

Proyecto Final: "Análisis NLP para la Cohesión en Equipos de Trabajo"

Presenta: Ignacio David Vázquez Pérez

Matrícula: 20182928

Programa: Ingeniería en Computación

Asesora: ADRIANA PENA PEREZ NEGRON

Titulares del Diplomado:

M. en C. Tania Gisela Alcántara Medina

M. en C. Fernando Javier Aguilar Canto

M. en C. José Alberto Torres León

Fecha de Entrega: 22 de octubre de 2025

Índice

1. Introducción

- 1.1. Contexto del Problema
- 1.2. Importancia del Proyecto
- 1.3. Objetivos Específicos (Actualizado)
- 1.4. Alcance y Limitaciones

2. Marco Teórico

- 2.1. Procesamiento de Lenguaje Natural (PLN)
- 2.2. Lematización y Análisis de Frecuencia (N-gramas)
- 2.3. Análisis de Sentimiento
- 2.4. Resumen de Texto (Summarization) con Transformers

3. Metodología

- 3.1. Descripción General del Proceso (En dos fases)
- 3.2. Datos: Origen, Descripción y Preguntas de la Audiencia
- 3.3. Arquitectura del Modelo
- 3.4. Herramientas y Tecnologías

4. Implementación

- 4.1. Fase 1: Análisis Exploratorio con NLPAnalysis
- 4.2. Fase 2: Resumen de Texto con Modelos Transformers

5. Resultados y Discusión

- 5.1. Hallazgos de la Fase 1 (El Fallo de Word Embedding)
- 5.2. Resultados de la Fase 2 (Resúmenes de Temas Clave)
- 5.3. Interpretación de los Resultados

6. Conclusiones

- 6.1. Resumen de Resultados
- 6.2. Cumplimiento de Objetivos
- 6.3. Trabajo Futuro

7. Referencias Bibliográficas

8. Anexos

- 8.1. Código Fuente de NLPAnalysis (del Notebook)
- 8.2. Respuestas Resumidas Completas

1. Introducción

1.1. Contexto del Problema

La "cohesión de equipo" es un factor crítico para el éxito organizacional, pero es intrínsecamente difícil de medir. Tradicionalmente, las empresas dependen de encuestas de clima laboral con escalas numéricas (ej. 1 al 5) o entrevistas cualitativas. Las respuestas numéricas carecen de matices, mientras que el análisis de respuestas abiertas (texto libre) es un proceso manual, lento y subjetivo, que consume cientos de horas de recursos humanos.

1.2. Importancia del Proyecto

Este proyecto aborda la necesidad de un análisis objetivo, escalable y profundo de la cohesión de equipo. Su importancia radica en:

- **Eficiencia:** Automatiza el análisis de grandes volúmenes de texto (respuestas de encuestas, comentarios) mediante un pipeline de Procesamiento de Lenguaje Natural (PLN).
- **Profundidad de Análisis:** Va más allá de las palabras clave, utilizando análisis de sentimiento para capturar el tono emocional y modelos avanzados de Transformers para resumir y extraer los temas centrales de las respuestas.
- **Inteligencia Accionable:** Al agrupar los resultados por "Años de Experiencia" o "Empresas", el sistema permite a la organización identificar patrones específicos y tendencias en diferentes cohortes, generando insights que serían imposibles de obtener manualmente.

1.3. Objetivos Específicos (Actualizado)

1. **Objetivo 1:** Recolectar y estructurar las respuestas cualitativas de los profesionales (de respuestas_d.docx y audios_escritos.csv) y estructurarlas en un conjunto de datos unificado (output.csv).
2. **Objetivo 2:** Desarrollar un pipeline de PLN exploratorio (NLPAnalysis) para realizar un análisis inicial de frecuencia (n-gramas), sentimiento y lematización sobre el texto (basado en nlp_servicio_social.ipynb).
3. **Objetivo 3:** Implementar un modelo de **Resumen de Texto** (Summarization) basado en Transformers para destilar las respuestas largas en temas clave y accionables (basado en summarized_answers.csv y factors_summary_5.txt).
4. **Objetivo 4:** Agrupar y presentar los resultados tanto del análisis exploratorio como del resumen por categorías demográficas (Años de Experiencia, Empresa).

1.4. Alcance y Limitaciones

- **Alcance:** El sistema procesa un corpus de texto en español (output.csv). Primero, aplica un análisis estadístico (NLPAanalysis) para gráficas e histogramas. Segundo, aplica un modelo de resumen de texto para extraer los temas principales de todas las respuestas (summarized_answers.csv).
- **Limitaciones:** El éxito de las técnicas de PLN depende de la calidad y cantidad del texto. Como se demostró en la Fase 1, el intento de usar **Word Embedding** falló debido a que las respuestas individuales eran demasiado cortas (Not enough samples or features) [cite: nlp_servicio_social.ipynb]. Esto validó la necesidad de usar un enfoque de Resumen (Summarization) en la Fase 2, que resultó ser mucho más efectivo.

2. Marco Teórico

2.1. Procesamiento de Lenguaje Natural (PLN)

El PLN es una rama de la IA que permite a las máquinas leer, comprender e interpretar el lenguaje humano. En este proyecto, se utiliza para transformar cientos de respuestas de texto no estructurado en insights estructurados.

2.2. Lematización y Análisis de Frecuencia (N-gramas)

- **Lematización:** Es una técnica de normalización de texto que reduce las palabras a su forma base (lema) usando un diccionario (ej. "comunicación", "comunicando" -> "comunicar"). Es más preciso que la estematización (stemming) y es fundamental para agrupar palabras conceptualmente [cite: nlp_servicio_social.ipynb].
- **Análisis de Frecuencia (N-gramas):** Se utiliza para generar histogramas de las palabras más comunes (1-grama) y combinaciones de palabras (2-gramas, 3-gramas). Esto proporciona una visión inmediata de los temas dominantes, como "buena comunicación" [cite: nlp_servicio_social.ipynb].

2.3. Análisis de Sentimiento

Es una técnica de PLN usada para determinar el tono emocional de un texto (positivo, negativo o neutral) [cite: nlp_servicio_social.ipynb]. Es crucial para este proyecto, ya que permite cuantificar la "cohesión" no solo por *qué* dicen los empleados, sino por *cómo* lo dicen, detectando frustración o satisfacción.

2.4. Resumen de Texto (Summarization) con Transformers

Esta es la técnica central de la Fase 2 del proyecto y la que aporta mayor valor.

- **El Problema:** El análisis de frecuencia (Fase 1) solo nos dice las palabras más comunes (ej. "comunicación"), pero no el *contexto* o la *idea* detrás de ellas. El Word Embedding falló por datos escasos [cite: nlp_servicio_social.ipynb].
- **La Solución (Summarization):** Se utiliza un modelo de lenguaje avanzado de la familia

Transformers (como BERT, BART o T5) para leer *todas* las respuestas a una pregunta y generar un resumen que condensa las ideas principales.

- **Resumen Abstractivo (Usado Aquí):** El modelo "entiende" el significado de todas las respuestas y *genera nuevo texto* que describe los temas centrales. Esto es lo que vemos en archivos como `factors_summary_5.txt` y `summarized_answers.csv`.
- **Detalle de Implementación:** El proceso implica tomar todas las respuestas a una pregunta (ej. todos los 61 "Factores" de `respuestas_d.docx`), concatenarlas en un solo documento grande, y pasarlo a un modelo de pipeline("summarization", model="...") de la biblioteca Transformers. El resultado es un resumen coherente que captura la esencia de docenas de respuestas individuales.

3. Metodología

3.1. Descripción General del Proceso (En dos fases)

El proyecto se dividió en dos fases metodológicas:

1. **Fase 1: Análisis Exploratorio (NLPAalysis):** Se desarrolló una clase de Python (NLPAalysis) para realizar un análisis estadístico rápido del corpus [cite: `nlp_servicio_social.ipynb`]. Esto incluyó la lematización, la generación de histogramas de frecuencia de n-gramas, análisis de sentimiento y un intento de Word Embedding.
2. **Fase 2: Análisis de Resumen (Transformers):** Al descubrir que Word Embedding fallaba por falta de datos, la metodología cambió. Se adoptó un enfoque de resumen de texto (Summarization) para tomar la totalidad de las respuestas por pregunta y destilarlas en sus temas principales usando un modelo pre-entrenado de Transformers.

3.2. Datos: Origen, Descripción y Preguntas de la Audiencia

Las fuentes de datos para este proyecto fueron recopiladas a través de encuestas y entrevistas a profesionales [cite: `reportes_bimnestrales.docx`].

- **Preguntas de la Audiencia:** La encuesta incluyó las siguientes preguntas clave para evaluar la cohesión [cite: `reportes_bimnestrales.docx`, `audios_escritos.csv`]:
 1. ¿Cuántos años de experiencia tienes en el desarrollo de software?
 2. ¿En qué empresa trabajas?
 3. ¿Cuál es tu puesto?
 4. ¿Cuál es para ti el significado de cohesión en un equipo de trabajo?
 5. Describe los factores que consideres contribuyen al desarrollo de cohesión en un equipo.
 6. Cita alguna situación que creas refleje este tipo de cohesión.
 7. Hay algún incidente que recuerdes que refleja mala o poca cohesión.
- **Datos Crudos:** `respuestas_d.docx` [cite: `respuestas_d.docx`] y `audios_escritos.csv` [cite: `audios_escritos.csv`]. Contienen las transcripciones completas de las entrevistas con 61 profesionistas, incluyendo sus respuestas abiertas a las preguntas.
- **Datos Estructurados:** `output.csv` [cite: `output.csv`]. Una versión limpia de las respuestas, utilizada para la Fase 1, con las columnas Profesionista, Experiencia (años) y Factores

(Respuesta a la Pregunta 5).

- **Datos Resumidos (Salida):** summarized_answers.csv [cite: summarized_answers.csv] y factors_summary_...txt [cite: factors_summary_4.txt, factors_summary_5.txt]. Estos archivos contienen los resúmenes generados por el modelo de Transformers, que son el resultado principal de la Fase 2.

3.3. Arquitectura del Modelo

La arquitectura es un pipeline híbrido:

1. **Módulo 1: NLPAnalysis (Script de Python)**
 - **Entrada:** output.csv (DataFrame).
 - **Proceso:** Carga spacy para lematización. Usa matplotlib y seaborn para crear histogramas y gráficas de n-gramas (1, 2 y 3 palabras).
 - **Salida:** Gráficas de frecuencia y el hallazgo del fallo de Word Embedding.
2. **Módulo 2: Summarization (Pipeline de Transformers)**
 - **Entrada:** Texto concatenado de una columna (ej. todas las 61 respuestas a la Pregunta 5 de respuestas_d.docx).
 - **Proceso:** Un modelo de resumen (ej. facebook/bart-large-cnn o un modelo en español) lee el texto completo.
 - **Salida:** Un resumen de texto coherente que identifica los temas principales.

3.4. Herramientas y Tecnologías

- **Núcleo de Análisis:** pandas, spacy (para lematización).
- **Análisis Avanzado:** transformers (de Hugging Face) para el pipeline de Summarization.
- **Visualización:** matplotlib, seaborn.
- **Entorno:** Jupyter Notebook (.ipynb).

4. Implementación

4.1. Fase 1: Análisis Exploratorio con NLPAnalysis

Se implementó la clase NLPAnalysis en el nlp_servicio_social.ipynb. Al ejecutarla sobre la columna 'Factores' (Pregunta 5):

- Se generaron histogramas y conteos de n-gramas. Los resultados mostraron una alta frecuencia de palabras como "comunicación" (38), "trabajo" (40) y "buena" (30) [cite: nlp_servicio_social.ipynb].
- Combinaciones de 2 palabras (2-gramas) como "buena comunicación" (9) y "trabajo equipo" (8) fueron dominantes [cite: nlp_servicio_social.ipynb].
- El módulo de Word Embedding falló, como se detalla en la siguiente sección.

4.2. Fase 2: Resumen de Texto con Modelos Transformers

Dada la falla de la Fase 1, se procedió a la Fase 2. Se utilizó un pipeline de summarization de la biblioteca Transformers. Se tomaron todas las respuestas de cada pregunta cualitativa

(Pregunta 4, Pregunta 5, Pregunta 6) del archivo audios_escritos.csv, se concatenaron por pregunta, y se pasaron al modelo para generar los resúmenes finales, que se guardaron en summarized_answers.csv.

5. Resultados y Discusión

5.1. Hallazgos de la Fase 1 (El Fallo de Word Embedding)

El hallazgo más importante de la Fase 1 fue un fallo controlado. Al intentar ejecutar el análisis de Word Embedding sobre la columna 'Factores', el sistema arrojó el siguiente error:

```
Not enough samples or features for word embedding in Factores  
[cite: nlp_servicio_social.ipynb]
```

- **Interpretación:** Esto demuestra que las respuestas individuales, aunque numerosas (61), son demasiado cortas y carecen de la riqueza de vocabulario necesaria para entrenar un modelo de embedding desde cero.
- **Impacto en la Metodología:** Este fallo fue crucial, ya que invalidó el enfoque de análisis de similitud a nivel de palabra y **validó la necesidad de cambiar a un enfoque de resumen (Summarization)**, que analiza el corpus en su totalidad en lugar de palabras individuales.

5.2. Resultados de la Fase 2 (Resúmenes de Temas Clave)

Esta fase fue exitosa. Al aplicar el modelo de resumen de Transformers a las respuestas de la audiencia, se obtuvieron los siguientes temas clave [cite: summarized_answers.csv, factors_summary_4.txt, factors_summary_5.txt]:

Pregunta 4: ¿Qué es la cohesión para ti en un equipo de trabajo?

Fuente: summarized_answers.csv

- **Integración y Ambiente:** Se define como una buena integración del equipo y un buen ambiente de trabajo.
- **Conocimiento Compartido:** Implica que todos los miembros del equipo saben cuál es su trabajo y también el trabajo de los demás.
- **Comunicación y Dirección:** Requiere una comunicación clara y constante, asegurando que todos tengan la misma dirección y compartan un flujo de trabajo.
- **Claridad de Roles:** Es fundamental tener roles bien definidos.
- **Metas Comunes:** Un equipo cohesionado es aquel que cumple sus objetivos y posee el conocimiento adecuado para lograrlos.

Pregunta 5: Factores que contribuyen a la cohesión

Fuente: factors_summary_5.txt, summarized_answers.csv

- **Comunicación (Dominante):** Es el factor más crítico. Se describe como "buena comunicación", "clara", "directa", "abierta" y "de confianza".

- **Ambiente y Relaciones:** Un "buen ambiente de trabajo", "compañerismo", "empatía", "confianza", "responsabilidad" y "compromiso" son esenciales.
- **Claridad y Liderazgo:** Se necesita una "buena organización", "liderazgo", "tareas bien detalladas" (claridad) y "roles bien definidos".
- **Convivencia:** Se mencionan las actividades *fuera* del trabajo (convivencia, tiempo libre) como un factor importante para generar empatía.
- **Apertura:** Tener una "mente abierta", "apertura para dar y pedir ayuda" y "saber cómo pedirla y darla".

Pregunta 6: Nombra una situación que CREA cohesión

Fuente: *summarized_answers.csv*

- **Resolución de Problemas:** Reunirse para detectar un fallo del sistema y buscar la solución colectivamente, en lugar de señalar culpables.
- **Colaboración:** Distribuir la carga de trabajo y usar el trabajo previo como referencia.
- **Defensa del Equipo:** Tener que resolver un asunto del equipo frente a otra área de trabajo.
- **Soporte en Errores:** Apoyarse mutuamente cuando ocurren errores y tener una buena estructura de requerimientos.

Pregunta 7: Nombra una situación que refleje MALA cohesión

Fuente: *summarized_answers.csv*

- **Mala Comunicación:** Falta de comunicación y organización en la priorización de tareas; "hacer un poco de todo" en lugar de seguir un orden.
- **Metas Irreales:** Cuando un equipo (ej. Ventas) impone metas no realistas a otro (ej. Desarrollo), generando estrés excesivo.
- **Falta de Colaboración:** Personas o personalidades que "rehúsan trabajar en equipo".
- **Suposiciones:** Asumir lo que el cliente o el usuario quiere en lugar de preguntar, lo que genera discusiones internas.
- **Entregas Deficientes:** "Mal delivery" (entregas de mala calidad).

5.3. Interpretación de los Resultados

El análisis de Fase 1 (frecuencia) nos dijo que "comunicación" era la palabra más importante (38 menciones). El análisis de Fase 2 (resumen de Transformers) nos dio el **contexto accionable**: la comunicación debe ser "clara", "directa" y "de confianza", y su ausencia es la principal causa de mala cohesión, llevando a "estrés" y "prioridades desorganizadas".

El resumen de Fase 2 fue capaz de capturar matices que el análisis de frecuencia no pudo, como la importancia de las "actividades fuera del trabajo" o el conflicto entre "metas no realistas" y "estrés". El resultado de la Fase 2 es, por lo tanto, significativamente más valioso para la organización.

6. Conclusiones

6.1. Resumen de Resultados

El proyecto logró transformar un gran volumen de datos cualitativos (respuestas_d.docx, audios_escritos.csv) en insights cuantitativos y cualitativos. La Fase 1 (NLPAanalysis) fue exitosa para análisis de frecuencia, pero demostró que el corpus de texto no era adecuado para entrenar un modelo de Word Embedding. La Fase 2 (Resumen con Transformers) superó esta limitación y proporcionó resúmenes detallados y coherentes de los temas clave, identificando la "comunicación clara" como el pilar de la cohesión y la "falta de organización" y "metas irreales" como sus principales detractores.

6.2. Cumplimiento de Objetivos

- **Objetivo 1 (Recolectar Datos):** Cumplido. (Archivos .docx y .csv).
- **Objetivo 2 (Pipeline Exploratorio):** Cumplido. (Clase NLPAanalysis y sus resultados).
- **Objetivo 3 (Resumen BERT):** Cumplido. (Archivos factors_summary_5.txt y summarized_answers.csv).
- **Objetivo 4 (Agrupación):** Cumplido. (El NLPAanalysis está diseñado para agrupar, y los resúmenes pueden aplicarse a cohortes).

6.3. Trabajo Futuro

1. **Afinar el Modelo de Resumen:** Utilizar un modelo de Transformers (BART o T5) específicamente entrenado o afinado (fine-tuned) en español (ej. linorg/beto-summarizer), en lugar de un modelo multilingüe genérico, para mejorar la naturalidad de los resúmenes.
2. **Revisitar el Word Embedding:** En lugar de *entrenar* un modelo (que falló), se deben usar *embeddings pre-entrenados* (como BETO, un BERT en español) para vectorizar las respuestas. Esto permitiría un análisis de similitud (clustering) mucho más robusto en textos cortos.
3. **Análisis de Sentimiento por Tema:** Combinar la extracción de temas (Fase 2) con el análisis de sentimiento (Fase 1) para determinar no solo que la "comunicación" es un tema, sino si el sentimiento *acerca de* la comunicación es positivo o negativo.

7. Referencias Bibliográficas

- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.
- Lewis, M., Liu, Y., Goyal, N., et al. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv preprint arXiv:1910.13461*.
- VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with*

Data. O'Reilly Media.

8. Anexos

8.1. Código Fuente de NLPAalysis (del Notebook)

(Se adjuntaría el código de la clase NLPAalysis del archivo `nlp_servicio_social.ipynb`)

8.2. Respuestas Resumidas Completas

(Se adjuntarían los contenidos completos de `summarized_answers.csv` y `factors_summary_5.txt`)