index 2025: economic fragility a leading threat to press freedom" *RSF Reporters Without Borders* [article en línia] [Data de consulta: 6 de maig del 2025] <https://rsf.org/en/rsf-world-press-freedom-index-2025-economic-fragility-leading-threat-press-freedom?year=2025&data_type=general>, es pot observar que la llibertat de premsa mundial decau, per tant, com més mitjans de comunicació hi participin millor.

Figura 9: Press Freedom index in Europe in 2025



Font: World Press Freedom Index, Link to share ©Reporters Without Borders

Pel que fa al seguiment de preocupacions entre països, amb la visualització actual resulta difícil donar resposta, tal com es comentava a l'abstracte, per saber quins països no tenen absència escolar i fer una investigació addicional a tall de modificar el rumb del nostre en cas que sí que en tingués. Així mateix, ni que finalment s'hagués agrupat els països per preocupacions similars, no es creu que s'hagués obtingut fàcilment resposta en aquesta qüestió o similars. El principal motiu podria ser a causa que les característiques es tracten en format paraula, i per facilitat el seguiment es podria haver incorporat *bigrams* i *trigrams*, ja que actualment es podria trobar la paraula '*absence'* i '*school*' però no 'school absence', així i tot, l'extracció de les característiques s'haguessin ampliat exponencialment i els recursos locals haguessin sigut insuficients. Per tant, aquesta millora quedaria pendent en entorns més professionals i escalables.

Referent al model d'agrupació de països segons similitud de característiques en el mateix període de temps, finalment no s'ha implementat a causa que solament es tenen notícies antigues dels Estats Units, i no s'ha pogut construir cap altre model amb notícies antigues d'altres països, per tant, el model de predicció entrenat amb les notícies dels Estats Units és l'únic que s'utilitza per a predir la resta de països.

Power BI web ha donat problemes tècnics per enllaçar els conjunts de dades directament generats a Ubuntu en local, com a opció alternativa es podien pujar els fitxers, però aquesta opció no permet actualitzar fàcilment les actualitzacions diàries dels jocs de dades, i tampoc s'ha pogut connectar en local directament al sistema distribuït de fitxers de *Hadoop*. Per tant, el producte final és local, tot i que la visualització es pot compartir en públic, no es podrà actualitzar les dades del model de forma automatitzada. Pel que fa a l'actualització diària automàtica dels fitxers que contenen el model, *Power BI* d'escriptori no permet la programació d'horari automatitzat, havent-se de refrescar manualment mitjançant l'opció de refrescament/actualització del menú. En tot cas amb *Power BI* web un cop es publica la visualització, tampoc permet un refrescament de les dades de forma automatitzada a la visualització pública.

Si bé s'havia considerat alguns d'aquests últims punts durant la fase de planificació del treball com a riscos, i intentar utilitza *Tableau* per solucionar-los, finalment no s'ha considerat aquesta opció, a causa dels problemes amb les connexions directes amb les dades, per tant, es deixa aquesta millora per a noves versions on es pugui connectar directament al sistema distribuït de fitxers i també on la visualització es pugui fer pública i s'actualitzi automàticament.

4. Resultats

Amb relació a l'objectiu principal, pronòstic de la posició política dels resultats electorals, utilitzant notícies de diferents mitjans de comunicació de cada país produïdes en temps real, les quals són ingerides per un sistema d'ingesta massiva de dades, amb una finestra de 24 hores, dades que flueixen per ser netejades, transformades i de les quals s'extrau característiques a través de varis *pipelines,* donant com a resultat la predicció de la posició política cada dia, es pot dir que, durant l'entrenament els resultats de l'avaluació han estat bons, ja que amb el conjunt de test de característiques diàries obtenim un F1-Score de 69%, i per a períodes electorals el recompte de classes predites dona un F1-Score del 93%.

Recapitulant, el conjunt de mostres de les característiques diàries, esdevenen de l'agrupació diària de les notícies de cada país, processades i transformades per ser ingerides per al model predictiu, afegint-hi prèviament la classe política com a *label* d'aquell dia concret.

Però el que realment interessa, no és predir les notícies de cada dia, sinó l'evolució d'aquestes. Per aquest motiu s'ha afegit com a característica extra el pes, el qual ve definit per la proximitat al dia de les eleccions, donant més rellevància a les notícies més pròximes al dia de les eleccions i menys pes com més temps fa que es van publicar, simulant així l'oblit en el temps. Per aquest motiu es realitza el recompte de classes per país i període electoral, i tal com s'ha vist amb el conjunt de test els resultats superen en 20% les prediccions diàries.

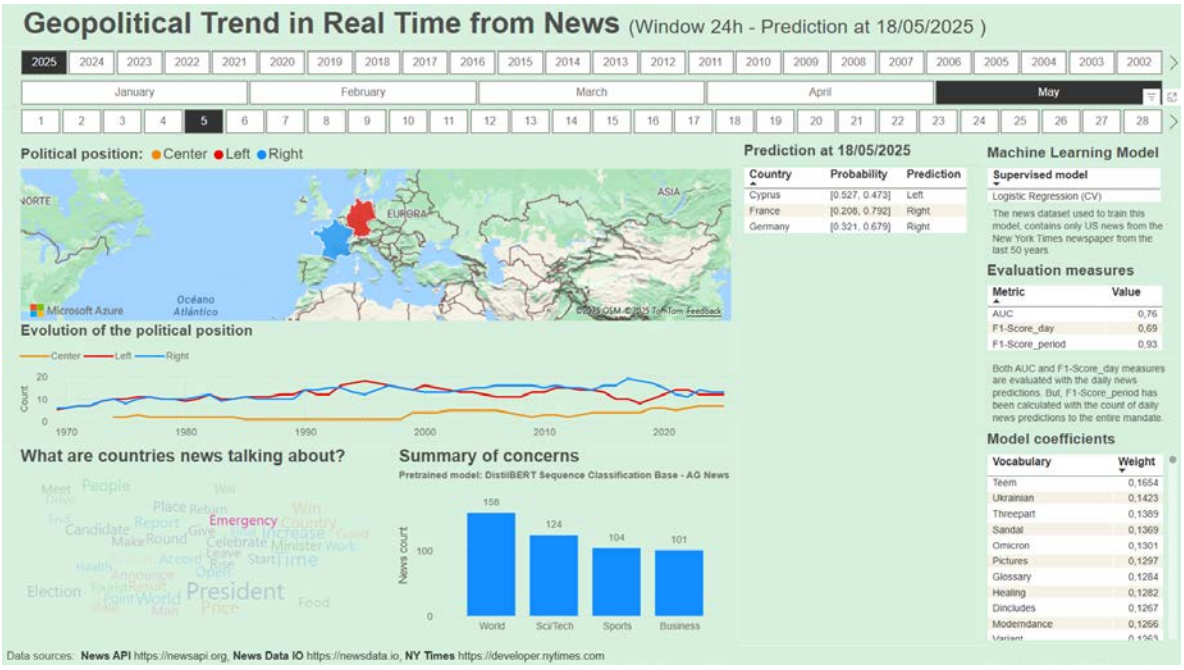
Per tant, si bé són uns resultats excel·lents, s'ha de dir que ho són per al país dels Estats Units, ja que la mostra de notícies amb la que s'ha dut l'entrenament prové únicament d'aquest país.

En aquest sentit, tenim que per a la predicció de la posició política de l'evolució de notícies de la resta de països de la Unió Europea més el Regne Unit, no es tenen les notícies des de les últimes eleccions de cada país, per tant, es contemplen únicament des que es van començar a ingerir al sistema, a mitjans de març, fins al dia actual per a fer la predicció, on també si hi incorpora la característica temporal de pes, i que addicionalment pateixen una primera transformació de traducció a l'idioma anglès des dels diferents idiomes de cada país. On aquests fets poden afeblir la bondat de les prediccions de l'avaluació del model.

Pel que fa a l'objectiu secundari, derivat de la dificultat d'obtenir jocs de dades de diferents idiomes, agrupació dels països segons preocupacions, a tall d'utilitzar un model o altre segons similituds de preocupacions, finalment no s'ha desenvolupat, ja que no s'han obtingut dades antigues d'altres països. En tot cas, sí que es pot fer un seguiment de les preocupacions o termes més freqüents dels diferents països interactuant amb la visualització final, i on també es mostra el vocabulari del model amb els seus coeficients, definint així les paraules més rellevants per a les posicions polítiques. Tot i que en primera instància no es pot filtrar per una paraula en concret pròpiament dit, sí que es pot seleccionar del núvol de paraules i observar en aquell dia quins països la nomenaven a les notícies. Tal com es

mostra a continuació al seleccionar *Emergency*, únicament queden remarcats els països de *Cyprus*, *France* i *Germany*:

Figura 10: Word Cloud selecció de paraula filtra països



Font: Elaboració pròpia

Si es vol donar resposta a les divergències històriques de la línia temporal de l'evolució de la posició política, s'ha de tenir present que únicament es tenen notícies antigues dels Estats Units. En tot cas els punts on convergeixen i s'intercanvien les línies serien interessants d'estudiar.

En referència a l'evolució del mapa coroplètic amb el pas dels anys, tot i que s'ha de fer manual clicant a cada any, es pot observar visualment l'intercanvi de posicions polítiques, i en aquest sentit, si es tinguessin notícies antigues dels últims 50 anys de tots els països intervinents, es podria investigar des del *Word Cloud* els termes d'aquell període, esbrinant així les preocupacions d'aquell moment especial.

Tot i això, tal com s'ha comentat a la secció de limitacions, seria interessant per fer un seguiment de les preocupacions més exhaustivament, afegint en noves versions del treball bigrames, trigrames, i també afegir un filtre avançat per a la cerca de preocupacions, on es detalli com a resultat any i països per tal de millorar el seguiment.

Per tant, amb tot el descrit s'obté un sistema adaptable i configurable, del qual resulta un producte de visualització de dades amb potencial d'exploració, el que introdueix credibilitat i robustesa als resultats, mitjançant la secció més tècnica que es mostra a la dreta de la visualització.

5. Conclusions i treballs futurs

En el primer apartat d'aquest capítol s'inclou una descripció de les conclusions del treball, així com una reflexió crítica sobre l'assoliment dels objectius establerts a l'inici. També es raona sobre el compliment de la definició de la planificació i metodologia establerta. A continuació se sospesen els impactes previstos i no previstos amb relació als a les diferents dimensions de sostenibilitat, ètic-social i de diversitat. En el segon apartat es descriuen noves línies de treball futures i millores.

5.1 Conclusions

Es conclou que els resultats obtinguts, han estat sorprenents, pel fet que l'avaluació del model de predicció mostra un AUC del 76% i un F1-Score del 69% per a mostres diàries, i per a períodes complets dels mandats el F1-Score ha estat del 93% i, per tant, fa pensar que el component d'oblit augmenta l'avaluació del model en els períodes complets dels mandats, donant més importància a les qüestions de les notícies més pròximes a les dates de les eleccions. Tanmateix, aquest model ha estat únicament entrenat amb mostres dels Estats Units i d'una única font de dades, el *New York Times*, per tant, aquest fet pot reduir la bondat de les prediccions dels països Europeus, incloent-hi l'inconvenient de les traduccions automàtiques de les notícies de països no angloparlants, ja que poden ser traduïdes de forma que no reflecteixin correctament el seu significat, fent que baixi la precisió de les prediccions encara més.

Sobre l'objectiu principal, pronòstic de la posició política amb temps real de les eleccions presidencials segons les notícies emeses per mitjans de comunicació, s'ha assolit amb èxit. Tot i que com a objectiu secundari es pretenia crear un model de classificació dels països amb clústers, segons la similitud de les preocupacions dels països, per aplicar-hi un model predictiu o un altre, similar en preocupacions al país del clúster, no s'ha fet, ja que únicament s'ha aconseguit dades antigues d'un únic país.

Amb relació a la visualització perseguida, sí que mostra els pronòstics amb probabilitats, a més de l'històric de posicions polítiques i també les preocupacions, o més ben dit paraules freqüents de les notícies emeses. Permeten interaccions com selecció de dates, filtres per tema que mostren els països als quals aplica, a més mostra termes rellevants amb els seus coeficients provinents de la funció del model de regressió logística, i detalls tècnics que li donen credibilitat a la predicció. Tot i que té la dificultat per fer el seguiment de preocupacions de bigrames o trigrames, ja que en aquesta versió no s'han considerat.

A propòsit de la planificació, les tasques relacionades amb la preparació de l'entorn s'han allargat més del que s'havia previst a causa de la falta de coneixements, en concret la connexió de *Kafka* amb *PySpark* i les proves d'accés de les dades amb *Hive*, que al final s'ha fet directament amb HDFS i fitxers locals. També s'ha requerit més temps del previst per a la cerca de les dades inicials, en primer lloc, per a l'extracció de dades antigues per crear el model de predicció de diferents països, on únicament se n'han trobat per als

Estats Units, en segon lloc, també s'ha trobat dificultat per extraure els presidents i les seves posicions polítiques de tots els països intervinents, feina que s'ha fet de forma completament manual. Per tant, per a complir amb els *timings* s'ha hagut de reduir el temps d'alguna de les següents tasques, encara que, el dia establert de l'entrega s'ha fet el lliurament previst.

Sobre la metodologia àgil seguida, ha estat l'adequada tenint present que s'havien suposat diversos riscos, i les possibles adaptacions. Així i tot, s'han introduït canvis per garantir l'èxit del treball, derivats en gran mesura per les limitacions de recursos i la dificultat d'obtenir prou dades antigues d'altres països de forma gratuïta. En concret no s'ha pogut utilitzar directament *Hive* per accedir a les dades, i tampoc s'ha pogut crear el model d'agrupació de països segons preocupacions, ja que no es tenia prou dades antigues de la resta de països, fet que ha obligat a fer servir l'únic model de predicció de posició política, entrenat amb notícies dels Estats Units provinents del *New York Times*, per a la predicció de la resta de països.

Amb relació als impactes del punt 1.4, per a la dimensió de sostenibilitat ja es va aclarir que depenia majoritàriament de l'ús que fes l'usuari dels resultats, seria positiu si l'ús millora els ODS de la dimensió, i negatiu si prioritza motius personals en detriment del benefici global. En relació amb la impremta ecològica del projecte amb aquest prototip, és mínima, no obstant això, a escala global augmentaria, tal com es preveia tot i utilitzar eficientment les eines disponibles de *Big Data* i *Green Data Centers*, entre d'altres.

Pel que fa a la dimensió de ètic-social, si bé el producte final no s'ha pogut publicar i fer funcionar en temps real per a que tothom i pugui accedir, aquest es considera un prototip amb potencial per millorar i facilitar-hi l'accés en noves versions, per tant, a priori l'impacte positiu no s'ha pogut implementar, però és factible, i estaria llest amb una inversió inicial. Ara bé, el que faltaria es promocionar-lo per donar-li visibilitat, tot i que encara avui dia hi ha països en els quals és molt complicat accedir a internet i a més sense restriccions, qüestió bastant més complicada de solucionar.

Respecte a la dimensió de diversitat, dels impactes previstos sí que s'han aconseguit mitigar els negatius, eliminant en gran mesura les entitats nombrades com noms propis i d'organitzacions, i fent la lematització de les paraules. Tot i que no serà suficient i s'haurà de revisar, ja que es preveu que hi podran haver impactes no previstos, els quals poden aparèixer en futures versions del projecte si s'utilitzen bigrames i trigrames. On determinar per exemple el significat de "*bad girl*" serà més complicat. Malgrat això, s'ha de considerar que les fonts de dades parteixen de mitjans de comunicació oficials i no de comentaris anònims de xarxes socials, per consegüent, es creu que seran els mínims casos, tot i que caldria fer-hi una pensada per intentar minimitzar possibles casuístiques, com per exemple avaluar durant algun temps la professionalitat dels mitjans de comunicació que proveeixen les notícies, o bé investigar si existeix algun certificat de reputació de mitjans de comunicació i intentar utilitzar únicament aquests com a fonts de dades.

5.2 Línies de treball futures i millores

A més de les millores pendents derivades de les limitacions trobades descrites a l'apartat 3.7, a continuació es presenten noves línies de treballs i possibles millores:

- Seria recomanable incorporar a la visualització l'opció de: en seleccionar un país mostrar la seva línia temporal indicant els canvis de posició política i en seleccionar la intersecció mostrar les paraules més freqüents d'aquell darrer any, i també les paraules que tenen un pes més important del vocabulari del model de coeficients més alts i baixos, on aquests coeficients poden donar color gradual de vermell a blau en aquest segon núvol de paraules, mostrant així dos *Word Cloud*, un amb les paraules més freqüents, i en un altre les més rellevants per al model, d'aquell país, d'aquell dia..., en les que és possible que les paraules mostrades no siguin coincidents, ja que un mostra freqüència i l'altre rellevància.

Figura 11: Word Cloud comparativa freqüència/rellevància

Font: Elaboració pròpia

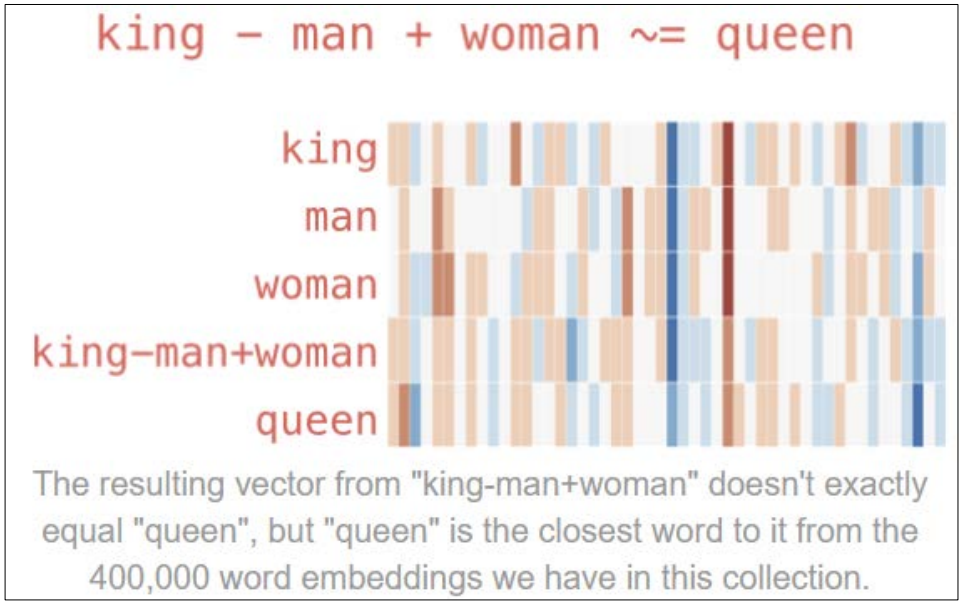
- Per tal de no fer ús de notícies monopolitzades d'un únic mitjà de comunicació, seria interessant poder controlar-ho, mostrant en un diagrama de pastís el percentatge de les notícies extretes de cada mitjà de comunicació de cada país. Així mateix, en haver-hi més pluralitat de mitjans de comunicació, els resultats de la predicció serien més representatius, tot i que es corre el risc d'obtenir notícies similars o repetides procedents de diversos mitjans, no obstant això, seria del que s'estaria parlant, per tant, també arribaria a més públic.

- Incorporar l'autoavaluació del model en temps real, es a dir, a mesura que s'ingesten dades de nous resultats electorals dels diferents països, que el mateix sistema avaluï el model amb les prediccions generades per país i es mostrin els resultats de l'avaluació a la visualització. D'aquesta forma si en un país els resultats de la predicció són incorrectes s'hauria de reflexionar sobre el perquè per aquest país el model no funciona correctament.

- Utilitzar Word2vec, representació vectorial de paraules, per l'agrupació de països segons similituds de preocupacions, com a millora i nova via d'investigació, ja que la vectorització de les paraules inclou la definició numèrica com es pot observar a la

figura 12, a diferència del que s'ha fet al treball d'aquest document on el vocabulari s'ha indexat.

Figura 12:The Illustrated Word2vec



Font: Jay Alammar https://jalammar.github.io/illustrated-word2vec/

- També seria interessant estudiar el vessant de l'evolució de les prediccions diàries per observar la tendència dia a dia, i per això s'haurien d'anar guardant les prediccions diàries i no solament la del dia actual. En concret observar el progrés de probabilitats de cada classe, i mostrar-les a la visualització com gràfics de línies des de les últimes eleccions acompanyades de les notícies antigues de cada país, així també s'observaria si en un moment en concret hi ha un canvi significatiu, es podria estudiar les notícies dels dies anteriors.

- Donat que en la visualització el desplaçament del filtre temporal és manual, és a dir, no mostra automàticament l'evolució de la posició política del mapa coroplètic per exemple fent clic en un botó de *play*, aquesta millora en noves versions ajudaria a visualitzar algun tipus de patró de correlació entre preocupacions, derivat dels canvis en la posició política, iniciats per un país i on els països limítrofs s'hi van adscrivint a mesura que es celebren les eleccions, esbrinant així els països que marquen tendència i també el perquè estudiant les preocupacions.

6. Glossari

Apache Hadoop

És un *framework* de codi obert, que ajuda a processar i emmagatzemar grans quantitats de dades.

Apache Hive

És un sistema d'emmagatzematge de dades per a *Apache Hadoop*. *Hive* fa possible el resum de les dades, les consultes i l'anàlisi de dades. Les consultes de *Hive* s'escriuen a *HiveQL*, llenguatge de consulta similar a SQL.

Apache Kafka

És una plataforma distribuïda per a la transmissió de dades que permet no només publicar, emmagatzemar i processar fluxos d'esdeveniments de forma immediata, sinó també subscriure-s'hi. Està dissenyada per administrar els fluxos de dades de diverses fonts i enviar-los a diferents usuaris.

Apache Spark

Apache Spark és un motor de processament de dades a gran escala que integra mòduls per a SQL, *streaming*, aprenentatge automàtic i processament de grafs.

API

Prové de l'anglès *Application Programming Interface*. Interfície que permet accedir a un conjunt de biblioteques estàndard que faciliten al programador el desenvolupament d'aplicacions.

AUC

Prové de l'anglès *area under the curve*, i és una mètrica resum que quantifica el rendiment global del model, el seu valor està comprès entre 0 i 1.

BDA in Streaming

Prové de l'anglès *Big Data Analysis in Streaming* (Anàlisis de dades massives en temps real).

Bigrames

Els bigrames són dues paraules que contenen un significat diferent quan s'utilitzen juntes.

Broker

A Kafka, els brokers són els servidors que emmagatzemen dades i gestionen totes les sol·licituds de transmissió de dades.

Cluster

Els clústers de Kafka són un grup de brokers de Kafka interconnectats que treballen conjuntament per gestionar els fluxos de dades que entren i surten d'un sistema Kafka.

Consumidor

Els consumidors són aplicacions client que llegeixen missatges d'esdeveniments dels temes del clúster Kafka.

F1-Score

És una mètrica d'avaluació d'aprenentatge automàtic que combina mesures de precisió i de record.

HDFS

Prové de l'anglès *Hadoop Distributed File System*. Sistema distribuït de fitxers de *Hadoop*.

ML

Prové de l'anglès *Machine learning*. Aprenentatge automàtic. És una branca de la intel·ligència artificial (IA) centrada a entrenar ordinadors i màquines per imitar la manera com aprenen els humans, realitzar tasques de forma autònoma i millorar el seu rendiment i precisió a través de l'experiència i l'exposició a més dades.

NLP

Prové de l'anglès *Natural language processing*. Processament de llenguatge natural. És un subcamp de la informàtica i la intel·ligència artificial (IA) que utilitza el *ML* per permetre que els ordinadors entenguin i es comuniquin amb el llenguatge humà

ODS

Objectius de desenvolupament sostenibles.

Outliers

Valor atípic. És una observació anormal i extrema distant de la resta.

Pan

Moviment horitzontal o vertical d'una imatge.

Pipeline

Canalització amb seqüència d'etapes o processos connectats on la sortida d'una etapa és l'entrada de la següent.

Productor

Els productors són aplicacions client que escriuen missatges d'esdeveniments als temes del clúster Kafka.

PySpark

API de *Python* per a *Apache Spark*, un motor de computació distribuïda molt potent dissenyat per processar grans volums de dades de forma ràpida i escalable.

Spark ML

És la llibreria de *Machine Learning (ML)* de *Apache Spark*, dissenyada per treballar de manera eficient amb grans volums de dades en entorns distribuïts.

Tòpic

És una categoria o tema on les seqüències de dades són publicades.

Trigrames

Els trigrames són tres paraules que contenen un significat diferent quan s'utilitzen juntes.

Word2vec

Tècnica que representa numèricament el significat semàntic d'una paraula en vectors.

WSL

Prové de l'anglès *Windows Subsystem for Linux*. És una característica de Windows que permet executar un entorn Linux a la màquina Windows, sense necessitat d'una màquina virtual independent o arrencada dual.

Zoom

Ampliar o apropar-se i reduir o allunyar-se d'una imatge.

7. Bibliografia

[1] Alzamora Bisbal, J. [Jaume]. (2008): *Espigolant dins l'antigor. Refranys i dites de la nostra terra*. Editorial Moll. https://pccd.dites.cat/obra/Alzamora_Bisbal%2C_Jaume_%282008%29%3A_Espigolant_dins_l%27antigor._Refranys_i_dites_de_la_nostra_terra

[2] Naciones Unidas. Objetivos de desarrollo sostenible. https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/

[3] López Córdoba, D. [David], Cazorla Martín, A. [Ángel], Martín-Lagos, A. [Ángel]. (2024). Medición psicofisiológica de las emociones políticas. Un análisis de sus antecedentes y propuesta metodológica. *RIPS: Revista De Investigaciones Políticas Y Sociológicas,* 23(1). https://doi.org/10.15304/rips.23.1.9796

[4] SUZUKI, K. [Kei], LAOHAKANGVALVIT, T. [Tipporn], MATSUBARA, R. [Ryota], SUGAYA, M. [Midori]. (2021). Constructing an Emotion Estimation Model Based on EEG/HRV Indexes Using Feature Extraction and Feature Selection Algorithms. *Sensors* 21(9), 2910. https://doi.org/10.3390/S21092910

[5] Donnini, Z. [Zachary], Louit, S. [Sydney], Wilcox, S. [Shelby], Ram, M. [Mukul], McCaul, P. [Patrick], Frank, A. [Arianwyn], Rigby, M. [Matt], Gowins, M. [Max], Tranter, S. [Scott]. (2024). Election Night Forecasting With DDHQ: A Real-Time Predictive Framework. *Harvard Data Science Review,* 6(4). https://doi.org/10.1162/99608f92.ccb395f0

[6] Denicia-Carral, M. C. [María Claudia], Ballinas-Hernández, A. L. [Ana Luisa], Minquiz-Xolo, G. M. [Gustavo Manuel], Medina-Cruz, H. [Héctor]. (2025). Análisis de sentimientos en la red social X para la evaluación del posicionamiento de candidatos en elecciones políticas. *Revista Científica de Sistemas e Informática*, 5(1), e763. https://doi.org/10.51252/rcsi.v5i1.763

[7] Topîrceanu, A. [Alexandru]. (2025). Macro-Scale Temporal Attenuation for Electoral Forecasting: A Retrospective Study on Recent Elections. *Mathematics*. 2025; 13(4): 604. https://doi.org/10.3390/math13040604

[8] Alaminos-Fernández, A. F. [Antonio Francisco], Alaminos, A. [Antonio]. (2023). *Métodos y Modelos para la Predicción Electoral: Una Guía Práctica*. Alicante: Limencop. ISBN 978-84-09-50283-7, 145 p. https://rua.ua.es/dspace/handle/10045/138240

[9] Web oficial de la Unión Europea. (2025). *Protección de Datos conforme al reglamento RGPD - Reglamento general de protección de datos*

https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_es.htm

[10] Institut Cartogràfic i Geològic de Catalunya. (2013). Mapa de coropletes. https://www.icgc.cat/ca/Publicacions/Diccionaris/Mapa-de-coropletes

[11] Shapira, G. [Gwen], Palino, T. [Todd], Sivaram, R. [Rajini], Petty, K. [Krit]. (2024). Capítulo 4. Consumidores de Kafka: Lectura de datos de Kafka. *Kafka: La Guía Definitiva, 2ª Edición*. . O'Reilly Media, Inc. https://learning.oreilly.com/library/view/kafka-la-guia/9781098181673/

[12] DistilBERT Sequence Classification Base - AG News (distilbert_base_sequence_classifier_ag_news) https://sparknlp.org/2021/11/21/distilbert_base_sequence_classifier_ag_news_en.html

[13] Thomas, A. [Alex]. (2020). 5. Processing Words Natural Language Processing with *Spark NLP*. O'Reilly Media, Inc. https://learning.oreilly.com/library/view/natural-language-processing/9781492047759/

[14] English berttest BertForTokenClassification from RtwC https://sparknlp.org/2025/01/29/berttest_en.html

[15] Nokeri, T. C. [Tshepo Chris]. (2021). 5. Nonlinear Modeling With Scikit-Learn, PySpark, and H2O. *Data Science Solutions with Python: Fast and Scalable Models Using Keras, PySpark MLlib, H2O, XGBoost, and Scikit-Learn.* Apress. https://learning.oreilly.com/library/view/data-science-solutions/9781484277621/html/Cover.xhtml

[16] Tandon, A. [Akash], Ryza, S. [Sandy], Laserson, U. [Uri], Owen, S. [Sean], Wills, J. [Josh]. (2024). Capítulo 6. Comprender Wikipediacon LDA y Spark NLP. *Analítica avanzada con PySpark*. O'Reilly Media, Inc. https://learning.oreilly.com/library/view/analitica-avanzada-con/9781098196844/

[17] Kolokolov, A. [Alex], Zelensky, M. [Maxim]. (2024). *Data Visualization with Microsoft Power BI.* O'Reilly Media, Inc. https://learning.oreilly.com/library/view/data-visualization-with/9781098152772/

8. Annexos

Annex 1 - Manual de configuració de l'entorn

Pas 1 - Instal.lar Ubuntu a Windows

```
# Instal·lar Ubuntu en Windows amb WSL obrir PowerShell i executar:
wsl --install
# Tancar PowerShell i obrir Ubuntu des del menú Inici de Windows.

# Desactivar ipv6 per conflicte en Ubuntu.
sudo sysctl -w net.ipv6.conf.all.disable_ipv6=1
sudo sysctl -w net.ipv6.conf.default.disable_ipv6=1
```

Pas 2 - Instal·lar Java versió 11

```
# Instal·lar Java versió 11.
wget -O - https://apt.corretto.aws/corretto.key | sudo apt-key add -
sudo add-apt-repository 'deb https://apt.corretto.aws stable main'
sudo apt-get update; sudo apt-get install -y java-11-amazon-corretto-jdk

# Si dona problemes de DNS editar fitxer resolv.conf, i tornar a intentar
instal·lar Java, modificant les línies següents perquè resolgui i perquè no
reescrigui el fitxer, i descomentar-les:
sudo nano /etc/resolv.conf
[network]
generateResolvConf = false
nameserver 8.8.8.8

# Comprovar versió de Java.
java --version

# Afegir a .bashrc variables d'entorn de Java (ja instal·lat prèviament).
echo 'export JAVA_HOME=/usr/lib/jvm/java-11-amazon-corretto' >> ~/.bashrc
echo 'export PATH=$JAVA_HOME/bin:$PATH' >> ~/.bashrc

# Aplicar els canvis.
source ~/.bashrc
```

Pas 3 - Instal·lar Kafka

```
# Instal·lar Kafka (descarregar tgz).
wget https://dlcdn.apache.org/kafka/3.9.0/kafka_2.13-3.9.0.tgz
```

```
# Extreure contingut i moure'l al directori principal.
tar -xvzf kafka_2.13-3.9.0.tgz
mv kafka_2.13-3.9.0 ~

# Editar fitxer de configuració .bashrc afegint PATH per no posar la ruta sencera
de les instruccions de Kafka o executar-les des de la carpeta bin.
nano .bashrc
PATH="$PATH:~/kafka_2.13-3.9.0/bin"

# Aplicar els canvis.
source ~/.bashrc

# Validar modificació path.
cat .bashrc

############### Iniciar clúster de Kafka ###############
# Crear nou Cluster Kafka amb ID aleatori.
kafka-storage.sh random-uuid

# Configurar directori Logs (amb l'ID aleatori retornat al pas anterior).
kafka-storage.sh format -t ID -c ~/kafka_2.13-3.9.0/config/kraft/server.properties

# Inicialitzar Kafka en deamon mode
kafka-server-start.sh ~/kafka_2.13-3.9.0/config/kraft/server.properties &

# Per parar el servei de Kafka
kafka-server-stop.sh

############### Fini inici Kafka ###############
```

Pas 4 – Instal.lar Visual Stucio Code a Ubuntu

```
# Per parar el servei de Kafka
code .
```

Pas 5 – (Opcional) Executar Jupyter Notebook a VSC

```
# En cas de voler executar  jupyter notebook des de Visual Code en entorn (Ubuntu).
sudo apt update
sudo apt install python3-ipykernel
```

Pas 6 – Instal.lar Python i Spark

```
# Configuració Python + Spark.
sudo apt install python3-pip
pip install findspark --break-system-packages
```

```
# Instal·lar Spark.
wget https://dlcdn.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz
tar -xvzf spark-3.5.5-bin-hadoop3.tgz
mv spark-3.5.5-bin-hadoop3 spark

# Afegir variables entorn de Spark i  Python al fitxer de configuració .bashrc .
echo 'export SPARK_HOME=$HOME/spark' >> ~/.bashrc
echo 'export PATH=$SPARK_HOME/bin:$PATH' >> ~/.bashrc
echo 'export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH' >> ~/.bashrc
echo 'export PYSPARK_PYTHON=python3' >> ~/.bashrc
source ~/.bashrc

sudo apt install python3-pandas
sudo apt install python3-kafka
```

Pas 7 – Crear un entorn virtual i installar llibreries

```
# Crear entorn virtual, activar-lo i afegir llibreries.
sudo apt install python3-venv
python3 -m venv myenv
source myenv/bin/activate
pip3 install kafka
pip3 install --upgrade --force-reinstall kafka-python six
pip3 install requests
pip3 install setuptools # error Pandas no module named 'distutils'
pip3 install pandas
pip3 install spark-nlp==5.5.3
pip3 install scikit-learn

# https://sparkbyexamples.com/pyspark/pyspark-importerror-no-module-named-py4j-java-gateway-error/
# https://support.datastax.com/s/article/Spark-hostname-resolving-to-loopback-address-warning-in-spark-worker-logs
```

Pas 8 – Instal.lar HIVE i HADOOP (HDFS)

```
# Instal·lar primer Hadoop (HDFS).
wget https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
tar -xzvf hadoop-3.4.1.tar.gz
sudo mv hadoop-3.4.1 /usr/local/hadoop

# Afegir a .bashrc variables d'entorn Hadoop.
echo 'export HADOOP_HOME=/usr/local/hadoop' >> ~/.bashrc
echo 'export HADOOP_INSTALL=$HADOOP_HOME' >> ~/.bashrc
echo 'export HADOOP_MAPRED_HOME=$HADOOP_HOME' >> ~/.bashrc
echo 'export HADOOP_COMMON_HOME=$HADOOP_HOME' >> ~/.bashrc
echo 'export HADOOP_HDFS_HOME=$HADOOP_HOME' >> ~/.bashrc
```

```
echo 'export HADOOP_YARN_HOME=$HADOOP_HOME' >> ~/.bashrc
echo 'export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native' >> ~/.bashrc
echo 'export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin' >> ~/.bashrc
echo 'export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"' >>
~/.bashrc
echo 'export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop' >> ~/.bashrc

# Aplicar els canvis.
source ~/.bashrc

# Confirmem la instal·lació de Hadoop mostrant la versió
hadoop version

# Validem HDFS llistant el directori
hdfs dfs -ls /

# Descarreguem Hive i instal·lem
wget https://downloads.apache.org/hive/hive-4.0.1/apache-hive-4.0.1-bin.tar.gz
tar -xzvf apache-hive-4.0.1-bin.tar.gz
sudo mv apache-hive-4.0.1-bin /usr/local/hive

# Afegir a .bashrc variables d'entorn HIVE.
echo 'export HIVE_HOME=/usr/local/hive' >> ~/.bashrc
echo 'export PATH=$HIVE_HOME/bin:$PATH' >> ~/.bashrc
# Aplicar els canvis.
source ~/.bashrc

# Confirmem la instal·lació de Hive mostrant la versió.
hive --version

# Descomentar i modificar 'HADOOP_HEAPSIZE' de 1024 a 2048.
cp /usr/local/hive/conf/hive-env.sh.template /usr/local/hive/conf/hive-env.sh
nano /usr/local/hive/conf/hive-env.sh
export HADOOP_HEAPSIZE=2048
source /usr/local/hive/conf/hive-env.sh

# Instal·lar servei SSH i crear Keygen i assignar a usuari.
sudo apt install openssh-server –y

# Validar servei instal·lat.
dpkg -l | grep ssh

# Verificar port 22 de SSH.
grep Port /etc/ssh/sshd_config

# Crear psw ssh i assignar als usuaris.
ssh-keygen -t rsa -b 4096
```

```
# Copiar key a public place (demana psw ubuntu).
ssh-copy-id -i ~/.ssh/id_rsa.pub roser@Roser-Dell

# Validar que s'ha copiat la key a authorized_keys.
cat ~/.ssh/authorized_keys

# Donar permisos
cd ~/.ssh/
chmod 600 id_rsa
chmod 600 authorized_keys
chmod 700 ~/.ssh

# Verificar configuració SSH per usar public key.
sudo nano /etc/ssh/sshd_config
# descomentar:
PubkeyAuthentication yes
PasswordAuthentication yes
AuthorizedKeysFile .ssh/authorized_keys

# Reiniciar servei SSH
sudo systemctl restart sshd

## INICI Configuració de Hadoop, es consulta:
# https://medium.com/@madihaiqbal606/apache-hadoop-3-3-6-installation-on-ubuntu-22-
04-2-lts-wsl-for-windows-bb57ed599bc6

# Configurar Java environment variables.
# Editar fitxer Hadoop conf.
nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
# Descomentar línia i afegir ruta d'instal·lació de JAVA.
export JAVA_HOME=/usr/lib/jvm/java-11-amazon-corretto

# Obrir core-site.xml.
sudo nano $HADOOP_HOME/etc/hadoop/core-site.xml
# Afegir les següents línies entre <Configuration> </Configuration>.
<configuration>
 <propiedad>
      <nombre>fs.defaultFS</nombre>
      <valor>hdfs://0.0.0.0:9000</valor>
      <descripción> El URI del sistema de archivos predeterminado</descripción>
   </propiedad>
   <property>
      <name>fs.default.name</name>
      <value>hdfs://0.0.0.0:9000</value>
    </property>
</configuration>
```

```
# Crear directori per emmagatzemar les metadades de node:
sudo mkdir -p /home/hadoop/hdfs/{namenode,datanode}

# Canviar el propietari del directori (roser) user:
sudo chown -R roser:roser /home/hadoop/hdfs

# Editar el fitxer hdfs-site.xml:
sudo nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
# Afegir les següents propietats entre <Configuration> </Configuration> i guardar:
<property>
      <name>dfs.replication</name>
      <value>1</value>
   </property>
   <property>
      <name>dfs.name.dir</name>
      <value>file:///home/hadoop/hdfs/namenode</value>
   </property>
   <property>
      <name>dfs.data.dir</name>
      <value>file:///home/hadoop/hdfs/datanode</value>
   <property>
      <name>dfs.webhdfs.enabled</name>
      <value>true</value>
   </property>
</property>

# Editar el fitxer mapred-site.xml:
sudo nano $HADOOP_HOME/etc/hadoop/mapred-site.xml
# Afegir les següents propietats entre <Configuration> </Configuration> i guardar.
<property>
      <name>mapreduce.framework.name</name>
      <value>yarn</value>
</property>

# Editar el fitxer yarn-site.xml:
sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
# Afegir les següents propietats entre <Configuration> </Configuration> i guardar.
<property>
      <name>yarn.nodemanager.aux-services</name>
      <value>mapreduce_shuffle</value>
</property>

# Finalment, validar la configuració de Hadoop el format de HDFS NameNode:
hdfs namenode -format

# Iniciar el Clúster de Hadoop:
```

```
start-dfs.sh

# Iniciar el node mànager i els recursos mànager:
start-yarn.sh

#Verificar que els serveis estan iniciats:
jps
# S'ha de veure així:
"""
2178 NodeManager
1845 SecondaryNameNode
2054 ResourceManager
423519 Jps
811 Kafka
1484 NameNode
1614 DataNode
"""

# Access the Namenode:
http://localhost:9870

# Access the Hadoop Resource Manager:
http://localhost:8088

# Tancar els serveis de Hadoop:
stop-yarn.sh
stop-dfs.sh
```

Pas 9 – Descarregar models preentrenats de SparkNLP

```
# Descarregar models SparkNLP
sudo apt update && sudo apt install unzip -y
# https://sparknlp.org/2025/01/29/berttest_en.html
#copiar a:
cd TFM/models
mkdir -p ~/models_npl/berttest
unzip berttest_en_5.5.1_3.0_1738112596101.zip -d ~/models_npl/berttest

#
https://sparknlp.org/2021/11/21/distilbert_base_sequence_classifier_ag_news_en.html
#copiar a:
cd TFM/models
mkdir -p ~/models_npl/distilbert
unzip distilbert_base_sequence_classifier_ag_news_en_3.3.3_3.0_1637503060617.zip -d
~/models_npl/distilbert
```