**Validación manual por muestreo aleatorio**

La validación manual por muestreo aleatorio es un método utilizado en ciencia de datos, machine learning, procesamiento de lenguaje natural (NLP) y desarrollo de sistemas inteligentes. Consiste en seleccionar una muestra representativa de la población de datos, de forma razonablemente aleatoria y de tamaño suficientemente grande (habitualmente entre el 20% y 40%).

En este caso, la población total consta de 132 preguntas detectadas automáticamente. Se seleccionó una muestra de 44 preguntas, lo que representa el 33,33% del total (44 /132 \*100).

Para realizar la selección, se utilizó un intervalo sistemático de 3 (132 ÷ 44). Se eligió aleatoriamente un número entre 1 y 3 para comenzar (en este caso, el número 2), y a partir de ahí se tomaron las posiciones: 2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59, 62, 65, 68, 71, 74, 77, 80, 83, 86, 89, 92, 95, 98, 101, 104, 107, 110, 113, 116, 119, 122, 125, 128, 131.

La revisión se hizo manualmente, con criterios claros y previamente definidos para evaluar tanto las preguntas como las respuestas asociadas.

**Criterios de evaluación de detección de preguntas**

|  |  |  |
| --- | --- | --- |
| Tipo de evaluación | Descripción | Ejemplo |
| BIEN | El mensaje es efectivamente una pregunta (académica o administrativa), clara y completa (no depende de mensajes anteriores o posteriores). | “Buenas! ¿Qué tal? Quería saber si iban a subir el código de lo que hizo Eze hoy en clase.” |
| REGULAR | El mensaje es una pregunta, pero necesita contexto para entenderse (es dependiente de otros mensajes) o puede no ser útil como parte de una base de conocimiento. | “¿La única forma de hacerlo es creando una lista desde antes?” |
| MAL | El mensaje no es una pregunta, sino un agradecimiento, afirmación o fragmento de conversación. | “Sí, eso también va en el final.” |

**Criterios de evaluación de respuestas asociadas**

|  |  |  |
| --- | --- | --- |
| Tipo de evaluación | Descripción | Ejemplo |
| BIEN | Al menos una de las respuestas detectadas responde correctamente el contenido principal de la pregunta. Se acepta que haya respuestas adicionales irrelevantes. | Una pregunta con tres respuestas, pero una responde adecuadamente. |
| REGULAR | Las respuestas son parciales, poco claras o necesitan demasiado contexto para comprenderse. | El docente agrega un comentario útil después de que el alumno ya había cerrado la conversación. |
| MAL | La pregunta no tiene respuesta, o las respuestas detectadas no contestan la pregunta. | A una consulta técnica sobre Git se responde solo con “Gracias, era eso!”. |

**Métricas de calidad evaluadas**

En esta etapa se calculó únicamente la precisión, tanto en la detección de preguntas como de respuestas:

* + Precisión en preguntas =
  + Precisión de respuestas =

Donde "respuestas correctas" hace referencia a aquellas que efectivamente responden a la pregunta detectada. Esta evaluación se realiza por hilo de conversación, no por mensaje individual. Se considera una respuesta correcta si contesta a la pregunta, aunque haya otras respuestas irrelevantes alrededor.

No se calculó recall ya que no se conoce la cantidad total real de preguntas o respuestas existentes en los archivos JSON.

* + - * Recall =

F1-score tampoco puede calcularse porque requiere combinar precisión y recall:

* + - * F1-score =

Resultados

La validación se realizó el 30 de abril de 2025, sobre el archivo “ log\_preguntas\_respuestas” generado por el sistema de detección automática.

* + Cantidad de preguntas evaluadas: 44
  + Precisión de preguntas: 98,86%
  + Precisión de respuestas: 89,77%
  + Precisión combinada (preguntas con sus respuestas): 94,32%

Tendencias observadas

* Alta precisión en la detección de preguntas: solo se pierde 0,5 puntos en toda la muestra.
* Buena precisión en respuestas: aunque algunas preguntas quedaron sin respuesta o fueron parcialmente respondidas, se justifica por la naturaleza de las conversaciones en Discord: hay idas y vueltas de mensajes de distintos usuarios.

Casos complejos observados:

* #3: Respuesta incompleta porque el docente agregó más comentarios luego de que el alumno cerró la conversación con alguna frase como “muchas gracias”.
* #7: La pregunta debió haber sido parte de un hilo o conversación como una respuesta, pero fue agrupada incorrectamente y el error se replica en las respuestas.
* #41: La pregunta no tuvo respuesta de un docente, lo que llevó a una falsa asociación de respuesta.