Análisis de Textos

Grupo PLN – InCo
2024

Problemas, métodos y evaluación en PLN

Problemas

- Traducción
- Resumen
- Extracción de información (entidades, relaciones, correferencias, ...)
- Categorización de documentos (clasificación / clustering)
- Búsqueda de respuestas (Q&A)
- Análisis de subjetividad: sentimiento (polaridad, emociones), humor, ironía/sarcasmo, discurso de odio, ...
- Diálogo (chatbots)
- Desarrollo de herramientas de análisis lingüístico
- Desarrollo de métodos (word embeddings, machine learning, deep learning)
- ...

Problemas

- Traducción
- Resumen
- Extracción de información (entidades, relaciones, correferencias, ...)
- Categorización de documentos (clasificación / clustering)
- Búsqueda de respuestas (Q&A)
- Análisis de subjetividad: sentimiento (polaridad, emociones), humor, ironía/sarcasmo, discurso de odio, ...
- Diálogo (chatbots)
- Desarrollo de herramientas de análisis lingüístico
- Desarrollo de métodos (word embeddings, machine learning, deep learning)
- ...

Extracción de información

- Named Entity Recognition (NER)
- Extracción de relaciones
- Extracción de eventos
- Template filling

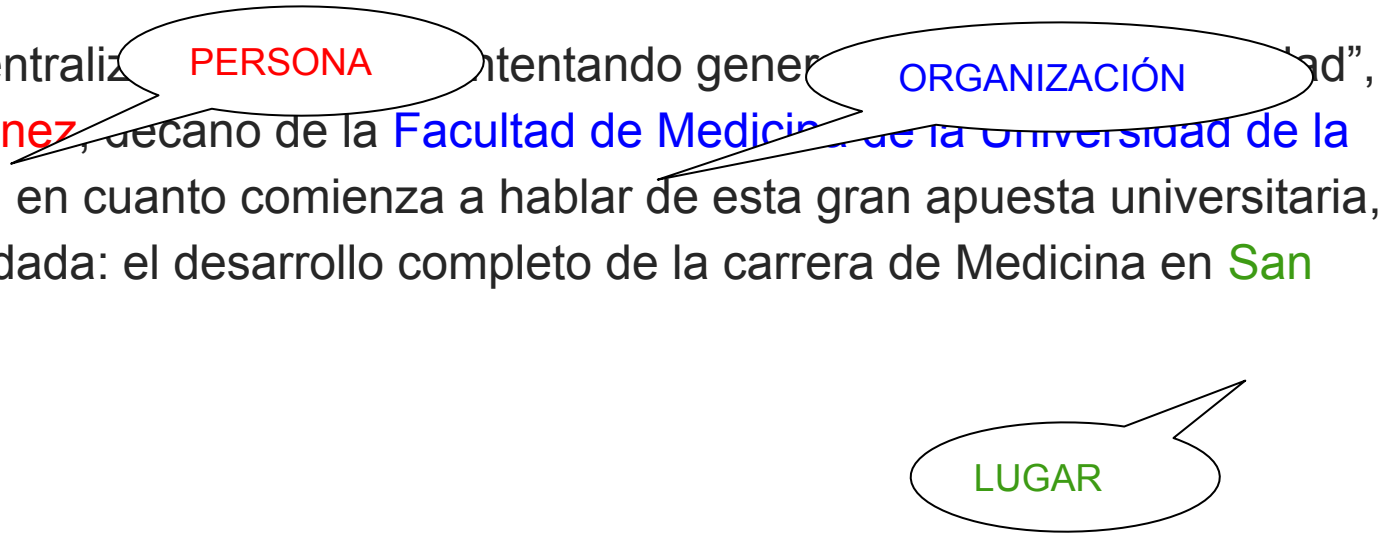
Extracción de información

- **Named Entity Recognition (NER)**

“No estamos descentralizando, estamos intentando generar una nueva centralidad”, afirma Miguel Martínez, decano de la Facultad de Medicina de la Universidad de la República (Udelar), en cuanto comienza a hablar de esta gran apuesta universitaria, largamente demandada: el desarrollo completo de la carrera de Medicina en el San José.

Extracción de información

- **Named Entity Recognition (NER)**



“No estamos descentraliz ntentando gener ad”, afirma Miguel Martínez, decano de la Facultad de Medicin de la Universidad de la República (Udelar), en cuanto comienza a hablar de esta gran apuesta universitaria, largamente demandada: el desarrollo completo de la carrera de Medicina en San José.

Extracción de información

- **Named Entity Recognition (NER)**

PERSONA

ORGANIZACIÓN

"No estamos descentralizando, estamos intentando generar una nueva centralidad", afirma Miguel Martínez, decano de la Facultad de Medicina de la Universidad de la República (Udelar), en cuanto comienza a hablar de esta gran apuesta universitaria, largamente demandada: el desarrollo completo de la carrera de Medicina en San José.

LUGAR

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	**Mt. Sanitas** is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states	**Palo Alto** is raising the fees for parking.

Figure 8.5 A list of generic named entity types with the kinds of entities they refer to.

(Jurafsky et al., 3rd edition draft)

Extracción de información

- **Named Entity Recognition (NER)**
  - Clasificación de secuencias
  - Esquema B-I-O

Extracción de información

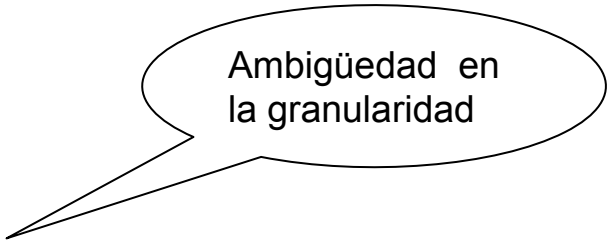
… afirma Miguel Martínez, decano de la Facultad de Medicina de la Universidad de la República (Udelar), en cuanto ...

afirma	O
Miguel	B-PER
Martínez	I-PER
,	O
decano	O
de	O
la	O
Facultad	B-ORG
de	I-ORG
Medicina	I-ORG
de	I-ORG
la	I-ORG
Universidad	I-ORG
de	I-ORG
la	I-ORG
República	I-ORG
(UdelaR)	I-ORG
,	O
en	O

Extracción de información

… afirma Miguel Martínez, decano de la Facultad de Medicina de la Universidad de la República (Udelar), en cuanto ...

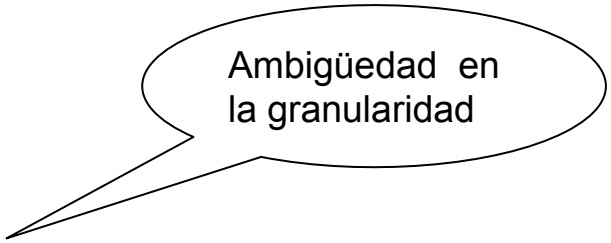
afirma	O
Miguel	B-PER
Martínez	I-PER
,	O
decano	O
de	O
la	O
Facultad	**B-ORG**
de	**I-ORG**
Medicina	**I-ORG**
de	**O**
la	**O**
Universidad	**B-ORG**
de	**I-ORG**
la	**I-ORG**
República	**I-ORG**
(UdelaR)	**I-ORG**
,	O
en	O



Extracción de información

… afirma Miguel Martínez, decano de la Facultad de Medicina de la Universidad de la República (Udelar), en cuanto ...

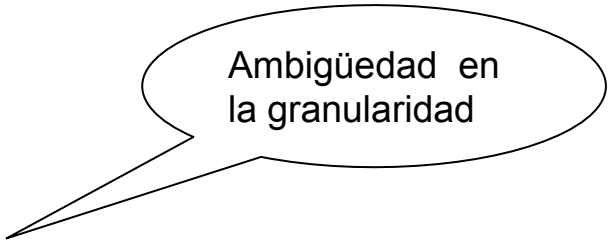
afirma	O
Miguel	B-PER
Martínez	I-PER
,	O
decano	O
de	O
la	O
Facultad	B-ORG
de	I-ORG
Medicina	I-ORG
de	O
la	O
Universidad	B-ORG
de	I-ORG
la	I-ORG
República	I-ORG
(UdelaR)	**B-ORG**
,	O
en	O



Extracción de información

… afirma Miguel Martínez, decano de la Facultad de Medicina de la Universidad de la República (Udelar), en cuanto ...

afirma	O
Miguel	B-PER
Martínez	I-PER
,	I-PER
decano	I-PER
de	I-PER
la	I-PER
Facultad	I-PER
de	I-PER
Medicina	I-PER
de	I-PER
la	I-PER
Universidad	I-PER
de	I-PER
la	I-PER
República	I-PER
(UdelaR)	I-PER
,	O
en	O



Extracción de información

Evolución de métodos (clasificación secuencial):

- Reglas
- ML basado en atributos: CRF
- Redes neuronales: Bi-LSTM (+CRF)
- Transformers (modelo neuronal + fine tuning)

Usados en sistemas comerciales

A partir de representaciones vectoriales, sin atributos

Extracción de información

Métodos para NER y NEC (clasificación secuencial):

- ML basado en atributos: CRF (Conditional Random Fields)
- Atributos usuales:

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
gazetteer features

Figure 8.15 Typical features for a feature-based NER system.

(Jurafsky et al., 3rd edition draft)

Extracción de información

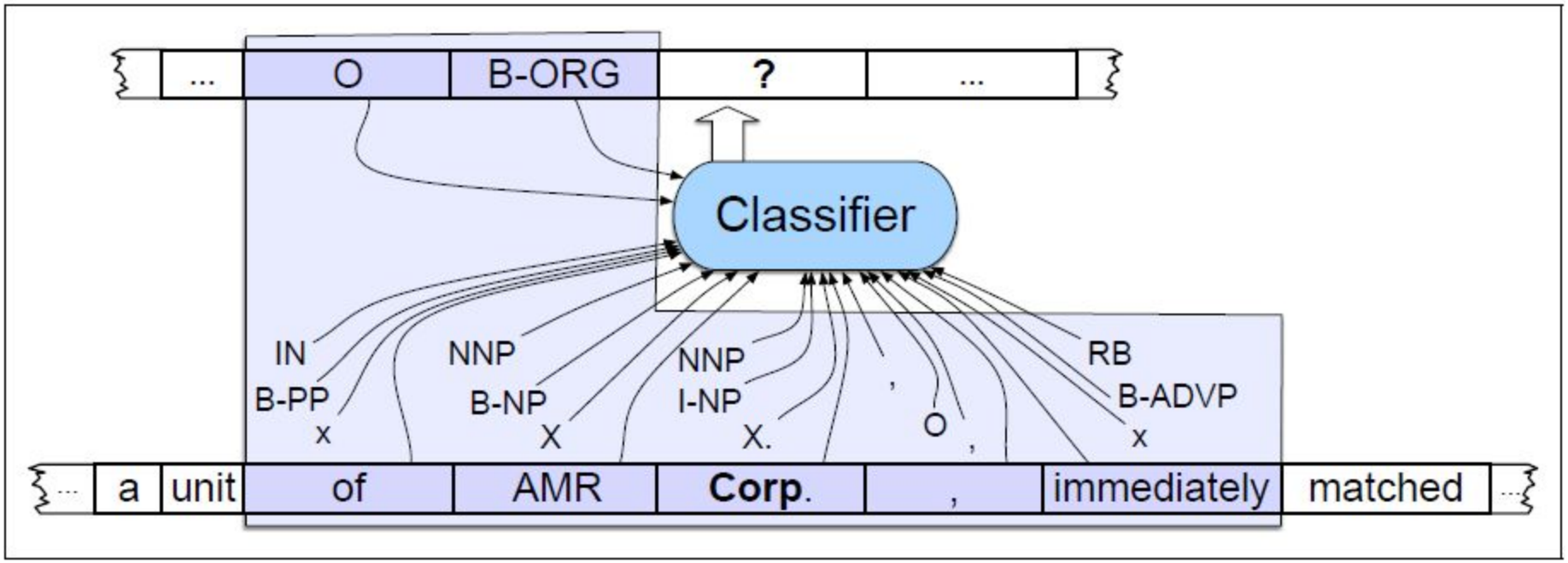
Words	POS	Short shape	Gazetteer	BIO Label
Jane	NNP	Xx	0	B-PER
Villanueva	NNP	Xx	1	I-PER
of	IN	x	0	O
United	NNP	Xx	0	B-ORG
Airlines	NNP	Xx	0	I-ORG
Holding	NNP	Xx	0	I-ORG
discussed	VBD	x	0	O
the	DT	x	0	O
Chicago	NNP	Xx	1	B-LOC
route	NN	x	0	O
.	.	.	0	O

Figure 8.16 Some NER features for a sample sentence, assuming that Chicago and Villanueva are listed as locations in a gazetteer. We assume features only take on the values 0 or 1, so the first POS feature, for example, would be represented as $1\{\text{POS} = \text{NNP}\}$.

(Jurafsky et al.,
3rd edition draft)

Extracción de información

Se usa la clase asignada a palabras anteriores para clasificar la palabra actual



Información disponible para el clasificador (en el ejemplo, la palabra a clasificar es *Corp.*)
(Jurafsky et al., 3rd edition draft 2019 (no disponible)

Extracción de información

- **Redes neuronales: Bi-LSTM (+CRF)**

(Jurafsky et al. 3rd edition draft 2019, no disponible)

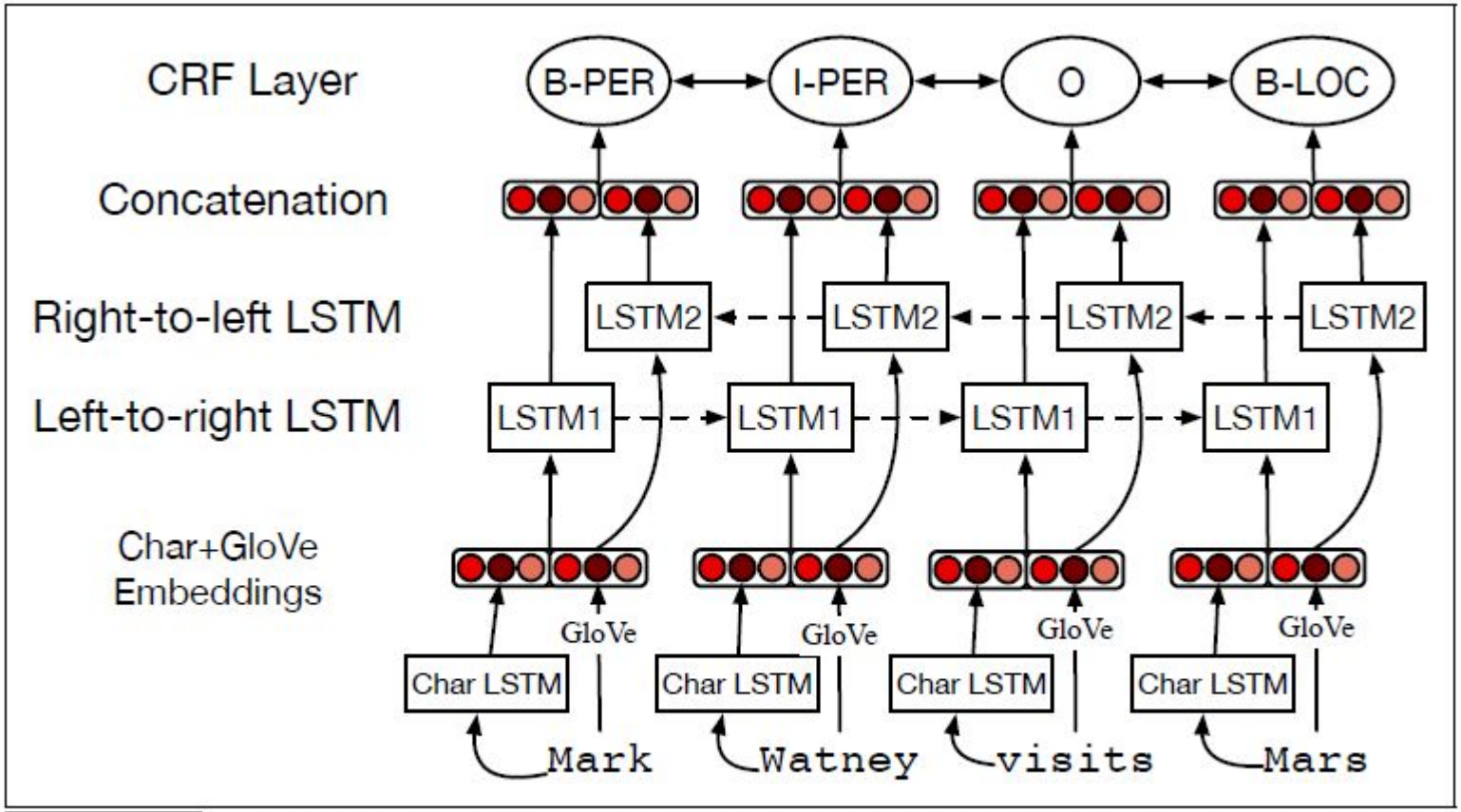
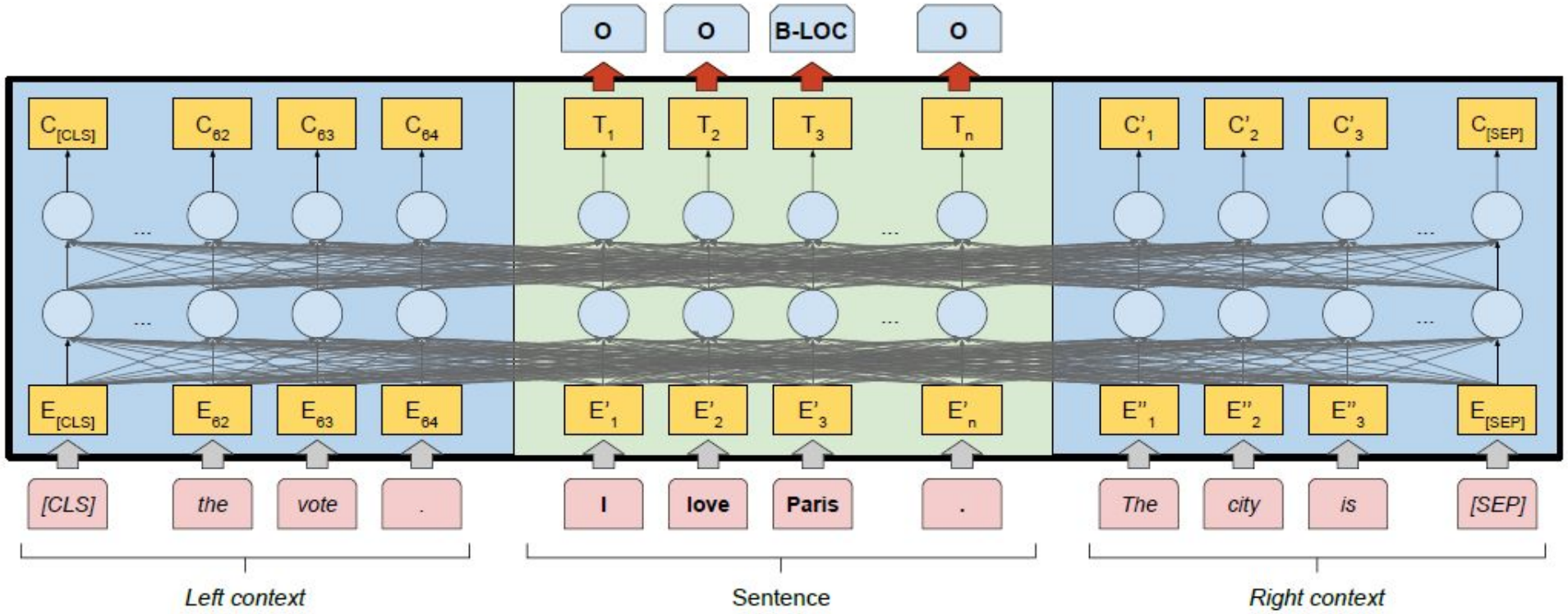


Figure 17.8 Putting it all together: character embeddings and words together a bi-LSTM sequence model. After (Lample et al., 2016)

Extracción de información

- **Modelo de lenguaje neuronal basado en Transformers**
  - **modelo de lenguaje + fine tuning**

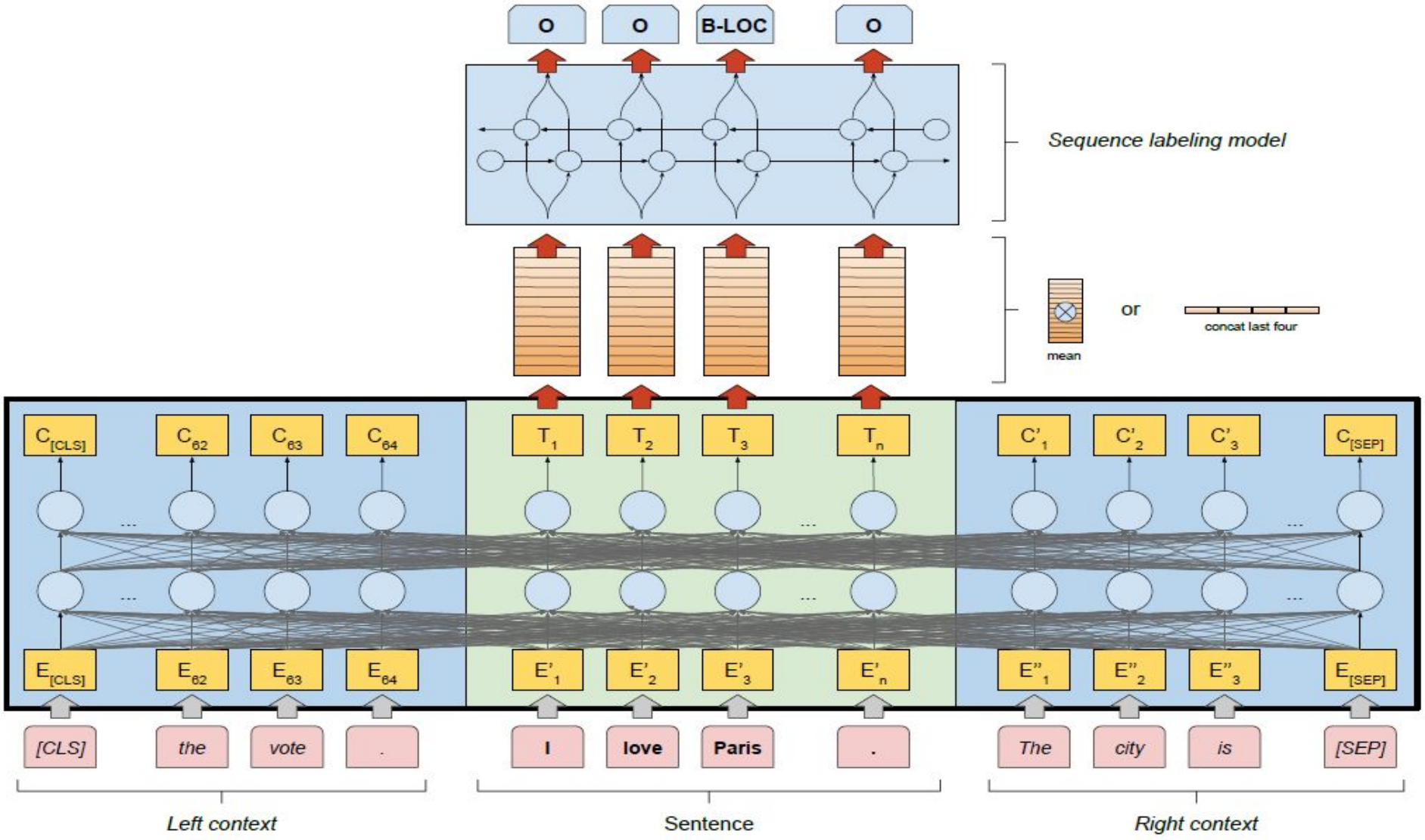


Stefan Schweter and Alan Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition. https://arxiv.org/pdf/2011.06993v2.pdf

Extracción de información

- **Modelo de lenguaje neuronal basado en Transformers**
  - modelo de lenguaje + (LSTM+CRF)

    -> solo se entrenan capas LSTM y CRF



Stefan Schweter and Alan Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition. https://arxiv.org/pdf/2011.06993v2.pdf

Extracción de información

Problemas asociados a NER:

- Correferencias:

  El día miércoles dos el funcionario del Registro de Empresas José Marchese fue enviado al Banco de la República a hacer efectivo tres cheques. En el Banco de la República se le rechazaron los cheques. Los tres cheques pertenecen a una libreta que fue retirada por el Sr. Marchese en el año 2020. El banco le solicitó al funcionario que concurriera al día siguiente para resolver el incoveniente. Él indicó que no le era posible concurrir.

- Entity linking: relacionar con entidades del mundo real.
- Problemas asociados a dominios específicos: entidades en dominio biomédico, dominio legal, documentos antiguos, etc.
  - Las herramientas genéricas no funcionan tan bien en estos dominios particulares.

Extracción de información

- NER y NEC
- Extracción de relaciones
- Extracción de eventos
- Template filling

Extracción de información

Extracción de relaciones

“No estamos descentralizando, estamos intentando generar una nueva centralidad”, afirma Miguel Martínez, decano de la Facultad de Medicina de la Universidad de la República (Udelar), en cuanto comienza a hablar de esta gran apuesta universitaria, largamente demandada: el desarrollo completo de la carrera de Medicina en San José.

Extracción de información

Extracción de relaciones

“No estamos descentralizando, estamos intentando generar una nueva centralidad”, afirma Miguel Martínez, decano de la Facultad de Medicina de la Universidad de la República (Udelar), en cuanto comienza a hablar de esta gran apuesta universitaria, largamente demandada: el desarrollo completo de la carrera de Medicina en San José.

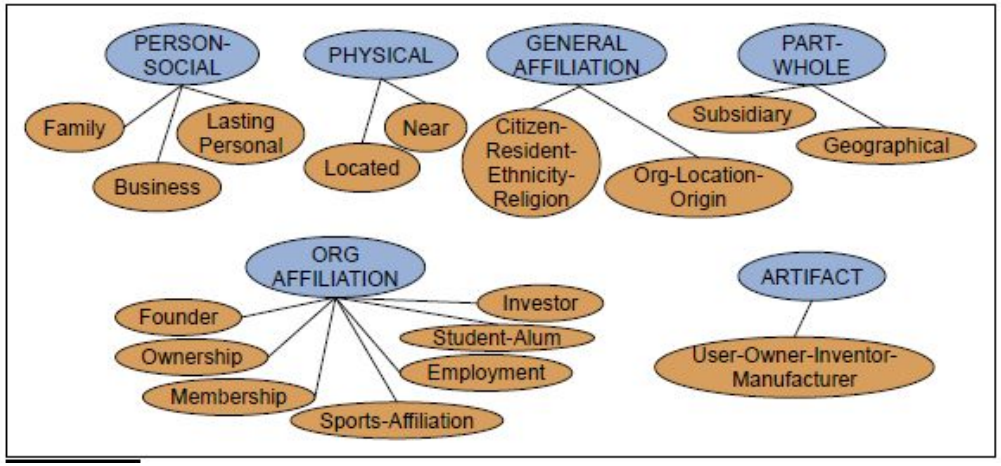
Entidad 1	**Relación**	**Entidad 2**
Miguel Martínez	empleado (decano)	Facultad de Medicina
Facultad de Medicina	parte de	Univ. de la República
Carrera de Medicina	ubicación	San José

Extracción de información

Extracción de relaciones

“No estamos descentralizando, estamos intentando generar una nueva centralidad”, afirma Miguel Martínez, decano de la Facultad de Medicina de la Universidad de la República (Udelar), en cuanto comienza a hablar de esta gran apuesta universitaria, largamente demandada: el desarrollo completo de la carrera de Medicina en San José.

Relaciones utilizadas en ACE
(Automatic Content Extraction task)
(Jurafsky et al, 3rd edition draft)



Extracción de información

Extracción de relaciones

- Las relaciones extraídas pueden ser almacenadas como triplas RDF (Resource Description Framework): tupla entity-relation-entity (subject-predicate-object).

- Existen grandes bases de relaciones:
  - DBpedia: ontología derivada de Wikipedia con alrededor de 2 billones de triplas RDF.
  - Freebase: creada a partir de información estructurada (*infoboxes*) de Wikipedia.

- Métodos:
  - Patrones manuales
  - ML supervisado
  - Métodos semi-supervisados: bootstrapping, distant supervision
  - Métodos no supervisados (Open IE)

Extracción de información

- NER y NEC
- Extracción de relaciones
- Extracción de eventos
- Template filling

Extracción de información

- Extracción de eventos
  - Identificar eventos y sus argumentos.
  - Ubicarlos temporalmente (identificar expresiones temporales).
  - Saber si ocurrieron o no (factualidad).

- Template filling
  - Extraer información de interés en dominios particulares.
  - Completar templates o slots predefinidos.

Extracción de información

Bibliografía

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, 3rd edition draft. Stanford. 2024. [https://web.stanford.edu/~jurafsky/slp3/ed3bookfeb3_2024.pdf Acceso: junio 2024].

Capítulo 8: Sequence labeling for Parts of Speech and Named Entities

Capítulo 19: Information Extraction: Relations, Events, and Time

Análisis de sentimiento

Análisis de Sentimiento o Minería de Opiniones
[Turney, 2002; Wiebe et al., 2005; Pang & Lee, 2008; Liu, 2020]

Diferentes tareas:

- Clasificación de textos (tweets, opiniones sobre películas, hoteles, noticias, etc.) según su polaridad (positivos o negativos) o emociones (alegría, tristeza, enojo, sorpresa, ...).
- Extraer opiniones a partir de textos (por ejemplo, prensa).
- Para cada opinión:
  - Autor
  - Polaridad (P, N, P+, N+, Neu, etc.)
  - Aspectos: celular -> batería P, pantalla N
  - ...

Análisis de sentimiento

Qué bien que marcaron a Messi

Gracias a la gente de Brooksfield por su excelente atención. https://t.co/r38i4q7clV,P

pues a mi me ha encantado #sherlock

@user top secret: el #psoe de andalucía despilfarra otros 90 millones de euros en subvenciones. http://t.co/aiwhnxrj

Análisis de sentimiento

Qué bien que marcaron a Messi

Gracias a la gente de Brooksfield por su excelente atención. https://t.co/r38i4q7clV,P

pues a mi me ha encantado #sherlock

@user top secret: el #psoe de andalucía despilfarra otros 90 millones de euros en subvenciones. http://t.co/aiwhnxrj

Análisis de sentimiento

***Qué** bien que marcaron a Messi* 

Gracias a la gente de Brooksfield por su excelente atención. https://t.co/r38i4q7clV,P 

pues a mi me ha encantado #sherlock 

@user top secret: el #psoe de andalucía despilfarra otros 90 millones de euros en subvenciones. http://t.co/aiwhnxrj 

Análisis de sentimiento

Entró la magia al camp nou vamo #Messi

Uruguay sin mucho fútbol pero con abundante marca y actitud

volvió la #bestia... volvió Lionel Andrés #Messi!!! #GIGANTE #FCBDEP

Qué fantasma este pelado! Roja inexistente.

Después de esto como para que no lo echaran a Sanmartino

ajaja muy bueno https://t.co/i4mfeb1qdY

Análisis de sentimiento

Entró la magia al camp nou vamo #Messi 

Uruguay sin mucho fútbol ***pero*** *con abundante marca y actitud* 

volvió la #bestia... volvió Lionel Andrés #Messi!!! #GIGANTE #FCBDEP 

Qué fantasma este pelado! Roja inexistente. 

Después de esto como para que no lo echaran a Sanmartino ¿   ?

ajaja muy bueno https://t.co/i4mfeb1qdY 

¿pero de qué se está hablando?

Análisis de sentimiento

Tareas previas al análisis del contenido del texto:

- Obtener los textos de interés: tweets, comentarios en foros o blogs, mails, etc.
- Limpiarlos: eliminar o sustituir símbolos, enlaces, etiquetas.
- Guardarlos: definir el modelo de datos y los campos relevantes.

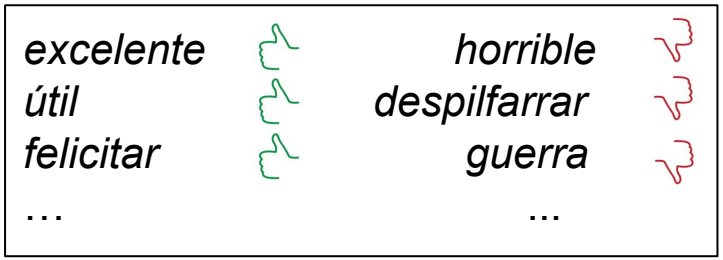
Análisis de sentimiento

Primer enfoque: reglas y patrones manuales

- Escritura de reglas en base al análisis de ejemplos.
- Las reglas incluyen conocimiento lingüístico:
  - Lematización.
  - POS tagging (categorías gramaticales de palabras).
  - Parsing (estructura sintáctica de oraciones).
- También conocimiento específico del dominio o del problema:
  - Léxicos afectivos para análisis de sentimiento.
  - Negadores e intensificadores.

Análisis de sentimiento

- **Léxicos afectivos**



Análisis de sentimiento

- Léxicos afectivos
- Lematización

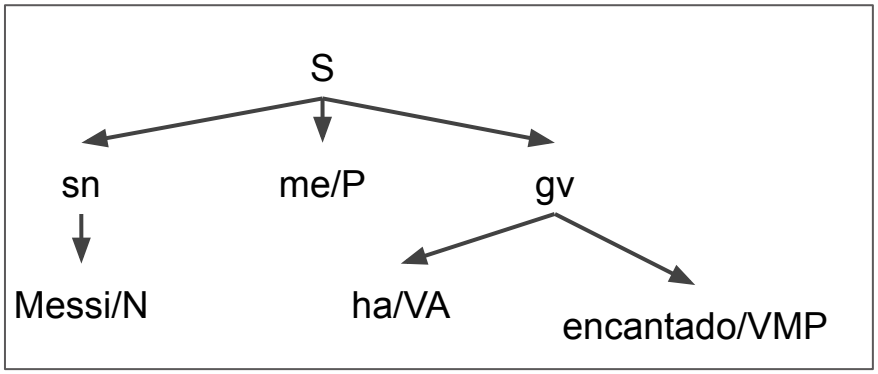
sugerencia/sugerencias → sugerencia
despilfarra/despilfarraron/despilfarrando → despilfarrar
buena/bueno/buenas/buenos → bueno

Análisis de sentimiento

- Léxicos afectivos
- Lematización
- POS-tagging

Análisis de sentimiento

- Léxicos afectivos
- Lematización
- POS-tagging
- Parsing



Análisis de sentimiento

Gracias a la gente de Brooksfield por su excelente atención.

Posible regla:

```
busco palabras en léxicos afectivos
si cantidad_palabras_positivas > cantidad_palabras_negativas
    el tweet es positivo
si cantidad_palabras_positivas < cantidad_palabras_negativas
    el tweet es negativo
si no
    el tweet es neutro
```

Análisis de sentimiento

Gracias a la gente de Brooksfield por sus excelentes sugerencias.

Posible regla:

```
busco lemas en léxicos afectivos
si cantidad_lemas_positivos > cantidad_lemas_negativos
    el tweet es positivo
si no
    el tweet es negativo
```

Análisis de sentimiento

*La atención de Brooksfield es excelente, **pero** los productos son de mala calidad.*

Posible regla:

```
busco lemas en léxicos afectivos

si cantidad_lemas_positivos = 0 y cantidad_lemas_negativos = 0
   el tweet es neutro

si no
    si cantidad_lemas_positivos > cantidad_lemas_negativos
       el tweet es positivo
    si no, si cantidad_lemas_positivos < cantidad_lemas_negativos
       el tweet es negativo
    si no
       el tweet es mixto
```

Análisis de sentimiento

*La atención de Brooksfield es excelente, **pero** los productos son de mala calidad.*

Posible regla:

```
busco lemas en léxicos afectivos

si cantidad_lemas_positivos = 0 y cantidad_lemas_negativos = 0
   el tweet es no_sent

si no
    si cantidad_lemas_positivos > cantidad_lemas_negativos
       el tweet es positivo
    si no, si cantidad_lemas_positivos < cantidad_lemas_negativos
       el tweet es negativo
    si no
       el tweet es neutro
```

Análisis de sentimiento

*La atención de Brooksfield es excelente, **pero** los productos son de mala calidad.*

Posible regla:

```
busco lemas en léxicos afectivos
análisis sintáctico para tomar el segmento introducido por pero
si cantidad_lemas_positivos > cantidad_lemas_negativos
    el tweet es positivo
si no, si cantidad_lemas_positivos < cantidad_lemas_negativos
    el tweet es negativo
si no
    el tweet es neutro
```

Análisis de sentimiento

*La atención de Brooksfield es excelente, **pero** los productos **no** son buenos.*

Posible regla:

```
busco lemas en léxicos afectivos
invierto valor afectivo de palabras negadas
si cantidad_lemas_positivos > cantidad_lemas_negativos
    el tweet es positivo
si no, si cantidad_lemas_positivos < cantidad_lemas_negativos
    el tweet es negativo
si no
    el tweet es neutro
```

Análisis de sentimiento

- Es muy difícil abarcar todos los casos escribiendo reglas manuales.
- Se amplía la cobertura con grandes cantidades de ejemplos y métodos de aprendizaje automático supervisado, también con métodos híbridos.
- Esto es posible si se cuenta con conjuntos de datos anotados.
- Actualmente, los grandes modelos de lenguaje logran buenos resultados, incluso sin corpus anotados específicamente para la tarea.

Análisis de sentimiento

Segundo enfoque: Aprendizaje automático basado en *features* manuales

- Corpus de tweets
  - Cada uno con su clasificación (positivo, negativo, etc...)
  - Etiquetados a mano

- Aprender una función que prediga la clase dado el tweet:
  - Entradas de la función: conjunto de atributos (*features*)
  - Salidas: la clase predicha

- En estos casos hablamos de aprendizaje automático supervisado

Análisis de sentimiento

Algunos atributos posibles para análisis de sentimiento:

- **word embeddings**
- palabras (*bag of words*)
- lemas
- categorías gramaticales
- cantidad de palabras positivas/negativas
- presencia de negación
- información sintáctica
- ...

Análisis de sentimiento

Tercer enfoque: aprendizaje profundo

Deep learning

- word embeddings como entrada de la red
- diferentes arquitecturas
- bi-LSTM mejores resultados

Análisis de sentimiento

Tercer enfoque: aprendizaje profundo

- modelo de lenguaje neuronal (transformers) + fine tuning

Análisis de sentimiento

Análisis de sentimiento en tweets en español: TASS (SEPLN)

[García Vega et al, 2020]

(organizado por la Sociedad Española para el Procesamiento de Lenguaje Natural)

- La organización distribuye datos para entrenamiento y validación a los participantes.
- 1 mes y medio para desarrollo de sistemas.
- Se publican datos de testeo (sin anotaciones).
- Los participantes envían los resultados sobre testeo, que son evaluados contra el gold standard y rankeados.
- Medidas de evaluación: Macro-F (promedio de Medida F por cada clase) y Accuracy.

Análisis de sentimiento

Conjunto de etiquetas del TASS (hasta 2019): P, N, NEU, NONE

“@user top secret: el #psoe de andalucía despilfarra otros 90 millones de euros ensubvenciones. http://t.co/aiwhnxrj” **N**

“pues a mi me ha encantado #sherlock” **P**

“el principio ha sido súper iconic pero con tanto movimiento iba muerta la pobre #vmas” **NEU**

“hoy conoceremos datos definitivos de 2011 del padrón municipal. datos ine:en españa hay casi un 1% más de mujeres q de hombres” **NONE**

Análisis de sentimiento

pysentimiento

- Entrenamiento de un modelo de lenguaje basado en Roberta (basado a su vez en BERT) con un gran corpus de tweets en español: RroBERTuito (Pérez et al., 2022).
- Mejoras significativas en una herramienta para análisis de subjetividad, que incluye análisis de sentimiento: pysentimiento (Pérez et al, 2021).
- La mejor configuración alcanza un 70.7% de Macro F para análisis de sentimiento, sobre el dataset de TASS 2020, usando la unión de todas las variantes.

Análisis de sentimiento

- Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424.
- Wiebe, J., Wilson, T., Cardie, C., 2005. Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation, Volume 39, Issue 2-3, pp 165-210.
- Pang, B., Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval Vol. 2, No 1-2, 1–135.
- Liu, B. 2020. Sentiment Analysis. Mining Opinions, Sentiments, and Emotions, 2nd ed.; Cambridge University Press: New York, NY, USA.
- García Vega, M. et al.. 2020. Overview of TASS 2020: Introducing Emotion Detection . Proceedings of TASS 2020: Workshop on Semantic Analysis at SEPLN (TASS 2020), co-located with 36th SEPLN Conference (SEPLN 2020), Malaga, Spain.
- Pérez, JM., Furman, D., Alemany, L., Luque, F. RoBERTuito: a pre-trained language model for social media text in Spanish. Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022.
- Pérez, JM., Giudici, JC., Luque, F. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks, arXiv cs.CL, 2021. https://arxiv.org/abs/2106.09462

Evaluación

Evaluar los resultados de nuestros métodos sobre textos nuevos es fundamental, ya que las herramientas de PLN no suelen ser 100% efectivas.

Evaluación

Métricas usuales

Medidas para cada clase

Precision: P = VP / (VP + FP)

Recall: R = VP / (VP + FN)

Combinación de las dos medidas: Medida F1 = 2.P.R / (P + R)

Promedio entre todas las clases:

- Macro F1
- Weighted F1 (promedio ponderado por cantidad de ocurrencias)

Acierto general (correctos de todas las clases / total)

Accuracy= (VP + VN) / (VP + VN + FP + FN)

No es una buena medida para datos desbalanceados

VP: verdaderos positivos
VN: verdaderos negativos
FP: falsos positivos
FN: falsos negativos

Evaluación