Avaluació de la qualitat dels sistemes de reconeixement de sentiments

PID_00257776

Joaquim Moré

Temps mínim de dedicació recomanat: 1 hora

Universitat Oberta de Catalunya

Joaquim Moré

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Jordi Casas (2019)

Primera edició: setembre 2019
Autoria: Joaquim Moré

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

Introducció

La interpretació d'un text i la identificació dels sentiments en les opinions es produeixen sense intervenció humana, aplicant processament del llenguatge natural (PLN). Això es pot veure, per exemple, en la predicció de quines notícies de *The New York Times* provocaran més comentaris, en la predicció del sentiment expressat en una opinió i en la classificació d'opinions falses.

Ara bé, la interpretació i classificació per mitjans plenament automàtics no és infal·lible. Quan un sistema processa textos per predir i classificar, cal avaluar la qualitat d'aquest sistema. La qualitat del sistema depèn fonamentalment de dos factors: les dades processades i el mètode utilitzat per predir i classificar amb aquestes dades.

Les dades poden ser bones, però el mètode pot no ser l'adequat i viceversa. Per això, l'avaluació dels dos factors ha de ser contínua durant el procés d'elaboració del sistema. Cal avaluar la idoneïtat de les dades que es tenen i explorar, també, la introducció de dades noves per millorar els resultats. D'altra banda, cal avaluar els resultats obtinguts amb diferents mètodes (així fins a arribar a un nivell òptim d'adequació entre les dades i el mètode).

En aquest material presentarem primer algunes de les mètriques més utilitzades per fer l'avaluació i ensenyarem a interpretar-les. Posteriorment, s'ensenyarà a aplicar aquestes mètriques en l'avaluació d'un sistema que classifica una opinió com a favorable o desfavorable. A més, aquestes mètriques s'aplicaran en dos casos d'ús que hem vist. El primer cas d'ús és la classificació de titulars de notícies que provoquen comentaris i el segon cas és la detecció d'opinions falses.

A més de les mètriques, s'ensenyarà a avaluar per mitjà de diferents mètodes de classificació i predicció.

1. Mètriques d'avaluació

La predicció i classificació automàtica s'ha fet entrenant el sistema amb dades representatives de les classes que ha de predir o classificar. Ho fa amb un corpus en el qual s'assigna a cada ítem (frase, document, opinió, etc.) una etiqueta de la classe que cal predir o classificar. L'etiqueta la posa una persona, o un conjunt de persones, amb la qual cosa s'acredita que aquest ítem pertany a una classe. Per exemple, una persona etiqueta una opinió com a falsa quan té evidències que és així.

Després arriba el moment en què es deixa que el sistema «faci ell sol» la tasca per a la qual s'ha entrenat. Ho fa amb un corpus de test que també ha estat etiquetat manualment, però el sistema etiqueta aquest corpus sense saber-ho. L'avaluació consisteix principalment a comparar l'assignació d'etiquetes del sistema amb l'assignació d'etiquetes de l'expert. Com més semblant sigui el resultat, més bo serà el sistema.

1.1. La matriu de confusió

La matriu de confusió (a partir d'ara, ***confusion matrix***) és una taula en la qual es mostra la relació d'encerts i errors del sistema. Anomenarem **encert** la coincidència amb l'expert en l'assignació automàtica d'una etiqueta i anomenarem **error** la no coincidència amb l'expert.

En una *confusion matrix* es pren en consideració el caràcter **positiu** i **negatiu**. És positiu quan l'ítem s'ha assignat «positivament» a la classe; és a dir, que s'ha considerat l'ítem com a pertanyent a una classe de referència. En el cas de la detecció d'opinions falses, la classe de referència podria ser *Opinió Falsa* i totes les opinions etiquetades pel sistema amb la classe *Opinió Falsa* serien opinions amb una assignació positiva.

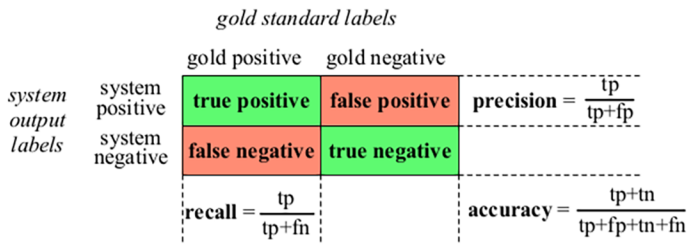
Ara bé, una cosa és que les opinions s'etiquetin positivament i una altra és que ho hagin fet coincidint amb l'etiquetatge de l'expert. Aquí és on entren les nocions de **fals positiu** (*false positive*), **fals negatiu** (*false positive*), **veritable positiu** (*true positive*) i **veritable negatiu** (*true negative*).

La qualificació de **fals** indica que no hi ha hagut coincidència amb l'expert. Així doncs, en la classificació d'opinions falses —que a partir d'ara anomenarem ***fake*** per no confondre'ns amb els sentits de la paraula *fals*—, un **fals positiu** és una opinió classificada com a *fake* que no ha estat etiquetada així per l'expert; en canvi, un **veritable positiu** és una opinió classificada com a *fake* que coincideix amb la classificació de l'expert. D'altra banda, un **fals negatiu** és una opinió que no ha estat classificada com a *fake*, però que ha estat etiquetada com a tal per l'expert, i un **veritable negatiu** és una opinió que el sistema no ha classificat com a *fake*, i l'expert, tampoc.

1.2. Mètriques a partir de la matriu de confusió

A partir de la *confusion matrix* es poden obtenir tres mètriques importants: la **precisió**, l'**exactitud** i la **cobertura**. Per referir-nos a l'exactitud i a la cobertura emprarem les denominacions en anglès *accuracy* i *recall*, respectivament, que són les més habituals. En la figura 1 s'il·lustra la relació entre la precisió, l'*accuracy* i *recall* en la *confusion matrix*.

Figura 1. *Confusion matrix*, precisió i *accuracy*



Font: Jurafsky i Martin (2018). *Speech and Language Processing* (cap. 4, pàg. 73)

1.2.1. *Accuracy*

A partir de l'exemple de la classificació d'opinions *fake*, l'***accuracy*** és simplement el percentatge de totes les opinions que el sistema ha classificat coincidint amb l'etiquetador expert. És, per tant, la suma de les *true positive* i *true negative* dividit per la suma d'opinions en total.

Imaginem ara que les opinions *fake* tenen un percentatge molt petit respecte al total d'opinions i que el classificador classifica opinions que no són *fake*. El percentatge de coincidència, que seria molt gran, validaria el sistema? En realitat, no gaire; sobretot perquè, malgrat la seva alta *accuracy*, el sistema no tindria prou dades per saber quins trets distingeixen les opinions *fake* de les no *fake*. Si les mostres de les dues classes estiguessin equilibrades, llavors sí.

1.2.2. Precisió

La precisió pren com a referència les opinions que el sistema ha classificat com a pertanyents a la classe i que coincideixen amb la classificació de l'expert. La precisió es calcula dividint el nombre de positius veritables per la suma de positius veritables i positius falsos.

1.2.3. Cobertura (*recall*)

La cobertura dona el percentatge d'opinions del corpus d'avaluació que el sistema ha classificat correctament. Es calcula dividint els positius veritables per la suma dels positius veritables més els falsos negatius (és a dir, les vegades que s'ha equivocat).

Així, encara que el sistema encerti a detectar les opinions no *fake*, si no encerta a classificar alguna opinió com a *fake*, la cobertura és 0 i el sistema no passa l'avaluació.

1.2.4. *F-measure* i F1

Generalment, la precisió i la *recall* es combinen en una sola mètrica, l'anomenada *F-measure*, en la qual un pot establir el pes que tindrà la precisió o la *recall* en el càlcul. Quan totes dues tenen el mateix pes, aleshores la *F-measure* s'anomena *F1*.

Si P és la precisió i R és la *recall*, la F1 es calcula així:

$$F_1 = \frac{2PR}{P+R}$$

El valor va del 0 a l'1. Com més a prop de l'1 és, més bo és el sistema.

1.3. La sensibilitat de les dades de la matriu de confusió

Segons l'aplicació que té el sistema, les dades de la matriu de confusió són altament sensibles. Imaginem que classifiquem pacients amb un risc alt de càncer que haurien de fer-se una revisió. En aquest cas, conflueixen dos aspectes crítics. El primer és que cal evitar els falsos negatius; és a dir, pacients que no estan classificats com que haurien de fer-se una revisió quan, de fet, sí que la necessiten. La manera d'evitar el patiment i la pèrdua de vides que suposen els falsos negatius seria que tots els pacients es fessin la revisió, però aquí hi entra el segon factor: fer anàlisis a una multitud de persones que no les necessiten és una solució inviable.

Si es decideix que han de passar la revisió els pacients amb una alta probabilitat de tenir càncer, com més alt posem el llistó més falsos negatius hi haurà, i com més baix el posem més despesa hi haurà en anàlisi per a persones que no en necessiten. Cal una eina que ens digui quin és el valor de probabilitat òptim per discriminar els dos grups.

Un altre exemple de sensibilitat en el tractament de falsos positius i falsos negatius és la classificació automàtica de les opinions sobre restaurants com a favorables o desfavorables. A partir de la classe «opinió favorable» com a referència, els falsos positius es prenen com a bones valoracions del restaurant, quan en realitat no ho són. La conseqüència és que la reputació del restaurant millora injustament i perjudica la competència. A més, produeix una frustració en el client que ha confiat en l'opinió favorable i comprova amb irritació que la qualitat no es correspon amb les seves expectatives. A més, els falsos negatius són devastadors. Les opinions classificades com a «desfavorables» que en realitat no ho són poden produir una mala reputació injustificada que és letal en un negoci en el qual és molt important la recomanació per captar nous clients.

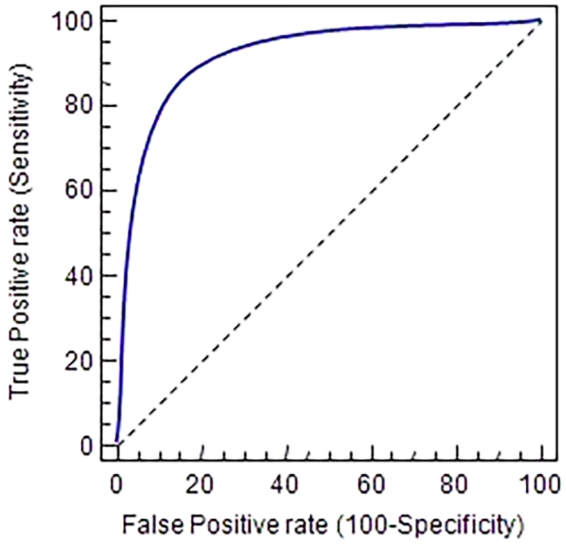
Si el classificador servís per fer un seguiment de les opinions dels clients amb la finalitat d'analitzar quins aspectes del restaurant caldria millorar, els falsos positius serien veritablement perjudicials, perquè enganyarien l'encarregat, que pot tenir l'apreciació errònia que tot va bé i no veure les mancances que els clients assenyalen.

1.3.1. La corba ROC

L'eina que ens permet trobar el valor de probabilitat òptim per classificar un ítem és la que calcula l'anomenada **corba ROC**.

Especificant un llindar de probabilitat, es classifiquen les dades i es calcula la ràtio de falsos positius i veritables positius que resulten de la classificació. El resultat és un punt en una gràfica en què a les x hi ha la ràtio de falsos positius i a les y hi ha la ràtio de falsos negatius. A mesura que es va baixant el llindar, els punts es van disposant de manera que, si s'uneixen els punts, es traça una corba com la de la figura 2.

Figura 2. Representació de la corba ROC



Font: https://www.quora.com/whats-roc-curve

Com veieu, a partir d'un valor de probabilitat de veritables positius la corba es va fent més horitzontal. Per això, com més a prop sigui el valor cap a l'esquerra, més fiable serà el valor de predicció. En el cas de la figura 2, seria un valor de 0.40.

2. Avaluació dels casos d'ús

En aquest apartat explicarem l'aplicació i els resultats de l'avaluació de tres sistemes:

1) El sistema que classifica opinions com a favorables o desfavorables.

2) El sistema que és el classificador que prediu els titulars de les notícies que generen més comentaris.

3) El sistema que és el classificador d'opinions falses.

2.1. Predictor d'una opinió com a favorable o desfavorable

El sistema classifica una opinió publicada a la plataforma Yelp com a favorable o desfavorable. Yelp és una plataforma destinada a recomanar negocis locals als usuaris recollint-ne les opinions.

El cas de Yelp

El cas d'ús està inspirat en un entrada publicada en el blog *The Tensorist* que porta per títol «Sentiment analysis for Yelp review classification». Disponible a: https://medium.com/tensorist/classifying-yelp-reviews-using-nltk-and-scikit-learn-c58e71e962d9

Les opinions tenen una etiqueta indicativa de la valoració de l'usuari. Aquesta etiqueta és un nombre de l'1 al 5, que correspon al nombre d'estrelles que guanya el negoci segons l'opinió de l'usuari. L'etiqueta «1 estrella» correspon a la valoració més baixa, i l'etiqueta «5 estrelles», a la més alta. Tot seguit hi ha el text de l'opinió en què s'argumenta la valoració.

El corpus per desenvolupar el classificador és de 10.000 opinions, de les quals el sistema aprèn a classificar les de cinc estrelles, representatives de les opinions molt favorables, i les d'una estrella, que corresponen a les opinions molt desfavorables. El corpus d'opinions de cinc estrelles i una estrella consta d'unes 4.000 opinions.

En el *notebook* PLA-3 1 es pot veure com s'apliquen els mètodes de preprocessament, entrenament, classificació amb un corpus de test i, finalment, com s'obtenen les dades d'avaluació de la classificació. És interessant veure que un classificador basat en la *logistic regression* obté uns resultats notables amb un preprocessament molt simple. Un preprocessament que obté com a *features* els valors de tf.idf dels *tokens* dels comentaris sense signes de puntuació. La taula 1 mostra les mitjanes ponderades de la precisió, *recall* i F1 de les opinions classificades.

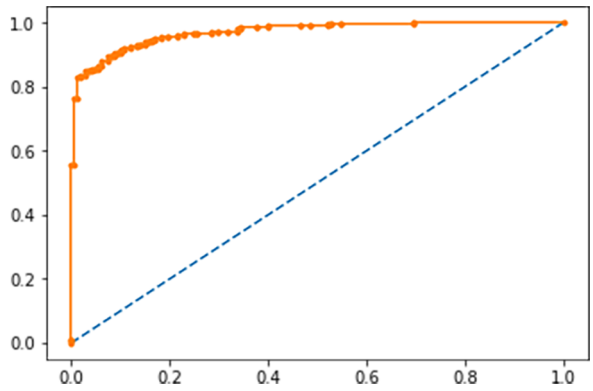
Taula 1. Resultats de l'avaluació del classificador d'opinions Yelp

Precisió	Recall	F1
0.89	0.97	0.92

Són, en efecte, resultats bastant bons, tot i que cal tenir en compte que el corpus recull moltes més opinions de cinc estrelles (2.794) que d'una estrella (646). Hauria estat millor que el nombre d'opinions dels dos tipus fos més equilibrat; no obstant això, són les dades reals i la tendència (o *bias*) de les dades cap a una classe determinada sovint no es pot controlar.

A continuació es mostra la corba ROC corresponent.

Figura 3. Representació de la corba ROC del classificador d'opinions de Yelp



2.2. Avaluació del classificador de titulars del NYT

Recordem que els membres del grup de PLN de l'empresa S&S van classificar els titulars prenent com a *features* els valors de tf.idf dels termes lematitzats dels titulars (vegeu *notebook* PLA-1, apartat 5).

2.2.1. Avaluació del mètode de classificació

A la taula següent presentem els resultats dels titulars classificats en funció de si motiven o no motiven comentaris. Aquests són els resultats obtinguts aplicant un classificador basat en la *logistic regression model* (vegeu *notebook* PLA-3 2.1).

Taula 2. Resultats amb un classificador basat en la *logistic regression model*

Precisió	Recall	F1
0.70	0.84	0.72

Aquests resultats es prenen com a referència per veure si es poden obtenir resultats millors amb altres tipus de classificadors. En Henry decideix comparar els resultats obtinguts amb un classificador bayesià, un classificador SVM (*support vector machine*) i un classificador Random Forest. La taula 3 en recull els resultats. En el *notebook* PLA-3 1.3 hi ha els *scripts* que permeten calcular-los.

Taula 3. Resultats segons el tipus de classificador

Mètode	Precisió	Recall	F1
Logistic regression	0.70	0.84	0.72
Bayesià	0.65	0.83	0.70
SVM	**0.71**	**0.81**	**0.74**
Random Forest	0.68	0.81	0.72

Els resultats indiquen que la classificació de titulars no depèn del mètode de classificació utilitzat. Els resultats són similars en els quatre mètodes. Si es vol millorar el classificador, s'hauran de fer millorant els *features* que el classificador ha d'aprendre.

2.2.2. Nous *features*

En Henry i la Beth plantegen si és possible millorar els resultats afegint *features*. La Beth recorda a en Henry una impressió que va tenir quan va veure els termes agrupats en temes. Es va adonar que, a l'entorn del terme *Trump* hi havia termes que suggerien violència (*trade war, fight, gun*). A en Henry li interessa aquest detall i suggereix afegir al model de titulars una matriu que tingui en compte la connotació positiva i negativa de les paraules del titular. Així doncs, es torna a entrenar el sistema amb el nou model, es deixa que el sistema torni a classificar els titulars i després es comprova si milloren els resultats. El nou classificador es coneixerà com a CLASSIFICADOR-2, per distingir-lo del classificador anterior (CLASSIFICADOR-1).

El vectoritzador per crear aquesta nova matriu tindrà un *analyzer* que consultarà, per a cada *token* dels titulars, si es troba en un diccionari d'*opinion words* amb valors de polaritat. El diccionari utilitzat és l'AFINN . A la taula 4 se'n pot veure una mostra.

Taula 4. Mostra del diccionari *AFINN*

vulnerability	–2
vulnerable	–2
walkout	–2
walkouts	–2
wanker	–3

want	1
war	–2
warfare	–2
warm	1
warmth	2
warn	–2
warned	–2
warning	–3

El vocabulari serà el conjunt de termes que tenen un valor de negativitat i el vectoritzador posarà els seus valors als índexs corresponents.

Els resultats obtinguts comparant mètodes de classificació són els següents:

Taula 5. Comparació de resultats del CLASSIFICADOR-1 i del CLASSIFICADOR-2 segons els mètodes de classificació

CLASSIFICADOR 1

Mètode	Precisió	Recall	F1
Logistic regression	0.70	0.84	0.72
Bayesià	0.65	0.83	0.70
SVM	**0.71**	**0.81**	**0.74**
Random Forest	0.68	0.81	0.72

CLASSIFICADOR 2

Mètode	Precisió	Recall	F1
Logistic regression	0.71	0.84	0.72
Bayesià	0.72	0.85	0.73
SVM	**0.70**	**0.81**	**0.74**
Random Forest	**0.71**	**0.82**	**0.74**

Com es veu a la taula 5, la introducció dels *features* de polaritat ha millorat els resultats en els mètodes de *logistic regression*, bayesià i Random Forest. El mètode en què la millora és notable és el mètode bayesià. Ara el classificador Random Forest és amb l'SVM com a classificadors millors.

2.3. Avaluació del classificador d'opinions falses

La Beth es pregunta si les característiques de les opinions falses que han trobat suposaran fer un classificador a la mesura d'aquestes característiques. Però en Henry creu que amb un vectoritzador com el de les opinions de Yelp, el classificador aprendrà les característiques de les opinions falses.

Així doncs, obtenen els resultats que es mostren a la taula 6, després d'haver entrenat el sistema amb els quatre mètodes de classificació.

Taula 6. Comparació de resultats del classificador d'opinions falses segons els mètodes de classificació

	CLASSIFICADOR		
Mètode	**Precisió**	**Recall**	**F1**
Logistic regression	**0.80**	**0.80**	**0.80**
Bayesià	0.78	0.75	0.74
SVM	0.80	0.79	0.79
Random Forest	0.60	0.59	0.59

Com es pot veure, el millor resultat s'obté amb un classificador basat en la *logistic regression*. Amb tot, el classificador Random Forest sembla que és el menys indicat.

2.4. Anàlisi dels resultats i nous reptes

En Henry i la Beth presenten a en Peter, el coordinador del grup, els resultats de l'avaluació de la classificació de titulars i d'opinions falses. Els valors de F1 de tots dos sistemes de classificació poden arribar al voltant del 0.8. A en Peter els resultats li semblen correctes.

Com a mínim, aquests resultats s'han obtingut evitant l'*overfitting*.

> L'***overfitting*** es produeix quan el sistema és capaç de classificar molt bé les dades amb què s'ha entrenat, però no classifica bé una dada nova.

Precisament, la preparació per separat d'un corpus d'entrenament i de test s'ha fet per evitar-ho i, en el cas d'haver-se produït, els valors de F1 haurien estat sospitosament superiors.

Els resultats d'una tasca feta amb un aprenentatge automàtic haurien de comparar-se amb els resultats de la mateixa tasca feta per una persona externa al projecte. És interessant fer aquesta comparació, quan és possible, per comprovar que en la classificació humana hi ha marges d'error comparables als

del sistema. En Peter creu que si una persona classifiqués les opinions falses, per exemple, aplicaria criteris subjectius i els nivells de coincidència amb l'etiquetatge de referència serien semblants.

Encara que els resultats siguin acceptables, l'anàlisi dels titulars i les opinions falses suggereixen noves preguntes; preguntes que en Peter es planteja i que presentem a continuació per als que vulguin desenvolupar-ne una resposta. Algunes d'aquestes preguntes són:

1) Hi ha una manera de classificar els titulars de *The New York Times* motivadors de comentaris que no depenguin de si Trump és el president?

2) Atès que la sola referència a una persona provoca comentaris i controvèrsia, es podria considerar aquesta referència com a *opinion word*? És a dir, les referències a Hitler, Trump, Aznar o Mandela no incideixen ja en la polaritat d'un text?

3) Quins serien els resultats si s'apliqués un classificador neuronal aplicant *word embeddings*?

4) Milloraríem els resultats de la classificació de titulars si afegíssim com a *features* les distàncies semàntiques entre els termes segons ConceptNet?

5) Per què el mètode de Random Forest té uns resultats sensiblement inferiors als altres mètodes en la classificació d'opinions falses i no en la classificació de titulars?

6) És cert que si un de nosaltres hagués de classificar un titular o una opinió falsa tindria un nivell de qualitat semblant al del sistema?

Resum

En aquest mòdul hem ensenyat com cal avaluar un sistema que fa una tasca relacionada amb el *sentiment analysis* de manera automàtica. Ho hem exemplificat avaluant un sistema que classifica opinions com a favorables o desfavorables i amb sistemes preparats per a altres casos d'ús. Hem avaluat un sistema que classifica titulars motivadors de comentaris i també hem avaluat un sistema de classificació d'opinions veritables i falses.

Hem explicat les nocions bàsiques en l'avaluació de sistemes que aprenen una tasca automàticament i també hem comprovat com s'obtenen resultats diferents segons l'algorisme adoptat per entrenar el sistema. A més, hem ensenyat com s'han d'afegir dades a l'entrenament i com cal comprovar en quina mesura la introducció de noves dades millora (o no) el sistema.

Els resultats obtinguts han estat acceptables. Ara bé, com tot, sempre es poden millorar. Animem els estudiants a pensar i a proposar procediments i algorismes que millorin els resultats. A més, hem presentat unes preguntes, que també poden ser interpretades com a reptes, per aprofundir encara més en el *sentiment analysis* i en el processament de llenguatge natural. Seria desitjable que la cerca de respostes sigui igual de fascinant que les noves preguntes que, probablement, es plantejaran els estudiants.