|  |  |
| --- | --- |
| **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/)

**Llicències alternatives (triar alguna de les següents i substituir la de la pàgina anterior)**

**A) Creative Commons:**

[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/)

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

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

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

[Llicència de Creative Commons](http://creativecommons.org/licenses/by-sa/3.0/es/)  
Aquesta obra està subjecta a una llicència de [Reconeixement-CompartirIgual 3.0 Espanya de Creative Commons](http://creativecommons.org/licenses/by-sa/3.0/es/)

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

**B) GNU Free Documentation License (GNU FDL)**

Copyright © ANY EL-TEU-NOM.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no

Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.

A copy of the license is included in the section entitled "GNU Free Documentation License".

**C) Copyright**

© (l'autor/a)

Reservats tots els drets. Està prohibit la reproducció total o parcial d'aquesta obra per qualsevol mitjà o procediment, compresos la impressió, la reprografia, el microfilm, el tractament informàtic o qualsevol altre sistema, així com la distribució d'exemplars mitjançant lloguer i préstec, sense l'autorització escrita de l'autor o dels límits que autoritzi la Llei de Propietat Intel·lectual.

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 8](#_Toc192836761)

[Llista de Taules 9](#_Toc192836762)

[1. Introducció 10](#_Toc192836763)

[1.1 Context i justificació del Treball 10](#_Toc192836764)

[1.2 Explicació de la motivació personal 11](#_Toc192836765)

[1.3 Objectius del Treball 11](#_Toc192836766)

[1.4 Impacte en sostenibilitat, ètic-social i de diversitat 12](#_Toc192836767)

[1.5 Enfocament i mètode seguit 14](#_Toc192836768)

[1.6 Planificació del Treball 16](#_Toc192836769)

[1.7 Breu sumari de productes obtinguts 22](#_Toc192836770)

[1.8 Breu descripció dels altres capítols de la memòria 23](#_Toc192836771)

[2. Estat de l’art 24](#_Toc192836772)

[3. Materials i mètodes 25](#_Toc192836773)

[4. Resultats 26](#_Toc192836774)

[5. Conclusions i treballs futurs 27](#_Toc192836775)

[6. Glossari 28](#_Toc192836776)

[7. Bibliografia 29](#_Toc192836777)

[8. Annexos 30](#_Toc192836778)

# Llista de Figures

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

Figura 2: Planificació temporal - Diagrama Gantt 20

# Llista de Taules

Taula 1: Recursos necessaris 16

Taula 2: Descripció de les tasques 16

# 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 al 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 a nivell 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.

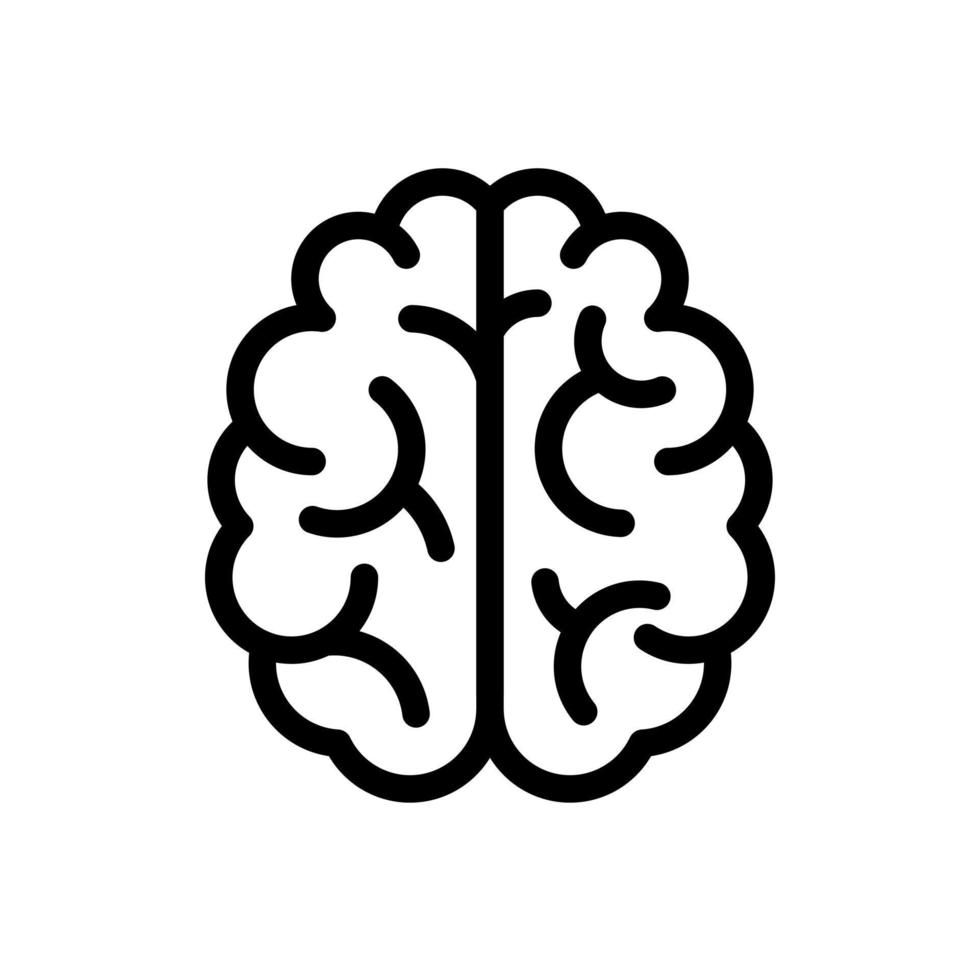
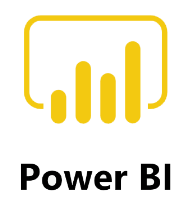
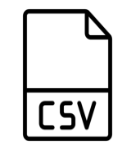
## Enfocament i mètode seguit

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

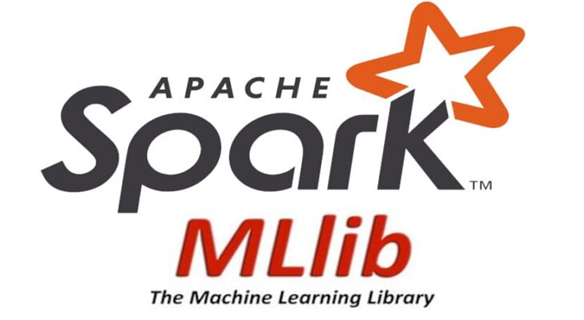
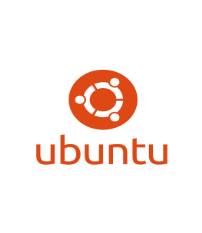
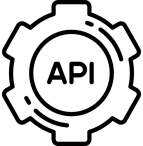
Extracció

Transformació

Emmagatzematge

Visualització

Models



*Front: 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 |
| ***Apache Hive*** | 4.0.1 |

*Front: 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* i *Hive*** | 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, preparar connexió Apache Hive. | 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 *Apache Hive***  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 *Apache Hive* 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 *Apache Hive***  Crear la connexió de les dades a l’*Apache Hive* 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** |

*Front: 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).

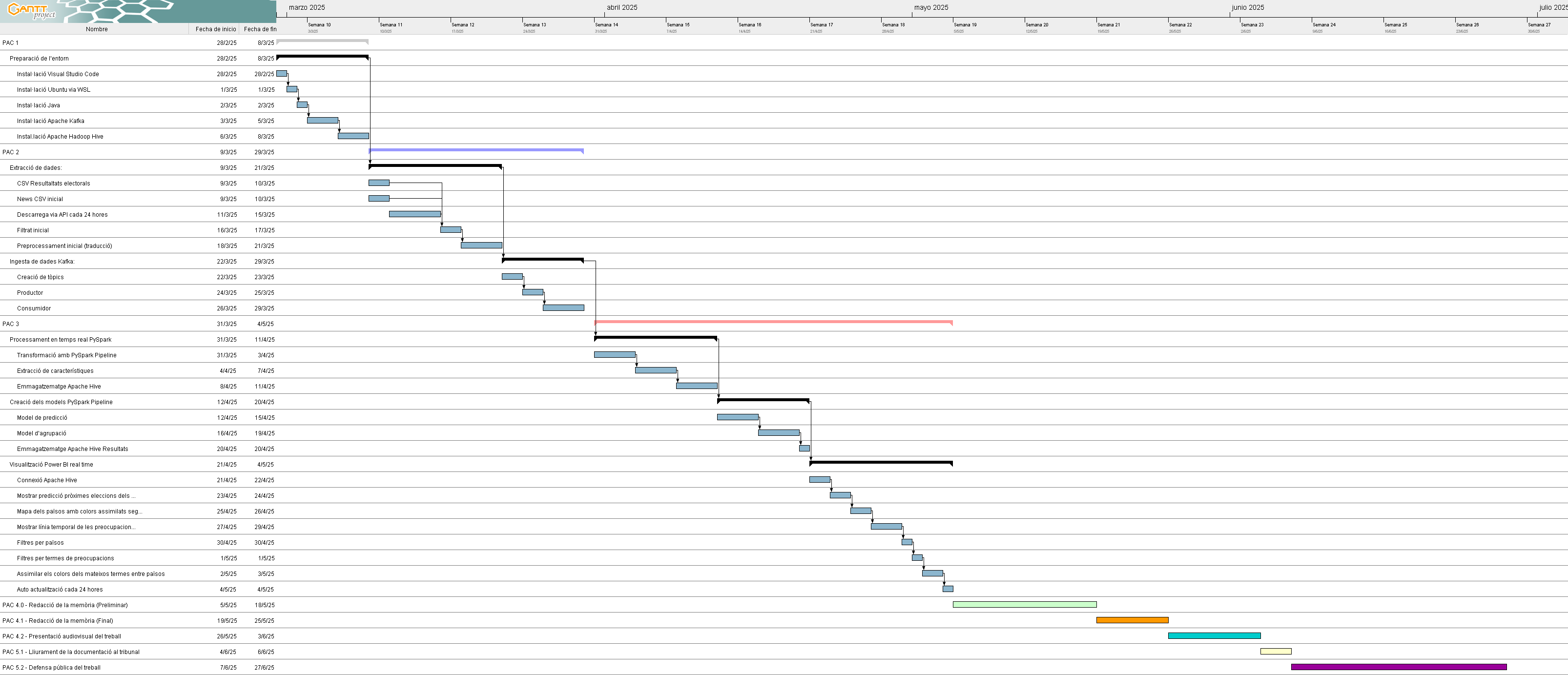
**PAC 4.1** - Redacció de la memòria (Final).

**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*



*Front: 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 no 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 relacíó 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.

............................ 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 erera 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 aquests capítols, cal descriure:

* els aspectes més rellevant del disseny i desenvolupament del treball
* la metodologia triada per a fer aquest desenvolupament, descrivint les alternatives possibles, les decisions preses, i els criteris utilitzats per prendre aquestes decisions.
* descripció dels productes obtinguts.

**L’estructuració dels capítols pot variar segons el tipus de treball.**

En cas que s’escaigui, s’inclourà un apartat de “Valoració econòmica del treball”. Aquest apartat indicarà les despeses associades al desenvolupament i manteniment del treball, així com els beneficis econòmics obtinguts i una anàlisi final sobre la viabilitat del producte.

................. SUBCAPITOLS ........... SELECCIÓ DE DADES I PROCESSAT INICIAL

SEGUIR EL ESQUEMA DE LA GUIA – APARTAT 4.1 I FER-HI REFERÈNCIA SI ESCAL

## El conjunt de dades

Els conjunt de dades principals, dels quals extraurem les característiques provenen de diferents interfície de programació d'aplicacions (API), de les que s’extrauran a diari les noticies dels diferents països per a ser processades, a mes de les descarregades com a 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.

Amb relació a les etiquetes, 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. 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.

# Resultats

Detalleu en aquest apartat els resultats obtinguts utilitzant la metodologia descrita a l’apartat anterior.

# 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.

# 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>

[x]

[x] Shapira, G. [Gwen], Palino, T. [Todd], Sivaram, R. [Rajini], Petty, K. [Krit]. (2024). Kafka: La Guía Definitiva, 2ª Edición. . O'Reilly Media, Inc.

<https://learning.oreilly.com/library/view/kafka-la-guia/9781098181673/>

[x] Tandon, A. [Akash], Ryza, S. [Sandy], Laserson, U. [Uri], Owen, S. [Sean], Wills, J. [Josh]. (2024). *Analítica avanzada con PySpark*. O'Reilly Media, Inc. <https://learning.oreilly.com/library/view/analitica-avanzada-con/9781098196844/>

[x] Thomas, A. [Alex]. (2020). *Natural Language Processing with Spark NLP*. O'Reilly Media, Inc.

<https://learning.oreilly.com/library/view/natural-language-processing/9781492047759/>

[x] Nokeri, T. C. [Tshepo Chris]. (2021). *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>

[x] Robert, J. [Johnson]. (2025*). Apache Hive Handbook: Query, Analyze, and Optimize Big Data*. HiTeX Press. <https://www.booktopia.com.au/apache-hive-handbook-robert-johnson/ebook/6610000731251.html?srsltid=AfmBOoqLsAoS2pg_d6Z6uPzTFH4sdmi-9WyhWqLZPdTMxQF5g2MKy6GG>

[x] 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

Llistat d’apartats que són massa extensos per incloure dins la memòria i tenen un caràcter auto contingut (per exemple, manuals d’usuari, manuals d’instal·lació, etc.)

Depenent del tipus de treball, és possible que no calgui afegir cap annex.