Transcripción y análisis de sentimientos de las cartas de Máximo Gómez

Adrian Hernández Santos
Alejandro Alvarez Lamazares
Frank Pérez Fleitas
Eisler Valles Rodríguez
Rafael Acosta Márquez

Facultad de Matemática y Computación
Universidad de La Habana

January 30, 2025

1 Descripción General

Nuestro problema consiste en extraer el texto de una colección de las cartas del Generalísimo Máximo Gómez y realizar un análisis de sentimientos a dichos resultados.

2 Características de los datos

Los datos provistos están conformados por fotos de cartas históricas escritas a mano por Gómez. Al ser documentos históricos, algunos están dañados por quemaduras, desgaste del papel, rasgados, por lo que algunas presentan incompletitud en los propios datos a procesar. También cabe destacar que el color del papel no es constante, hay tonos más claros, más oscuros, con manchas de humedad, incluso con hojas rayadas (con los renglones escritos). Además, existe un subconjunto de cartas mecanografiadas y algunas completadas por terceras personas. La cantidad de cartas es alrededor de 160, junto con las cartas personales enviadas a Lola Tió.

3 Análisis del estado del arte

Se realizó un análisis del estado del arte respecto al tema: "Reconocimiento de texto manuscrito y análisis de sentimiento". Se encontraron diversos enfoques, desde el uso de *CNN* (redes neuronales de convolución) y *SVM* para la extracción de texto, hasta *Naive Bayes* y *SVM* nuevamente en el análisis de sentimiento. Estas propuestas fueron descartadas debido a la falta de datos suficientes para entrenar adecuadamente estos modelos. Finalmente, se decidió utilizar *YOLOv11* y *TrOCR* para la extracción del texto y un modelo basado en *RoBERTa* para el análisis de sentimientos, empleando modelos preentrenados.

4 Extracción de texto manuscrito

4.1 Selección de características

Dada la variabilidad de los datos, se decidió convertir las imágenes a escala de grises para minimizar el sesgo introducido por variaciones en el color del fondo y la tinta. Se optó por detectar palabras individuales mediante el entrenamiento de un modelo *YOLOv11* en un subconjunto de los datos, como una reducción de complejidad al problema inicial. Se usó la plataforma *Roboflow* para anotar las *"bounding boxes"* con un único label *"word"*, entrenando el modelo para detectar bloques contiguos de escritura.

4.2 Evaluación del modelo *YOLOv11*

Para evaluar el modelo *YOLOv11*, se emplearon las métricas *mAP*, *Accuracy*, *Precision* y *Recall*. La métrica *mAP* se utiliza en los modelos de detección de objetos para evaluar el ajuste del modelo a la detección de los *bounding boxes* y la clasificación de las mismas. *Precision* representa la razón de Verdaderos Positivos *(True Positives)* con respecto al total de positivos inferidos por el modelo *(Verdaderos + Falsos)*, se puede formular como: "¿Del total de predicciones, cuántas fueros correctas?". Recall representa la razón de los Verdaderos Positivos con respecto a los Verdaderos Positivos y Falsos Negativos, de lo cual se puede formular: "¿Del total de positivos, cuántos el modelo fue capaz de identificar correctamente?". Los resultados fueron los siguientes:

- mAP: 0.94
- Precision: 0.89
- Recall: 0.942

Estos valores reflejan un buen rendimiento en la detección de bloques de texto manuscrito. Se ajustaron los hiperparámetros con un umbral de confianza del 20% y un solapamiento del 45% para mejorar la diferenciación entre imágenes con líneas estrechas y letras con trazos alargados.

4.3 Extracción del texto de las palabras

Una vez extraídas las imágenes de las palabras, se utilizó el modelo TrOCR de HuggingFace: `microsoft/trocr-small-handwritten`. Dicho modelo esta basado en transformers, con un encoder de imagen y un decoder de texto, lo que permite extraer el texto de las imágenes, este modelo específicamente texto manuscrito. Se intentó hacer *fine tuning* con las imágenes anotadas, pero no fue posible debido a limitaciones de procesamiento y tiempo, por lo que se provee la evaluación de los resultados utilizando el modelo preentrenado base.

4.4 Evaluación del modelo TrOCR

Se emplearon las métricas *Precision*, *Recall*, *CER* (Character Error Rate), *WER* (Word Error Rate), *F1* y *Accuracy*. *CER* representa el porcentaje de error con respecto a los valores esperados con respecto a los caracteres, *WER* de manera similar pero con respecto a las palabras. La puntuación *F1* utiliza una media armónica para balancear los resultados de *Precision* y *Recall*, bastante útil cuando existe un desbalance entre estas o cuando es necesario tener en cuenta tanto los Verdaderos Positivos como los Falsos Negativos. *Accuracy* representa el porciento de aciertos del modelo con

respecto a los valores correctos anotados en el dataset. La evaluación arrojó los resultados:

- Accuracy: 0.0000
- Precision: 0.0000
- Recall: 0.0000
- F1: 0.0000
- CER: 14.4978
- WER: 9.7353

Dichas métricas nos dan a entender que el modelo es pésimo ajustándose a nuestros datos, pues no pudo acertar en ningún caso, de ahí las scores de 0, además que se espera que los valores de *CER* y *WER* estén entre 0 y 1 (representando un porciento), pero al ser mayores que 1, implica que la tasa de error no es ni siquiera válida, pues los fallos son enormes. Esperamos que si se pudiera hacer *fine tuning* estos resultados mejoren o al menos logren un mejor ajuste a nuestros datos.

5 Análisis de sentimientos

Dado que nuestro conjunto de datos es bastante reducido, optamos por el uso de un modelo preentrenado, basado en *RoBERTa*, una versión de BERT optimizada. Dicho modelo fue entrenado utilizando varios datasets en español, principalmente conteniendo tweets, reviews de productos, comentarios de películas, etc. Es sabido que el lenguaje moderno difiere bastante tanto gramatical como contextualmente del usado en documentos históricos, pero a falta de algo más ajustado, se decidió aceptar dicho sesgo. Cabe agregar que dichos datasets no son públicos y a pesar de solicitar acceso a ellos, hasta la fecha no hemos recibido respuesta.

5.1 Evaluación del modelo de análisis de sentimientos

En la evaluación del modelo se utilizaron nuevamente las métricas *Accuracy*, *Precision*, *Recall* y *F1*, ya que las clases son discretas, aunque el modelo también devuelve resultados continuos, pero decidimos solamente utilizar los resultados discretos, puesto que una anotación manual de los sentimientos de las cartas de forma continua es una tarea bastante tediosa y abierta a introducir sesgos innecesarios. Los resultados fueron:

- Accuracy: 0.7857
- Precision: 0.8190

- Recall: 0.7857
- F1: 0.7852

Estos resultados indican que la elección del modelo fue correcta, aunque un *fine tuning* con algún dataset de textos históricos mejor ajustado podría mejorar su desempeño.

6 Conclusiones

A pesar de no poder realizar el *fine tuning* del último paso de la extracción de texto y por tanto, no fue posible obtener resultados finales, las métricas de los modelos utilizados con respecto a los datos provistos de entrada fueron alentadoras. Esto nos da una idea de que con un poco de más tiempo y mejor equipo de cómputo el problema puede ser resuelto satisfactoriamente. Además, esperamos que nuestra experiencia sea de utilidad en problemas de dominios similares, especialmente a la Oficina del Historiador de la Ciudad, que puede poseer colecciones de cartas o manuscritos históricos, de distintos autores, que por falta de herramientas que solucionen decentemente el problema, solamente tienen en su mano las imágenes de dichos documentos. Cabe destacar que el resultado no va a ser perfecto, por lo que se sugiere también el uso de alguna herramienta basada en *Grandes Modelos de Lenguaje (LLMs)* para ayudar a corregir la salida y mejorar la calidad de la transcripción.