Percepción del consumidor en el sector asegurador colombiano

Caro Jimenez Edwin Antonio, Corzo Vargas Oscar Gabriel, Zafra Rodríguez Santiago

Resumen

*En la presente entrega se desarrolla un flujo completo de análisis de percepción sobre el corpus obtenido de los comentarios hacía las aseguradoras colombianas en Twitter, combinando técnicas de procesamiento de lenguaje natural (PLN) aprendidas en la asignatura y el uso de análisis supervisado y no supervisado, se limpia y no*rmaliza el corpus y *se identificaron patrones léxicos relevantes mediante n-gramas y MWEs* que no se lograron identificar en la primera entrega.

También se aplicó un pipeline que incluye TF-IDF con K-Means para descubrir temas recurrentes como Pagos/Siniestros, Atención/Trámite y Cobertura/Condiciones. Posteriormente, se implementaon dos enfoques de análisis de sentimiento: uno basado en modelos preentrenados (pysentimiento en español) y otro léxico-reglado (ML-SentiCon ajustado al dominio asegurador). Finalmente, se integraron los resultados para obtener una visión temática y emocional por marca donde se destaca el

I. INTRODUCCIÓN

Para abordar el análisis de la percepción de las aseguradoras se partió de la base de 15.469 tweets extraídos a través de la herramienta Brandwatch, que permite descargar información relativa a temas de interés general, en este caso, aseguradoras del sector colombiano, a partir de este corpus se propone integrar técnicas de PLN e identificar la percepción que el público tiene en la red social "X" mediante las diferentes metodologías y herramientas aprendidas en la asignatura.

II. CORRECCIONES Y MODIFICACIONES

En comparación con la primer entrega realizada en el proyecto se realizan ajustes asociados al pre procesamiento que se realiza sobre el corpus al considerar que los tweets tienden a ser textos cortos, para realizar una limpieza inicial de los tweets se realiza el siguiente pipeline:

1) Llevar todo el texto a minúsculas
2) Quitar URLs
3) Conservar hashtags sin el carácter #
4) Normalizar espacios
5) Eliminación de tildes y diacríticos

Una vez se realiza la limpieza inicial se filtran los tweets creados por las cuentas de las aseguradoras o de empresas del mismo holding empresarial ya que estos tweets no reflejan la imagen que tienen las personas, por el contrario se centran en temas informativos y pueden sesgar el ejercicio. Además, se calcula el TTR (Type-Token Ratio), esta medida calcula la diversidad del vocabulario de un texto, y el hapax ratio; el cual mide la proporción de palabras que aparecen una sola vez respecto al total de palabras distintas del texto. Por último, se analizan los n gramas para identificar cómo se mencionan las aseguradoras y las principales palabras asociadas a cada aseguradora, a continuación, un detalle de las actividades realizadas.

- Preparación de la data para clasificación de temas:

1. Se constituye un diccionario de palabras que se pueden asociar a cada una de las aseguradoras del mercado, se separaron las formas en las que se puede mencionar una aseguradora a través de un listado inicial de palabras de dominio para situar dentro del tweet las palabras que rodean estas menciones.

2. Se determinaron los bigramas y trigramas y se enriquecen los diccionarios de las aseguradoras, posteriormente, se realizaron varias iteraciones hasta lograr una buena clasificación de los tweets por marcas, la figura 1 y 2 contienen los principales bigramas y trigramas obtenidos respectivamente.

FORMAS DE MENCIÓN (BIGRAMAS con marca) - Top

ngram	freq	df
seguros bolivar	2342	1750
de grupo_sura	1902	1372
seguros sura	1115	729
del grupo_sura	861	622
grupo_sura y	843	660
axa colpatria	657	483
el grupo_sura	648	491
y grupo_sura	618	460
en grupo_sura	347	266
liberty seguros	327	261
grupo bolivar	314	263
sura y	303	241
de sura	293	236
eps sura	247	159
grupo_sura en	246	185

Fig. 1. Top bigramas obtenidos en el corpus (Fuente: Elaboración Propia).

Top 20 TRIGRAMAS (TEMAS, filtrado corporativo/CM):

ngram	freq	df
https t co	5835	2462
davidracero positivacol colombiacompra	291	146
positivacol colombiacompra wradiocolombia	290	145
dcoronell davidracero positivacol	210	105
aseguramiento integral salud	176	85
senor supersalud positivacol	112	57
primero semestre 2024	109	78
value and risk	103	52
nacional gestion riesgo	91	84
2024 patinaje velocidad	82	41
nuevo modelo salud	81	61
modelo sostenible aseguramiento	78	39
solventar caso manera	76	38
caso manera rapido	76	38
poder solventar caso	76	38
seguridad salud trabajo	75	46
manera rapido eficaz	74	37
dej alguno dato	74	37
atenderte manera directo	74	37

Fig. 2 Top trigramas obtenidos en el corpus (Fuente: Elaboración Propia).

3. Con las marcas identificadas apropiadamente, las palabras se transformaron en **stopwords de dominio** que permitían explorar los tweets dando más peso a

palabras importantes en su contenido, que no fuera la mención a las aseguradoras estudiadas.

- Argumentación de los cambios realizados

Considerando los comentarios realizados por el docente y las sugerencias realizadas por el docente asociadas a la naturaleza de los tweets y su corta longitud, se identificaron técnicas adecuadas de PLN.

En principio, no se conoce de forma explícita la aseguradora de la que se habla, el tema del que se hace mención, ni el sentimiento que engloba los comentarios, en adición a las transformaciones de limpieza de texto, se implementan N-Gramas para facilitar puntos de partida para entender cómo se mencionan las marcas y por otro lado, facilita excluir estas menciones de los tweets para permitir que los algoritmos posteriores: TF-IDF, K-Means, puedan operar mejor buscando patrones adecuadamente.

Finalmente, se etiquetan aquellas respuestas estándar de noticias corporativas (dadas por las aseguradoras) inferidas a partir de los N-Gramas y noticias de community manager, con el fin de centrarse más en temas de servicio.

III. ANÁLISIS EXPLORATORIO

1. **Identificación de temas generales abordados en el tweet.**
    a. Se crean vectores TF-IDF sobre los tweets limpios, sin marcas, stopwords, tokenizados y lematizados para identificar las palabras claves en los tweets.
    b. Con estos vectores, se ejecuta un K-Means que identifica tweets similares en K centroides, luego, se revisan las palabras más comunes en esos clúster y en función de eso se asocian nombres a esos clúster.
    c. Con los tweets etiquetados, se genera un algoritmo de clasificación supervisado (dado que ya se tienen las etiquetas) y se almacenan los parámetros de modelo para probar sobre tweets no etiquetados e identificar el tema del que tratan.

2. **Identificación de temas generales abordados en el tweet.**

Para este punto, se utiliza el lexicón ML-SentiCon, este es un *lexicón multilingüe de polaridades* desarrollado para tareas de análisis de sentimiento en distintos idiomas (incluido el español), cada palabra del lexicón viene en formato XML y puede incluir:

- **Lema:** Forma base de la palabra, por ejemplo “atender” en vez de “atiendo”.
- **Polaridad semántica:** Se determina si es positiva, neutra o negativa, a partir de una escala de -1 a 1.
- **Parseo cuidadoso del XML:** Se extraen y transforman todas las combinaciones de lema → polaridad.
- **Tokenización:** Se tokeniza de manera simple cada tweet convirtiendo el texto en una lista de palabras.
- **Relación token-lexicón:** Se busca cada token en el lexicon y se suma su polaridad.
- **Ajustes de puntuaciones**: Se revisan las puntuaciones a partir de palabras que nieguen o afirmen, por ejemplo:
    - *Negadores:* palabras como “no” o “nunca” invierten el signo de las palabras siguientes.
    - *Intensificadores:* Palabras como “muy” o “súper” multiplican la intensidad del sentimiento.
- **Score definitivo por tweet:** El resultado es un score total por tuit, por ejemplo:
    - +1.2 → Sentimiento positivo.
    - −0.8 → sentimiento negativo
    - 0.0 → neutro
    - Luego se etiqueta en lex_label como pos, neu, o neg.

3. **Diversidad léxica**

Es importante determinar el comportamiento de los textos y sus principales características, al revisar la distribución de la longitud de los comentarios se cumple el fenómeno evidenciado por el docente de la corta longitud, tenemos una media de 30 palabras por tweet y un 53% de los tweets que no superan las 35 palabras, esto es importante ya que implica mantener los stopwords y evitar perder contexto e información de cada comentario, la figura 3 contiene el histograma de longitudes de los tweets.

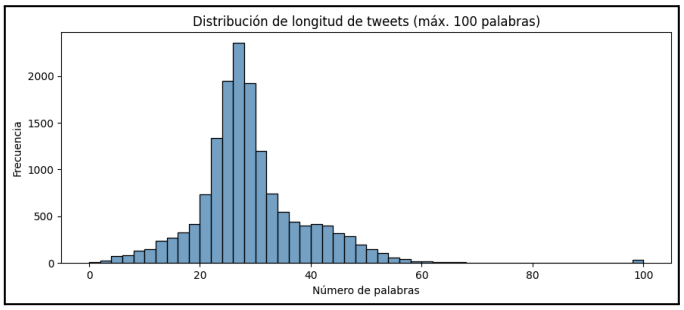

Fig. 3 Histograma de longitud de tweets(Fuente: Elaboración Propia).

Posteriormente se revisa la diversidad léxica de los tweets, usualmente en textos cortos se tiende a evidenciar tasas más altas, en este caso los tweets mezclan mensajes cortos y variados con mensajes repetitivos, la figura 4 muestra que el pico de distribución se centra en 0.5 y en en un rango de 0.8 - 1, estas dos distribuciones evidencian estilos altamente diferenciados, se cree este comportamiento se debe a palabras o expresiones como hashtags o emoticones.

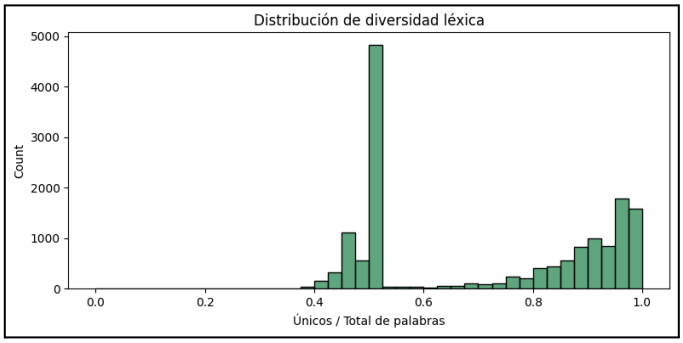

Fig. 4 Histograma de diversidad léxica (Fuente: Elaboración Propia).

Al revisar los términos más frecuentes se realiza la exclusión de los nombres de las aseguradoras con el propósito de evidenciar las palabras y tópicos que más se mencionan, se destacan palabras como **"junta directiva", "cementos argos", "empresas", "colombia", "año", "país", "negocio" y "millones"**, lo que sugiere que gran parte de los mensajes hacen referencia a **noticias corporativas, resultados financieros y movimientos empresariales**.

También aparecen expresiones como **"mensaje directo", "mensaje privado", "envíenos"**, que indican **respuestas automáticas o interacciones de atención al cliente**, aunque en menor proporción, en resumen conjunto, la imagen evidencia dos grandes núcleos temáticos que reflejan una mezcla entre publicaciones corporativas y reacciones del público hacia dichas entidades.:

1. **Contenido institucional o financiero** (empresas, juntas, resultados, economía).

2. **Comunicación y servicio al usuario** (mensajes, contacto, atención).

Fig. 5 Nube de palabras de los tweets (Fuente: Elaboración Propia).

Con el propósito de empezar a analizar de manera particular y revisar en que se centran los tweets asociados a cada aseguradora se revisa a qué aseguradora se dirige cada uno de los comentarios de los usuarios, la figura 6 evidencia que del total de los tweets se encuentra que el 90% se concentra en las aseguradoras "Seguros Bolivar", "Sura", "Fiduprevisora", "Positiva" y "Mapfre" con 35%, 19%, 14%, 12% y 10% respectivamente.

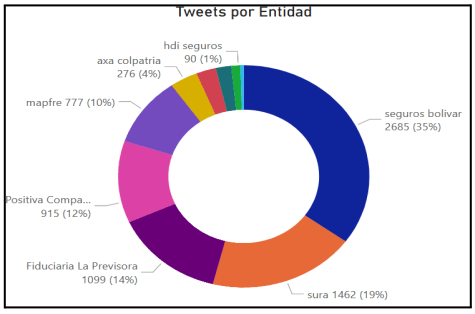

Fig. 6 Cantidad de tweets asociados a cada aseguradora (Fuente: Elaboración Propia).

Con base en esto se revisan las palabras que más se repiten cuando se mencionan cada una de las 5 aseguradoras mencionadas, y se evidencia lo siguiente.

- Seguros Bolívar: Los términos más frecuentes asociados a Seguros Bolívar evidencian una alta presencia de interacciones de atención al cliente, reflejada en palabras como "mensaje", "favor", "hola", "privado" y "responder". La fuerte coocurrencia con "Davivienda" sugiere una vinculación continua en las conversaciones digitales, posiblemente por productos o servicios compartidos. La presencia de "sfcsupervisor" y "sicsuper" indica también que los usuarios mencionan entidades de supervisión, lo cual podría relacionarse con reclamos o quejas sobre servicios financieros.

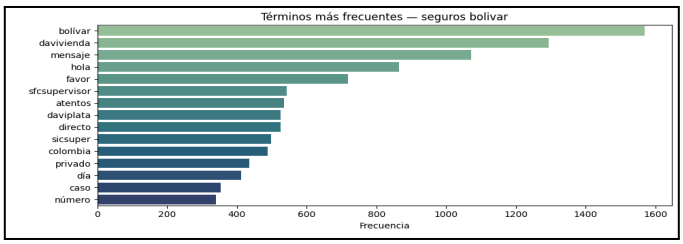

Fig. 7 Términos más frecuentes cuando se menciona a Seguros Bolívar (Fuente: Elaboración Propia).

- SURA: En el caso de SURA, los términos dominantes ("argos", "colombia", "gruposura", "salud", "empresas", "presidente", "gilinski") apuntan a un discurso corporativo y noticioso. Las menciones recurrentes a "acciones", "escisión" y "millones" sugieren que los usuarios comentan noticias financieras, movimientos accionariales o temas de mercado relacionados con la reestructuración del grupo empresarial. Esto refleja una conversación centrada en temas económicos y de gobernanza corporativa más que en servicio al cliente.

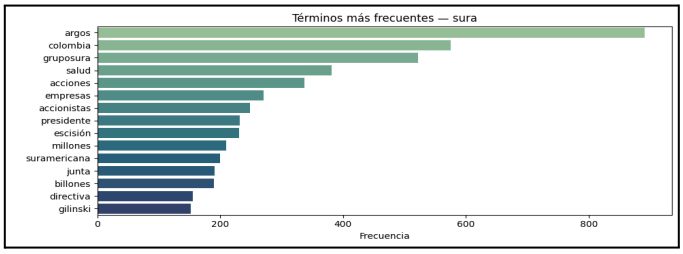

Fig. 8 Términos más frecuentes cuando se menciona a Sura (Fuente: Elaboración Propia).

- Fiduciaria La Previsora: Los términos más destacados (“salud”, “modelo”, “fiduciaria”, “ungrd”, “carrotanques”, “petrogustavo”, “aseguramiento”) muestran un fuerte enfoque en temas del sector público y de gestión estatal. Las menciones a “somosprevisora”, “fomag” y “nacional” sugieren que la conversación gira en torno a la administración de recursos, la cobertura de salud y la operación de programas del gobierno. Se evidencia un papel institucional, vinculado a contratos, proyectos y política pública.

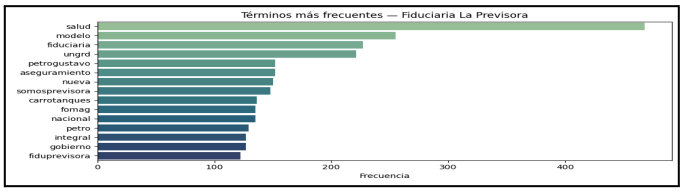

Fig. 9 Términos más frecuentes cuando se menciona a Fiduprevisora(Fuente: Elaboración Propia).

- Positiva Compañía de Seguros: En el caso de Positiva, las palabras más frecuentes (“positivacol”, “davidracero”, “dcoronell”, “wradiocolombia”, “supersalud”, “patinaje”) revelan una presencia mediática ligada a debates políticos y de opinión pública. La coocurrencia con figuras y medios nacionales sugiere que gran parte de la conversación proviene de la cobertura noticiosa y redes sociales más que de usuarios particulares. Se asocia, además, con temas de salud, deporte y gestión institucional.

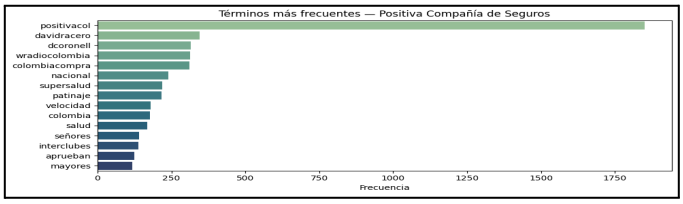

Fig. 10 Términos más frecuentes cuando se menciona a Positiva Compañía de Seguros (Fuente: Elaboración Propia).

- MAPFRE: Las menciones hacia MAPFRE se concentran en un discurso financiero y de negocio, con palabras como “millones”, “póliza”, “crecimiento”, “beneficio” y “riesgo”. La combinación de términos como “mercado”, “empresas” y “latinoamérica” apunta a un enfoque de comunicación más corporativo y regional, con baja aparición de expresiones ligadas a reclamos o atención al cliente. Esto sugiere que la percepción digital se relaciona más con su desempeño económico que con interacciones directas con usuarios.

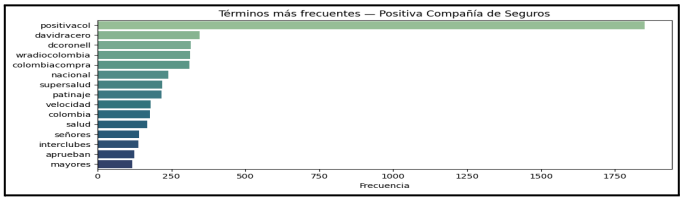

Fig. 11 Términos más frecuentes cuando se menciona a Mapfre (Fuente: Elaboración Propia).

IV. REPRESENTACIÓN DE LOS TEXTOS

A. Seleccionar y justificar las técnicas de representación más adecuadas para el proyecto (BoW, TF-IDF, Embeddings, etc)

Como se mencionó anteriormente, para poder crear una etiqueta sobre los temas de los que habla la opinión pública dentro de los tweets, se generan vectores **TF-IDF (Term Frequency–Inverse Document Frequency)** a partir del texto preprocesado de cada tweet (text_topics), con el objetivo de representar matemáticamente el contenido textual de forma que las palabras más relevantes de cada documento (tweet) tengan mayor peso que las palabras comunes, el proceso fue el siguiente:

1. **Corpus de entrada:** Se tomó el texto limpio y lematizado de los tweets (text_topics), excluyendo mensajes corporativos o respuestas de community manager (flag_corp_news y flag_cm_reply).
2. **Vectorización:** Se utilizó TfidfVectorizer de scikit-learn con los siguientes parámetros,
   - *ngram_range* = 1 y 2 para usar unigramas y bigramas.
   - *min_df* = 5 para ignorar términos que aparezcan menos de 5 veces.
   - *max_df* = 0.6 para ignorar términos que aparezcan en más del 60% de los documentos.
   - *max_feautures* = 50.000 para contemplar máximo 50.000 tokens.
   - *lowercase* = False ya que anteriormente se normalizó y transformó todo el texto a minúscula..
   - *token_pattern* = ‘(?u)\b\w+\b’ para traer palabras alfanuméricas.
   - *stopwords* = list(stop_es) para eliminar las palabras personalizadas.
3. **Stopwords personalizadas:** Se combinaron las stopwords españolas de NLTK con palabras neutras o no informativas del dominio, por ejemplo: {"https", "t", "co", "él", "la", "los", "las"}, esto ayuda a eliminar ruido y concentrar el modelo en términos con carga semántica.
4. **Representación resultante:** Cada documento (tweet) se representó como un vector esparso (sparse vector) en un espacio de **~9.000 a 10.000 dimensiones**. Cada dimensión representa un término o combinación de términos (unigrama/bigrama), y el valor de cada celda refleja la **importancia relativa del término** en ese documento con respecto al resto del corpus.
5. **Uso posterior:**
   Estos vectores TF-IDF fueron la entrada para el modelo **MiniBatchKMeans**, que agrupó los tweets en temas basados en la similitud coseno entre sus vectores.

B. Mostrar ejemplos de cómo los documentos fueron transformados en vectores o representaciones numéricas.

La figura 12 muestra una representación matricial de la transformación de los documentos en vectores numéricos mediante la técnica TF-IDF en el eje vertical se ubican los documentos del corpus y en el eje horizontal los términos más representativos identificados por el modelo, cada celda refleja el peso o importancia de un término dentro de un documento.

Valores más altos (en tonos más oscuros) indican que el término es relevante en ese texto y aparece con mayor frecuencia relativa, mientras que los valores bajos o cercanos a cero (en tonos claros) representan palabras poco frecuentes o ausentes esta visualización permite comprender cómo el modelo cuantifica la relevancia de las palabras y cómo cada documento adquiere una firma numérica única que luego se utiliza para tareas como la agrupación o clasificación temática.

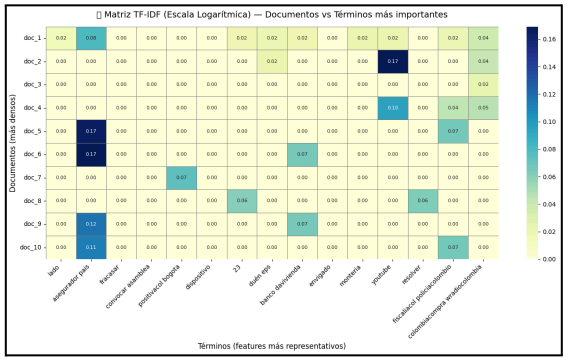

Fig.12 Matriz de TF-IDF de los documentos (Fuente: Elaboración Propia).

Para hacer la definición de temas sobre los que se habla en los tweets, se realiza un proceso de clusterización para asignar cada documento a un grupo tentativo, esto dará a entender cada marca a qué tipo de conversaciones se encuentra mayoritariamente relacionado, los grupos tentativos de establecieron mediante el siguiente proceso:

1. Remoción de stopwords y nombres de las aseguradoras para prevalecer el contenido semántico del tweet.
2. Creación de vectores TF-IDF a partir de los lemas creados anteriormente para representar cada tweet mediante los vectores.
3. Modelación de un mini batch K-means con diferentes cantidades de clusters para identificar las cercanías entre los tweets.
4. Se identifican las palabras más comunes dentro de cada cluster y con base en esto se generan las etiquetas.

Esto permitió clasificar los tweets dentro de siete categorías diferentes:

- Política - Economía - Medio ambiente
- SOAT - Fraudes - Carrotanques
- Respuestas automáticas
- Quejas - Reclamos - Oportunidades de mejora
- Empleo - Salarios - Oportunidades
- Salud - EPS - Retiro
- Movilidad - Pico y placa

La figura 13 muestra cómo se distribuyen los tweets dentro de estas categorías, el 58% se concentran en la etiqueta de Política - Economía - Medio ambiente, un objetivo de la siguiente entrega es seguir fortaleciendo el modelo e identificar las tendencias de los comentarios. Por último, la figura 14 muestra cómo se distribuyen los tweets por aseguradora clasificados dentro de estas categorías,

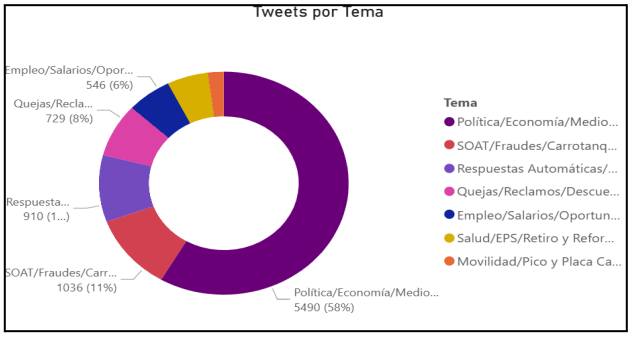

Fig.13 Distribución de tweets por tema (Fuente: Elaboración Propia).

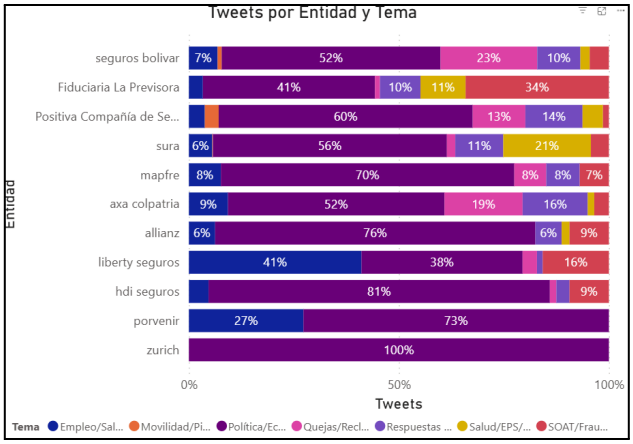

Fig.14 Distribución de tweets por tema según la asegura a la que se asocia el tweet (Fuente: Elaboración Propia).

V. PROPUESTA METODOLÓGICA DE LOS MODELOS

A. Definir qué tipo de enfoques de modelado se planea usar en la fase final (Clasificación Supervisada, Clustering, Topic Modeling, Análisis de Sentimiento)

En la fase final se implementarán enfoques combinados de Clustering, Clasificación Supervisada (En desarrollo) y Análisis de Sentimiento. Inicialmente, se aplicó Topic Modeling mediante técnicas de agrupamiento no supervisado (K-Means sobre representaciones TF-IDF) para identificar patrones temáticos en los textos y agrupar los documentos en función de su contenido semántico. Posteriormente, se desarrolló un modelo de Clasificación Supervisada basado en regresión logística, con el fin de asignar automáticamente nuevos textos a los temas previamente definidos. Finalmente, se complementa con un Análisis de Sentimiento apoyado en el léxico ML-SentiCon, para determinar la polaridad de las opiniones expresadas hacia las aseguradoras.

B. Justificar por qué esa metodología es adecuada para el objetivo

La metodología seleccionada permite una comprensión integral del corpus textual. El uso del Clustering y el Topic Modeling facilita la detección de tópicos emergentes sin necesidad de etiquetas previas, lo que es esencial para explorar grandes volúmenes de información no estructurada. La Clasificación Supervisada asegura consistencia y escalabilidad en la asignación temática, permitiendo automatizar la categorización de nuevos datos. Finalmente, el Análisis de Sentimiento complementa el enfoque al ofrecer una dimensión evaluativa de las percepciones de los usuarios, diferenciando entre comentarios positivos, negativos y neutros hacia cada marca. En conjunto, estos métodos brindan una visión tanto descriptiva como predictiva del comportamiento discursivo en redes sociales.

C. Especificar si se plantea comparar más de un modelo y bajo qué criterios

Nuestra propuesta busca comparar distintos modelos supervisados —entre ellos Regresión Logística, SVM y Random Forest— utilizando métricas de desempeño como precisión, recall, F1-score y matriz de confusión. Estos indicadores permitirán evaluar la capacidad de cada modelo para clasificar correctamente los textos según su tema o polaridad, seleccionando finalmente el que ofrezca el mejor equilibrio entre interpretabilidad y rendimiento.