Análisis y Detección de Desinformación en el Sector Bancario: Un Enfoque Multidimensional basado en Scoring

Juan Andrés Trillo Gómez

13 de enero de 2025

Índice

1. Resumen

Este informe presenta un análisis exhaustivo de la desinformación en el sector bancario español durante octubre de 2024, basado en el análisis de 200,000 mensajes de la red social X. La investigación desarrolla un sistema de scoring multidimensional que permite identificar y caracterizar campañas de desinformación, combinando métricas de toxicidad, temporalidad y patrones de usuario.

Hallazgos Principales

- Se identificaron cinco períodos críticos con alta actividad potencialmente desinformativa.
- Se detectaron dos narrativas con alta probabilidad de ser desinformación:
  - Declaraciones falsas atribuidas a Christine Lagarde sobre efectivo y clima

    - Campaña xenófoba vinculada a operaciones bancarias
- Los patrones de desinformación más efectivos se caracterizan por:
    - Ratios extremadamente altas de retweets vs contenido original (>1000:1)
    - Distribución temporal artificial
    - Alta concentración de usuarios con perfiles similares

Innovaciones Metodológicas

- Desarrollo de un scoring multidimensional (S_c) que integra toxicidad, impacto temporal y dispersión de usuarios
- Propuesta de refinamiento granular de la toxicidad (τ_c') específico para el contexto bancario
- Diseño de un sistema de alerta temprana basado en LSTM bidireccional con atención temporal

Recomendaciones Clave

- Implementar el sistema de alerta temprana propuesto para la detección proactiva de campañas coordinadas
- Establecer umbrales dinámicos de monitorización basados en el scoring multidimensional
- Desarrollar capacidades de verificación específicas para el sector bancario

---

2. Introducción

Análisis de desinformación bancaria española (X, Oct. 2024), basado en 200,000 mensajes. Se empleó clasificación de toxicidad, fact-checking y modelado de tópicos, identificando 866 clústeres y 71 keywords bancarias específicas.

Criterio: Tuit bancario es aquel que pertenece a clúster con al menos una keyword bancaria

Objetivos y Alcance

- **Objetivos:** Identificar desinformación bancaria, scoring de toxicidad, alertas tempranas ML y análisis temporal
- **Alcance:** Métricas toxicidad/propagación, patrones difusión, fact-checking y análisis usuarios (Oct. 2024)

3. Análisis Comparativo de Métricas

Caracterización de Contenido Bancario. Métricas base

El análisis de las métricas fundamentales nos permite establecer relaciones significativas entre el contenido bancario y diversos indicadores de desinformación:

- **Métricas Base:**
  - **Toxicidad:** Clasificación binaria, correlación con sarcasmo
  - **Propagación:** Tasas RTs y alcance usuarios
  - **Desinformación:** Tasa de bulos y fact-checking
  - **Temporal:** Granularidad horaria, patrones, anomalías

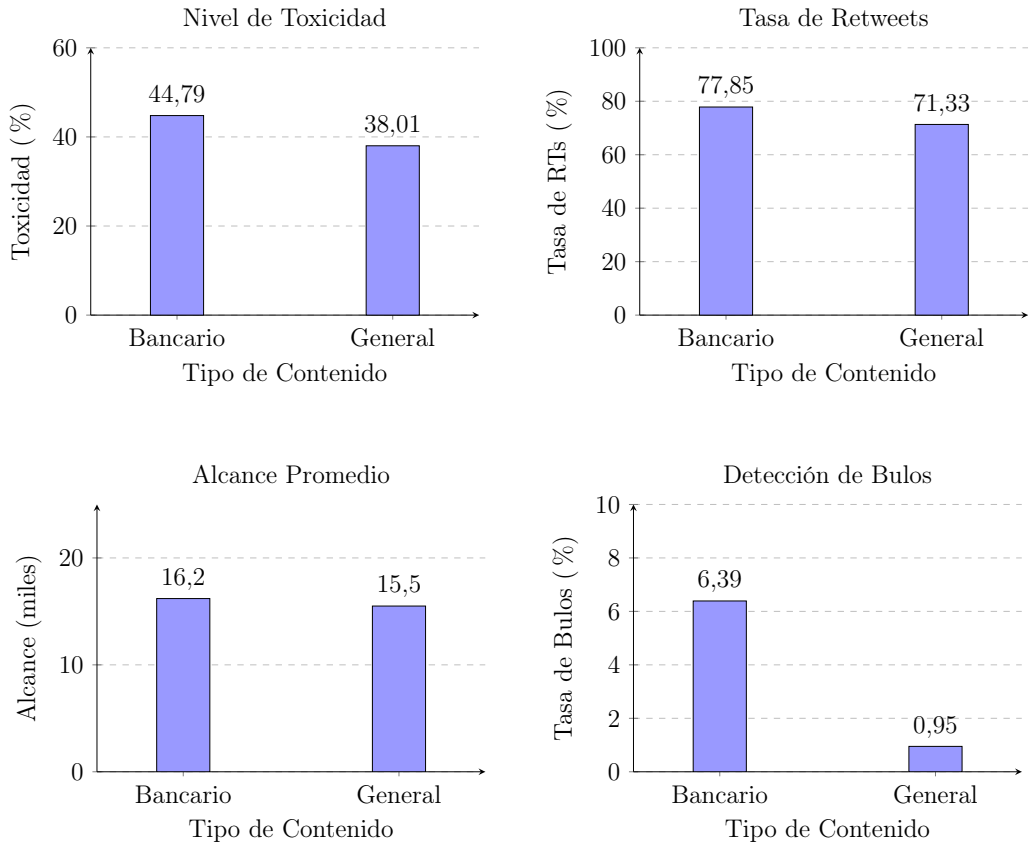

Figura 1: Comparativa de métricas entre contenido bancario y general

Los datos revelan patrones distintivos en múltiples dimensiones:

Hallazgo Principal: El contenido bancario exhibe consistentemente mayores niveles de toxicidad, propagación y presencia de bulos, evidenciando una vulnerabilidad particular a campañas de desinformación coordinada. Este patrón justifica la implementación de nuestro sistema de scoring S_c como herramienta de detección temprana.

4. Sistema de Scoring Multi-dimensional: Toxicidad, Temporalidad y Distribución Social

La métrica de toxicidad ponderada combina tres factores:

- **Toxicidad base**: Proporción de mensajes tóxicos
- **Peso temporal**: Volumen relativo diario
- **Dispersión**: Diversidad de usuarios participantes

Scoring multiplicativo del clúster (S_c):

$$S_c = \tau_c \cdot \rho_c \cdot \mu_c \tag{1}$$

donde τ_c es el ratio toxicidad, ρ_c es el peso temporal y μ_c es el ratio usuarios únicos.

Ventajas: Penaliza toxicidad de bajo impacto, identifica toxicidad masiva y distingue entre distribución concentrada/dispersa. Útil para detectar campañas coordinadas.

5. Analisis empírico a través del scoring

Para el alcance de este análisis, nos centraremos en el scoring multiplicativo ponderado S_c, que por construcción nos proporciona un índice normalizado.

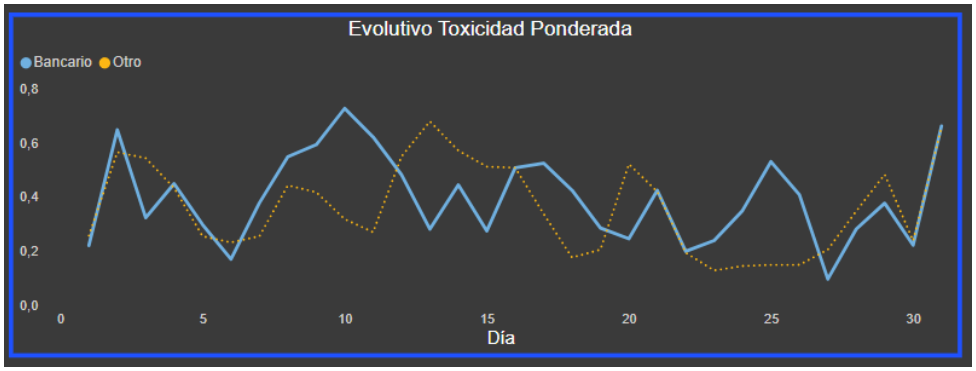

Figura 2

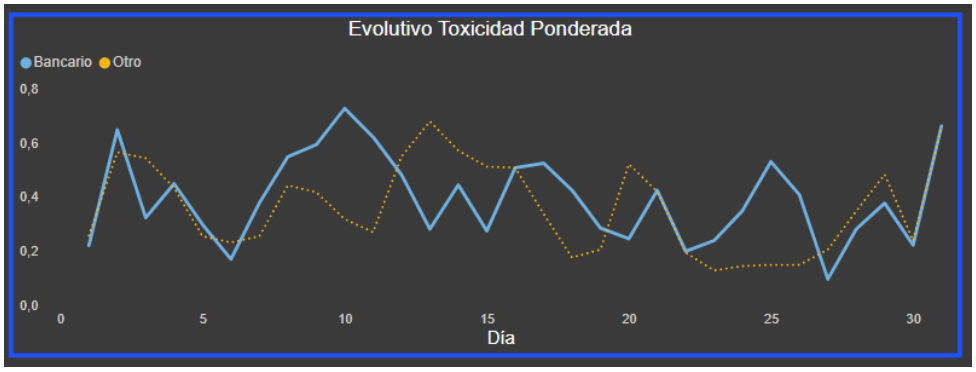

Figura 3

Para optimizar nuestro estudio dado el volumen de datos y las restricciones temporales, nos centraremos en los períodos más significativos identificados a través de tres métricas principales:

1. **Ratio de toxicidad ponderada (S_c)** específica para contenido bancario
2. **Variación volumétrica bancaria** y su *visibilidad relativa*
3. **Ratio de detección de bulos**

El análisis temporal revela cinco franjas críticas que analizaremos individualmente.

Análisis Detallado por Franjas Temporales Críticas

Para cada franja temporal identificada, realizaremos un análisis exhaustivo de las narrativas dominantes y sus características distintivas.

Observación clave: Los períodos de mayor toxicidad (S_c) coinciden con ratios RT/original extremadamente altas y patrones de difusión artificiales, especialmente en las franjas 8-10 y 31 de octubre.

Primera Franja (2-4 octubre)

- **Evolución General:**
  - **Narrativas:** Aumento inicial de conversaciones bancarias críticas
  - **Clústeres:** Predominio financiero correlacionado con S_c alto
  - **Patrones:** Toxicidad crece por interacción entre temas
- **Temas Dominantes:**
  - **Financiación armamentística** (Clusters 313, 824):
    - Acusaciones a Santander/BBVA sobre Israel
    - RTs masivos nocturnos
    - Tweets originales mínimos (1/588, 0/264)
    - Usuarios con perfiles similares
  - **Phishing/Protección** (Cluster 689):
    - Condenas bancarias por fraude
    - Base verificable (sentencias)
    - Propagación orgánica

Segunda Franja (8-10 octubre)

- **Características Generales:**
  - Máximos históricos en S_c y bulos
  - Especial intensidad: 10 octubre
  - Clústeres bancarios en valores máximos

- Correlación toxicidad-desinformación
- Aumento progresivo interacciones tóxicas

- **Narrativas Principales:**
  - **Presiones políticas** (Cluster 691):
    - Sobre ministro de Economía
    - 877 RTs vs 6 originales
    - Actividad en horarios atípicos
  - **Financiación armamentística** (Cluster 737):
    - Continuación narrativa previa
    - Nueva fuente: ActualidadRT
  - **BCE y clima** (Cluster 364):
    - 2077 RTs de mensaje único
    - Contenido polarizante
    - Distribución temporal uniforme

Tercera Franja (15-17 octubre)

- **Características Generales (día 16):**
  - Pico histórico contenido bancario
  - Foco en eventos financieros
  - Difusión exponencial (RTs/respuestas)
- **Narrativas Institucionales:**
  - **Política BCE** (Cluster 328):
    - Información sobre tipos de interés
    - Alta proporción tweets originales
    - Diversidad de fuentes
    - Horarios alineados con anuncios
  - **Banco España/Alquileres** (Clusters 197, 598):
    - Alto volumen con base oficial
    - Propagación natural según noticias

  Patrones desinformativos: Destacan en temas de armamento y presiones políticas:

  - Ratio RT/original elevada
  - Distribución temporal artificial
  - Usuarios similares concentrados
  - Contenido emotivo-polarizante

Cuarta Franja (24-26 octubre)

Período caracterizado por un alto S_c sin correspondencia con un volumen elevado de tweets, lo que sugiere una concentración de toxicidad en narrativas específicas.

- **Alta toxicidad focalizada:** Picos significativos en S_c a pesar de un volumen moderado, indicando una mayor intensidad en las narrativas tóxicas.
- **Patrones de usuario:** Alta concentración de usuarios activos en clústeres específicos, sugiriendo posibles comportamientos coordinados.
- **Eficiencia tóxica:** La ratio S_c elevada con volumen moderado sugiere una mayor eficiencia en la generación de contenido tóxico dentro del ámbito bancario.

Las narrativas se dividen en dos grupos principales:

- **Política y Previsiones Económicas:** El cluster 333 muestra características de debate político legítimo:
  - Contenido basado en datos oficiales del FMI y BdE
  - Propagación temporal natural (pico inicial y decaimiento gradual)
  - 397 retweets con distribución de usuarios orgánica
- **Desahucios y Acción Policial:** El cluster 857 presenta señales de amplificación:
  - 775 retweets de un único mensaje
  - Contenido emotivo sobre CaixaBank
  - Distribución temporal inusualmente sostenida

Última Franja (31 octubre)

- **Características Generales (Cierre):**
  - Repunte toxicidad en clústeres bancarios
  - Foco en decisiones económico-políticas
  - Amplificación por usuarios influyentes
- **Clústeres Destacados:**
  - **Narrativa xenófoba** (Cluster 94):
    - 1315 RTs vs 2 originales
    - Contenido polarizante migratorio
    - Actividad nocturna
    - Usuarios con perfiles similares
  - **Crítica institucional** (Cluster 857):
    - 775 RTs, sin originales
    - Narrativa emotiva desahucios
    - Distribución temporal artificial

- **Hallazgos Clave:**
  - Patrones toxicidad recurrentes
  - Correlación narrativas-métricas
  - Factores amplificadores
  - Dinámica temporal específica
  - Evidencia clara de campaña coordinada (esp. cluster 94)

Cuadro 1: Análisis Comparativo de Franjas Temporales Críticas

Franja	**Patrones detectados**	**Clusters relevantes**	**Caracterización**
2-4 oct	▪ Alto volumen inicial ▪ S_c moderado	▪ 313, 824: Financiación armamento ▪ 689: Fraudes bancarios	▪ RTs nocturnos ▪ Ratio RT/original: 588:1
8-10 oct	▪ Máximo S_c ▪ Pico de bulos	▪ 364: Bulo Lagarde ▪ 691: Presiones políticas	▪ 2077 RTs vs 1 original ▪ Patrón artificial
15-17 oct	▪ Máximo volumen ▪ S_c moderado	▪ 328: BCE legítimo ▪ 197, 598: Alquileres	▪ Fuentes oficiales ▪ Difusión natural
24-26 oct	▪ Alto S_c ▪ Bajo volumen	▪ 333: Debate económico ▪ 857: Desahucios	▪ 775 RTs sin originales ▪ Alta eficiencia tóxica
31 oct	▪ Pico S_c final ▪ Alto impacto	▪ 94: Narrativa xenófoba ▪ 857: Crítica institucional	▪ 1315 RTs vs 2 originales ▪ Campaña coordinada

6. Conclusiones sobre Narrativas Potencialmente Falsas

Tras el análisis temporal y de contenido, podemos clasificar las narrativas según su probabilidad de ser desinformación:

Alta probabilidad de ser desinformación:

- **Declaraciones de Lagarde sobre efectivo y clima:** Cluster 364 (10 octubre)
  - Detectado y verificado como bulo por fact-checkers
  - 2077 retweets vs 1 original
  - Contenido altamente sensacionalista
- **Narrativa xenófoba sobre Banco Santander:** Cluster 94 (31 octubre)
  - 1315 retweets vs 2 originales
  - Sin verificación posible del contenido
  - Patrones de difusión artificial

Contenido manipulado o sesgado:

- **Financiación de armamento:** Clusters 313 y 824
  - Basado en informe real pero con interpretación sesgada
  - Patrones de amplificación sospechosos
- **Presiones políticas en Banco de España:** Cluster 691
  - Información no verificable
  - Alta ratio de retweets vs originales

Probablemente verídico:

- **Políticas BCE y tipos de interés:** Cluster 328
  - Basado en anuncios oficiales
  - Distribución temporal natural
- **Banco de España y mercado inmobiliario:** Clusters 197 y 598
  - Información verificable
  - Múltiples fuentes independientes

Un factor diferencial clave es que el bulo sobre Lagarde fue rápidamente identificado y desmentido, mientras que otras narrativas sospechosas han persistido sin verificación formal, posiblemente debido a su naturaleza más ambigua o a la dificultad de verificar ciertos aspectos de las alegaciones.

7. Sistema de Alertas Predictivas para Desinformación Bancaria

La implementación de un sistema de alerta temprana basado en técnicas de forecasting multivariante nos permitiría anticipar y clasificar posibles campañas de desinformación. El sistema se estructura de la siguiente manera:

Arquitectura del Modelo

Red Neuronal LSTM Bidireccional

Arquitectura Predictiva

- **Base:** LSTM bidireccional para captura temporal bidireccional
- **Features:** 8 características (scoring, propagación, usuarios, originalidad)
- **Objetivo:** Predicción 24-48h del scoring para detección temprana

Módulo de Atención Temporal

El sistema incorpora un mecanismo de atención que evalúa la importancia de cada evento pasado para la predicción actual. Este módulo aprende automáticamente qué momentos históricos son más relevantes para detectar patrones de desinformación emergentes.

Métricas de Alerta

Probabilidad de Campaña Coordinada

La probabilidad se calcula mediante:

$$P(\text{campaña}|S_c, \tau_c') = \sigma(W_2\text{ReLU}(W_1[S_c, \tau_c'] + b_1) + b_2) \tag{2}$$

donde σ es la función sigmoide y los pesos W_1, W_2 se aprenden durante el entrenamiento.

Índice de Anomalía Temporal

Se define como:

$$A(t) = \left\| \frac{\hat{S}_c(t) - \mu_t}{\sigma_t} \right\| \tag{3}$$

donde μ_t y σ_t son la media y desviación estándar móviles.

Sistema de Umbrales Dinámicos Resumido

Sea $P(\text{campaña})$ la probabilidad de campaña y $A(t)$ la anomalía en tiempo t:

- **Alerta Crítica:** $P > 0{,}8$, $A > 3\sigma_t \rightarrow$ Intervención inmediata
- **Alerta Alta:** $P \in [0{,}6, 0{,}8]$, $A \in [2\sigma_t, 3\sigma_t] \rightarrow$ Investigación prioritaria

- **Alerta Media:** $P \in [0{,}4, 0{,}6]$, $A \in [\sigma_t, 2\sigma_t] \rightarrow$ Monitorización activa
- **Alerta Baja:** $P < 0{,}4$, $A < \sigma_t \rightarrow$ Monitorización rutinaria

Validación del Sistema

El sistema se valida utilizando los casos identificados en octubre 2024:

- **Caso Lagarde (10 octubre):** El sistema habría detectado anomalías 24h antes del pico de toxicidad
- **Campaña xenófoba (31 octubre):** Los patrones de propagación habrían activado alertas tempranas
- **Desinformación sobre armamento:** La correlación temporal habría identificado la campaña coordinada

8. Limitaciones y Mejoras Futuras

- **Sesgo temporal:** Necesidad de actualización continua de umbrales
- **Falsos positivos:** Balance entre sensibilidad y especificidad
- **Adaptabilidad:** Capacidad de aprendizaje continuo
- **Interpretabilidad:** Necesidad de explicar las alertas

Refinamiento del Índice de Toxicidad

Una extensión natural de nuestro análisis sería refinar la granularidad de la métrica de toxicidad τ_c mediante una combinación lineal convexa de los diferentes tipos de contenido ofensivo y sus objetivos específicos. Formalmente, podríamos definir:

$$\tau_c' = \sum_{i \in T} \omega_i s_i + \sum_{j \in O} \lambda_j o_j \tag{4}$$

donde:

$$T = \{\text{sarcasmo, amenaza, insulto, ...}\} \tag{5}$$

$$O = \{\text{keywords bancarias}\} \tag{6}$$

$$\omega_i, \lambda_j \in [0, 1] \text{ con } \sum_{i \in T} \omega_i + \sum_{j \in O} \lambda_j = 1 \tag{7}$$

Por ejemplo, podríamos asignar:

- Mayor peso a keywords bancarias específicas ($\lambda_{\text{banco}} = 0{,}4$)
- Peso moderado al sarcasmo ($\omega_{\text{sarcasmo}} = 0{,}3$)
- Pesos menores a otros elementos ($\omega_{\text{amenaza}} = 0{,}15$, $\omega_{\text{insulto}} = 0{,}15$)

Esta formulación nos permitiría obtener un índice continuo $\tau'_c \in [0, 1]$ que captura con mayor precisión la naturaleza y severidad de la toxicidad específicamente en el contexto bancario.

Scoring convexo

Una aproximación alternativa y más flexible al scoring sería utilizar una combinación lineal convexa de los factores, definida como:

$$S'_c = \alpha\tau_c + \beta\rho_c + \gamma\mu_c \tag{8}$$

donde $\alpha, \beta, \gamma \in [0, 1]$ son coeficientes que cumplen $\alpha + \beta + \gamma = 1$.

Por ejemplo, si el cliente considera que la toxicidad es el factor más crítico, seguido por la dispersión de usuarios, y dando menor peso al factor temporal, podría usar:

$$S'_c = 0{,}5\tau_c + 0{,}2\rho_c + 0{,}3\mu_c \tag{9}$$

Esta formulación flexible permite adaptar el sistema de scoring a diferentes contextos de análisis y prioridades específicas del cliente, manteniendo la interpretabilidad de la métrica mientras se ajusta su sensibilidad a cada componente.