Herramienta de Visual Analytics para Análisis Geoespacial de Sentimientos en Twitter

Autor Nombre
Afiliación Institucional
Ciudad, País
correo@institucion.edu

***Abstract*—**

I. Contexto, Motivación y Justificación

Las redes sociales, en particular Twitter, se han convertido en una de las principales fuentes de "datos de opinión pública" a escala global. Durante los últimos años, la comunidad académica ha desarrollado múltiples sistemas de *visual analytics* para explorar la evolución del sentimiento y las emociones en Twitter en relación con eventos globales (por ejemplo, pandemias, desastres naturales o crisis políticas). Entre ellos destacan [1], un sistema interactivo para detectar sentimientos sobre COVID-19; [2], que permite comparar emociones entre regiones y tiempos; y el conjunto de datos TSGI [3], que propone un índice geográfico diario de sentimiento como proxy de bienestar subjetivo.

II. Problema

Las emociones humanas en redes sociales varían entre regiones y momentos; sin embargo, *no existen suficientes herramientas interactivas que permitan comparar y visualizar estas diferencias de forma efectiva*.

III. Objetivos

A. Objetivo General

Desarrollar una *plataforma de visual analytics interactiva* que permita comparar, a nivel diario, los valores agregados de sentimiento extraídos de Twitter para distintas regiones geográficas (país, provincias/departamentos, ciudades/condados), de modo que investigadores y responsables de políticas puedan identificar patrones, correlaciones y anomalías en el ánimo colectivo de la población a lo largo del tiempo.

IV. Descripción del Dataset

A. Contexto

El *Twitter Sentiment Geographical Index Dataset (TSGI)* se basa en el *Geotweet Archive v2.0* del Harvard Center for Geographic Analysis, que agrupa miles de millones de tuits con metadatos de geolocalización (coordenadas o *bounding box* de lugar). TSGI focaliza la recolección a partir de 2019 y clasifica cada tuit con un modelo BERT multilingüe, asignándole una probabilidad de "sentimiento positivo". Posteriormente, se agregan diariamente esas probabilidades por las unidades administrativas definidas en GADM v3.6 (país, estado/provincia, condado/ciudad), generando índices diarios que sirven como proxy de bienestar subjetivo.

B. Objeto o Entidad de Estudio

El objeto de estudio son los *índices diarios de sentimiento* de Twitter en distintas divisiones administrativas georreferenciadas. Cada registro representa la combinación de:

- Una **fecha** específica (variable temporal).
- Una **unidad administrativa** en tres niveles (país, provincia/departamento, ciudad/condado).
- Dos métricas agregadas calculadas a partir de todos los tuits georreferenciados que caen dentro de esa unidad el día en cuestión:
  - `SCORE`: Valor promedio de probabilidad de sentimiento positivo (float $\in [0, 1]$).
  - `N`: Número de tuits georreferenciados usados para el cálculo (entero ≥ 0).

C. Atributos del Dataset

A continuación se describen los principales atributos del archivo CSV agregado por año (por ejemplo, `num_posts_and_sentiment_summary_2023.csv`):

- **DATE**
  - Descripción: Fecha del índice.
  - Tipo: `YYYY-MM-DD` (cadena de caracteres, *string*).
  - Rango de valores: Desde `2019-01-01` hasta la fecha más reciente disponible.
- **NAME_0**
  - Descripción: Nombre del país (nivel ADMIN 0).
  - Tipo: Cadena de caracteres (string).
  - Ejemplos: "Peru", "United States", "Spain".
- **NAME_1**
  - Descripción: Nombre de la subdivisión de nivel 1 (estado/provincia/región según GADM).
  - Tipo: Cadena de caracteres (string).
  - Ejemplos (Perú): "Lima", "Cusco", "Arequipa".
- **NAME_2**
  - Descripción: Nombre de la subdivisión de nivel 2 (ciudad/condado/provincia secundaria).
  - Tipo: Cadena de caracteres (string).
  - Ejemplos (Perú): "Lima Province", "Cusco Province", "Arequipa Province".
  - Notas: Puede estar vacía si no existen datos georreferenciados a nivel 2 en esa fecha/ubicación.
- **SCORE**

  - Descripción: Valor promedio diario de probabilidad de "sentimiento positivo" para todos los tuits georreferenciados dentro de la unidad administrativa y fecha especificadas.
  - Tipo: Número de punto flotante (float).
  - Rango de valores: $[0.0, 1.0]$ (valores más cercanos a 1 indican mayor proporción de tuits con sentimiento positivo).
- **N**
  - Descripción: Número total de tuits georreferenciados (originales) utilizados para calcular el `SCORE` en esa fecha y unidad administrativa.
  - Tipo: Entero (int).
  - Rango de valores: $[0, +\infty)$ (un valor de 0 indica que no hubo tuits, en cuyo caso puede omitirse la fila en algunos niveles).

D. Cuadro Resumen de Atributos

TABLE I
RESUMEN DE ATRIBUTOS DEL TSGI (POR REGISTRO DIARIO).

Atributo	**Descripción**	**Tipo**	**Rango / Ejemplos**
DATE	Fecha del índice	`string`	"2023-02-15", "2021-07-01"
NAME_0	Nombre del país (ADMIN 0)	`string`	"Peru", "United States", "Spain"
NAME_1	Nombre del estado/provincia (ADMIN 1)	`string`	"Lima", "Cusco", "Arequipa"
NAME_2	Nombre de la ciudad/condado (ADMIN 2)	`string`	"Lima Province", "Cusco Province"
SCORE	Promedio diario de sentimiento positivo	`float`	$[0.0, 1.0]$
N	Número de tuits utilizados	`int`	$[0, +\infty)$

AGRADECIMIENTOS

Se agradece al Harvard Center for Geographic Analysis por proveer el *Geotweet Archive* y al equipo de Panacea Lab por publicar el *Twitter Sentiment Geographical Index Dataset* de acceso libre.

REFERENCES

[1] X. Yu, M. D. F. Ferreira, F. V. Paulovich, *Senti-COVID19: An Interactive Visual Analytics System for Detecting Public Sentiment and Insights Regarding COVID-19 From Social Media*, *Proceedings of Scientific Visualization*, vol. 13, no. 4, pp. 50–58, 2021.

[2] I. Nemtsov, J. Jahan, C. Yan, S. R. Humayoun, *Visual Exploration of Emotion Feelings Comparison in Tweet Data*, *EG UK Theory and Practice of Computer Graphics*, pp. 53–61, 2024.

[3] Y. Chai, D. Kakkar, J. Palacios, S. Zheng, *Twitter Sentiment Geographical Index Dataset*, *Scientific Data*, vol. 10, Art. no. 572, 2023. DOI:10.1038/s41597-023-02572-7.