PUCE Virtual

CLASE

1

PROCESAMIENTO DEL LENGUAJE NATURAL

Bases del texto y del dataset en ciberseguridad

QS WORLD UNIVERSITY RANKINGS

Educación de calidad

INTRODUCCIÓN DE LA CLASE

GRADO / POSGRADO / TECNOLOGÍAS

Estudia a tu tiempo
son interrupciones

EDUCACIÓN EN LÍNEA DE CALIDAD

El análisis automático de texto en ciberseguridad parte de una comprensión fundamental: la mayor parte de la información crítica en un entorno de seguridad no está estructurada. Correos electrónicos, tickets de soporte, logs narrativos de incidentes, mensajes en foros internos o reportes técnicos contienen datos en formato libre, con variaciones lingüísticas, ruido, ambigüedad y estrategias persuasivas. A diferencia de los datos estructurados tradicionales (tablas, registros numéricos), estos textos requieren ser interpretados y transformados antes de poder ser procesados por un sistema computacional. El Procesamiento de Lenguaje Natural (PLN) proporciona los marcos conceptuales y técnicos necesarios para representar el lenguaje humano en una forma manipulable por algoritmos.

En ciberseguridad para detectar amenazas reales es necesario comprender qué es un corpus, qué constituye ruido textual, cómo funcionan los tokens o qué papel cumplen las stopwords. Además, los ataques digitales presentan características discursivas particulares que deben ser reconocidas como patrones lingüísticos recurrentes. Paralelamente, es necesario comprender cómo se estructura un dataset en un entorno SOC o CSIRT, esto permite transformar el texto crudo en información analizable.

Bases del texto y del dataset en ciberseguridad.

Naturaleza del texto no estructurado en ciberseguridad: correos, tickets, logs narrativos.

Texto no estructurado

Es todo contenido cuyo significado está en lenguaje natural y no viene organizado en campos fijos (como una tabla). A diferencia de un log estructurado (e.g. JSON con claves constantes), el contenido es variable, con estilos distintos, abreviaturas, errores, y fragmentos técnicos mezclados con lenguaje común. Esto incluye comunicaciones humanas y narrativas de incidentes.

Tipos de texto más comunes

En un entorno real de ciberseguridad, la información relevante para la detección de amenazas no se encuentra únicamente en registros estructurados, sino principalmente en cuatro tipos de textos no estructurados que circulan dentro de la organización: correos electrónicos, tickets de soporte, logs narrativos de incidentes y mensajes en chats o foros corporativos. Cada uno de estos formatos posee características discursivas, técnicas y contextuales distintas, pero todos comparten un rasgo común: contienen indicios lingüísticos y técnicos (Forma típica) que pueden revelar comportamientos maliciosos o situaciones de riesgo (Rasgos clave). Comprender la naturaleza de estos tipos de texto constituye el primer paso para poder representarlos computacionalmente y analizarlos mediante técnicas de PLN orientadas a la ciberseguridad.

Tipo de texto	Forma típica	Rasgos clave
Correos electrónicos (phishing / spear phishing)	asunto + cuerpo + firma + enlaces + adjuntos referenciados	persuasión (urgencia, autoridad) + solicitud de acción
Tickets de soporte (help desk / ITSM)	descripción del usuario + historial de conversación + notas técnicas	mezcla de síntomas (“no puedo entrar”) con contexto operativo
Logs narrativos de incidentes (informes escritos por analistas)	relato secuencial: “se detectó…”, “se aisló…”, “se confirmó…”	lenguaje semitécnico + cronología
Mensajes en chats/foros	Frases cortas, jerga, emojis, capturas	Significado contextual

Texto no estructurado es “difícil” para una máquina

Un sistema automático se enfrenta a problemas recurrentes:

- Variación léxica: “urgente”, “inmediato”, “ASAP”, “ahora”.
- Ambigüedad: “revisa esto” puede ser legítimo o malicioso.
- Ruido: firmas, disclaimers legales, cadenas de correo, HTML.
- Entidades mixtas: nombres, dominios, URLs, hashes, códigos, IPs.
- Intención oculta: el ataque puede estar “disfrazado” como solicitud normal.

  Esta complejidad explica por qué el primer reto del curso inicia con comprender el texto antes de modelar.

Componentes internos del texto o “capas”

En un texto de incidente, conviene separarlo mentalmente en capas:

- Contenido lingüístico (qué se dice): verbos de acción, tono, urgencia.
- Contenido técnico (qué referencia): URLs, dominios, IPs, adjuntos.
- Metadatos (si existen): fecha, remitente, canal, categoría, prioridad.
- Estructura discursiva: saludo → petición → justificación → llamada a acción.

Análisis de texto no estructurado

Demostrativamente se analiza un correo electrónico sospechoso (phishing simulado), y un ticket de soporte o legítimo (también simulado). Basado en (Jurafsky et al, 2023) y (Eiseinstein,2019). Este ejercicio permite:

- Comprender la naturaleza del texto no estructurado.
- Diferenciar lenguaje persuasivo vs. técnico.
- Preparar el terreno para limpieza y tokenización (Semana 2).
- Vincular lenguaje con riesgo operativo en un SOC.
-

Elemento	Caso 1 - Phishing	Caso 2 - Ticket legítimo
Ejemplo	Asunto: URGENTE: Verificación inmediata de su cuenta corporativa Estimado usuario, Hemos detectado actividad inusual en su cuenta. Para evitar la suspensión inmediata, confirme sus credenciales aquí:	Buen día, desde ayer no puedo conectarme a la VPN corporativa. Aparece el mensaje “Authentication failed”. Mi usuario es jramirez. ¿Podrían revisar si mi cuenta

	http://secure-verification-login.com Atentamente, Departamento de Seguridad TI	está bloqueada? Gracias.
Ruido textual	"Atentamente", firma genérica	Saludo simple
Señales lingüísticas	"URGENTE", "suspensión inmediata", apelación a miedo	Solicitud descriptiva, tono neutral
Entidades técnicas	URL sospechosa	Mensaje de error específico
Intención aparente	Obtener credenciales	Resolver problema técnico

Conceptos básicos de PLN: token, corpus, ruido, stopwords, n-gramas (Jurafsky et al, 2023).

A continuación se definen cada uno de estos conceptos y se los ejemplifica con el correo del Caso 1 mostrado en la Tabla 2.

Token

Es una unidad mínima de texto que el sistema puede procesar. Generalmente corresponde a palabras individuales. También pueden ser números, símbolos o URL.

En el texto original: "Hemos detectado actividad inusual en su cuenta".

Los tokens posibles son:

["Hemos", "detectado", "actividad", "inusual", "en", "su", "cuenta"]

Así el sistema deja de ver oraciones y comienza a ver listas de unidades manipulables. Este proceso se llama "tokenización". Para profundizar sobre tokens revise el capítulo 1 del PDF del artículo (Delso et al, 2024) en https://epsir.net/index.php/epsir/article/view/782/958.

Corpus

Es el conjunto organizado de textos que se utilizará para analizar o entrenar modelos. En ciberseguridad, un corpus puede estar compuesto por:

- 20.000 emails (legítimos y phishing).
- Cientos o miles de tickes de soporte
- Cientos o miles de logs narrativos.

El correo de ejemplo es parte de un corpus etiquetado como “sospechoso” o “phishing”.

Ruido

Es todo elemento que no aporta valor a la tarea específica. Depende del objetivo del análisis. Ejemplos:

- Analizar solo patrones lingüísticos.
- La información crítica en dominios sospechosos.
- Mayúsculas excesivas.
- Firmas automáticas.
- Encabezados repetidos.
- HTML.

Stopwords

Son palabras frecuentes que aportan poco significado semántico, por lo que suelen eliminarse. En el ejemplo, al eliminar los stopwords queda: [“detectado”, “actividad”, “inusual”, “cuenta”].

n-gramas

Es una secuencia de n tokens consecutivos.

- Unigramas (1 palabra): ["urgente", "verificación"]
- Biogramas (2 palabras): ["verificación inmediata", "actividad inusual"]. En phishing estos biogramas se pueden convertir en patrones detectables por el modelo.
- Trigramas (3 palabras): ["suspensión inmediata confirme"]

Características lingüísticas de amenazas: urgencia, suplantación, lenguaje persuasivo.

El objetivo del PLN en ciberseguridad es identificar la dimensión discursiva del ataque. Muchas amenazas textuales no se distinguen únicamente por contenido técnico, sino por estrategias discursivas. Los atacantes diseñan el lenguaje para activar respuestas automáticas en el usuario, explotando atajos cognitivos (ver Ilustración 1):

- **Miedo** → términos como “suspensión”, “actividad inusual”, “riesgo de pérdida” reducen el pensamiento analítico.
- **Obediencia / autoridad** → referencias a “Seguridad TI”, “Administración”, “Soporte” inducen cumplimiento.
- **Urgencia** → expresiones como “inmediato”, “último aviso”, “24 horas” presionan a actuar sin verificar.

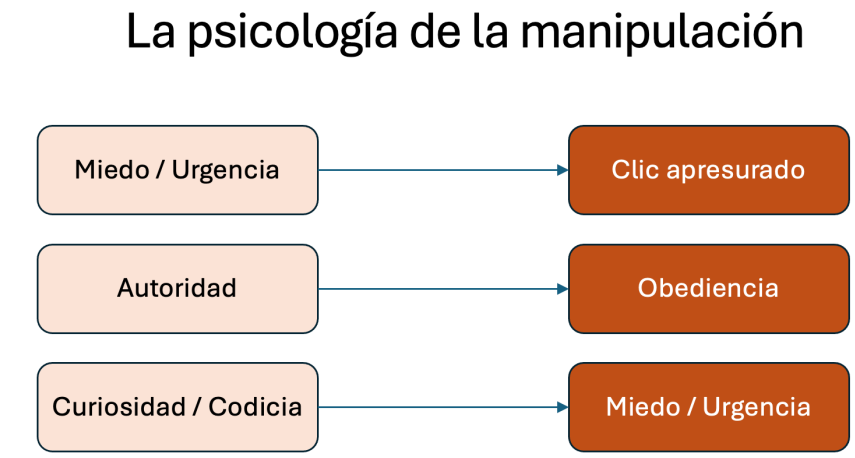

Ilustración 1. Manipular para la acción (creación de autor Rafael Melgarejo, Adaptado de Jurafsky,1023)

En PLN aplicado a ciberseguridad, estas estrategias se traducen en señales lingüísticas (imperativos, marcadores temporales, suplantación institucional) que pueden modelarse y detectarse: ya sea como tokens (urgente, verifique, [URL]), n-gramas ("suspensión inmediata"), o features semánticas (intención, presión emocional).

Estas estrategias pueden clasificarse en tres categorías:

- Urgencia artificial
- Suplantación de autoridad
- Persuasión emocional o coercitiva

Urgencia artificial

La urgencia busa reducir el tiempo de reflexión crítica del usuario, favoreciendo decisiones impulsivas. En informática, palabras asociadas a urgencia pueden convertirse en: features léxicas, n-gramas relevantes, indicadores semánticos.

En el ejemplo, los n-gramas asociados con urgencia son: "URGENTE", "suspensión inmediata", "verificación inmediata". Se puede observar:

- Uso de mayúsculas
- Adverbios de inmediatez
- Plazos implícitos o amenazas

Suplantación de autoridad

La suplantación se detecta lingüísticamente cuando: el dominio del enlace no coincide con la institución, la firma es genérica, y no hay personalización real del destinatario. Aquí se combinan análisis lingüístico + análisis técnico.

En el ejemplo "Departamento de seguridad de TI", combina estrategias de:

- Referencia a una entidad institucional
- Uso de términos corporativos formales
- Construcción de legitimidad aparente.

Lenguaje persuasivo y manipulación emocional

El verbo en imperativo (e.g. “confirme”) es clave. En PLN, los verbos de acción pueden convertirse en: Indicadores de intención, marcadores pragmáticos, señales de ingeniería social.

“Hemos detectado actividad inusual en su cuenta”, demuestra las siguientes estrategias:

- Introducción de un problema ambiguo
- Generación de incertidumbre
- Llamada directa a la acción (“confirme”).

Síntesis conceptual

Basado en Eisenstein (2019), la Tabla 1 sintetiza las estrategias respecto del objetivo del atacante, resaltando la señal lingüística y la característica PLN.

Tabla 1. Estrategias PLN frente a ataques (adaptado de Eisenstein 2019)

Estrategia	Objetivo del atacante	Señal lingüística	Feature PLN
Urgencia artificial	Evitar reflexión	Tiempo + imperativo	n-gramas temporales
Suplantación de autoridad	Generar obediencia	Léxico institucional	entidades nombradas
Persuasión emocional	Manipular decisión	carga emocional	análisis semántico

Comparación con texto legítimo: ejemplificación

En el ticket legítimo simulado “Desde ayer no puedo conectarme a la VPN corporativa”, se observa: tono descriptivo, ausencia de presión, e información verificable. Ilustración 1 muestra diferencias entre texto amenazante y legítimo.

Dimensión	Amenaza	Texto legítimo
Tono	Persuasivo / coercitivo	Descriptivo / colaborativo
Tiempo	Inmediato / urgente	Referencia temporal normal
Verbos	Imperativos	Narrativos
Emoción	Miedo / presión	Neutral
Técnica	Enlaces sospechosos	Errores verificables
Contexto	Ambiguo	Específico

Ilustración 2. Comparación entre amenaza y texto legítimo (adaptado de - Eisenstein, 2019)

Formalización para PLN

Estas características pueden convertirse en variables computables:

- Conteo de términos de urgencia
- Detección de imperativos
- Análisis de polaridad emocional
- Identificación de estructuras persuasivas

En etapas posteriores relacionadas con el Reto 3, estas señales pasarán de ser solo léxicas a ser semánticas y contextuales. Ilustración 3 conecta psicología humana, lenguaje y procesamiento computacional, miestra el puente conceptual entre Semana 1 (comprensión del lenguaje en ataques), Semana 2 (preparación del texto), Reto 3 (análisis semántico) y reto 4 (sistema inteligente de alerta).

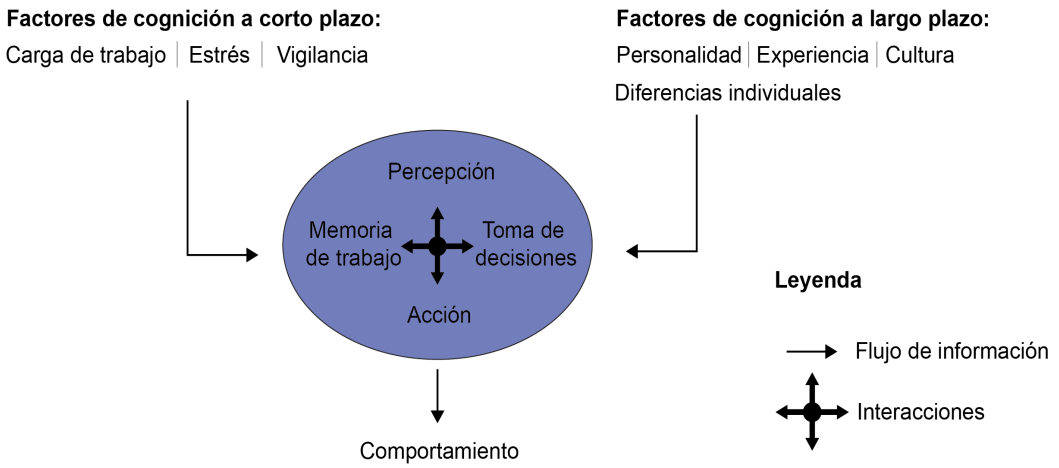

Ilustración 3. De Cognición Humana a PLN (creación de autor Rafael Melgarejo)

Fuentes de datos textuales en SOC/CSIRT.

Contextualmente:

- **SOC (Security Operations Center):** Unidad encargada de monitorear, detectar y responder a incidentes de seguridad en tiempo real.
- **CSIRT (Computer Security Incident Response Team):** Equipo especializado en la gestión, análisis y resolución de incidentes de seguridad.
- **CERT:** Equipo de respuesta de seguridad informática.

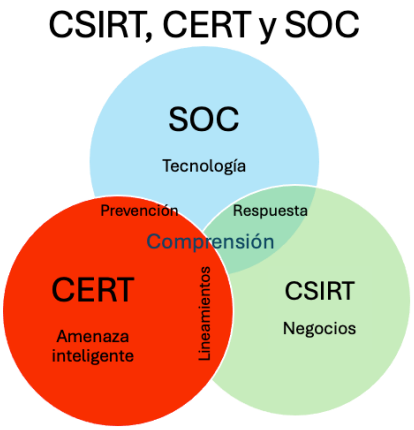

Ilustración 4. SOC - CERT – CSIRT (creación autor Rafael Melgarejo)

En ambos entornos, una parte significativa de la información crítica está contenida en texto no estructurado, que debe ser analizado para apoyar la toma de decisiones. El texto en ciberseguridad proviene de múltiples canales. Cada fuente tiene rasgos lingüísticos propios. La preparación del dataset depende de la fuente.

Principales fuentes textuales en un entorno real

En un SOC/CSIRT, los textos provienen de múltiples canales. Las más relevantes para PLN se encuentran en Tabla 2.

Tabla 2. Principales fuentes textuales (creación de autor Rafael Melgarejo, basado en Eisenstein,2019)

Canal	Ejemplo	Características
Correos electrónicos corporativos	Reportes de usuarios sobre mensajes sospechosos. Correos potencialmente maliciosos (phishing). Comunicaciones internas durante incidentes.	ingeniería social, robo de credenciales.
Tickets de soporte (plataformas ITSM)	Descripciones de fallas. Notificaciones de comportamiento inusual. Seguimiento de incidentes.	un ticket aparentemente técnico puede revelar una intrusión
Logs narrativos de incidentes	Reportes redactados por analistas. Descripción secuencial de eventos. Justificación de decisiones técnicas.	contienen conocimiento experto implícito.

Chats corporativos y foros internos	Conversaciones rápidas entre equipos. Alertas informales. Compartición de indicadores de compromiso (IOCs).	lenguaje breve, abreviado y altamente contextual.
Reportes de inteligencia de amenazas	Informes externos sobre nuevas campañas. Indicadores técnicos acompañados de narrativa explicativa.	combinan lenguaje técnico + análisis estratégico.

Diferencias estructurales entre fuentes

Cada fuente tiene características particulares (Ilustración 5). Estas diferencias afectan: El tipo de preprocesamiento necesario, la presencia de ruido, el tipo de patrones lingüísticos detectables. El análisis PLN se enfoca dependiendo las particularidades.

Fuente	Nivel de formalidad	Nivel técnico	Estructura	Análisis en PLN
Correo	Media–alta	Variable	Asunto + cuerpo	Detectar urgencia
Ticket	Media	Técnica	Campos + descripción	Detectar síntomas técnicos
Log narrativo	Alta	Alta	Secuencia cronológica	Identificar secuencias temporales
Chat	Baja	Variable	Conversacional	Contextual

Ilustración 5. Diferencias entre fuentes (adaptado de Eisenstein 2019)

Estructura del dataset: campos, metadatos y etiquetas.

El texto no estructurado debe convertirse en un dataset organizado para poder ser procesado por modelos de PLN. Esto implica transformar documentos individuales en registros estructurados que contengan: el texto, información contextual, una posible etiqueta de clase. Sin esta estructura, el modelo no puede entrenarse ni evaluarse adecuadamente.

Campos principales de un dataset

Para que un conjunto de datos pueda ser identificado y analizado debe contener mínimo:

- **id:** identificador único
- **Fuente:** tipo de texto (correo, ticket, log, chat)
- **Texto:** Contenido textual limpio o crudo
- **Fecha:** Fecha del evento
- **Canal:** Medio de recepción (email, sistema, ITSM, chat)

Metadatos

Aportan información valiosa para el analásis, como: Dominio del remitente, nivel de prioridad del ticket, departamento afectado, severidad asignada, usuario reportante. Estos elementos pueden ser variables adicionales en modelos posteriores.

Etiquetas (labels)

Las etiquetas son fundamentales cuando se trabaja con aprendizaje supervisado (Reto 2). Los tipos de etiquetas posibles son:

- phishing
- malware
- ingenieria_social
- legitimo
- incidente_tecnico

Buenas prácticas de estructuración

- Mantener formato consistente (CSV o JSON).
- Evitar celdas vacías innecesarias.
- Documentar criterios de etiquetado.
- Versionar el dataset.
- Registrar fecha de creación y modificaciones.

A continuación Ilustración 6 muestra JSON aplicado al correo del ejemplo, de esta manera la transición queda:

Texto no estructurado → Registro Estructurado → Dataset analizable

```
{
  "id": "001",
  "fuente": "correo",
  "texto": "Hemos detectado actividad inusual en su cuenta...",
  "fecha": "2025-05-10",
  "dominio_remitente": "secure-verification-login.com",
  "severidad": "Alta",
  "etiqueta": "phishing"
}
```

Ilustración 6. Estructuración JSON aplicado al email (creación de autor Rafael Melgarejo)

Herramientas para PLN

Rubrix es un framework de Python listo para producción que permite explorar, anotar y gestionar datos en proyectos de PLN. Rubrix es código abierto, compatible con las principales bibliotecas de PLN (Transformers, Hugging Face, spaCy, Stanford Stanza, entre

otros). Es recomendable usar Rubrix como apoyo automático al etiquetado manual. Descárguela en: https://rubrix.readthedocs.io/en/v0.4.1/.

Ilustración 7. Logo de Rubrix

Problemas éticos y de privacidad.

Las fuentes textuales pueden contener datos sensibles, por lo tanto se debe tomar precauciones. Los datos reales pueden ser: nombres, correos electrónicos, direcciones IP, datos sensibles que identifican personas o aspectos de la organización. Por ello, es obligatorio anonimizar, respetar las políticas institucionales, y documentar el tratamiento del dato.

Riesgos al trabajar con textos reales

- **Exposición de datos personales:** Un correo de phishing puede incluir nombres reales, cargos, teléfonos o direcciones internas.
- **Fuga de información sensible:** Logs narrativos pueden describir vulnerabilidades o configuraciones internas.
- **Reidentificación:** Incluso si se elimina el nombre, puede identificarse a una persona por contexto (cargo, departamento, incidente específico).
- **Uso indebido del corpus:** El dataset podría ser reutilizado fuera del propósito académico o sin autorización.

Anonimización

Consiste en eliminar o transformar datos personales para impedir la identificación de individuos, como en Ilustración 8.

Original	Anonimizado
Juan Pérez del departamento financiero confirmó sus credenciales.	[USUARIO] del [DEPARTAMENTO] confirmó sus credenciales.
jperez@empresa.com	[USUARIO]@[DOMINIO]
192.168.0.10	[IP_INTERNA]

Ilustración 8. Ejemplos de anonimización (creación de autor Rafael Melgarejo)

Minimización del dato

Implica que el dataset debe contener solo la información necesaria para el objetivo del análisis. Ejemplo: Para detectar phishing no es necesario almacenar el nombre completo del remitente, puede bastar con almacenar el dominio.

Documentación ética del dataset

Un dataset académico responsable incluye un archivo README o una ficha técnica del corpus, detallando:

- Fuente de los datos.
- Procedimiento de anonimización.
- Fecha de recopilación.
- Restricciones de uso.
- Responsable del tratamiento.

Educación de calidad

REFERENCIAS

Eisenstein, J. (2019). Introduction to natural language processing. MIT Press. https://cdn.jsdelivr.net/gh/it-ebooks-0/it-ebooks-2018-04to07/Natural%20Language%20Processing%20%28Jacob%20Eisenstein%29.pdf

Jurafsky, D., & Martin, J. H. (2023). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed., draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/.

SomosNLP. (2022). Etiquetado de datos para PLN con Rubrix, con Daniel Vila Suero - Hackathon de PLN en Español. https://www.youtube.com/live/UR6MHBbPA3g?si=Iy_zUrmD-9LZIEQay

Delso Vicente, A. T., Carvajal Camperos, M., & Corral De La Mata, D. Ángel. (2024). La evolución del procesamiento del lenguaje natural y su influencia en la inteligencia artificial: Una revisión y líneas de investigación futura. European Public & Social Innovation Review, 10, 1–23. https://doi.org/10.31637/epsir-2025-782

PUCE Virtual

síguenos en:

www.pucevirtual.puce.edu.ec