1. **Introducción**

Los Agentes Conversacionales son sistemas capaces de entender y generar lenguaje natural. Un ejemplo de estos agentes, se conoce como Sistemas de Búsqueda de Respuestas (o Sistemas de Question-Answering), los cuales se dedican exclusivamente a proveer respuestas de forma adecuada a partir de una pregunta formulada en lenguaje natural [2]. Analizar la pregunta entrante es una componente fundamental para el éxito de estos sistemas por lo que, entre otras tareas, es necesario realizar la Clasificación de la Pregunta. Esta tarea se define como el subproceso de derivar el tipo de respuesta esperada a partir de la pregunta entrante [3], permitiendo mejorar el desempeño de los Sistemas de Búsqueda de Respuestas al reducir el espacio de búsqueda de respuestas potenciales [4]. Sin embargo, los recursos en idiomas diferentes al inglés siguen siendo escasos, difíciles y costosos de adquirir [5], recurriendo principalmente a la traducción desde el inglés a distintos idiomas como el español [6, 7, 8, 9] o bien a la construcción de datasets propios aplicando diferentes técnicas de etiquetado [10, 11, 12]. En la Tesis de Grado [13] se llevó a cabo un etiquetado manual de la colección con el fin de resolver el problema de clasificación de preguntas considerando el tipo de respuesta. Entre las técnicas utilizadas, el modelo BETO destacó al demostrar eficiencia en la clasificación de la mayoría de las categorías. Resultó interesante observar que el modelo enfrentó desafíos en etiquetas que también fueron problemáticas durante el proceso de etiquetado. Si bien se realizó un estudio completo, el principal problema fue la cantidad limitada de muestras de la colección de preguntas. En este contexto, el surgimiento de modelos de lenguajes generativos como GPT-2 y GPT-3 de OpenAI [19, 20], LaMDA y PaLM2 de Google [21, 22], LLaMa de Meta [23], como así también la familia de modelos MarIA[24]**,** podrían ser de gran utilidad.

De esta manera, el objetivo general de la pasantía fue: *contribuir a la formación académica del pasante mediante el uso de modelos de lenguaje generativos para la aumentación de datos*. Para ello, se definieron los siguientes objetivos específicos:

* Complementar la formación académica del pasante aplicando los conceptos de Aprendizaje Automático.
* Adquirir conocimientos y fundamentos vinculados a: agentes conversacionales, aumentación de datos, modelos generativos, entre otros.
* Experimentar con modelos generativos implementando una herramienta de software de generación de texto.
* Introducir al pasante en prácticas de investigación: lectura y comprensión de artículos científicos, escritura y presentación de resultados, entre otras.

La pasantía se desarrolló en un plazo de 8 semanas, a partir del siguiente cronograma de actividades:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tarea** | **Semanas** | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **1** | **X** |  |  |  |  |  |  |  |
| **2** |  | **X** | **X** |  |  |  |  |  |
| **3** |  |  | **X** | **X** | **X** | **X** |  |  |
| **4** |  |  |  |  |  | **X** | **X** |  |
| **5** |  |  |  |  |  |  | **X** | **X** |

1. Lectura de material bibliográfico vinculado a aprendizaje automático supervisado, agentes conversacionales y modelos generativos.
2. Explorar el uso de diversos modelos para la generación de texto en lenguaje natural.
3. Implementación de una herramienta de software para la generación de muestras etiquetadas y aplicarla para aumentar la colección de preguntas de AFIP usando un modelo de lenguaje.
4. Interpretación de los resultados.
5. Elaboración del Informe de Pasantía.

En las siguientes secciones se explica el desarrollo completo de la pasantía, abordando la metodología utilizada, resultados alcanzados y conclusiones. Además, se enuncian las relaciones entre los resultados y los objetivos planteados originalmente.

1. **Metodología 🡺 TODO LO QUE HICISTE**

hola

(Tareas 1 y 2) Se utilizó *Bard*, un modelo generativo que utiliza Palm 2 [CITA Palm2]

Inicialmente se hizo una lectura del material bibliográfico relacionado al tema de la pasantía. Fue necesario comprender en profundidad el problema de la Clasificación de Preguntas. La Tesis de Grado [13] explica cada una de las componentes de los Sistemas de Búsqueda de Respuestas, así como los desafíos que requiere la Clasificación de Preguntas. Se comprendió cómo fue el proceso de etiquetado y se hizo una revisión en profundidad de la colección de preguntas. En resumen, se tienen las siguientes clases de preguntas:

* Explicar las clases de preguntas… (tesis de grado)
* Mostrar el desbalance entre las clases

Esto fue necesario para….

Algo de los modelos de clasificación que viste en la tesina (BETO, que luego lo usaste)

Modelos de Lenguajes Generativos…

* Que tipos averiguaste? (incluir alguno en ingles) ventajas y desventajas.
* Cual aplicaste de todos ellos? Porque lo elegiste? (razones… porque son para español, limitaciones… etc)

(Tarea 3)

Se realizó un software para la generación automática de preguntas utilizando el modelo generativo Bard y considerando las clases definidas para resolver el problema de la Clasificación de Preguntas según el Tipo de Respuesta. A continuación, se mencionan los detalles procedimentales.

**Descripción del software propuesto**.

Se dividió en dos componentes:

1. Generación de contenido

* Breve descripción de lo que hace
* Configuración inicial
* Describir “cómo hace lo que hace” (mencionar que haces un prompting y vas generando resultados)
* Cual es la salida del software (todo el contenido lo vas guardando)
* Limitaciones

1. Extracción de preguntas a partir del contenido generado

Acá podes describir que de todo lo que se generó, fuiste capturando/extrayendo las preguntas y armando el dataset.

1. **Resultados alcanzados 🡺 TODO LO QUE OBTUVISTE**

(Tarea 4)

A partir de las muestras generadas por el software propuesto, se realizó un análisis de las preguntas obtenidas por el modelo generativo.

A través de una revisión manual, se detectaron un gran número de preguntas duplicadas…..

* Cuantas preguntas (con los duplicados)
* Razones de porque se duplicaban?

Se hizo la eliminación de preguntas duplicadas… Obtuvimos ….. # nuevas preguntas (sin duplicados)

Se decidió analizar si las preguntas generadas están efectivamente dentro de la clase correcta. Debido a la cantidad elevada de nuevas muestras, sería inviable analizar pregunta a pregunta si clase es la correcta…. Por tal motivo, utilizando el modelo BETO (mejor clasificador propuesto en [13]) el cual fue fine-tuneado para resolver el problema de clasificación de preguntas usando el corpus original. (contar cual fue la idea de esto…..)

De esta manera obtuvimos estos resultados (1° matriz de confusión)

Analizar estos resultados…

* Poner ejemplos de clases correctas (diagonal) preguntas que sean interesantes!!! En tema!!
* Decir que pasó con lo que esta mal? (razones, ejemplos, hablar de la clase Entidad y que tenia pocas muestras)

A partir de estos resultados, que se hizo?? Cual fue la idea? Explicar… mostrar nuevamente la diagonal de las preguntas mal clasificadas.

Cual fue el dataset final?

Mostrar resultados finales… La estadística con las barras de cada clase donde se vea (antes y después)

De esta manera generamos muestras, aseguramos de que sean de tal clase … (frase linda para terminar resultados).

1. **Conclusiones**

Cerrar dando conclusiones finales.. Deben responder a los objetivos..

Al final podes cerrar con alguna frase que resuma lo que aprendiste.

1. **Relación entre los resultados obtenidos durante el desarrollo de la pasantía con los objetivos planteados originalmente**

Mencionar que se cumplieron todos los objetivos.. etc..

1. **Bibliografía**

**[2]**

**[3]**

**[9]**

**[CITA Palm2]** Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.