Evaluación de la calidad de los sistemas de reconocimiento de sentimientos

PID_00257777

Joaquim Moré

Tiempo mínimo de dedicación recomendado: 1 hora

Universitat Oberta de Catalunya

Joaquim Moré

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Jordi Casas (2019)

Primera edición: septiembre 2019
Autoría: Joaquim Moré
Licencia CC BY-SA de esta edición, FUOC, 2019
Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción

La interpretación de un texto y la identificación de los sentimientos en las opiniones se producen sin intervención humana, aplicando procesamiento del lenguaje natural (PLN). Se puede ver, por ejemplo, en la predicción de qué noticias del *New York Times* provocarán más comentarios, en la predicción del sentimiento expresado en una opinión y en la clasificación de opiniones falsas.

Ahora bien, la interpretación y clasificación por medios plenamente automáticos no es infalible. Cuando un sistema procesa textos para predecir y clasificar, hay que evaluar la calidad de este sistema. La calidad del sistema depende fundamentalmente de dos factores: los datos procesados y el método utilizado para predecir y clasificar con estos datos.

Los datos pueden ser buenos pero el método no ser el adecuado y viceversa. Por ello, la evaluación de los dos factores debe ser continua durante el proceso de elaboración del sistema. Hay que evaluar la idoneidad de los datos que se tienen y explorar también la introducción de datos nuevos para mejorar los resultados. Por otro lado, hay que evaluar los resultados obtenidos con distintos métodos. Así hasta llegar a un nivel óptimo de adecuación entre los datos y el método.

En este material presentaremos primero algunas de las métricas más utilizadas para hacer la evaluación y enseñaremos a interpretarlas. Posteriormente, se enseñará a aplicar estas métricas en la evaluación de un sistema que clasifica una opinión como favorable o desfavorable. Además, estas métricas se aplicarán en dos casos de uso que hemos visto. El primer caso de uso es la clasificación de titulares de noticias que provocan comentarios, y el segundo caso es la detección de opiniones falsas.

Además de las métricas, se enseñará a evaluar probando distintos métodos de clasificación y predicción.

1. Métricas de evaluación

La predicción y clasificación automática se ha hecho entrenando el sistema con datos representativos de las clases que tiene que predecir o clasificar. Lo hace con un corpus en el que a cada ítem (frase, documento, opinión, etc.) se asigna una etiqueta de la clase que hay que predecir o clasificar. La etiqueta la pone una persona, o un conjunto de personas, con lo cual se acredita que ese ítem pertenece a una clase. Por ejemplo, una persona etiqueta una opinión como falsa cuando tiene evidencias de que es así.

Luego viene el momento en que al sistema «se le deja solo» haciendo la tarea para la cual se ha entrenado. Lo hace con un corpus de test que también ha sido etiquetado manualmente, pero el sistema etiqueta este corpus sin saberlo. La evaluación consiste principalmente en comparar la asignación de etiquetas del sistema con la asignación de etiquetas del experto. Cuanto más parecido sea el resultado, mejor será el sistema.

1.1. La matriz de confusión

La matriz de confusión (a partir de ahora, ***confusion matrix***) es una tabla en la que se muestra la relación de aciertos y errores del sistema. Llamaremos **acierto** a la coincidencia con el experto en la asignación automática de una etiqueta y llamaremos **error** a la no coincidencia con el experto.

En una *confusion matrix* se toma en consideración el carácter **positivo** y **negativo**. Positivo se refiere a que el ítem se ha asignado «positivamente» a la clase; es decir, que se ha considerado el ítem como perteneciente a una clase de referencia. En el caso de la detección de opiniones falsas, la clase de referencia podría ser *Opinión Falsa* y todas las opiniones etiquetadas por el sistema con la clase *Opinión Falsa* serían opiniones con una asignación positiva.

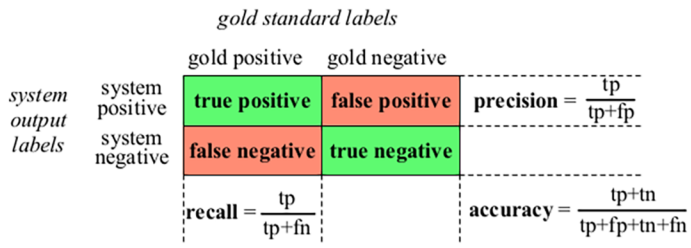
Ahora bien, una cosa es que las opiniones se etiqueten positivamente y otra es que lo hayan hecho coincidiendo con el etiquetaje del experto. Aquí es donde entran las nociones de **falso positivo** (*false positive*), **falso negativo** (*false positive*), **verdadero positivo** (*true positive*) y **verdadero negativo** (*true negative*).

La calificación de **falso** indica que no ha habido coincidencia con el experto. Así pues, en la clasificación de opiniones falsas —cuya clase, a partir de ahora, llamaremos ***fake*** para no confundirnos con los sentidos de la palabra falso—, un **falso positivo** es una opinión clasificada como *fake* que no ha sido etiquetada así por el experto; en cambio, un **verdadero positivo** es una opinión clasificada como *fake* que coincide con la clasificación del experto. Por otro lado, un **falso negativo** es una opinión que no ha sido clasificada como *fake*, pero que ha sido etiquetada como tal por el experto; y un **verdadero negativo** es una opinión que el sistema no ha clasificado como *fake* y el experto tampoco.

1.2. Métricas a partir de la matriz de confusión

A partir de la *confusion matrix* se pueden obtener tres importantes métricas: la **precisión**, la **exactitud** y la **cobertura** (*recall*). Para referirnos a la exactitud y a la cobertura usaremos las denominaciones en inglés *accuracy* y *recall*, respectivamente, que son las más habituales. En la figura 1 se ilustra la relación entre la precisión, la *accuracy* y la *recall* en la *confusion matrix*.

Figura 1. *Confusionmatrix*, precisión y *accuracy*



Fuente: Jurafsky y Martin (2018). *Speech and Language Processing* (cap. 4, pág. 73)

1.2.1. *Accuracy*

Si tomamos el ejemplo de la clasificación de opiniones *fake*, la ***accuracy*** es simplemente el porcentaje de todas las opiniones que el sistema ha clasificado coincidiendo con el etiquetador experto. Es, por tanto, la suma de las *true positive* y *true negative* dividido por la suma de opiniones en total.

Imaginemos ahora que las opiniones *fake* tuvieran un porcentaje muy pequeño respecto al total de opiniones y que el clasificador clasificara opiniones que no son *fake*. ¿El porcentaje de coincidencia, que sería muy grande, validaría el sistema? En realidad, no mucho; sobre todo porque, a pesar de su alta *accuracy*, el sistema no tendría suficientes datos para saber qué rasgos distinguen las opiniones *fake* de las no *fake*. Si las muestras de las dos clases estuvieran equilibradas, entonces sí.

1.2.2. Precisión

La precisión toma como referencia las opiniones que el sistema ha clasificado como pertenecientes a la clase y que coinciden con la clasificación del experto. La precisión se calcula dividiendo el número de positivos verdaderos por la suma de positivos verdaderos y positivos falsos.

1.2.3. Cobertura (*recall*)

La cobertura da el porcentaje de opiniones del corpus de evaluación que el sistema ha clasificado correctamente. Se calcula dividiendo los positivos verdaderos por la suma de los positivos verdaderos más los falsos negativos (es decir, las veces que se ha equivocado).

Así, aunque el sistema acierte en detectar las opiniones no *fake*, si no acierta en clasificar alguna opinión como *fake*, la cobertura es 0 y el sistema no pasa la evaluación.

1.2.4. *F-measure* y F1

Generalmente, la precisión y la *recall* se combinan en una sola métrica, la llamada *F-measure*, en la que uno puede establecer el peso que tendrá la precisión o la *recall* en el cálculo. Cuando ambas tienen el mismo peso, entonces a la *F-measure* se le llama F1.

Siendo P la precisión y R la *recall*, la F1 se calcula de la siguiente manera:

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

El valor va del 0 al 1. Cuanto más cerca del 1 está, más bueno es el sistema.

1.3. La sensibilidad de los datos de la matriz de confusión

Según la aplicación que tiene el sistema, los datos de la matriz de confusión son altamente sensibles. Imaginemos que clasificamos pacientes con un alto riesgo de cáncer que deberían hacerse una revisión. En este caso, confluyen dos aspectos críticos. El primero es que hay que evitar los falsos negativos, esto es, pacientes que no están clasificados como que deberían hacerse una revisión cuando en realidad sí que la necesitan. La forma de evitar todo el sufrimiento y la pérdida de vidas que suponen los falsos negativos sería que todos los pacientes se hicieran la revisión, pero aquí entra el segundo factor: realizar los análisis a una multitud de personas que no los necesitan es una solución inviable.

Si se decide que tienen que pasar la revisión los pacientes con una alta probabilidad de tener cáncer, cuanto más alto pongamos el listón más falsos negativos habrá, y cuanto más bajo lo pongamos más gasto en análisis para personas que no lo necesitan. Necesitamos una herramienta que nos diga cuál es el valor de probabilidad óptimo para discriminar los dos grupos.

Otro ejemplo de sensibilidad en el tratamiento de falsos positivos y falsos negativos es la clasificación automática de las opiniones sobre restaurantes como favorables o desfavorables. Tomando la clase «opinión favorable» como referencia, los falsos positivos se toman como buenas valoraciones del restaurante, cuando en realidad no lo son. La consecuencia es que la reputación del restaurante mejora injustamente y perjudica a la competencia. Además, produce una frustración en el cliente que ha confiado en la opinión favorable y comprueba con irritación que la calidad no se corresponde con sus expectativas. Por su parte, los falsos negativos son devastadores. Las opiniones clasificadas como «desfavorables» que en realidad no lo son pueden producir una mala reputación injustificada que es letal en un negocio en el que es tan importante la recomendación para la captación de nuevos clientes.

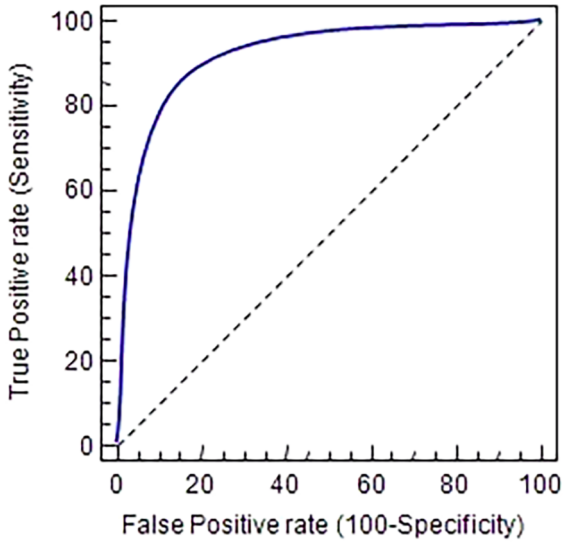
Si el clasificador sirviera para hacer un seguimiento de las opiniones de los clientes con el fin de analizar qué aspectos del restaurante habría que mejorar, los falsos positivos serían verdaderamente perjudiciales, porque engañarían al encargado, el cual puede tener la apreciación errónea de que todo va bien y no ver las carencias que los clientes demandan.

1.3.1. La curva ROC

La herramienta que nos permite encontrar el valor de probabilidad óptimo para clasificar un ítem es la que calcula la llamada **curva ROC**.

Especificando un umbral de probabilidad, se clasifican los datos y se calcula el ratio de falsos positivos y verdaderos positivos que resultan de la clasificación. El resultado es un punto en una gráfica donde en las x hay el ratio de falsos positivos y en las y el ratio de falsos negativos. A medida que se va bajando el umbral, los puntos se van disponiendo de modo que se si se unen los puntos, se traza una curva como en la figura 2.

Figura 2. Representación de la curva ROC



Fuente: https://www.quora.com/Whats-ROC-curve

Vemos que a partir de un valor de probabilidad de verdaderos positivos la curva se va aplanando. Por ello, cuanto más cerca esté el valor hacia la izquierda, más fiable será el valor de predicción. En el caso de la figura 2, sería un valor de 0.40.

2. Evaluación de los casos de uso

En este apartado explicaremos la aplicación y los resultados de la evaluación de tres sistemas:

1) El sistema que clasifica opiniones como favorables o desfavorables.

2) El sistema que es el clasificador que predice los titulares de las noticias que generan más comentarios.

3) El sistema que es el clasificador de opiniones falsas.

2.1. Predictor de una opinión como favorable o desfavorable

Caso Yelp

El caso de uso está inspirado en un entrada publicada en el blog *The Tensorist* titulado «Sentiment Analysis for Yelp review classification». Disponible en: https://medium.com/tensorist/classifying-yelp-reviews-using-nltk-and-scikit-learn-c58e71e962d9

El sistema clasifica una opinión publicada en la plataforma Yelp como favorable o desfavorable. Yelp es una plataforma destinada a recomendar negocios locales a los usuarios recogiendo sus opiniones.

Las opiniones tienen una etiqueta indicativa de la valoración del usuario. Esta etiqueta es un número del 1 al 5, que corresponde al número de estrellas que merece el negocio según la opinión del usuario. La etiqueta «1 estrella» corresponde a la valoración más baja y la etiqueta «5 estrellas» a la más alta. Seguidamente, está el texto de la opinión donde se argumenta la valoración.

El corpus para desarrollar el clasificador es de 10.000 opiniones, de las cuales el sistema aprende a clasificar las de 5 estrellas, representativas de las opiniones muy favorables, y las de 1 estrella, que representan las opiniones muy desfavorables. El corpus de opiniones de 5 estrellas y 1 estrella consta de unas 4.000 opiniones.

En el *notebook* PLA-3 1 se puede ver cómo se aplican los métodos de preprocesado, entrenamiento, clasificación con un corpus de test y, por último, cómo se obtienen los datos de evaluación de la clasificación. Es interesante ver que un clasificador basado en la *logistic regression* obtiene unos resultados notables con un preprocesado muy simple. Un preprocesado que obtiene como *features* los valores de tf.idf de los *tokens* de los comentarios sin signos de puntuación. La tabla 1 muestra las medias ponderadas de la precision, *recall* y F1 de las opiniones clasificadas.

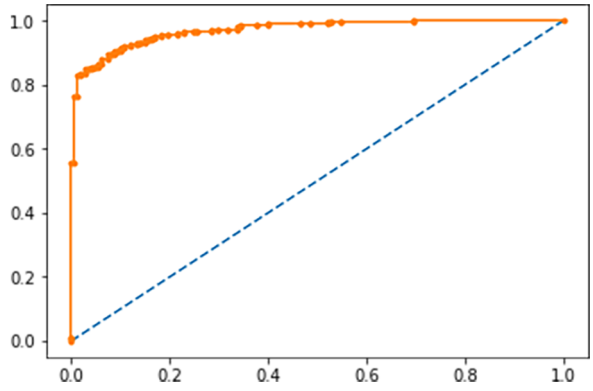
Tabla 1. Resultados de la evaluación del clasificador de opiniones Yelp

Precisión	Recall	F1
0.89	0.97	0.92

Son, en efecto, resultados bastante buenos, aunque hay que tener en cuenta el hecho de que el corpus recoge muchas más opiniones de 5 estrellas (2.794) que de 1 estrella (646). Hubiera sido mucho mejor que el número de opiniones de los dos tipos estuviera más equilibrado; sin embargo, son los datos reales y la tendencia (o *bias*) de los datos hacia una determinada clase a menudo no se puede controlar.

A continuación se muestra su curva ROC.

Figura 3. Representación de la curva ROC del clasificador de opiniones de Yelp



2.2. Evaluación del clasificador de titulares del NYT

Recordemos que los miembros del grupo de PLN de la empresa S&S clasificaron los titulares tomando como *features* los valores de tf.idf de los términos lematizados de los titulares (véase *notebook* PLA-1, apartado 5).

2.2.1. Evaluación del método de clasificación

En la siguiente tabla presentamos los resultados de los titulares clasificados en función de si motivan o no comentarios. Estos son los resultados obtenidos aplicando un clasificador basado en la *logistic regression model* (véase *notebook* PLA-3 2.1).

Tabla 2. Resultados con un clasificador basado en la *logistic regression model*

Precisión	Recall	F1
0.70	0.84	0.72

Estos resultados se toman como referencia para ver si se pueden obtener resultados mejores con otros tipos de clasificador. Henry decide comparar los resultados obtenidos con un clasificador bayesiano, un clasificador SVM (*Support Vector Machine*) y un clasificador Random Forest. En la tabla 3 se pueden ver los resultados. En el *notebook* PLA-3 1.3 están los *scripts* que los calculan.

Tabla 3. Resultados según el tipo de clasificador

Método	Precisión	Recall	F1
Logistic regression	0.70	0.84	0.72
Bayesiano	0.65	0.83	0.70
SVM	**0.71**	**0.81**	**0.74**
Random Forest	0.68	0.81	0.72

Los resultados indican que la clasificación de titulares no depende del método de clasificación utilizado. Los resultados son similares en los cuatro métodos. Si se quiere mejorar el clasificador, se tendrá que hacer mejorando los *features* que el clasificador tiene que aprender.

2.2.2. Nuevos *features*

Henry y Beth plantean si es posible mejorar los resultados añadiendo *features*. Beth le recuerda a Henry una impresión que tuvo cuando vio los términos agrupados en temas. Se dio cuenta de que, en torno al término *Trump*, había términos que sugerían violencia (*trade war, fight, gun*). A Henry le interesa este detalle y sugiere añadir al modelo de titulares una matriz que tenga en cuenta la connotación positiva y negativa de las palabras del titular. Así que se vuelve a entrenar el sistema con el nuevo modelo, se deja al sistema que vuelva a clasificar los titulares y luego se comprueba si mejoran los resultados. El nuevo clasificador se conocerá como CLASIFICADOR-2, para distinguirlo del clasificador anterior (CLASIFICADOR-1).

El vectorizador para crear esta nueva matriz tendrá un *analyzer* que consultará, para cada *token* de los titulares, si se encuentra en un diccionario de *opinion words* con valores de polaridad. El diccionario utilizado es el *AFINN*. En la tabla 4 se puede ver una muestra.

Tabla 4. Muestra del diccionario *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 vocabulario será el conjunto de términos que tienen un valor de negatividad y el vectorizador pondrá sus valores en los correspondientes índices.

Los resultados obtenidos comparando métodos de clasificación son los siguientes:

Tabla 5. Comparación de resultados del CLASIFICADOR-1 y CLASIFICADOR-2 según métodos de clasificación

CLASIFICADOR 1

Método	Precisión	Recall	F1
Logistic regression	0.70	0.84	0.72
Bayesiano	0.65	0.83	0.70
SVM	**0.71**	**0.81**	**0.74**
Random Forest	0.68	0.81	0.72

CLASIFICADOR 2

Método	Precisión	Recall	F1
Logistic regression	0.71	0.84	0.72
Bayesiano	0.72	0.85	0.73
SVM	**0.70**	**0.81**	**0.74**
Random Forest	**0.71**	**0.82**	**0.74**

Como se ve en la tabla 5, la introducción de los *features* de polaridad ha mejorado los resultados en los métodos de *logistic regression*, bayesiano y Random Forest. El método donde la mejora es mayor es el método bayesiano. Ahora el clasificador Random Forest está con el SVM como los clasificadores mejores.

2.3. Evaluación del clasificador de opiniones falsas

Beth se pregunta si las características de las opiniones falsas que han encontrado supondrán hacer un clasificador a la medida de estas características. Henry, sin embargo, cree que con un vectorizador como el de las opiniones de Yelp, el clasificador aprenderá las características de las opiniones falsas.

Así pues, obtienen los resultados que se muestran en la tabla 6 después de haber entrenado el sistema con los cuatro métodos de clasificación.

Tabla 6. Comparación de resultados del clasificador de opiniones falsas según métodos de clasificación

	CLASIFICADOR		
Método	**Precisión**	**Recall**	**F1**
Logistic regression	**0.80**	**0.80**	**0.80**
Bayesiano	0.78	0.75	0.74
SVM	0.80	0.79	0.79
Random Forest	0.60	0.59	0.59

Como pueden ver, el resultado mejor se obtiene con un clasificador basado en la *logistic regression*. El clasificador Random Forest, sin embargo, parece ser el menos indicado.

2.4. Análisis de los resultados y nuevos retos

Henry y Beth presentan a Peter, el coordinador del grupo, los resultados de la evaluación de la clasificación de titulares y de opiniones falsas. Los valores de F1 de ambos sistemas de clasificación pueden llegar a alrededor del 0.8. A Peter los resultados le parecen correctos.

Al menos estos resultados se han obtenido evitando el *overfitting*.

> El ***overfitting*** se produce cuando el sistema es capaz de clasificar muy bien los datos con los que se ha entrenado, pero no clasifica bien un dato nuevo.

Precisamente, la preparación por separado de un corpus de entrenamiento y de test se ha hecho para evitarlo y, en el caso de haberse producido, los valores de F1 habrían sido sospechosamente superiores.

Los resultados de una tarea hecha con un aprendizaje automático deberían compararse con los resultados de la misma tarea hecha por una persona externa al proyecto. Es interesante hacer esta comparación, cuando es posible, para

comprobar que en la clasificación humana hay márgenes de error comparables a los del sistema. Peter cree que si una persona clasificara las opiniones falsas, por ejemplo, aplicaría criterios subjetivos y los niveles de coincidencia con el etiquetaje de referencia serían parecidos.

Aunque los resultados son aceptables, el análisis de los titulares y las opiniones falsas sugieren nuevas preguntas. Preguntas que Peter se plantea y que presentamos a continuación para quien quiera desarrollar una respuesta. Algunas de estas preguntas son:

1) ¿Hay una manera de clasificar los titulares del *New York Times* motivadores de comentarios que no dependa de que Trump sea el presidente?

2) Visto que la sola referencia a una persona provoca comentarios y controversia, ¿se podría considerar esta referencia como un *opinion word*? Es decir, ¿referencias a Hitler, Trump, Aznar o Mandela no inciden ya en la polaridad de un texto?

3) ¿Cuáles serían los resultados si se aplicara un clasificador neuronal aplicando *word embeddings*?

4) ¿Mejoraríamos los resultados de la clasificación de titulares si añadiéramos como *features* las distancias semánticas entre los términos según ConceptNet?

5) ¿Por qué el método de Random Forest tiene unos resultados sensiblemente inferiores a los demás métodos en la clasificación de opiniones falsas y no en la clasificación de titulares?

6) ¿Es cierto que si alguno de nosotros tuviera que clasificar un titular o una opinión falsa tendría un nivel de calidad parecido al del sistema?

Resumen

En este módulo hemos enseñado cómo evaluar un sistema que realiza una tarea relacionada con el *sentiment analysis* de forma automática. Lo hemos ejemplificado evaluando un sistema que clasifica opiniones como favorables o desfavorables y con sistemas preparados para otros casos de uso. Hemos evaluado un sistema que clasifica titulares motivadores de comentarios y también hemos evaluado un sistema de clasificación de opiniones verdaderas y falsas.

Hemos explicado las nociones básicas en la evaluación de sistemas que aprenden una tarea de forma automática y también hemos comprobado cómo se obtienen resultados distintos según el algoritmo adoptado para entrenar el sistema. Por otra parte, hemos enseñado cómo añadir datos en el entrenamiento y comprobar en qué medida la introducción de nuevos datos mejora (o no) el sistema.

Los resultados obtenidos han sido aceptables. Ahora bien, como todo, siempre se pueden mejorar. Animamos a los alumnos a pensar y proponer procedimientos y algoritmos que mejoren los resultados. Además, hemos presentado unas preguntas, que pueden ser también interpretadas como retos, para profundizar aún más en el *sentiment analysis* y en el procesamiento del lenguaje natural. Lo deseable es que la búsqueda de respuestas sea igual de fascinante que las nuevas preguntas que, seguro, los alumnos se van a plantear.