Anàlisi de sentiments d'usuaris d'aerolínies mitjançant tècniques d'aprenentatge automàtic

Anna Espona Ninou

Resum– Les xarxes socials han evolucionat de tal manera que es deixa en segon pla la comunicació, i passa a ser un espai on compartir opinions i experiències. Les empreses s'aprofiten d'aquest canvi analitzant les valoracions per avaluar l'èxit d'un nou producte, identificar quines característiques agraden més i les necessitats del mercat, entre d'altres. Aquest procés s'anomena mineria de dades i té certes complicacions que cal tenir en compte per obtenir uns resultats acurats. Les dades obtingudes de les xarxes socials són heterogènies i contenen un alt percentatge d'ironia, acrònims i caràcters especials, específics de cada idioma i que s'han de valorar per categoritzar adequadament cada comentari. En aquest treball es fa una anàlisi automàtica de les opinions de clients extretes de la xarxa social Twitter. S'utilitza un conjunt de dades en brut que cal netejar i adaptar als algoritmes de detecció de sentiments. Un cop aplicats els algoritmes es pot veure quin d'ells aconsegueix un percentatge d'encert més elevat i se'n pot analitzar la causa.

Paraules clau– Anàlisi de sentiments, Support Vector Machines (SVM), Grid Search, Twitter, tokenització, stopwords, stemming, algoritme VADER, aproximació amb diccionaris.

Abstract– Social media has evolved in a way that leaves the communication on a secondary role, and has become a place where you can share opinions and experiences. Businesses take advantage of this change analyzing the reviews to evaluate the success of a new product, identifying which features are the most liked and the market needs, among others. All this process is called opinion mining. This process has some challenges that need to be taken into account to obtain accurate results. Social network data are heterogeneous and contains a high percentage of irony, acronyms and special characters specific to each language, this must be considered to categorize each review properly. In this paper, an automatic analysis of the opinions of clients extracted from the Twitter social network is made. A set of raw data is used, and must be cleaned and adapted to the sentiment analysis algorithms. Once the algorithms have been applied, one can see which of them achieves a better percentage of success and the cause can be analyzed.

Keywords– Sentiment analysis, Support Vector Machines (SVM), Grid Search, Twitter, tokenization, stopwords, stemming, VADER algorithm, dictionary approach.

1 INTRODUCCIÓ

PER les empreses és essencial analitzar dades dels seus clients o del mercat per adaptar el producte o servei als canvis constants. A causa de la gran quantitat de dades que es reben, és necessari tractar les dades de manera que s'obtingui coneixement útil d'elles. Utilitzant software per analitzar els patrons en grans conjunts de dades, les empreses poden aprendre més sobre els seus consumidors per desenvolupar estratègies de màrqueting millors, augmentar vendes i disminuir costos.

La mineria de dades (*data mining*) és el procés utilitzat per les companyies per transformar dades en brut en informació útil. Aquest procés consta d'una recollida efectiva de dades, un correcte emmagatzematge i el processament adequat a les dades.

Per a l'emmagatzematge de les dades es pot utilitzar un *data warehouse*, que és un repositori unificat per totes les dades que es recullen dels diversos sistemes d'una empre-

• E-mail de contacte: 96esponaanna@gmail.com
• Menció realitzada: Tecnologies de la Informació
• Treball tutoritzat per: Jordi Casas Roma (DEIC)
• Curs 2018/19

sa. D'aquesta manera les companyies poden centralitzar les dades. Amb un *data warehouse*, una organització pot separar segments de dades per usuaris específics, analitzar-les i utilitzar-les [1].

Independentment de com les companyies i altres entitats organitzen les dades, les dades s'utilitzen per donar suport als processos de presa de decisions de la direcció. Els programes de mineria de dades analitzen les relacions i els patrons de les dades en funció del que demana l'usuari. En altres casos, els analistes de dades troben clústers d'informació basats en relacions lògiques o observen associacions i patrons seqüencials per treure conclusions sobre les tendències de comportament del consumidor.

El procés de mineria de dades consta de 5 passos:

1. Les organitzacions recullen dades i es carreguen a la base de dades.

2. Es guarda i es gestionen les dades, dins del *data warehouse* o en el *cloud*.

3. Els analistes, equips de gestió i professionals de les tecnologies de la informació accedeixen a les dades i determinen com les volen organitzar.

4. El *software* ordena les dades en funció dels resultats dels usuaris.

5. L'usuari final presenta les dades en un format fàcil de compartir, com ara gràfics o taules [2].

La mineria de text (*text mining*) és el procés d'explorar i analitzar grans quantitats de dades de text no estructurades amb l'ajuda de software que pot identificar conceptes, patrons, tòpics, paraules clau i altres atributs dins les dades. El *text mining* s'ha convertit en una pràctica més utilitzada a causa del desenvolupament de plataformes de *big data* i algoritmes de *deep learning* que poden analitzar conjunts massius de dades no estructurades. Minar i analitzar textos ajuda a les organitzacions a trobar informació comercial valuosa en documents corporatius, e-mails, etc.

El *text mining* és similar al *data mining*, però focalitzat a texts en comptes de dades més estructurades. Un dels primers passos en el procés de *text mining* és organitzar i estructurar les dades de manera que es pugui sotmetre a anàlisis quantitatives i qualitatives. Aquest procés implica l'ús de processament de llenguatge natural (*Natural Language Processing, NLP*), que aplica principis de la lingüística computacional per analitzar i interpretar conjunts de dades.

En aquest projecte es treballa sobre una aplicació del *text mining* bastant utilitzada, com és l'anàlisi de sentiments. Analitza les dades de text subjectives per identificar rols habituals que apunten a sentiments positius o negatius per part dels clients. Aquesta informació es pot utilitzar per arreglar problemes dels projectes, millorar el servei al client i planejar noves campanyes de màrqueting [3].

Aquest document està dividit en diverses seccions. Inicialment a les seccions 2 i 3, s'expliquen els objectius i l'estat de l'art d'aquest treball, respectivament. A la secció 4, s'especifiquen les metodologies i eines emprades. Seguidament, a la secció 5, corresponent al desenvolupament, s'explica com són les dades analitzades, el preprocessament que cal aplicar i els tres algoritmes d'anàlisi de sentiments que s'utilitzen. A la següent secció es mostren els resultats obtinguts amb els tres algoritmes de *sentiment mining*, i finalment en les seccions 7 i 8 hi ha les conclusions i el treball futur, respectivament.

2 OBJECTIUS

L'objectiu principal d'aquest treball és l'anàlisi de sentiments en comentaris de *Twitter*. Primer de tot cal seleccionar un conjunt de dades suficientment extens per obtenir informació més fiable a l'hora de la detecció de sentiments. A més, si les dades són etiquetades, facilitarà la utilització d'algoritmes d'aprenentatge automàtic supervisat.

Pel fet que les dades seran en format text, s'ha de fer un processament previ a l'anàlisi per netejar i adequar les dades als algoritmes, ja que aquests no tracten amb text sinó amb valors numèrics. Per tant, s'haurà d'investigar quines tècniques de preprocessament existeixen i seleccionarne les més adequades.

Un cop les dades estiguin preparades per l'anàlisi, s'hauran de buscar algoritmes per detectar els sentiments. Primerament es pot utilitzar l'estratègia més bàsica per poder tenir una referència inicial, i a partir d'aquí, augmentar la precisió dels algoritmes per arribar al màxim percentatge d'encert. Per cada una de les aproximacions utilitzades, es calcularà el percentatge d'encerts i es raonarà la possible causa d'aquest valor. Finalment amb aquestes dades, es podran veure les diferències entre les diverses estratègies i determinar quina d'elles és la més eficient.

3 ESTAT DE L'ART

3.1 Mètodes per l'anàlisi de sentiments

A l'hora de fer una anàlisi dels sentiments d'un text, ens trobem amb dues aproximacions principals, utilitzant lexicons o algoritmes d'aprenentatge automàtic. Les aproximacions basades en lexicons consisteixen en llistes de paraules manualment etiquetades amb una polaritat positiva o negativa, i amb un valor de polaritat. El lexicó construït s'utilitza per calcular el sentiment general d'un text. Un avantatge d'aquestes aproximacions és que no és necessari entrenar les dades. Els lexicons són molt utilitzats en texts convencionals com fòrums i blogs. Per altra banda, no són tan utilitzats en dades extretes de xarxes socials, a causa del format desestructurat de les dades, ja que contenen peculiaritats.

Quant a les aproximacions basades en algoritmes de *machine learning*, en podem trobar de dos tipus, algoritmes d'aprenentatge supervisat i no supervisat. L'objectiu de l'aprenentatge supervisat és crear una funció capaç de preveure el valor corresponent a qualsevol objecte d'entrada, després d'haver entrenat. Tot seguit s'expliquen, a grans trets, alguns dels algoritmes que actualment s'utilitzen per aquesta classificació [4].

3.1.1 Arbres de decisió

L'arbre de decisió és un mètode supervisat utilitzat en classificació i regressió. L'objectiu és crear un model que predigui correctament a partir de l'aprenentatge de regles de decisió simples. En els arbres de decisió els nodes repre-

senten els atributs, les branques les decisions i les fulles són els resultats obtinguts [5].

3.1.2 Random forest

El *random forest* és un mètode supervisat de classificació i regressió basat en un conjunt d'arbres de decisió. En comptes de buscar la característica més important mentre separa un node de l'arbre, busca la millor característica dins d'un conjunt aleatori de característiques [6].

3.1.3 Support Vector Machines

Els *support vector machines* són un conjunt de mètodes d'aprenentatge supervisat utilitzats per classificació, regressió i detecció d'anomalies. L'objectiu d'aquest algoritme és trobar un hiperplà en un espai n-dimensional que classifica les dades [7].

3.1.4 Xarxes neuronals

Les xarxes neuronals són una eina computacional que pretén simular l'arquitectura i les operacions internes del cervell humà i el seu sistema nerviós. Una xarxa neuronal consta d'uns elements de procés, que representen les neurones, i que estan connectats uns als altres, formant una xarxa. Cadascun d'aquests elements de procés pot tenir diverses entrades i emet un únic senyal de sortida. En el procés d'entrenament es determinen els diferents pesos que s'aplicaran als senyals d'entrada [8].

4 METODOLOGIA

S'ha utilitzat una metodologia *Agile* per tal de mantenir un progrés constant del treball. Hi ha diverses metodologies *Agile* [10], la que s'ha utilitzat és *Extreme programming*, ja que es basa en un feedback continu i en constants entregues per tal de poder fer un seguiment i testeig més rigorós [11]. Així doncs, s'ha fet una reunió setmanal amb el tutor del treball per avaluar els canvis realitzats i poder seguir amb el desenvolupament des d'una base correcte. A més, s'ha establert la feina requerida per la següent reunió i s'han resolt possibles dubtes per ambdues parts.
S'ha fet servir una eina de versionatge, per tal de mantenir còpies i un registre de la feina realitzada al llarg del treball. Per a tal cosa s'ha utilitzat un repositori a GitHub on s'ha anat actualitzant el document final [12].
El treball s'ha realitzat amb el llenguatge de programació Python, ja que hi ha diverses llibreries creades per l'anàlisi de dades que faciliten la implementació de funcions i càlculs [13].
Concretament per aplicar les tècniques de preprocessament de text s'ha utilitzat el *framework Natural Language Toolkit, NLTK*, ja que implementa la majoria d'algoritmes de processament de llenguatge natural, *NLP*.

5 DESENVOLUPAMENT

5.1 Conjunt de dades

S'ha utilitzat un *dataset* [14] amb 14.600 comentaris a l'aplicació de *Twitter* sobre diferents aerolínies dels Estats Units. S'ha escollit aquest *dataset* perquè és prou gran per a poder entrenar els algoritmes de *machine learning* amb efectivitat. A més, per cada un dels *tweets* es té la classificació de sentiment, és a dir, és un conjunt de dades etiquetat, de manera que es pot utilitzar per entrenar algoritmes supervisats.

5.2 Preprocessament

Per realitzar el projecte s'ha utilitzat processament del llenguatge natural, *NLP*. *NLP* és una branca de la intel·ligència artificial que permet entendre, interpretar i manipular textos en el llenguatge humà. D'aquesta manera la comunicació entre humans i màquines és més propera. En essència, *NLP* trenca el text en peces elementals i intenta entendre les relacions entre elles i com funcionen juntes per donar-li un significat [15]. Està format per un conjunt de tècniques de *machine learning* que permeten treballar amb documents de texts, considerant la seva estructura interna i la distribució de les paraules.

Per aquest treball s'ha utilitzat l'estratègia *bag-of-words*, la qual no té en compte l'ordre de les paraules en una sentència sinó la freqüència d'aparença de les paraules en el text. Per a l'anàlisi de sentiments no és essencial l'ordre semàntic de les paraules, sinó el seu significat. Bé és cert que en algunes ocasions dues paraules juntes tenen significat més rellevant o diferent que si no ho estiguessin, això també es tindrà en compte més endavant. L'objectiu acaba sent el mateix, maximitzar la informació d'un document reduint la mida del vocabulari eliminant els termes que són massa freqüents, com determinants i adverbis.
El procés complet d'aquesta estratègia consta de quatre fases: tokenització, eliminació de stopwords, *stemming* i vectorització. Tot seguit s'explicaran en detall les tècniques que s'han aplicat, juntament amb una representació gràfica per a una major comprensió.

5.2.1 Tokenització

La primera tècnica que s'ha aplicat és la tokenització, la qual consisteix a separar un text per paraules, sentències o conjunt de paraules. En aquest cas, com que els *tweets* són frases curtes, s'ha separat per paraules. S'ha hagut de tenir en compte que les abreviacions en anglès, com la negació *didn't*, es pot trobar separada en una barra lateral, com ara: *didn/'t*, per tant s'ha d'aplicar una tokenització amb expressions regulars, per tenir en compte aquests patrons i que no separi la paraula. En el primer cas, de la paraula *didn/'t* obtindríem *didn* i *'t* per separat. Amb les expressions regulars obtenim la paraula sencera sense la barra lateral. Com es pot apreciar a la figura 1, en aquest procés també s'eliminen tots els caràcters que no siguin alfabètics.

5.2.2 Eliminació de stopwords

Un cop tenim el conjunt de paraules per separat, cal eliminar-ne les que són molt freqüents però no tenen cap informació semàntica, com ara els determinants i preposicions. Aquestes paraules s'anomenen *stopwords* i el *framework NLTK* proporciona una llista per a diversos idiomes, per tant es pot aplicar amb facilitat al text per eliminar-ne les paraules que apareixen a la llista. A la figura 1 es pot veure com varia el text després d'extreure'n els *stopwords*.

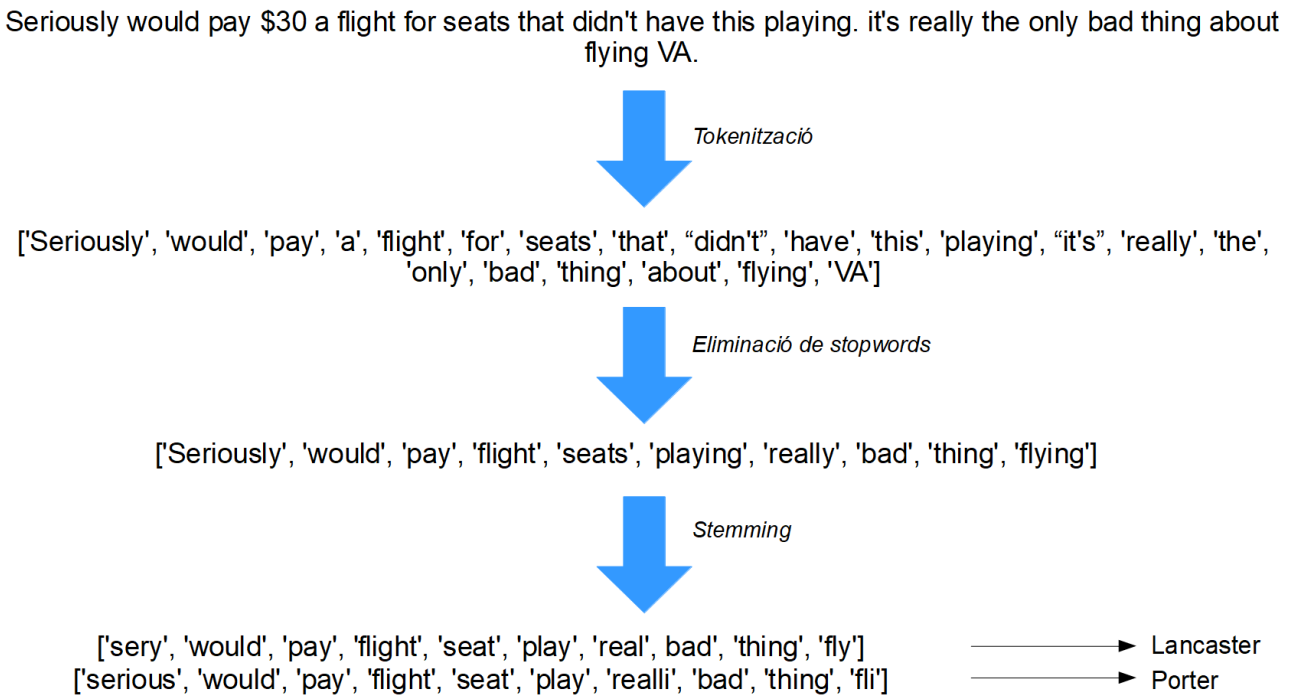

Fig. 1: Procés de l'estratègia *bag-of-words*

5.2.3 Stemming

La següent tècnica que s'ha d'aplicar és *stemming*. És el procés de transformar les paraules a la seva forma arrel. D'aquesta manera mantenim la semàntica i unifiquem les paraules derivades a una única. Dins del *framework NLTK* trobem diverses implementacions d'aquest procés, a la imatge 1 es veu el procés d'aquesta tècnica amb dues d'aquestes. Com es pot veure ambdues són molt semblants i tenen alguns resultats erronis, per tant les dues ens serveixen per igual. En aquest cas s'ha escollit el *Lancaster*.

5.2.4 Vectorització

La gran majoria, si no tots, els algoritmes de *machine learning* requereixen que les dades d'entrada estiguin en format numèric. És per això que treballant amb dades de text cal transformar-les en un vector numèric. Aquest procés és el que s'anomena vectorització. En aquest procés també hi ha diverses estratègies a seguir, s'ha escollit *Count Vectorizer*. Es basa en la representació en funció del nombre d'aparicions en el document. És a dir, s'han de processar tots els *tweets* per determinar quantes paraules úniques hi ha i la seva freqüència. D'aquesta manera, cada *tweet* es transforma en un vector de mida fixa amb valors de 0 o 1. Si és 0, vol dir que la paraula no està present, mentre que 1 és el contrari [16].

5.3 Anàlisi de sentiments

L'anàlisi de sentiments és un tipus de processament de llenguatge natural que implica construir un sistema per recollir i categoritzar opinions sobre un producte, és a dir, determinar si un text és positiu, negatiu o neutre. Un sistema d'anàlisi de sentiment combina processament de llenguatge natural (*NLP*) amb algoritmes de *machine learning* [17]. L'anàlisi de sentiments ajuda a l'anàlisi de dades de les grans empreses a conduir recerques de mercat, analitzar l'opinió pública, etc.
Hi ha diversos reptes a l'hora d'analitzar els sentiments. Primerament hi ha paraules que es poden considerar positives en un context i negatives en un altre. Per exemple, la paraula "llarg". Si un client diu que la vida de la bateria del telèfon mòbil és llarga, aleshores seria una opinió positiva. Per contra, si es digués que el temps d'encesa del telèfon mòbil és llarg, seria negativa. Aquestes diferències fan que un sistema entrenat per les opinions d'un producte o les seves característiques, no funcioni del tot bé en un altre producte. Un altre desafiament és tenir en compte les contradiccions. La majoria de valoracions tindran comentaris positius i negatius alhora, cosa que es pot solucionar tractant les frases per separat. Malgrat això, com més informal és el mitjà de comunicació, més es tendeix a combinar opinions diverses en la mateixa frase [18].

Hi ha dos possibles enfocaments per l'anàlisi de sentiments: basat en la polaritat i basat en valència. El primer d'aquests simplement classifica el text en positiu o negatiu, mentre que en el segon cas es té en compte la intensitat del sentiment. Per exemple les paraules 'bé' i 'excel·lent' es tractarien igual en el cas que sigui basat en polaritat però si és per valència, la paraula 'excel·lent' es considera més positiva [17]. S'han aplicat tres algoritmes diferents, per detectar els sentiments de diversos *tweets* sobre aerolínies dels Estats Units. Els algoritmes utilitzats són: basat en diccionari, VADER i *Support Vector Machines* (*SVM*). Tot seguit s'expliquen en detall cadascun d'ells.

5.3.1 Diccionari

El primer algoritme que s'ha aplicat es basa en diccionaris de paraules positives i negatives, és un clar exemple d'anàlisi basat en la polaritat. En general consisteix en diccionaris que relacionen paraules amb el seu sentiment. En

el nostre cas, en comptes de diccionaris es tracta de dos llistats, un de termes positius i un de negatius [19]. Per cada paraula del *tweet* es comprova si existeix dins d'alguna de les llistes. Si es troba a la llista positiva, s'augmenta el valor de l'estat del *tweet*, i si es troba a la negativa, es disminueix. Quan s'han comprovat totes les paraules del *tweet* es revisa el valor de l'estat i se sap si és positiu, negatiu o neutre.

5.3.2 VADER

La següent anàlisi s'ha fet amb *'Valence Aware Dictionary and sEntiment Reasoner'* (VADER), una eina d'anàlisi de sentiments basada en valències. Per tant, totes les paraules estan definides amb un valor significatiu en funció de si són més positives o més negatives. Quan VADER analitza un *tweet* genera quatre mètriques. Mostra els percentatges positiu, negatiu i neutre del *tweet* i el compost, que és la suma dels valors de les paraules que s'han trobat dins del diccionari de VADER. D'aquesta manera, el resultat es mira únicament del valor del compost per saber si és positiu, negatiu o neutre. Aquests valors estan normalitzats en un rang de -1 a 1.
Analitzem un *tweet* d'exemple per veure-ho més clar.

@VirginAmerica and it's a really big bad thing about it

Mètriques de sentiment	Valor
Positiu	0.0
Negatiu	0.321
Neutre	0.679
Compost	-0.5829

Fig. 2: Exemple resultats VADER

Fins aquí aquesta eina no es diferencia gaire a l'anterior, ja que puntua les paraules negatives i positives i el sumatori d'aquest és el resultat. Però VADER està més enfocat a textos de xarxes socials tenint en compte diversos factors. A les xarxes socials s'acostuma a mostrar el sentiment a partir d'una emoticona, si no es té en compte, la frase pot semblar neutral en comptes de negativa o positiva, així que VADER afegeix les emoticones al seu diccionari. D'aquesta manera, si a la frase anterior afegim una emoticona representativa, el sentiment varia.

@VirginAmerica and it's a really big bad thing about it :(

{'neg': 0.455, 'neu': 0.545, 'pos': 0.0, 'compound': -0.7703}

Fig. 3: Exemple resultats VADER amb emoticones

També es tenen en compte les paraules majúscules, ja que les paraules es consideren amb més èmfasi. D'aquesta manera, si la paraula *'bad'* del primer exemple s'escriu en majúscules, el valor del compost augmenta de 0.5829 a 0.6717.
Els signes d'exclamació també es comptabilitzen com un indicatiu d'increment d'intensitat. Es valora diferent si n'hi ha un, dos, tres o quatre signes seguits. A partir del quart, no augmenta el valor. Partint de l'exemple inicial, amb un signe d'exclamació augmenta fins a 0.6212, mentre que amb quatre signes augmenta fins a 0.7137.

@VirginAmerica and it's a really big bad thing about it!!!!

{'neg': 0.382, 'neu': 0.618, 'pos': 0.0, 'compound': -0.7137}

Fig. 4: Exemple resultats VADER amb signes d'exclamació

Addicionalment es tenen en consideració els adverbis que modifiquen la intensitat dels adjectius. Es comptabilitza diferent *'extremadament bo'* que *'una mica bo'*. D'aquesta manera, si a l'exemple ometem la paraula *'really'* disminueix a 0.5423. En el context de les xarxes socials s'utilitza bastant la conjunció *'però'*. És important tenir-ho en compte, ja que normalment la frase abans d'aquesta paraula té un sentiment diferent de la posterior. VADER valora amb més importància la frase que es troba després de *'però'*.

@VirginAmerica and it's a really big bad thing about it, but it's ok.

{'neg': 0.16, 'neu': 0.641, 'pos': 0.199, 'compound': 0.1406}

Fig. 5: Exemple resultats VADER amb conjunció

Com es pot veure a l'exemple de la figura 5, si afegim *'but it's ok'* darrere del *tweet* original, el sentiment general del *tweet* canvia de ser molt negatiu a positiu [20].

5.3.3 Support Vector Machines

Aquest mètode d'anàlisi de sentiments és un algoritme de *machine learning* supervisat que es pot utilitzar tant per problemes de classificació com de regressió. El nostre cas es tracta d'un problema de classificació. L'objectiu de les SVM és trobar un hiperplà en un espai de N dimensions, on N és el nombre de característiques. Els hiperplans són límits que ajuden a classificar les dades. Per separar les dues classes de dades, hi ha molts possibles hiperplans que es podrien escollir. L'objectiu és trobar un pla que tingui el màxim marge, és a dir, la distància màxima entre punts de dades de classes diferents. Maximitzar aquesta distància, ens proporciona més seguretat en futures classificacions de dades.

La dimensió de l'hiperplà depèn del nombre de característiques. Si el nombre de característiques d'entrada és 2, aleshores l'hiperplà només és una línia. Però si el nombre de característiques d'entrada és tres, aleshores l'hiperplà és un pla de dues dimensions. El que s'anomenen *support vectors*, són punts de dades que estan propers a l'hiperplà i influencien en la posició i orientació d'aquest. Utilitzant aquests vectors, maximitzem el marge. Si s'eliminen els *support vectors*, la posició de l'hiperplà canviarà [21].

Primer de tot es representa cada element de les dades en un espai n-dimensional. El valor de cada característica, passa a ser una coordenada en particular. A partir d'aquí s'a-

consegueix la classificació trobant l'hiperplà que diferencia les classes. Ens podem trobar amb el cas que no es pugui utilitzar un hiperplà lineal per separar les classes, en aquests casos es pot utilitzar la tècnica *kernel trick* de les SVM, que afegeix una nova característica, és a dir, una nova dimensió i s'observen les dades des d'eixos diferents [22]. Els *kernels* són unes funcions que s'utilitzen per reordenar els elements, per tal d'arribar a un hiperplà lineal, transformen l'espai dimensional d'entrada en un espai dimensional major [23].

Per aplicar l'algoritme primer de tot s'han de separar les dades en dos conjunts, un d'entrenament de l'algoritme i un per testejar-lo. Com que es té un conjunt de dades prou gran, s'ha agafat un 80 % de les dades per l'entrenament. Els models de *machine learning* tenen dos tipus de paràmetres. Per un costat hi ha els paràmetres del model que són les propietats de les dades d'entrenament, que s'aprenen durant l'entrenament. Per l'altra banda hi ha els hiperparàmetres que defineixen les propietats que dirigeixen el procés d'entrenament. Aquests paràmetres s'han d'escollir adequadament per tal d'obtenir els millors resultats possibles [24]. Primerament s'ha aplicat l'algoritme amb els paràmetres per defecte, per tenir una idea base dels resultats. D'aquesta manera, no hi ha una millora significativa de l'*accuracy* respecte a l'algoritme VADER, així que s'ha aplicat una tècnica d'optimització d'hiperparàmetres per trobar els valors òptims.

Hi ha diverses tècniques per optimitzar els hiperparàmetres, la que s'ha utilitzat és *Grid Search*. Aquesta tècnica consisteix a entrenar l'algoritme amb totes les combinacions possibles entre els valors dels hiperparàmetres que volem testejar. És a dir, a cada hiperparàmetre li assignem els valors pels quals volem que se'n comprovi l'eficiència, i aquesta tècnica prova totes les combinacions per trobar la que retorna el valor d'*accuracy* més elevat [25]. El *Grid Search* optimitza els paràmetres de SVM utilitzant la tècnica de *cross validation*, CV, com a mètrica de rendiment. L'objectiu és identificar una bona combinació d'hiperparàmetres de tal manera que el classificador pugui predir dades desconegudes acuradament [26].

Per problemes de SVM els hiperparàmetres que es tenen en compte són: C, *kernel*, *gamma* i *degree*. El *kernel* selecciona el tipus d'hiperplà que s'utilitzarà per separar les dades, hi ha tres possibilitats: lineal, RBF o polinòmic. El *kernel* lineal, utilitza un hiperplà lineal, i els *kernel* RBF i polinòmic un hiperplà no lineal. En funció del *kernel* s'han de tenir en compte diferents paràmetres. Pel *kernel* lineal només s'ha de donar valor a l'hiperparàmetre C, per un *kernel* RBF es té en compte l'hiperparàmetre C i *gamma*, i pel polinòmic s'hi afegeix el *degree* [27].

Amb les SVM es busca un valor de marge elevat, i una taxa de classificació incorrecta baixa. Però aquestes dues coses són contradictòries. Si incrementem el marge, acabarem amb una taxa de classificació incorrecta elevada. No hi ha una regla general per escollir el valor de C, depèn de les dades d'entrenament. L'opció més viable és provar diferents valors de C i escollir el que dona menys classificació incorrecta en les dades d'entrenament. Un valor alt de C, significa un marge petit, i viceversa [28].

L'hiperparàmetre *gamma*, únicament és pels hiperplans no lineals, és a dir, RBF i polinòmic. A mesura que s'augmenta el valor de *gamma*, es condueix a l'*overfitting*. L'*overfitting* és l'efecte de sobre entrenar un algoritme sobre un conjunt específic de dades, de manera que el classificador intenta ajustar-se perfectament amb les dades d'entrenament. Això provoca que l'algoritme no funcioni amb eficiència amb altres conjunts de dades que tinguin característiques diferents. Finalment l'hiperparàmetre *degree*, indica el grau del polinomi utilitzat per trobar l'hiperplà. Per tant, utilitzar un *degree* amb valor 1, és el mateix que utilitzar un *kernel* lineal [27].

6 RESULTATS

Un cop s'han aplicat els diferents algoritmes d'anàlisi de sentiments s'ha calculat el percentatge de *tweets* detectats com a positius, negatius i neutres, per cada un dels algoritmes. En la figura 6 es pot veure una comparació d'aquest percentatge amb els percentatges esperats, en el cas de la primera estratègia, amb diccionaris.

Com es pot observar s'han detectat el *tweets* majoritàriament com a positius, deixant un baix valor d'encert en el cas de polaritat negativa. Concretament s'ha detectat un 70 % de comentaris positius, quan realment només hi ha un 16 %. Aquest valor tan extrem pot ser degut als diccionaris utilitzats. Com s'ha explicat anteriorment, hi ha paraules que poden tenir significats tant positius com negatius, en funció del context. Aquest algoritme no té en compte l'entorn de la paraula, per tant si es troba en una frase negativa, es pot haver comptat com a positiva.

En general, aquest algoritme ha obtingut un percentatge d'encert, també conegut com a *accuracy*, de 30,60 %.

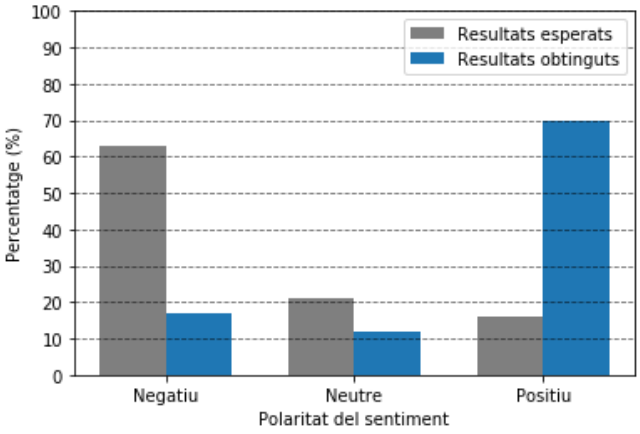

Fig. 6: Gràfica de comparació aplicant diccionaris

Aplicant la segona estratègia, VADER, s'ha millorat el percentatge d'encert total a un 48,99 %. En aquest cas s'ha provat d'aplicar l'algoritme sense haver fet un processament previ a les dades i el valor d'encert augmenta fins a 54,65 %. Això és a causa que l'algoritme VADER té en compte els signes de puntuació i les emoticones, que en el preprocessament s'eliminen, i els adverbis es modifiquen a la seva forma arrel.

A la figura 7 es pot veure una gràfica comparant els resultats originals amb els de l'algoritme sense aplicar preprocessament. Com es pot veure, el percentatge de *tweets* positius detectats ha disminuït respecte a l'anterior algoritme. Així mateix, ha augmentat el valor de *tweets* detectats com a negatius i neutres.

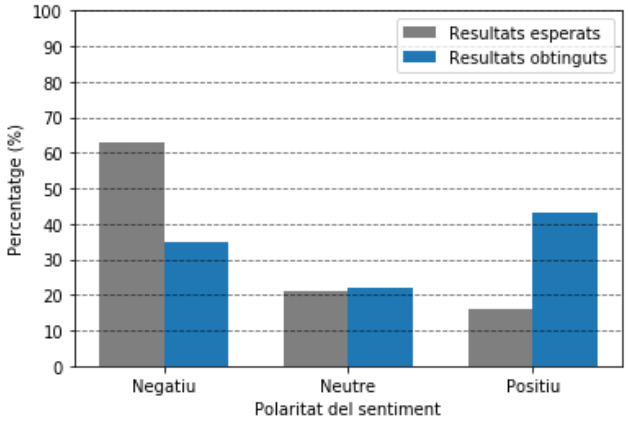

Fig. 7: Gràfica de comparació aplicant VADER

Aplicant l'algoritme de *machine learning*, SVM, amb els paràmetres per defecte l'*accuracy* és de 64,51 %. Després d'aplicar la tècnica *Grid Search* augmenta a 79,54 %. A la següent imatge es veuen els resultats comparats amb els originals. Es pot observar com el marge d'error disminueix significativament en comparació als dos anteriors algoritmes.

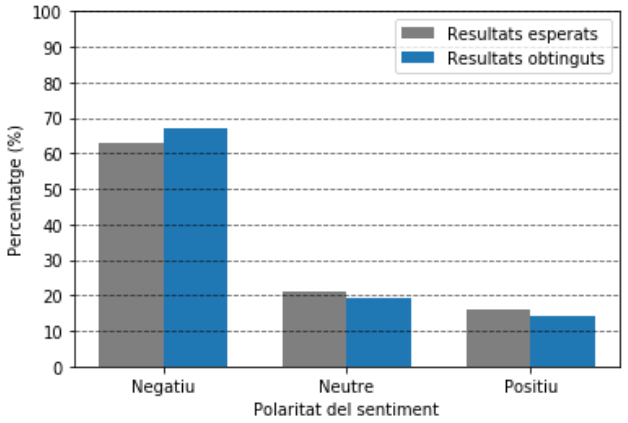

Fig. 8: Gràfica de comparació aplicant SVM

A la taula 1 es pot apreciar una taula comparativa entre els tres algoritmes, així com els valors *accuracy* i d'error. En aquesta es mostren els valors de la *true positive rate, TPR* per a cada un. És a dir, els comentaris detectats correctament dins de cada polaritat de sentiment. Per exemple, hi ha un total de 2363 *tweets* positius dins del conjunt de dades, i l'aproximació basada en diccionaris, n'ha detectat 10260. De tots aquests, només 2050 han estat detectats correctament, el qual porta a un 86 % d'encerts respecte al total de positius, no de tot el conjunt de dades. Aquesta comparació és interessant, ja que es pot veure la diferència d'encerts entre algoritmes, per cada polaritat de sentiment.

Amb l'estratègia basada en diccionaris s'aconsegueix un 86 % en el cas de la TPR en *tweets* positius, com s'ha indicat en l'exemple anterior. Si es compara amb el valor de l'*accuracy*, el TPR és molt elevat. Com s'ha vist a la figura 6, la majoria de comentaris han estat classificats com a positius. Això provoca que per estadística, s'encerti un gran nombre de classificacions. Per altra banda, tant el TPR per *tweets* negatius com per *tweets* neutres és molt baix. Per tant, dels pocs comentaris que s'han detectat com a negatius i neutres, la gran majoria han estat mal classificats.

Amb l'estratègia VADER, s'obté aproximadament el mateix valor de TPR en *tweets* positius. En aquest cas, encara que el percentatge de classificacions amb polaritat positiva ha disminuït fins a prop d'un 40 %, segueix sent més del doble de la realitat. Es podria pensar que la causa d'un valor de TPR tan elevat sigui l'estadística com en el cas anterior, però si observem el TPR de *tweets* negatius i neutres, veiem com han augmentat considerablement respecte a l'anterior algoritme. Per tant, encara que hi pugui jugar un paper important, és innegable que aquest algoritme ha millorat la classificació.

Per últim, l'algoritme SVM obté el valor més reduït de TPR per *tweets* positius, encara que el valor de l'*accuracy* sigui més elevat. Això és causat pel fet que s'han detectat més comentaris positius dels que hi ha en realitat. Per tant, encara que tots haguessin estat ben classificats, no s'obtindria un 100 %, però és clar que no tots els comentaris classificats com a positius, han estat correctament detectats. De la mateixa manera passa amb el valor de TPR de *tweets* neutres. Per altra banda, el TPR de *tweets* negatius augmenta considerablement en comparació als dos anteriors algoritmes.

Com s'ha vist a la figura 8, el percentatge de comentaris detectats com a negatius, és molt semblant al percentatge real. Amb el 68 % de comentaris detectats com a negatius, i el 89 % d'aquests, classificats correctament, el valor de l'*accuracy* queda justificat, ja que la majoria de deteccions han estat encertades.

7 Conclusions

S'han pogut assolir tots els objectius proposats inicialment. El conjunt de dades utilitzat, amb referència a valoracions d'usuaris d'aerolínies dels Estats Units, ha resultat efectiu per l'anàlisi. En tractar-se d'un conjunt de dades extens i etiquetat, ha permès comprovar els resultats obtinguts en les diferents estratègies de detecció de sentiments, i el correcte entrenament de l'algoritme supervisat d'aprenentatge automàtic.

També s'ha pogut estudiar amb èxit les diferents tècniques de processament de texts, i s'han escollit les més adequades pel propòsit del projecte. D'aquesta manera, també s'han contemplat diverses opcions d'algoritmes d'anàlisi de sentiments, i s'han aplicat de manera seqüencial de menys a més acurats. D'aquesta manera es pot veure amb més facilitat les diferències entre ells.

La primera tècnica utilitzada, corresponent a la utilització de diccionaris, no dona un percentatge molt elevat d'encerts, i s'ha utilitzat com a referència base per la comparació amb els següents algoritmes. A partir d'aquesta tècnica, s'ha analitzat quines millores necessita per incrementar el percentatge, i s'ha seleccionat l'algoritme posterior en funció d'aquests paràmetres. La següent tècnica també es basa en diccionaris, però especialment enfocats a texts de xarxes socials, per tant, es tenen en compte l'escriptura peculiar d'aquest entorn i elements tipogràfics que en el primer cas s'ometen, com per exemple, els signes d'exclamació.

Finalment s'ha utilitzat un algoritme d'aprenentatge automàtic, el qual s'esperava una millora considerable, i els resultats obtinguts, han estat els esperats.

	TPR *tweets* positius (%)	TPR *tweets* negatius (%)	TPR *tweets* neutres (%)	% *accuracy*	% error
Diccionaris	86.75	22.17	12.78	30.61	69.39
VADER	87.39	50.41	42.24	54.65	45.35
SVM	67.97	88.99	57.93	79.54	20.46

TAULA 1: TAULA COMPARATIVA ENTRE ELS ALGORITMES DE DETECCIÓ DE SENTIMENTS

8 TREBALL FUTUR

Actualment s'utilitza la tècnica *Grid Search* per l'optimització de paràmetres de l'algoritme SVM. Es pot aplicar aquesta mateixa tècnica amb diferents combinacions de paràmetres per veure'n les diferències i si s'obté alguna millora. Així com aplicar la tècnica *Random Search*, que és una alternativa a *Grid Search*. D'aquesta manera es pot comprovar si dona diferents resultats i es poden comparar ambdues tècniques.

Addicionalment es poden seguir aplicant algoritmes de *machine learning* per veure'n la diferència entre ells i trobar quin és el millor algoritme per analitzar sentiments en texts extrets de xarxes socials.

Quant als propòsits del treball, a partir de trobar el valor emocional dels *tweets* es poden analitzar els texts per trobar els tòpics més significatius de queixa i lloança sobre aquest conjunt de dades, d'aquesta manera es trobarien els principals problemes de les aerolínies, i es podria classificar segons estació de l'any, països, etc. Per altra banda, en comptes de classificar els texts en positiu, negatiu o neutre, es pot obrir el ventall de sentiments a diverses emocions.

REFERÈNCIES

[1] "Data Warehouse: todo lo que necesitas saber sobre almecenamiento de datos,"*PowerData*. [En línia]. Disponible a:https://www.powerdata.es/data-warehouse. [Accedit Febrer 5, 2019].

[2] Will Kenton, "Data Mining,"*Investopedia*. [En línia]. Disponible a: https://www.investopedia.com/terms/d/datamining.asp. [Accedit Gener 24, 2019].

[3] Margaret Rouse, "Text mining (text analytics),"*SearchBusinessAnalytics*. [En línia]. Disponible a: https://searchbusinessanalytics.techtarget.com/definition/text-mining. [Accedit Gener 24, 2019].

[4] Shahid Shayaa, Noor Ismawati Jaafar, Shamshul Bahri, Ainin Sulaiman, Phoong Seuk Wai, Yeong Wai Chung, Arsalan Zahid Piprani, Mohamed Ali Algaradi, *Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges*. University of Malaya: 2018.

[5] "Decision Trees,"*Scikit-learn*. [En línia]. Disponible a: https://scikit-learn.org/stable/modules/tree.html. [Accedit Febrer 6, 2019].

[6] Niklas Donges, "The Random Forest Algorithm,"*Medium Corporation*. Disponible a: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd. [Accedit Febrer 6, 2019].

[7] "Support Vector Machines,"*Scikit-learn*. [En línia]. Disponible a: https://scikit-learn.org/stable/modules/svm.html#svm-outlier-detection. [Accedit Febrer 6, 2019].

[8] "Introducció a les xarxes neuronals,"*Universitat Politècnica de Catalunya*. [En línia]. Disponible a: https://upcommons.upc.edu/bitstream/handle/2099.1/3326/55875-6.pdf?sequence=6&isAllowed=y. [Accedit Febrer 6, 2019].

[9] Margaret Rouse, "Big data,"*SearchDataManagement*. [En línia]. Disponible a: https://searchdatamanagement.techtarget.com/definition/big-data. [Accedit Gener 26, 2019].

[10] "What is Agile Methodology?,"*CollabNET VersionOne*. [En línia]. Disponible a: https://resources.collab.net/agile-101/agile-methodologies. [Accedit Octubre 6, 2018].

[11] Kent McDonald, "What is Extreme Programming (XP) ?,"*Agile Alliance*. [En línia]. Disponible a: https://www.agilealliance.org/glossary/xp/. [Accedit Octubre 6, 2018].

[12] Anna Espona, "Treball de fi de grau d'Anna Espona,"*GitHub*. [En línia]. Disponible a: https://github.com/AnnaEspona/TFG.git. [Accedit Desembre 23, 2018].

[13] Paula Rochina, "Python vs R para el análisis de datos,"*Revistadigital*. [En línia]. Disponible a: https://revistadigital.inesem.es/informatica-y-tics/python-r-analysis-datos/. [Accedit Octubre 5, 2018].

[14] Akash, "Ariline sentiment,"*Kaggle*. [En línia]. Disponible a: https://www.kaggle.com/welkin10/airline-sentiment. [Accedit Octubre 5, 2018].

[15] "Natural Language Processing,"*SAS*. [En línia]. Disponible a: https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html. [Accedit Novembre 10, 2018].

[16] Giuseppe Bonaccorso, *Machine Learning Algorithms*. Birmingham, United Kingdom: Packt, 2017.

[17] "Sentiment Analysis Explained,"*Lexalytics*. [En línia]. Disponible a: https://www.lexalytics.com/technology/sentiment-analysis. [Accedit Gener 19, 2019].

[18] Margaret Rouse, "Opinion mining (sentiment mining),"*SearchBusinessAnalytics*. [En línia]. Disponible a: https://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining. [Accedit Gener 15, 2019].

[19] Andrew Yue Xie, "Sentiment Analysis Dictionary,"*Kaggle*. [En línia]. Disponible a: https://www.kaggle.com/andyxie/sentiment-analysis-dictionary#positive-words.txt. [Accedit Gener 18, 2019].

[20] "Using VADER to handle sentiment analysis with social media text,"[En línia]. Disponible a: http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html. [Accedit Desembre 23, 2018].

[21] Rohith Gandhi, "Support Vector Machine - Introduction to Machine Learning Algorithms,"*Medium Corporation*. [En línia]. Disponible a: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. [Accedit Gener 24, 2019].

[22] Sunil Ray, "Understanding Support Vecto MAchine algorithm from examples (along with code),"*Analytics Vidhya*. [En línia]. Disponible a: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/. [Accedit Gener 24, 2019].

[23] "Support Vector Machines (SVM) Introductory Overview,"*Statsoft*. [En línia]. Disponible a: http://www.statsoft.com/textbook/support-vector-machines. [Accedit Gener 23, 2019].

[24] Prabhu, "Understanding Hyperparameters and its Optimisation techniques,"*Medium Corporation*. [En línia]. Disponible a: https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568. [Accedit Gener 23, 2019].

[25] Usman Malik, "Cros Validation and Grid Search for Model Selection in Python,"*Stack Abuse*. [En línia]. Disponible a: https://stackabuse.com/cross-validation-and-grid-search-for-model-selection-in-python/. [Accedit Gener 23, 2019].

[26] Iwan Syarif, Adam Prugel-Benett i Gary Wills. SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance. *Universitas Ahmad Dahlan*, 2016.

[27] Mohtadi Ben Fraj, "In Depth: Parameter tuning for SVC,"*Medium Corporation*. [En línia]. Disponible a: https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769. [Accedit Gener 25, 2019].

[28] Pushkar Mandot, "What is the Significance of C value in Support Vector Machine?,"*Medium Corporation*. [En línia]. Disponible a: https://medium.com/@pushkarmandot/what-is-the-significance-of-c-value-in-support-vector-machine-28224e852c5a. [Accedit Gener 25, 2019].