# UNIVERSIDAD DE LAS FUERZAS ARMADAS – ESPE

## DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

## Carrera de Software

**Modalidad Presencial**

**Tema:**

**“Modelo basado en Minería de textos y Procesamiento del Lenguaje Natural (NLP) para la gestión de artículos científicos – caso de estudio Nanopartículas”**

**Autor:**
CERÓN ÑAUÑAY MARÍA BELÉN

**Tutor:** ING. GUALOTUÑA ÁLVAREZ, TATIANA MARISOL, PHD

**SANGOLQUÍ, AGOSTO 2022**

# *Agenda*

- **INTRODUCCIÓN**
- **OBJETIVOS**
- **IMPLEMENTACIÓN**
- **RESULTADOS**
- **VALIDACIÓN DEL MODELO**
- **CONCLUSIONES Y RECOMENDACIONES**

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

# INTRODUCCIÓN

# Antecedentes

Compartir información desde cualquier parte del mundo

ProQuest · ProQuest Ebook Central · Science Direct · IEEE · Scopus · Ovid · pivot · ACM DL Digital Library · Springer · PubMed · WEB OF SCIENCE™

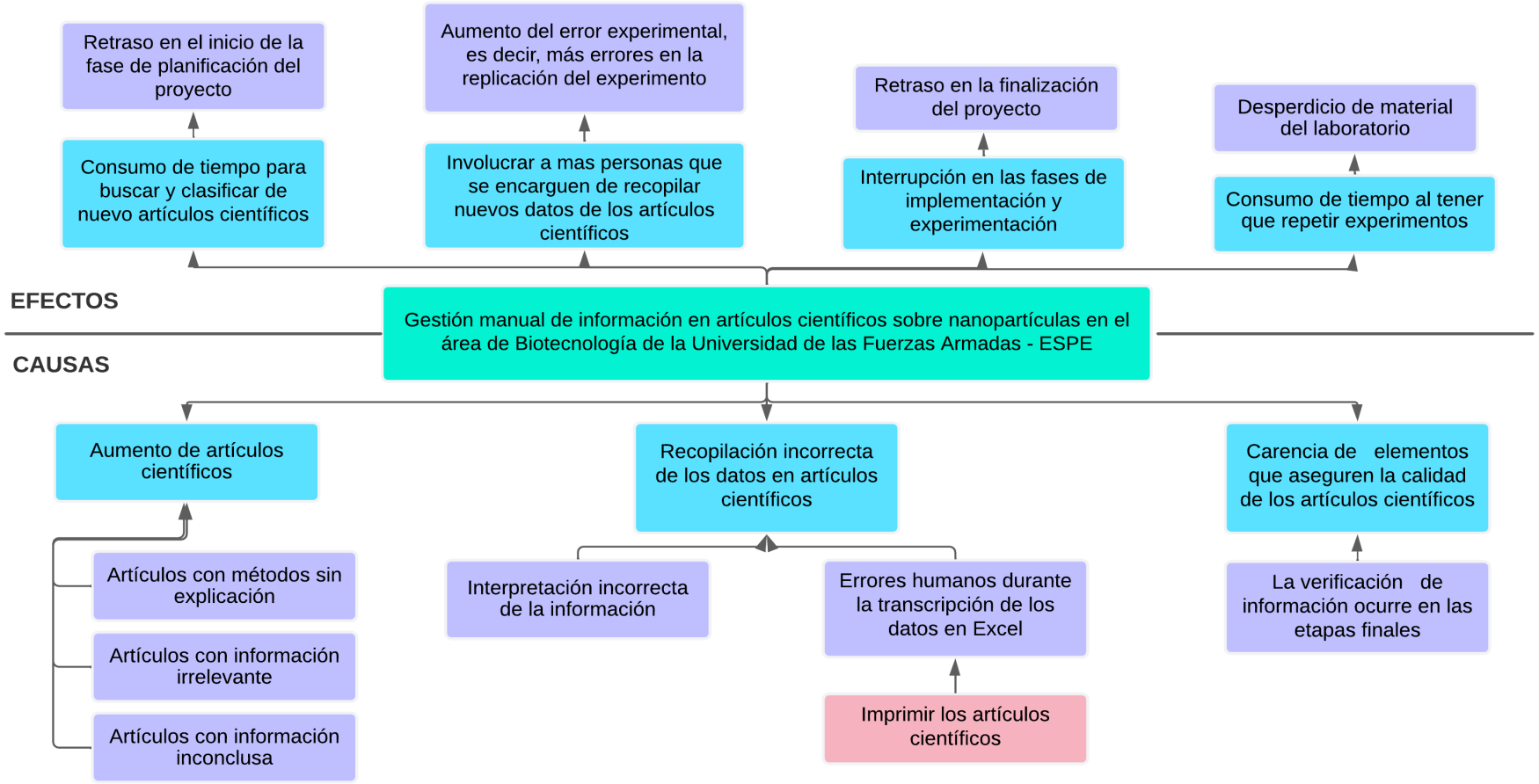
Bases de datos académicas


Artículos Científicos

Investigaciones científicas


Sobreabundancia de información

# Problemática



# OBJETIVOS

# *Objetivo general*

Desarrollar un modelo machine learning para la gestión de artículos científicos a través del cual se pueda establecer patrones de información que aporten al conocimiento científico

# *Objetivos específicos*

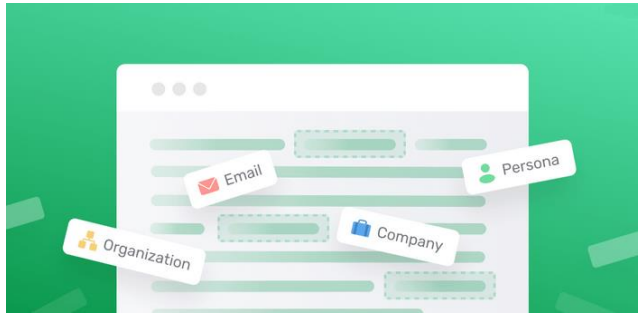
Realizar una revisión de literatura que permita determinar técnicas de minería de textos y Procesamiento del Lenguaje Natural (NLP), adecuadas para el procesamiento, etiquetado y extracción de datos.

Definir e implementar el modelo machine learning para la generación de patrones de comportamiento en las investigaciones relacionadas con nanopartículas.

Realizar la validación del modelo utilizando métricas de evaluación de rendimiento de clasificadores en tareas de extracción de datos.

# *Estado del Arte*

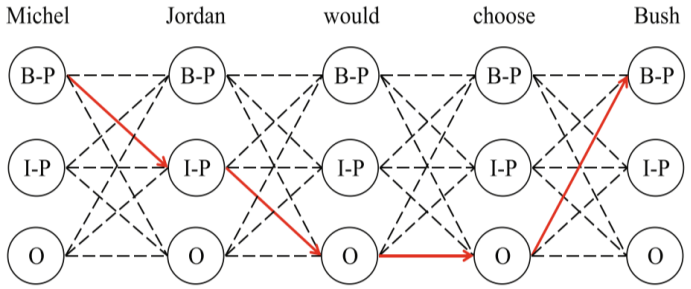
## NER




## Soluciones presentadas por los EP

| Modelo | Estudios |
|---|---|
| **BiLSTM** | Tres investigaciones: EP1, EP4, EP5. |
| **CRF** | Dos investigaciones: EP4, EP7. |
| **BERT** | Dos investigaciones: EP4, EP6. |
| **RNN** | Una investigación: EP7. |

## BIO encoding



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

# IMPLEMENTACIÓN

# *Design Science Research (DSR)*

# *Arquitectura de la solución*

# *Origen de los datos*

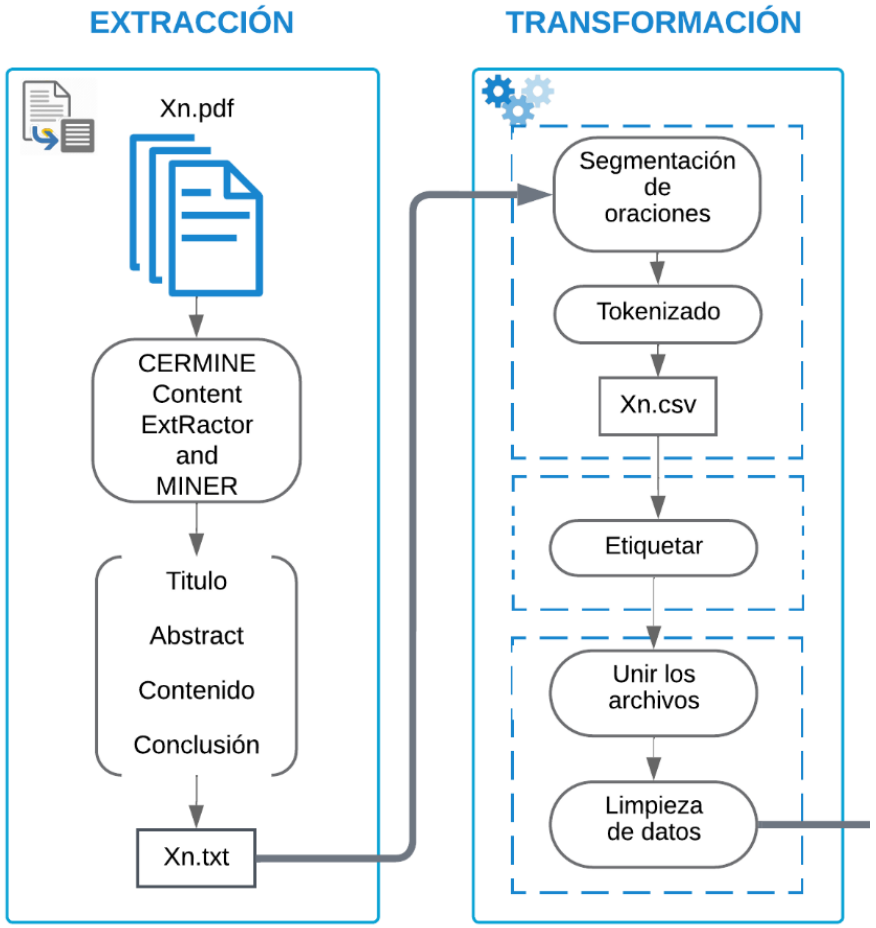
Comprensión de los datos → Recolección de datos → Análisis de los datos

- 141 entidades

| Material | Cantidad | Detalle |
|---|---|---|
| **Artículos sin anotaciones** | 100 | Artículos descargados directamente de la base de datos Scopus sin ninguna modificación. |
| **Artículos con anotaciones** | 100 | Artículos donde se encuentra subrayado la información necesaria a extraer. |
| **Base de datos** | 1 | Base de datos en Excel con la información extraída de los artículos, esta información se puede comparar con los artículos subrayados. |

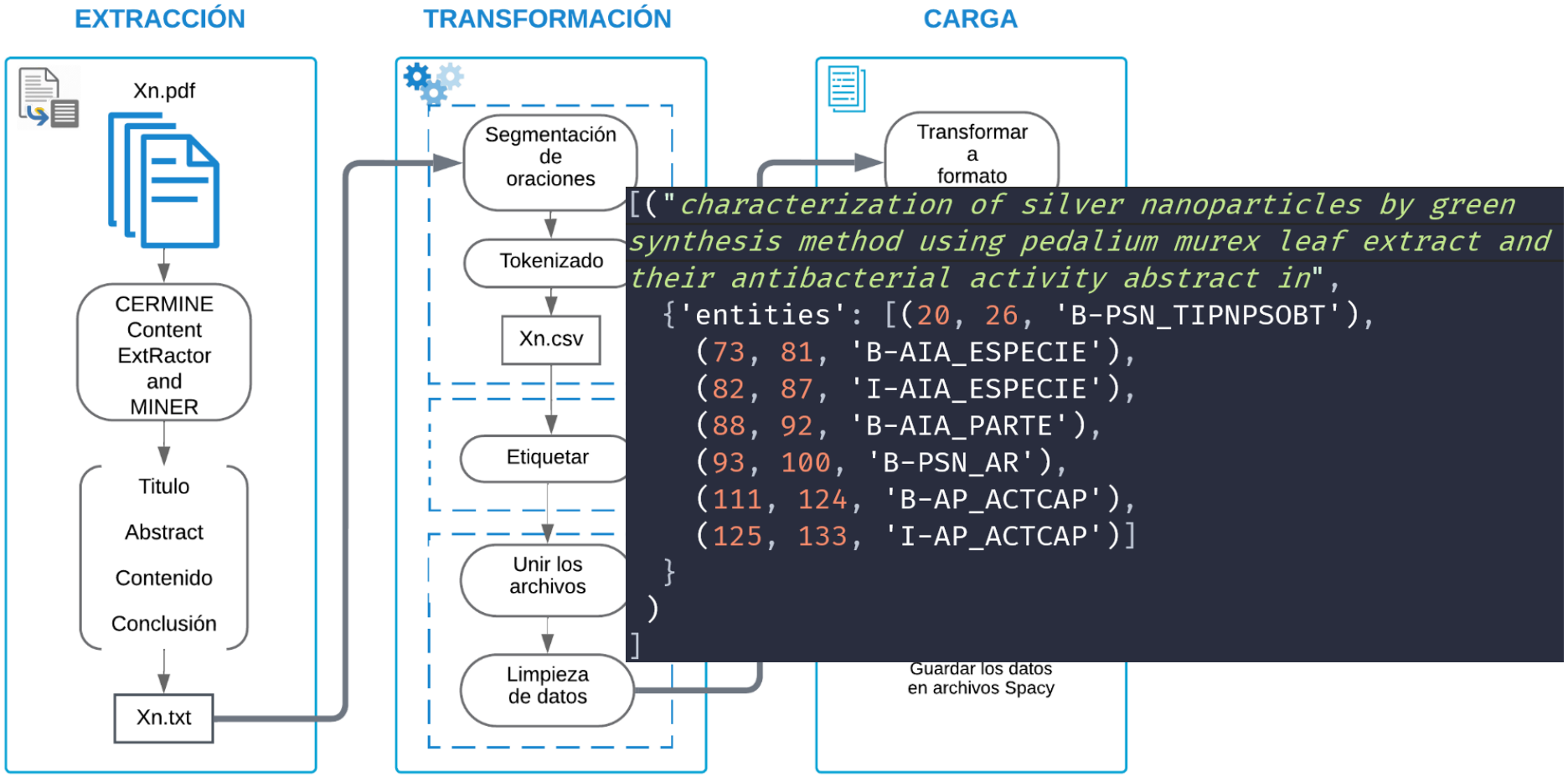
- 45 entidades
- 88 artículos científicos

# Proceso ETL




| id_paper | id_sentence | words | tag |
|---|---|---|---|
| X9.txt | 0 | characterizat | O |
| X9.txt | 0 | of | O |
| X9.txt | 0 | silver | B-PSN_TIPNPSOBT |
| X9.txt | 0 | nanoparticle | O |
| X9.txt | 0 | by | O |
| X9.txt | 0 | green | O |
| X9.txt | 0 | synthesis | O |
| X9.txt | 0 | method | O |
| X9.txt | 0 | using | O |
| X9.txt | 0 | pedalium | B-AIA_ESPECIE |
| X9.txt | 0 | murex | I-AIA_ESPECIE |
| X9.txt | 0 | leaf | B-AIA_PARTE |
| X9.txt | 0 | extract | B-PSN_AR |
| X9.txt | 0 | and | O |
| X9.txt | 0 | their | O |
| X9.txt | 0 | antibacterial | B-AP_ACTCAP |
| X9.txt | 0 | activity | I-AP_ACTCAP |
| X9.txt | 1 | abstract | O |
| X9.txt | 2 | in | O |

# Proceso ETL




```
[("characterization of silver nanoparticles by green synthesis method using pedalium murex leaf extract and their antibacterial activity abstract in",
  {'entities': [(20, 26, 'B-PSN_TIPNPSOBT'),
    (73, 81, 'B-AIA_ESPECIE'),
    (82, 87, 'I-AIA_ESPECIE'),
    (88, 92, 'B-AIA_PARTE'),
    (93, 100, 'B-PSN_AR'),
    (111, 124, 'B-AP_ACTCAP'),
    (125, 133, 'I-AP_ACTCAP')]
  }
 )
]
```

# *Entrenamiento del modelo*

Elección del modelo

Entrenamiento del módulo NER

SpaCy

Redes Neuronales Convolucionales

RoBERTa

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

# RESULTADOS

# *Evaluación del Rendimiento del modelo*

## Métricas de evaluación

**Precisión**

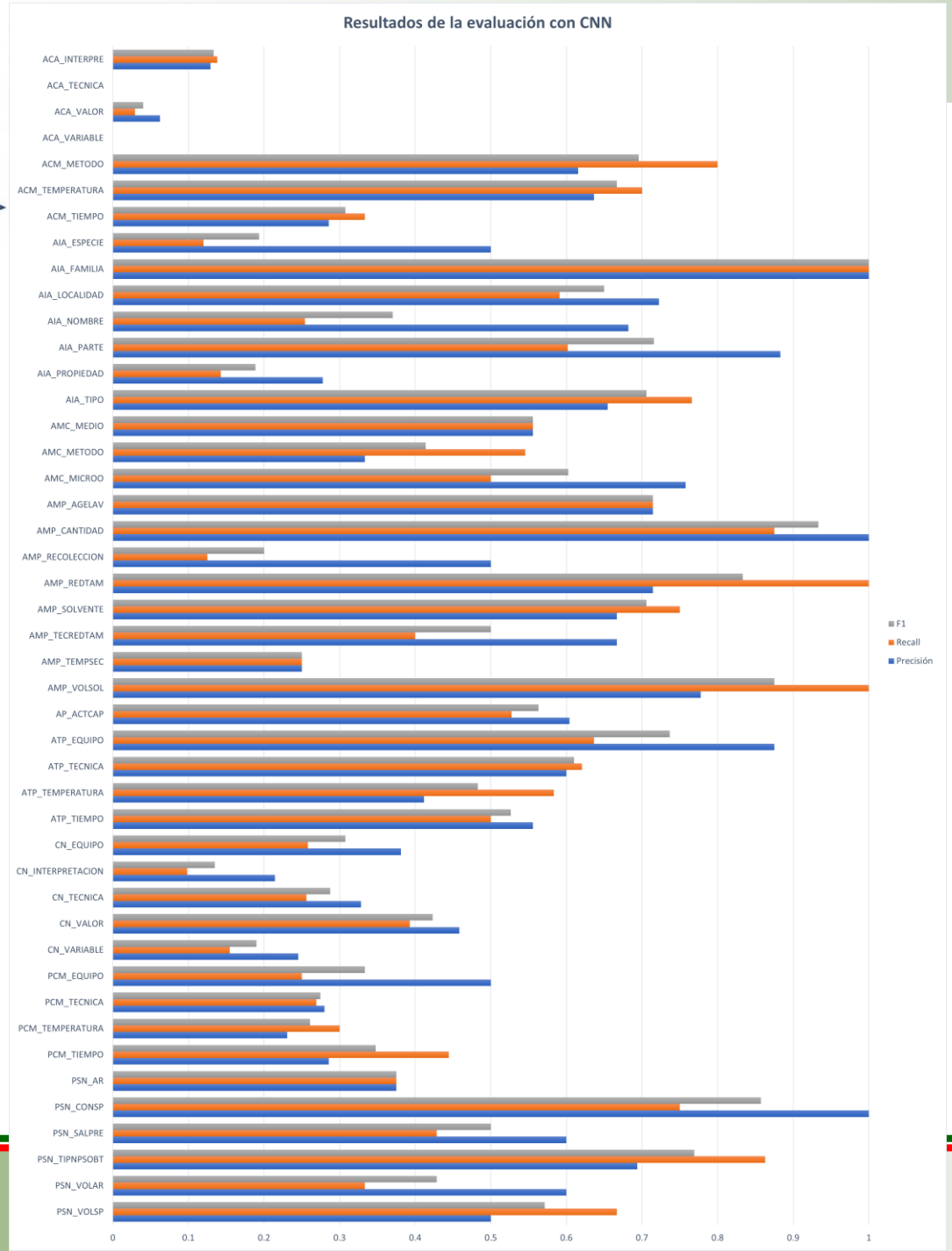
$$P = \frac{TP}{TP + FP}$$

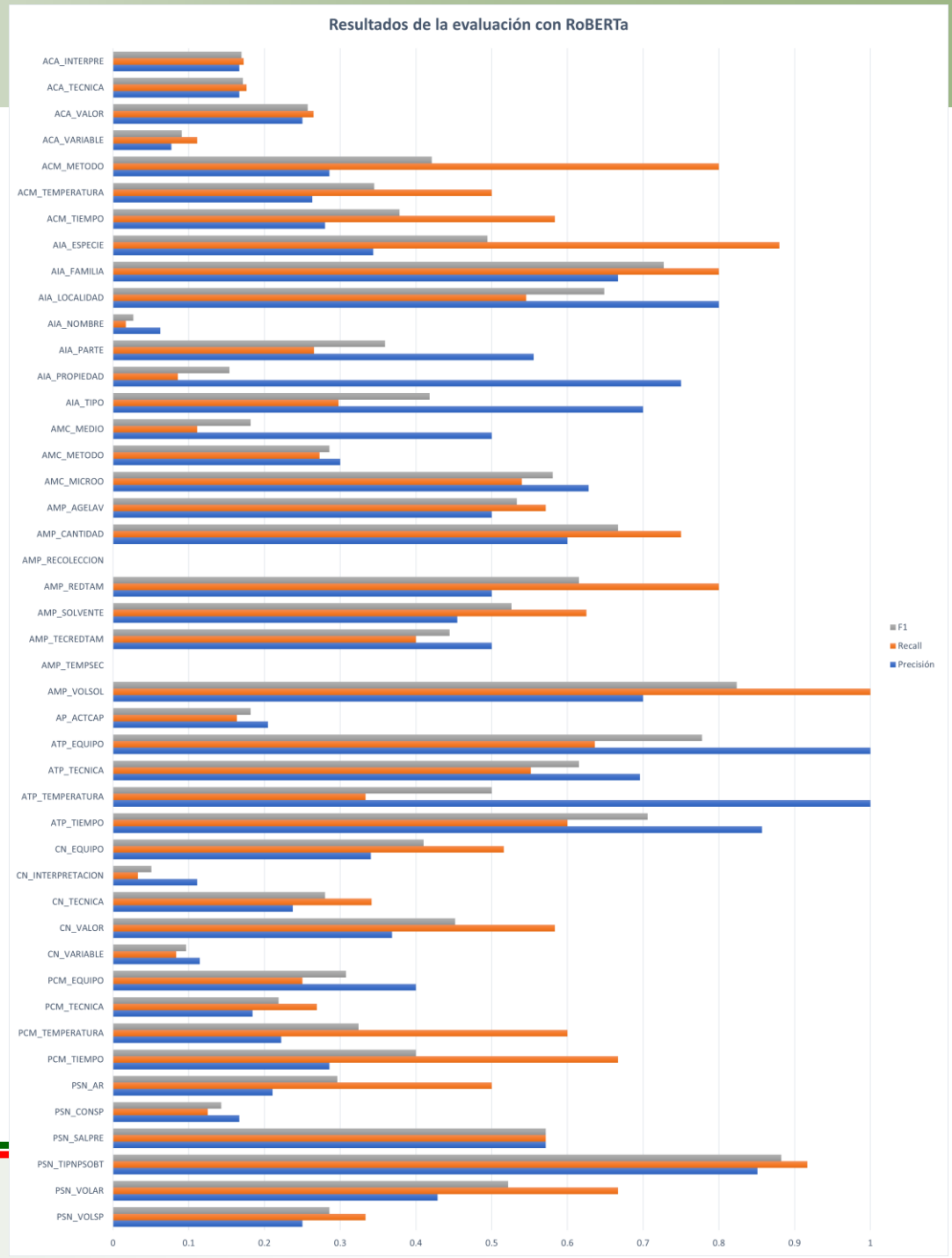
**Recall**

$$R = \frac{TP}{TP + FN}$$

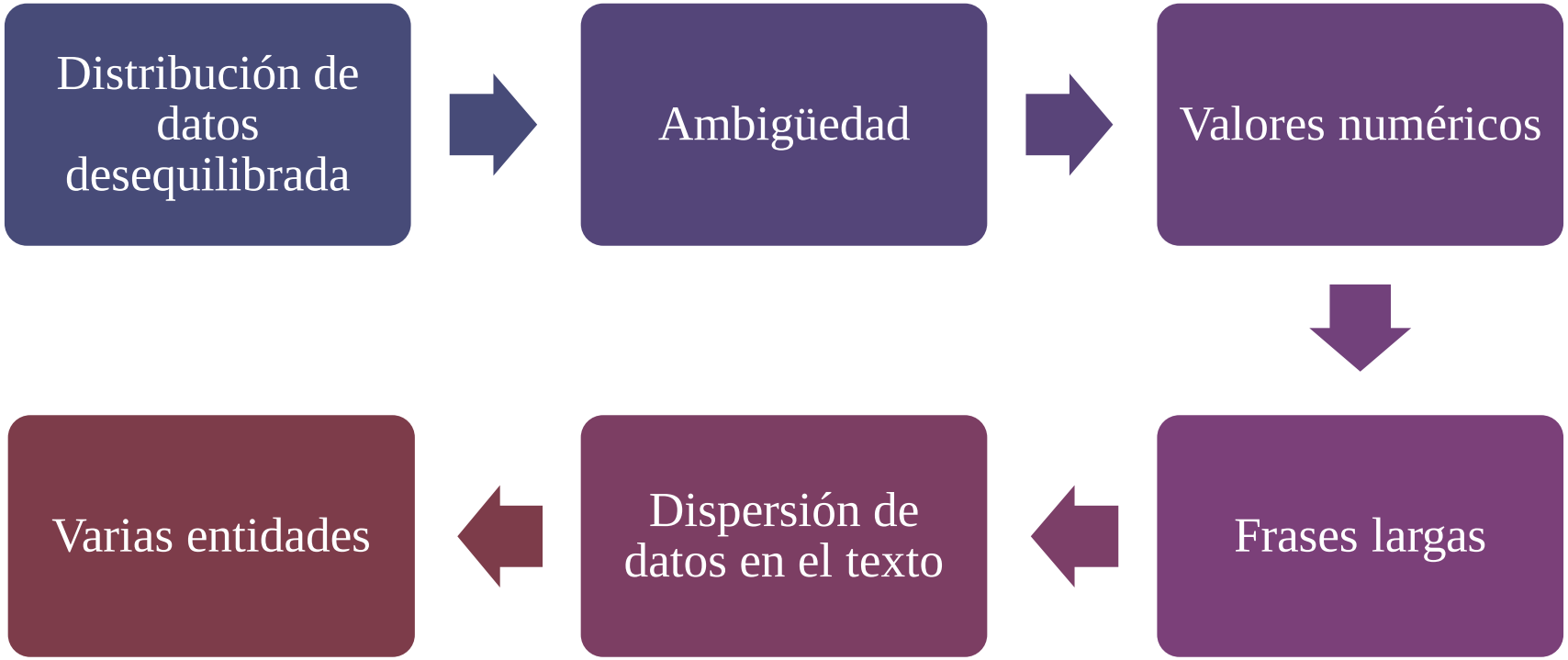
**Puntuación F1**

$$F1 = 2\frac{P \times R}{P + R}$$


Resultados de la evaluación con CNN
ACA_INTERPRE
ACA_TECNICA
ACA_VALOR
ACA_VARIABLE
ACM_METODO
ACM_TEMPERATURA
ACM_TIEMPO
AIA_ESPECIE
AIA_FAMILIA
AIA_LOCALIDAD
AIA_NOMBRE
AIA_PARTE
AIA_PROPIEDAD
AIA_TIPO
AMC_MEDIO
AMC_METODO
AMC_MICROO
AMP_AGELAV
AMP_CANTIDAD
AMP_RECOLECCION
AMP_REDTAM
AMP_SOLVENTE
AMP_TECREDTAM
AMP_TEMPSEC
AMP_VOLSOL
AP_ACTCAP
ATP_EQUIPO
ATP_TECNICA
ATP_TEMPERATURA
ATP_TIEMPO
CN_EQUIPO
CN_INTERPRETACION
CN_TECNICA
CN_VALOR
CN_VARIABLE
PCM_EQUIPO
PCM_TECNICA
PCM_TEMPERATURA
PCM_TIEMPO
PSN_AR
PSN_CONSP
PSN_SALPRE
PSN_TIPNPSOBT
PSN_VOLAR
PSN_VOLSP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
F1
Recall
Precisión


Resultados de la evaluación con RoBERTa
ACA_INTERPRE
ACA_TECNICA
ACA_VALOR
ACA_VARIABLE
ACM_METODO
ACM_TEMPERATURA
ACM_TIEMPO
AIA_ESPECIE
AIA_FAMILIA
AIA_LOCALIDAD
AIA_NOMBRE
AIA_PARTE
AIA_PROPIEDAD
AIA_TIPO
AMC_MEDIO
AMC_METODO
AMC_MICROO
AMP_AGELAV
AMP_CANTIDAD
AMP_RECOLECCION
AMP_REDTAM
AMP_SOLVENTE
AMP_TECREDTAM
AMP_TEMPSEC
AMP_VOLSOL
AP_ACTCAP
ATP_EQUIPO
ATP_TECNICA
ATP_TEMPERATURA
ATP_TIEMPO
CN_EQUIPO
CN_INTERPRETACION
CN_TECNICA
CN_VALOR
CN_VARIABLE
PCM_EQUIPO
PCM_TECNICA
PCM_TEMPERATURA
PCM_TIEMPO
PSN_AR
PSN_CONSP
PSN_SALPRE
PSN_TIPNPSOBT
PSN_VOLAR
PSN_VOLSP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
F1
Recall
Precisión

# *Análisis de resultados*



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

# VALIDACIÓN DEL MODELO

# *Caso de Estudio*

1. Reunión con el personal del área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE
2. Comparación entre los datos obtenidos con los modelos y la base de datos en Excel que contenía la extracción manual de los datos.

**Correcto:** Los modelos y la base de datos coinciden.

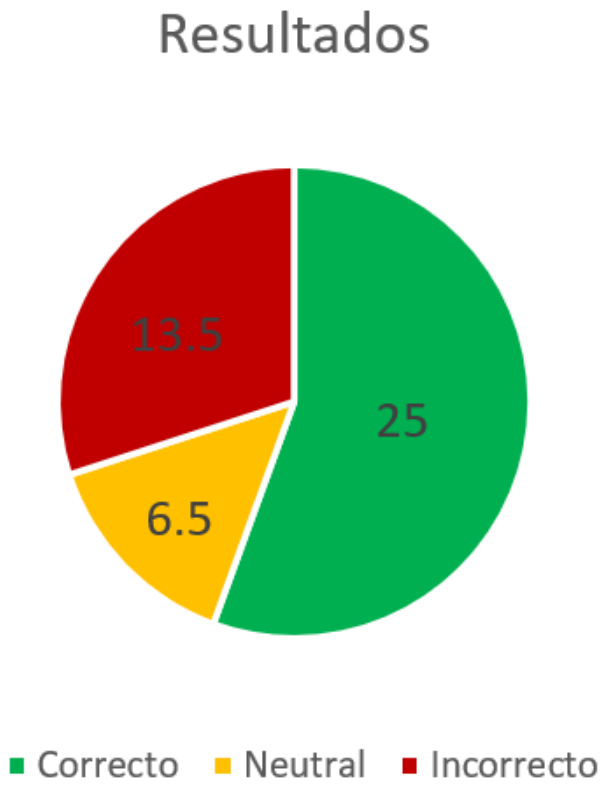
**Neutral:** Los modelos y la base de datos coinciden parcialmente, es decir, los modelos no lograron extraer por completo la información.

**Incorrecto:** Los modelos y la base de datos no coinciden.

| N | Etiquetas | X88 | X167 | X120 | X9 |
|---|---|---|---|---|---|
| 1 | AIA_TIPO | Correcto | Correcto | Correcto | Correcto |
| 2 | AIA_NOMBRE | Correcto | Correcto | Correcto | Correcto |
| 3 | AIA_ESPECIE | Incorrecto | Correcto | Incorrecto | Correcto |
| 4 | AIA_FAMILIA | Correcto | Incorrecto | Incorrecto | Correcto |
| 5 | AIA_PARTE | Incorrecto | Incorrecto | Correcto | Correcto |
| 6 | AIA_LOCALIDAD | Correcto | Correcto | Incorrecto | Correcto |
| 7 | AIA_PROPIEDAD | Incorrecto | Correcto | Incorrecto | Correcto |
| 8 | AMP_RECOLECCION | Correcto | Correcto | Incorrecto | Correcto |
| 9 | AMP_AGELAV | Incorrecto | Correcto | Incorrecto | Incorrecto |

# Caso de Estudio

| Términos | X88 | X167 | X120 | X9 |
|---|---|---|---|---|
| Correcto | 30 | 28 | 10 | 32 |
| Neutral | 3 | 8 | 8 | 7 |
| Incorrecto | 12 | 9 | 27 | 6 |



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

# CONCLUSIONES Y RECOMENDACIONES

# *Conclusiones*


Utilización de NER

Necesidad de contar con mayor cantidad de textos etiquetados

La solución se validó por expertos del área de Biotecnología

# *Recomendaciones*


Continuar con estudios sobre el tema de extracción de textos automatizada, específicamente para el caso de estudio de nanopartículas

Enfoque semi supervisado


Elaborar una interfaz gráfica que facilite su uso para los investigadores de Biotecnología


***Gracias por su atención***

E S P E
ESCUELA POLITÉCNICA DEL EJÉRCITO
CAMINO A LA EXCELENCIA