Entrega 1 – Corpus, Idea de Proyecto y Textometría

Oscar Gabriel Corzo Vargas
Edwin Antonio Caro Jiménez
Santiago Zafra Rodríguez

1. Idea de proyecto

Área de aplicación / dominio:

El presente proyecto busca a través del monitoreo de comentarios en **redes sociales** extraer información sobre la **percepción del consumidor** en el **sector asegurador colombiano**.

Esto con el fin de **elaborar una idea general** para las compañías **de la opinión que tienen sobre su servicio** los actores del mercado y comprender, fuera de la información cotidiana que pueden captar a través de sus canales, la idea que tienen los consumidores en entornos donde pueden expresarse de forma más crítica y flexible.

En este sentido, la información que los usuarios publican en plataformas como Twitter, foros y blogs constituye una fuente valiosa para comprender cómo perciben a las aseguradoras y al sector en general.

Problema por abordar:

El problema central es **caracterizar la percepción del mercado hacia las principales aseguradoras** en Colombia. Se busca usar técnicas de Procesamiento de Lenguaje Natural (PLN) para analizar sentimientos, detectar temas recurrentes y extraer patrones de opinión en textos cortos (tweets, comentarios).

Justificación:

La relevancia de este proyecto radica en que las aseguradoras necesitan conocer la **opinión real y espontánea** de los consumidores para mejorar sus procesos de atención, ajustar productos y fortalecer la confianza del cliente. A diferencia de encuestas estructuradas, el análisis de redes sociales captura **percepciones inmediatas** y no solicitadas, permitiendo detectar problemas de servicio, expectativas y oportunidades de innovación.

Preguntas de investigación iniciales:

1. ¿Cuáles son los principales temas o atributos que los clientes mencionan al referirse a las aseguradoras? Por ejemplo, pagos, demoras, atención o cobertura.

Para identificar los atributos más relevantes en las conversaciones de los consumidores, se empleará un **análisis de frecuencias léxicas y de coocurrencias** en el corpus. El objetivo es detectar **qué términos emergen con mayor fuerza** en torno a las aseguradoras y cómo estos configuran **categorías significativas** como *demoras, pagos, atención, cobertura* o *siniestros*. De esta manera, el análisis busca abstraer los tópicos dominantes que estructuran la experiencia del usuario en la relación con su aseguradora.

Este procedimiento permitirá distinguir entre menciones recurrentes y vocabulario accesorio, **delimitando un conjunto de palabras clave que representan percepciones centrales.** Dichos atributos servirán posteriormente como **ejes de interpretación para comparar aseguradoras** y comprender en qué aspectos se concentra la atención de los clientes.

2. ¿La percepción expresada es mayoritariamente positiva, negativa o neutra hacia cada aseguradora?

La evaluación de la polaridad del discurso se plantea a través de técnicas de **análisis de sentimiento en español**, con el propósito de clasificar las menciones en **positivas**, **negativas** o **neutras**. Esta aproximación no se limita a cuantificar emociones aisladas, sino a caracterizar la orientación afectiva que acompaña las referencias a los servicios de cada aseguradora.

De esta manera, la metodología busca generar un **panorama agregado de la percepción del mercado,** expresado en proporciones de opiniones positivas, negativas y neutras. Este balance es crucial para **identificar fortalezas y debilidades reputacionales**, así como para anticipar riesgos en la relación con los clientes.

3. ¿Existen diferencias significativas entre las aseguradoras en cuanto a los tópicos más discutidos?

Para examinar las diferencias entre aseguradoras se propone **un análisis comparativo que cruce los tópicos identificados con la marca mencionada**. El propósito es establecer **qué atributos concentran mayor atención en cada empresa** y si existen **patrones diferenciadores en la conversación pública.**

El valor de esta comparación radica en poder caracterizar la percepción del mercado no solo de manera general, sino en relación con la identidad de cada aseguradora. Así, se posibilita un entendimiento más fino que evidencia, por ejemplo, si una compañía es más asociada a problemas de demora mientras otra lo es a la atención al cliente, aportando insumos estratégicos para la toma de decisiones.

---

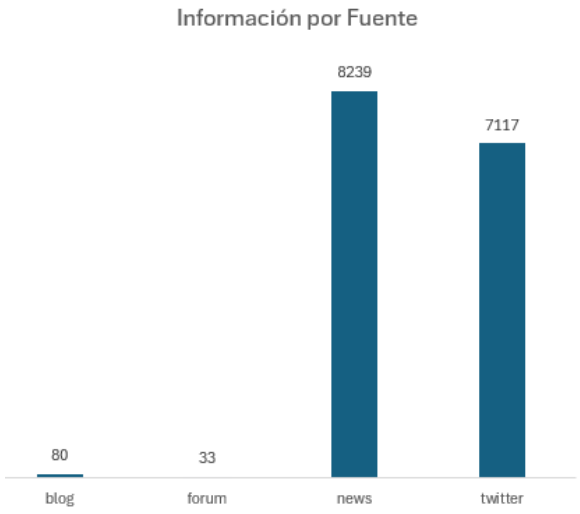
2. Corpus (descripción + evidencia)

Fuente del corpus:

El corpus proviene de una **exportación de Brandwatch**, que recopila publicaciones en redes sociales (principalmente Twitter). Se limitó a contenido en español y a menciones relacionadas con las aseguradoras líderes en Colombia (SURA, Seguros Bolívar, AXA Colpatria, Mapfre, Allianz, HDI, Positiva, Colmena, MetLife, La Previsora, Seguros Alfa).

Características del corpus:

- **Cantidad de documentos:** Se extrajeron 15.470 tweets, news, blogs y foros, generados durante el período de 2024-01-01 al 2025-08-28.
- **Número total de palabras:** 314.637
- **Número promedio de palabras por documento:** 20



Fuente elaboración propia

- **Idioma:** español.
- **Tipo de texto:** Inicialmente tweets, y noticias aunque se tiene en menor medida blogs y foros de internet.

Ejemplo de fragmentos:

- *"Si tienen la oportunidad de asegurar su vehículo con @SegurosBolivar*

PIERDANSELA, sobre todo si ud está ubicado en Medellín."

- *"Value and Risk mantuvo la calificación AAA (Triple A) al Riesgo de Contraparte de Fiduciaria La Previsora S.A*
- *Tengan cuidado con las llamadas masivas de @SegurosBolivar con una información de un seguro para los clientes de @Davivienda que se transforma inmediatamente en aceptación y sin sonrojarse empiezan a debitar de su cuenta el monto del tal seguro sin autorización. Fraude bancario?"*
- *"El colmo! @Davivienda cobrando un servicio que no autoricé de servicios Bolívar @SegurosBolivar dan un número que no atiende mi y nadie responde ya hay varios casas @SFCsupervisor"*

Formato de almacenamiento:

Actualmente el corpus se guarda en **Excel (.xlsx)** y, tras la limpieza, se exporta a **CSV** para su análisis en Python.

---

3. Textometría básica (exploración preliminar)

Limpieza básica aplicada:

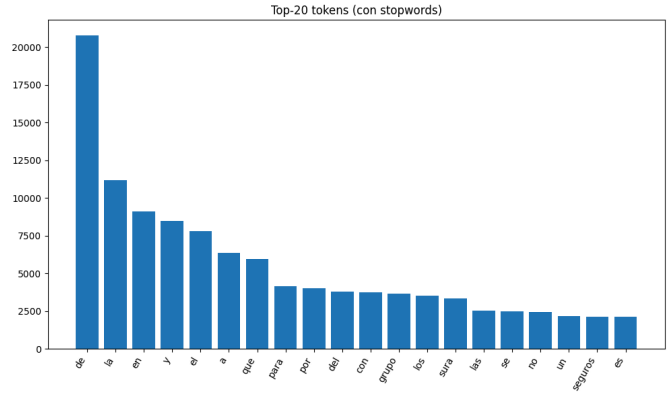
- Conversión a minúsculas.
- Eliminación de URLs, menciones (@usuario), hashtags y caracteres especiales.
- Uso de tokenizador específico para Twitter (TweetTokenizer), adaptado a lenguaje informal.
- Definición de *stopwords de dominio* para evitar que los nombres de aseguradoras dominen los conteos de palabras.

Medidas textuales:

- Número de documentos: X (según corpus filtrado en español).
- Número total de palabras: Y.
- Promedio de palabras por documento: Z.

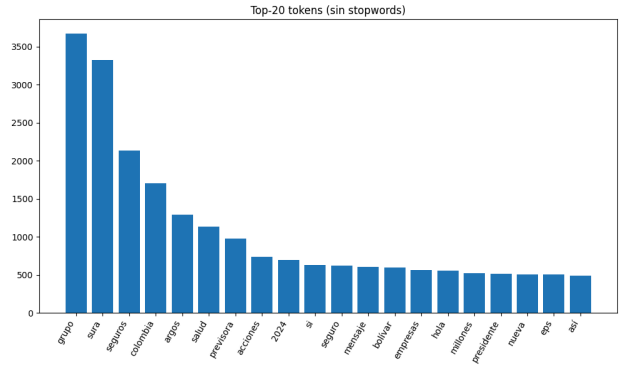
Frecuencias:

- **Top-20 palabras más frecuentes (con stopwords):** aparecen conectores comunes y nombres de aseguradoras.



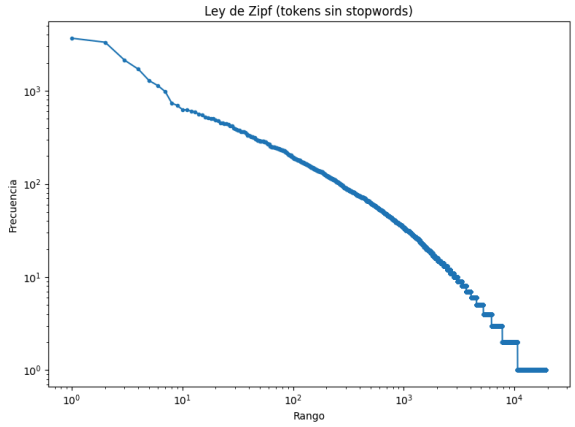
Fuente elaboración propia

- **Top-20 palabras más frecuentes (sin stopwords):**



Fuente elaboración propia

- Curva de Zipf (log-log), mostrando el patrón esperado de distribución de frecuencias léxicas.



Fuente elaboración propia

- Nubes de Palabras (Por Aseguradora)

Fuente elaboración propia

---

4. Próximos Pasos:

Depurar el foco del contenido

Eliminar menciones y metadatos no útiles para la percepción del servicio (p. ej., referencias a presidentes, confirmación o conformación de juntas directivas). El objetivo es concentrar el análisis exclusivamente en textos que describen la **experiencia de uso** (tiempos, siniestros, autorizaciones, atención, pagos, canales, etc.).

Extracción "pura y cruda" de percepciones

Ajustar la limpieza y tokenización para priorizar términos que capturen fricciones y atributos de servicio (demora, respuesta, trámite, cita, reembolso, etc.). Mantener dos vistas:

- **Con stopwords de dominio** removidas (para resaltar atributos).
- **Sin removidas** (cuando se requiera medir visibilidad/volumen de marca).

Categorización de sentimiento por tuit (positivo / negativo / neutral)

Implementar una primera aproximación de sentimiento en español (lexicón o modelo supervisado) y generar métricas por **aseguradora** y **tema**. Incluir matriz de confusión en una muestra etiquetada manualmente para validar.

Cobertura de marcas (validar faltantes)

Revisar si hay aseguradoras por incluir y ampliar el diccionario de alias/handles/hashtags en los patrones de detección de marca. Documentar la lista final y la fecha de corte.

Revisión y ampliación de la lista de stopwords

Auditar la lista para agregar términos que enmascaran el contenido pero no aportan semántica del servicio. Propuesta inicial para evaluar/incorporar (según contexto del corpus):

- **Entidad/branding genérico:** "grupo", "bolívar/bolivar", "aseguradora(s)", "compañía", "empresa".
- **Términos distractores frecuentes:** "usura" (si aparece como ruido contextual), "noticia", "prensa", "rt", "link", "www", "http/s".

  Mantener un control de versiones de stopwords y comparar top-términos antes vs. después para no perder señales relevantes.

5. Link GitHub:

[zafrar0926/Proyecto-PLN]