Métricas y métodos de evaluación en Procesamiento del Lenguaje Natural (PLN)

Elena Lloret Pastor

elloret@dlsi.ua.es

Materiales creados a partir de los materiales de los profesores Yoan Gutiérrez y José Ignacio Abreu

experto en procesamiento del lenguaje natural

Universitat d'Alacant Universidad de Alicante
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
UJI UNIVERSITAT JAUME I
valgrAI

Índice

1. Evaluación en PLN
2. Tipos de evaluación
3. Esquemas generales de evaluación de modelos
4. Métricas de evaluación
5. Protocolo de evaluación
6. Caso práctico
7. Bibliografía de referencia

experto en procesamiento del lenguaje natural

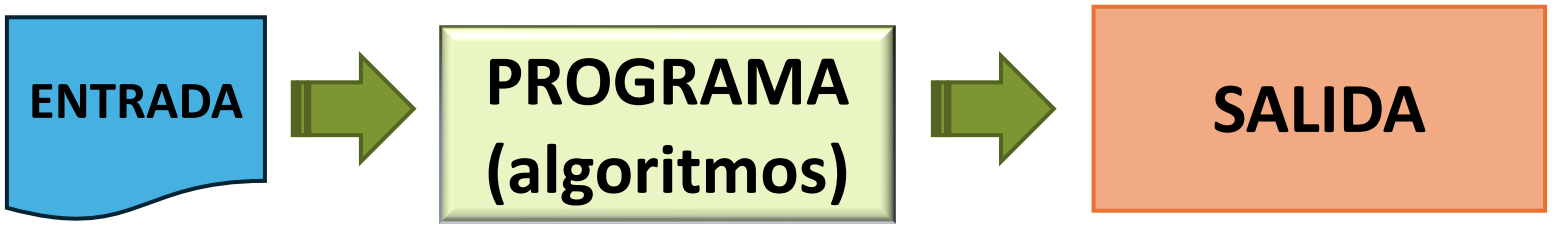
Índice

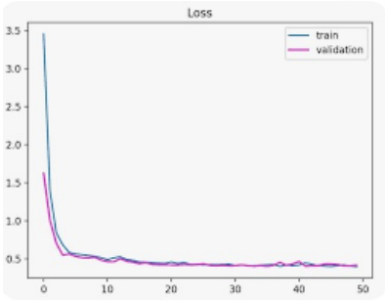
1. **Evaluación en PLN**
2. Tipos de evaluación
3. Esquemas generales de evaluación de modelos
4. Métricas de evaluación
5. Protocolo de evaluación
6. Caso práctico
7. Bibliografía de referencia

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Evaluar
  - Estimar, apreciar, calcular el valor de algo (RAE, 2024)

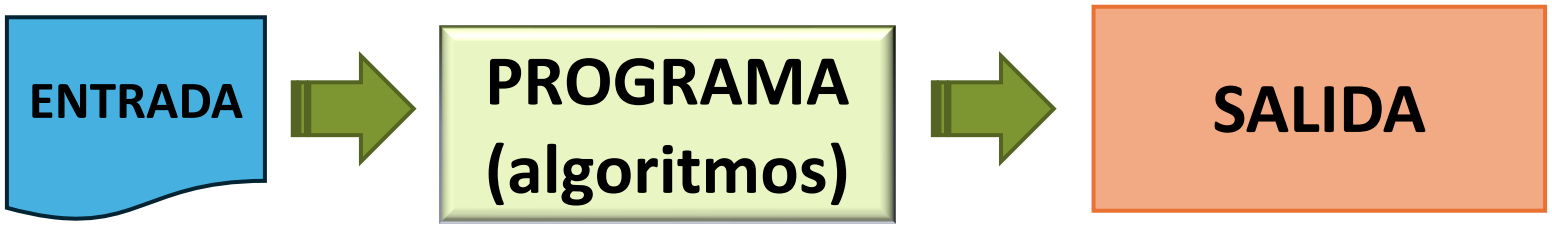


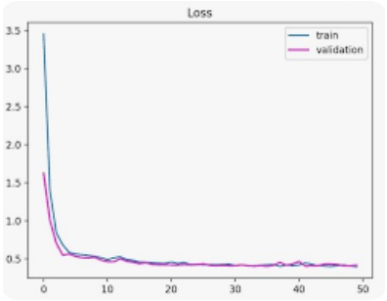


experto en procesamiento del lenguaje natural

Evaluación en PLN

- Evaluar
  - Estimar, apreciar, calcular el valor de algo (RAE, 2024)



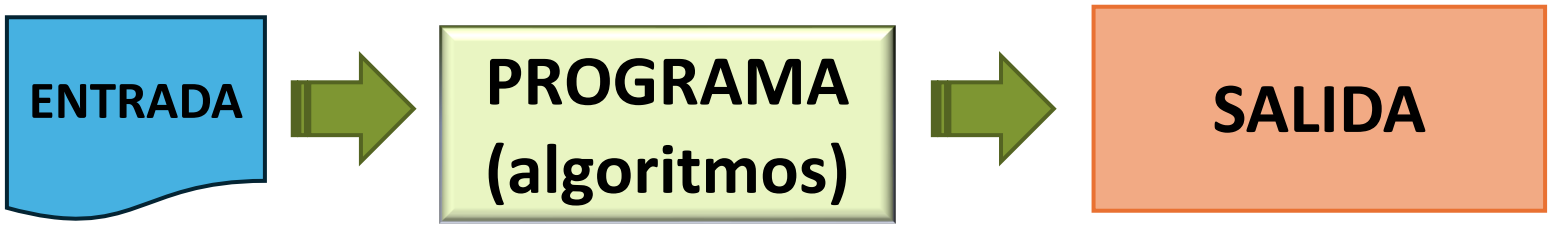


Determinar la calidad y el rendimiento de los modelos/enfoques de PLN

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Evaluar
  - Estimar, apreciar, calcular el valor de algo (RAE, 2024)

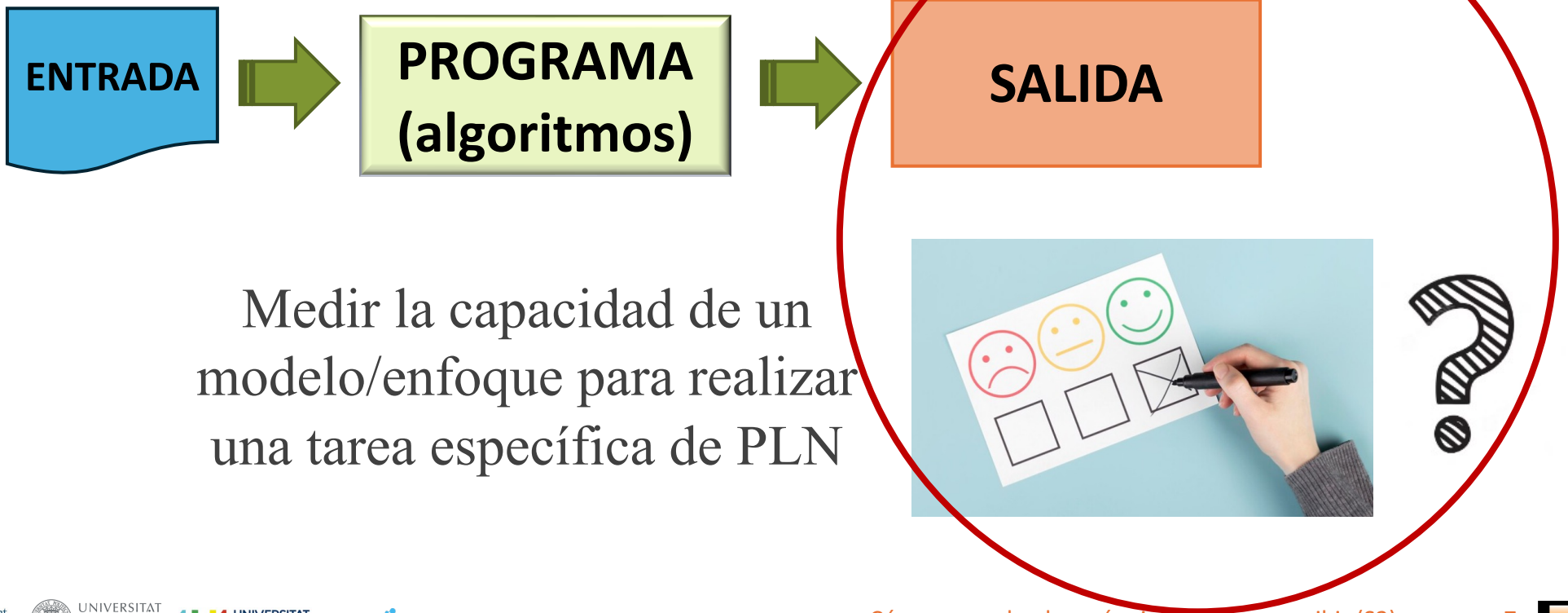


Medir la capacidad de un modelo/enfoque para realizar una tarea específica de PLN

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Evaluar
  - Estimar, apreciar, calcular el valor de algo (RAE, 2024)

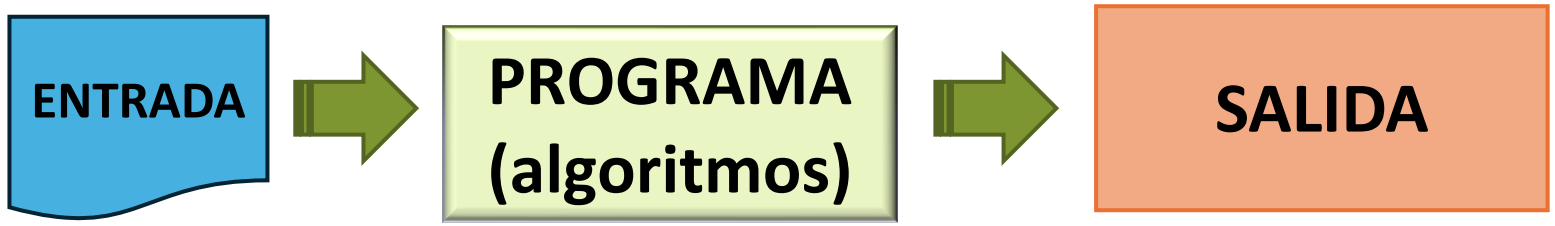


Medir la capacidad de un modelo/enfoque para realizar una tarea específica de PLN

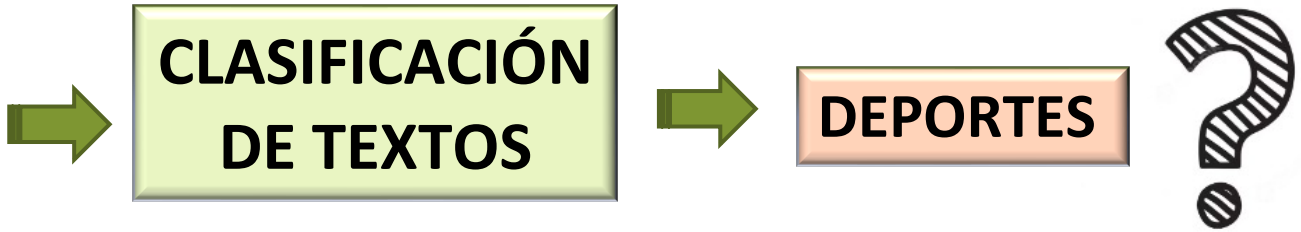
experto en procesamiento del lenguaje natural

Evaluación en PLN

- Evaluar



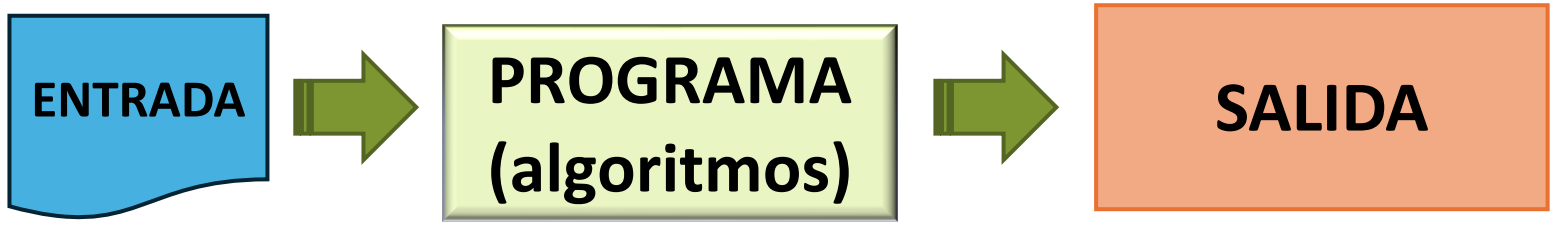
No afloja Carlos Alcaraz, disfrutón y a la vez serio. Lanzado. A lomos todavía de esa fantástica ola que cogió en Indian Wells y que sigue surfeando una semana después, en Miami. Tiemblan las pelotas, despeluchadas por la humedad y, sobre todo, por los constantes zurriagazos del murciano, tan violento como sofisticado en el golpeo: tan pronto revienta la bola como traza una dejada milimétrica y delicada, o bien detiene el tiempo. Todo lo hace, y casi todo lo hace bien en este sábado en el que a Roberto Carballés, gran tenista y magnífico tipo, le salta la risa floja mientras esprinta para intentar devolver uno de esos globos. Piernas y más piernas, porque es bien veloz, pero nada que hacer. La parábola es perfecta, él asiente y se pregunta: ¿Cómo demonios lo hace? Después, el marcador desemboca en el definitivo 6-2 y 6-1 (en 1h 25m) y el murciano saborea otra victoria. La valiosa rutina de estos días.



experto en procesamiento del lenguaje natural

Evaluación en PLN

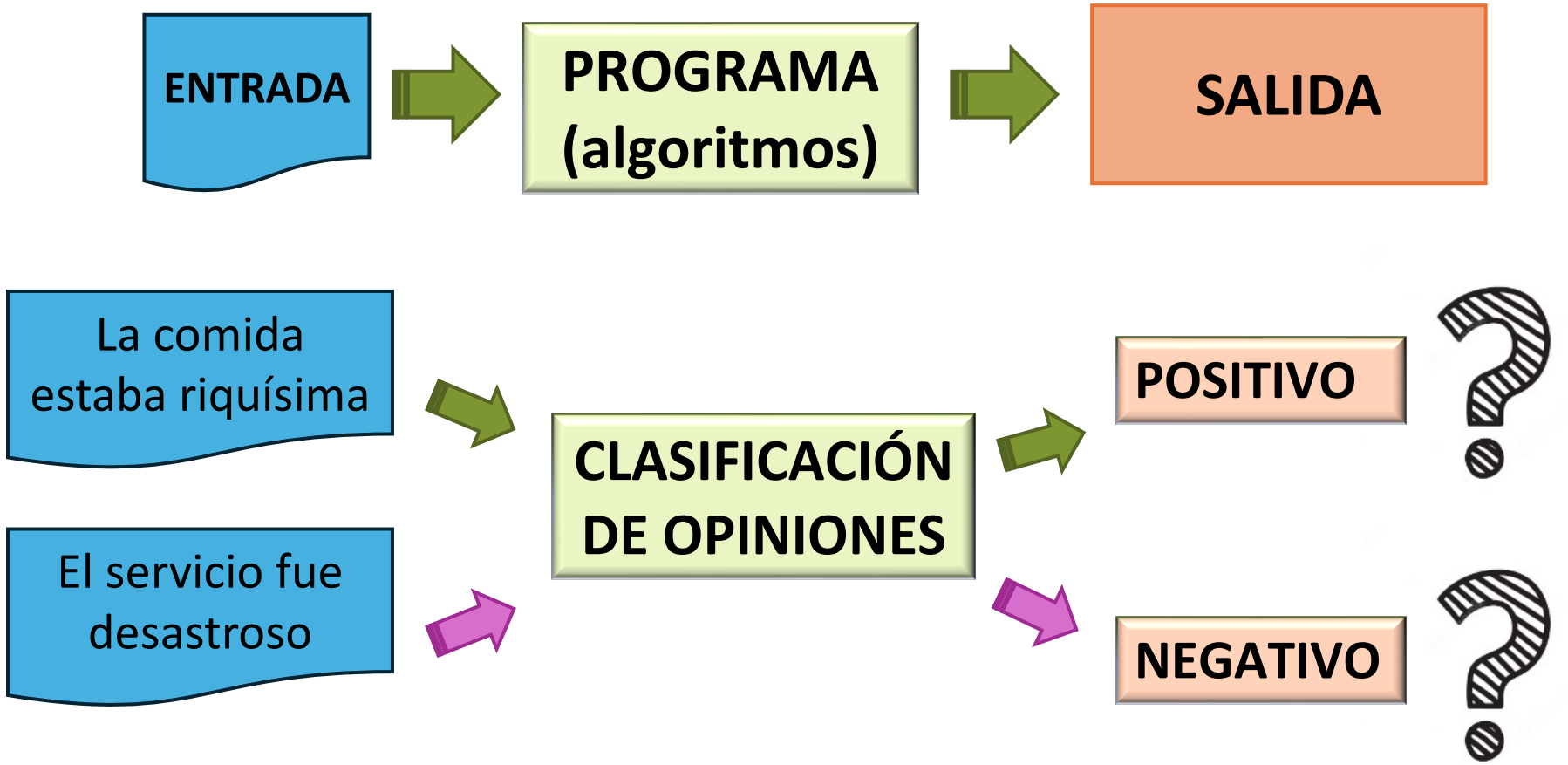
- Evaluar



experto en procesamiento del lenguaje natural

Evaluación en PLN

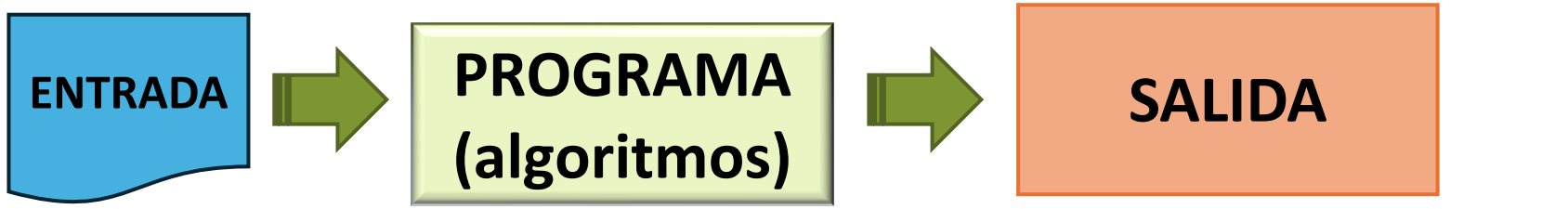
- Evaluar

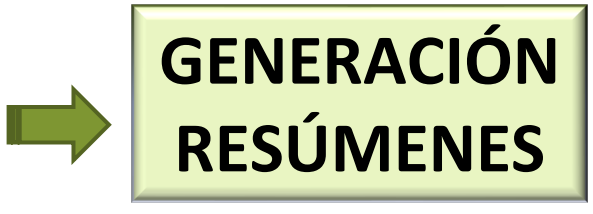


experto en procesamiento del lenguaje natural

Evaluación en PLN

- Evaluar





Big Ben is the name of the bell in the clock tower of Westminster Palace. The clock was installed in 1854, but was not fully operational until 1859. Big Ben itself, the bell, was cast in 1856. The tower is 316 feet high. The name Big Ben applied originally only to the bell but was probably after Sir Benjamin Hall the first Commissioner of Works. Each year a lot of money is spent to keep the clock running. The clock tower is not open for public visitors. It is located at the Thames and can be seen from the Westminster Bridge.

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Conceptos fundamentales

GOLD STANDARD (REFERENCE) → Referencia para el algoritmo de PLN que nos permite evaluar su rendimiento. Conjunto de respuestas correctas para una tarea de PLN.

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Conceptos fundamentales

GOLD STANDARD (REFERENCE) → Referencia para el algoritmo de PLN que nos permite evaluar su rendimiento. Conjunto de respuestas correctas para una tarea de PLN.

Ejemplo (clasificación de opiniones):

La comida estaba riquísima → positivo
El servicio fue desastroso → negativo
Le película me encantó → positiva
La película duró 90 minutos → neutra

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Conceptos fundamentales

GOLD STANDARD (REFERENCE) → Referencia para el algoritmo de PLN que nos permite evaluar su rendimiento. Conjunto de respuestas correctas para una tarea de PLN.

Ejemplo (generación de resúmenes):
http://elpais.com/deportes/2010/01/01/actualidad/1262334114_850215.html

Resumen de referencia:
El español supera a su compatriota Ferrer en el primer torneo de exhibición del año, mientras que el sueco se deshace de Federer.

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Conceptos fundamentales

BASELINE → Método obvio y simple que define el rendimiento mínimo que debe mejorarse para que se considere que el método/enfoque de PLN propuesto es válido.

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Conceptos fundamentales

BASELINE → Método obvio y simple que define el rendimiento mínimo que debe mejorarse para que se considere que el método/enfoque de PLN propuesto es válido.

Ejemplos de *baselines*:

- Rendimiento de elegir una respuesta/etiqueta al azar.
- Rendimiento de elegir siempre la etiqueta más frecuente.
- Método previo que se haya propuesto en la literatura.

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Conceptos fundamentales

BASELINE → Método obvio y simple que define el rendimiento mínimo que debe mejorarse para que se considere que el método/enfoque de PLN propuesto es válido.

Ejemplos de *baselines*:

- Rendimiento de elegir una respuesta/etiqueta al azar.
- Rendimiento de elegir siempre la etiqueta más frecuente.

} Comprobación de la realidad

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Conceptos fundamentales

BASELINE → Método obvio y simple que define el rendimiento mínimo que debe mejorarse para que se considere que el método/enfoque de PLN propuesto es válido.

Ejemplos de *baselines*:

- Rendimiento de elegir una respuesta/etiqueta al azar.
- Rendimiento de elegir siempre la etiqueta más frecuente.

¿Qué pasa si nuestro método no mejora el baseline?

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Conceptos fundamentales

BASELINE → Método obvio y simple que define el rendimiento mínimo que debe mejorarse para que se considere que el método/enfoque de PLN propuesto es válido.

Ejemplos de *baselines*:

- Rendimiento de elegir una respuesta/etiqueta al azar.
- Rendimiento de elegir siempre la etiqueta más frecuente.

¿Qué pasa si nuestro método no mejora el *baseline*?

- Revisar el planteamiento del método
- Buscar métodos alternativos

experto en procesamiento del lenguaje natural

Evaluación en PLN

- Conceptos fundamentales

UPPER BOUND → Mejor resultado que obtendría los seres humanos resolviendo la tarea de PLN.

Ejemplo de *upper bound*:

- Resultado que obtendríamos de evaluar resúmenes humanos sobre un conjunto de datos

"llegar al tope en cuanto al rendimiento que puede obtener un método/tarea de PLN"

experto en procesamiento del lenguaje natural

Índice

1. Evaluación en PLN
2. **Tipos de evaluación**
3. Esquemas generales de evaluación de modelos
4. Métricas de evaluación
5. Protocolo de evaluación
6. Caso práctico
7. Bibliografía de referencia

experto en procesamiento del lenguaje natural

Tipos de evaluación

Evaluar una tarea de PLN, ¿es fácil o es difícil?

experto en procesamiento del lenguaje natural

Tipos de evaluación

Evaluar una tarea de PLN, ¿es fácil o es difícil?

DEPENDE

experto en procesamiento del lenguaje natural

Tipos de evaluación

Evaluar una tarea de PLN, ¿es fácil o es difícil?

DEPENDE

DE LA TAREA EN SÍ

experto en procesamiento del lenguaje natural

Tipos de evaluación

Evaluar una tarea de PLN, ¿es fácil o es difícil?

DEPENDE

DE LA TAREA EN SÍ

DE LA DISPONIBILIDAD DE "GOLD STANDARDS"

experto en procesamiento del lenguaje natural

Tipos de evaluación

Evaluar una tarea de PLN, ¿es fácil o es difícil?

DEPENDE

DE LA TAREA EN SÍ

DE LA DISPONIBILIDAD DE "GOLD STANDARDS"

DEL TIPO DE EVALUACIÓN QUE QUERAMOS

experto en procesamiento del lenguaje natural

Tipos de evaluación

Evaluar una tarea de PLN, ¿es fácil o es difícil?

DEPENDE

DE LA TAREA EN SÍ

DE LA DISPONIBILIDAD DE "GOLD STANDARDS"

DEL TIPO DE EVALUACIÓN QUE QUERAMOS

experto en procesamiento del lenguaje natural

Tipos de evaluación

- Intrínseca vs. extrínseca
- Formativa vs. sumativa
- Cuantitativa vs. cualitativa
- Manual vs. automática
- Componentes vs. sistema completo

experto en procesamiento del lenguaje natural

Tipos de evaluación

Intrínseca vs. extrínseca

- Intrínseca → mide el rendimiento del método/modelo de PLN. ¿Cómo de bien funciona?
- Extrínseca → mide cómo de bien el método/modelo de PLN propuesto para conseguir el objetivo de otra tarea. ¿Cómo ayuda el método/modelo a mejorar otra tarea de PLN?

experto en procesamiento del lenguaje natural

Tipos de evaluación

Formativa vs. sumativa

- Formativa → suelen producirse durante el desarrollo de los sistemas de PLN. Su objetivo principal es informar sobre si se está avanzando hacia los objetivos previstos de esa tarea. Suelen ser ligeras (rapidez) e iterativas.
- Sumativa → suelen realizarse una vez que el sistema está completo (o ha alcanzado un hito importante en su desarrollo). Su objetivo es valorar si se han alcanzado los objetivos previstos del sistema.

experto en procesamiento del lenguaje natural

Tipos de evaluación

Cuantitativa vs. cualitativa

- Cuantitativa → orientada a valores numéricos como resultado de la evaluación
- Cualitativa → orientada al análisis manual (p.ej. análisis DAFO: fortalezas, debilidades, oportunidades, amenazas…)

experto en procesamiento del lenguaje natural

Tipos de evaluación

Manual vs. automática

- Manual →sujetos humanos evalúan los resultados del sistema de PLN según unos criterios predeterminados.
- Automática → se evalúan los sistemas a través de métricas que “imitan el comportamiento de los evaluadores humanos”. Tienen que correlacionar con evaluaciones humanas.

experto en procesamiento del lenguaje natural

Tipos de evaluación

Componentes vs. sistema completo

- Componentes → mide el rendimiento de cada uno de los módulos/etapas que intervienen en un sistema de PLN.
- Sistema completo →mide el rendimiento de la salida producida por el sistema completo de PLN.

experto en procesamiento del lenguaje natural

Índice

1. Evaluación en PLN
2. Tipos de evaluación
3. **Esquemas generales de evaluación de modelos**
4. Métricas de evaluación
5. Protocolo de evaluación
6. Caso práctico
7. Bibliografía de referencia

experto en procesamiento del lenguaje natural

Esquemas generales de evaluación de modelos

- Entrenamiento-Prueba
- Entrenamiento-prueba-validación
- Validación cruzada

experto en procesamiento del lenguaje natural

Esquemas generales de evaluación de modelos

Entrenamiento-Prueba

- Consiste en dividir el conjunto de datos que tenemos para una tarea en dos conjuntos disjuntos: uno de entrenamiento (para entrenar el modelo de PLN) y otro de prueba (para validar el modelo de PLN).

experto en procesamiento del lenguaje natural

Esquemas generales de evaluación de modelos

Entrenamiento-Prueba

- Consiste en dividir el conjunto de datos que tenemos para una tarea en dos conjuntos disjuntos: uno de entrenamiento (para entrenar el modelo de PLN) y otro de prueba (para validar el modelo de PLN).

Entrenamiento-prueba-validación

- Consiste en dividir el conjunto de datos que tenemos para una tarea en tres conjuntos disjuntos: conjunto de entrenamiento (para entrenar el modelo); conjunto de validación (para optimizar y ajustar los parámetros) y conjunto de prueba (para validar el modelo).

experto en procesamiento del lenguaje natural

Esquemas generales de evaluación de modelos

Entrenamiento-Prueba

- Consiste en dividir el conjunto de datos que tenemos para una tarea en dos conjuntos disjuntos: uno de entrenamiento (para entrenar el modelo de PLN) y otro de prueba (para validar el modelo de PLN).

Entrenamiento-prueba-validación

- Consiste en dividir el conjunto de datos que tenemos para una tarea en tres conjuntos disjuntos: conjunto de entrenamiento (para entrenar el modelo); conjunto de validación (para optimizar y ajustar los parámetros) y conjunto de prueba (para validar el modelo).

Validación cruzada

- Consiste en dividir el conjunto de datos que tenemos para una tarea en diferentes conjuntos, donde en cada vez, se deja uno de esos conjuntos como prueba y el resto para entrenar el modelo.

experto en procesamiento del lenguaje natural

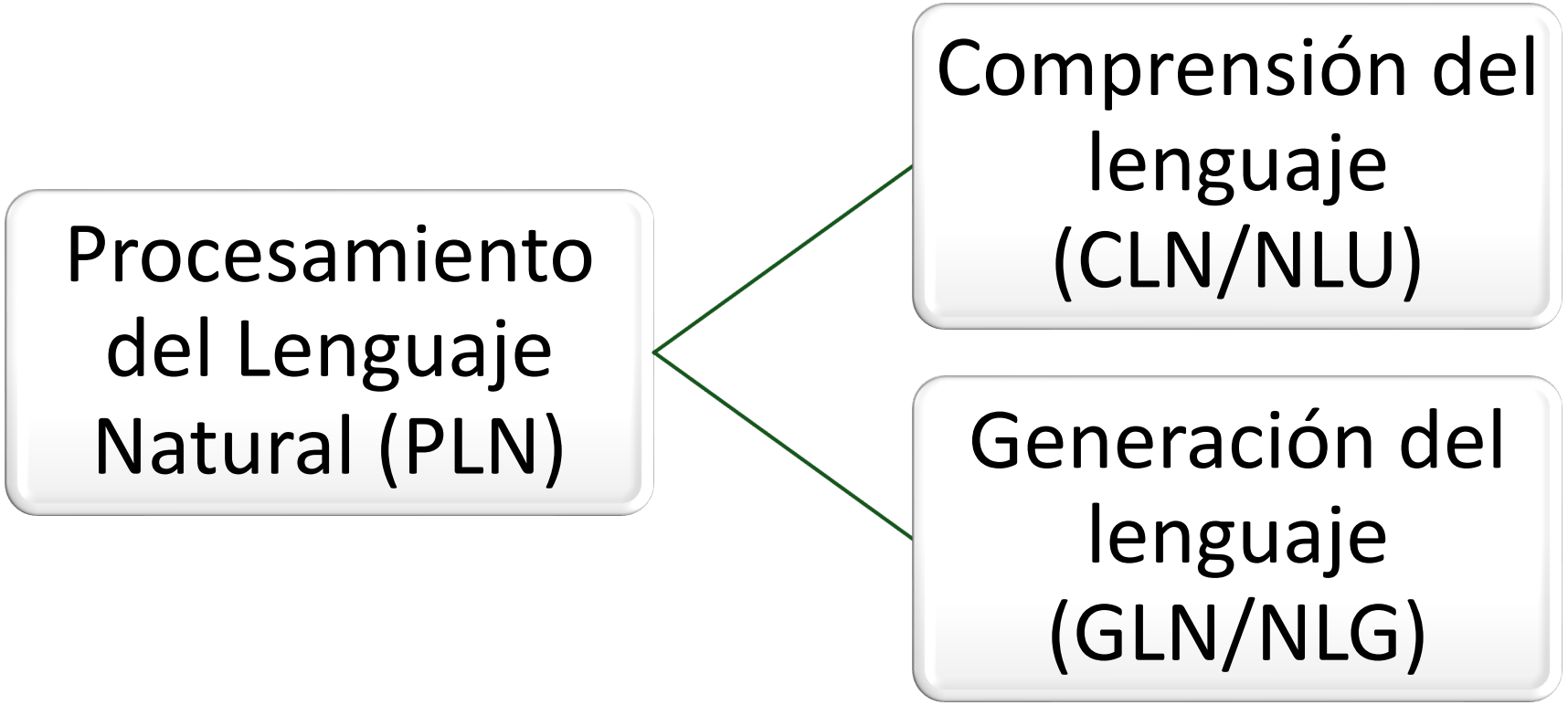
Índice

1. Evaluación en PLN
2. Tipos de evaluación
3. Esquemas generales de evaluación de modelos
4. **Métricas de evaluación**
5. Protocolo de evaluación
6. Caso práctico
7. Bibliografía de referencia

experto en procesamiento del lenguaje natural

Métricas de evaluación

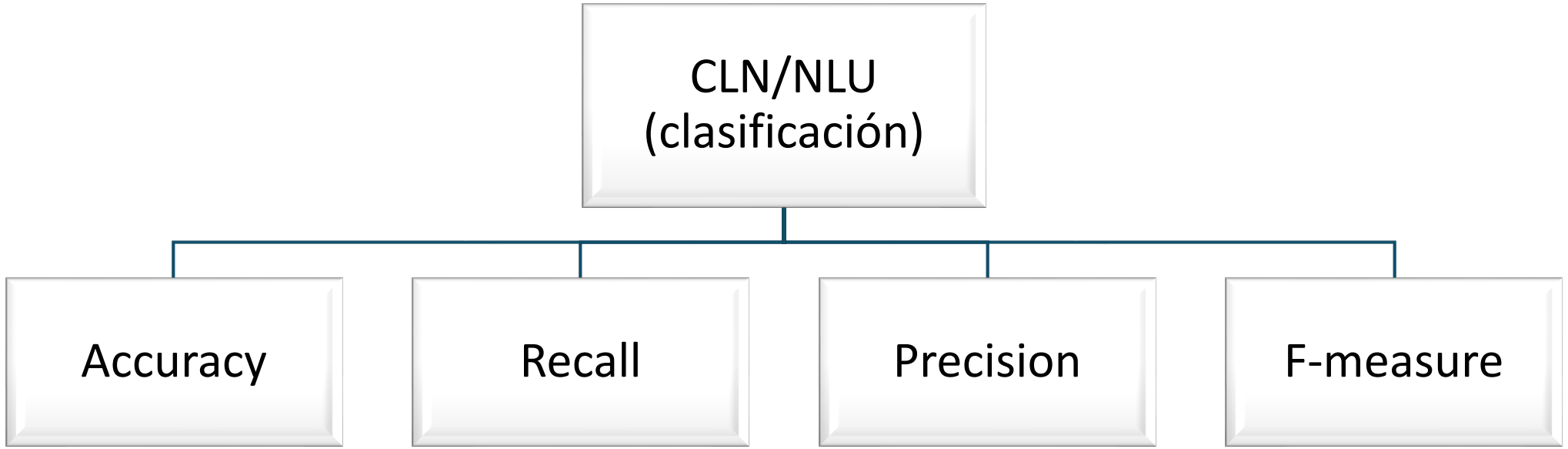
- Según el tipo de problema/tarea que queramos resolver en PLN, predominan un tipo u otro de métricas



experto en procesamiento del lenguaje natural

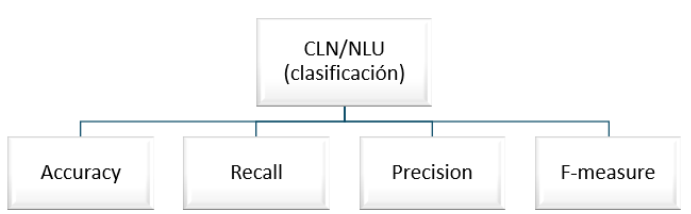
Métricas de evaluación

- Medir la eficacia de los modelos que clasifican texto



experto en procesamiento del lenguaje natural

Métricas de evaluación

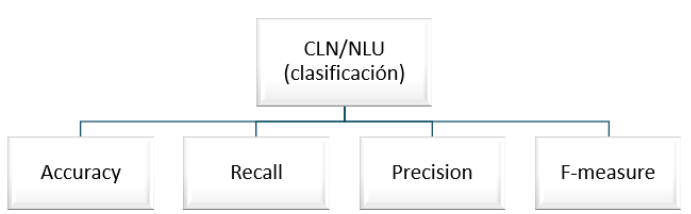


- Clasificación → instancias o ejemplos
  - Instancias **correctamente clasificadas**
    - Mide el número o porcentaje de instancias correctamente clasificadas
  - Instancias **incorrectamente clasificadas**
    - Mide el número o porcentaje de instancias mal clasificadas

MATRIZ DE CONFUSIÓN

experto en procesamiento del lenguaje natural

Métricas de evaluación

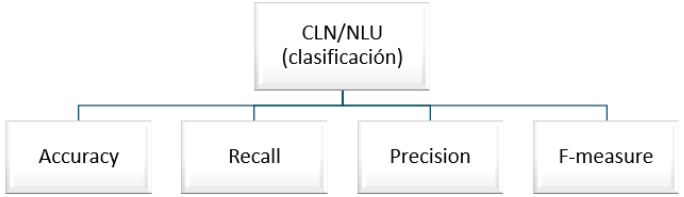


- Matriz de confusión (clasificación binaria)

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

experto en procesamiento del lenguaje natural

Métricas de evaluación



- Matriz de confusión (clasificación binaria)

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

VP (TP) → cantidad de instancias positivas que fueron clasificadas correctamente como positivas por el modelo.

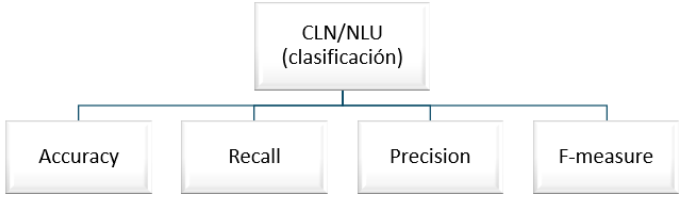
VN (TN)→ cantidad de instancias negativas que fueron clasificadas correctamente como negativas por el modelo.

FN → cantidad de instancias positivas que fueron clasificadas incorrectamente como negativas.

FP → cantidad de instancias negativas que fueron clasificadas incorrectamente como positivas.

experto en procesamiento del lenguaje natural

Métricas de evaluación



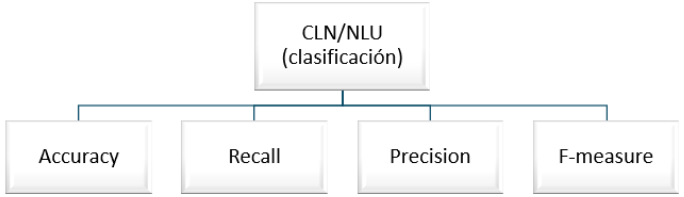
- Matriz de confusión (clasificación binaria)

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

experto en procesamiento del lenguaje natural

Métricas de evaluación



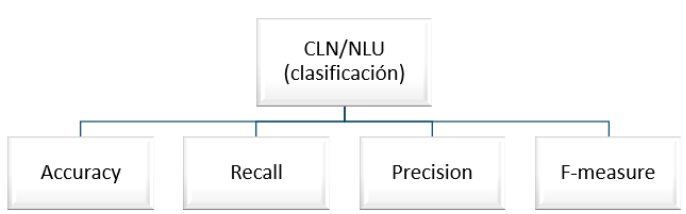
- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	**ECONOMÍA**
Observación	**DEPORTES**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	ECONOMÍA	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

experto en procesamiento del lenguaje natural

Métricas de evaluación



- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	ECONOMÍA	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Ejemplo:

Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

VP (TP) → cantidad de **textos de deportes que fueron clasificados correctamente en la categoría de deportes**

VN (TN)→ cantidad de **textos de economía que fueron clasificados correctamente en la categoría de economía.**

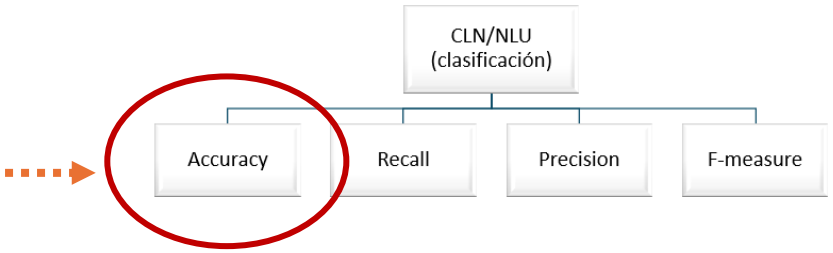
FN → cantidad **de textos que son de deportes pero que fueron clasificadas incorrectamente en la categoría de economía.**

FP → cantidad de **textos que son de economía pero que fueron clasificados incorrectamente en la categoría de deportes**

experto en procesamiento del lenguaje natural

Métricas de evaluación

VALORES ENTRE 0 y 1
(1 mejor valor)



- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	4219 TP	0 FN
	ECONOMÍA	54 FP	6 TN

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

Accuracy (exactitud)

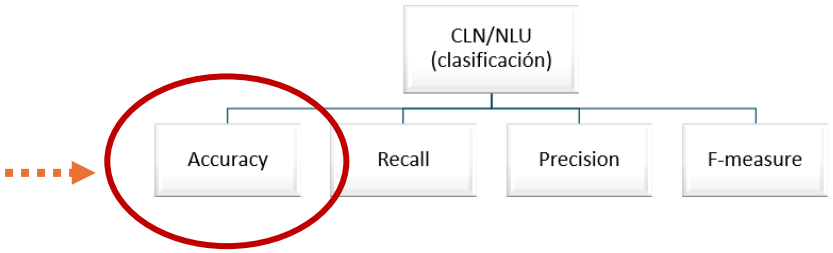
- Representa la frecuencia con la que el clasificador realiza la predicción correcta
- No es muy apropiada cuando los datos están desbalanceados

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

experto en procesamiento del lenguaje natural

Métricas de evaluación

VALORES ENTRE 0 y 1
(1 mejor valor)



- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	4219 TP	0 FN
	ECONOMÍA	54 FP	6 TN

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

Accuracy (exactitud)

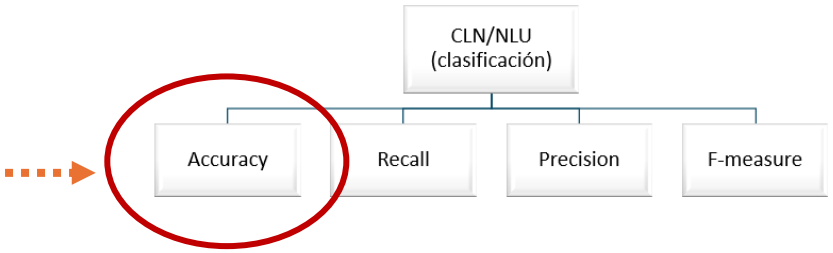
- Representa la frecuencia con la que el clasificador realiza la predicción correcta
- No es muy apropiada cuando los datos están desbalanceados

$$Accuracy = \frac{4219+6}{4219+54+6+0} = 0.98$$

experto en procesamiento del lenguaje natural

Métricas de evaluación

VALORES ENTRE 0 y 1
(1 mejor valor)



- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	4219 TP	0 FN
	ECONOMÍA	54 FP	6 TN

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

Balanced Accuracy (exactitud balanceada)

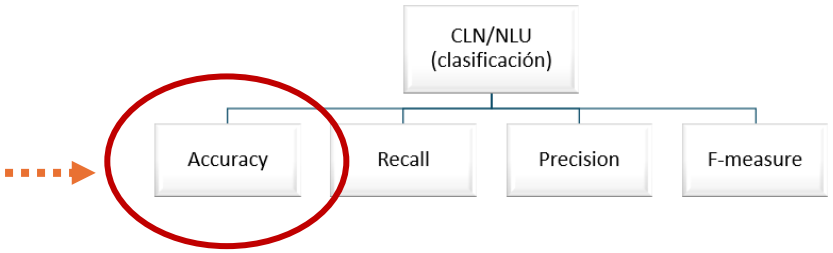
- Más adecuada cuando los datos de prueba están desbalanceados

$$BAccuracy = \frac{TP}{2 * (TP + FN)} + \frac{TN}{2 * (TN + FP)}$$

experto en procesamiento del lenguaje natural

Métricas de evaluación

VALORES ENTRE 0 y 1
(1 mejor valor)



- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	4219 TP	0 FN
	ECONOMÍA	54 FP	6 TN

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

Balanced Accuracy (exactitud balanceada)

- Más adecuada cuando los datos de prueba están desbalanceados

$$BAccuracy = \frac{4219}{2 * (4219 + 0)} + \frac{6}{2 * (6 + 54)}$$

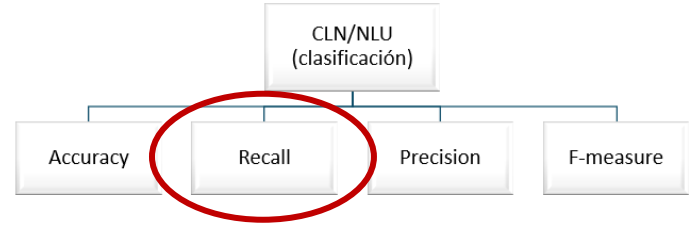
$$BAccuracy = 0.5 + 0.05$$

$$BAccuracy = 0.55$$

experto en procesamiento del lenguaje natural

Métricas de evaluación

VALORES ENTRE 0 y 1
(1 mejor valor)



- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	4219 TP	0 FN
	ECONOMÍA	54 FP	6 TN

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

Recall (cobertura)

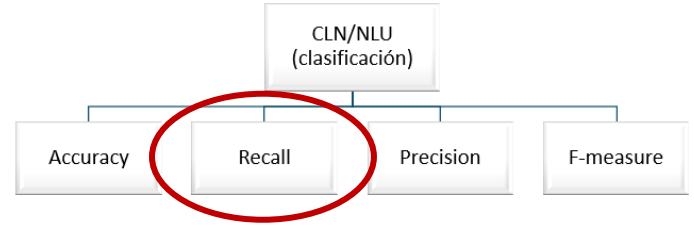
- Mide la proporción de los casos positivos reales que fueron correctamente identificados por el modelo (proporción de verdaderos positivos frente a falsos negativos).

$$Recall = \frac{TP}{(TP + FN)}$$

experto en procesamiento del lenguaje natural

Métricas de evaluación

VALORES ENTRE 0 y 1
(1 mejor valor)



- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	4219 TP	0 FN
	ECONOMÍA	54 FP	6 TN

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

Recall (cobertura)

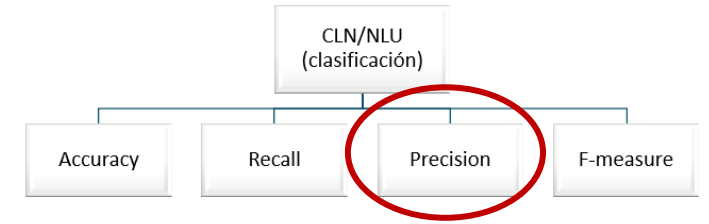
- Mide la proporción de los casos positivos reales que fueron correctamente identificados por el modelo (proporción de verdaderos positivos frente a falsos negativos).

$$Recall = \frac{4219}{4219+0} = 1.00$$

experto en procesamiento del lenguaje natural

Métricas de evaluación

VALORES ENTRE 0 y 1
(1 mejor valor)



- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	4219 TP	0 FN
	ECONOMÍA	54 FP	6 TN

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

Precision (precisión)

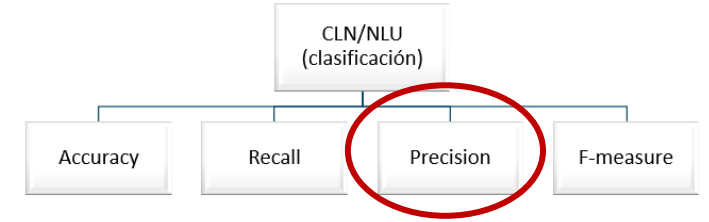
- Mide el número de predicciones correctas que se han hecho respecto a todas las predicciones hechas de esa clase.

$$Precision = \frac{TP}{(TP + FP)}$$

experto en procesamiento del lenguaje natural

Métricas de evaluación

VALORES ENTRE 0 y 1
(1 mejor valor)



- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	4219 TP	0 FN
	ECONOMÍA	54 FP	6 TN

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

Precision (precisión)

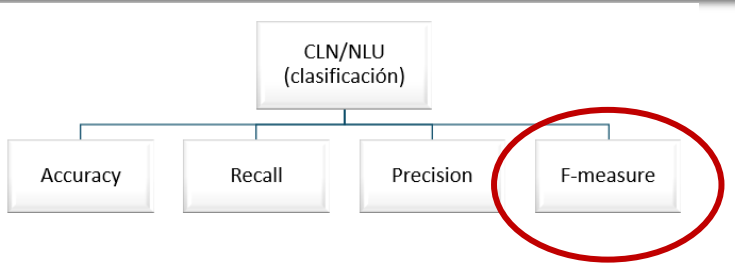
- Mide el número de predicciones correctas que se han hecho respecto a todas las predicciones hechas de esa clase.

$$Precision = \frac{4219}{4219 + 54}$$

$$Precision = 0.98$$

experto en procesamiento del lenguaje natural

Métricas de evaluación



VALORES ENTRE 0 y 1
(1 mejor valor)

- Matriz de confusión (clasificación binaria)

		Predicción	
		DEPORTES	ECONOMÍA
Observación	DEPORTES	4219 TP	0 FN
	ECONOMÍA	54 FP	6 TN

Ejemplo:
Queremos clasificar un texto según su tema principal, si pertenece a "**deportes**" o a "**economía**".

F-measure (F-score)

- Representa un promedio ponderado entre la precisión y la cobertura.
- Normalmente, beta=1 → medida F1

$$FMeasure = (1+\beta^2)\frac{Precision*Recall}{(\beta^2*Precision)+Recall}$$

experto en procesamiento del lenguaje natural

Tipos de evaluación

Calcular el valor para el F-measure de los datos anteriores, con el valor de beta=1 (5’)

Recall = 1

Precision = 0,98

experto en procesamiento del lenguaje natural

Tipos de evaluación

Calcular el valor para el F-measure de los datos anteriores, con el valor de beta=1 (5’)

Recall = 1
Precision = 0,98

F-measure = 0,99

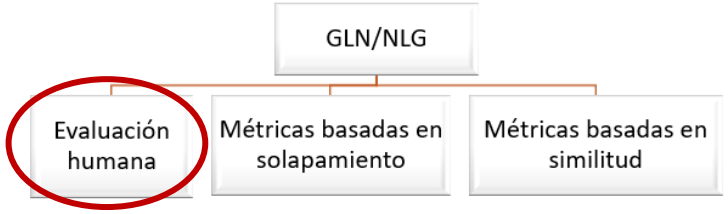
experto en procesamiento del lenguaje natural

Métricas de evaluación

- Medir la eficacia de los modelos que generan texto

experto en procesamiento del lenguaje natural

Métricas de evaluación



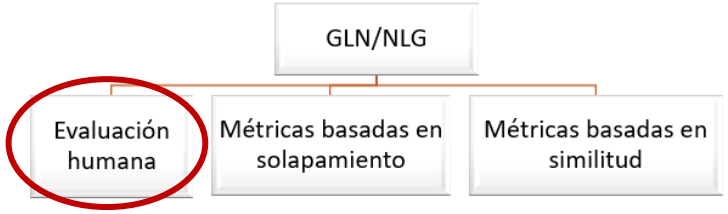
- Evaluación humana
  - Consiste en preguntar a humanos (encuestas)

  de manera general o en base a ciertos criterios (p.ej. coherencia, adecuación, estilo, sentido común, redundancia,…)
  - Importante definir qué evaluar y cómo.
  - Se suelen emplear escalas de Likert (3 puntos, 5 puntos; 7 puntos,…)
  - Acuerdo entre anotadores/evaluadores

experto en procesamiento del lenguaje natural

Métricas de evaluación

- Evaluación humana
  - Consiste en preguntar a humanos (encuestas) de manera general o en base a ciertos criterios (p.ej. coherencia, adecuación, estilo, sentido común, redundancia,…)
  - Importante definir qué evaluar y cómo.
  - Se suelen emplear escalas de Likert (3 puntos, 5 puntos; 7 puntos)….
  - Acuerdo entre anotadores/evaluadores



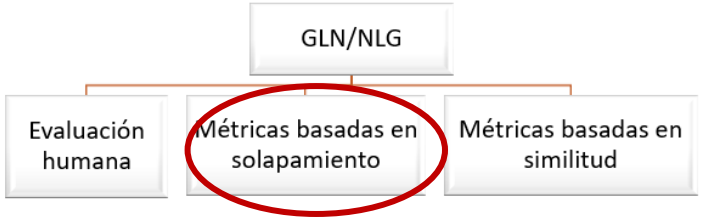
Ejemplo:

Pedimos a personas que evalúen la estructura y coherencia de un resumen: *"El resumen debe estar bien estructurado y organizado. El resumen no debe ser una mera recopilación de información relacionada, sino que debe construirse frase a frase hasta formar un cuerpo coherente de información sobre un tema".*

1. Muy pobre
2. Pobre
3. Apenas aceptable
4. Bueno
5. Muy bueno

experto en procesamiento del lenguaje natural

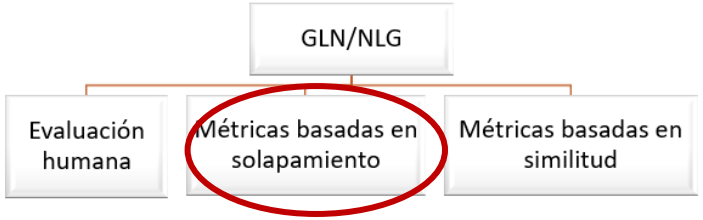
Métricas de evaluación



- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).
  - Elementos = n-gramas (unigramas, bigramas, trigramas, etc.)

experto en procesamiento del lenguaje natural

Métricas de evaluación



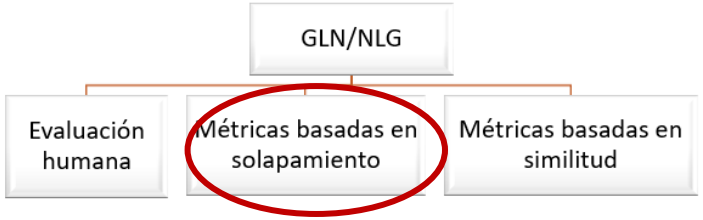
- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).
  - Elementos = n-gramas (unigramas, bigramas, trigramas, etc.)

Ejemplo de n-gramas:

The way to make people trustworthy is to trust them

experto en procesamiento del lenguaje natural

Métricas de evaluación



- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).
  - Elementos = n-gramas (unigramas, bigramas, trigramas, etc.)

Ejemplo de n-gramas:

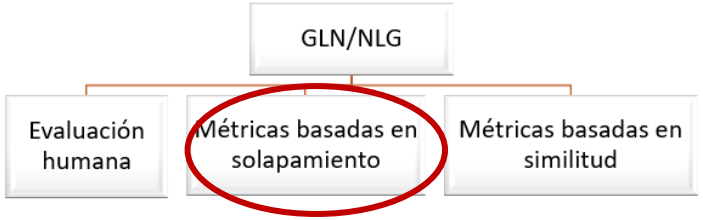
The way to make people trustworthy is to trust them.

Unigramas (n-gramas tamaño 1):

The way to make people trustworthy is to trust them

experto en procesamiento del lenguaje natural

Métricas de evaluación



- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).
  - Elementos = n-gramas (unigramas, bigramas, trigramas, etc.)

Ejemplo de n-gramas:

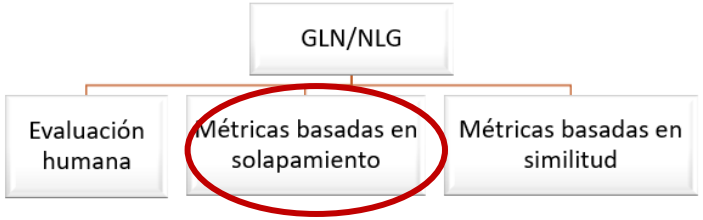
The way to make people trustworthy is to trust them

Bigramas (n-gramas tamaño 2):

The way
Way to
to make
make people
people trustworthy
trustworthy is
is to
is to
to trust
trust them

experto en procesamiento del lenguaje natural

Métricas de evaluación



- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).
  - Elementos = n-gramas (unigramas, bigramas, trigramas, etc.)

Ejemplo de n-gramas:

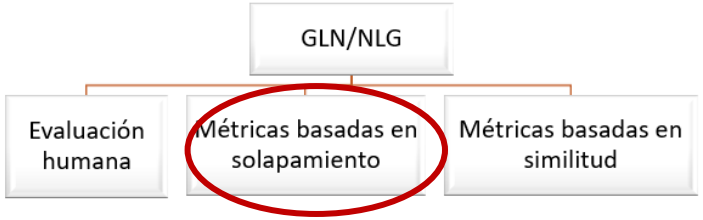
The way to make people trustworthy is to trust them

Trigramas (n-gramas tamaño 3):

The way to	trustworthy is to
way to make	is to trust
to make people	to trust them
people trustworthy is	

experto en procesamiento del lenguaje natural

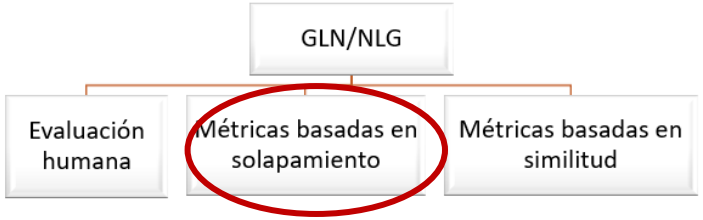
Métricas de evaluación



- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).
  - Elementos = n-gramas (unigramas, bigramas, trigramas, etc.)
  - Algunas métricas de este tipo: ROUGE, BLEU, METEOR, CIDER,…

experto en procesamiento del lenguaje natural

Métricas de evaluación



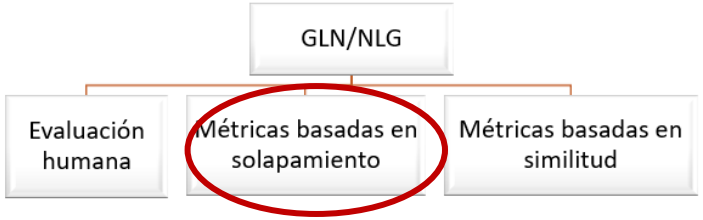
- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).

BLEU (BiLingual Evaluation Understudy)

- Se utiliza normalmente a nivel de frases
- Métrica orientada a la precisión: calcula la precisión en n-gramas entre una frase generada automáticamente y otra de referencia
- Se suelen emplear en tareas de traducción automática o generación de lenguaje

experto en procesamiento del lenguaje natural

Métricas de evaluación



- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).

BLEU (BiLingual Evaluation Understudy)

- Se utiliza normalmente a nivel de frases
- Métrica orientada a la precisión: calcula la precisión en n-gramas entre una frase generada automáticamente y otra de referencia
- Se suelen emplear en tareas de traducción automática o generación de lenguaje

Calcular BLEU online: https://huggingface.co/spaces/evaluate-metric/bleu

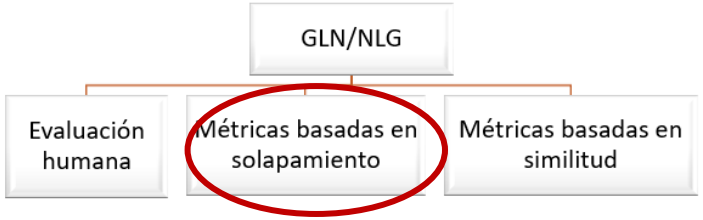
Reference	The way to make people trustworthy is to trust them.
Hypothesis	To make people trustworthy, you need to trust them.

{'bleu': 0.3392513407933634, 'precisions': [0.636363636363636, 0.5, 0.3333333333333333, 0.125], 'brevity_penalty': 1.0, 'length_ratio': 1.0, 'translation_length': 11, 'reference_length': 11}

BLEU = 0,33933
VALORES ENTRE 0 y 1
(+ alto → mejor valor)

experto en procesamiento del lenguaje natural

Métricas de evaluación



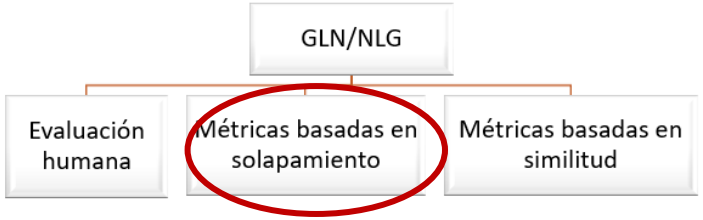
- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- Se utiliza normalmente a nivel de textos
- Métrica orientada a la cobertura: calcula recall, precisión y F-measure en función de n-gramas comunes entre un resumen generado automáticamente y uno o varios resúmenes de referencia
- Se suelen emplear en tareas de generación de resúmenes

experto en procesamiento del lenguaje natural

Métricas de evaluación



- Métricas basadas en solapamiento
  - Calcular una puntuación que indica el grado de coincidencia entre elementos del texto generado y el texto de referencia (gold-standard).

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- Se utiliza normalmente a nivel de textos
- Métrica orientada a la cobertura: calcula recall, precisión y F-measure en función de n-gramas comunes entre un resumen generado automáticamente y uno o varios resúmenes de referencia
- Se suelen emplear en tareas de generación de resúmenes

Calcular ROUGE online: https://huggingface.co/spaces/evaluate-metric/rouge

Reference	The way to make people trustworthy is to trust them.
Hypothesis	To make people trustworthy, you need to trust them.

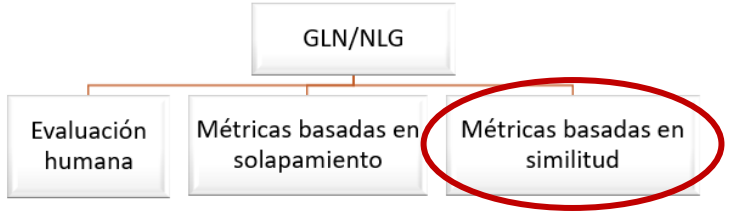
{'rouge1': 0.7368421052631577, 'rouge2': 0.5882352941176471, 'rougeL': 0.7368421052631577, 'rougeLsum': 0.7368421052631577}

ROUGE-1= 0,737; ROUGE-2=0,588

VALORES ENTRE 0 y 1
(+ alto → mejor valor)

experto en procesamiento del lenguaje natural

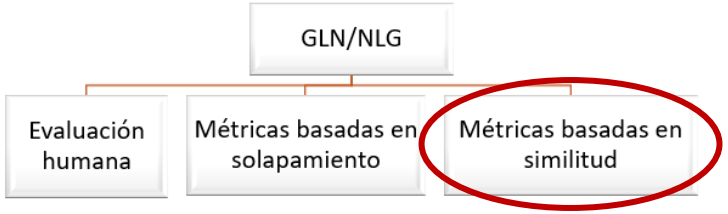
Métricas de evaluación



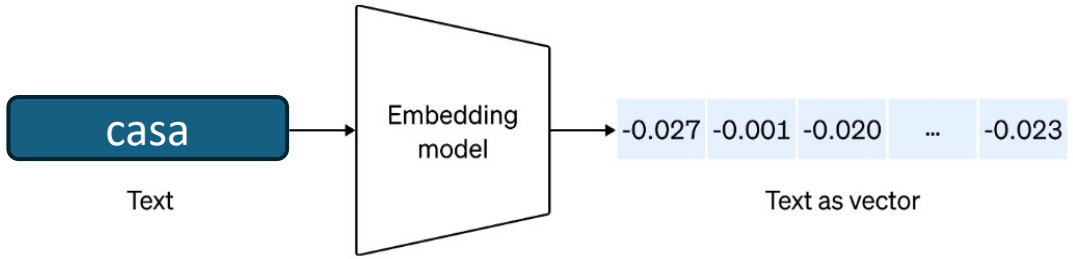
- Métricas basadas en similitud
  - Calcular un valor de similitud en entre el texto generado y el texto de referencia (gold-standard) previa representación vectorial de los mismos.

experto en procesamiento del lenguaje natural

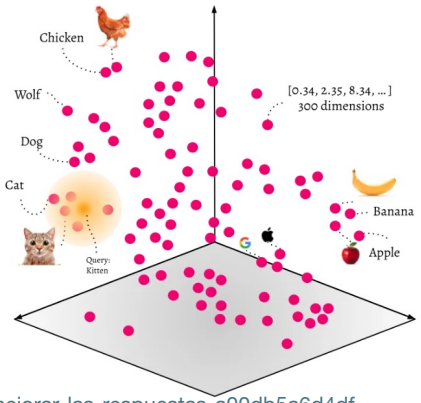
Métricas de evaluación



- Métricas basadas en similitud
  - Calcular un valor de similitud en entre el texto generado y el texto de referencia (gold-standard) previa representación vectorial de los mismos.
  - Representación vectorial = transformar las palabras en números (vectores) para poder realizar operaciones matemáticas. Una de estas representaciones vectoriales son los EMBEDDINGS, que capturan el significado semántico de la información.



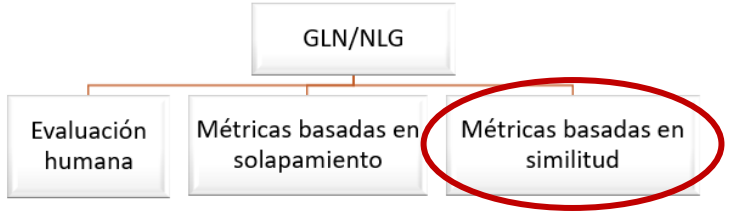
Transformando un concepto en un embedding.



Fuente: https://gustavo-espindola.medium.com/optimizando-los-modelos-de-lenguaje-c%C3%B3mo-utilizar-datos-vectoriales-para-mejorar-las-respuestas-a00db5a6d4df

experto en procesamiento del lenguaje natural

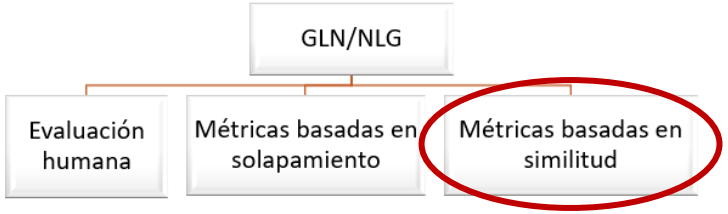
Métricas de evaluación



- Métricas basadas en similitud
  - Calcular un valor de similitud en entre el texto generado y el texto de referencia (gold-standard) previa representación vectorial de los mismos.
  - Representación vectorial = transformar las palabras en números (vectores) para poder realizar operaciones matemáticas
  - Algunas métricas de este tipo: similitud del coseno, BERTScore, distancia Euclídea, coeficiente de similitud de Jaccard,… https://huggingface.co/tasks/sentence-similarity

experto en procesamiento del lenguaje natural

Métricas de evaluación

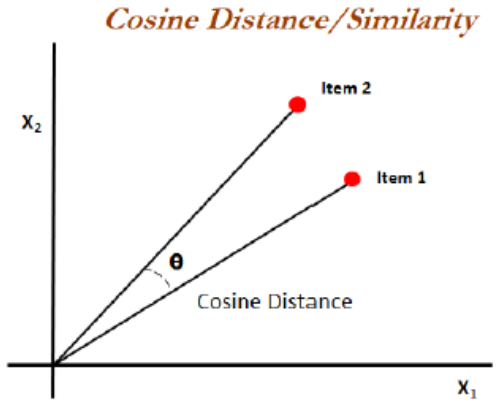


- Métricas basadas en similitud
  - Calcular un valor de similitud en entre el texto generado y el texto de referencia (gold-standard) previa representación vectorial de los mismos.

Similitud del coseno (cosine similarity)

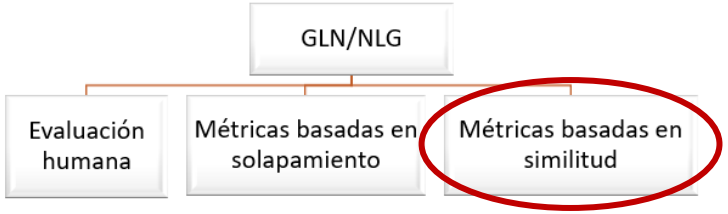
- Medida de la similitud existente entre dos vectores.
- Valor 1 si el ángulo comprendido es cero, es decir si ambos vectores apuntan a un mismo lugar

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$



experto en procesamiento del lenguaje natural

Métricas de evaluación

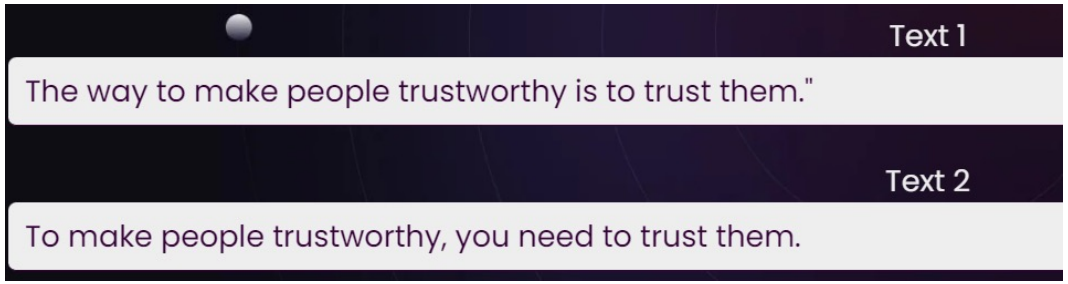


- Métricas basadas en similitud
  - Calcular un valor de similitud en entre el texto generado y el texto de referencia (gold-standard) previa representación vectorial de los mismos.

Similitud del coseno (cosine similarity)

- Medida de la similitud existente entre dos vectores.
- Valor 1 si el ángulo comprendido es cero, es decir si ambos vectores apuntan a un mismo lugar

Calcular sim. coseno online: https://tilores.io/cosine-similarity-online-tool

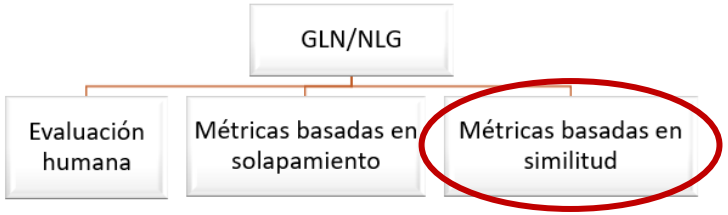


Similarity Algorithm	Similarity
Cosine	72.80%

VALORES ENTRE 0 y 1
(+ alto → + similitud)

experto en procesamiento del lenguaje natural

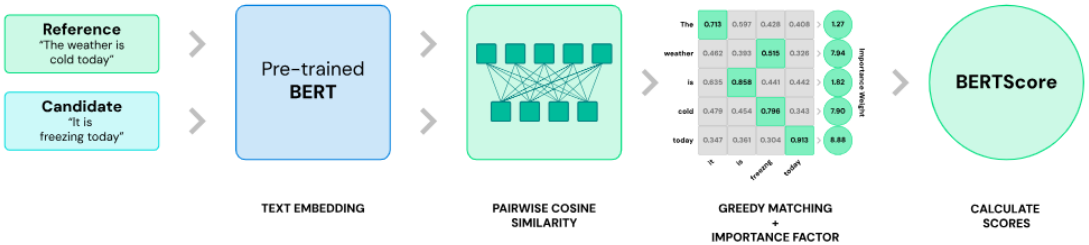
Métricas de evaluación

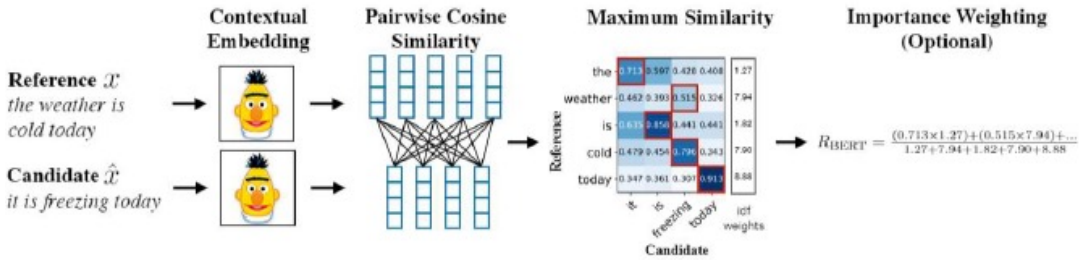


- Métricas basadas en similitud
  - Calcular un valor de similitud en entre el texto generado y el texto de referencia (gold-standard) previa representación vectorial de los mismos.

BERTScore

- calcula un valor de similitud para cada token de la frase generada de forma automática con cada token de la frase de referencia.
- Aprovecha los embeddings contextuales preentrenados de los modelos BERT y compara las palabras de las frases candidatas y de referencia por similitud coseno.
- BERTScore calcula la precisión, la cobertura y la medida F1.

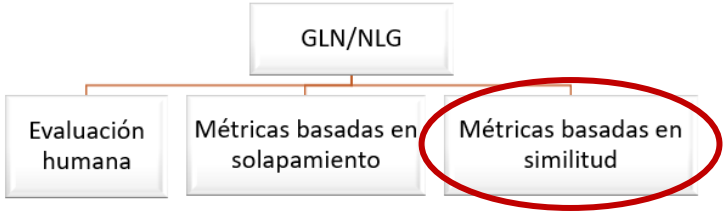




Fuentes: https://docs.kolena.com/metrics/bertscore/
https://arxiv.org/pdf/1904.09675.pdf

experto en procesamiento del lenguaje natural

Métricas de evaluación



- Métricas basadas en similitud
  - Calcular un valor de similitud en entre el texto generado y el texto de referencia (gold-standard) previa representación vectorial de los mismos.

BERTScore

- calcula un valor de similitud para cada token de la frase generada de forma automática con cada token de la frase de referencia.
- Aprovecha los embeddings contextuales preentrenados de los modelos BERT y compara las palabras de las frases candidatas y de referencia por similitud coseno.
- BERTScore calcula la precisión, la cobertura y la medida F1 (valores entre 0 y 1, **cuanto más alto, mejor**)

Candidate Text	Reference Text
The sun set behind the mountains, casting a warm, orange glow across the horizon.	As the mountains obscured the sun, a warm, orange glow painted the horizon.
She sipped her coffee and gazed out of the window, lost in thought on a rainy afternoon.	Lost in thought on a rainy afternoon, she sipped her coffee and stared out of the window.
The adventurous explorer trekked through the dense jungle, searching for hidden treasures.	In search of hidden treasures, the intrepid explorer ventured through the dense jungle.
Laughter echoed through the park as children played on the swings and slides.	Children's laughter filled the park as they enjoyed the swings and slides.
The old bookstore was filled with the scent of well-worn pages, a haven for book lovers.	A haven for book lovers, the old bookstore exuded the fragrance of well-read pages.

We get the following BertScores: $P_{\text{BERT}} = 0.9526, R_{\text{BERT}} = 0.9480, F_{\text{BERT}} = 0.9503$. This is in-line with human judgement, as the reference and candidate texts are semantically very similar.

However, if we were to calculate the BLEU score given these candidates and references, BLEU would yield a sub-optimal score of $\text{BLEU} = 0.2403$, despite the sentences being the same, semantically. This shows an advantage of embeddings-based metrics over n-gram-based metrics like BLEU.

Fuente: https://docs.kolena.com/metrics/bertscore/

experto en procesamiento del lenguaje natural

Métricas de evaluación

- A partir de las métricas basadas en similitud que usan las representaciones de embeddings, han surgido nuevas versiones de métricas para BLEU → BLEURT (Sellam et.al. 2020) o ROUGE → ROUGE-WE (Ng and Abrecht, 2015).

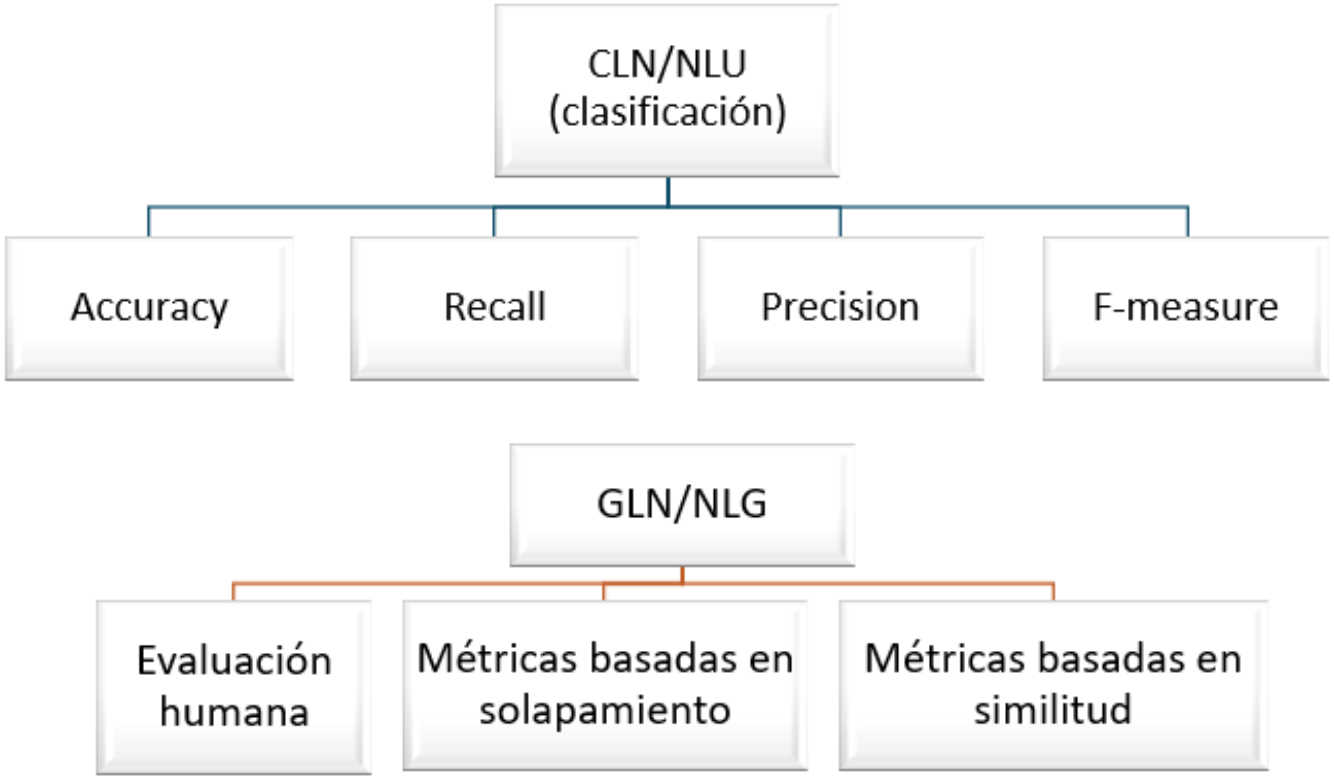
Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better Summarization Evaluation with Word Embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

experto en procesamiento del lenguaje natural

Métricas de evaluación

- Las métricas aprendidas, además de para evaluar, se pueden utilizar también como características de un modelo de aprendizaje automático.



experto en procesamiento del lenguaje natural

Índice

1. Evaluación en PLN
2. Tipos de evaluación
3. Esquemas generales de evaluación de modelos
4. Métricas de evaluación
5. **Protocolo de evaluación**
6. Caso práctico
7. Bibliografía de referencia

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
  3. Seleccionar las métricas de evaluación
  4. Recopilar y preparar los datos
  5. Ejecutar la evaluación y analizar los resultados
  6. Evaluar la generalización del método

experto en procesamiento del lenguaje natural

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación

experto en procesamiento del lenguaje natural

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
     - ¿Cuál es el propósito del modelo de PLN? ¿Cuáles son los resultados y beneficios esperados? ¿Cuáles son las limitaciones y los desafíos? ¿Cómo se define el éxito y el fracaso del modelo?
       - Si hemos desarrollado un chatbot para ser usado en un entorno real, es posible que quedamos evaluar el tiempo de respuesta, la precisión, la relevancia y la naturalidad de las frases generadas por el chatbot.

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
     - ¿Qué tipo de evaluación queremos: instrínseca o extrínseca? ¿Queremos evaluar cada componente o sólo el sistema global? ¿Vamos a llevar a cabo una evaluación manual/automática, cuantitativa/cualitativa?

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
  3. Seleccionar las métricas de evaluación

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
  3. Seleccionar las métricas de evaluación
     - ¿Qué métricas de evaluación son las más apropiadas para la tarea que estoy llevando a cabo? ¿Voy a usar sólo una métrica o varias?

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
  3. Seleccionar las métricas de evaluación
  4. Recopilar y preparar los datos

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
  3. Seleccionar las métricas de evaluación
  4. Recopilar y preparar los datos
     - ¿Qué datos tengo disponibles? ¿qué datos necesito? Los datos deben ser relevantes, representativos y confiables para el propósito y el dominio de la evaluación que queramos llevar a cabo. Idealmente, balanceados si queremos evaluar una tarea de clasificación.
       - Por ejemplo, si voy a evaluar un algoritmo de generación de resúmenes, tendré que recopilar corpus que contengan pares documento-resumen.

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
  3. Seleccionar las métricas de evaluación
  4. Recopilar y preparar los datos
  5. Ejecutar la evaluación y analizar los resultados

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
  3. Seleccionar las métricas de evaluación
  4. Recopilar y preparar los datos
  5. Ejecutar la evaluación y analizar los resultados
     - ¿Existen librerías (por ejemplo, en lenguaje de programación Python) que ya tengan implementadas las métricas de evaluación que queremos usar?
     - ¿Qué *baselines* existen para la tarea o cuáles puedo definir?
     - ¿Cómo comparo mis resultados con otros obtenidos con otros métodos?

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
  3. Seleccionar las métricas de evaluación
  4. Recopilar y preparar los datos
  5. Ejecutar la evaluación y analizar los resultados
  6. Evaluar la generalización del método

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Protocolo de evaluación

- ¿Qué procedimiento se debe seguir para evaluar un algoritmo/enfoque/modelo de PLN?
  1. Definir el objetivo y el alcance de la evaluación
  2. Determinar el tipo de evaluación
  3. Seleccionar las métricas de evaluación
  4. Recopilar y preparar los datos
  5. Ejecutar la evaluación y analizar los resultados
  6. Evaluar la generalización del método
     - ¿Cómo de generalizable es mi método/modelo? ¿Con qué otros conjuntos de datos puedo probarlo? ¿podría aplicar mi método a alguna otra tarea de PLN?

Protocolo definido inspirado en: https://es.linkedin.com/advice/1/what-best-ways-evaluate-natural-language-wekwc?lang=es

experto en procesamiento del lenguaje natural

Índice

1. Evaluación en PLN
2. Tipos de evaluación
3. Esquemas de evaluación
4. Métricas de evaluación
5. Protocolo de evaluación
6. **Caso práctico**
7. Bibliografía de referencia

experto en procesamiento del lenguaje natural

Caso práctico

Tareas de PLN:

- Clasificación de texto
- Extracción de entidades
- Generación de texto
- Resúmenes automáticos
- Respuestas a preguntas

experto en procesamiento del lenguaje natural

Caso práctico

Resumen de un texto:

Mecanismo para realizar una síntesis de un texto extenso, reduciendo su tamaño y extrayendo únicamente la información relevante.

Tipos de resúmenes automáticos:

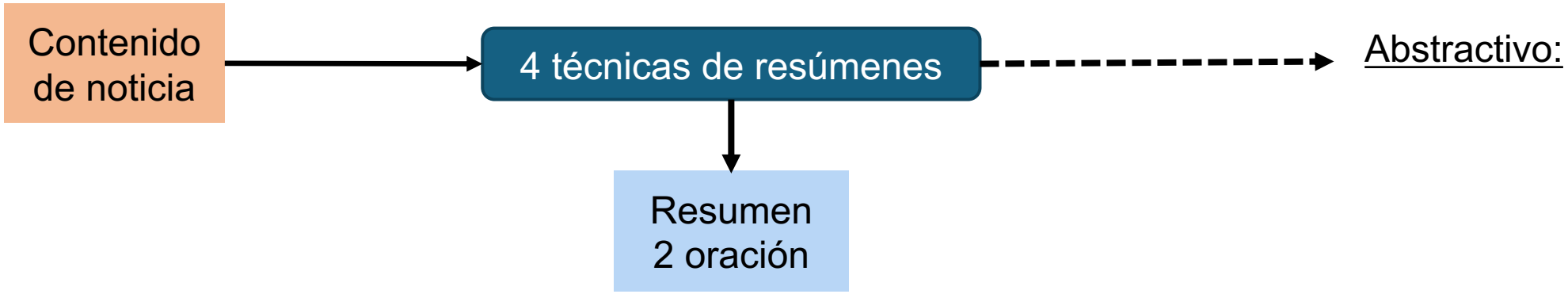
- Extractivos
- Abstractivos
- Híbridos

experto en procesamiento del lenguaje natural

Caso práctico

Dado un texto de uno noticia se ha resumido utilizando cuatro métodos de resúmenes abstractivos:

- Copilot (https://copilot.microsoft.com/)
- ChatGPT 3.5 (https://chat.lmsys.org/)
- Mixtral-8x7b-instruct-v0.1 (https://chat.lmsys.org/)
- ChatGPT 4 (Con subscription)



experto en procesamiento del lenguaje natural

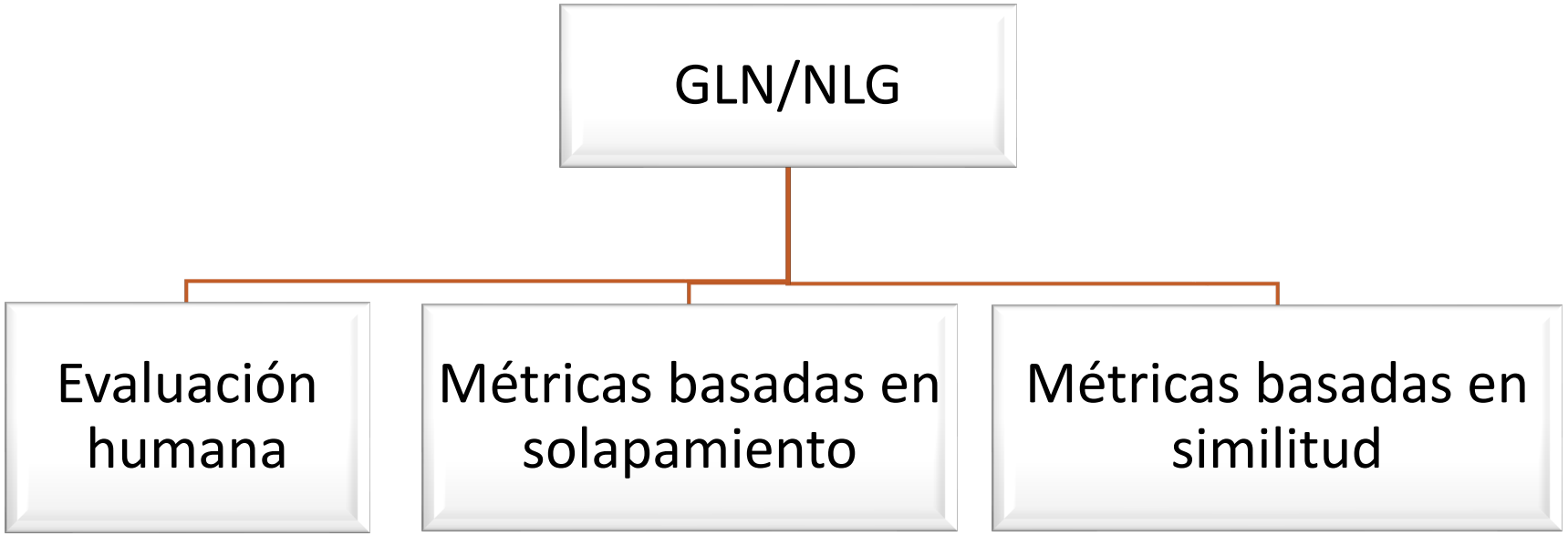
Caso práctico

Dataset: MLSUM (https://huggingface.co/datasets/mlsum)

- Texto
- Resumen humano
- Tema
- Titular
- Url
- Fecha

experto en procesamiento del lenguaje natural

Caso práctico



experto en procesamiento del lenguaje natural

Bibliotecas de evaluación

Evaluate

PYCARET

TorchMetrics

TensorFlow

scikit learn

experto en procesamiento del lenguaje natural

Índice

1. Evaluación en PLN
2. Tipos de evaluación
3. Esquemas de evaluación
4. Métricas de evaluación
5. Protocolo de evaluación
6. Caso práctico
7. **Bibliografía de referencia**

experto en procesamiento del lenguaje natural

Bibliografía de referencia

- The Amazing World of Neural Language Generation (tutorial del EMNLP 2020): https://nlg-world.github.io/Part-IV-Evaluation.pdf (a partir de 1h:37min:55 seg del video: https://virtual.2020.emnlp.org/tutorial_T7.html)
- PHILIP RESNIK AND JIMMY LIN. Evaluation of NLP Systems: https://home.cs.colorado.edu/~jbg/teaching/CMSC_773_2012/reading/evaluation.pdf
- https://clementbm.github.io/theory/2021/12/23/rouge-bleu-scores.html

experto en procesamiento del lenguaje natural

Métricas y métodos de evaluación en Procesamiento del Lenguaje Natural (PLN)

¡MUCHAS GRACIAS!

Elena Lloret Pastor

elloret@dlsi.ua.es

experto en procesamiento del lenguaje natural

Universitat d'Alacant
Universidad de Alicante
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
UJI UNIVERSITAT JAUME I
valgrAI