LLMs y Sentimientos

Influencia de los LLMs en el Procesamiento y Análisis de Sentimientos

Nadine Apellido, Manuel Garcés, Carlos Navarro, Marcos León

Universidad Politécnica de Madrid

13 de noviembre de 2025

Contenido

Introducción

Análisis de Sentimientos en la Era de los Grandes Modelos de Lenguaje (LLMs): Una Verificación de la Realidad

- **Introducción al Análisis de Sentimientos (SA):** El SA es un área de investigación establecida en el procesamiento del lenguaje natural (NLP). Busca estudiar las opiniones, sentimientos y emociones de las personas a través de métodos computacionales.
- **El Boom de los LLMs:** Los LLMs (como GPT-3.5/ChatGPT) han demostrado un rendimiento impresionante en varias tareas de NLP.
- **Pregunta de Investigación:** ¿Hasta qué punto se pueden aprovechar los LLMs actuales para las diferentes tareas de SA? Este estudio busca una "verificación de la realidad" sobre el estado actual del SA con LLMs.

Evaluación

Alcance de la Evaluación: 13 Tareas, 26 Datasets y Comparación con SLMs

- **Modelos Evaluados (LLMs):** Se consideraron modelos de código abierto (Flan-T5, Flan-UL2) y modelos de la serie GPT-3.5 (ChatGPT, InstructGPT/text-davinci-003).
- **Modelos de Comparación (SLMs):** Se utilizaron modelos de lenguaje más pequeños (Small Language Models - SLMs) como T5 (versión large, 770M), entrenados con conjuntos de datos específicos del dominio (in-domain labeled data).
- **Evaluación Integral:** La investigación cubrió 13 tareas distintas de SA en 26 conjuntos de datos.
- **Categorías de Tareas de SA Investigadas:**
  1. **Clasificación de Sentimientos (SC):** Clasificar la orientación sentimental de un texto (a nivel de documento, oración o aspecto).
  2. **Análisis de Sentimientos Basado en Aspectos (ABSA):** Análisis de sentimientos y opiniones a un nivel de aspecto más detallado.
  3. **Análisis Multifacético de Textos Subjetivos (MAST):** Tareas centradas en fenómenos subjetivos específicos (p. ej., detección de hate speech, ironía, reconocimiento de emociones).

Ilustración diferentes PROMPTs

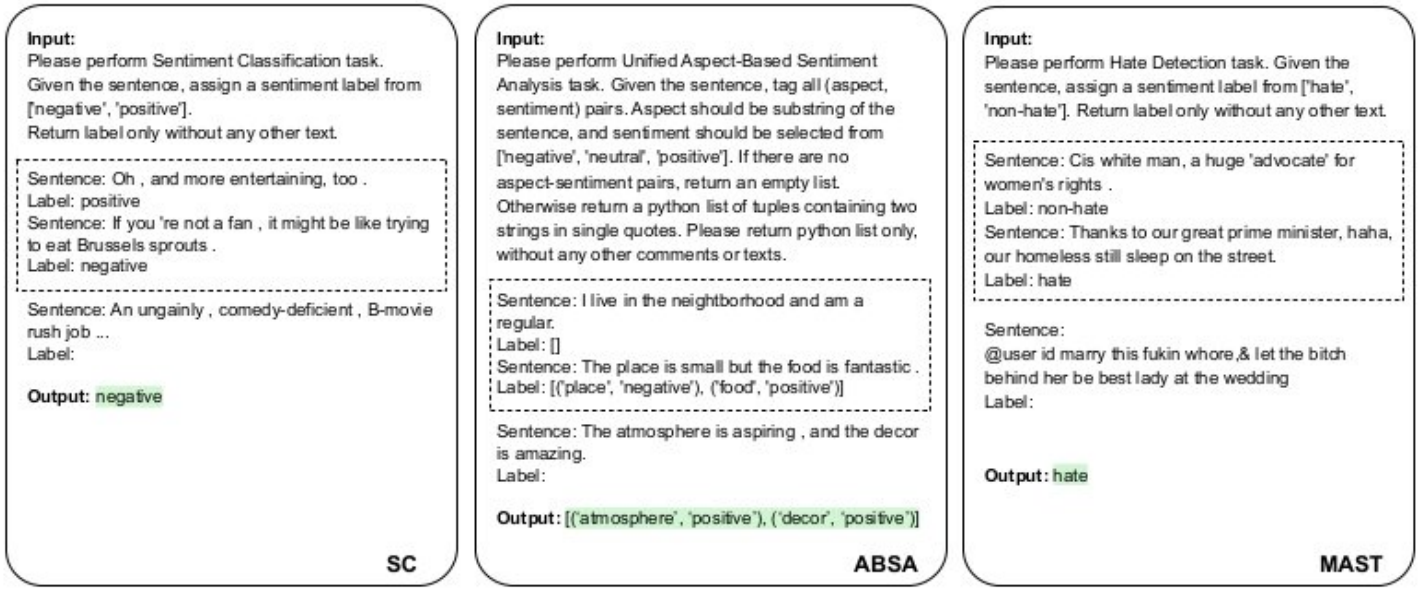

Figura: Prompt examples for SC, ABSA, and MAST respectively.

Hallazgo 1: Rendimiento Cero-Shot

LLMs vs. SLMs: Brechas de Rendimiento en Configuración Cero-Shot (Zero-Shot)

- **Tareas Simples (SC):** Los LLMs (como ChatGPT) muestran una fuerte capacidad de SA en configuraciones cero-shot. En algunas tareas SC, los LLMs pueden rendir a la par de los SLMs ajustados con el conjunto de entrenamiento completo.
  - *Ejemplo:* En promedio, el rendimiento de ChatGPT alcanzó el 97 % del rendimiento del modelo T5 ajustado en tareas SC.
- **Tareas Complejas (ABSA y MAST):** Los LLMs aún están por detrás de los SLMs entrenados con datos específicos del dominio en tareas que requieren un conocimiento más profundo o información estructurada.
  - *Desafío ABSA:* Tareas como la Extracción Unificada de Sentimientos Basada en Aspectos (UABSA) o la Extracción de Tripletes (ASTE) requieren información de sentimiento estructurada y de grano fino. Modelos como `text-davinci-003` solo alcanzaron alrededor del 54 % del rendimiento de un modelo T5 ajustado en tareas ABSA.
  - *Desafío MAST:* Los LLMs también van a la zaga en tareas MAST complejas, como el análisis de sentimiento implícito.

Hallazgo 2: Aprendizaje Few-Shot

Ventaja Estratégica: LLMs Superan a SLMs en Escenarios Few-Shot

- **Superioridad Consistente:** Los LLMs (con *in-context learning*) superan consistentemente a los SLMs entrenados con la misma cantidad limitada de datos en todas las categorías de tareas (**SC**, **ABSA**, **MAST**) en configuraciones *few-shot* (p. ej., 1-shot, 5-shot, 10-shot).
- **Implicación Práctica:** Esto sugiere que la aplicación de LLMs es ventajosa cuando los recursos de anotación son escasos.
- **La Brecha de Datos:** `ChatGPT` establece una base sólida que requiere que el modelo `T5` utilice casi cinco a diez veces más datos (p. ej., 50-shot o 100-shot) para lograr un rendimiento comparable en promedio.
- **Impacto de la Complejidad del Prompt:** El beneficio incremental de añadir más ejemplos (*shots*) es menos obvio para tareas **SC** más simples, pero aumenta considerablemente el rendimiento de los LLMs en tareas **ABSA** que exigen una comprensión más profunda y un formato de salida preciso.

Resultados FEW-SHOT

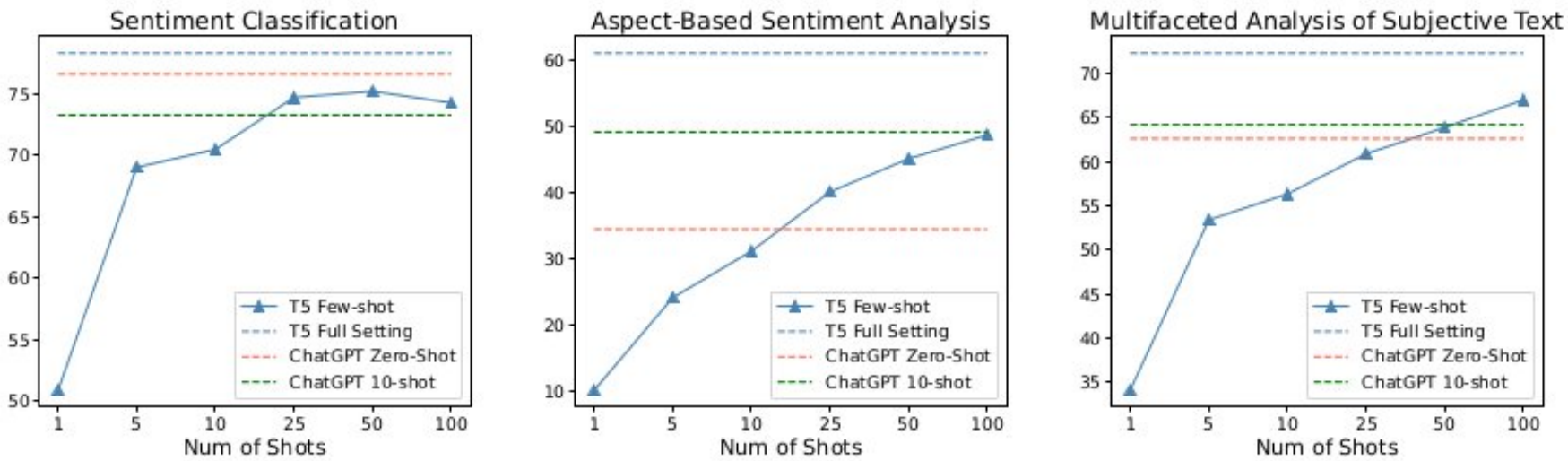

Figura: Averaged few-shot results on all datasets for each task type with an increasing number of different shots.

Desafio clave: Sensibilidad a Prompts

Fragilidad de los LLMs: La Sensibilidad al Diseño del Prompt

- **Variabilidad Observada:** El diseño de *prompts* adecuados es fundamental. Los LLMs pueden producir respuestas muy diferentes incluso con *prompts* semánticamente similares.
- **Impacto de la Complejidad de la Tarea:**
  - En tareas **SC** (clasificación simple), la elección del *prompt* parece tener un efecto menor.
  - En tareas que requieren una salida estructurada y de grano fino (**ABSA**), el rendimiento puede variar significativamente (mayor sensibilidad) dependiendo del diseño del *prompt*.
  - *Detalle:* Se encontró que los modelos pueden ser sensibles a ciertas palabras, como *analyze*, donde el modelo podría generar explicaciones largas a pesar de recibir instrucciones explícitas para no hacerlo.

Sensibilidad de PROMPTs

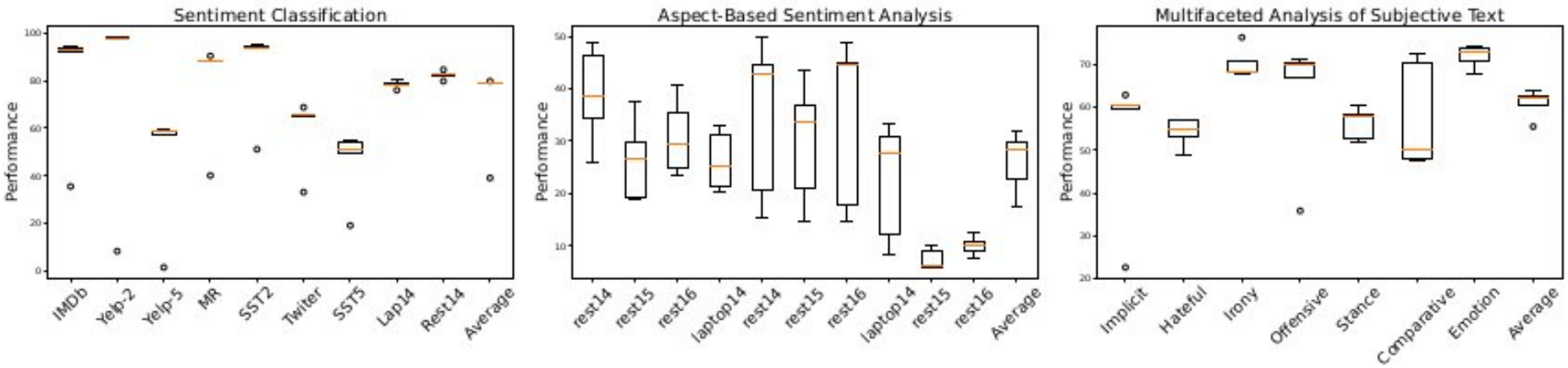

Figura: Sensitivity of different prompt designs on three types of SA tasks.

Prompt Estructurado

Ilustración de la Complejidad en Tareas ABSA

Ejemplo de Prompt para UABSA

Para la tarea UABSA (Unified Aspect-Based Sentiment Analysis), un LLM podría recibir un prompt como:
"Extrae los aspectos y clasifica el sentimiento de cada uno en el siguiente texto, proporcionando la salida en formato JSON estructurado con las claves 'aspect', 'sentiment' y 'evidence'."

Limitaciones y Benchmark

Limitaciones de la Evaluación Actual y el Nuevo Benchmark SENTIEVAL

- **Deficiencias de las Prácticas de Evaluación Actuales:** Las evaluaciones a menudo se centran estrechamente en tareas o conjuntos de datos específicos, y utilizan *prompts* inconsistentes entre estudios. Esto no logra capturar la amplitud total de las capacidades de **SA** de un **LLM**.
- **Propuesta SENTIEVAL:** Los autores proponen un nuevo *benchmark* llamado `SENTIEVAL` para una evaluación más integral y realista.
  - *Objetivo 1:* Romper la barrera entre las tareas individuales de **SA** para una evaluación más integral.
  - *Objetivo 2:* Evaluar el modelo utilizando instrucciones de lenguaje natural presentadas en varios estilos. Esto imita el uso real donde los humanos interactúan con el modelo de diversas maneras.
  - *Resultado:* `SENTIEVAL` requiere que el modelo comprenda diversos estilos de instrucciones y cumpla con formatos requeridos, un desafío donde `ChatGPT` demostró mayor robustez en comparación con `Flan-UL2`.

Conclusiones Clave

Conclusiones Clave para la Adopción Práctica de LLMs en SA

- **LLMs son Soluciones Viables para Tareas Simples:** Para clasificación binaria o trinaria simple, los LLMs pueden ser soluciones efectivas, incluso en modo cero-shot.
- **Ventaja en Escasez de Datos:** Cuando los recursos de anotación son limitados, los LLMs son una buena opción debido a su rendimiento superior en el aprendizaje *few-shot*.
- **Precaución con Tareas Estructuradas:** Para tareas que requieren una salida estructurada de sentimientos (**ABSA**), los LLMs pueden no ser la mejor opción y su rendimiento puede variar significativamente con diferentes *prompts*.
- **Tamaño del Modelo vs. Instruction-Tuning:** Los modelos más grandes no siempre garantizan un rendimiento superior (p. ej., `Flan-UL2` fue comparable a la serie `GPT-3.5` en algunos casos). El uso de *Instruction-Tuning* en modelos de tamaño razonable puede ser suficiente para aplicaciones prácticas de **SA**.

Contenido

Conclusiones

- Idea principal 1
- Idea principal 2
- Idea final inspiradora

LLMs y Sentimientos

Influencia de los LLMs en el Procesamiento y Análisis de Sentimientos

Nadine Apellido, Manuel Garcés, Carlos Navarro, Marcos León

Universidad Politécnica de Madrid

13 de noviembre de 2025