|  |  |
| --- | --- |
| **Tendència Geopolítica en Temps Real**    **Subtítol (si n’hi ha)** | |
|  | |
|  |  |
| **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 |
|  |

[Llicència de Creative Commons](http://creativecommons.org/licenses/by-nc-nd/3.0/es/)

Aquesta obra està subjecta a una llicència de [Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya de Creative Commons](http://creativecommons.org/licenses/by-nc-nd/3.0/es/)

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 MUDS |
| **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** | |
| A dia d’avui 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]](#Ref_1_B)  Si s’interpreta com que no es pot recordar tot el que s’aprèn ni la història que precedeix, en futures situacions similars es cauran 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, basant-se amb notícies emeses per varis 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, 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 que rebrem les noticies, per tant, es traduiran a l’anglès durant la ingesta.  Addicionalment, agrupar els països per preocupacions similars dona la 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 te absència escolar i el nostre sí, aquest fet dona lloc a una investigació addicional extra de com es va resoldre.  Si be la idea principal es pronosticar els resultats electorals, la visualització de series temporals de les principals preocupacions dels diferents països, en diferents 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]*](#Ref_1_B)  *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 analysed, 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, 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 9](#_Toc197362601)

[Llista de Taules 10](#_Toc197362602)

[1. Introducció 11](#_Toc197362603)

[1.1 Context i justificació del Treball 11](#_Toc197362604)

[1.2 Explicació de la motivació personal 12](#_Toc197362605)

[1.3 Objectius del Treball 12](#_Toc197362606)

[1.4 Impacte en sostenibilitat, ètic-social i de diversitat 13](#_Toc197362607)

[1.5 Enfocament i mètode seguit 15](#_Toc197362608)

[1.6 Planificació del Treball 17](#_Toc197362609)

[1.7 Breu sumari de productes obtinguts 24](#_Toc197362610)

[1.8 Breu descripció dels altres capítols de la memòria 25](#_Toc197362611)

[2. Estat de l’art 26](#_Toc197362612)

[3. Materials i mètodes 36](#_Toc197362613)

[3.1 Configuració de l’entorn 37](#_Toc197362614)

[3.2 Els conjunts de dades 38](#_Toc197362615)

[3.2.1 Conjunt de dades inicial 38](#_Toc197362616)

[3.2.2 Llistat de presidents i posició política (la etiqueta a predir) 39](#_Toc197362617)

[3.2.3 Conjunt de dades diari i ingesta Kafka 40](#_Toc197362618)

[3.3 Neteja de les dades (Spark NLP Pipeline) 45](#_Toc197362619)

[3.4 Extracció de característiques 48](#_Toc197362620)

[3.5 Creació dels models PySpark ML, Pipeline 51](#_Toc197362621)

[3.6 Visualització *Power BI real time (window 24h)* 56](#_Toc197362622)

[3.6.1 Model de la visualització 57](#_Toc197362623)

[3.6.2 La visualització 61](#_Toc197362624)

[3.7 Limitacions 67](#_Toc197362625)

[4. Resultats 70](#_Toc197362626)

[5. Conclusions i treballs futurs 72](#_Toc197362627)

[6. Glossari 73](#_Toc197362628)

[7. Bibliografia 74](#_Toc197362629)

[8. Annexos 76](#_Toc197362630)

[Annex 1 - Manual de configuració de l’entorn 76](#_Toc197362631)

# Llista de Figures

Figura 1: Arquitectura d’alt nivell, eines i flux principal. 15

Figura 2: Planificació temporal - Diagrama Gantt 20

Figura 3: Diagrama d’activitats 35

Figura 4: API NYT 36

Figura 5: News API 39

Figura 6: NewsDATA.IO API 39

Figura 7: Model de la visualització 56

Figura 8: Visualització Power BI 63

Figura 9: Press Freedom index in Europe in 2025 67

# Llista de Taules

Taula 1: Recursos necessaris 16

Taula 2: Descripció de les tasques 16

Taula 3: *Pipeline* Processament de Llenguatge Natural 44

# 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 es 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 tan a la planificació, com a la qualitat del resultat del model de predicció.

## 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 temen 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.

Que li preocupa a la societat per que prengui la decisió de votar un o altre partit polític? Quins son aquest termes (preocupacions) que fan decantar la balança a partits de dreta o esquerra. En certa mesura el que vol el poble es millorar la seva forma de vida, però si el coneixement i la saviesa únicament la tenen les persones que estan al capdavant; a cas no tenen avantatja sobre el poble? a cas la informació no es poder com be sempre s’ha sentit dir, doncs compartim-la!

En tot cas i a nivell individual, si be s’acostuma a gravar amb foc els successos traumàtics, que passa amb els que no ho son? Que passa amb les persones que no han tingut aquest trauma? Així i tot en ocasions similars poden evitar-ho? I si hi ha transferència de coneixement? Per exemple:

*“De ben segur que donada la desgràcia de les inundacions de València, moltes persones que vulguin comprar una casa demanaran, avanç de prendre la decisió, un informe de riscos d’inundacions. Però, perquè es va construir en terrenys inundables? I perquè es va permetre?“*

No obstant, 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 o 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 pressa de decisions.

## Explicació de la motivació personal

En record a les converses de persones que es preocupen pels seus sers 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 en ocasions es fan pesades, saps que es 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 es 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 següent passos en la evolució d’un país? Es podria millorar la pressa 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 mon?

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

## Objectius del Treball

Com objectiu principal s’estableix que segons les noticies del 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 España).

Com probablement no disposarem de tota la informació dels resultats electorals del diferents països dels que rebrem les notícies, s’intentarà extrapolar els resultats. Per tant, com objectiu secundari es farà una classificació de països mitjançant les similituds dels continguts dels titulars de les noticies, les quals es creu que reflecteixen les preocupacions de la població, i donat que hi intervenen dades de diferents països, per 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, els quals s’hagin pogut utilitzar durant l’entrenament del model.

Tan 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 per a que el pronòstic electoral tingui un bon percentatge d’encert.

## Impacte en sostenibilitat, ètic-social i de diversitat

Donat que l’objectiu principal es esbrinar les preocupacions dels diferents països i basant-se en aquestes pronosticar els resultats electorals, i el secundari es classificar els països segons les similituds en 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’hi inclouran els objectius i metes de desenvolupament sostenible (ODS) per al 2030, tal com es defineixen a les Nacions Unides [[2]](#Ref_2_B).

**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 son del seu gust, en situacions similar podran prendre la decisió de canviar-lo avanç de 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 del 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 extrau la saviesa del projecte.

Pel que fa a l’impacte negatiu, tret del mal us 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 hi que el desenvolupament d’aquest projecte es durà a terme amb una maquina local, tret de la visualització (*dashboard* que podrà ser consultat des de una URL pública), la idea final seria el ús de serveis *cloud* escalables per poder gestionar el màxim volum de noticies de tots els països del mon, i així tenir una visió mundial de la trajectòria de la humanitat.

Així mateix, en aquest context l’empremta ecològica augmentarà, tot hi que les pròpies eines que s’utilitzen per al desenvolupament d’aquest projecte Big Data, mes 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 be cada país evoluciona al seu ritme, aquest ritme es pot accelerar si tenim mostres de com han evolucionat països mes 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.

Però que passa quan el poble també te accés a la mateixa informació de poder evolucionar i millorar el seu futur? De ben segur que les persones dirigents hauran d’acabar fen un canvi a favor del poble, ja que al final el poble els supera en número, i de ven segur 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 be comú. Amb tot això en aquesta dimensió també hi intervenen tots els ODS definits:

* ODS 1 - Sense pobresa
* ODS 2 - Cero 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 us 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 hi que la idea principal del projecte es extreure les preocupacions de les noticies rebudes de diferents mitjans de comunicació de diferents països, aquestes noticies poden incloure desigualtat de gènere, diversitat o incompliment dels drets humans, degut probablement a la cultura en sí de cada país, i a la divulgació de dades personal, com per exemple noms propis, o ideologies subjacents del propi medi de comunicació. A mes la mateixa comparació 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 definit:

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

Per tal de minimitzar aquest impacte negatiu inherent a la pròpia solució, que aporta la mateixa font de dades, durant el procés de desenvolupament i transformació de les dades es procedirà a eliminar els noms propis, referencies a institucions, així com lematització de les paraules. Tot hi que aquestes solucions probablement no serà suficients per exhaurir el problema, s’implementarà una revisió constant per tal de seguir minimitzant aquest impacte establin noves mesures per intentar no ferir la sensibilitat de cap col·lectiu.

Pel que fa als impactes positius, com be s’ha dit que: la informació es poder, dependrà en sí mateix de la utilitat i sentit que li vulgui donar l’usuari final. Tot hi així tenir una visió amplia del que passa a d’altres països, sempre pot enriquir el coneixement i consegüentment millorar en futur.

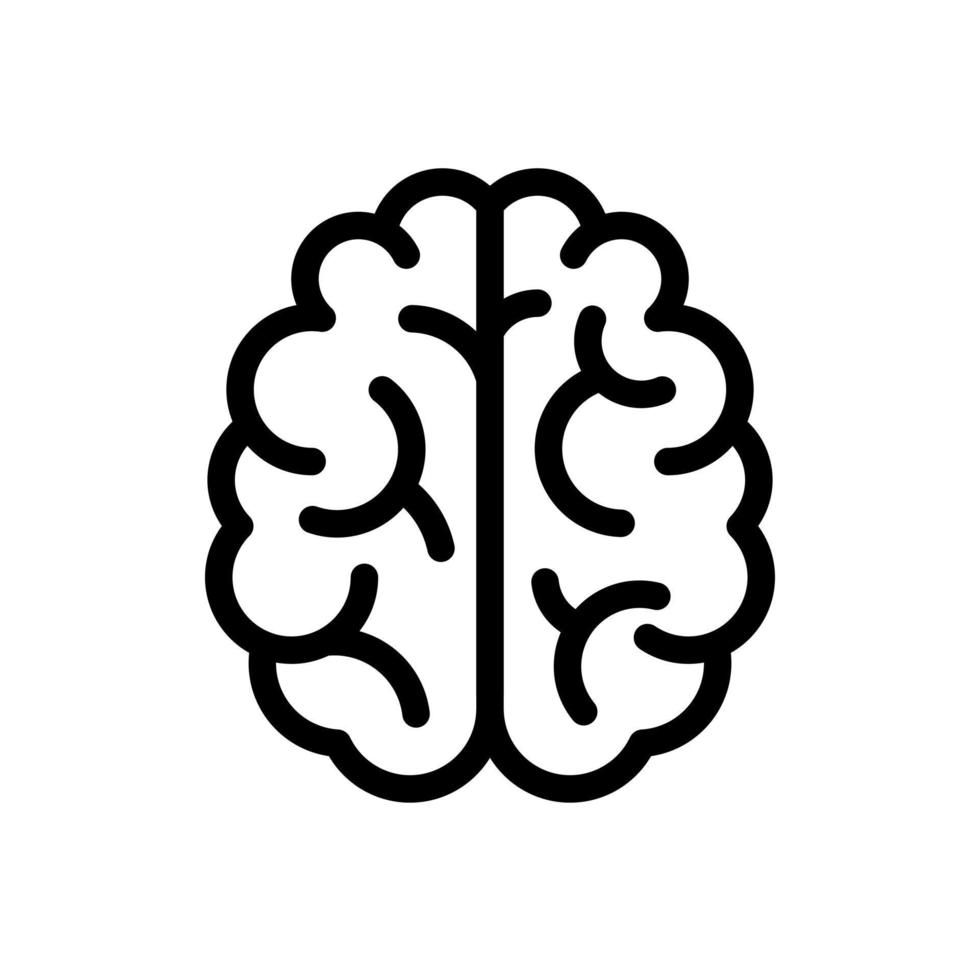
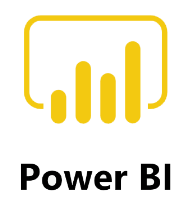
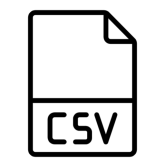
## Enfocament i mètode seguit

Pel que fa l’estratègia per al desenvolupament del projecte, d’entre les diferents 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 es 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 el mes aviat possible, el qual s’anirà millorant fins aconseguir el producte final desitjat.

Tot hi que l’objectiu final del producte es 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 a causa de que 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’utilitzaran i el flux de les dades. S’ha de tenir present que, com el desenvolupament es farà en local, tret de *Power BI*, únicament tindrem un broker de Kafka per tant no podrem tenir rèpliques de la informació.

*Figura 1: Arquitectura d’alt nivell, eines i flux principal.*



Ingesta

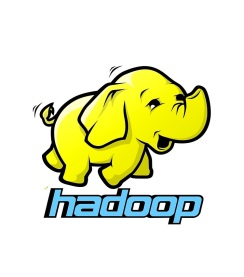
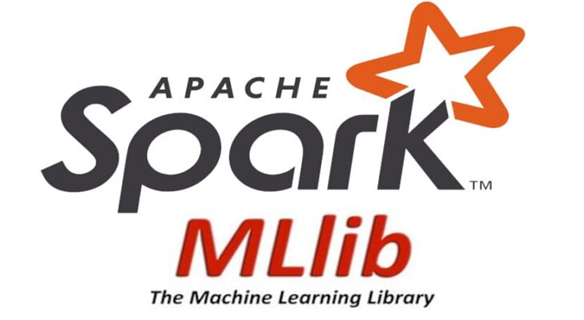
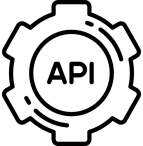
Extracció

Transformació

Emmagatzematge

Visualització

Models



*Font: Elaboració pròpia*

## 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 Estudio 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** | | **Duració** |
| --- | --- | --- | --- | --- |
| **Inici** | **Fi** |
| **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 fitxer amb resultats electorals de varis països dels quals rebrem les noticies des de la 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 noticies de diferents països amb estructura similar a les que rebrem de la API. Per tal de tenir un data set base prou gran com per entrenar els models de pronòstic i agrupació. | 09/03/25 | 10/03/25 | 2 |
| 1.3 | **Descarrega via API cada 24 hores.**  Desenvolupar codi per descarregar diàriament les notícies des de la API. | 11/03/25 | 15/03/25 | 5 |
| 1.4 | **Filtrat 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 noticies 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 preentreenat 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 noticies), expandir paraules, eliminar puntuacions, espais extra i *stopwords*, *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 |
| 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 auto actualització 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 la ampliació incremental de 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. No obstant, 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.

1. Preparació de l’entorn.

**PAC 02** - Extracció i Ingesta, i l’estudi del art.

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

**PAC 03** - Processament, models i visualització.

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

**PAC 4.0** - Redacció de la memòria (Preliminar).

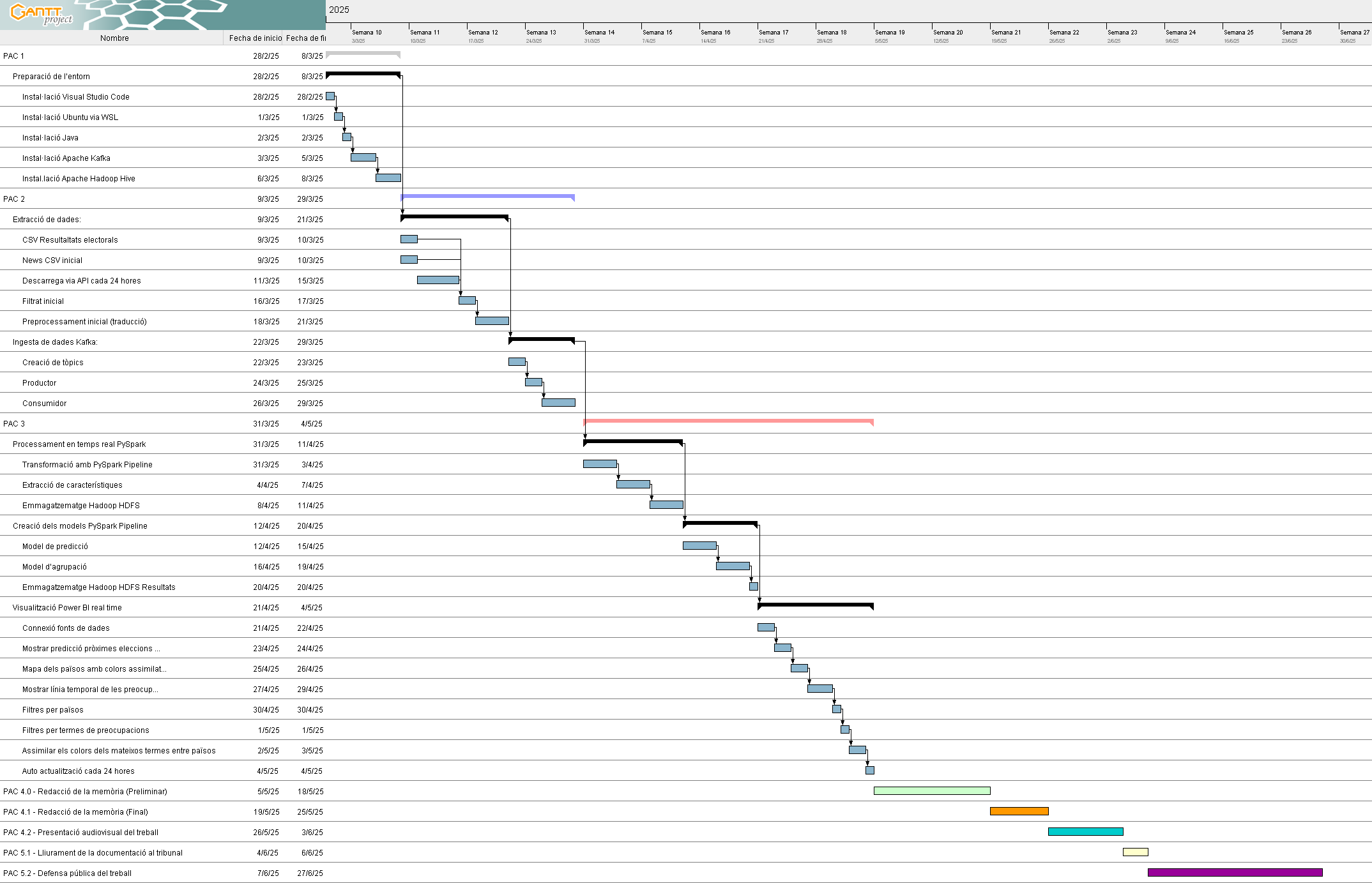
**PAC 4.1** - Redacció de la memòria (Final).

**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 detallem els possibles riscos que es poden presentar, els quals impedeixin seguir amb la planificació establerta, i que poden afectar.

* Incompatibilitats entre eines com per exemple que *Power BI* web no funcioni correctament amb la cerca de dades a la instal·lació de *Apache Hive* local. En tal cas es provaria a 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 es 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 de que l’avaluació no doni resultats prou significatius. En aquest cas es provaria a modificar el conjunt de dades, incloent o excloent mes categories de paraules (preocupacions i d’altre contingut). Així i tot si finalment els resultats no son rellevants, es a dir, el pronòstic es aleatori en aquells països on tenim les dades dels resultats electorals, donaria lloc a pensar en que amb els titulars de les noticies i una descripció no es 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 governant als països limítrofs, no obstant aquest últim repunt quedaria fora de l’abast del projecte actual ja que amb el *timing* actual no podem recuperar tota la historia 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 extraurem les noticies actuals amb perfil gratuït, per tan tenim un límit de descarrega diària, i a mes no sabem si prioritzen en enviar noticies de certs temes o certs mitjans de comunicació, per tant podrien estar esbiaixades. En tot cas aquest inconvenient no endarrereix la planificació però si la qualitat final dels models. En aquest cas la solució seria mes complicada, 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 mes possiblement en moltes webs no es permet i seria il·legal. També podríem fer un *upgrade* del perfil del les API i descarregar mes noticies diàries de cada país, així hi ha mes probabilitat d’ampliar les categories i la procedència de mitjans de comunicació, però la quota mensual queda fora d’abast. Per tant, confiarem en que les API ens retornaran noticies aleatòries.
* Un altre punt complicat pel que fa a l’extracció de preocupacions dels països, es que a les mateixes noticies de cada país no es parli únicament dels esdeveniments del mateix país, sinó també d’altres països. Aquest fet fa pensar que no son preocupacions en si del país que notifica la noticia, si no del que se’n parla. En tot cas es podria descartar les noticies que parlen d’altres països, però per no complicar la recuperació de les característiques a priori es consideraran preocupacions del país que notifica la noticia, 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, es té el risc afegit de que les traduccions automàtiques de les noticies no siguin prou precises, i tot hi que la traducció es fa de la noticia completa avanç d’extreure les característiques, per tal de que les paraules no perdin els significat contextual, una mala traducció pot afectar als resultats del model i no assimilar correctament els països per les preocupacions reals. En tot cas es podria buscar un altre traductor automàtic gratuït i comparar els resultats de les traduccions, o be fer servir un de pago, no obstant aquesta última solució augmentaria considerablement el cost del projecte, i per aquest motiu queda descartada.

## Breu sumari de productes obtinguts

No cal entrar en detall: la descripció detallada es farà a la resta de capítols.

## 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 un breu descripció del seus enfocament d’acord amb al conjunt 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 desenvolupaments del projecte, mostrant en primer lloc a mode introductori un diagrama d'activitat per entrar en context. També es descriu les parts mes rellevants del codi, com l'extracció dels conjunts de dades, la neteja, transformació i l'extracció de característiques amb NLP, el flux de les dades, l’emmagatzematge, com es genera el conjunt de dades per a l'entrenament del model de predicció i l'avaluació, creació i entrenament del model de predicció de la posició política, així com l’aplicació del model de classificació de les noticies. A continuació s’explica el model i la visualització final en *Power BI*. Finalment s’incorpora un detalla de les limitacions trobades durant el desenvolupament i com s’han afrontat.

............................ PENDENT D’ESCRIURE LA RESTA DE CAPÍTOLS .............................

# Estat de l’art

En aquest capítol es realitza una recerca existent de la bibliografia recent amb relacíó 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 noticies 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 varis treballs que aborden la predicció electoral des de diferents perspectives:

* En el treball (López, Cazorla i Martin, 2024)[[3]](#Ref_3_B) 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 perden 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 70, que es comenten a l’article i que expliquen el comportament polític, els quals definirien el vot entre republicà o demòcrata, es bases en:

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

1. La transferència d’afinitats partidistes induïdes pels familiars.
2. La racionalitat, benefici o el prejudici 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 mes important, en el pla explicatiu emocional, es 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 al 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, el que 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 auto referencials**: 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 del organisme i l’estat emocional se 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 demostrar 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 per el cervell i la medul·la espinal, i les emocions es mesuren amb electroencefalografia i les imatges de tomografia, tot hi que aquesta última tècnica es va descartar per la gran complexitat d’interpretar les imatges. Pel que fa a la 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]](#Ref_4_B) desenvolupat al *Instituto de Tecnología de Shibaura*.
3. **El reflexes**: on es mesuren amb electromiogrames la resposta elèctrica dels músculs del coll, esquena i el parpelleig. On es determina que la expressió facial es una forma molt efectiva de precisar les emocions.

Es comenta que com ús potencial es pot aplicar la teoria de la intel·ligència afectiva recolzada pel mesurament de les emocions, combinant mesuraments autor referencials i psicofisiologies, on s’explica que varies 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 auto referencials, dona lloc a la 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 ve la idea de incorporar l’avaluació de les emocions al model de predicció electoral es rellevament i afegeix una altra precisió als resultats de les predicció, s’observa com a limitació la metodologia de detectar les emocions. Doncs seria necessari realitzar qüestionaris i mesuraments psicofisiològiques per determinar les emocions i posterior predicció del vot, fet que inclou temps i equips cars.

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

* Tot hi que no es considera el mesurament de les emocions directes, s’incorporen al model de forma indirecta, mitjançant les preocupacions determinades pels titulars de les noticies emeses. Ja que les emocions son 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 d’aquest treball no requereix d’estudis de pacients i tot recau en la ingesta en temps real de les noticies. Tot hi que, si ve es cert que es requereix un volum substancial de dades, noticies mes 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 si no la evolució del seu vot en vers a les preocupacions, i en comparació es molt mes ràpid.
* En el treball (Donnin et al., 2024)[[5]](#Ref_5_B) 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 el 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, donada les acusacions de frau electoral succeïdes recentment.

La ingesta de dades ve donada per:

1. Durant la nit electoral, amb el recompte complert i parcial en directe els analistes busquen predir els resultats electorals.
2. Les condicions especials de recompte de cada comptat, diferencies entre els patrons de votació i el report d’informes. On quan el comptat ha reportat quasi el recompte complert se hi incorporen:

* Dades geogràfiques: regió d’origen dels candidats on solen tenir millors resultats combinat amb la distancia on resideixen, on la força estimada disminueix a mesura que augmenta la distancia entre el comptat d’origen i la residencia 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. El que permet una anàlisi més completa dels patrons de votació, per exemple els votants tendeixen a votar mes candidats amb similituds demogràfiques, afirmacions que es basen en nombrosos anàlisis realitzat per l’equip del model de Primàries en Viu.

Tot plegat proporciona una predicció mes confiable i precisa dels resultats finals. Tot hi 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 completes 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 total de comptats informats.

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

**Comentaris:**

Si ve es un model molt precís, en el qual intervenen diferents fonts de dades de diferents dimensionalitat (votacions en directe, dades geoespacials i informació demogràfica) el model es per fer servir durant les votacions, i te la limitació de 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 de que la predicció siguin 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 mig 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, es 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 es segueix pel mateix camí, i això es tenir informació d’avant mà, que per tant pot donar la oportunitat de rectificar la trajectòria, tenint en compte que a mes proximitat de les noves 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, ..., taxa 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, es tractar a les persones per igual, tot hi que les mateixes noticies poden inferir aquestes mateixes classificacions entre països, segons les preocupacions que anuncien les noticies, però tal com s’ha comentat amb anterioritat cada país va al seu ritme.

* En el treball (Denicia, Ballinas, Minquiz i Medina, 2025)[[6]](#Ref_6_B) 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 anàlisis 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 ve 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 mes 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*](https://docs.rapidminer.com/9.7/studio/installation/index.html), la qual permet extracció de publicacions establint criteris de cerca mitjançant el connector de *searchTwitter*. Per al processament de text i la anàlisis de sentiment es va utilitzar la extensió de *MeaningCourd* de *RapidMiner*.

Les etapes del procés d’investigació van ser la recol·lecció de missatges de X durant els 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 de un dispositiu mòbil.
4. Tots els *tweets* compartits en un episodi polític important.
5. Tots els tweets publicats en un episodi polític important des de un dispositiu mòbil.

Finalment es conclou que considerar únicament la polaritat de sentiment no es suficient per predir el 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 mes negatiu.

Addicionalment s’indica que durant l’anàlisi textual es van trobar moltes publicacions repetides, el que els dona lloc a pensar que hi va haver un us de bots tendenciós, i esbiaixen la polaritat sobre un tema.

**Comentaris:**

Tot hi que els resultats no van predir amb precisió els resultats electorals, l’anàlisi obra la oportunitat a aprofundir en l’extracció de sentiment combinant-ho amb altre variables. Per tant si ve el procés de predicció es 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 hi que aquest incorporació de classificació de sentiments inclouria afegir els candidats de tots els països que intervenen en el treball, que ara mateix esdevenen el Estats Units d’Amèrica i el països pertanyents a la Unió Europea, per tant, això implicaria que cada període electoral s’hauria d’actualitzar els candidats, i requeriria un manteniment del model final mes complex.

Addicionalment no tots els votants tenen compta a les xarxes socials, i en aquests situació s’estaria excloent a bona part d’ells i els seus sentiment. Per tant s’hauria de fer un estudi per saber la ponderació, en vers a la població que te compte de la xarxa social X respecte del total, amb la que s’haurien d’incorporar els resultats, al treball d’aquest document.

* En el treball (Topîrceanu, 2025)[[7]](#Ref_7_B) 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 la opinió a microescala i les teories de difusió d’epidèmies temporals, per tal de millorar la precisió dels pronòstics electorals fent us 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, fen us d’un conjunt de dades, enquestes públiques confiables, de 10 eleccions celebrades entre el 2020 y el 2024 de tot el mon, 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 son mes 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 recolzament a cada candidat en un sistema de múltiples candidats, mes opinions diàries que emplenaran els dies que no s’hagin realitzat enquestes, es a dir amb vectors de sondejos de 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, es cerca el comportament a macroescala, fent us de les teories de difusió epidèmiques adaptades, per tal d’observar la evolució de la dinàmica de la opinió després de cada injecció d’opinió, en els individus en el context de opinió electoral.

El model es prova i s’avalua amb diferents mètriques, i fa rellevància en que la precisió del pronòstic electorals es 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 ve es un model amb resultats notables de pronòstics electorals, que no fa us de dades demogràfiques ni econòmiques, ni de cap context polític, si que requereix d’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. No obstant la combinació de 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 d’altres models.

Si es compare 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 es centra en el període preelectoral i que la metodologia es diferent, aquest s’assimila a la evolució de les preocupacions que es volen extreure de les noticies 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]](#Ref_8_B) 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 relacíó a les teories i visions de l’estudi de la democràcia, incloent la democràcia liberal, participativa i la deliberativa. Així com el 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 la identitats social, les avaluacions econòmiques o les actituds polítiques.

Pel que fa a les fonts de dades utilitzades per a predir els resultats electores, es fa menció a les enquestes, models estadístics, mercats d’apostes, medis de comunicació i anàlisis de sentiment, i xarxes social, on s’expliquen els avantatges, complicacions de recol·lecció, i limitacions en vers a biaixos, errors de mostreig, influencia de factors externs, dades falses, respostes ambigües, ..., to hi 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 estandardització de variables,... on la representativitat i la qualitat tenen un impacte directe en el rendiment i l’avaluació dels models. Així com varies 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 els models estadístics i economètrics, on aquest 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 series temporals.

En la secció del 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 maquines de suport vectorial (SVM) o xarxes de neurones artificial (ANN) per a l mètodes supervisat, k-means per als no supervisats, i l’algoritme *Q-learning* i el 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 praxis del mantenim i monitorització dels models d’aprenentatge automàtic, a manera degarantir la precisió i la eficiència d’aquests. El que implica l’avaluació del model en temps real a mesura que s’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 de informar la incertesa i el marge d’error, per tal que els interessats puguin interpretar el 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]](#Ref_10_B), on els colors graduats mostrarien l’interval de confiança dels resultats i permeten visualitzar clarament on els pronòstics son mes 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 noticies, en el treball d’aquest document, tret la de les identitats socials, ja que es mes 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 referencia la capacitat del polítics electes que ha han de tenir per a adaptar-se a les demandes recents i canviants dels ciutadans. Demandes que poden quedar reflectides en les noticies emeses pels mitjans de comunicació, tal com es pretén esbrinar en el treball d’aquest document; si el resultat de predicció del model resulten concloents vers les preocupacions que s’extrauran de les noticies.

S’ha trobat especialment rellevant la descripció del 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 tendeixin a canviar el seu vot segons les preocupacions extretes de les noticies. 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 per tant, no els afectin les preocupacions actuals o be en facin cas omís.

Amb relació als medis de comunicació es comenta la seva influencia en vers a les actituds i participació del votants, i per tant es reflexiona sobre si les preocupacions que es volen extraure de les noticies son genuïnes i independents? o venen definides per influir? Si be es cert que es un aspecte a tenir en compte, les fonts de dades que es pretenen fer servir, extrauen les noticies de diversos mitjans de comunicació, en la majoria de 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 varis mitjans de comunicació. El que implicaria emmagatzemar també el mitja de comunicació de cada noticia a manera d’avaluar la varietat de procedència i evitar biaixos. En aquest sentit tenim les notícies dels últims 50 anys d’Amèrica, les quals procedeixen del diari *New York Times*, així que es podria considerar un biaix, tot hi que tenim entre 3.000 i 8.000 noticies mensuals, en pròximes versions s’hauria d’incorporar mes varietat de medis de comunicació pel que fa als Estats Units d’Amèrica.

Pel que fa las mètodes d’aprenentatge automàtic que es descriuen, destaquem el supervisat per als casos on es tinguin suficients dades per a l’entrenament, com per exemple per a els Estatus Units, on es tenen noticies de bastants anys enrere amb els seus presidents amb la seva posició. En el cas on no tinguem prou dades de noticies per realitzar un entrenament de predicció, com malauradament en els casos dels països de la Unió Europe, s’utilitzaran models d’aprenentatge no supervisat per classificar-los segons la similitud en períodes concrets amb el dades extretes de les noticies dels Estats Units, com per exemple els models d’agrupació o classificació de textos.

A propòsit del compliment de la protecció personals de les dades, en vers al reglament GDPR establert per a tots els ciutadans de la Unió Europe[[9]](#Ref_9_B), tot hi que a priori es recullen dades de diferents mitjans de comunicació les quals poden incloure dades personal, durant el processament d’aquestes es pretenen anonimitzar, eliminat noms propis, tret dels presidents de cada país, que en tot cas si fos necessari també s’eliminarien.

Tot hi 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 demostrar 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 a el text de les notícies, a causa de que no son escrites per el poble en general, si no 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 son rellevants, es consideraria afegir aquest nou estudi com a complement de predicció, això sí, enfocat amb la 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 mes robust.

# Materials i mètodes

En aquest capítol, es descriuran els aspectes més rellevant 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*

**Predict the winner political position in next presidential elections.**

Pre-process and apply classify model, save in local and HDFS

No msg in

5 minutes.

Get features daily news with NLP and save it to HDFS

Prepare dataset of last news to predict

Load the best model

Predict political position and save results

**Sleep till next day**

**Get old news**

Filter and Save in HDFS, and apply classify model

Get features old news with NLP and save it to HDFS

**Train models**

Save the best model

Prepare data set to train model

Save Concerns

(words and coefficients)

Save metrics

**Get daily news**

Send to topic producer

Translate

Initialize consumer

Get messages

*Font: Elaboració pròpia*

La documentació del codi es pot trobar a l’accés directe “*Documentation - index.html*” que s’adjunta amb el codi, la qual s’ha generat amb l’aplicació *DoxyGen*.

## Configuració de l’entorn

La configuració de l’entorn per al desenvolupament es realitza en local, el manual d’instal·lació s’inclou al [Annex 1](#Annex_1_B).

Un cop l’entorna esta configurat i iniciat, només cal iniciar els serveis, de *Kafka* i *Spark* tal com s’indica al manual.

Tot seguit s’ha de obtenir les *key* de les diferents API, donar-se d’alta a cada una, i intercanviar les *key* 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 descarregar les dades antigues executant el fitxer:

* get\_old\_news\_us.py

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

* train\_model.py

Finalment un cop tenim el model entrenat, podem iniciar la descarrega diària de les noticies, on es tracten i s’injecten al tòpic del productor, consumidor, processament, classificació i 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 esta llesta per capturar les dades a diari i mostrar-les. Prèvia actualització de la ruta dels conjunts de dades des de *Power Query* accedit 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

## Els conjunts de dades

Els conjunt de dades principals, dels quals extraurem les característiques provenen de diferents interfície de programació d'aplicacions (API), de les que es descarregaran a diari les noticies dels diferents països per a ser processades, a mes de la obtenció del conjunt de dades inicial amb milers de noticies per a poder utilitzar-lo com a conjunt d’entrenament en els models d’aprenentatge automàtic supervisat.

### 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 noticies 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 hi haver filtrat els atributs que retorna la consulta, únicament s’ha seleccionat el títol la descripció i la data de publicació, generant un *dataset* de 1,5 Gb, on per cada mes s’han acumulat entre 3000 i 8000 noticies.

Encara que les noticies venen únicament d’un únic mitjà de comunicació, indici d’un possible biaix polític, en processos posteriors es tractarà de minimitzar-lo eliminant persones, organitzacions, localitzacions, mitjançant el model entrenat [“*English Berttest BertForTokenClassification from RtwC*](https://sparknlp.org/2025/01/29/berttest_en.html)".

Aquest conjunt inicial juntament amb [els presidents dels Estats Units](https://www.whitehousehistory.org/the-presidents-timeline), el seu nomenament i la seva posició política, serviran per a entrenar el model supervisat de regressió logística, on la variable objectiu o depenent serà la posició política guanyadora de les eleccions presidencials, on la posició es simplificarà a binomial (esquerra o dreta).

Per tant aquest conjunt de dades juntament amb el model ajudarà a donar 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 extretes de les noticies (variables independents) i la posició política guanyadora de les eleccions (variable depenent).
* I el grau de magnitud d’aquesta relació, la que s’avaluarà amb diferents mesures.

### Llistat de presidents i posició política (la etiqueta a predir)

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

* Presidents Estats Units: <https://www.whitehousehistory.org/the-presidents-timeline>
* Presidents Espanya: <https://www.lamoncloa.gob.es/presidente/presidentes/Paginas/index.aspx>
* Presidents França: <https://es.wikipedia.org/wiki/Presidente_de_Francia>

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

En aquestes llistes de presidents s’han exclosos els presidents interins i el mandataris en període de dictadura, ja que el poble no els ha votat. Per tant s’ha decidit que, aquest períodes buits entren directament dins del 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 pera España 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 hi que per a ser utilitzats en el procés de generació de les etiquetes predictives i els períodes electorals, al generar el conjunt de dades de entrenament del model, i per a fer la predicció de les noticies actuals, els fitxers dels diferents països es combinen en un a diari, just avanç de generar els *datasets*, així si hi ha una actualització es te 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

En relació a la posició política, si be amb els presidents dels Estats Units hem trobat que tots son republicans o demòcrates, amb els dels països de la comunitat Europeo i el Regne Unit, hem troba posicions de centre i independents. En aquests últims casos s’han unificat per tant la etiqueta resultat de la combinació dels fitxers serà [centre,2], tot hi 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.

### 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 noticies dels diferents països te limitacions, es reparteix la càrrega entre dues API, una per a descarregar les últimes noticies dels Estats Units, i l’altra per a descarregar les últimes noticies dels diferents països de la comunitat Europea.

* 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   + Axios   + Associated Press   + ... |
| https://newsapi.org/v2/top-headlines?**country**=**us**&apiKey=yourkey&**pageSize**=**10** | |

*Font: Elaboració pròpia*

* Les últimes notícies dels països de la comunitat Europea, provenen 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   + The Local   + 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 noticies 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 al haver-les eliminat en el conjunt d’entrenament, i per tant podria tenir un impacte negatiu en la predicció. A mes les entitats d’un país poden no ser les de altres, i considerant que les noticies descarregades de països no angloparlants es tradueixen a l’anglès, es 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 noticies i utilitza el ISO del país i el Codi ISO de la llengua que s’hi parla, per descarregar les noticies 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 fa un *request* a les apis amb límit de noticies configurable, amb versió gratuïta unes 10, i establint l’idioma amb el que es volen rebre les noticies de cada país, ja que altrament pot ser diferent i complica la traducció.

Un cop descarregades es tradueixen una a una amb *Google Translate*, limitant prèviament el total de caràcters a 2500, ja que en ocasions si la descripció de la noticia l’excedia retornava error per *time-out* o excés, establint l’idioma origen i el destí, que encara que *Google Translate* dona la opció de reconèixer automàticament l’idioma origen, aquest ha donat problema, doncs en ocasions el castellà la interpretat com a portuguès en frases curtes, i per aquest motiu tant en els *requests* de les API com en la traducció s’estableix l’idioma origen en el que es vol la noticia i la tenim.

|  |
| --- |
| 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 mode de control al finalitzar la descarrega de noticies, traducció i enviament dels missatges al tòpic, es mostra un resum on es pot visualitzar el total de noticies tractades per país, on podem observa que els països angloparlants no han estat traduïts, i que en aquesta ocasió les consultes a les API han retornat notícies demanades, excepte per en el cas dels estats Units, configuració establerta a les constants, 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 noticies dels països definits, s’inicia el consumidor de Kafka[[11]](#Ref_11_B) 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 fitxer 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 exemple de 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 mes per cada bloc rebut s’extreu la classe de cada noticia 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 un format processable per Spark NLP.
2. **Tokenizer**: Retorna el document en *tokens*.
3. ***DistilBertForSequenceClassification***: finalment s’aplica el model entrenat *DistilBERT Sequence* *Classification* [[12]](#Ref_12_B), que prèviament durant la preparació de l’entorn s’ha descarregat i ara el carregem, ja que carregar-lo cada cop que s’usa requereix mes recursos. Amb aquest model es procedeix a identificar de quin tipus es cada noticia (Business, Sci/Tech, Sports, World). El model te una avaluació molt bona, doncs te un F1-Score de 0,90 per cada classe. A continuació es mostra un exemple del la classificació, que com a resum de preocupacions generals diàries de cada país es guardarà.

+--------+------------+-------------------------------------------------------------+----------+

|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.

  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

## 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]](#Ref_13_B), les quals son:

1. **DocmentAssembler**: Prepara les dades en un 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 pre-entrenat a l’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 eliminant les paraules inferiors a 3 caràcters.
5. **StopWordsCleaner**: elimina les paraules buides aplicant el model entrenat en angles anomenat *“stopwords\_en”*.
6. ***BertForTokenClassification***: finalment s’aplica el model entrenat English berttest BertForTokenClassification from RtwC[[14]](#Ref_14_B), que prèviament durant la preparació de l’entorn hem descarregat i ara el carregem, ja que carregar-lo cada cop que s’usa requereix mes recursos. Aquest model te una avaluació de la mesura F1-Score del 0,94 per tant el reconeixement d’entitats nombrades es molt bo, tot hi que per a les traduccions de les noticies d’altres idiomes es probable que no les identifiqui amb aquesta bonança. En tot cas es procedeix a identificar les entitats anomenades *NER*. (Tot hi que de tant en quant s’haurà d’anar descarregant versions recents per identificar millor les entitats.)

Un cop instanciat el *Pipeline* i les seves etapes es procedeix a entrenar el model i a transformar 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 noticia, on es podrà observar la 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] |
| Lema | [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 es seleccionen in combinen els *tokens* resultants de la etapa de *Stop Words* amb la classificació *NER*, 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 hi que com s’ha comentat poden ser diferents per cada país, i per tant al predir la posició política siguin irrellevants si no han estat considerades durant l’entrenament. En tot cas durant les proves realitzades si s’ha vist alguna entitat MISC internacional que te especial rellevància a dia d’avui com UCRANIAN.

Es converteixen a minúscules, ja que si es feia avanç del NER no reconeixia correctament algunes entitats. Tot seguit s’exclouen referencies temporals i 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 a diari 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 al haver de recuperar múltiples fitxers.)

A continuació es pot visualitzar una mostra de com queda la combinació de les noticies processades per a l dia 20 d’abril del 2025 per a España 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 |

## 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 noticies diàries que es reben.

Per a generar el conjunt de dades per a l’entrenament del model, es necessita incorporar l’etiqueta per a cada agrupació de noticies diàries, per tant, es necessitarà la llista de períodes electorals dels presidents dels estats units, dels quals s’extraurà la data inici i final de cada president i la posició política, a continuació es mostra un exemple, tot hi que com a punt rellevant, la data de nomenament dels presidents dels estats Units es aproximadament 76 dies mes tard del dia de les eleccions, per aquest motiu a la data de nomenament se li ha restat aquest dies. Així a l’hora de seleccionar les noticies publicades en cada període electoral aquestes no influiran en la predicció, ja que durant aquest 76 dies ja se sap, o be hi ha un alta certesa de qui es el guanyador, per aquest motiu s’exclouen durant el procés d’entrenament i test del model. Aquesta mesura no hagés sigut necessària si durant la recollida del conjunt de dades dels presidents dels estant units, s’hagés anotat el dia de les eleccions en lloc del nomenament. En tot cas es important excloure-les, ja que com s’explicarà en breu, les noticies pròximes al dia de les eleccions prenen mes rellevància al 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 podem observar també s’han calculat els dies de cada període en el que ha governat el president electe. Aquesta nova característica s’utilitzarà pera ponderar les noticies diàries, així segons si les noticies emeses estan mes pròximes al dia de les eleccions el model els hi donarà major importància, i a la inversa, com mes dies faci de les noticies, menys importants seran per al model, encara que si es una noticia 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 matis de ponderar les notícies segons la data de publicació s’ha incorporant 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 escoltat, i per aquest motiu s’han ponderat. A continuació es mostra un exemple del conjunt de dades de noticies antigues ponderat, remarcant les noticies agrupades per dia amb mes 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 del les noticies actuals, es a dir, des de les últimes eleccions fins a dia d’avui, es considerarà les dades de les últimes eleccions de cada país fins al dia actual, per calcular els dies transcorreguts i poder ponderar les noticies diàries de cada país. Per exemple per al dia 20 d’abril tenim que:

## Show days fom 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|

| ee| Alar Karis|2021-10-11| centre| 2|2025-04-20| 1287|

| es|Pedro Sánchez Pér...|2023-11-17| esquerra| 0|2025-04-20| 520|

| fi| Alexander Stubb|2024-03-01| dreta| 1|2025-04-20| 415|

| fr| Emmanuel Macron|2022-05-14| dreta| 1|2025-04-20| 1072|

| gb| Keir Starmer|2024-07-05| esquerra| 0|2025-04-20| 289|

| gr|Konstantinos Taso...|2025-03-13| dreta| 1|2025-04-20| 38|

| hr| Zoran Milanović|2020-02-19| esquerra| 0|2025-04-20| 1887|

| hu| Tamás Sulyok|2024-03-05| centre| 2|2025-04-20| 411|

| ie| Michael D. Higgins|2018-11-11| esquerra| 0|2025-04-20| 2352|

| it| Sergio Mattarella|2022-02-03| esquerra| 0|2025-04-20| 1172|

| lt| Gitanas Nausėda|2019-07-12| centre| 2|2025-04-20| 2109|

| lu| Luc Frieden|2023-11-17| dreta| 1|2025-04-20| 520|

| lv| Edgars Rinkēvičs|2023-07-08| dreta| 1|2025-04-20| 652|

| mt|Myriam Spiteri De...|2024-04-04| esquerra| 0|2025-04-20| 381|

| nl| Dick Schoof|2024-07-02| centre| 2|2025-04-20| 292|

| pl| Andrzej Duda|2020-08-06| dreta| 1|2025-04-20| 1718|

| 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 han 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 una millora en futures actualitzacions mostrar també la probabilitat de la predicció en la visualització i/o fer una classificació amb múltiples classes, ja que les posicions polítiques extremes també s’han agrupat.

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 de Àustria, remarcant dia amb mes pes ja que es el mes pròxim al dia actual, i el 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 upp 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 hi que actualment no es fa servir, s’ha decidit calcular les paraules diàries de les noticies agrupades per país, la columna *“word\_count”* en fa referència. Si ve s’ha pogut observar que per cada país el valor acostuma a ser similar, sí que s’ha vist en ocasions el que sembla *outliers*, no obstant no els tractem, ja que les noticies venen de diferents mitjans de comunicació, i cadascun deu tenir la seva forma d’explicar, o be en segons quin temes n’aprofundeixen mes. 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 noticies antigues, i per a predir les posicions de les noticies recents. A continuació s’explica com s’ha creat el model de predicció de les posicions polítiques.

## 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 entrenarem un model de Regressió logística[[15]](#Ref_15_b), establin com a model base el model *Naive Bayes*, classificador probabilista simple basat en teorema de *Bayes*, el que suposa independència entre característiques a diferència de la Regressió logística.

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

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

De les paraules agrupades de les noticies diàries de cada país amb els seus pesos, es crearan el conjunt d’entrenament i el de test, que com veurem estan esbiaixades en un 30 per cent 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 seqüencia d’etapes, en aquest cas per al model de Regressió Logística:

1. **Tokenizer**: Retorna el text en *tokens*.
2. **CountVectorizer****[[16]](#Ref_16_B)**: converteix la llista de *tokens* en un vector numèric on cada numero representa la freqüència de la paraula del vocabulari.
3. **IDF**[**[16]**](#Ref_16_B)**:** Calcular la freqüència inversa per donar mes rellevància a les paraules importants, reduint els pesos dels termes que apareixem amb mes freqüències.
4. **LogisticRegresion()**: crea model per a la cssificaicó que utlitzarà 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 un 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])

La optimització dels hiperparàmetres amb validació creuada es centra en la mida del vocabulari, que s’ha extret de la 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 cero, i el tipus de model “Multinomial” o “Gaussian”, el primer per a dades continues i el segon per a discretes, a mes 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 model per noticies diàries i agrupant aquestes per períodes electorals, mostren el millor AUC en el Model de Regressió Logística optimitzat amb la cerca de hiperparàmetres en la validació creuada, tal hi 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 hi que s’ha vist que el F1-Score per període es alt per al model de *Naive Bayes* optimitzat, i per al model de Regressió Logística base es igual al de la validació creuada, la resta de mesures son inferior, 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 hi que les mesures de AUC son similar, van de 5 punts, a mes la validació creuada es mes robusta.

Pel que fa al resums de mètriques de predicció de noticies diàries, s’observa que el model classifica millor amb una diferència de 15 punts la classe 1 (dreta), 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 hi que si s’observa l’agrupació dels resultats per període electoral el resum de mètriques informa que la classe millor classificada es també la 1 (dreta), però en aquest cas amb una diferència de 4 punts, aquest fet pot ser a ran dels pocs períodes electorals que es tenen al conjunt de dades, que esdevindrien per al conjunt de test 15, es a dir les noticies diàries seleccionades aleatòriament per al conjunt de test pertanyen a 15 períodes electorals, d’un total de 17 presidències des dels 1970.

Es podria pensar, i perquè no s’han agrupat les noticies per període electoral, doncs perquè únicament s’hagés obtingut un conjunt de dades amb 17 mostres, i no s’hagés pogut incorporar l’oblit en el temps, tal com s’ha implementat amb el pes que se li ha donat a les noticies mes o menys llunyanes respecte el 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 noticies de la resta de països seleccionats. Addicionalment es guarden les mètriques i les paraules del vocabulari amb els seus coeficient, per mostrar-ho a la visualització, a continuació es mostren les 10 mes 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|1980-11-05|1984-11-05| 1| 0.0| 77|

| us|1980-11-05|1984-11-05| **1| 1.0| 204|**

| us|1984-11-05|1988-11-05**| 1|** **1.0| 214|**

| us|1984-11-05|1988-11-05| 1| 0.0| 81|

| us|1988-11-05|1992-11-05| 0| 1.0| 140|

| us|1988-11-05|1992-11-05**| 0|** **0.0| 166|**

| us|1992-11-05|1996-11-05| 0| 1.0| 93|

| us|1992-11-05|1996-11-05**| 0|** **0.0| 180|**

| us|1996-11-05|2000-11-05| 1| 0.0| 74|

| us|1996-11-05|2000-11-05| **1| 1.0| 229|**

| us|2000-11-05|2004-11-05| 1| 0.0| 90|

| us|2000-11-05|2004-11-05**| 1|** **1.0| 197|**

| 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, a diari amb la descarrega de les noticies dels diferents països netes i processades es fa la predicció per a cada país de les noticies emeses des de l’últim nomenament fins al dia actual obtenint la predicció, la que es combina amb el fitxer de presidents CSV el 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

cy,PREDICTION,2025-04-22,dreta,1

cz,PREDICTION,2025-04-22,dreta,1

de,PREDICTION,2025-04-22,dreta,1

dk,PREDICTION,2025-04-22,dreta,1

ee,PREDICTION,2025-04-22,dreta,1

es,PREDICTION,2025-04-22,dreta,1

fi,PREDICTION,2025-04-22,dreta,1

fr,PREDICTION,2025-04-22,dreta,1

gb,PREDICTION,2025-04-22,dreta,1

gr,PREDICTION,2025-04-22,dreta,1

hr,PREDICTION,2025-04-22,esquerra,0

hu,PREDICTION,2025-04-22,esquerra,0

ie,PREDICTION,2025-04-22,dreta,1

it,PREDICTION,2025-04-22,dreta,1

lt,PREDICTION,2025-04-22,dreta,1

lu,PREDICTION,2025-04-22,dreta,1

lv,PREDICTION,2025-04-22,dreta,1

mt,PREDICTION,2025-04-22,dreta,1

nl,PREDICTION,2025-04-22,esquerra,0

pl,PREDICTION,2025-04-22,dreta,1

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 mode de curiositat, com es pot observar en la predicció del dia 22 d’abril, per als Estats Unit prediu que, si hi hagessin 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 a causa de que solament es tenen noticies antigues dels Estats Units, i no s’ha pogut construir altre model amb noticies antigues d’altres països, per tant el model de predicció entrenat amb les noticies dels Estats Units serà el que s’utilitza per a predir la resta de països.

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

Un cop desenvolupat el diagrama d’activitats, el qual recull l’extracció de dades, la neteja, tan com 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 al 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 mode d’introducció, tot i que s’ha anat esmentant en els punt anteriors, a continuació es descriuen els fitxers que s’utilitzaran en el model de la visualització, i el perquè d’algunes decisions preses, on tot es trobaran en local a \\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.

**Noticies**

* **concerns\_summary.csv**: Conté la classificació de les noticies diàries de cada país. Les classes son [Business, World, Sci/Tech, Sports] . (S’actualitza automàticament cada dia.)
* **news\_clean.csv**: Conté les paraules netes i agrupades per dia de les noticies extretes de cada país a diari. (S’actualitza automàticament cada dia.)
* **old\_news\_clean.csv**: Conté les paraules netes i agrupades per dia de les noticies antigues dels Estats Units des del 1970. (Es crea únicament al descarregar i processar les noticies antigues.)

**Model predictiu**

* **predictions.csv**: A mes de les prediccions de la posició política del dia actual de cada país, conté els presidents que van tenir des del 1970. S’ha de tenir en compte que pera a una correcta visualització de la 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 cada noticia de cada país, des del nomenament del president de les últimes eleccions, donant així una referència detallada de per quan guanya una posició de l’altra. En aquest apartat tot hi que el model de predicció únicament prediu 0 o 1 (esquerra o dreta), al fer el recompte si hi ha empat la predicció mostrarà un 2 el qual s’indicarà amb un color distint al mapa que es 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 distint 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 mes positives/negatives i estudiar-les. (Es crea únicament al 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 considerats i els seus valors. (Es crea únicament al finalitzar l’entrenament del model.)

### 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, en els objectes que ho requereixin, com per al mapa, taules il·lustratives i interactives.

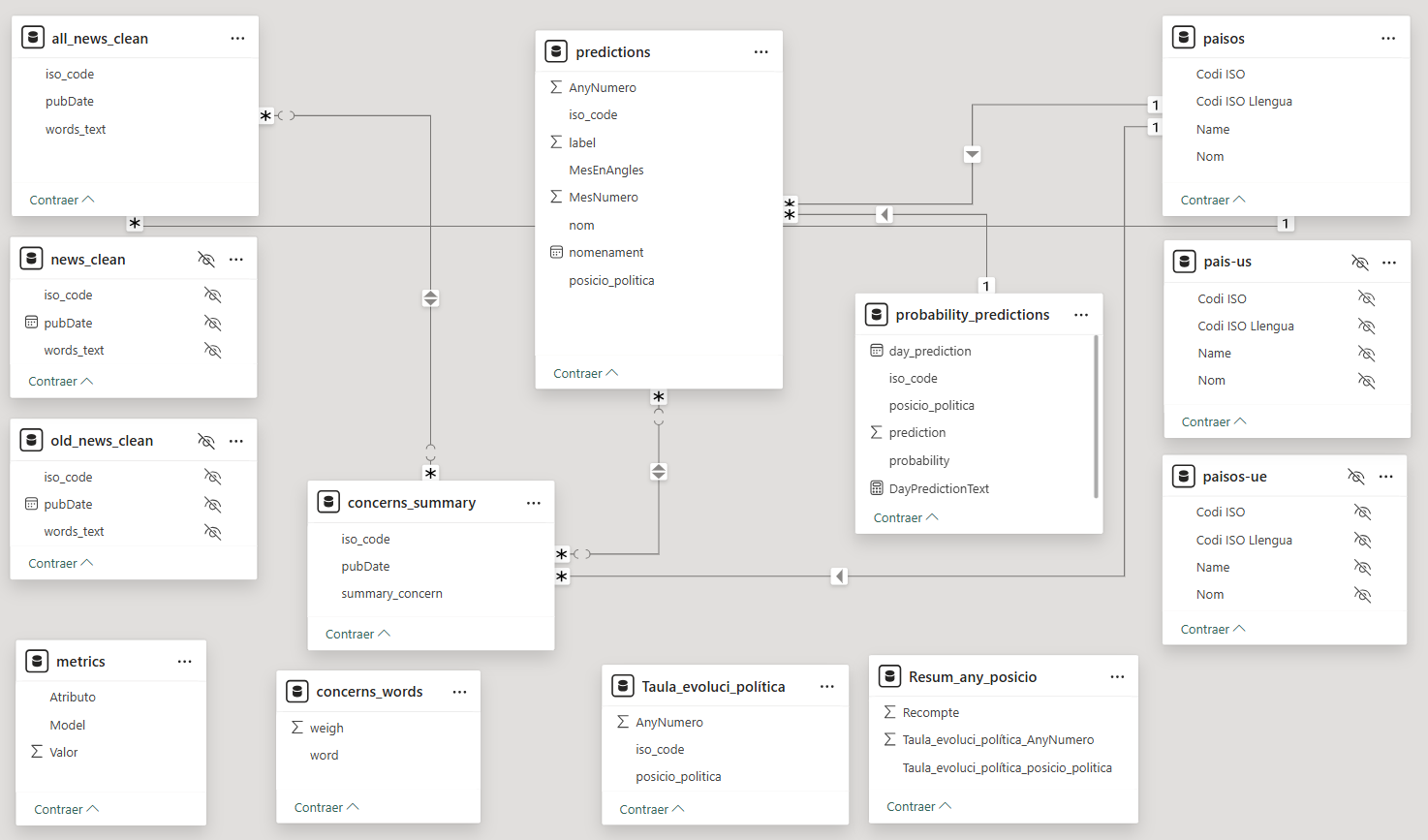
A mes també s’han combinat les noticies actuals amb les antigues, aquestes últimes son les que es van fer servir per a l’entrenament del model, així mateix es pot seguir la evolució de la trajectòria fàcilment seleccionant les dates. Aquesta unió quedarà reflectida al núvol de paraules on, mentre hi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

Pel que fa als a les taules que representen els fitxer de concerns\_words.csv i metrics.csv, no estan relacionades amb la resta a causa de que esdevenen una informació mes tècnica del model entrenat que s’utilitza per a la predicció. Tot hi que s’ha volgut mostrar per a que els usuaris mes experts, tinguin un detall mes 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. Aquestes taules no s’han relacionat amb la resta a ran de que la gràfica es 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, al 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 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 mes de 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 mes 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 en primer lloc per codi de país, any i posició política, excloent les prediccions per a que 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 progres amb una única selecció de data. Tot hi que es podia haver fet amb selecció de data inici i fi, aquest rang de dates podria incorporar dos nomenament en el mateix període d’un país, i en aquest cas si la posició política hages 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 la que donar-hi una pensada.

### La visualització

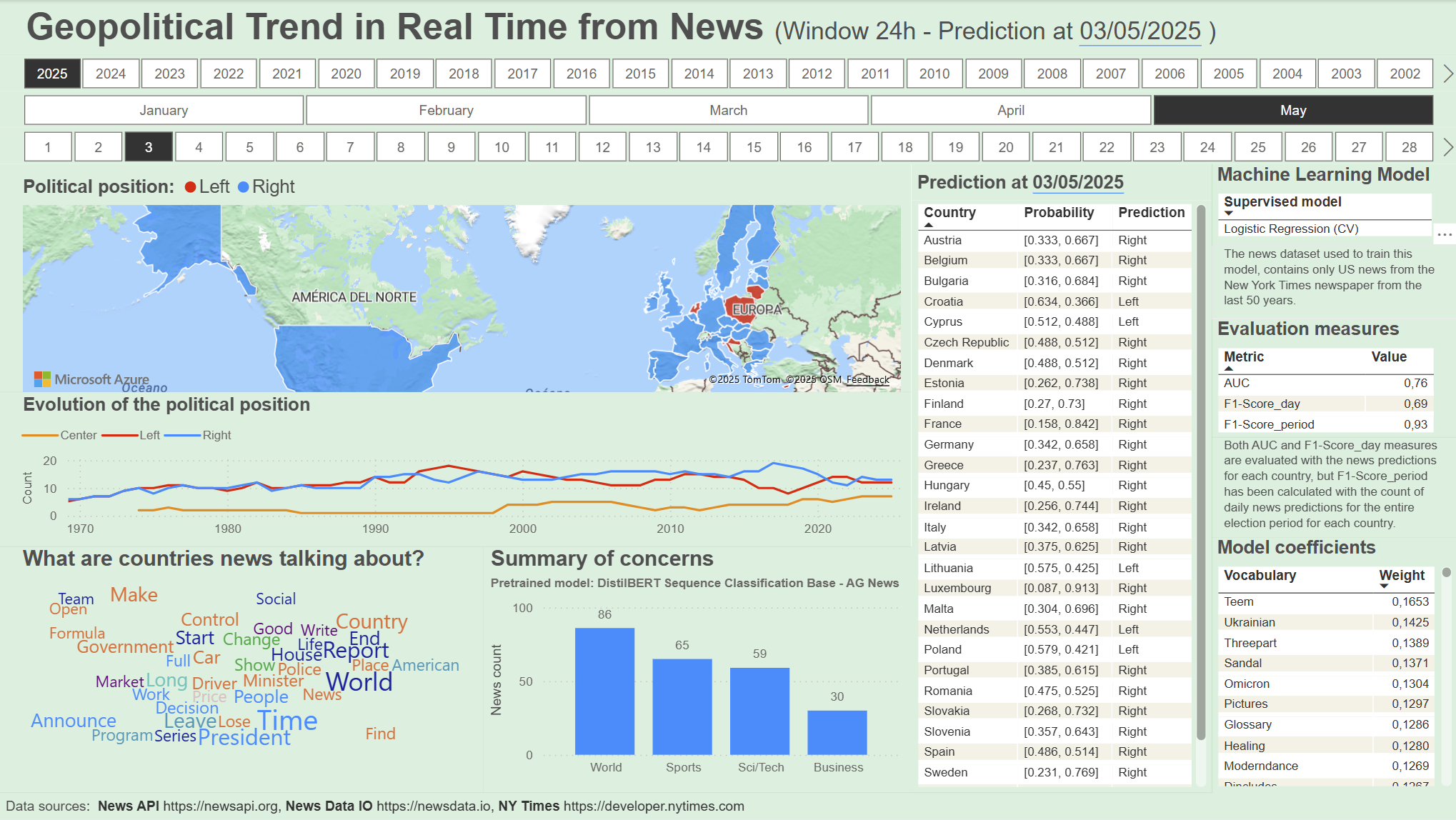
A continuació es mostra la visualització final realitzada amb *Power BI*[[17]](#Ref_17_B), en la que s’explicaran els diferents filtres, gràfics, taules i les seves interaccions. On per a la disposició dels diferents elements de la visualització s’ha seguit els criteris basics d’ordre natural:

* En primera posició a dalt i a la esquerra: El títol amb una grandària de font rellevant, que resumeix de que va la visualització, seguit d’un subtítol el qual fa mes 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 la evolució de les relacions entre posició política i paraules de les noticies.
* De dalt a baix: el cos principal que mostra els gràfics amb filtres secundaris i taules, 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 la 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 mes comentades de les noticies, i un diagrama de barres vertical que mostra el recompte de tipus de noticies.
* Als racons: fonts, llegendes i informació mes tècnica per als usuaris experimentats.

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

En 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 mes facilita la lectura horitzontal entre les diferents columnes.

*Figura 8: Visualització Power BI*



*Font: Elaboració pròpia*

Filtres temporals

S’han extret de les dates que inclouen la taula de prediccions, i es mostren en format botó l’any, el mes i el dia. Tot hi que s’ha trobat limitacions per mostrar-ho com una barra de progres seleccionat una única data, tal hi com s’ha comentat en l’apartat 3.6.1, la avantatja que s’hi es que es pot anar directament a la data interessada, en lloc de balancejar fins una data concreta.

La interacció amb aquest filtre afectarà al mapa coroplètic canviant de color els països en vers a la posició política del gover en aquella data, tret del dia actual de la predicció, en que mostrarà la etiqueta predita.

També afectarà al núvol de paraules mostrant aquelles mes freqüents de les noticies del dia seleccionat.

Igualment filtrarà el recompte del resum de preocupacions en el gràfic de barres, per al 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 mes 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 mes freqüents de les noticies del país seleccionat.
* **Gràfic de barres**: filtrarà mostrant únicament el recompte de les noticies 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, i amb aquest 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 be, de que 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 noticies 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 diferents països.

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

Si ve es cert que a mitjans de març es quan s’incorpora la resta de noticies dels països Europeus seleccionats, hi ha uns dies en que les noticies antigues descarregades i processades es solapen amb les diàries per al país dels Estats Units, per tant, per aquest dies encara que es seleccioni únicament els US, 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 noticies *DistilBert Sequence Classification Base – AG News*, mostra el recompte de les classes per a tots els països del dia seleccionat.

Al interactuar-hi 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 noticies 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 noticia de cada dia de cada país, i en el núvol de paraules es mostren les paraules resultants de la unió de les noticies de cada país per a cada dia. Per tant, al seleccionar la classe ciència/tecnologia, no es pot saber quines paraules intervenen en aquesta classe, ja que talc com s’ha comentat intervé la descripció sencera de la noticia per a ser classificada.

Si ve el títol d’aquest gràfic inclou la paraula preocupacions, tal hi com s’hi ha fet referència des de l’inici d’aquest document, que preocupa a les persones per a que 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 hi 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 de lo 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 aquest s’han de preocupar dels problemes dels ciutadans, per tant, preocupacions, problemes i interessos quedaria entrellaçat.

Taula de prediccions

La taula de les prediccions del dia actual, o mes ben dit de l’últim predit si encara no s’ha obtingut les últimes dades generades, en tot cas el dia predit es mostrarà sempre al títol de la taula i al títol de la visualització.

Com a interacció te que filtra igual que el mapa coroplètic, focalitzant-hi únicament el país seleccionat, mostrant únicament les paraules mes freqüents també del país seleccionat al *Word Cloud*, i resumint el recompte de preocupacions igualment.

Tot hi que inicialment no estava previst mostrar la probabilitat, es va considerar especialment rellevant ja que indica per quan es decanta per una posició política o per l’altra detallant així mes informació per 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 son només informatives, i no tenen cap tipus d’interacció amb la resta. Tot hi que per als usuaris mes experts donen confiança i credibilitat extra als resultats de les prediccions mostrades a la visualització.

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 mes positiu i mes negatiu, a mes 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.*

## Limitacions

La limitació mes 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.

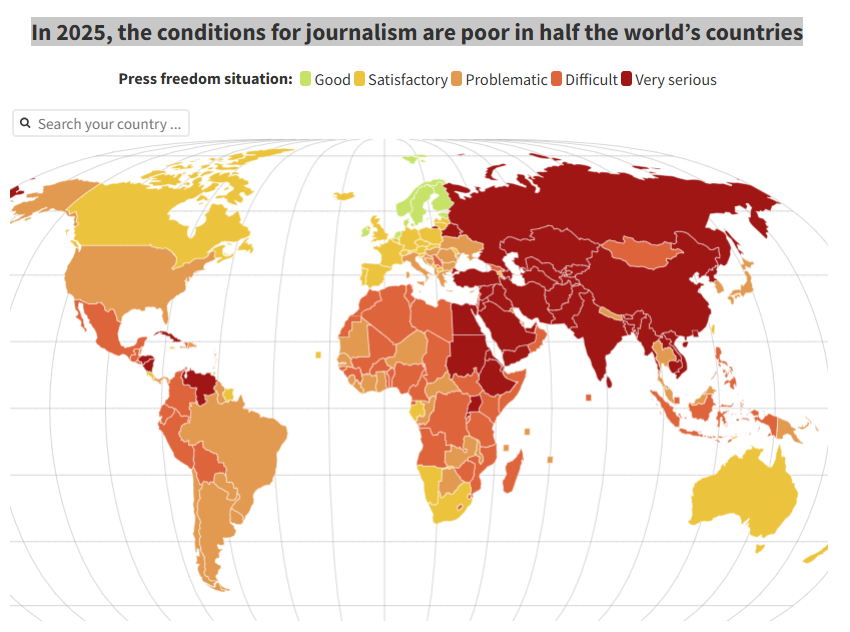
Tot hi així el volum de noticies descarregades per a l’entrenament del model ha estat tal, 1,5 Gb aproximadament, i que al generar l’extracció de característiques ha estat inviable el tractament, i per aquest motiu finalment s’ha seleccionat dels últims 50 anys únicament 10 noticies diàries, que resulten unes 300 mensuals. Tenint en comte que es van descarregar mensualment entre 3000 i 8000 noticies mensuals, s’ha descartat una quantitat immensa de mostres per limitacions de l’entorn. Tot hi així sembla que l’avaluació del model dona bastant bons resultats, fet que s’haurà d’anar contrastant a mesura que les eleccions dels països Europeus contemplats celebren eleccions.

En aquest sentit tot hi que es podia haver optat per un entorn *Cloud* pre-configurat i escalable com AWS, aquesta opció en local ha donat la oportunitat de resoldre diferents dificultats aprenent aspectes de sistemes no relacionats directament amb el treball d’aquest document, tot hi 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.

Amb relació a les versions gratuïtes de les API no han permès la descarrega de noticies antigues per paios, per tant no s’ha pogut validar la bondat del model entrenat únicament amb noticies dels Estats units d’Amèrica, amb noticies antigues d’altres països. Si be s’ha trobat noticies antigues d’Europa, o be 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. No obstant aquest fet, (model entrenat únicament amb noticies dels Estats Units), s'indicarà a la visualització per tal que l’usuari pugui interpretar correctament els resultats de les prediccions.

Addicionalment el us de API gratuïtes per a descarregar les últimes noticies de cada pais, amb un decalatge de 24 hores, no ha resultat real, doncs en ocasions al descarregar les últimes noticies 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 noticies descarregades no processades adequadament. Aquesta situació podria solucionar-se amb una versió de pagament, al poder escollir les noticies d’un dia concret, al tenir accés a l’històric. O be si no s’hagés tingut limitacions de memòria s’hagés 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 mes noticies per cada país, tenint així mes varietat de dades provinents de mes varietat de mitjanats 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 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 a el seguiment de preocupacions entre països, amb la visualització actual resulta difícil donar resposta, tal com es comentava al abstracte, a saber quins països no tenen absència escolar per fer una investigació addicional i modificar el rumb de nostre en cas que si en tingui. En tot cas ni que finalment s’hagés agrupat els països per preocupacions similars, finalment no es creu que s’hagés obtingut fàcilment resposta en aquesta qüestió o similars. El principal motiu podria ser a causa de que les característiques es tracten en format paraula, i per facilitat el seguiment es podria haver incorporat *bigrams* i *trigrams*, doncs actualment es podria trobar la paraula ‘absence’ i ‘school però no ‘school absence’, així i tot l’extracció de les característiques s’hagessin ampliat exponencialment i els recursos locals hagessin sigut insuficients. Per tant aquesta millora quedaria pendent en entorns mes 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 de que solament es tenen noticies antigues dels Estats Units, i no s’ha pogut construir cap altre model amb noticies antigues d’altres països, per tant el model de predicció entrenat amb les noticies dels Estats Units serà el 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 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 es local, tot hi que la visualització es pot compartir en públic, no es podrà actualitzar les dades del model. Pel que fa a l’actualització diària automàtica del fitxer que content el model, *Power BI* d’escriptori no permet la programació d’horària automatitzat, per tant s’hauran de refrescar manualment mitjançant la opció de refrescament/actualització del menú. En tot cas amb Power BI web un cop publiques les dades tampoc permet un refrescament a la visualització publica.

Si be s’havia considerat alguns d’aquest 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 publica i s’actualitzi automàticament.

# Resultats

En relació a l’objectiu principal, pronòstic de la posició política dels resultats electorals, utilitzant notícies de diferents mitjances de comunicació de cada país produïdes en temps real, les quals son ingerides per un sistema de ingesta massiva de dades, amb una finestra de 24 hores, on flueixen per el mateix sistema, netejant-les, transformant-les a traves de varis pipelines donant com a resultat la predicció de la posició política al dia actual, es pot dir que durant l’entrenament els resultats de l’avaluació han estat bons, doncs 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%.

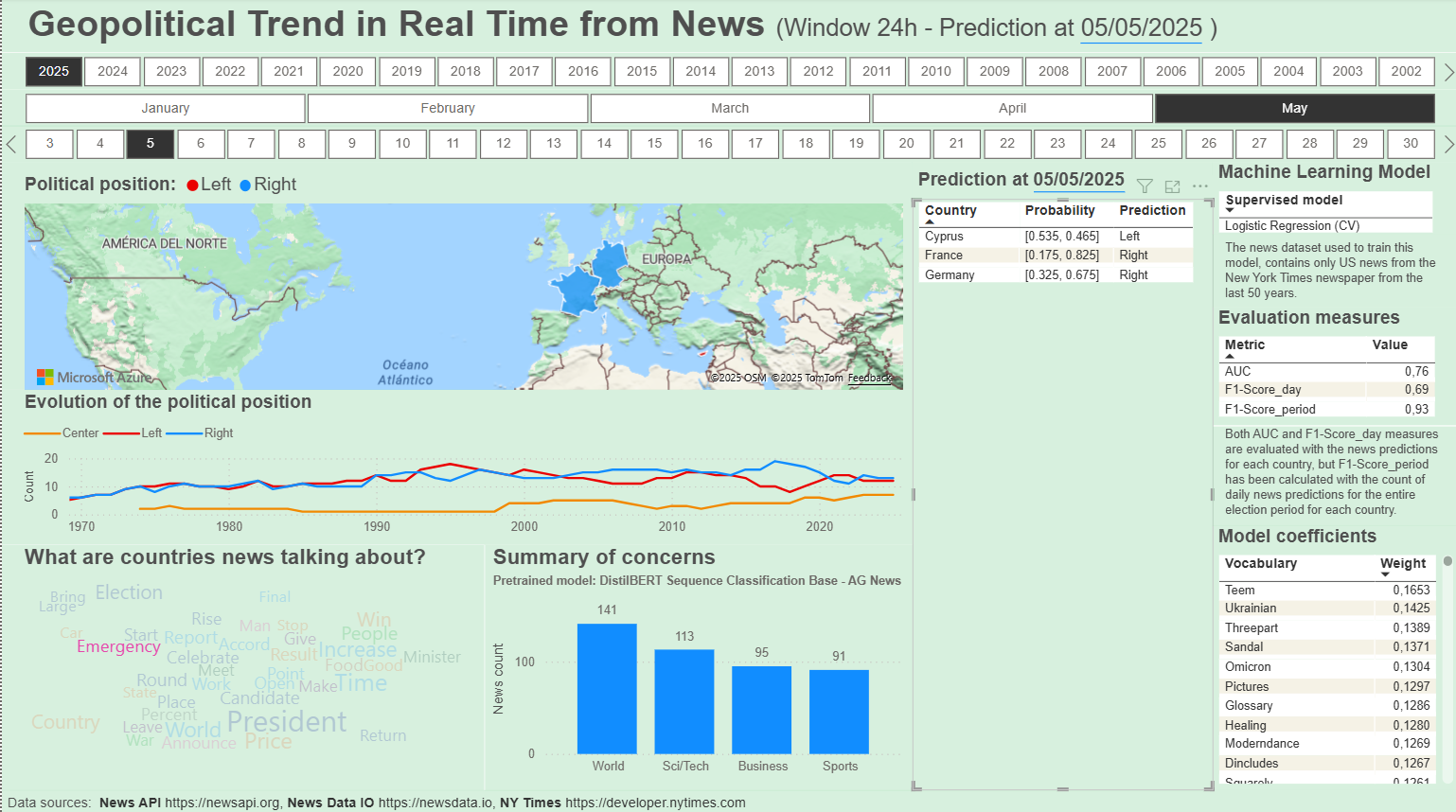
Recapitulant, el conjunt de mostres de les característiques diàries, esdevenen de l’agrupació diària de les noticies de cada país, processades i transformades per ser ingerides per al model predictiu, afegint la classe política com a label d’aquell dia concret.

Però el que realment interessa no es predir les noticies 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 mes rellevància a les noticies mes pròximes al dia de les eleccions i menys pes com mes 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 be son uns resultats excel·lents, s’ha de dir que ho son per al país dels Estats Units, ja que la mostra de noticies 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 noticies de la resta de països de la comunitat Europea mes el Regne Unit, no es tenen les noticies des de les últimes eleccions de cada país, per tant, es contemplen únicament des de 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 durant l’avaluació del model.

Pel que fa a l’objectiu secundari, derivat de la dificulta d’obtenir jocs de dades de diferents idiomes, agrupació dels països segons preocupacions, a mode 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, si es pot fer un seguiment de les preocupacions o termes mes 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 mes rellevants per a les posicions polítiques. Tot hi que en primera instancia no es pot filtrar per una paraula en concret, pròpiament dit, si que es pot seleccionar del núvol de paraules i observar en aquell dia quins països la nomenaven a les noticies. Tal i com es mostra a continuació al seleccionar Emergency, únicament queden remarcats els països de Cyprus, France i Germany:



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 hi que s’ha de fer manual clicant a cada any, es pot observar l’intercanvi de posicions polítiques visuals, i en aquest sentit si es tinguessis noticies antigues dels últims 50 anys de tots els països intervinents, es podria investigar des del Word Cloud els terme d’aquell període, esbrinant així les preocupacions d’aquell moment especial.

Tot hi això tal com s’ha comentat a la secció de limitacions, seria interessant per fer un seguiment de les preocupacions mes exhaustives i en noves versions del treball incorporar bigams, i també afegir un filtre avançat per a la cerca de preocupacions, on 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 mes tècnica que es mostra a la dreta de la visualització.

# Conclusions i treballs futurs

Aquest capítol ha d’incloure:

* Una descripció de les conclusions del treball:
  + Un cop s’han obtingut els resultats quines conclusions s’extreu?
  + Aquests resultats són els esperats? O han estat sorprenents? Per què?
* Una reflexió crítica sobre l’assoliment dels objectius plantejats inicialment:
  + Hem assolit tots els objectius? Si la resposta és negativa, per quin motiu?
* Una anàlisi crítica del seguiment de la planificació i metodologia al llarg del producte:
  + S’ha seguit la planificació?
  + La metodologia prevista ha estat prou adequada?
  + Ha calgut introduir canvis per garantir l’èxit del treball? Per què?
* Dels impactes previstos a 1.3 (ètic-socials, de sostenibilitat i de diversitat), avaluar/esmentar si s'han mitigat (si eren negatius) o si s'han aconseguit (si eren positius).
* Si han aparegut impactes no previstos a 1.3, avaluar/esmentar com s'han mitigat (si eren negatius) o què han aportat (si eren positius).
* Les línies de treball futur que no s’han pogut explorar en aquest treball i han quedat pendents.

Com a millora automatitzar el desplaçament temporal del filtre, es a dir podria posar a la visualització un botó play que mostri automàticament l’evolució del mapa coroplètic, així es veuria fàcilment els països que marquen tendència i els que els segueixen, en relació a la posició política.

# Glossari

Definició dels termes i acrònims més rellevants utilitzats dins la Memòria.

# Bibliografia

[[1]](#Ref_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]](#Ref_2) Naciones Unidas. Objetivos de desarrollo sostenible. <https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/>

[[3]](#Ref_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]](#Ref_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]](#Ref_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]](#Ref_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]](#Ref_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]](#Ref_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](#Ref_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]](#Ref_10) Institut Cartogràfic i Geològic de Catalunya. (2013). Mapa de coropletes. <https://www.icgc.cat/ca/Publicacions/Diccionaris/Mapa-de-coropletes>

[[11]](#Ref_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]](#Ref_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]](#Ref_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]](#Ref_14) English berttest BertForTokenClassification from RtwC <https://sparknlp.org/2025/01/29/berttest_en.html>

[[15]](#Ref_13) 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]](#Ref_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]](#Ref_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/>

# Annexos

### [Annex 1](#Annex_1) - Manual de configuració de l’entorn

Pas 1 - Instal.lar Ubuntu a Windows

# Instal·lar Ubuntu en Windows amb WSL obrir PowerShell y 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 per a que resolgui i per a que 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 el 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 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 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 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