Mineria de dades dels mitjans socials, tècniques per a l'anàlisi de dades massives

PID_00278307

Jordi Morales i Gras

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

Jordi Morales i Gras

Doctor en Sociologia per la Universitat del País Basc. Professor d'Anàlisi de xarxes, *machine learning* i *big data*, i sociodirector de Network Outsight, empresa especialitzada en l'anàlisi sociològica de dades massives.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Andrea Rosales Climent

Primera edició: octubre 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Autoria: Jordi Morales i Gras
Producció: FUOC

Índex

Introducció

A hores d'ara ja coneixem molt bé el paradigma de les dades massives, la mineria de dades i la seva cadena de valor. Ja sabem que el més habitual serà que les dades hauran estat recollides i registrades per algú diferent a l'analista, i amb un propòsit diferent al seu. Això implica que cedim autonomia en el disseny de l'instrument de captura de dades i que, a canvi, obtenim un gran volum de dades. Ara bé, també sabem que el més important no és el volum de dades, sinó la capacitat d'interpretar-les per generar coneixement i intel·ligència.

Heus aquí la importància de totes les estratègies que ens permeten esprémer les dades i afegir-los valor. Una d'aquestes tècniques és l'anàlisi de les xarxes socials, que ens permet generar coneixement d'acord amb l'estructura relacional de les interaccions virtuals (això és, interaccions entre usuaris i interaccions entre usuaris i altres objectes o plataformes). En l'assignatura «Analítica avançada en xarxes socials» s'estudia aquesta tècnica en profunditat.

En aquest mòdul ens centrarem en una estratègia clau per a la mineria de dades i la generació de valor en entorns de dades massives, que són els algoritmes que generen coneixement d'acord amb els continguts dels mitjans socials. Per fer-ho, entrarem en el món de la intel·ligència artificial, l'aprenentatge automàtic i el processament del llenguatge natural. Veurem la diferència entre els algoritmes d'aprenentatge supervisat i no supervisat, i també coneixerem els algoritmes híbrids i de conjunt, que barregen les lògiques dels dos tipus anteriors. Posteriorment, ens centrarem en algunes de les tècniques específiques per a l'anàlisi de textos, un dels continguts més abundants en els mitjans socials, i veurem quin tipus de coneixement és possible elaborar a partir de les dades. Per acabar, ubicarem els algoritmes de dades massives en la dimensió normativa del tractament just de les dades, que comprèn tant la identificació de biaixos i discriminacions en els sistemes de predicció i classificació automatitzada com el disseny conscient i mitigador d'aquestes desigualtats.

1. Intel·ligència artificial i *machine learning*

Per **intel·ligència artificial** (IA) hem d'entendre qualsevol tècnica que permet a un ordinador dur a terme una o diverses accions que aparentin o emulin alguna de les dimensions de la intel·ligència humana.

No cal que una IA es presenti en forma de robot «humanoide» i que es faci passar per una persona amb sentiments i il·lusions per considerar que una màquina és intel·ligent. En realitat, estem molt i molt lluny d'aquest tipus d'IA, malgrat que la literatura i el cinema ja hi hagin arribat fa una pila d'anys. De fet, l'exemple més habitual que podem tenir d'IA al nostre entorn són els braços robòtics que s'utilitzen en la indústria des de les acaballes dels anys setanta. A aquestes IA les anomenem *IA dèbils*, i només són capaces de dur a terme unes quantes tasques «intel·ligents». Contrasten amb les *IA fortes*, capaces d'aprendre qualsevol tasca humana, malgrat que, ara com ara, només existeixen en la ficció.

Moltes IA dèbils tenen la capacitat d'aprendre i fer cada cop millor aquella tasca en què són expertes.

L'**aprenentatge automàtic** o *machine learning* és la subdisciplina de l'IA que persegueix la millora del propi sistema mitjançant l'entrenament i l'experiència (Gopinath Rebala i altres, 2019).

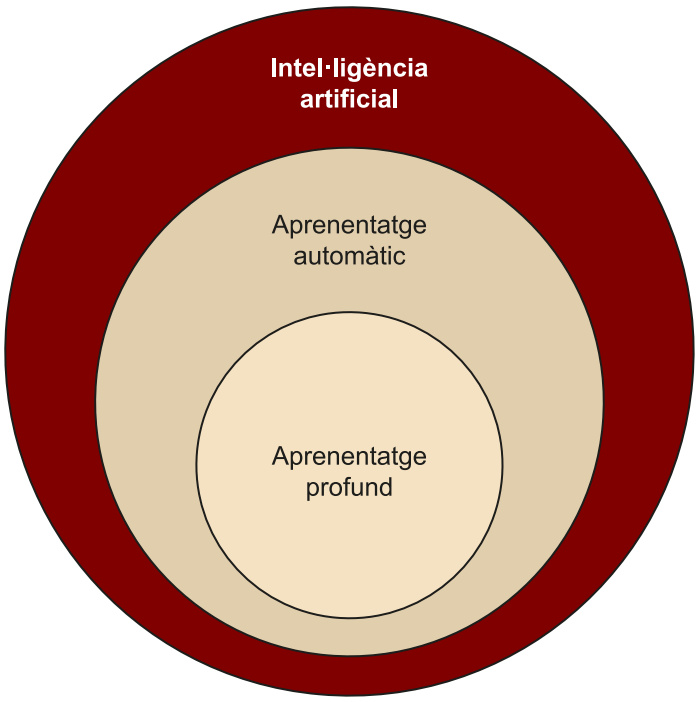
Això vol dir que, mitjançant l'aprenentatge automàtic, un ordinador no solament és capaç d'emular un comportament que sembli intel·ligent, sinó que és capaç de fer-ho cada cop amb més destresa i precisió. La majoria d'algoritmes que treballen amb dades massives formen part d'aquesta família d'algoritmes.

Per exemple, els algoritmes que ens proposen continguts a les xarxes socials en funció de les nostres amistats i de les nostres cerques, o els algoritmes que permeten dur a terme operacions de «remarketing» a plataformes com ara Google o Amazon, o fins i tot els filtres de contingut brossa més avançats, que són capaços d'adaptar-se a les particularitats de cada usuari, en funció del que aquest consideri correu brossa.

Tradicionalment, hi ha hagut dues tipologies d'algoritmes d'aprenentatge automàtic: l'**aprenentatge supervisat i el no supervisat**. Es tracta de tècniques preparades per resoldre problemes, com ara la classificació automàtica de casos, la predicció numèrica (això és, regressió) o de clusterització basant-se en les propietats intrínseques de les dades.

Avui, també hi ha una sèrie de **models híbrids** (per exemple, l'aprenentatge semisupervisat o l'aprenentatge reforçat) que combinen elements de totes dues tipologies i que aprofundeixen en el concepte d'autoaprenentatge per part de la mateixa màquina. En aquesta darrera categoria, trobem les tècniques de xarxes neuronals i d'aprenentatge profund o *deep learning* (vegeu figura 1), molt populars durant els últims anys i que actualment estan en fase de gran expansió i desenvolupament, tot i que també són tècniques discutides pel tipus de resultats que proporcionen i per la seva opacitat.

Figura 1. Intel·ligència artificial, aprenentatge automàtic i aprenentatge profund



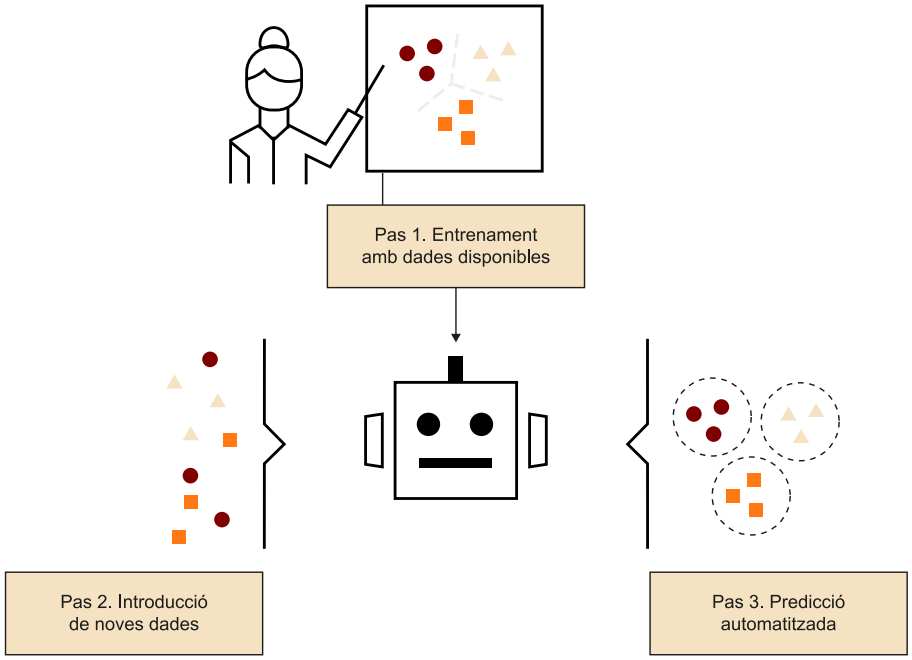
Font: elaboració pròpia.

1.1. Algoritmes d'aprenentatge supervisat

Els **algoritmes d'aprenentatge supervisat** persegueixen la predicció de resultats futurs en funció de sèries de dades conegudes.

L'entrenament d'un algoritme mitjançant un procés d'aprenentatge supervisat consisteix en el fet que l'analista proporcioni a l'ordinador tant les dades (això és, els conjunts de dades per entrenar i testar) com els resultats (això és, l'*output* esperat), i deixi que sigui l'ordinador el que elabori un model basant-se en alguna tècnica específica degudament parametritzada (això és, l'algoritme). És a dir, l'analista haurà d'alimentar el sistema amb una sèrie de casos coneguts i prèviament resolts, i la màquina haurà d'aprendre a resoldre'n de nous amb l'ajuda d'una seqüència d'operacions predefinida (vegeu figura 2).

Figura 2. Aprenentatge supervisat



Font: elaboració pròpia.

Els elements que intervenen en una tasca d'aprenentatge automàtic supervisat són tres, tot i que es poden presentar en diferents formats que requeriran modes d'articulació diferents:

1) En primer lloc, hi ha les **dades d'introducció**: les dades a partir de les quals l'algoritme haurà d'establir la predicció. Aquestes dades prenen diferents noms en la literatura científica, com ara «predictors», «característiques» (*features*), «variables d'entrada» o «variables independents».

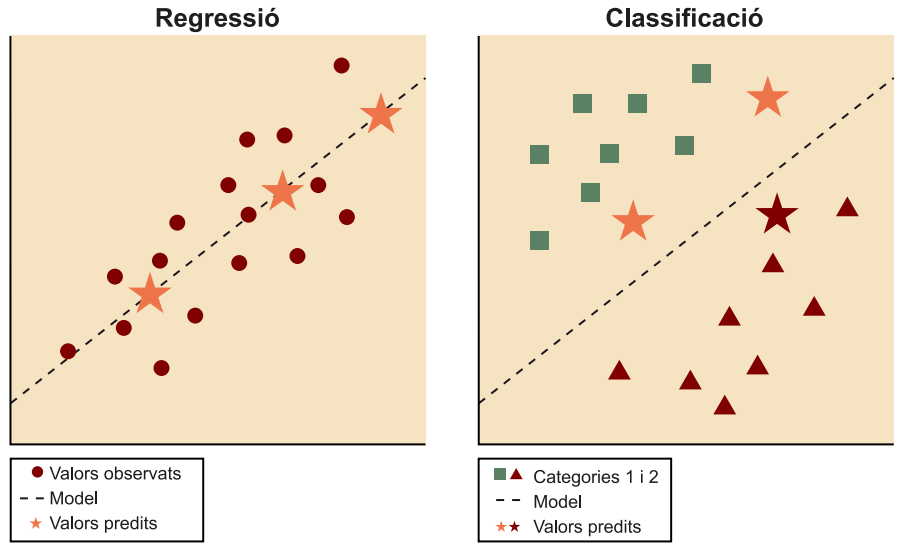
2) En segon lloc, hi ha el conjunt de **dades que volem predir**. Aquestes dades s'anomenen «objectiu» (*target*), «variable de sortida» o «variable dependent», i es correspondran amb la «variable categoria» (*label*) amb la qual s'haurà entrenat el model.

3) El tercer element és l'**algoritme d'aprenentatge automàtic** que utilitzarem per elaborar el model. Consisteix en un conjunt d'instruccions ordenades i amb un tipus de resultat previsible. En el context d'una tasca d'aprenentatge automàtic, a l'algoritme a vegades se l'anomena «aprenent» (*learner*) o «inductor».

Una de les tasques més habituals que haurà de dur a terme un algoritme és la classificació de casos, com per exemple, la separació dels correus desitjats del correu brossa. Per tal d'entrenar un ordinador per dur a terme la tasca classificatòria, caldrà comptar amb una base de dades amb correus amb tota la informació possible que pugui ajudar l'algoritme en el seu aprenentatge: text del correu, remitent, hora d'enviament, etcètera (això és, dades d'introducció). Addicionalment, haurem de proporcionar la resposta esperada per uns quants casos –com més millor, els algoritmes entrenats amb milers o milions de casos oferiran resultats més precisos– dividits entre correus desitjats o correus brossa (això és, dades objectiu). Finalment, haurem de decidir quin tipus d'algoritme d'aprenentatge fem servir per dur a terme l'entrenament i, posteriorment, per establir les prediccions (això és, algoritme inductor).

Els **algoritmes** són conjunts de tasques que ens ajuden a resoldre problemes matemàtics. En conseqüència, saber escollir i entrenar el millor algoritme possible per dur a terme una tasca determinada requereix entendre molt bé el tipus de problema matemàtic que cal resoldre. Un primer element que ho condiciona tot és el tipus de variable dependent. No és el mateix predir un nombre d'una sèrie d'una variable numèrica o quantitativa (per exemple, estimar el nombre de m'agrada que tindrà una publicació a Facebook) que predir una categoria d'una variable categòrica o qualitativa (per exemple, classificar un missatge segons el tema de què tracta). Per regla general, els problemes de predicció numèrica els resoldrem amb **models de regressió**, mentre que els problemes de predicció de categories (per exemple, el tema d'un missatge) els resoldrem amb **algoritmes classificatoris** (vegeu figura 3).

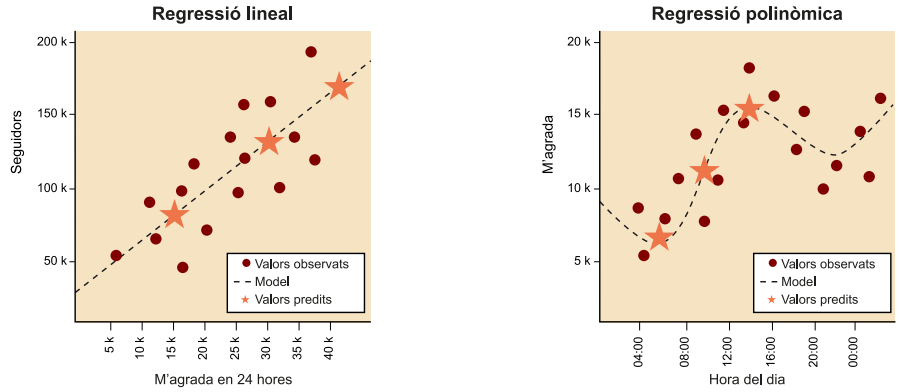
Figura 3. Regressió i classificació



Font: elaboració pròpia.

Un segon factor clau per escollir l'algoritme apropiat serà la pròpia distribució de les dades: el patró segons el qual les dades prenen els seus valors per mitjà de les variables de la base de dades. Quan ens trobem amb un problema de predicció numèrica o quantitativa, haurem de preguntar-nos quin és el tipus de funció que dibuixen les variables en qüestió. No és el mateix estimar un valor per a una funció lineal o per a una funció no lineal (vegeu figura 4). En el primer cas, més senzill, optarem per una **regressió lineal** (per exemple, predir els m'agrada en funció dels seguidors). En el segon cas, haurem de recórrer a mètodes més complexos, com pot ser una **regressió polinòmica** (per exemple, predir els m'agrada segons l'hora del dia).

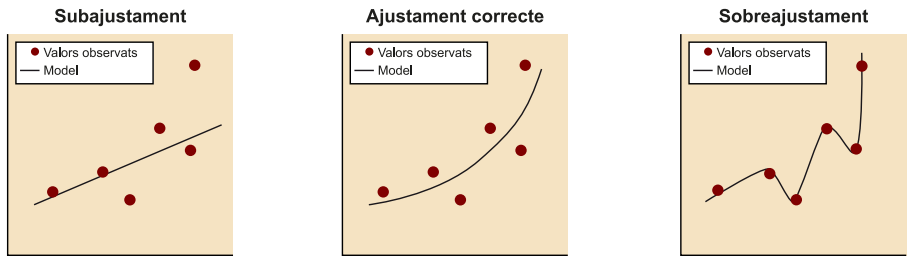
Figura 4. Regressió lineal i polinòmica



Font: elaboració pròpia.

Les regressions polinòmiques són més complexes d'implementar, però això no implica necessàriament que siguin millors ni que tinguin una major capacitat predictiva que les lineals. Un perill potencial d'aquest mètode és el sobreajustament (*overfitting*), que consisteix en el sobreentrenament d'un algoritme basant-se en unes dades particulars, que l'incapaciten per fer prediccions quan les dades són diferents. Per contra, el perill potencial de les regressions lineals és el subajustament (*underfitting*), que consisteix en la simplificació excessiva del model, comprometent així la seva capacitat predictiva. Aquests dos problemes (vegeu figura 5) són probablement les causes principals dels errors en l'aprenentatge automàtic i posen de manifest com n'és d'important la interpretació de les dades i les seves relacions.

Figura 5. Problemes de subajustament i sobreajustament

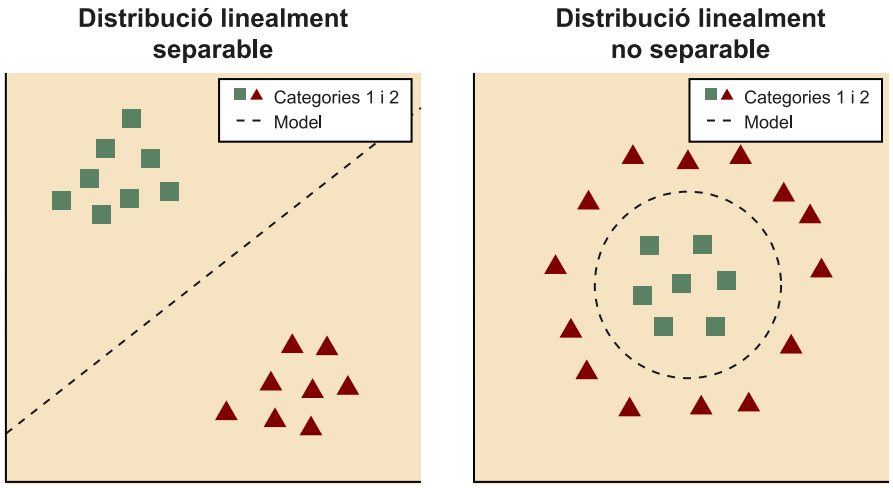


Font: elaboració pròpia.

En els problemes de predicció categòrica, quan volem classificar un cas nou en una o altra categoria d'una variable qualitativa, també haurem de distingir els casos que són linealment separables dels que no ho són (vegeu figura 6). Quan ens trobem amb el cas d'una distribució linealment separable (per exemple, si hem de categoritzar els missatges en funció del seu idioma) podem recórrer a algoritmes relativament simples, com ara la regressió logística o el classificador de veí més proper (això és, k-NN). En canvi, quan ens trobem amb distribucions linealment no separables, que sol ser l'habitual amb dades provinents dels mitjans socials (per exemple, si hem de discriminar el sentiment d'un text o decidir si un correu és un correu brossa o no), haurem de recórrer a algoritmes més complexos, com ara l'arbre de decisió (*decision tree*), o a algoritmes híbrids

–que incorporen elements no supervisats– com per exemple el bosc aleatori (*random forest*), l'algoritme d'impuls adaptatiu (*adaptive boosting* o *AdaBoost*) o les xarxes neuronals.

Figura 6. Categories linealment separables i no separables



Font: elaboració pròpia.

Per avaluar el **rendiment d'un algoritme** des d'un punt de vista matemàtic s'utilitzen tècniques estadístiques com ara la retenció (*holdout*) o la validació encreuada (*cross validation*).

Aquestes tècniques divideixen les dades en conjunts diferenciats que s'utilitzen com a dades d'entrenament o com a dades de prova, de manera que les dades d'entrenament serveixen per generar el model i les de prova per validar-lo. En realitat, la validació creuada és una evolució perfeccionada de la retenció. Mentre que el model de retenció divideix les dades només en dos blocs, en el model de validació encreuada les dades es divideixen en més grups i les proves són més robustes: s'obté més fiabilitat a canvi d'un major cost computacional.

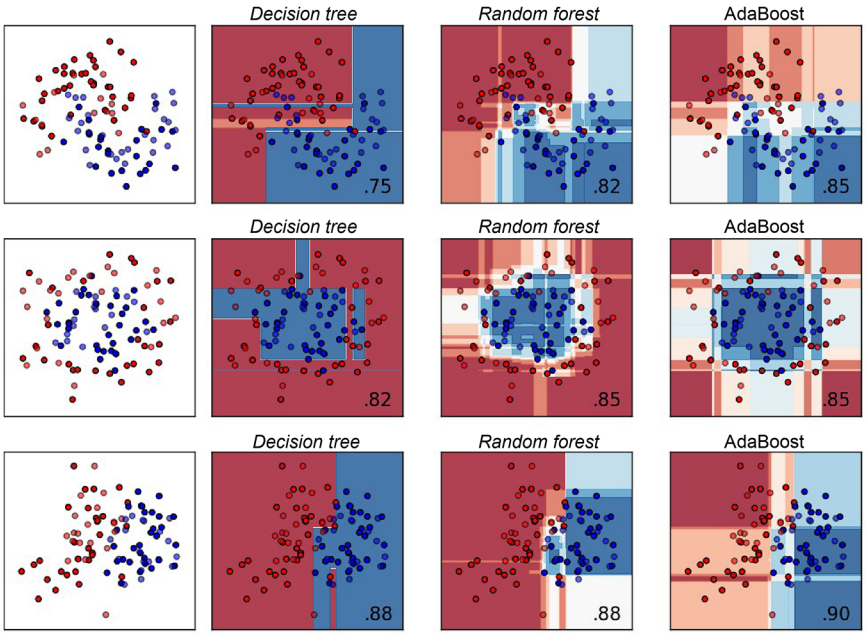
La principal mètrica que ens permet quantificar el **rendiment d'un model de regressió** és el coeficient de determinació (això és, R^2), que proporciona un valor entre –1 i 1 (això és, valors propers a 1 indiquen un bon ajustament de la predicció, valors propers a 0 indiquen absència de capacitat predictiva del model i valors propers a –1 indiquen que el model té una capacitat de predicció pitjor que una simple línia horitzontal) que ens informa de la capacitat del model per replicar els resultats obtinguts. En canvi, per quantificar el **rendiment d'un model de classificació**, farem servir mètriques com la precisió classificatòria (vegeu figura 7), que és el nombre mitjà de prediccions correctes fetes sobre les dades de prova, expressades com una porció del total. La mètrica obté valors entre 0 i 1 i ens informa sobre la precisió del model (això és, valors propers a 1 indiquen una capacitat predictiva propera al 100 % i a 0 % quan tendeixen a 0).

Els algoritmes que hem anomenat fins ara solament són uns quants dels molts que avui estan disponibles en el programari de *machine learning*. Molts d'aquests poden oferir solucions a problemes similars, en funció de les característiques del mateix problema (això és, el tipus de variable a predir i la distribució de les dades) i, per això, una pràctica recomanable és posar-ne diversos a competir entre si i avaluar quin és capaç d'oferir una millor solució a un mateix problema. Però el rendiment no ho és tot. L'analista farà bé d'escollir aquella solució que proporcioni uns bons resultats i, alhora, que en permeti una bona interpretació. Per exemple, i com en el cas de les prediccions numèriques, en les categòriques també ens podem trobar amb problemes de sobreajustament o subajustament. Aquests problemes adquireixen un major potencial quan més complex i difícil d'interpretar sigui un algoritme. El problema està, com veurem més endavant, en el fet que en el paradigma de l'aprenentatge automàtic, moltes vegades, a mesura que augmenta la capacitat predictiva, també augmenta l'opacitat algorítmica: com més complexos i ininterpretables són els algoritmes, més capacitat de predicció demostren.

Vegeu també

Vegeu aquestes qüestions en major profunditat en l'apartat «El tractament just de dades».

Figura 7. Precisió classificatòria de tres algoritmes per tres problemes



Font: https://martin-thoma.com/comparing-classifiers/.

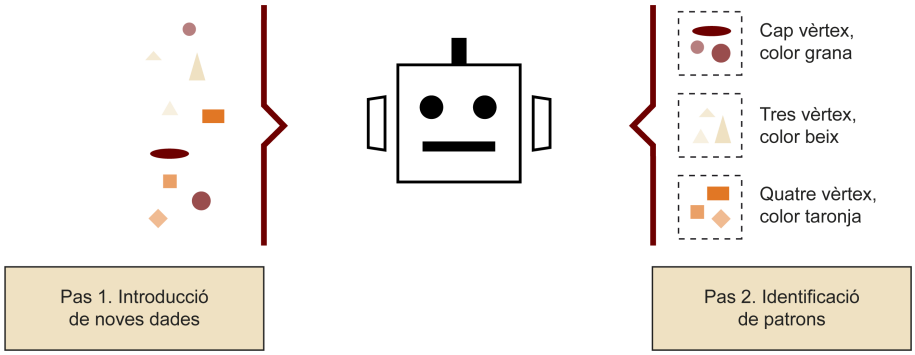
1.2. Algoritmes d'aprenentatge no supervisat

Una de les possibilitats més interessants del paradigma de les dades massives és que, gràcies al volum i a la capacitat computacional dels sistemes actuals, podem partir de models d'investigació inductius, en què l'observació va per davant de la teoria. Aquest és l'esperit que hi ha darrere dels algoritmes d'aprenentatge no supervisat, l'objectiu dels quals és classificar les dades en funció de les seves propietats intrínseques, sense partir d'un model predictiu preentrenat.

Els **algoritmes d'aprenentatge no supervisat** són capaços d'identificar patrons en les dades sense rebre instruccions específiques per part de l'analista.

Aquí no es tracta que l'investigador prepari l'ordinador per dur a terme una classificació o una predicció numèrica, sinó que sigui la mateixa màquina la que descobreixi patrons en les dades per si mateixa (vegeu figura 8) i que, d'aquesta manera, visibilitzi o faci emergir una sèrie de propietats de les dades que l'analista podria haver passat per alt.

Figura 8. Aprenentatge no supervisat



Font: elaboració pròpia.

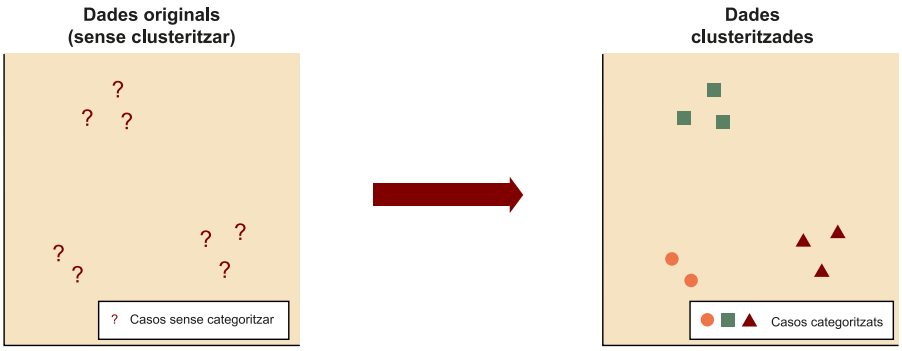
L'aprenentatge no supervisat parteix d'un tipus de raonament que els éssers humans coneixem molt bé, el **mètode inductiu**, que és una estratègia que forma part del mètode científic però que mai no ha estat, fins avui, l'estratègia més utilitzada pels científics. Tradicionalment, aquests han utilitzat el mètode deductiu. La diferència entre tots dos tipus de raonaments està en la relació entre la teoria i l'observació. Quan utilitzem el mètode inductiu, no partim d'una teoria o model de relació entre les variables que volem validar, sinó que partim de l'observació empírica i provem de sintetitzar una teoria o model. En la versió computacional d'aquest tipus de raonament hi ha un nombre més reduït d'elements que en els algoritmes d'aprenentatge supervisat, ja que no haurem de diferenciar entre les variables dependents i independents, ni entre dades d'entrenament i dades de prova. Per una banda, tindrem les dades que voldrem ordenar o simplificar i, per altra banda, tindrem l'algoritme que durà a terme la tasca en qüestió.

Per escollir el millor algoritme d'aprenentatge no supervisat, com en el cas de l'aprenentatge supervisat, haurem d'entendre molt bé quin és el problema matemàtic que volem resoldre i quin tipus de dades hi intervenen. La tasca més senzilla que podem dur a terme mitjançant un procés no supervisat consisteix a ordenar les dades i agrupar els casos similars. Aquesta tasca s'anomena **clusterització** (vegeu figura 9) i podem dur-la a terme mitjançant tècniques com l'algoritme k-Means o l'algoritme Louvain multinivell, que és enormement eficient en dades reticulars o de xarxes. Altres tasques més complexes que es poden resoldre amb aquest tipus d'algoritmes són la **reducció dimen-**

sional, que voldrem dur a terme quan ens trobem amb moltes variables o dimensions, o també la **resolució de problemes de regles** i **associació entre variables**. Tècniques com l'anàlisi de components principals (això és, PCA) o l'algoritme Eclat poden ajudar-nos a dur a terme aquestes tasques.

Tant important com el problema matemàtic o la tasca a resoldre de manera no supervisada és el tipus de dades amb què comptem, especialment en el paradigma de les dades massives, caracteritzat per una enorme variabilitat de formats. Avui en dia, hi ha algoritmes que ens ajuden a classificar dades de qualsevol tipus: textos, imatges, vídeos, àudios, etc. En el cas dels corpus documentals, són molt importants els algoritmes de **modelatge temàtic** (*topic modelling*), que identifiquen grups de paraules utilitzades conjuntament en textos (això és, temes). Pel que fa a les imatges, vídeos i àudios, el més habitual és identificar-hi els patrons de manera no supervisada després d'haver-los transformat en vectors numèrics mitjançant algoritmes híbrids anomenats ***encaixos*** (*embeddings*).

Figura 9. Clusterització (no supervisada)



Font: elaboració pròpia.

El procediment d'avaluació dels resultats obtinguts per un algoritme d'aprenentatge no supervisat és força diferent a la d'un algoritme supervisat, perquè no implica tècniques de validació creuada.

En les tasques de clusterització, és habitual comprovar com de correctament s'han classificat els casos, particularment i en conjunt, mitjançant mètriques com el **valor «silhouette»** o l'**estadístic de modularitat**.

Totes dues mètriques proporcionen una xifra entre –1 i 1 que ens informa de com de ben agrupats estan els casos en les categories identificades (això és, valors propers a 1 indiquen que els casos agrupats s'assemblen entre si i són molt diferents de la resta, valors propers a 0 indiquen que els casos agrupats s'assemblen tant entre si com amb la resta, i valors propers a –1 indiquen que els casos agrupats s'assemblen més a la resta que entre si). Tanmateix, és important entendre que no es tracta de mètriques de validació pures, ja que molts algoritmes les incorporen com a procediments d'optimització dels

mateixos processos. Aquesta diferència, lluny de ser una debilitat del mètode inductiu, és una de les seves principals virtuts, ja que permeten que l'analista es recolzi en mètriques en la construcció de les categories analítiques.

Totes les virtuts de l'aprenentatge no supervisat cal contrastar-les amb les seves limitacions. Per començar, és un error pensar que aquest tipus de models fan obsoletes les teories científiques, tal com han afirmat alguns observadors excessivament optimistes (Chris Anderson, 2008). L'aprenentatge no supervisat actualment disponible és un molt bon assistent per a l'elaboració teòrica, que acompanya i ajuda l'analista, però en cap cas substitueix les seves funcions. Moltes vegades, tal com veurem més endavant, els algoritmes d'aprenentatge no supervisat proporcionen resultats massa evidents o redundants. En el pitjor dels casos, els errors metodològics comesos en les fases de preparació de les dades o la falta de supervisió humana poden conduir al fet que una màquina adquireixi comportaments discriminatoris i indesitjables.

Un bon exemple és l'algoritme de selecció de personal d'Amazon que discriminava les dones, i un altre, el *chatbot* Tay de Microsoft que, a Twitter, va adquirir un llenguatge racista i una «ideologia» neonazi en només dos dies d'autoaprenentatge basat en converses de joves entre divuit i vint-i-quatre anys (Metz, 2016). Per tot això, és important entendre que es tracta de processos que poden requerir un esforç interpretatiu elevat i que moltes vegades seran difícils de sistematitzar.

Vegeu també

Vegeu el mòdul «Dades massives i mineria de dades socials, conceptes i eines bàsiques».

1.3. Algoritmes híbrids i de conjunt

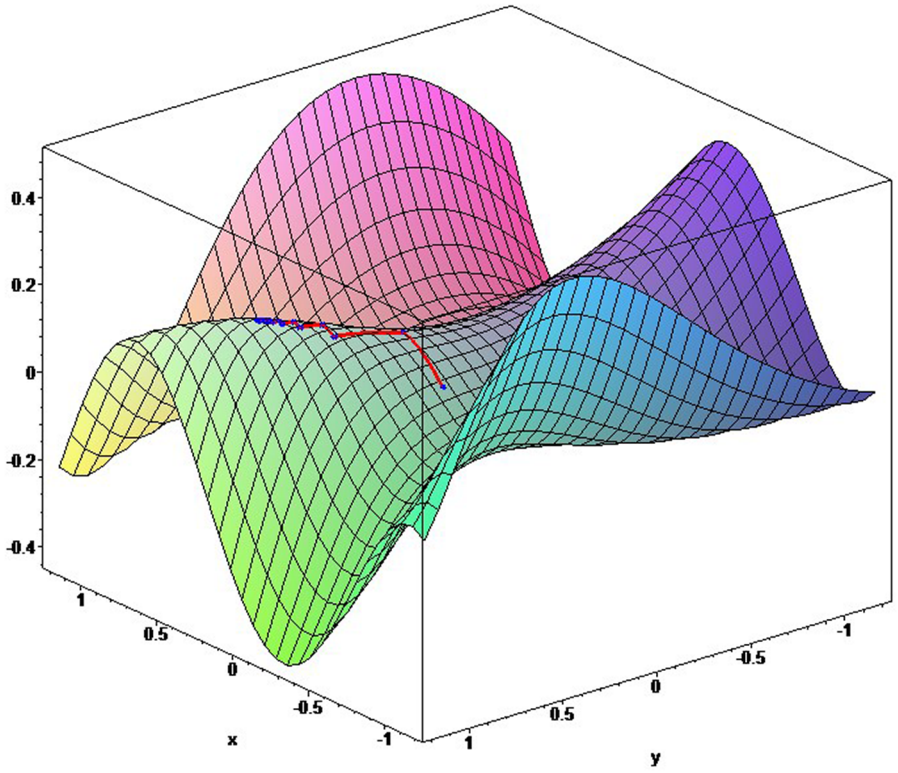
Els **algoritmes híbrids i de conjunt** combinen diversos aspectes dels algoritmes que hem vist anteriorment, amb l'objectiu de millorar el poder predictiu d'un model.

Es tracta d'opcions molt interessants quan es vol predir un resultat en funció de diferents i múltiples causes.

Per exemple, per predir els m'agrada d'una publicació en funció del tema, la longitud del text, el seu contingut emocional, l'hora del dia, el tipus d'imatge o vídeo, la seva tonalitat cromàtica, etc., o qualsevol altre aspecte mesurable.

Algunes d'aquestes tècniques són els algoritmes de **gradient descendent** (*gradient descent*), d'**impuls adaptatiu** (*adaptive boosting* o AdaBoost) i de **boscos aleatoris** (*random forests*). També els algoritmes més populars dels últims anys, les **xarxes neuronals profundes**, formen part d'aquesta categoria. La principal virtut d'aquestes tècniques és la seva capacitat per establir correlacions entre variables en múltiples dimensions, fet que els permet assolir una gran capacitat predictiva, però que en dificulta enormement la interpretació per part de l'analista (vegeu figura 10).

Figura 10. Representació d'una tasca multidimensional resolta per un algoritme de gradient descendent



Font: adaptada de Wikicommons (imatge lliure de drets).

La funció principal i més explotada d'aquest tipus d'algoritmes és la **resolució de problemes clàssics de categorització o de predicció numèrica** mitjançant la incorporació d'elements d'autoaprenentatge no supervisat. És a dir, l'algoritme parteix d'un entrenament implementat per l'analista, però és capaç de seguir aprenent de manera autònoma d'acord amb els seus propis èxits i fracassos. Com ja hem vist, aquest tipus de procediments corren el perill d'entrenar-se basant-se en dades esbiaixades i de produir resultats igualment esbiaixats i de mala qualitat, de vegades, inclús, discriminatoris: heus aquí els casos ja esmentats de l'algoritme d'Amazon per a la selecció de personal o el *chatbot* racista de Microsoft.

Durant la darrera dècada, el camp d'estudi més prometedor de l'aprenentatge automàtic han estat els **algoritmes de xarxes neuronals i d'aprenentatge profund**. Es tracta d'algoritmes de caixa negra, totalment orientats a la predicció, i que no pretenen, en cap cas, ser utilitzats per un analista que pretengui entendre o comprendre la relació entre les variables d'un model. És en aquest punt on l'aprenentatge automàtic pren major distància respecte de l'estadística inferencial, que té com a objectiu principal la formalització i la interpretació de les relacions entre les variables. Una altra característica de l'aprenentatge profund és el fet de desentendre's totalment del principi de parsimònia, que és reemplaçat per la capacitat de computació.

Principi de parsimònia i capacitat de computació

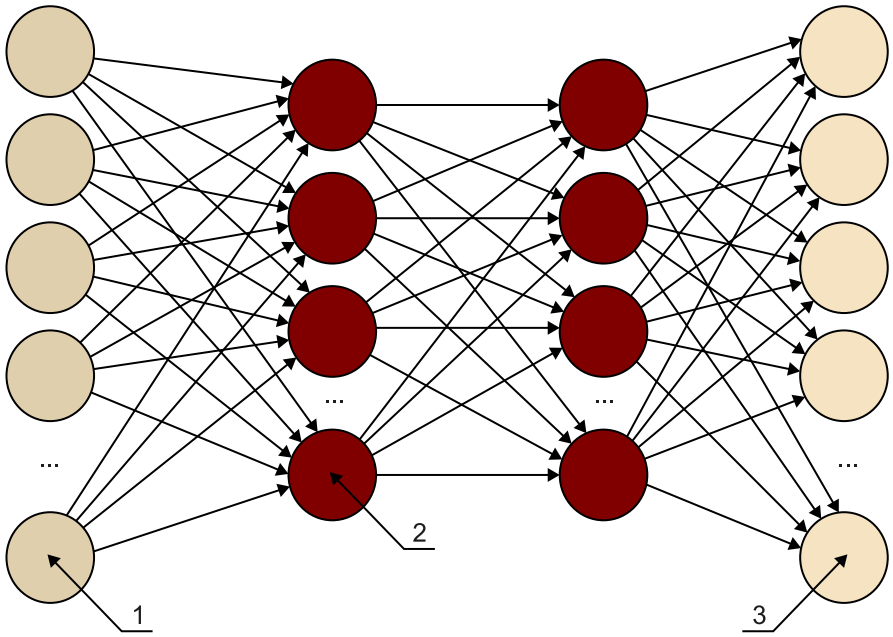
El **principi de parsimònia** és el principi segons el qual les explicacions senzilles de poques variables són superiors a les complexes si tenen capacitats explicatives semblants.

Capacitat de computació: és poc important el pes específic d'una o altra variable mentre la màquina ho pugui processar en un temps assumible.

L'estudi de les relacions entre les variables i els seus pesos específics que anhela l'estadística inferencial resulta del tot impossible quan ens trobem davant d'un model de caixa negra. Com el seu nom indica, els models de caixa negra

consisteixen en processos interns que elaboren les màquines i que no poden ser interpretats per l'analista. Les xarxes neuronals profundes (vegeu figura 11) estableixen diferents nivells de coneixement de manera oculta, de manera que l'analista només té accés a l'*input* i l'*output* del sistema. Com que són algoritmes que aprenen sols, són capaços d'assolir una enorme capacitat predictiva amb el temps, però, per altra banda, són algoritmes força complicats de parametritzar, i encara més difícils d'analitzar, i que tendeixen a assumir i integrar biaixos no corregits en les dades que utilitzen per entrenar-se.

Figura 11. Model d'una xarxa neuronal profunda



Font: Wikicommons (imatge lliure de drets).

Els algoritmes de caixa negra plantegen problemes als analistes de dades, però també presenten problemes importants d'opacitat i poden, fins i tot, adquirir comportaments discriminatoris, com el cas de l'algoritme entrenat per Amazon per a la selecció de personal, que vàrem veure al mòdul 1, o com el *chatbot* Tay de Microsoft. Cada cop són més els experts en intel·ligència artificial que recomanen no utilitzar-los (Alex Campolo i altres, 2017). L'alternativa als algoritmes de caixa negra són els **algoritmes anomenats de «caixa blanca» o transparents**: algoritmes fàcils d'interpretar i que permeten explorar les relacions establertes entre les variables, com pretén l'estadística. En realitat, més que de blancs i negres, es tracta d'una escala de grisos: mentre que les regressions i els arbres de decisions són més o menys senzills d'interpretar, la complexitat creix quan es tracta d'interpretar algoritmes *random forest* o AdaBoost, especialment adequats per a la classificació de distribucions no lineals.

Per tant, decidir quin algoritme utilitzar depèn de diversos factors. Com hem vist, és important tenir en compte criteris estrictament matemàtics i vinculats al tipus de problema que volem resoldre: el tipus de dades, la seva distribució o el seus formats. També dependrà, en gran mesura, del plantejament metodològic d'una investigació: l'aprenentatge supervisat és útil per als plantejaments hipoteticodeductius i el no supervisat per als inductius. Tanmateix, ca-

da cop hi ha més veus que reclamen que el rendiment i la capacitat predictiva no haurien de ser les úniques variables en joc. També és important que els algoritmes siguin transparents i interpretables, ja que el fet que no ho siguin compromet totalment l'objectiu final de qualsevol investigació, que no és cap altre que la creació de valor. Un cop més, és responsabilitat de l'analista mantenir un equilibri entre la transparència (això és, la facilitat d'interpretació dels resultats) i la precisió (això és, la capacitat predictiva) dels models.

2. Algoritmes per al processament del llenguatge natural

Els continguts dels mitjans socials són, conjuntament amb les relacions que s'hi estableixen, una enorme font de dades massives. Entre els formats de publicació, a més dels textos, hi ha imatges, vídeos o àudios.

La subdisciplina de la intel·ligència artificial, la informàtica i la lingüística que s'ocupa de l'anàlisi de textos i documents és el **processament del llenguatge natural** (PLN).

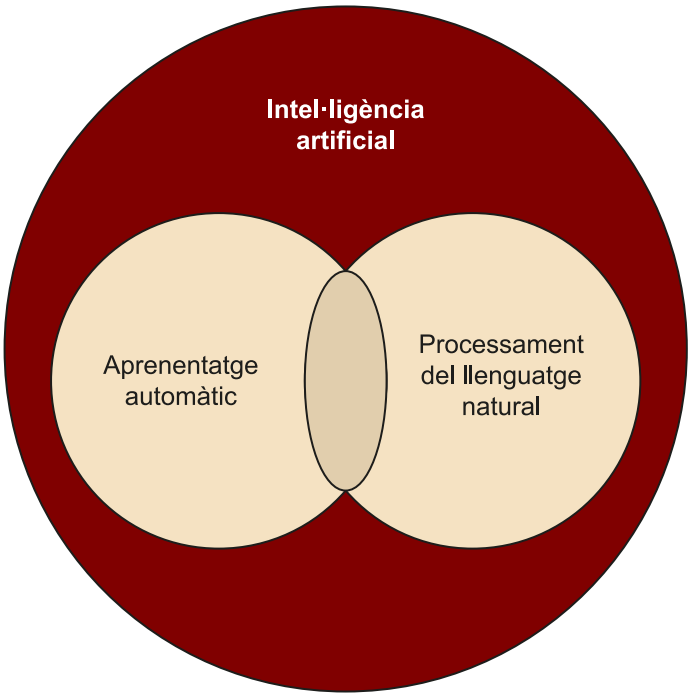
Es tracta d'un camp d'estudi molt ampli que cobreix tasques que van de des de la traducció automàtica o la predicció de textos fins a les vinculades a l'avaluació de discursos (Steven Bird i altres, 2009). El millor exemple d'aquest tipus de tècniques són les funcionalitats de traducció i de text predictiu de cercadors com ara Google o Bing, que els darrers anys han millorat enormement.

Es pot dir que el camp d'anàlisi del PLN és tan ampli i multidimensional com ho és el mateix llenguatge humà i, per tant, es nodreix d'elements tant diferents com la morfologia, la gramàtica, el lèxic, la semàntica, la pragmàtica, l'entonació, la fonètica, etc. El PLN és un camp de l'IA, però no de l'aprenentatge automàtic (vegeu figura 12). Això és així perquè el PLN comprèn processos i algoritmes de diferent naturalesa i no tots impliquen models en què la màquina aprèn basant-se en un entrenament o per si mateixa. Per una banda, és important distingir les aproximacions basades en **regles heurístiques** (això és, seqüències d'operacions que proporcionen solucions «suficientment bones» per als objectius del procediment i en un temps curt) i, per l'altra, **solucions d'aprenentatge automàtic**: supervisat, no supervisat i híbrid.

Nota

En aquest mòdul no aprofundirem en les tècniques d'anàlisi específiques d'aquests formats –que generalment s'analitzen després de ser convertits en vectors numèrics anomenats encaixos o *embeds*, mitjançant tècniques semblants a algunes de les que veurem aquí– sinó que ens centrarem en els continguts escrits, que són, sens dubte, els més abundants i transversals, ja que els trobem a totes les plataformes.

Figura 12. IA, aprenentatge automàtic i PLN



Font: elaboració pròpia.

2.1. Processos de les regles heurístiques

Les tasques més senzilles que es duen a terme en un PLN són d'ordre sintàctic i lèxic, i solen ser subprocessos o passos preparatoris d'algoritmes més complexos. Es tracta d'operacions computacionalment molt eficients que es duen a terme basant-se en regles preestablertes i que, per tant, no impliquen un procés d'aprenentatge automàtic com a tal. En aquests casos, la màquina disposa d'una guia o d'un diccionari dissenyat manualment, identifica cada cas o situació i aplica la solució manualment predefinida. Les següents són algunes de les **operacions més habituals** d'un PLN basades en regles heurístiques:

1) Tokenització. És la segmentació del text en parts més petites anomenades *tokens* (això és, normalment paraules, però també poden ser frases, paràgrafs o altres parts d'un text), segons uns espais i signes de puntuació. En la majoria de llengües del nostre entorn (això és, romàniques) aquesta tasca no implica cap dificultat. La dificultat augmenta en llengües amb característiques aglutinants com ara el japonès, el finès o l'èuscar.

2) Filtratge. Conjuntament amb la tokenització, és habitual dur a terme un procés d'eliminació de paraules innecessàries o inadequades per a l'anàlisi. Les anomenades *paraules buides* són les més comunes d'una llengua, com ara les preposicions i els articles. És convenient que l'analista conegui bé el llistat de paraules que utilitza, ja que no hi ha cap criteri unificat i pot ser que una

paraula sigui innecessària en un context i necessària en un altre. El procés de filtratge també es pot aplicar en positiu, deixant que només un conjunt determinat de paraules passi el filtre.

3) Lematització. Consisteix a associar totes les paraules d'un text al seu lema o forma canònica, prèviament indexada (per exemple, «serem», «éreu», «ets» = «ser»). El procés heurístic implica una anàlisi morfològica de cada paraula i un diccionari detallat de l'idioma o idiomes en què està escrit el text.

4) Bossa de paraules. És un mètode que es pot aplicar després de la tokenització, el filtratge i la lematització, i que consisteix a agrupar les paraules en «bosses», de manera que cada text queda associat a les diferents paraules que conté, ignorant el seu ordre original. Típicament, les matrius de dades processades amb aquest mètode tenen tantes files com peces de text i tantes columnes com paraules.

5) Valor TF-IDF. És una tècnica similar a la bossa de paraules, però que pondera cada paraula en funció del nombre de cops que apareix en un document en relació amb la presència de la paraula en el conjunt de documents. D'aquesta manera, les paraules més idiosincràtiques de cada document (això és, les que el distingeixen de la resta de documents d'un corpus) prenen més protagonisme. És una tècnica molt útil si es vol classificar documents maximitzant les seves diferències.

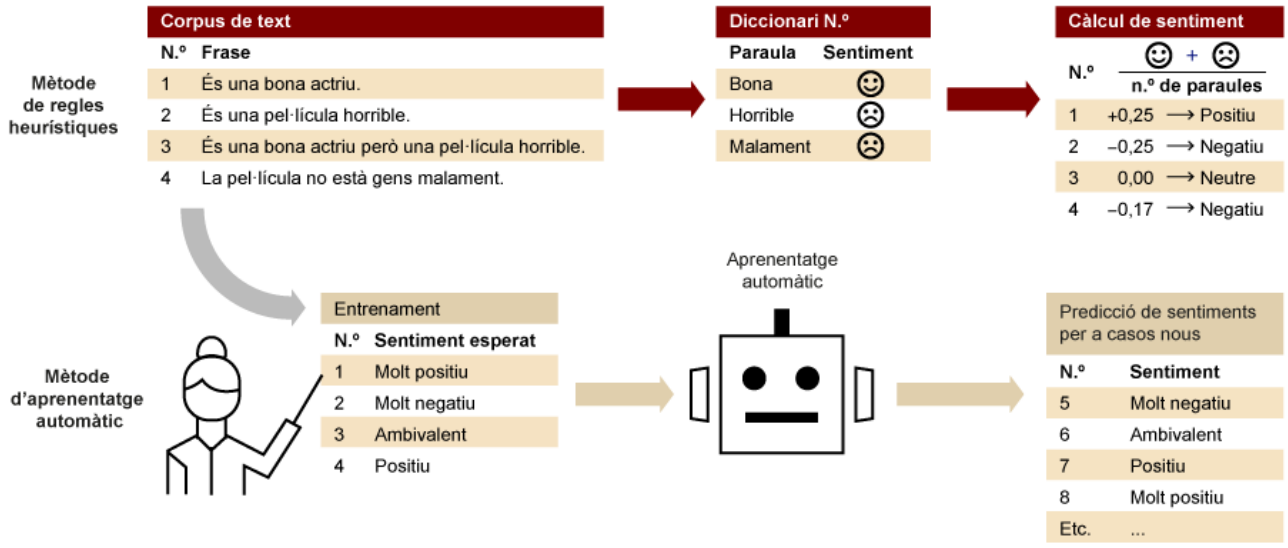
Les tècniques anteriors són algunes de les més habituals en qualsevol anàlisi de textos, atès que constitueixen precondicions per a multitud d'anàlisis. No sempre serà recomanable aplicar-les, ni fer-ho de la mateixa manera. Per exemple, en textos molt curts (per exemple, piulades) moltes vegades serà preferible no lematitzar les paraules, ja que molts algoritmes necessiten comptar amb una certa diversitat morfològica per tenir un bon rendiment, especialment els algoritmes no supervisats i els d'aprenentatge profund. A més a més, algunes d'aquestes tècniques són mútuament excloents, com ara la bossa de paraules i la tècnica TF-IDF.

Una de les tècniques més populars del PLN, l'**anàlisi de sentiment**, és un altre exemple de procediment basat en regles heurístiques. En realitat, el seu funcionament és molt senzill: es parteix d'un llistat de paraules anomenades *positives* i d'un llistat de paraules anomenades *negatives*, i es procedeix a fer un recompte sobre el corpus del text en qüestió. El text es considerarà «positiu» o «negatiu» en funció del tipus de paraules més abundants, i «neutre» quan hi hagi un empat. Aquest tipus d'anàlisi pot ser molt útil en els mitjans socials, ja que ens permet automatitzar la detecció de crisis reputacionals, o aproximar-nos a l'opinió del conjunt d'una audiència sobre una marca o producte. Ara bé, és un procediment amb limitacions molt importants, sobretot vincu-

lades al context (per exemple, l'adjectiu «freda» podrà ser adequat per a una beguda i inadequat per a una pizza) o als elements pragmàtics del llenguatge (per exemple, la ironia o el sarcasme).

Durant els últims anys, s'han desenvolupat diverses aproximacions a l'anàlisi de sentiment i a altres problemes clàssics del PLN des de paradigmes algorítmics d'aprenentatge supervisat i híbrid, especialment des de les xarxes neuronals profundes (*deep learning*). Aquests models han ofert millores substancials respecte dels models basats en regles heurístiques (vegeu figura 13). També s'han desenvolupat recentment models mixtes, que combinen les dues aproximacions i ofereixen resultats força bons (Paramita Ray i Chakrabartib, 2019). De tota manera, en matèria d'anàlisi de sentiment encara estem lluny de poder dissenyar procediments i algoritmes que proporcionin resultats homologables als que obté la cognició humana. Es tracta d'una àrea del PLN força subdesenvolupada, sobretot si la comparem amb els enormes avenços dels últims anys en altres problemes clàssics, com ara les tasques de traducció de textos o de predicció de paraules.

Figura 13. Anàlisi de sentiment amb regles heurístiques i amb aprenentatge automàtic



Font: elaboració pròpia.

2.2. Aprenentatge supervisat aplicat a textos

La millor manera d'obtenir bons resultats amb un algoritme és entrenar-lo específicament perquè dugui a terme la tasca o les tasques que volem que faci. D'aquesta manera, mitjançant models específics d'aprenentatge, és possible superar aspectes com pot ser la forta dependència contextual del llenguatge (per exemple, una beguda «freda» serà probablement adequada, i una pizza «freda» serà probablement inadequada). Moltes de les limitacions dels sistemes basats en regles heurístiques es poden superar amb models d'aprenentatge supervisat, com en el cas de l'anàlisi de sentiment. Al cap i a la fi, identificar el sentiment d'una frase, o identificar si conté un tipus de llenguatge, o si pertany a una llista predeterminada de temes és un problema clàssic de catego-

rització, que es pot resoldre mitjançant un algoritme de l'estil d'una regressió logística o d'un arbre de decisió. La flexibilitat dels algoritmes d'aprenentatge supervisat permeten que qualsevol analista pugui crear el seu propi algoritme per categoritzar textos nous, si es disposa de suficients textos prèviament categoritzats.

Els **elements necessaris** per crear un algoritme de classificació de textos no són tan diferents dels que podem aplicar en qualsevol problema a resoldre mitjançant l'aprenentatge supervisat:

1) Les **dades d'introducció** seran els textos a partir dels quals s'entrenarà l'algoritme. Poden ser textos llargs (per exemple, novel·les, notícies, webs, articles de la Viquipèdia, etc.) o curts (per exemple, publicacions de Facebook, piulades, etc.), i constituiran la matèria primera per crear les nostres variables d'entrada (això és, el corpus). Generalment, aplicarem els processos de les regles heurístiques sobre el corpus per generar les variables independents del nostre model, com ara les paraules contingudes a cada text (per exemple, model de bossa de paraules o TF-IDF).

2) Les **dades a predir** podran ser de dos tipus: quantitatives o qualitatives. Per exemple, es pot provar de predir el nombre de compartits o m'agrada d'una publicació de Facebook (això és, variable numèrica) en funció de les paraules que hi apareixen; o es pot predir l'autor d'un llibre o la temàtica (això és, variable categòrica) en funció del mateix criteri. En qualsevol cas, l'analista haurà de proporcionar les respostes a una sèrie de casos per tal d'entrenar l'ordinador, creant així un algoritme únic.

3) Finalment, l'**algoritme inductor** utilitzat haurà d'ajustar-se als estàndards de la predicció numèrica o categòrica (per exemple, regressió lineal, logística, arbre de decisions, *random forest*, etc.) i obtenir un bon rendiment en les proves de validació creuada. El més recomanable és entrenar-ne diversos i quedar-se amb aquell que proporciona uns millors resultats, també tenint en compte el delicat equilibri entre la capacitat predictiva i la transparència desitjable en aquest tipus d'algoritmes.

A causa de la gran complexitat del llenguatge humà, els algoritmes com les xarxes neuronals profundes, que estableixen relacions entre variables (per exemple, paraules, posicions en la frase, distància respecte a altres paraules, etc.) a diferents nivells de profunditat, han demostrat una capacitat predictiva molt superior a algoritmes més senzills. El preu a pagar és una parametrització complicada i una elevada opacitat en la interpretació dels seus resultats, fet que dificulta la fiscalització del rendiment de l'algoritme per part de l'analista. És important que tots aquests elements estiguin sobre la taula, conjuntament amb el cas d'anàlisi particular, en el moment que l'analista decideix decantar-se per una o altra tècnica.

2.3. Aprenentatge no supervisat aplicat a textos

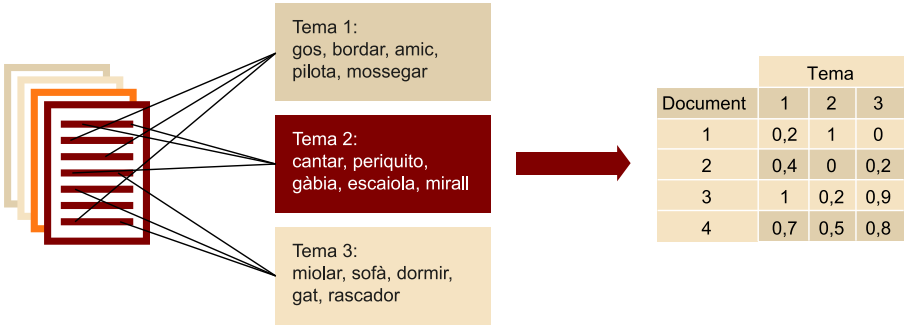
Pel que fa als models d'aprenentatge no supervisat cal destacar, per una banda, que molts models generals es poden aplicar a l'anàlisi del corpus dels documents. Per exemple, cada cop més analistes i investigadors apliquen tècniques com ara l'**anàlisi de xarxes** per identificar grups de paraules en documents (Martin Gerlach i altres, 2018). Per altra banda, també hi ha algoritmes no supervisats propis del PLN, com ara les tècniques de **modelatge temàtic** (*topic modelling*) basades en la distribució probabilística de Dirichlet (David Andrzejewski i altres, 2009).

Matriu de distàncies

Una base de dades de doble entrada que proporciona una mètrica sobre les relacions entre els seus elements.

Per dur a terme aquest tipus de procés, cal processar els textos mitjançant tècniques basades en regles heurístiques que permetin disposar de variables per aplicar-hi els algoritmes no supervisats. Típicament, el que voldrem és generar una base de dades que es pugui interpretar com una matriu de distàncies i, així, poder estudiar les relacions entre les paraules i agrupar-les en funció d'una sèrie de criteris. Els algoritmes de modelatge temàtic identifiquen grups de paraules que s'utilitzen de manera conjunta en documents o en grups de documents. Per exemple, podem utilitzar-los per identificar els temes més importants d'una conversa o per descobrir grups de missatges semblants emesos per usuaris diferents. Aquest tipus d'algoritmes assumeixen que cada document és una col·lecció de temes, i que cada tema és una col·lecció de paraules. D'aquesta manera, s'identifiquen els temes en els conjunts de documents i s'obté una xifra que quantifica la vinculació de cada document amb cada tema (vegeu figura 14).

Figura 14. Modelatge temàtic



Font: elaboració pròpia.

Actualment, hi ha disponibles una gran varietat d'algoritmes de modelatge temàtic. Entre els més populars hi ha l'**assignació latent de Dirichlet** (LDA) o l'**anàlisi semàntica latent** (LSA o LSI). Es tracta d'algoritmes que proposen solucions diferents per a problemes classificatoris similars. Mentre que el model LDA només agrupa les paraules utilitzades conjuntament, el model LSA identifica aquelles que mai s'utilitzen de manera conjunta. L'elecció d'un o altre algoritme dependrà d'una quantitat important de factors que poden alterar el seu resultat i que poden ser complexos de controlar, com ara la longitud dels textos o la seva diversitat semàntica i temàtica. Segons les característiques del

corpus documental, caldrà prendre decisions sobre el preprocessament (per exemple, lematitzar sí o no, bossa de paraules o TF-IDF) o inclús sobre el nombre òptim de temes a identificar.

L'avaluació del **rendiment d'un modelatge temàtic** s'ha de dur a terme considerant una sèrie de qüestions diferents, tant d'ordre matemàtic com interpretatiu i pràctic. Mètriques com la coherència temàtica (això és, una xifra entre el 0 i l'1 que quantifica del grau de similitud semàntica que hi ha entre les paraules més vinculades a cada tema) o la probabilitat temàtica marginal (això és, una xifra amb una magnitud que varia segons el procediment utilitzat, que quantifica com de representat està un tema en un text) poden servir de guia per a l'analista, i per ajudar-lo a decidir quin procediment és més adequat i quin resultat és millor. Altres algoritmes com ara l'anàlisi de components principals o l'algoritme Louvain multinivell també poden ajudar a determinar el nombre òptim de comunitats. Però, en qualsevol cas, cal que l'analista interpreti els resultats i els avaluï des d'una perspectiva aplicada al cas d'estudi, preguntant-se si les comunitats obtingudes són fenomenològicament rellevants: si aporten informació important i interessant sobre el fenomen contingut en els documents que s'estan analitzant.

Finalment, és important entendre que la definició de «tema» que elabora una màquina mitjançant un procediment no supervisat pot ser substancialment diferent del que una persona entengui com a «tema». Per a un ordinador, qualsevol patró és susceptible de ser-ne un, de tema. Per exemple, un text mal preprocessat de ben segur que produirà resultats massa obvis o absurds, com ara l'agrupació d'articles i determinants com a paraules clau d'un tema. Per aquesta raó, és important entendre molt bé les passes que es duen a terme en aquest tipus d'anàlisi, i també, el tipus de resultat que pot esperar-se i com interpretar-los.

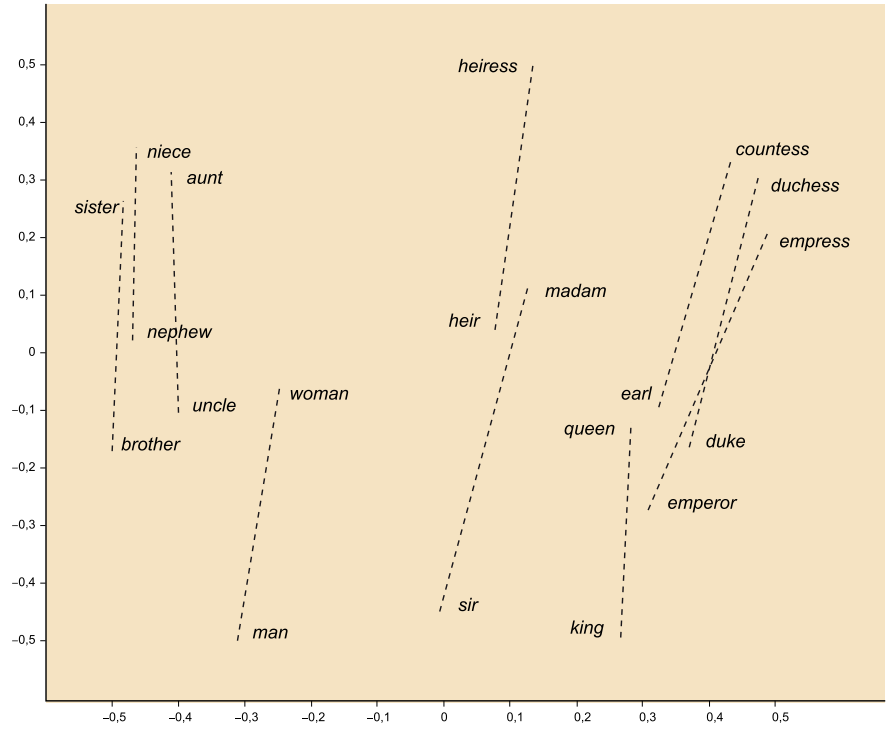
2.4. Algoritmes híbrids per a l'anàlisi de textos

Ja hem pogut veure que l'anàlisi de textos mitjançant algoritmes d'aprenentatge supervisat i no supervisat poden plantejar problemes importants. Per una banda, ens trobem amb el problema que els algoritmes fàcils d'interpretar i entendre, sovint, poden resultar excessivament senzills per processar la complexitat del llenguatge humà. Per altra banda, hi ha la qüestió que molts algoritmes poden ser complicats d'ajustar, i tot i així, és possible que ofereixin resultats poc sorprenents o interessants des de la perspectiva de la creació de valor. Per tot això, els algoritmes profunds i de conjunt (això és, híbrids, que combinen una lògica supervisada i no supervisada) són enormement populars en PLN.

El tipus de tècniques més populars durant els últims anys han estat les **xarxes neuronals basades en encaixos de paraules** o *word embeddings* (Ronan Collobert i altres, 2008). Mitjançant aquests processos, podem classificar missatges en funció de models preentrenats, i que són capaços d'identificar dife-

rents aspectes, com la temàtica o l'emocionalitat. El procés d'encaix consisteix a transformar una paraula en un nombre mitjançant un criteri predefinit (això és, una xarxa neuronal profunda prèviament entrenada) i que permet establir relacions matemàtiques entre paraules de l'estil «germà-germana = rei-reina» i, per tant, establir processos de predicció mitjançant equacions de l'estil «germà-germana + rei = reina» (vegeu figura 15). La xarxa neuronal Word2Vec (Mikolov i altres, 2013), desenvolupada al mateix cor de Google, és l'encaix de paraules més popular avui i és utilitzat per una gran quantitat de programari, inclòs el mateix cercador de Google i la seva funcionalitat de text predictiu.

Figura 15. Encaixos de paraules vinculades amb el gènere



Font: adaptada d'https://nlp.stanford.edu/projects/glove/.

Un analista que disposi dels coneixements necessaris podrà entrenar una xarxa neuronal profunda, optimitzant els encaixos per a la tasca particular de classificació o predicció numèrica que vulgui dur a terme. Una altra opció és utilitzar encaixos preentrenats, com ara GloVe o ELMo, que són xarxes neuronals profundes de codi obert entrenades amb textos de Twitter, de la Viquipèdia o grans compilacions de webs amb milers de milions de casos.

Les **xarxes neuronals preentrenades** garanteixen una precisió classificatòria que, en la majoria de casos, milloren substancialment els resultats d'un algoritme d'aprenentatge supervisat, ja que proporcionen una major contextualització i punts de referència perquè els algoritmes estableixin prediccions.

El **procés de predicció** consisteix a incrustar cada paraula en els paràmetres del model (p. ex. GloVe utilitza xarxes amb fins a tres-cents encaixos) i establir les correlacions oportunes entre les paraules i els grups de paraules en funció de les puntuacions obtingudes en cadascun dels encaixos (vegeu taula 1).

Taula 1. Exemple d'encaixos de paraules amb dades aleatòries

Paraula	Encaix				
	1	2	3	4	5
amic	0,92	0,61	0,63	0,66	0,82
bordar	0,74	0,38	0,33	0,20	0,90
cantar	0,35	0,74	0,27	0,57	0,57
dormir	0,02	0,63	0,24	0,31	0,35
escaiola	0,80	0,19	0,47	0,33	0,33

Font: elaboració pròpia.

De la mateixa manera que hi ha xarxes neuronals profundes preentrenades per facilitar encaixos de paraules, també n'hi ha que permeten l'encaix de dades en altres formats, com ara vídeos, àudios i imatges. Mitjançant aquesta tècnica, qualsevol format de dada es pot transformar en una sèrie de vectors numèrics a partir dels quals establir criteris de similitud o algebraics i, basant-se en aquests vectors, dur a terme operacions de clusterització no supervisada, o de resolució de problemes de categorització o de predicció numèrica típics dels algoritmes supervisats.

La principal limitació d'aquestes eines és, com ja s'ha comentat, la impossibilitat d'interpretar les relacions entre les variables establertes en una caixa negra multidimensional i que opera sobre distintes capes de coneixement computacional autoadministrat i autogestionat.

3. El tractament just de les dades

La intel·ligència artificial i els algoritmes formen part de les nostres vides. Ens faciliten moltes de les operacions i tasques que duem a terme durant el nostre dia a dia. Aquests algoritmes aprenen de nosaltres, de les nostres preferències i dels nostres patrons de consum. També, com és lògic, aprenen dels nostres defectes, de les nostres fòbies i dels nostres prejudicis. Tots aquests elements són absorbits pels algoritmes i constitueixen allò que anomenem «**biaix algorítmic**».

Operacions i tasques cotidianes

Per exemple, quan el mòbil ens proposa una ruta alternativa per arribar a la feina per estalviar-nos la retenció provocada per un accident, o quan YouTube ens proposa continguts que no hem vist i que probablement ens agradaran.

Un algoritme pot incorporar aquest tipus de biaixos fins i tot quan el seu creador no ho pretén i, per tant, pot romandre ocult i indetectable durant anys, reproduint tots aquells factors de desigualtat i totes les discriminacions que ha après i reforçat. Això és el que succeeix quan l'algoritme aprèn d'una sèrie de dades internament correlacionades, sense que ningú hagi avaluat aquestes correlacions i les hagi analitzat des d'un punt de vista normatiu i prescriptiu.

Per exemple, si l'algoritme darrere d'un portal de cerca de feina aprèn que la categoria «directiu» correlaciona amb «home», i que la categoria «mitja jornada» correlaciona amb «dona», oferirà més feines de directius als homes i més feines de mitja jornada a les dones. L'algoritme es converteix, d'aquesta manera, en un agent acrític de reproducció social, que pot generar una gran quantitat de situacions injustes i discriminatòries perfectament evitables (per exemple, en processos de selecció de personal, decisions financeres, avaluacions de risc, adjudicacions de subvencions, etc.).

Els algoritmes no solament poden aprendre dels nostres propis biaixos i dels biaixos continguts en les dades, sinó que són capaços d'amplificar-los i magnificar-los, donant lloc a una sèrie de problemes que la comunitat de la mineria de dades ja fa anys que n'és conscient (Nicholas DiFonzo, 2011).

En són bons exemples els problemes de classificació esbiaixada i discriminatòria (per exemple, la discriminació cap a les dones, persones migrants o altres col·lectius en la publicació d'ofertes de feina en els portals especialitzats d'internet) o els problemes en la recomanació i prescripció esbiaixada de continguts (per exemple, la propagació de notícies falses a les xarxes socials i la generació d'estats globals de desinformació mitjançant la presentació de notícies que s'ajusten a les creences prèvies dels individus). Des de la mateixa comunitat de la mineria de dades s'han proposat diferents tipus de solucions per a aquesta mena de problemes, tant per a la identificació o el descobriment d'aquests biaixos, com per a la sinterització d'algoritmes i processos orientats a un tractament just de les dades (Hajian i altres, 2016; Jordi Morales i Gras, 2020).

Pel que fa a la identificació i el descobriment de biaixos algorítmics, diversos autors han proposat **models d'identificació de discriminacions ocultes** en algoritmes basats en processos heurístics i en procediments d'enginyeria inversa. No tots els algoritmes són igualment accessibles i avaluables. Com ja hem comentat abans, com més opac i poc transparent sigui un algoritme, més complicada serà la interpretació de les relacions establertes entre les variables i també la detecció dels seus biaixos. Per això, cada vegada són més les veus que demanen que es deixin d'utilitzar els algoritmes de caixa negra com ara les xarxes neuronals en tasques de classificació que poden comportar elements discriminatoris (Alex Campolo i altres, 2017). Quant a la prevenció del biaix i dels seus efectes discriminatoris i socialment indesitjables, s'han desenvolupat tres tipus d'estratègies diferents, que actuen sobre els diferents elements comuns en diversos algoritmes amb finalitats classificatòries o de recomanació de continguts (Bora Edizel i altres, 2020).

1) Intervencions sobre les dades d'entrenament. El primer tipus de mesures són aquelles que actuen sobre les dades a partir de les quals s'entrenarà un algoritme, a mode d'operació de preprocessament de dades, com si es tractés d'una operació de tokenització o de lematització.

2) Intervencions sobre l'algoritme inductor. El segon tipus de mesures són les que actuen sobre el procediment heurístic que s'aplica sobre les dades introduïdes, i són capaces de corregir, com a mínim, una part dels seus biaixos durant la mateixa operació de processament de les dades.

3) Intervencions sobre els resultats de l'algoritme. L'últim grup de mesures són aquelles que actuen sobre l'*output* de l'algoritme: sobre la predicció, la classificació o la recomanació. Es tracta d'operacions de postprocessament de les dades que corregeixen en aquest punt els resultats discriminatoris d'un algoritme.

Serà en funció d'una sèrie de qüestions diferents i interrelacionades que haurem d'optar per un o altre disseny preventiu: la disponibilitat de dades històriques, el grau de coneixement sobre les correlacions que contenen, la transparència de l'algoritme inductor, etc.

El disseny d'algoritmes conscients de les discriminacions estructurals que hi ha a les nostres societats i que contribueixin a mitigar-los és un camp d'estudi en expansió i que avença de la mà del mateix progrés social. El seu desenvolupament depèn totalment de l'existència de consensos socials, d'acord amb un codi ètic i normatiu, com ara la igualtat entre homes i dones, la protecció dels drets dels infants o la protecció de les minories i dels grups socials no dominants. Els algoritmes poden ser agents de reproducció de les desigualtats presents en una societat, però també poden servir per mitigar el seu efecte i les

seves conseqüències indesitjables. Precisament per això no es pot deslligar el disseny d'aquests instruments de predicció i classificació automatitzada d'una anàlisi social en clau normativa, sobre què és desitjable i què és indesitjable.

Bibliografia

Anderson, Chris (2008). «The end of theory: The data deluge makes the scientific method obsolete». *Wired magazine* (vol. 16, núm. 7, pàg. 16-07). Nova York: Condé Nast Publications.

Andrzejewski, David; Zhu, Xiaojin; Craven, Mark (2009). «Incorporating domain knowledge into topic modeling via Dirichlet forest priors». A: A. Danyluk, L. Bottou i M. Littman. *Proceedings of the 26th annual international conference on machine learning* (pàg. 25-32). Canadà: ICML.

Bird, Steven; Klein, Ewan; Loper, Eduard (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. Massachusetts: O'Reilly Media.

Campolo, Alex; Sanfilippo, Madelyn; Whittaker, Meredith; Crawford, Kate (2017). *AI now 2017 report*. Nova York: AI Now Institute.

Collobert, Ronan; Weston, Jason; Bottou, Léon; Karlen, Michael; Kavukcuoglu, Koray; Kuksa, Pavel (2011). «Natural language processing (almost) from scratch». *Journal of machine learning research* (vol. 12, pàg. 2493-2537). Nova Jersey: AI Now Institute.

DiFonzo, Nicholas (2011). «The echo-chamber effect». *New York Times* (vol. 12, pàg. 2493-2537). Nova York: New York Times Company.

Edizel, Bora; Bonchi, Francesco; Hajian, Sara; Panisson, André; Tassa, Tamir (2020). «FaiRecSys: Mitigating algorithmic bias in recommender systems». *International Journal of Data Science and Analytics* (vol. 9, núm. 2, pàg. 197-213). Nova York: Springer.

Gerlach, Martin; Peixoto, Tiago P.; Altmann, Eduardo G. (2018). «A network approach to topic models». *Science Advances* (vol. 4, núm. 7). Pennsilvània: American Association for the Advancement of Science.

Hajian, Sara; Bonchi, Francesco; Castillo, Carlos (2016). «Algorithmic bias: From discrimination discovery to fairness-aware data mining». *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pàg. 2125-2126). Nova York: Association for Computing Machinery.

Metz, Rachel (2016, agost). «Why Microsoft Accidently Unleashed a Neo-Nazi Sexbot». *MIT Technology Review* [en línia]. Disponible a: <https://www.technologyreview.com/2016/03/24/161424/why-microsoft-accidentally-unleashed-a-neo-nazi-sexbot/>

Morales i Gras, Jordi (2020). «Cognitive Biases in Link Sharing Behavior and How to Get Rid of Them: Evidence from the 2019 Spanish General Election Twitter Conversation». *Social Media + Society* (vol. 6, núm. 2, pàg. 1-4). Nova York: SAGE Publications.

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey (2013). «Efficient estimation of word representations in vector space». *arXiv preprint arXiv* (pàg. 1301-3781). Arizona: International Conference on Learning Representations.

Ray, Paramita; Chakrabarti, Amlan (2019). «A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis». *Applied Computing and Informatics*. Amsterdam: Elsevier BV.

Rebala, Gopinath; Ravi, Ajay; Churiwala, Sanjay (2019). *An Introduction to Machine Learning*. Nova York: Springer.