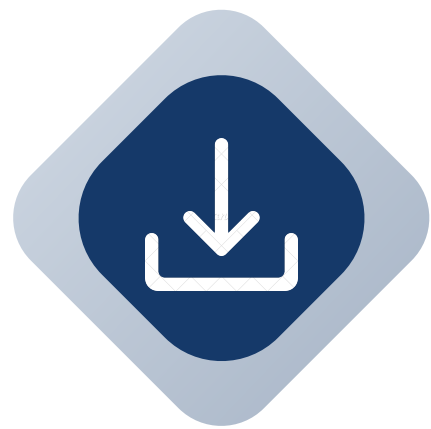


Análisis de sentimientos en mensajes de WhatsApp

En el proyecto anterior...



Exportar los chats



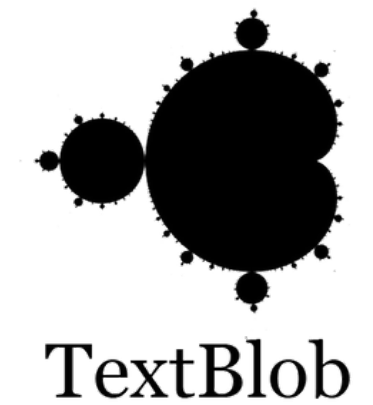
Limpiar la data



Extraer patrones



Analizar sentimientos



Pregunta de investigación



¿Qué sentimientos predominan en los mensajes de texto del grupo de WhatsApp de la generación de 31 estudiantes de Ciencias Actuariales de la Universidad de Costa Rica, identificados mediante algoritmos de clasificación?

Marco metodológico



Preprocesamiento de los datos



Modelo principal

- Mapas autoorganizados



Métodos complementarios

- K-medias, K-medianas, Clusterización jerárquica, DBSCAN, Modelos de mezcla gaussianos y no gaussianos; Agrupamiento espectral y Mean shift.



Evaluación del agrupamiento

- Coeficiente de Silhouette, Índice de Calinski-Harabasz e Índice de Davies-Bouldin

- Se le agrega una etiqueta a cada mensaje.
- Se eliminan los mensajes indeseados.
- Se procede a tokenizar los mensajes.
- Se utilizó el *NRC Word-Emotion Association Lexicon* para cuantificar las emociones presentes en los mensajes.
- Cada palabra extraída fue comparada contra el lexicon en español.
- A cada palabra se le asigna un conteo binario respecto a ocho emociones básicas y dos polaridades:

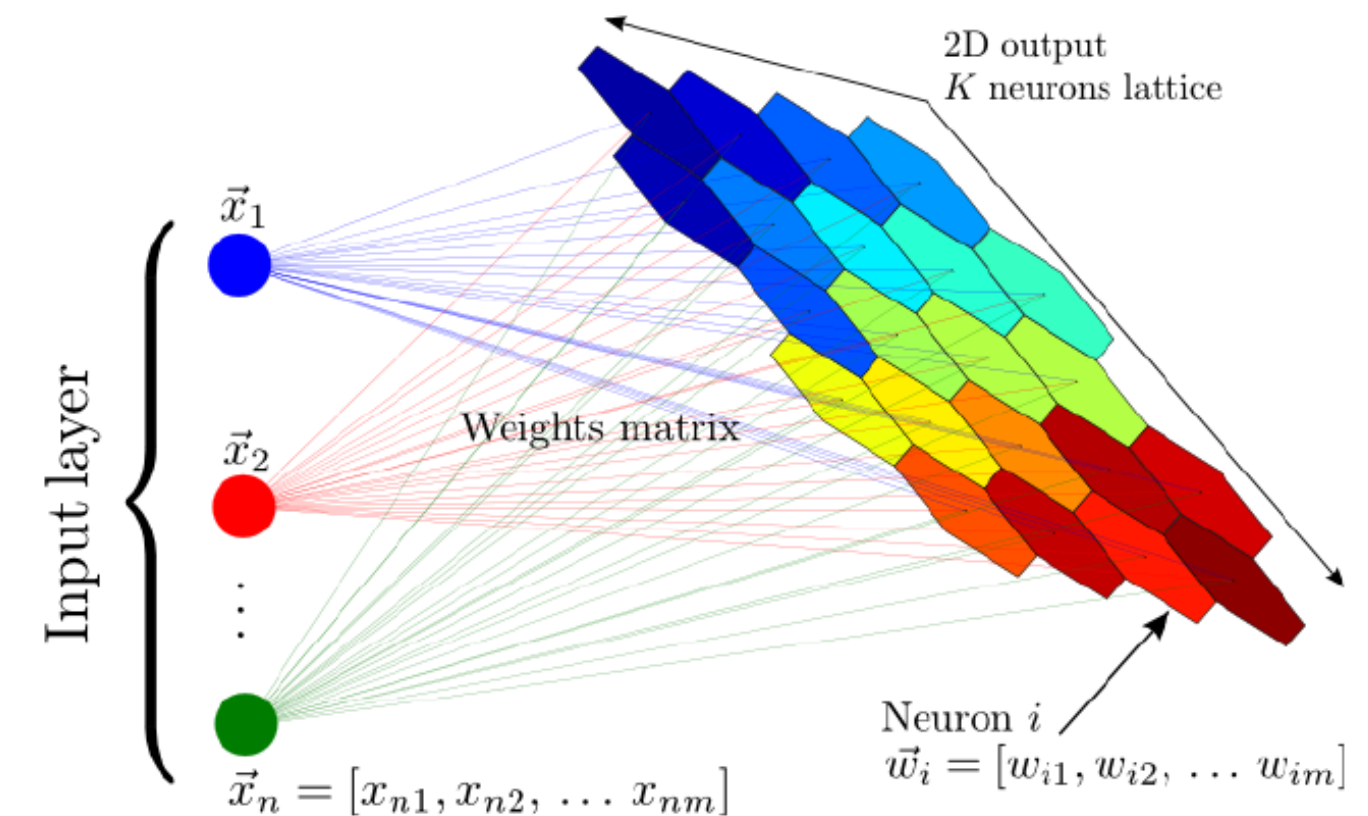
$$X(m_j) = (x_{e_1}(m_j), x_{e_2}(m_j), \dots, x_{e_{10}}(m_j)) \in \mathbb{R}^{10}$$



Mapas Autoorganizados

Análisis de datos I

- Técnica de aprendizaje no supervisado para reducción de dimensionalidad y visualización de datos complejos.
- Transforma datos multidimensionales en un mapa bidimensional que preserva relaciones topológicas.
- Cada nodo del mapa tiene un vector de pesos que representa un grupo de datos similares.
- Útil para analizar y visualizar estructuras complejas, como emociones en textos o segmentos en grandes bases de datos.



Fuente: https://dcain.etsin.upm.es/~carlos/bookAA/O6_mapasAutoorganizativosIntro.html

1. Para cada dato \mathbf{x} , se encuentra el nodo ganador i^* .
2. Se actualizan los pesos del nodo ganador y sus vecinos para ajustar el mapa a la distribución de los datos.
3. Nodos cercanos en el mapa representan datos similares, facilitando la identificación de patrones y clústeres.

*Se seleccionaron 9 clústeres tras un análisis visual de la topología del mapa y la evaluación con métricas internas, logrando un equilibrio entre nivel de detalle emocional y facilidad de interpretación.

$$i^* = \arg \min_i \|\mathbf{x} - \mathbf{w}_i\|$$

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \alpha(t) \cdot h_{i,i^*}(t) \cdot (\mathbf{x} - \mathbf{w}_i(t))$$

$$h_{i,i^*}(t) = \exp \left(-\frac{\text{dist}^2(i, i^*)}{2\sigma(t)^2} \right)$$

Ajuste del modelo

- **Coeficiente de Silhouette**

- Mide qué tan bien está asignado cada punto a su clúster.
- **Valores cercanos a 1** indican una buena separación entre clústeres.
- $a(i)$ es la distancia media a su clúster y $b(i)$ al clúster más cercano.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- **Índice de Calinski-Harabasz**

- Evalúa la proporción entre la dispersión entre clústeres y la dispersión interna.
- **Valores altos** indican clústeres bien separados y compactos.

$$CH = \frac{\text{tr}(B_k)}{\text{tr}(W_k)} \cdot \frac{n - k}{k - 1}$$

- **Índice de Davies-Bouldin**

- Compara la similitud entre cada clúster y su clúster más similar.
- **Valores bajos** indican mejor separación y menor dispersión interna.

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right)$$

Resultados

Análisis de datos I

- Mean Shift obtuvo el valor más alto de Silhouette (0.313), indicando buena cohesión interna y separación entre clústeres.
- Clustering jerárquico alcanzó el mayor valor de CH (171.65), reflejando clústeres muy compactos y bien separados.
- Clustering jerárquico también obtuvo el índice DB más bajo (1.17), lo cual es deseable porque indica baja similitud entre clústeres.
- Se aplicó una eliminación progresiva de los métodos con menor rendimiento según las tres métricas internas
- Los mapas autoorganizativos lograron un buen balance general.

Cuadro 1: Comparación de los métodos de clusterización.

| método | clusters | silhouette | cal.har | dav.bou |
|------------|----------|------------|-----------|---------|
| kmeans | 5 | 0.16273 | 85.97041 | 1.98684 |
| kmedians | 5 | 0.09895 | 81.14052 | 1.98691 |
| jerarquico | 5 | 0.19617 | 171.65244 | 1.16791 |
| dbscan | 5 | 0.11451 | 49.25579 | 1.66379 |
| gmm | 7 | 0.15318 | 68.59909 | 2.48287 |
| som | 9 | 0.23832 | 123.10330 | 1.47087 |
| spectral | 5 | 0.22614 | 108.12655 | 1.60120 |
| teigen | 3 | 0.26243 | 93.88190 | 1.31970 |
| meanshift | 5 | 0.31323 | 76.44206 | 1.74662 |

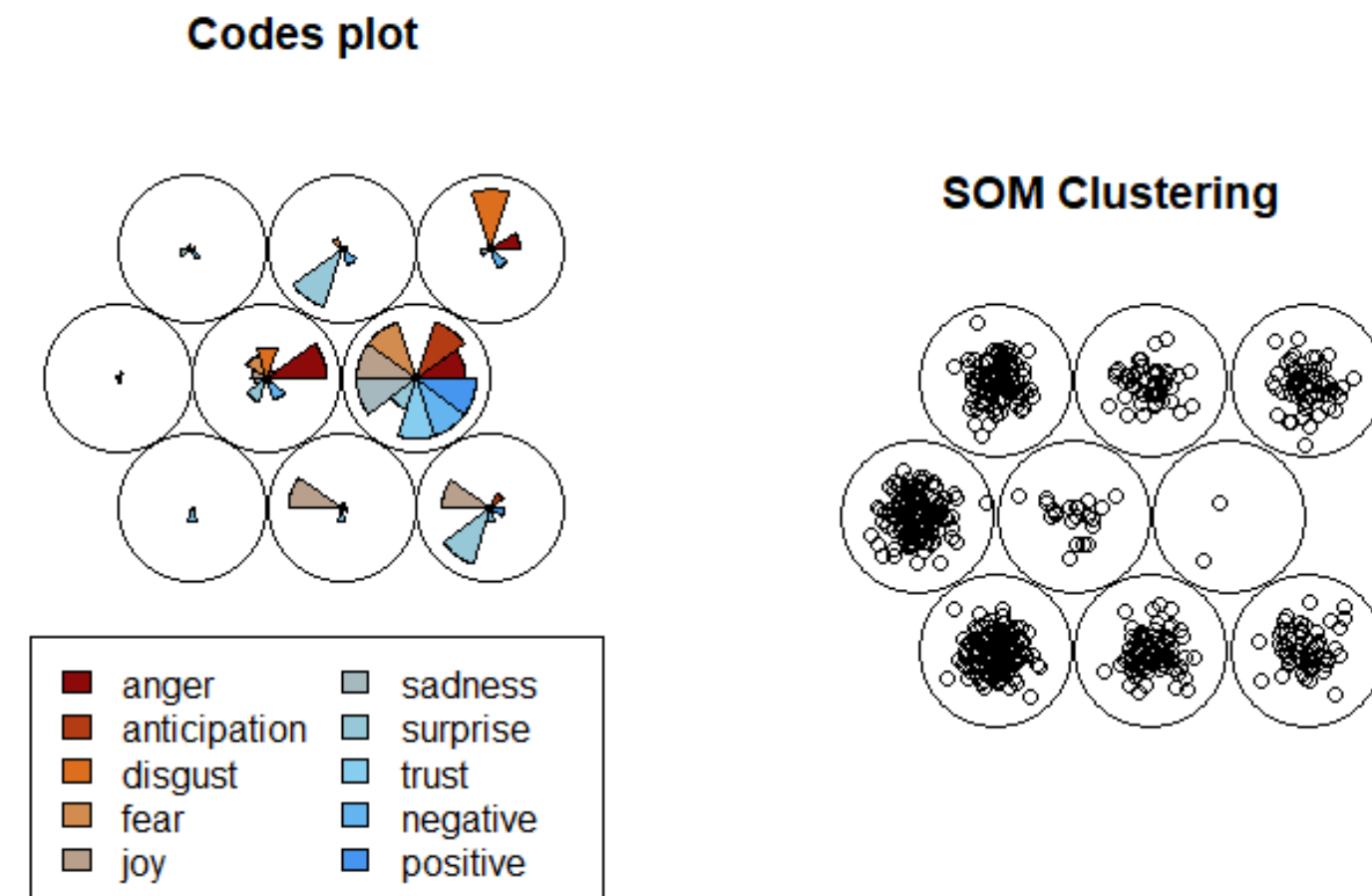
Fuente: Elaboración propia.

Resultados

Análisis de datos I

- El gráfico muestra cómo los mensajes se agrupan en una grilla bidimensional, donde cada nodo representa un clúster con características emocionales similares.
- Cada grupo identificado por el SOM está representado con un color distinto, lo que facilita distinguir visualmente las regiones emocionales predominantes.
- Se observa la densidad de mensajes en cada nodo, indicando cuáles clústeres concentran más interacciones del grupo y cuáles son menos frecuentes.
- La representación gráfica ayuda a interpretar de manera intuitiva una estructura emocional de 10 dimensiones en solo 2, haciendo más comprensible el análisis para públicos no técnicos.

Figura 1: Visualización de la distribución de observaciones por clúster en el mapa auto-organizativo (SOM).

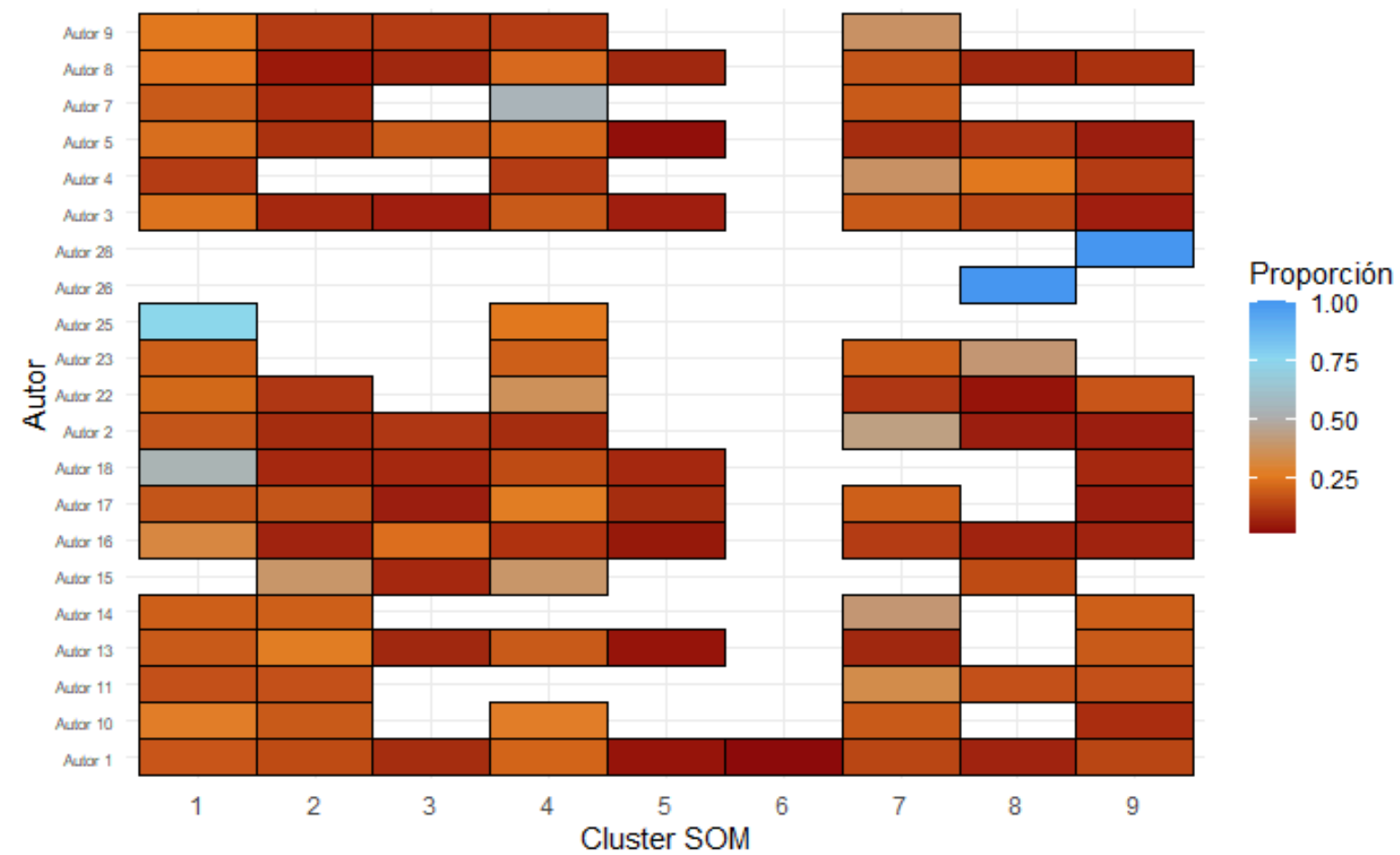


Fuente: Elaboración propia.

Resultados

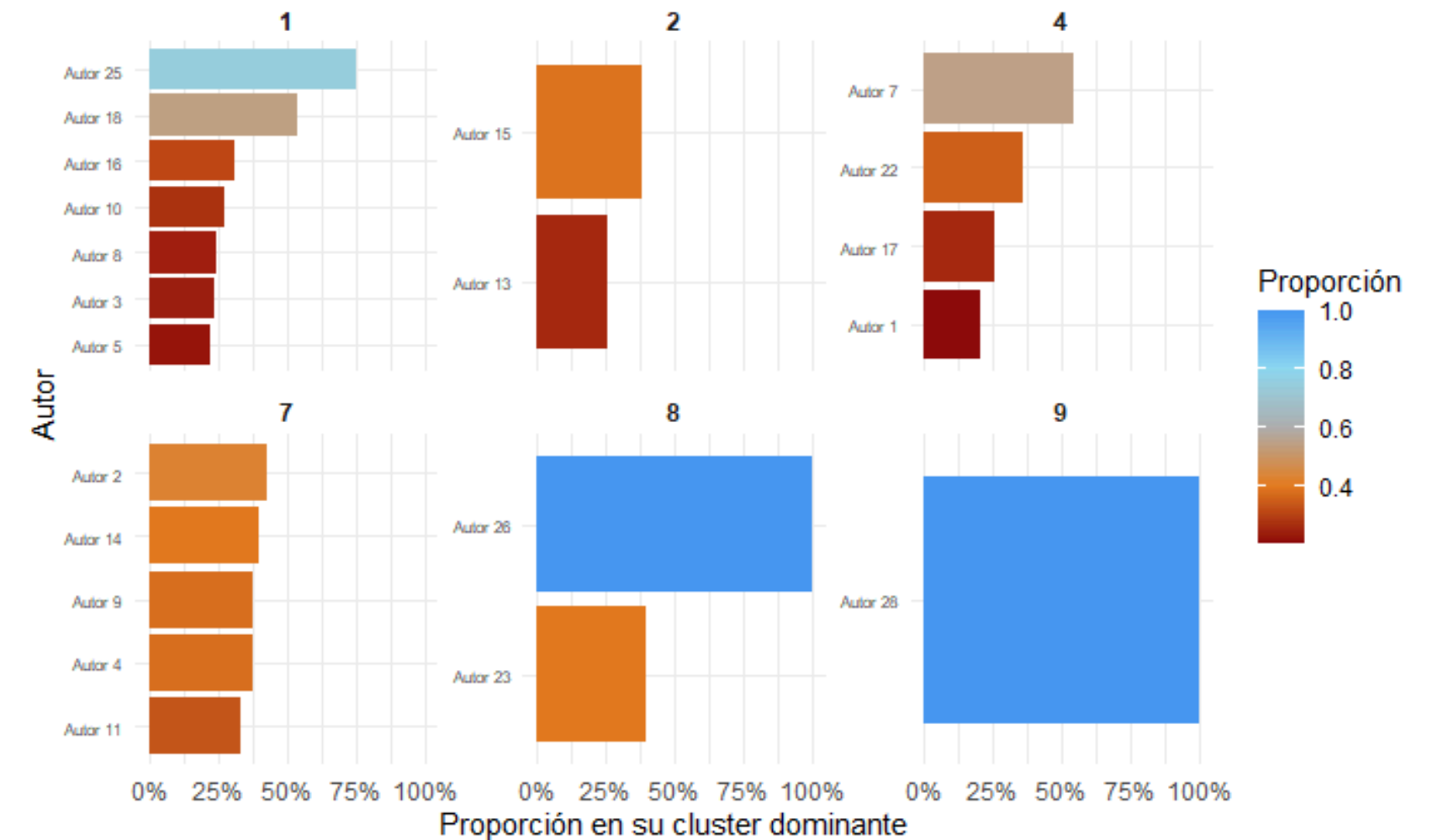
Análisis de datos I

Figura 2: Proporción de mensajes de cada autor por clúster.



Fuente: Elaboración propia.


Figura 3: Clúster dominante de cada autor.




Fuente: Elaboración propia.

Conclusiones, limitaciones y recomendaciones


Análisis de datos I




Los Mapas Autoorganizados destacaron como el modelo más efectivo, combinando buen desempeño en métricas internas con una visualización clara de la estructura emocional.



El análisis mostró que ciertos clústeres presentan emociones predominantes y que la participación de los autores varía en concentración emocional, lo cual aporta información valiosa sobre dinámicas individuales y grupales.



La traducción de diccionarios emocionales del inglés al español puede haber afectado la precisión semántica, ya que algunas palabras no tienen equivalentes exactos.



Para mejorar la detección de matices, se sugiere analizar unidades textuales más amplias (como frases completas) y utilizar recursos lingüísticos diseñados específicamente para el español.

Análisis de sentimientos en mensajes de WhatsApp