UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

PREDICCIÓN DE LA SIGUIENTE PALABRA EN UN TABLERO DIGITAL DE COMUNICACIÓN PARA NIÑOS CON PARÁLISIS CEREBRAL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

KONRAD IVELIC CORREA

PROFESOR GUÍA:
FRANCISCO GUTIÉRREZ FIGUEROA

MIEMBROS DE LA COMISIÓN:
JOSÉ PINO URTUBIA
JAVIER BUSTOS JIMÉNEZ

SANTIAGO DE CHILE
2024

PREDICCIÓN DE LA SIGUIENTE PALABRA EN UN TABLERO DIGITAL DE COMUNICACIÓN PARA NIÑOS CON PARÁLISIS CEREBRAL

La comunicación es un proceso central dentro del entendimiento entre las personas de nuestra sociedad. Para la mayoría de las personas, expresarse de manera comunicativa ocurre de forma natural e intuitiva. Sin embargo, para parte de la población que posee algún tipo de impedimento del habla, como lo son las personas que padecen parálisis cerebral (PC), la tarea de comunicarse es muy compleja e incluso imposible.

Para mejorar la comunicación de las personas que padezcan enfermedades que afectan su comunicación, existen Sistemas Alternativos y Aumentativos de Comunicación (SAAC), como lo son los tableros digitales en base a pictogramas. En estos, el usuario debe seleccionar iconos asociados a palabras en forma secuencial para así hilar oraciones.

En este trabajo de título se busca integrar un sistema de predicción de la palabra siguiente a un tablero digital de comunicación basado en pictogramas para niños con parálisis cerebral. Esto con el fin de aumentar la utilidad y el valor que percibe el usuario al utilizar la aplicación. Permitiendo una comunicación más eficaz y fluida para el usuario.

Se desarrollan tres modelos de predicción distintos, uno en base a Frecuencias de las palabras, uno basado en procesos de cadena de Markov y uno basado en Transformadores. Los cuales serán evaluados de manera intrínseca y extrínseca para determinar si es que agregan utilidad y valor al tablero ya existente y en particular, para determinar cual de estos tres modelos es el que mejor se adapta a la solución.

De forma intrínseca se muestra como resultado que el modelo de Markov y el de Transformadores producen predicciones mucho más acertadas que el modelo básico que utiliza frecuencias relativas para estimar la probabilidad de la palabra siguiente.

Al realizar una comparativa extrínseca entre los modelos de predicción permite llegar a la conclusión que todos los modelos diseñados mejoran la comunicación al ser integrados en la aplicación. Además se muestra un desempeño transversalmente más eficiente para el modelo de Markov y el de Transformadores.

Al considerar el contexto particular y las limitaciones entre en las cuales se enmarca este trabajo de título, se determinó que el modelo que entrega una utilidad y valor agregado mayor a la aplicación es el basado en las cadenas de Markov.

Para el niño de 13 años
que le gustaban las matemáticas.

Se logra mucho al hacer lo que te gusta

Agradecimientos

Quiero agradecer primero a mi familia. A mi papá porque desde muy chico me enseño que la perseverancia, trabajo duro siempre dan frutos, sobretodo cuando se sigue el buen camino. A mi mamá por siempre demostrar que la creatividad es un atributo importante en cualquier profesión y ámbito diario, además que el cariño todo lo puede. A mis hermanos Krasna, Branko y Tonka que me sacaron sonrisas cuando no sabián que las necesitaba y fueron la gran razón por la que siempre me gusta volver a mi casa. A mis gatos y perritos por la serotonina diaria.

A mis amigos que me acompañan desde la étapa escolar: Baeza, Chino, Emiliano, Lechuga, Lohmann, Mondaca, Pita, Sanhueza, Angi, Ferni, Flori, Fran, Isi, Javi, Sophia y Vale. Han estado en todas las etapas de mi vida distrayéndome de la misma, siempre me ayudaron a disfrutar de la vida y me enseñaron que es una amistad. Sobretodo gracias por aguantar mis excusas de que estaba escribiendo este trabajo.

A mis amigos que me acompañaron durante toda la universidad, que hicieron que estos 6 años hayan sido de los mejores de mi vida, gracias Cristobal, Crocco, Jaime, Jose, Hernando, Martín, Mendez, Michi, Musso, Saavedra, Pancha, Pepe y Rodri por acogerme en su grupo desde primer año. En particular gracias Yoyo por ser mi mejor amiga durante muchos de estos años, siempre supiste como hacer que me sintiera a gusto en la U.

Gracias a mis amigos furritos, que desde que entré al DCC hicieron que el toqui se volviese un lugar agradable y hacer que el paso por la especialidad sea la experiencia más grata posible. Gracias a Asu, Albani, Cebolla, Champi, Cholo, Coti, Julia, Oportus, Paula, Pablo, Roberto, Tomi y Vale por acogerme en su lista y en el grupo. Agradecimiento especial a Bruno y Diego que nos acompañamos mutuamente haciendo la memoria y sin ellos no lo habría logrado. Otro saludo especial a la Cami, que fue casi una hermana para mi desde que entramos al DCC, siempre me contuviste y apoyaste, fuiste una muy buena amiga y compañera.

No sería yo sin el basket y tengo dos equipos muy lindos que son básicamente una familia para mí. Gracias a mi equipo de ingeniería, fue un agrado ser su capitán durante los últimos 2 años. A la selección de la U que me llevaron a competir conmigo y con ellos todos los días, empujándome a ser la mejor versión de mi como basquetbolista y como persona (Imposible nombrarlos a todos así que sientansé aludidos por completo). Gracias a Anaís, Chasca, Clemente, Pepo, Monti y Monzó, por ser mi circulo más cercano dentro y fuera del -3.

Tengo que agradecer de forma especial al Seba, que fue un hermano dentro y fuera de la cancha. Desde el primer día de la U fue un compañero esencial para estos últimos 6 años, apoyándome y felicitándome en todos mis logros y corrigiéndome en mis errores. Siempre

confiaste en todo lo que hice y fue un pilar para mi vida universitaria, gracias Seba.

Considerando la memoria misma. Quiero darle las gracias a Francisco, quien confío en mí desde el primer semestre de 3er año como su auxiliar y luego guío de gran manera para desarrollar este trabajo de título. Gracias a Matías Bahamonde por sentar las bases de un gran trabajo que yo pude extender y también por su inmensa ayuda para poder entender y desarrollar la solución que se presentará. Gracias también a la doctora Gabriela Hidalgo y todo su equipo de Teleton, tuvieron una disposición increíble a escuchar, evaluar y ayudarme a que este trabajo saliera de la mejor forma posible.

Quiero darle las gracias a la Popis, fuiste un espacio de confort durante todo el semestre, ayudándome aún cuando no lo necesitaba ni quería, sin ti la memoria hubiera sido un proceso mucho más pesado, gracias por estar siempre.

Por último quiero agradecer a todos mis entrenadores y profesores que he tenido en el colegio y la universidad. A parte de formarme académica y deportivamente, me ayudaron y guiaron a ser la persona que soy. Son parte de la razón por la que elegí Ingeniería y también la razón por la que la terminé.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

La comunicación es un proceso de transmisión de mensajes entre un emisor y un receptor. Es un eje central dentro del entendimiento entre las personas de nuestra sociedad. A lo largo de nuestro día a día ocurren muchas situaciones donde necesitamos comunicarnos, por ejemplo durante una conversación casual con un amigo, mediante una pregunta en la sala de clases o al momento de pagar en un almacén. En todos estos momentos se necesita de la comunicación para expresarnos correctamente.

Para la mayoría de las personas, esto ocurre de forma natural e intuitiva. Sin embargo, para parte de la población la tarea de comunicarse es muy compleja e incluso imposible. Este grupo lo conforman las personas que poseen algún tipo de impedimento del habla (el cual puede ser un problema tanto físico como cognitivo). Un ejemplo de un problema de salud que provoca trastornos motores en el usuario es la parálisis cerebral (PC), siendo esta la principal causa de discapacidad infantil. Los desórdenes motores de la PC frecuentemente se acompañan de alteraciones en la comunicación del usuario [1]. Para intentar mejorar la comunicación de las personas que padezcan enfermedades que afectan su comunicación, existe un área de investigación que busca crear Sistemas Alternativos y Aumentativos de Comunicación (SAAC), los cuales apuntan a aumentar la capacidad de comunicación de las personas con impedimentos del habla (de ahí el aumentativo) y compensar los obstáculos comunicativos de las personas que integran esta área con formas alternativas de interlocución (de ahí lo alternativo). [1].

Un tipo de sistema que intenta resolver este problema es el que se conoce como tablero de comunicación, en donde se presenta una matriz con iconos o letras en cada elemento. En estos, el usuario deberá seleccionar pictogramas en secuencia para así hilar oraciones. Un ejemplo de esto se puede ver en la figura mostrada a continuación:

[1] Centro Aragonés para la Comunicaciónn Aumentativa y Alternativa (ARASAAC) https://arasaac.org/aac/es

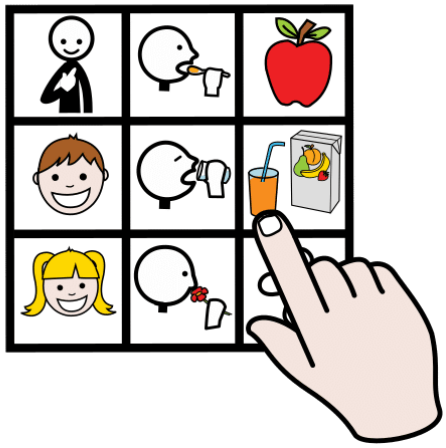


Figura 1.1: Ejemplo de tablero de comunicación. Imagen tomada del sitio de ARASAAC: https://arasaac.org/pictograms/es/16157

Este tipo de solución fue introducida por primera vez en la década del 50 por *Goldstein* [2], mientras el primer uso de estos tableros a la población que padece parálisis cerebral se vino a dar en la década siguiente, donde se publicó una guía de creación y utilización de un tablero de comunicación físico específico para personas con esta enfermedad [3]. En estos tableros se tienen distintos iconos o letras que el usuario deberá ir seleccionando para ir hilando la palabra o frase que desea transmitir. Por ejemplo, en la figura 1.1 se muestra un tablero de comunicación que tiene distintos conceptos representados por imágenes, de los cuales el usuario puede seleccionar los que necesite para comunicar una oración.

A lo largo de los años, los SAAC se han ido beneficiando de los avances tecnológicos de la época, creando sistemas digitales desde el inicio de los 70 y logrando crear sistemas portables al final de la misma década [4]. En la actualidad, se ha logrado digitalizar los tableros de comunicación alternativa, existiendo varias alternativas en el mercado como por ejemplo: *Proloquo2Go*[2], *TDSnap*[3], *TouchChat*[4], etc. Estas aplicaciones puede ser descargadas desde la *AppStore* y *PlayStore*, permitiendo que la solución sea más portable y ayude al usuario con impedimento del habla a comunicarse en cualquier lugar, sin tener como limitación algún aspecto físico-espacial. Sin embargo, estas aplicaciones poseen un alto costo. Es por esta razón que el 2022 el estudiante de la Universidad de Chile Matías Bahamonde creó un tablero de comunicación digital para niños con parálisis cerebral que sea accesible tanto de forma funcional como económica [5].

Uno de los problemas principales de estos sistemas es que, si bien asisten a mejorar la comunicación de personas con impedimentos del habla, no logran cerrar completamente la brecha comunicativa existente. Esto se demuestra analizando el dato entregado por *Newell* en 1997, el cual dice que las personas con impedimento del habla, aún utilizando SAAC, poseen una tasa de comunicación de 10 palabras por minuto, lo cual es bajo al compararse con las 150-200 palabras escritas en el mismo tiempo por una persona sin impedimentos del habla [6].

[2] https://apps.apple.com/us/app/proloquo2go/id308368164
[3] https://apps.apple.com/us/app/td-snap/id1257753762?ls=1
[4] https://apps.apple.com/us/app/touchchat-hd-aac-wordpower/id412351574

Es por esta razón que en este trabajo de título se busca mejorar la fluidez y rapidez de la comunicación de niños y niñas con parálisis cerebral que utilicen el tablero digital de comunicación alternativa creado por Bahamonde.

Se buscará simplificar la entrada de texto al tablero, utilizando un modelo de lenguaje, propio del área de procesamiento de lenguaje natural (NLP por sus siglas en inglés), que se encargue de predecir la palabra siguiente basandose en toda la oración tipeada hasta el momento. Este modelo le asignará una probabilidad a cada palabra posible dentro del vocabulario, utilizando las palabras anteriores como evidencia condicional, y luego entregará las cuatro palabras con la mayor probabilidad de que ocurran a continuación [7]. De esta forma, se busca que la predicción entregada por el sistema le sirva al usuario para comunicarse de una forma más natural y rápida. Existen aplicaciones de estos modelos a sistemas de comunicación alternativa, los cuales se revisarán en detalle en la sección 2.2. Se espera que este trabajo finalice con la integración de un modelo de predicción de palabra a un SAAC y que esta aplicación otorgue utilidad extra al usuario para su comunicación.

1.1. Objetivos

A continuación se presenta el objetivo a cumplir al término del presente trabajo de título.

1.1.1. Objetivo General

Mejorar la utilidad percibida de un tablero digital de comunicación aumentativa y alternativa mediante el diseño e implementación de un sistema de predicción de palabra siguiente.

1.1.2. Objetivos Específicos

- Evaluar el desempeño de un sistema de predicción de palabra siguiente sobre un vocabulario extremadamente limitado.
- Evaluar el desempeño del sistema al incluir datos personales al modelo, como datos horarios, localización, selección de predicciones anteriores, etc.
- Integrar los modelos de texto predictivo sin degradar la experiencia de usuario y usabilidad percibida del tablero de comunicación.
- Evaluar el tradeoff entre desempeño de un sistema de NLP y la utilidad percibida por el usuario al aplicarlo al tablero de comunicación digital.

1.2. Solución desarrollada

La solución desarrollada como resultado de este trabajo de título tiene como objetivo principal mejorar la rapidez y fluidez de la comunicación utilizando un tablero digital en base a pictogramas al implementar un sistema de predicción.

La solución desarrollada obtiene como resultados:

- Un uso más eficiente de la aplicación al integrarle cualquier modelo de predicción.

- Una usabilidad no degradada con respecto al modelo base, al integrarle un modelo de predicción.
- Una utilidad y valor percibido mayores al preferir un modelo de predicción de Markov por sobre un modelo más simple de basado en frecuencias y uno más complejo basado en transformadores.

Es por esto que se considera el trabajo como finalizado de forma exitosa.

Capítulo 2

Trabajo relacionado

2.1. Situación Actual

En el siguiente apartado se describe lo que se tiene actualmente al momento de iniciar este trabajo de título. Esto vendría siendo el tablero digital de comunicación alternativa y aumentativa creado el año pasado. También se incluye una síntesis de trabajos relacionados a esta memoria, como lo vendrían siendo técnicas de NLP aplicadas a SAAC y técnicas ocupadas para resolver la tarea de predecir palabras.

2.1.1. Tablero digital de comunicación

El tablero digital de comunicación diseñado por Bahamonde, está en estado de producción y su usabilidad fue validada por expertos del dominio de trata de pacientes, fisiatras, fonoaudiólogos y terapeutas ocupacionales de Coaniquem y Teleton. Además fue sometido a una validación de usabilidad con usuarios utilizando métricas de usabilidad de Wilson y Wixon [8] y también usando un cuestionario basado en el método NASA-TLX, la cual mide la carga mental de trabajo sobre el usuario al usar la aplicación[5]. Se obtuvo como resultado que el tablero diseñado es usable por usuarios con el perfil de cuidador de niños, y además es robusto a errores provocados por los mismos [5].

El tablero creado por Bahamonde permite a niños y niñas con parálisis cerebral comunicarse mediante la construcción de frases usando un vocabulario predefinido. El vocabulario utilizado se construyó a partir de las categorías semánticas de Feldman[6], las cuales definen las palabras básicas que debiesen ser manejadas por niños de un rango etario especificado. Bahamonde en conjunto con expertos de dominio de trata de pacientes, decretaron que el nivel que mejor se adaptaba a la solución era el que abarca desde los 2 años a los 3 años y 11 meses.

Cada una de las palabras presentes en el vocabulario tiene asociado un icono. Además, está ligada a una categorización semántica y a un color, existiendo actualmente las siguientes combinaciones: Personas/Nombres propios: amarillo, Verbos: verde, Descriptivos: azul, Sustantivos: naranja y Sociales: rosado/morado (apreciables en la figura 2.2.a). Los códigos de color de los pictogramas siguen un estándar llamado *"Clave Fitzgerald"*, creado con el fin de

[5] Método NASA-TLX, Ministerio del Trabajo y Previsión Social: https://ergomedia.isl.gob.cl/app_ergo/nasatlx/

[6] https://www.academia.edu/36432198/Categorias_Feldman

corregir errores gramáticos y sintácticos mediante una ayuda visual. Este código se utiliza desde su creación en 1926 por Edith Mansford Fitzgerald. A continuación se muestra una imagen del tablero digital producto del trabajo de título de Bahamonde.

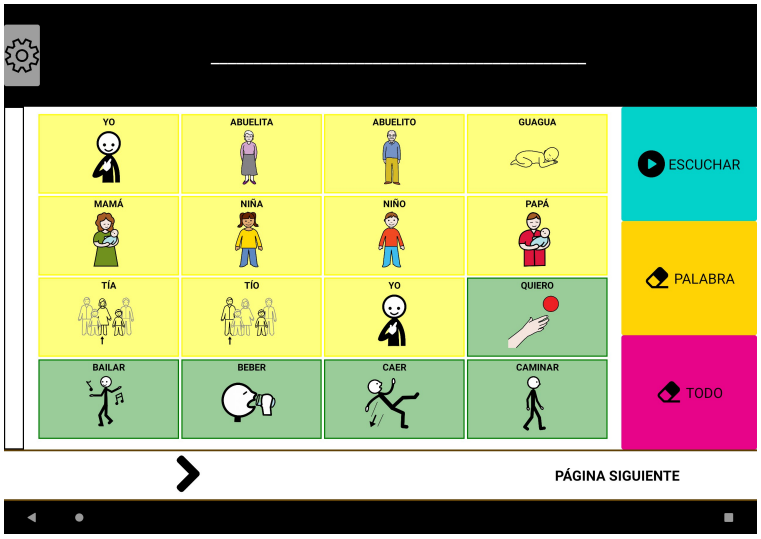


Figura 2.1: Tablero digital a modificar

Utilizando las palabras de este tablero, los usuarios pueden crear oraciones limitadas y sin conectores, lo cual difiere con las frases creadas durante una conversación normal. Por ejemplo, un niño que utilice este tablero podría usar los pictogramas para generar la frase:

Yo jugar auto de juguete

Mientras que en una conversación natural, la frase sería hilada de la siguiente forma:

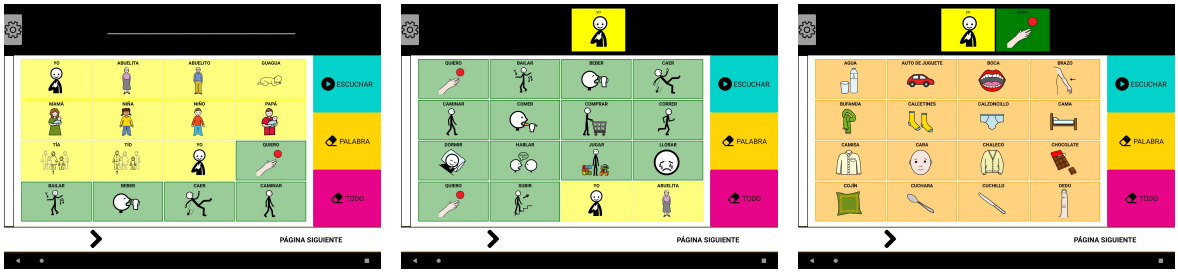
Yo juego con un auto de juguete

En el ejemplo anterior, el subrayado corresponde a preposiciones, artículos y cambios morfológicos a las palabras de la oración inicial. Esta forma de construir oraciones provoca que el lenguaje generado sea único. Es debido a esto que su modelado es un desafío a abordar durante este trabajo de título, ya que nunca se ha ocupado el vocabulario de este tablero para crear representaciones de lenguajes, provocando que las soluciones existentes en el modelado de lenguajes a base de iconos [9, 10] deban ser adaptadas en el caso de ser utilizadas.

El tablero posee actualmente un sistema de reordenación del teclado que facilita la escritura de la siguiente palabra. Si bien no intenta predecir la siguiente palabra, sí se realiza una predicción de la categoría a la cual puede pertenecer y deja en primer lugar los pictogramas de las palabras de esta categoría. Esto presenta una base para un sistema de predicción ya integrado a la aplicación y se evaluará que la solución a implementar entregue más utilidad que este sistema. Podemos ver el reordenamiento del teclado al escribir una frase en la siguiente figura:

Las reglas seguidas por este sistema de ordenación se guían por la estructura sintáctica de las oraciones:

- Si la palabra anterior es un pronombre (amarillo), las palabras que aparecerán primero serán los verbos (verdes), luego los pronombres (amarillos) y finalmente los sustantivos (naranjo).



(a) Input vacío (b) Sujeto escrito (c) Verbo escrito

Figura 2.2: Ordenación de pictogramas según categoría de la palabra anterior

- Si la palabra anterior es un verbo (verde), las palabras que aparecerán primero serán los sustantivos (naranjo), luego los verbos (verde) y finalmente los pronombres (amarillo).
- Si la palabra anterior es un sustantivo (naranjo), las palabras que aparecerán primero serán los sustantivos (naranjo), luego los verbos (verde) y finalmente los pronombres (amarillo).

2.2. Revisión bibliográfica

En el siguiente apartado se revisan las publicaciones y libros que presentan trabajo considerado como relevante para el desarrollo de la solución resultante de esta memoria.

2.2.1. Procesamiento del lenguaje natural aplicado en sistemas de comunicación alternativa y aumentativa

Los sistemas que asisten a las personas con impedimentos del habla se crean con el fin de mejorar la habilidad de las personas con impedimentos del habla para comunicarse con otras personas de forma sencilla y natural [6]. Para acelerar el progreso de los SAAC, se ha comenzado a implementar herramientas de inteligencia artificial, como por ejemplo modelos pertenecientes al área de procesamiento del lenguaje natural, con el fin de intentar reducir esta brecha comunicativa [11].

Este cruce de áreas ha llevado a incluir tareas conocidas dentro del área de NLP al mundo de los SAAC, como por ejemplo *word completition*, la cual trata sobre completar una palabra luego de que el usuario haya tipeado las letras iniciales [12]. Esto se ha aplicado a tableros digitales de comunicación que funcionan letra a letra, logrando aumentar la velocidad de comunicación del usuario y reducir el número de errores al momento de escribir [13, 14].

Otra tarea que se ha implementado dentro de los SAAC es la de *sentence construction*, la cual genera una frase coherente gramaticalmente a partir de palabras seleccionadas continuamente [12]. Esto permite una reducción del tiempo entre el input de un usuario y el output deseado, además de conseguir un mejor entendimiento por parte de ambos usuarios involucrados en la conversación [15].

2.2.2. Predicción de la siguiente palabra

La tarea que nos interesa implementar durante este trabajo de título es la de *next word prediction*, la cual intenta predecir qué palabra vendrá a continuación, usando las palabras anteriores y/o contexto del usuario como evidencia. Dentro de la disciplina de NLP existen variados acercamientos para resolver este problema. A continuación se presentan los modelos actualmente utilizados para resolver esta tarea en un contexto de comunicación normal.

Se ha utilizado, con resultados satisfactorios, modelos de N-gramas basados en Modelos oculto de Markov (HMM) [16, 17], los cuales son modelos probabilísticos que utilizan los estados anteriores para estimar la probabilidad del actual. En el caso de NLP, los modelos de N-gramas, utilizan las n-1 palabras anteriores para estimar la probabilidad de la palabra en la posición n.

Recientemente, el área de procesamiento del lenguaje natural ha adoptado dentro de sus herramientas el uso de modelos de *deep learning*. En 2017 se introdujeron los modelos que utilizan una arquitectura basada en transformadores, la cual desecha toda estructura recurrente o convolucional (previamente considerada como estado del arte para resolver problemas relacionados con la modelación del lenguaje) para depender únicamente de la atención [18]. Esta nueva arquitectura alcanzó resultados del estado del arte en tareas clásicas de NLP, entre ellas *next word prediction*.

Aplicando la arquitectura recientemente descrita, la representación de codificador bidireccional basada en transformadores (BERT por sus siglas en ingles), consiguió resultados superiores a todas las representaciones de modelos de lenguaje anteriormente existentes en 2018 [19]. Se considera que este tipo de modelos corresponden al estado del arte en esta disciplina.

2.2.3. Predicciones en sistemas de comunicación alternativa y aumentativa

Existen diversas formas en las que los sistemas de comunicación alternativa y aumentativa se han beneficiado de la predicción de palabras. Algunos sistemas existentes aprovechan el contenido sintáctico y semántico de una oración al crear predicciones de pictogramas, creando así las gramáticas semánticas, un sistema que logra aprovechar ambas estructuras para generar predicciones más adecuadas y precisas [20]. Una integración de estos sistemas hacia en un tablero de pictogramas para personas con trastornos del espectro autista (TEA) obtiene como resultados unas predicciones que ayudan al usuario generar mensajes sintáctica y semánticamente correctos [21].

Estudios recientes utilizan modelos de lenguaje estadísticos basados en n-gramas para predecir el pictograma siguiente, en diversos sistemas de comunicación alternativa y aumentativa, como por ejemplo en uno especializado para personas con TEA [22]. En el tablero presentado en ese trabajo, se utiliza la cantidad de veces que un pictograma ha sido utilizado, en conjunto con la categoría sintáctica asociada del mismo para realizar una predicción. Siguiendo la misma linea, *Eugénio* es un sistema de AAC en portugués que además de aplicar un sistema de recomendación en base a n-gramas, añade también como información previa

la localización del individuo, consiguiendo así unas predicciones más precisas [23].

Con el fin de lograr adecuar el estado del arte en el modelado de lenguaje al contexto de los SAAC, es que se realizó una adaptación del modelo encoder-decoder BERT para la tarea de predecir el pictograma siguiente. Inicialmente se relaciona cada pictograma con un concepto de WordNet [24] (diccionario léxico online del inglés para uso computacional), y usando estos se intenta predecir el siguiente concepto para luego transformarlo de vuelta al pictograma y entregárselo a la aplicación [10]. Este modelo, llamado pictoBERT, logra de una manera efectiva aplicar un sistema del estado del arte de predicción de palabra a un contexto de uso enmarcado dentro del área de los SAAC. Sin embargo, solamente realiza una evaluación intrínseca, mostrando desempeño técnico del modelo independientemente de su aplicación. Este trabajo no presenta resultados con respecto a como los usuarios perciben la solución al momento de aplicarla a un tablero digital de comunicación. Es parte del trabajo de título presente, el realizar una evaluación intrínseca y extrínseca, debido a su importancia para determinar la utilidad de una aplicación [7].

Capítulo 3

Desarrollo de la solución

La solución desarrollada para resolver el problema planteado en este trabajo de título tiene como objetivo diseñar un sistema de predicción de la palabra siguiente, tal que al integrarlo en un tablero digital de comunicación alternativa y aumentativa le otorgue utilidad extra.

Lo que se busca es modelar probabilísticamente el lenguaje compuesto por combinaciones de 0 o más palabras provenientes del vocabulario del tablero, las que en conjunto constituyen a una oración. Este vocabulario poseía inicialmente 78 conceptos, cada uno acompañado de un pictograma que lo representa. Durante la realización del trabajo de título, en una de las reuniones que se tuvo con Gabriela Hidalgo, fisiatra de Teletón[7] y Coaniquem[8], se comentó que la lista de pictogramas incluidos en el tablero no daba abasto para poder cubrir todas las necesidades de comunicación de un niño con PC. En un esfuerzo de poder cubrir mejor estas necesidades es que se agregó el pictograma de *"dolor"*, debido a que permitía comunicar sensaciones físicas importantes para pacientes que padeciesen PC. Este pictograma se incluyó luego de una validación semántica, que la palabra elegida representase correctamente el significado de dolor, y semiótica, que el icono elegido representase correctamente el concepto de dolor, con expertos de dominio. El vocabulario completo, categorizado por categoría sintáctica (categorías de Feldman), se puede encontrar en la figura A.1 del anexo.

Un modelo de lenguaje corresponde a una distribución de probabilidad que le asigna una probabilidad a cualquier frase escrita utilizando palabras presentes en el vocabulario. De tal forma, permite estimar la probabilidad de la palabra en posición i de una frase, y por ende, se puede usar para calcular la probabilidad de la palabra $n+1$ en una oración de largo n. En otras palabras, un modelo de lenguaje que entrega una distribución de probabilidad que permite estimar la palabra que se escribirá a continuación, dadas las anteriores como contexto. Formalmente, la tarea a resolver se define como:

$$\mathbb{P}(w_{n+1} \mid w_{0:n}) = \{P(w_{n+1}|w_{0:n}) \ , \ \forall w_{n+1} \in V\}$$

$$\text{donde: } w_i \in V \ \forall i \in \{0, 1, ..., n\}\}$$
$$w_{0:n} = w_0, w_1, ..., w_n$$

Donde $\mathbb{P}(w_{n+1} \mid w_{0:n})$ es un vector de valores que representa una distribución de proba-

[7] https://www.teleton.cl
[8] https://coaniquem.cl/es/

bilidad condicional. Esto es una asignación de probabilidad para cada valor posible de la variable w_{n+1} condicionada por $w_{0:n}$.

Para la aplicación generada en este trabajo de título, los modelos deben usar la distribución de probabilidad expuesta para entregar un conjunto que contenga las 4 palabras con mayor probabilidad de ser elegidas para completar la oración. Se eligió este número debido a que el tablero posee filas de 4 elementos y se quiere mantener este estándar para no sacrificar usabilidad cambiando la aplicación de manera innecesaria. Formalmente, lo que se debe estimar es:

$$\begin{aligned} pred &= \{p1, p2, p3, p4\} \\ p1 &= max\{\mathbb{P}(w_{n+1} \mid w_{0:n})\} \\ p2 &= max\{\mathbb{P}(w_{n+1} \mid w_{0:n}) \mid p2 \neq p1\} \\ p3 &= max\{\mathbb{P}(w_{n+1} \mid w_{0:n}) \mid p3 \neq p1, p2\} \\ p4 &= max\{\mathbb{P}(w_{n+1} \mid w_{0:n}) \mid p4 \neq p1, p2, p3\} \end{aligned} \tag{3.1}$$

Previo a estimar la palabra que se escribirá a continuación, los modelos deben pasar por una etapa de entrenamiento. En esta etapa, los modelos usarán un conjunto de oraciones prediseñado para aprender las relaciones existentes entre las palabras que componen una frase. Lo primero que se debe hacer, entonces, es construir un corpus compuesto por una gran cantidad de oraciones para que los modelos puedan aprender la mayor cantidad de información posible de ellas y así lograr generar predicciones más precisas.

Para la construcción de este corpus se requiere de la ayuda de expertos del dominio. Estos deben crear oraciones sintáctica y semánticamente correctas que simulen las que podría escribir un niño con parálisis cerebral, usando el tablero digital. Esto es necesario para que el modelo pueda aprender las relaciones especificas que existen entre palabras del lenguaje usado en la aplicación y pueda así realizar predicciones aterrizadas a este contexto.

Se envió un formulario de Google para poder construir el corpus de oraciones para la etapa de entrenamiento del modelo. En este se pedía a cada experto que ingresara los siguientes datos: género del niño, edad del niño, hora del registro, título del primer pictograma, título del segundo pictograma, título del tercer pictograma, título del cuarto pictograma, título del quinto pictograma. Los primeros datos corresponden a la caracterización del paciente, lo que servirá para realizar predicciones adecuadas a cada individuo. Por otro lado, la hora servirá para modificar las predicciones realizadas según la hora del día en la que se utilizará la aplicación. Finalmente, los 5 pictogramas seleccionados constituirán a una frase escrita por un niño que use la aplicación.

El uso de este formulario fue descartado en pos de conseguir un corpus más grande de datos, se decidió así tomar un acercamiento más directo, en donde la doctora Gabriela Hidalgo escribió 104 oraciones utilizando 3 a 5 pictogramas del vocabulario para su construcción.

Para conseguir mayor variedad de pictogramas dentro del corpus se realizó un proceso de *data augmentation*, en donde se reemplazaron ciertas palabras por sinónimos dentro del vocabulario, consiguiendo así 624 frases. Para realizar esto, se procesaron iterativamente todas

las oraciones dentro del corpus de entrenamiento y se volvieron a agregar 4 veces, pero cambiando cada palabra que la compone por un sinónimo. De esta forma se consiguió aumentar el tamaño del corpus y también la información que cada modelo podrá aprender de él. Este fue el corpus final utilizado para entrenar los modelos.

A continuación se presentan en detalle los tres modelos creados.

3.1. Modelo de frecuencias

El primer modelo de lenguaje creado es uno que calcula la probabilidad de cada palabra usando su frecuencia relativa con respecto a un corpus. Formalmente:

$$P_{freq}(w_{n+1} \mid w_{0:n}) = \frac{conteo(w_{n+1})}{|Cw|} \tag{3.2}$$

donde: Cw son todas las palabras en el corpus de entrenamiento

Si ocupamos la ecuación 3.2 para calcular la probabilidad de todas las palabras dentro del vocabulario, conseguimos una distribución de probabilidad que nos permite representar el modelo de frecuencias de la siguiente manera:

$$\mathbb{P}_{freq}(w_{n+1} \mid w_{0:n}) = \left\{ \frac{conteo(w_{n+1})}{|Cw|} \;\mid\; \forall w_{n+1} \in V \right\} \tag{3.3}$$

De esta forma, las palabras más frecuentemente usadas serán las que tengan la probabilidad más alta dentro de la distribución.

En un intento de generar predicciones más útiles sin complejizar de sobremanera el modelo, se reutilizó la idea que aplicó Bahamonde en su sistema de reordenación de teclado (para más detalles ver 2.1.1). Siguiendo la misma metodología, se utilizó la categoría sintáctica de la palabra anterior para filtrar el conjunto de probabilidades y entregar predicciones compuestas únicamente por palabras que pertenezcan a la categoría siguiente. Ya fue demostrado por Bahamonde que realizar esta división es beneficioso para que los usuarios puedan escribir de forma más eficiente y eficaz [5], por ende, se aprovechará esta información para generar recomendaciones cumpliendo el mismo fin.

Podemos notar que este modelo de lenguaje no utiliza el contexto anterior ni posterior a la palabra. Solamente utiliza la cantidad de veces que esta aparece dentro de las oraciones del corpus, en conjunto con la categoría de la palabra anterior, para calcular su probabilidad. Es debido a esto que podemos decir que la probabilidad de la palabra siguiente, depende solamente de la palabra escrita con anterioridad y es independiente al contexto previo a esta.

Considerando estos puntos, la nueva estimación de la probabilidad de una palabra, queda definida de la siguiente forma:

$$P_{freq}(w_{n+1}|w_{0:n}) = P_f(w_{n+1}|w_n) \tag{3.4}$$

$$P_{freq}(w_{n+1}|w_n) \begin{cases} \frac{conteo(w_{n+1})}{|Cw_{amarillo}|} \text{ ; si } categoria(w_n) = \text{None} \\ \frac{conteo(w_{n+1})}{|Cw_{verde}|} \text{ ; si } categoria(w_n) = \text{“amarillo”} \\ \frac{conteo(w_{n+1})}{|Cw_{naranjo}|} \text{ ; si } categoria(w_n) \in \{\text{“verde”, “naranjo”}\} \end{cases} \tag{3.5}$$

donde $Cw_{amarillo}$ representa todas las palabras de la categoría sujetos, Cw_{verde} los verbos y $Cw_{naranjo}$ los sustantivos.

Usando estas nuevas ecuaciones que permiten estimar la probabilidad de la siguiente palabra, podemos definir nuevamente la distribución de probabilidad para todo el vocabulario. De tal forma, el espacio de donde la aplicación extraerá los 4 valores más grandes se define como:

$$\mathbb{P}_{freq}(w_{n+1} \mid w_n) = \{Pfreq(w_{n+1} \mid w_n) \mid \forall w_{n+1} \in V\} \tag{3.6}$$

Utilizando estas reglas para calcular las probabilidades de la palabra siguiente, cada predicción se calcula de la siguiente manera:

$$p_i = max(\mathbb{P}_{freq}(w_{n+1}|w_n)) \; \forall i \in \{1, 2, 3, 4\} \tag{3.7}$$

en donde w_{n+1} es la palabra que posiblemente estará escrita a continuación y w_n es la última palabra escrita por el usuario.

Cabe recalcar, que tal como se explico en 3.1, al entregar la predicción $i + 1$, no se debe considerar la predicción i como una palabra posible a predecir.

3.2. Modelo de estados de Markov

El segundo modelo implementado corresponde a un modelo basado en estados, donde los estados anteriores son usados como evidencia para predecir el valor del actual.

Conocido dentro del estudio de probabilidades y procesos estocásticos como proceso de Markov [7], el modelo se compone por una serie de estados variables catalogados como X_i, donde $i \in \{0, ..., t\}$ representa a la posición del estado dentro de la cadena, usualmente una medida de tiempo discreta. La distribución de probabilidad de cada estado utiliza como información todos los estados que lo preceden en la cadena. Formalmente:

$$\mathbb{P}(X_t) = \mathbb{P}(X_t \mid X_{0:t-1})$$

El problema de un modelo de este estilo, es que el crecimiento de la información de dependencia se vuelve insostenible a medida que se avanza en el tiempo. Es por esto que se suelen hacer suposiciones de independencia que permiten simplificar el modelo, conocidas como suposiciones de Markov, en donde asumimos que el estado actual solamente depende

de un número finito de estados precedentes [25].

Realizar una suposición de este tipo nos posibilita calcular la probabilidad de forma efectiva sin perder tanta información de relevancia. Utilizaremos en este trabajo procesos de Markov que apliquen suposiciones de primer y segundo orden. Esto significa que utilizan 1 o 2 estados anteriores como evidencia, respectivamente.

Proceso de Markov de primer orden:

$$\mathbb{P}(X_t) = \mathbb{P}(X_t \mid X_{0:t-1}) = \mathbb{P}(X_t \mid X_{t-1}) \tag{3.8}$$

Proceso de Markov de segundo orden:

$$\mathbb{P}(X_t) = \mathbb{P}(X_t \mid X_{0:t-1}) = \mathbb{P}(X_t \mid X_{t-1}, X_{t-2}) \tag{3.9}$$

De esta forma, el estado actual es independiente de todo el contexto anterior precedido por tres o más estados. En particular, para el caso de un proceso de Markov de primer orden, el estado actual depende solamente del estado anterior, mientras que para el proceso de segundo orden, dependerá de los dos estados que lo preceden.

Un modelo de cadenas de Markov aplicado a NLP se conoce como un modelo de N-gramas, en donde los estados son cada posición i dentro de una oración y las variables de cada estado son las palabras escritas en tal posición. Por ejemplo:

yo	quiero	jugar	muñeca	mamá
X_0	X_1	X_2	X_3	X_4

En este caso para el estado X_2 el valor que toma es *jugar*, la cual es una palabra dentro de las 79 posibles del vocabulario.

Al aplicar las suposiciones de Markov a los modelos de N-gramas, limitamos las N palabras anteriores que son usadas como evidencia a un número finito y definido de palabras. En nuestro caso particular, utilizaremos los modelos de unigrama, bigrama y trigrama, en donde se realiza una suposición de Markov de primer orden para calcular la distribución de probabilidad en los bigramas y una de segundo orden para los trigramas [25]. Considerando estas suposiciones, las probabilidades estimadas para la palabra siguiente, considerando la frase escrita hasta el momento, serían las siguientes:

Probabilidad de la siguiente palabra en Unigramas

$$P_{uni}(w_{n+1} \mid w_{0:n}) = P(w_{n+1}) \tag{3.10}$$

Probabilidad de la siguiente palabra en Bigramas

$$P_{bi}(w_{n+1} \mid w_{0:n}) = P(w_{n+1} \mid w_n) \tag{3.11}$$

Probabilidad de la siguiente palabra en Trigramas

$$P_{tri}(w_{n+1} \mid w_{0:n}) = P(w_{n+1} \mid w_n, w_{n-1}) \tag{3.12}$$

donde $w_{-1} =$ " %" y $w_{-2} =$ " % %" son caracteres especiales necesarios para definir de forma válida las probabilidades iniciales para los modelos de bigramas ($P(w_0) = P(w_0 \mid$ " %")) y trigramas ($P(w_0) = P(w_0 \mid$ " % %", " %")).

Las suposiciones de independencia de Markov nos permiten estimar la palabra siguiente utilizando solamente las dos palabras que se escribieron con anterioridad. Considerando que el vocabulario se compone de 79 palabras, para una oración de 5 palabras, logramos reducir el espacio muestral de los estados anteriores desde 79^5 (3 mil millones de combinaciones posibles) a solamente 79^2 (6.241 combinaciones posibles). Esta simplificación, si bien compromete evidencia previa que no se utiliza, sigue siendo un modelo de lenguaje con un buen desempeño de forma particular para esta aplicación.

Procederemos ahora a estimar las probabilidades para cada uno de los modelos presentados, para eso utilizaremos el estimador de máxima similitud (MLE por sus siglas en ingles). MLE estima los parámetros del modelo tal que la probabilidad se maximice. En el caso de los modelos de N-gramas, la probabilidad que se busca maximizar es:

$$\begin{aligned} P(w_{0:n}) &= P(w_1) * P(w_2 \mid w_1) * P(w_3 \mid w_{1:2}) * ... * P(w_n \mid w_{1:n-1}) \\ &= \prod_{i=1}^{n} P(w_i \mid w_{1:i-1}) \end{aligned} \tag{3.13}$$

Lo que constituye a la probabilidad de la cadena de palabras $w_1, w_2, ..., w_n$ utilizando el modelo de N-gramas.

Conseguimos el estimador MLE para los párametros del modelo de N-gramas al computar el conteo de esa secuencia particular dentro un corpus y normalizándolos tal que quede entre 0 y 1 para poder así conseguir una distribución probabilidad válida y bien definida. Al hacerlo conseguimos la siguiente estimación para la secuencia $w_{1:n}$ [7]:

$$P(w_n \mid w_{n-N+1:n-1}) = \frac{conteo(w_{n-N}, w_{n-(N-1)}, ..., w_{n-1}, w_n)}{conteo(w_{n-N}, w_{n-(N-1)}, ..., w_{n-1})} \tag{3.14}$$

donde: N corresponde al modelo de N-gramas utilizado

De esta forma, conseguimos una ecuación que nos permite estimar la probabilidad de los modelos de N-gramas, utilizando la frecuencia relativa conseguida al dividir el conteo de una secuencia particular por el conteo de su prefijo.

Si utilizamos la ecuación 3.14 para redefinir las probabilidades para los tres modelos de Markov a utilizar (mostradas en 3.10 3.11 3.12), podemos conseguir lo siguiente:

$$P_{tri}(w_{n+1} \mid w_n, w_{n-1}) = \frac{conteo(w_{n+-1}, w_n, w_{n+1})}{conteo(w_{n-1}, w_n)} \tag{3.15}$$

$$P_{bi}(w_{n+1} \mid w_n) = \frac{conteo(w_n, w_{n+1})}{conteo(w_n)} \tag{3.16}$$

$$P_{uni}(w_{n+1}) = \frac{conteo(w_{n+1})}{\mid C_w \mid} \tag{3.17}$$

donde Cw son todas las palabras en el corpus de entrenamiento

De esta forma logramos estimar la probabilidad de la siguiente palabra, considerando la frase escrita anteriormente, utilizando las oraciones presentes en el corpus de entrenamiento.

Ahora que tenemos definido correctamente la forma de calcular la probabilidad de la palabra siguiente, en cada modelo, procederemos a describir como construir la distribución de probabilidad para cada uno de ellos.

Distribución de probabilidad para modelo de unigramas:

$$\mathbb{P}_{uni}(w_{n+1} \mid w_{0:n}) = \{P_{uni}(w_{n+1}) \mid \forall w_{n+1} \in V\} = \mathbb{P}_{uni}(w_{n+1}) \tag{3.18}$$

Distribución de probabilidad condicional para modelo de bigramas:

$$\mathbb{P}_{bi}(w_{n+1} \mid w_{0:n}) = \{P_{bi}(w_{n+1}, w_n) \mid \forall w_{n+1} \in V\} = \mathbb{P}_{bi}(w_{n+1} \mid w_n) \tag{3.19}$$

Distribución de probabilidad condicional para modelo de trigramas:

$$\mathbb{P}_{tri}(w_{n+1} \mid w_{0:n}) = \{P_{tri}(w_{n+1}, w_n, w_{n-1}) \mid \forall w_{n+1} \in V\} = \mathbb{P}_{tri}(w_{n+1} \mid w_n, w_{n-1}) \tag{3.20}$$

Se puede notar que para todos los modelos es cierto que en el caso de que alguna combinación de palabras no exista dentro del corpus de entrenamiento, la probabilidad estimada será 0. Esto es un problema particularmente para el caso de los modelos de bigrama y trigrama, debido a que muchas de las 6.241 o 493.039 combinaciones posibles de palabras no estarán presentes en el corpus, debido a que juntas no forman una frase léxica ni gramaticalmente correcta, haciéndolas improbables. Esto quiere decir que para la distribución de probabilidad muchos valores serán 0, produciendo un problema conocido como escasez de datos o *sparseness*. Esta característica de escasez invalida la probabilidad de cualquier oración en donde se presente una combinación de palabras que no se haya visto antes, particularmente anulando la probabilidad de cualquier oración que no se presente en el corpus de entrenamiento. Esto resulta en un modelo de lenguaje que no generaliza bien el problema y que no tendrá un buen desempeño en cualquier aplicación donde se integre [25]. Para el caso de los unigramas, para que una palabra tenga probabilidad 0, debe no ser usada nunca dentro del corpus de entrenamiento. Considerando las 79 palabras posibles, hay solamente 11 que nunca fueron usadas por la doctora Hidalgo en la construcción de oraciones, por ende el problema de *sparsness* no está tan presente en este modelo. Podemos asumir entonces, que para todas las 68 palabras presentes, la probabilidad será no nula, independiente de su contexto.

Con el fin de evitar que muchos valores de la distribución de probabilidad sean 0, se realizará una técnica de interpolación, en donde se ponderará la probabilidad entregada por cada modelo por un hiperparámetro distinto. Al realizar esto, permitimos que combinaciones de palabras inexistentes dentro del corpus de entrenamiento tengan asignada una probabilidad distinta de 0, permitiéndonos diferir entre dos combinaciones poco probables utilizando la

frecuencia individual de sus componentes[9].

Considerando esto, la probabilidad de interpolación de los tres modelos se define de la siguiente forma:

$$P_{int}(w_{n+1} \mid w_n, w_{n-1}) = \lambda_{tri} * P_{tri}(w_n, w_{n-1}) + \lambda_{bi} * P_{bi}(w_n) + \lambda_{uni} * P_{uni} \tag{3.21}$$

Usando esta definición, podemos definir la distribución de probabilidad para el modelo completo mediante la siguiente fórmula:

$$\mathbb{P}_{int}(w_{n+1} \mid w_n, w_{n-1}) = \{P_{int}(w_{n+1} \mid w_n, w_{n-1}) \mid \forall w_{n+1} \in V\} \tag{3.22}$$

Donde λ_{tri},λ_{bi} y λ_{uni} pasan a ser hiperpárametros del modelo. Estos deben ser entrenados para maximizar alguna métrica asociada con la precisión del modelo, para nuestro caso particular el valor a maximizar fue el de Top-4, usado en otras aplicaciones de predicción de siguiente palabra en AAC [10]. (Para mayor detalle referirse a la sección 5.2). Al realizar el cálculo de la métrica recientemente referida 100 veces y promediando su resultado, los valores de cada λ que maximizan el rendimiento del modelo son los siguientes:

$$\lambda_{tri} = 0.65\ ,\ \lambda_{bi} = 0.3\ ,\ \lambda_{uni} = 0.05 \tag{3.23}$$

Utilizando estas reglas para calcular las probabilidades de la palabra siguiente, cada predicción se calcula de la siguiente manera:

$$p_i = max(\mathbb{P}_{int}(w_{n+1} \mid w_n, w_{n-1}))\ ,\ \forall i \in 1, 2, 3, 4 \tag{3.24}$$

en donde w_n es la última y w_{n-1} la penúltima palabra escrita por el usuario.

Cabe recalcar, que tal como se explico en 3.1, al entregar la predicción $i + 1$, no se debe considerar la predicción i, ni ninguna de las anteriores, como una palabra posible a predecir.

3.3. Modelo de transformadores

El tercer y último modelo que se implementó durante el desarrollo de este trabajo de título, es uno que utiliza una arquitectura de transformadores para crear un modelo no recurrente que consigue resultados del estado del arte en múltiples tareas de NLP, en particular en predecir la palabra siguiente, abordada en este trabajo de título (para más detalles ver sección 2.2.2).

Estos modelos ocupan un diseño de encoder-decoder (ver figura 3.1) igual que las, previamente consideradas estado del arte, redes *long short term memory loss* (LSTM) [26].

[9] Course notes for NLP by Michael Collins, Columbia University http://www.cs.columbia.edu/~mcollins/cs4705-spring2020/

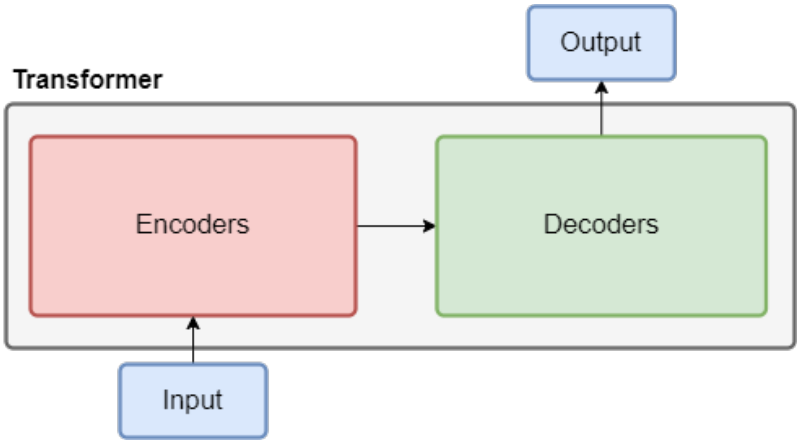


Figura 3.1: Arquitectura transformer

En el caso de BERT [19] (modelo que adaptaremos para nuestra solución), cada bloque se compone de 6 encoders o decoders respectivamente, los cuales están internamente conectados de forma secuencial, tal que cada elemento de cada tipo esta conectado al predecesor del mismo tipo. Además cada decoder esta conectado al output generado por la última componente en la stack de encoders[10].

De la misma forma que las LSTM, BERT utiliza la atención para extraer y utilizar información de contextos de largo arbitrariamente grande, pero a diferencia que las redes antiguas, no la pasa por capas recurrentes intermediarias. Este proceso recurrente provoca pérdida de información y de rendimiento durante el entrenamiento. En vez de esto, utiliza capas de atención para traspasar la información secuencialmente a través de los bloques que componen un transformer. A continuación se muestra una capa de atención con cuatro componentes:

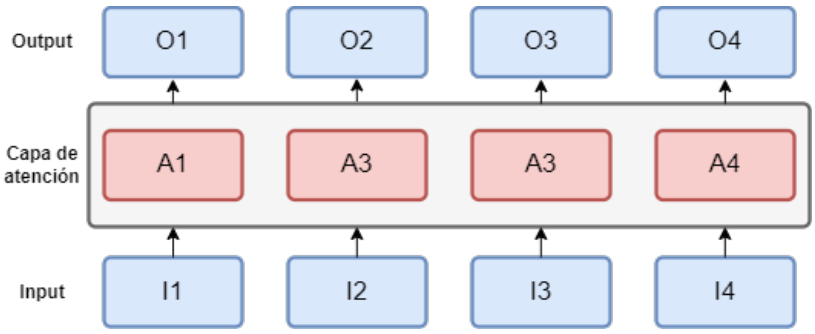


Figura 3.2: Capa de atención

Este mecanismo permite al modelo comparar el elemento actual con los que lo preceden, utilizando un componente de atención para determinar cuales son los más relevantes para el contexto del elemento enfocado y así generar el output. En particular cada output O_i se genera de la siguiente forma:

[10] Course notes for NLP by Michael Collins, Columbia University http://www.cs.columbia.edu/~mcollins/cs4705-spring2020/

$$O_i = \sum_{j=1}^{i} a_{ij} I_j \tag{3.25}$$

donde el valor a_{ij} corresponde a un ponderador de atención para el par i,j [7]. Tanto la forma en la que se calcula a_{ij}, como el origen de los valores I_j y O_i se explicarán a continuación.

Lo que diferencia a BERT de las LSTM y los anteriores modelos basados en redes recurrentes, es la forma en la que utiliza la atención. Este modelo implementa tres capas de atención distinta. Una capa de atención encoder-decoder que conecta cada componente decoder con el output final de la stack de encoders (esta ya estaba presente en los modelos antiguamente mencionados), una capa de auto atención (*self-attention layer* en inglés) en los encoders y una capa de auto atención en los decoders [18].

En cada una de estas capas se aplica la misma formula descrita en 3.25. Lo que varía entre las distintas capas es el rol que toma cada elemento al calcular la atención. Los roles que puede tomar una variable son los siguientes:

- **Query:** Elemento que se esta revisando en ese momento, se denomina foco de atención.
- **Llave:** Elemento que se esta usando como contexto para evaluar la *query* actual.
- **Valor:** Usado para computar el output del foco de atención usando las llaves como contexto.

En el caso de los transformadores, las llaves y los valores siempre son representaciones que se obtienen del mismo lugar [7].

Al conocer los roles que puede tomar cada elemento en el cálculo de la atención mostrado en 3.25, podemos definir en detalle como se computa la atención en las capas de los modelos basados en transformadores.

Considerando q_i como el elemento i proveniente de la *Query* actual, k_j como el elemento j proveniente de la Llave actual y v_j como el elemento j proveniente del Valor actual, podemos definir el componente de atención a_{ij} (presentado en 3.25) como:

$$a_{ij} = \frac{exp(q_i \cdot k_j)}{\sum_{k=1}^{p} exp(q_i \cdot k_j)} \tag{3.26}$$

Este valor presentado en 3.26, representa la relevancia que tiene el componente j de las llaves para generar el componente i de las queries [7]. El índice superior p de la sumatoria, presente en el denominador, varía según la capa de atención se este utilizando.

Considerando esta definición de atención presentada en la ecuación 3.26 y los roles que puede tomar cada elemento dentro del cálculo de la atención, es que se vuelve a definir la ecuación de atención descrita en 3.25 como:

$$O_i = \sum_{j=1}^{p} a_{ij} v_j \tag{3.27}$$

En la ecuación 3.27 es que se muestra como los valores $v_1, v_2, ..., v_p$, en conjunto de la relevancia que tiene la llave i para la query j permiten computar el valor del elemento i del output.

Veremos como se definen las queries, llaves y valores para cada capa de atención presente en los modelos de transformadores.

Para la capa de atención encoder-decoder (ver figura 3.3), las *queries* vienen de la capa decoder anterior, mientras que las llaves y los valores vienen desde el output del stack completo de encoders. Para esta capa el valor que toma p en la ecuación 3.27 es igual al tamaño de elementos del input. En la figura 3.3, vendría siendo cuatro.

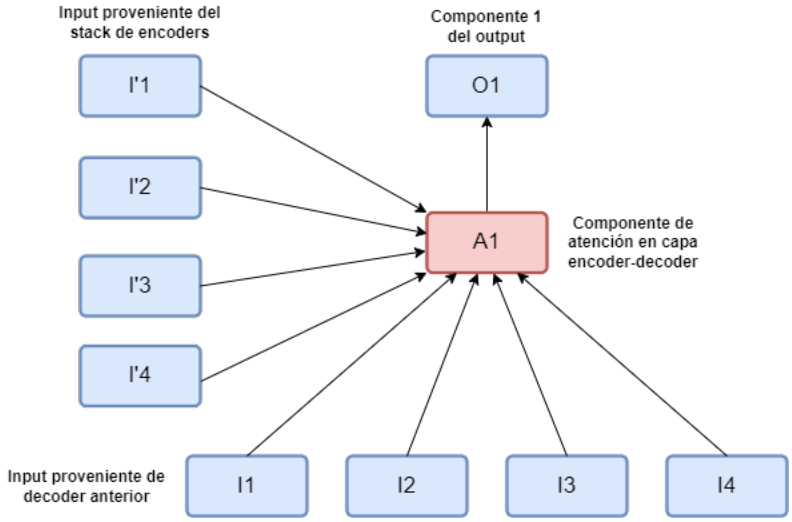


Figura 3.3: Componente de capa de atención encoder-decoder

Para ambas capas de auto atención presentes en el modelo, las queries, llaves y valores provienen del mismo output producido por el componente anterior del mismo tipo.

En la capa de auto atención de los encoders (ver figura 3.4), cada componente Ai computa la atención considerando como contexto toda la secuencia de output generado por el encoder anterior. De esta forma permite un comportamiento bidireccional, utilizando las palabras anteriores y las siguientes para predecir la actual. Para esta capa el valor que toma p en la ecuación 3.27 es igual al tamaño de elementos del input. En la figura 3.4, vendría siendo cuatro.

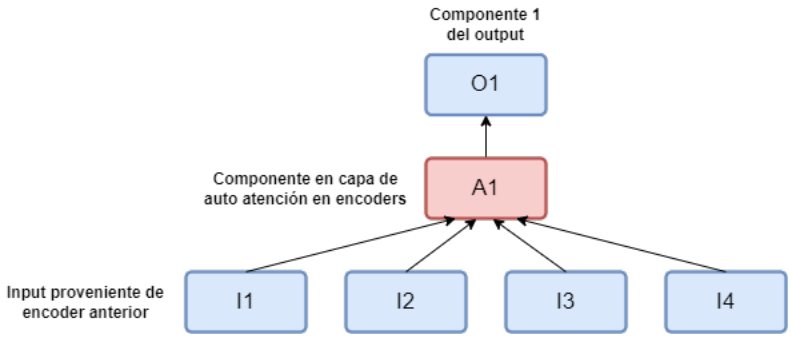


Figura 3.4: Componente en capa de auto atención de encoder

Similarmente, para la capa de auto atención de los decoders (descrita en la figura 3.5) se utiliza el output del decoder anterior como contexto para computar el valor de atención para cada posición. Sin embargo, a diferencia que los encoders, cada componente solamente puede usar como contexto el input hasta la posición actual. Esto quiere decir que las palabras siguientes no son consideradas en este caso [18]. Para esta capa el valor que toma p en la ecuación 3.27 es igual a la posición del elemento del output que se este computando. En la figura 3.5, vendría siendo uno y tres respectivamente.

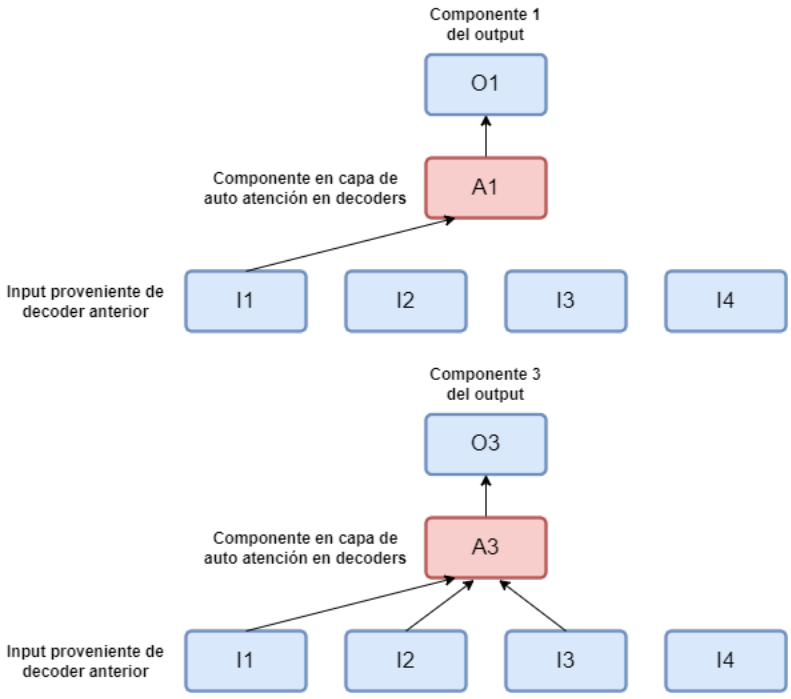


Figura 3.5: Componentes 1 y 3 en capa de auto atención de decoder

Combinando capas de atención con capas de *feed-forward* es que el modelo construye bloques de encoders y decoders. Al concatenarlos entre sí se logra la arquitectura de transformadores, mostrada en 3.1.

Luego de saber como es que BERT realiza la transformación desde input a output, debemos saber como es que se representa particularmente cada cadena entregada y recibida.

El input debe seguir un formato específico para que el modelo logré entenderlo y a partir de este, generar un output. En particular para BERT, uno de los tokens del input debe ser sustituido por "[MASK]" para que el modelo lo prediga. Mientras que el resto de palabras, tanto a la derecha como a la izquierda, serán usados como contexto para poder estimar la probabilidad de este token enmascarado. Es debido a esta representación que el modelo implementado se conoce como un modelo de lenguaje de máscaras (MLM por sus siglas en ingles).

En el modelo base, el autor presentó una estrategia para enmascarar el input, en donde se elige un 15 % de las posiciones de forma aleatoria. En caso de que el elemento i se escoja, ese token es reemplazado por: el token "[MASK]" un 80 % de las veces, por un token seleccionado al azar un 10 % de las veces y por el mismo token en esa posición un 10 % de las veces (dejando la oración igual a la original) [19].

Sin embargo, esta estrategia no se adapta de forma correcta a nuestra aplicación. Esto se debe principalmente a dos puntos:

- No tenemos suficientes datos de entrenamiento. En particular, cada una de las frases se compone de 3 palabras en promedio. Si es que enmascaramos un 15 % de los datos, existirían muchas frases sin token especial "[MASK]" y el modelo no haría predicciones para estas.
- No queremos predecir cualquier token dentro de una oración, sino que queremos siempre predecir el siguiente dados los anteriores como contexto.

Es por estas razones que se aplicó la siguiente metodología para procesar el input:

Algorithm 1 Formateo de dataset

for *oracion* en *dataset* **do**
    Separar *oracion* en tokens ▷ Los tokens son palabras en este caso
    Elegir una posicion i aleatoria entre 0 y el largo de *oracion*
    Guardar palabra en posicion i
    Truncar *oracion* hasta la posicion i
    Agregar $''[MASK]''$ al final de la *oracion*
    Juntar *oracion*
    Codificar *oracion* agregando $''[CLS]''$ al inicio y $''[SEP]''$ al final
    Guardar *oracion* codificada
    Guardar *attention mask* de la *oracion* ▷ Diferencia palabras de padding
end for

Los tokens "[CLS]" y "[SEP]" son caracteres especiales que le indican al modelo en que posición se empieza la frase y en cual se termina. Luego de tener todo el dataset formateado según lo descrito en el algoritmo 1, la arquitectura completa del modelo aplicado a nuestra aplicación queda definida de la siguiente forma:

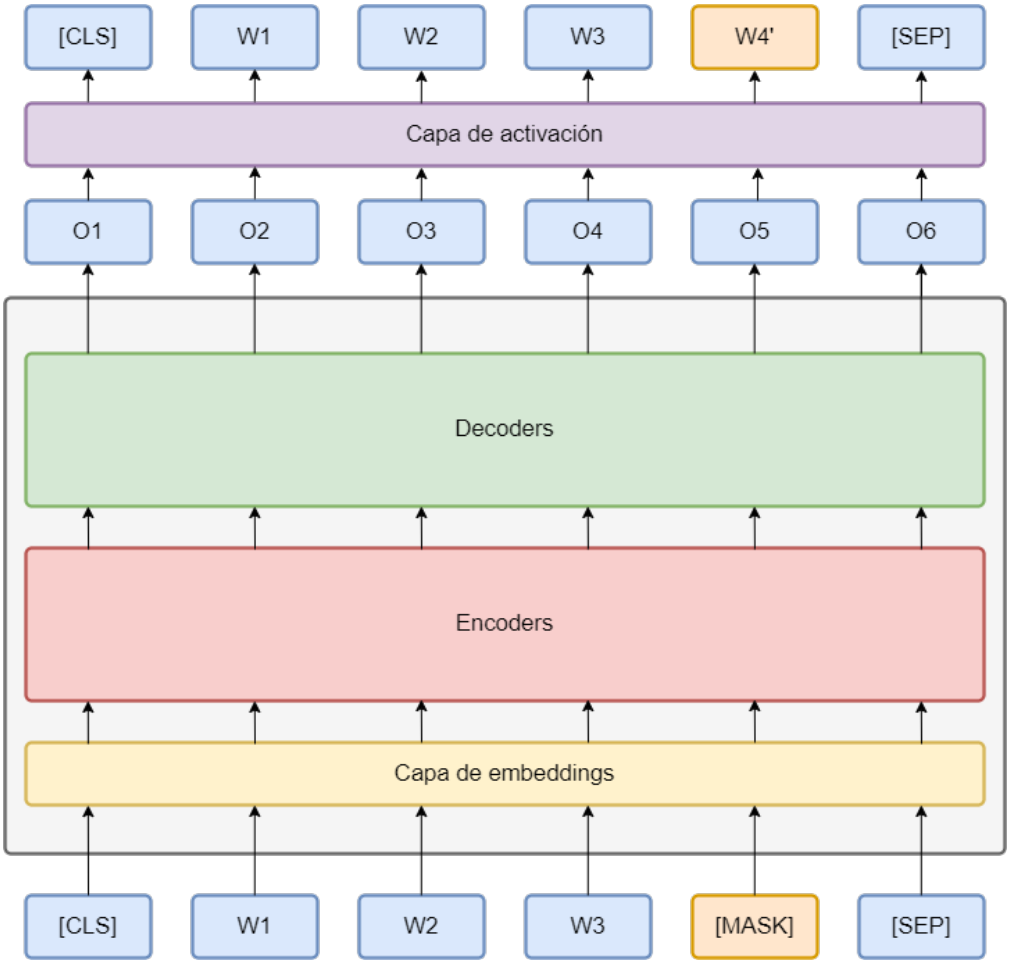


Figura 3.6: Arquitectura del modelo de BERT aplicado a nuestro problema

Tal como se describió en la sección 2.2.2, este modelo consigue resultados del estado del arte en múltiples tareas de NLP. Sin embargo, todos estos se consiguieron aplicandosé en datasets en inglés. Lo mismo ocurre con PictoBERT [10], el cual presenta buenos resultados intrínsecos para una base grande de pictogramas, utilizando descripciones extraidas de Wordnet [24]. Una de las características, según el autor, más relevante de este modelo aplicado a pictogramas, es la buena adaptabilidad que tiene, permitiendo aplicar la estrategia de *transfer learning* para pasar el conocimiento aprendido por un modelo robusto a uno distinto, ubicado en otro contexto distinto, pero similar [27].

Se intentó aplicar esta estrategia para transferir el conocimiento conseguido por pictoBERT, adaptándolo al contexto del tablero descrito en la sección 2.1.1. Sin embargo, no se logró debido a las diferencias semánticas y sintácticas presentes entre español e inglés, lo cual no permite generalizar lo aprendido por pictoBERT a un contexto de pictogramas en español.

Debido a estas razones, el modelo usado como red de pre-entrenamiento fue BETO, un modelo que sigue la misma arquitectura de BERT descrita en 3.6, pero que fue entrenado sobre un corpus en español. Con este modelo, el Departamento de Ciencias de la Computación de la Universidad de Chile logró conseguir resultados del arte en diversas tareas de NLP sobre corpus escritos en español [28].

Este modelo se encuentra presente en la librería *HuggingFace*[11], la cual alberga diversos modelos pre-entrenados basados en transformadores para resolver tareas de muchas modalidades, en particular de NLP. Es de esta librería que se consiguió la red pre-entrenada *uncased-beto*, la cual entrena BETO sobre un corpus de datos sin mayúsculas en español, lo cual coincide con el tipo de textos presentes en nuestra aplicación.

De esta forma, se usó la estrategia de *transfer-learning* para destilar el conocimiento de *uncased-beto* y aplicarlo en nuestra aplicación para predecir la palabra siguiente usando las escritas anteriormente como contexto. Cabe notar que al momento de realizar la inferencia sobre las palabras escritas con anterioridad por el usuario, se utilizaron como contexto solamente las 3 palabras previamente escritas. Si bien esto restringe la información a utilizar por el modelo al momento de predecir una palabra, no se considera que la pérdida de información es significativa porque en el corpus de oraciones escritas por la doctora Hidalgo no existen frases que se compongan de más de 4 palabras. Como se asume que este conjunto de oraciones es una buena representación del problema, se dictamina que la suposición realizada no afectará la evaluación intrínseca ni extrínseca del resultado (para más detalles ver las secciones 5.2 y 5.1).

[11] https://github.com/huggingface/transformers

Capítulo 4

Integración de modelos al tablero de comunicación

Luego de tener listo los sistemas de predicción, es necesario integrarlos al tablero de comunicación digital. Esto significa un cambio en la interfaz, debido a que estamos incluyendo información nueva —las predicciones— al tablero. Se busca que el diseño implementado en este cambio no rompa con el esquema ya establecido y validado. En particular, que cumpla con los principios mínimos de accesibilidad y usabilidad obtenidos como resultados en la memoria de Bahamonde [5].

4.1. Creación de nueva interfaz

La creación de la nueva interfaz se guió por los siguientes objetivos:

- Las recomendaciones debían destacarse del teclado normal.
- El usuario debe darse cuenta que la nueva sección eran sugerencias sin previo tutorial.
- No se debe modificar el diseño general de la aplicación.

El primer punto es importante para que el usuario no piense que son pictogramas normales, o bien otra sección del teclado original. Esto asegurará que se logren visualizar correctamente las predicciones realizadas por el modelo y que sean suficientemente llamativas para que el usuario desee utilizarlas.

Por otro lado, el segundo objetivo es clave para asegurar que la curva de aprendizaje al usar la aplicación se mantenga al mínimo y que agregar información nueva no requiera que el usuario re-aprenda a usar la aplicación. Bahamonde ya obtuvo como resultado que el producto desarrollado le permite al usuario entender cada uno de los objetos de la interfaz sin antes haber interactuado con ella. Es importante que esta característica se mantenga al agregar las predicciones al tablero [5].

El último eje es fundamental debido a que la interfaz de la aplicación ya fue testeada y validada de manera exhaustiva por Bahamonde, y por ende, no se quiere romper con el esquema diseñado al integrar la nueva información. Es por esta razón que se decidió entregar 4 recomendaciones al usuario, este número corresponde a la cantidad de pictogramas presentes

en una fila de el tablero de comunicación. Al hacer esto, se mantiene el esqueleto dentro de la aplicación, aún agregando información nueva.

Siguiendo estos tres ejes claves, se diseñó el siguiente prototipo:

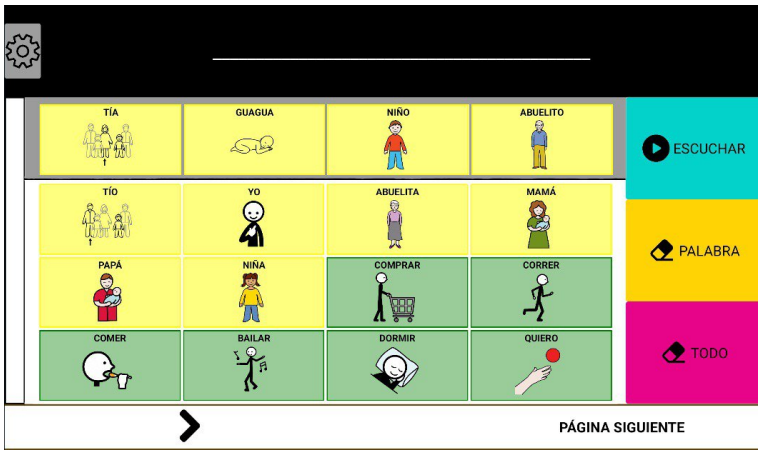


Figura 4.1: Prototipo interfaz nueva

En la primera fila del tablero mostrado en la figura 4.1 se muestran las recomendaciones entregadas por los modelos. Estos cuatro pictogramas se separan del teclado principal por una distancia mayor a la que existe entre cada una de las filas del tablero. Además, para que destaquen con respecto al resto, se coloreó el fondo de un color gris oscuro, distinto al blanco que posee el resto de la aplicación. El diseño de esta interfaz se basó en el teclado predictivo que se implementó en el sistema operativo de Apple, iOS 4[12]. Luego de esta actualización, cuando el usuario comienza a escribir algo, los dispositivos móviles presentan tres sugerencias que corresponden a una predicción de la palabra que se escribirá a continuación. Estas se muestran encima del teclado corriente en un color distinto al original. En la figura 4.2 se presenta el diseño de el teclado predictivo recientemente descrito.

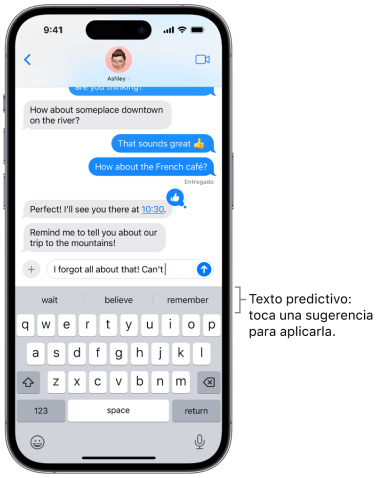


Figura 4.2: Teclado predictivo del Iphone. Imagen tomada del sitio oficial de Apple: https://support.apple.com/es-us/guide/iphone/iphd4ea90231/ios

[12] https://web.archive.org/web/20180510185141/https://www.imore.com/ios-4-features-spellcheck-text-replace

Se organizó una reunión con un experto de dominio en el área de HCI y diseño de sistemas e interfaces interactivas, en donde este comentó que el prototipo mostrado necesitaba destacar más las recomendaciones para que los usuarios notarán su presencia.

Uno de los comentarios recibidos indicaba que el cambio de fondo no era apropiado, ya que no denotaba una sección distinta dentro de la aplicación. Sin embargo, sí lo era la idea de destacar las recomendaciones con un color distinto, ya que esto permitiría que el usuario las identifique con mayor rapidez, cumpliendo así con el primer objetivo. Para lograr abarcar ambos puntos, se decidió en un borde de color más llamativo que envolviese a los pictogramas entregados por la aplicación como sugerencias. De esta forma, se hipotetiza que el usuario logrará verlos con mayor facilidad sin sacrificar la usabilidad del tablero.

Otro cambio que se implementó luego de la inspección de usabilidad con el experto de dominio de HCI, fue la disminución del tamaño de las recomendaciones con respecto al resto de pictogramas del teclado. Esto se hizo para aportar al cumplimiento del objetivo número 1. El tamaño distinto ayuda a que el usuario logre identificar la primera fila como algo separado al resto del teclado, concordando así con el primer eje presentado previamente.

Algo que previamente era considerado como un error era el hecho de que en las recomendaciones se incluyesen pictogramas también presentes en el teclado normal. Se pensó que los duplicados podrían generar confusión en el usuario, dándole a entender que un pictograma aparecía dos veces y por ende, un error. Sin embargo, el experto comentó al respecto que, el tener duplicados dentro de una misma página aportará a que no se rompa el esquema previamente establecido dentro del usuario. Como ya sabe utilizar la aplicación, estará también acostumbrado al orden en el que se encuentran distribuidos los pictogramas y eliminar uno del teclado para ponerlo en las recomendaciones podría provocar que este no sea encontrado por el niño.

Por otro lado, la repetición de pictogramas aporta al objetivo número 2 del diseño de la interfaz, debido a que muestra que existe una sección distinta dentro de la aplicación que tiene pictogramas de esa página y posiblemente de las siguientes, dándole a entender al usuario, tanto al nuevo como al experimentado, que la primera fila de pictogramas son sugerencias entregadas por el modelo.

Luego de aplicar las sugerencias entregadas por el experto de dominio, se diseñó la interfaz presentada en la figura 4.3

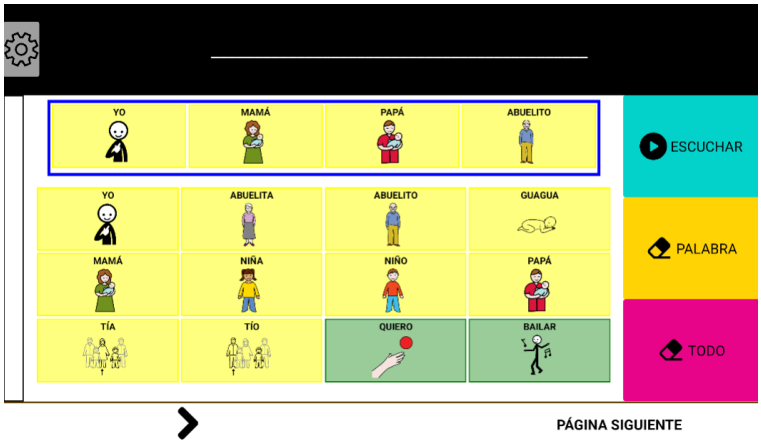


Figura 4.3: Diseño final de interfaz nueva

Capítulo 5

Evaluación de la solución

Considerando que la solución desarrollada se enmarca en el cruce de las áreas de interacción humano computador e inteligencia artificial, es decir, "Intelligent User Interfaces", de acuerdo a la Association for Computing Machinery—ACM, la evaluación de la solución se realizó buscando maximizar indicadores de desempeño, tanto en términos de la calidad del modelo de NLP a generar, como de usabilidad y utilidad percibida por parte de usuarios finales (es decir, niñas y niños con secuelas de parálisis cerebral y/o problemas motoros en sus extremidades superiores). En otras palabras, el proceso de evaluación implica identificar potenciales concesiones a abordar en la calidad de la solución final.

5.1. Evaluación intrínseca

Primero, se realizó una evaluación intrínseca del modelo, lo cual se define como medir la calidad del sistema de predicción independiente de cualquier aplicación [7]. Para esto se utilizaron métricas propias del área de aprendizaje de máquinas, como lo son las siguientes:

- *Top-N*: Porcentaje de veces que la palabra correcta se encuentra en las N primeras predicciones.
- *Perplexity*: Medida que indica qué tanto el modelo de lenguaje logra comprender el funcionamiento de un lenguaje en un corpus.

Considerando que nuestra evaluación se basa en las probabilidades entregadas para frases y palabras, es importante distinguir entre corpus de entrenamiento y de testing. Esto debido a que si hay frases que sirven para testear el modelo que también se encuentran dentro del corpus de entrenamiento, el modelo le asignará una probabilidad artificialmente alta y no implica que logre modelar bien el problema, sino que se acomodó de sobremanera a los datos (este fenómeno se conoce como overfitting) [7, 25].

Es por esto que, previo a realizar cualquier evaluación, se dividió el corpus completo en 80 % para entrenamiento y 20 % para testing. Como se dijo previamente, se computarán las métricas solamente en el subgrupo menor.

5.1.1. Top-N

Al utilizar la métrica de Top-N, se busca evaluar con qué frecuencia el modelo entrega la palabra correcta dentro de las N primeras predicciones. En el caso de nuestro problema particular, utilizaremos un $N \in \{1, 2, 3, 4\}$. Esto debido a que, según lo explicado en la sección 4.1, la aplicación le presentará al usuario 4 recomendaciones y es importante verificar que la predicción correcta se encuentre dentro de estas sugerencias.

Para poder realizar la evaluación se requiere que en cada una de las oraciones que componen el corpus, se escoja una al azar que será reemplazada por el token especial "[MASK]", el cual representará la palabra a predecir (para más detalle ver sección 3.3). Además se deberá guardar la oración truncada hasta el token especial y también la palabra correcta que anteriormente se encontraba en la posición seleccionada.

Tabla 5.1: Todas las opciones de formateo para la frase "yo quiero jugar muñeca"

Frase original:	"yo quiero jugar muñeca"
Frase truncada:	"[MASK]"
Predicción correcta:	"yo"
Frase truncada:	"yo [MASK]"
Predicción correcta:	"quiero"
Frase truncada:	"yo quiero [MASK]"
Predicción correcta:	"jugar"
Frase truncada:	"yo quiero jugar [MASK]"
Predicción correcta:	"muñeca"

El modelo calculará las predicciones utilizando la frase truncada como evidencia ($w_{0:n}$) y se determinará si la predicción correcta se encuentra dentro de las palabras entregadas como sugerencias.

Para cada una de las frases se computarán entonces las siguientes métricas:

- *Top-1*: Si la palabra correcta se encuentra en la primera posición de las predicciones
- *Top-2*: Si la palabra correcta se encuentra en la primera o segunda posición de las predicciones
- *Top-3*: Si la palabra correcta se encuentra en la primera, segunda o tercera posición de las predicciones
- *Top-4*: Si la palabra correcta se encuentra en cualquier posición de las predicciones

Estos cuatro valores nos permitirán evaluar el desempeño de los modelos al generar recomendaciones para completar oraciones construidas por expertos del dominio de AAC.

5.1.2. Perplexity

Al usar la métrica de *Perplexity (PPL)*, se busca evaluar cómo reacciona el modelo ante frases no vistas durante el entrenamiento.

Este valor se define como la inversa de la probabilidad que el modelo le asignó al corpus de testing, normalizado por la cantidad de palabras únicas en el vocabulario. Podemos notar que si el modelo le entrega una probabilidad alta al corpus de testeo, significa que la data vista ahí no le parece nueva y la logra representar correctamente, mostrando una buena comprensión de cómo funciona el lenguaje [7]. En conclusión, al tener un valor menor de perplexity, un modelo será mejor representando el lenguaje.

Dado un corpus de testeo $W = w_1, w_2, w_3, ..., w_n$, se define la métrica PPL, como:

$$\begin{aligned} PPL(W) &= P(w_1, w_2, w_3, ..., w_n)^{-\frac{1}{n}} \\ &= \sqrt[n]{\frac{1}{P(w_1, w_2, w_3, ..., w_n)}} \qquad (5.1) \\ &= \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i \mid w_1, w_3, ..., w_{i-1})}} \qquad (5.2) \end{aligned}$$

El paso desde 5.1 a 5.2 se consigue al aplicar la regla de la cadena en probabilidades, la que se define formalmente como:

Dados los eventos $A_1, A_2, ..., A_n$, con intersección no nula. Es cierto que:

$$P(A_1, A_2, ..., A_n) = \prod i^n P(A_i \mid A_{1:i-1}) \qquad (5.3)$$

Al utilizar la ecuación 5.2 para definir el cálculo de perplexity para los modelos de frecuencia y Markov definidos en las secciones 3.1 y 3.2 respectivamente, se consiguen las siguientes formulas:

Cálculo de perplexity para modelo de frecuencias usando la probabilidad definida en 3.5:

$$PPL_{freq}(W) = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P_{freq}(w_i \mid w_{i-1})}} \qquad (5.4)$$

Cálculo de perplexity para modelo de Markov usando la probabilidad definida en 3.21:

$$PPL_{markov}(W) = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P_{int}(w_i \mid w_{i-2}, w_{i-1})}} \qquad (5.5)$$

Podemos notar que en el caso de que el modelo le otorgue probabilidad 0 a una frase, el calculo de perplexity no va a estar correctamente definido, debido a que estaremos realizando una división por cero.

Para que esto ocurra, una opción es que en el corpus de testeo se utilicen palabras que no se han visto durante el entrenamiento; si consideramos el vocabulario finito y cerrado

utilizado, esto no debería pasar. Sin embargo, tenemos un número muy pequeño de oraciones de las cuales seleccionar y utilizar para el entrenamiento. En el corpus existen palabras con una frecuencia que llega a un valor mínimo de 2. Si es que durante la división de los textos en entrenamiento y testing, alguna palabra termina con todas sus ocurrencias apareciendo en el corpus de testeo, la probabilidad de la cadena que la contenga se estimará como 0, invalidando el cálculo de perplexity para este modelo.

Con el fin de cubrir este caso, se realizó un ajuste tal que el cálculo de perplexity no se invalidará, pero también manteniendo la principal característica de un buen modelo de lenguaje, la cual es que le debe entregar otorgar una alta probabilidad a una frase que tenga sentido lógico y una muy baja a una que no lo tenga. De esta forma, la probabilidad para una palabra no vista no será 0, sino que un número muy bajo. El valor que se determinó como apropiado fue $\frac{1}{n}$, donde n son todas las palabras presentes en el corpus de testeo. Esta frecuencia relativa es apropiada para representar palabras desconocidas, debido a que mantiene la validez de la métrica y a la vez es fidedigna representando palabras poco comunes con un valor muy bajo de probabilidad.

Si bien logramos definir una forma válida de calcular la métrica de perplexity para los modelos descritos, este cálculo causa un overflow, debido a que el denominador dentro de la raíz se vuelve muy pequeño al expandir el producto. Para corregirlo, se tuvo que hacer el siguiente ajuste, utilizando 5.1 como referencia:

$$\begin{aligned} PPL &= \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i \mid w_1, w_3, ..., w_{i-1})}} \\ &= exp\left(log\left(\sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i \mid w_1, w_3, ..., w_{i-1})}}\right)\right) \\ &= exp\left(\frac{1}{n} * log\left(\prod_{i=1}^{n} \frac{1}{P(w_i \mid w_1, w_3, ..., w_{i-1})}\right)\right) \\ &= exp\left(\frac{1}{n} * \sum_{i=1}^{n} log\left(\frac{1}{P(w_i \mid w_1, w_3, ..., w_{i-1})}\right)\right) \end{aligned} \tag{5.6}$$

Considerando que tanto la función $log()$ como la función $exp()$ son funciones monótonamente crecientes, el aplicarlas para simplificar su calculo computacional no cambia el resultado de lo que se quiere minimzar, en este caso perplexity. Esta fue la fórmula final usada para calcular los valores de perplexity para los dos primeros modelos.

Sin embargo, esta métrica no está bien definida para modelos de MLM como BERT (para más detalles revisar 3.3), debido a que el modelo realiza predicciones solamente sobre los tokens enmascarados y la probabilidad de una cadena $w_1, w_2, ..., w_n$ no está bien definida.

Para poder evaluar cómo BERT modela el lenguaje, utilizaremos la definición de perplexity que se deriva del concepto de la teoría de la información conocido como entropía [7].

La entropía es una medida que permite calcular la información mínima requerida para definir una pieza de información. Al ser aplicada a probabilidades, la entropía de una variable aleatoria X, con x siendo un valor posible de X, con probabilidad $P(X)$ se define como:

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x) \tag{5.7}$$

Esto permite describir la información mínima requerida para representar una distribución de probabilidad $\mathbb{P} = \{P(x) \mid \forall x \in X\}$.

Si volvemos al problema principal, se quiere calcular qué tan bien un modelo m logra representar una data. Para tal, se utiliza la métrica de *cross-entropy loss*, la cual utiliza una comparación entre la entropía real requerida para representar un set de datos y la entropía de la representación realizada por el modelo. Entre menor sea la diferencia entre ambos valores, mejor será la capacidad del modelo de representar a la variable. Si consideramos una secuencia de palabras W, pertenecientes a un set de testeo, y un modelo $M = P(w_i \mid wi - n + 1 : i - 1)$ que utiliza la probabilidad condicional P para estimar la data de testeo. Podemos definir la *cross-entropy loss* del modelo al representar W como [7]:

$$H(W) = -\frac{1}{n} \log_2 P(w_1, w_2, w_3, ..., w_n) \tag{5.8}$$

Esta medida nos permite calcular el desempeño de un modelo al representar un conjunto de frases W. Utilizando esta definición, perplexity queda definido de la siguiente forma:

$$PPL(W) = 2^{H(W)} = 2^{-\frac{1}{n} \log_2 P(w_1, w_2, w_3, ..., w_n)} \tag{5.9}$$

Como BERT realiza predicciones solamente sobre los tokens enmascarados, la *cross entropy loss* calculada por defecto también representa la forma en que el modelo estima las palabras a predecir y no las oraciones completas. Para poder computar cómo el modelo se comporta sobre cadenas de palabras, debemos realizar un cálculo de la *cross entropy loss* en el corpus de testing, pero sin reemplazar ningún token por "[MASK]" [10].

Esto es lo que finalmente se hizo para medir cómo se comporta BERT prediciendo frases no vistas durante el entrenamiento.

5.2. Evaluación extrínseca

El buen desempeño de un modelo en métricas intrínsecas como lo son las descritas en la sección anterior debe ser siempre validado por una evaluación extrínseca que mida cuanto mejora una aplicación luego de la integración de un modelo de lenguaje [7].

En este caso, se busca diseñar y ejecutar un estudio de usuarios de naturaleza experimental (prueba A/B) para comparar el rendimiento de la aplicación en su estado de desarrollo en el contexto de este trabajo de título (es decir, con distintas variantes de teclado predictivo, utilizando los modelos de NLP especificados en la sección 3).

Se realizó una evaluación basada en tareas con usuarios que cumplan con el perfil de cuidador de niño con PC, esto incluye a padres/madres, fonoaudiólogos, terapeutas ocupacionales

y educadores. Esto se hizo debido a que una evaluación directamente sobre usuarios finales trae consigo una complejidad logística que puede provocar que los tiempos ocupados en crear la muestra escapen del tiempo destinado para este trabajo de título. Es importante testear la usabilidad de la aplicación para este perfil de cuidador, debido a que son usuarios indirectos de la aplicación, al encontrarse en constante contacto, y por ende, comunicación, con niños con PC.

Se siguió un protocolo de evaluación empírica de tipo prueba de concepto. En esta prueba experimental el usuario debe realizar seis tareas descritas por el evaluador, utilizando la aplicación. La evaluación tiene como objetivo determinar diferencias en usabilidad y eficiencia entre cada uno de los modelos predictivos presentados en la sección 3. Para realizar esta comparación, el evaluador midió las métricas de usabilidad de Wilson y Wixon [8], utilizando el tiempo que toma cada usuario en completar la tarea para medir la eficiencia del modelo predictivo, el número de errores cometido por tarea, el número de errores por unidad de tiempo en cada tarea, el número de ayudas necesarias para completar la tarea y un valor de verdad que represente si el usuario logró completar o no la tarea. Luego de que el usuario haya completado todas las tareas, este debió responder un cuestionario utilizando el método NASA-TLX para medir la carga que sintió durante la evaluación.

Para que todos los modelos predictivos puedan ser evaluados, se dividió la muestra en tres grupos del mismo tamaño. A cada usuario se le asignó uno de los modelos dependiendo de la cantidad de personas que se habían evaluado hasta el momento. En caso de que fuese el primer usuario en ser evaluado, se le realizaba la evaluación con el modelo de frecuencias integrado en la aplicación, en caso de ser el segundo, se utilizaba el modelo de Markov, si era el tercero, se utilizaba BETO, si era el cuarto, se utilizaba el modelo de frecuencias, y así siguiendo el patrón presentado hasta lograr el número final deseado para la muestra.

Cabe destacar que el proceso cumple con los protocolos éticos de la universidad, viéndose sometido a una auditoría por parte del comité de ética en investigación de la facultad. Los protocolos se encuentran en conocimiento de los usuarios, mediante un consentimiento verbal entregado previo a la evaluación.

5.2.1. Participantes

¿Quiénes conformaron la muestra?

Usuarios que cumpliesen con el perfil de cuidador de niño con parálisis cerebral. Esto incluye a familiares directo del niño como padre, madre, abuelo, tía, etc, y también a personal considerado como experto de dominio, es decir personas especializadas en terapias con usuarios objetivo. Cabe destacar que los familiares no estaban necesariamente relacionados con un niño con PC, sino que se simuló el perfil para poder asegurar la usabilidad y utilidad de la aplicación en entornos cercanos al usuario final.

¿Cómo se reclutaron los participantes?

Se utilizó la técnica de muestreo no probabilístico, “muestreo deliberado”. Esta técnica implica que los investigadores seleccionen cuidadosamente a cada individuo para que forme

parte de la muestra utilizando un conocimiento o juicio previo. Para conseguir personas que cumpliesen con el perfil de cuidador, el investigador fue a un establecimiento deportivo (con previo acuerdo con el club dueño del gimnasio) y reclutó a padres/madres de niños que hubiesen recién terminado su partido. Se consiguió un total de 60 evaluaciones, las cuales eran las esperadas para esta evaluación.

Criterios de inclusión y exclusión

Para que un usuario fuese incluido en la muestra, este debía cumplir con el perfil de cuidador de un niño con parálisis cerebral y/o con el de persona experta en el dominio. El criterio de exclusión, corresponde a cualquier otro usuario que no cumpliera con alguno de los perfiles expuestos.

Categorización

- Género: Cualquiera.
- Edad: Mayor a 18 años.
- Conocimiento previo de la aplicación: Nulo.

Verificabilidad de la muestra

La verificación de la muestra se realizó mediante una pregunta por parte del investigador, en donde el usuario debía responder si cumplía con alguno de los dos perfiles expuestos.

Consideraciones éticas

Durante la evaluación el investigador no presionó al usuario a completar ninguna tarea ni utilizar la aplicación de cierta forma, explicando que el objetivo es evaluar el tablero y no al usuario. Previo al inicio de la evaluación se le indicó al usuario que datos iban a ser recolectados mediante el experimento y se le pidió consentimiento verbal de que estos fuesen usados para la investigación propuesta en esta. Cabe destacar que este protocolo experimental fue sometido a una auditoría por parte del comité de ética y bioseguridad en investigación de la facultad.

Posibles sesgos en la muestra

Que los entrevistados entreguen una respuesta que no corresponde a la realidad al preguntarles si pertenecen a alguno de los dos perfiles anteriormente descritos.

Tamaño de la muestra

El tamaño mínimo para una muestra en una evaluación basada en entrevistas con usuarios es de 30, número planteado por Holzinger [29]. Debido a la naturaleza de esta evaluación, se puede tratar a cada modelo como una solución separada, es decir que el número deseado de participantes es 90, 30 por modelo. Debido a que el tiempo para conseguir usuarios estaba acotado por los límites de trabajo de una memoria, se dictaminó que 60 era un tamaño muestral suficiente para determinar como significante el estudio. Hubo 4 evaluaciones que no

se contaron dentro del final debido a errores en la toma de resultados, dejando un total de 56 entrevistas (19 se evaluaron con el modelo de markov, 19 utilizaron el modelo de BERT y 18 usaron el modelo de frecuencia). Este fue el tamaño final de la muestra.

5.2.2. Materiales

¿Qué se usó para ejecutar el estudio?

Se utilizó una tablet de 8" de marca Samsung con sistema operativo Android y que tuvieran instalada la aplicación *Expo Go* para poder portar el tablero digital. Todos los usuarios realizaron la evaluación con el mismo dispositivo móvil.

Condiciones de control y variación

Control: El hardware utilizado fue provisto por el equipo de investigación. Las especificaciones del hardware son:

- Samsung Galaxy Tab A SM-T290 (lanzado 05 de julio de 2019):
  - Pantalla: 8", 1280x800(WXGA)
  - CPU:Qualcomm Snapdragon 429 SDM429 (12 nm), Quad-core, 1950 MHz, ARM Cortex-A53, 64-bit
  - RAM: 2GB
  - OS: Android 11

Variación: Se varió el modelo utilizado para generar las recomendaciones cada usuario, rotando entre todos de forma equitativa. De esta forma, se logró que el tamaño de la muestra para modelo fuese similar. Debido a que se usó la misma aplicación y tablet entre cada persona del mismo grupo, la única variación presente en la evaluación recae en cómo se predijeron las palabras sugeridas al usuario.

Plausibilidad del estudio

El diseño de la aplicación y el protocolo experimental está validado por expertos de dominio, con experiencia en HCI y estudios de usabilidad en la población. Los métodos de medición de usabilidad basados en tareas están validados por Wilson y Wixon. [8]

Replicabilidad del estudio

El hardware utilizado está especificado en la sección de condiciones de control. El código fuente del software se encuentra disponible en GitHub [13]. Utilizando este repositorio se puede montar la aplicación evaluada en la tablet.

5.2.3. Definición del experimento

El experimento es del tipo prueba AB entre sujetos, considerando tres grupos independientes dentro de la evaluación. La aplicación con el modelo de frecuencias integrado, la aplicación

[13] https://github.com/konrad-ivelic/SAAC-Cerebral-Palsy-Predictive

con el modelo de Markov integrado y la aplicación con el modelo basado en transformadores integrado (para más detalle sobre los modelos, revisar la sección 3). Esta división responde al objetivo de querer evaluar cómo cada modelo de predicción se comporta al ser integrado a la aplicación, midiendo específicamente la usabilidad, eficiencia y utilidad percibida por el usuario.

El proceso de asignación de grupos no consideró antecedentes previos del usuario. Cada usuario tenía un modelo diferente que el anterior y el predecesor del anterior. De esta forma se construyó la muestra de forma equitativa para cada modelo, sin dejar ninguno sobrerrepresentado.

Proceso de experimentación

Considerando que los grupos A, B y C deben ser independientes, se tuvo especial consideración que ninguna persona que haya dado o vaya a dar la evaluación estuviese presente en otra evaluación. Esto permitió eliminar bias experimental al tener personas con conocimiento previo de la aplicación al momento de la evaluación.

Se comenzó la evaluación con un saludo y una explicación de la investigación y la aplicación, describiendo el tablero digital y los usuarios finales. "Este proyecto es para una memoría para optar al título de ingeniero civil en la Universidad de Chile, consiste en un tablero de comunicación digital en base a pictogramas para niños con páralisis cerebral. En donde ellos utilizarán íconos asociados a una palabra como bloques para escribir frases. Se está buscando evaluar cómo se adapta esta aplicación a un contexto real con usuarios que cumplan con el perfil de cuidador de niños y medir la utilidad de este tablero. Se le va a pedir que escriba algunas frases y que vaya diciendo todo lo que siente al usar la aplicación. No se sienta presionado/a durante la evaluación porque lo que se esta evaluando es el desempeño de la aplicación y no a usted".

Luego de esta descripción y que el usuario diera consentimiento verbal para el uso de sus datos de forma anónima en la presente investigación, se dio comienzo a la realización de las seis tareas. Las tareas se describen a continuación.

1. Primera tarea:
   - Frase a escribir: *"yo quiero agua"*
   - Acciones a ejecutar:
     a) Con el tablero en estado vacío, es decir sin ninguna palabra escrita, el usuario debe presionar los pictogramas dentro del tablero formando la frase.
     b) Borrar toda la frase.
2. Segunda tarea:
   - Frase a escribir: *"yo quiero jugar"*
   - Acciones a ejecutar:
     a) Con el tablero en estado vacío, es decir sin ninguna palabra escrita, el usuario debe presionar los pictogramas dentro del tablero formando la frase.

b) Borrar toda la frase.

3. Tercera tarea:

   - Frase a escribir: *'mamá quiero galletas jugo"*
   - Acciones a ejecutar:
     a) Con el tablero en estado vacío, es decir sin ninguna palabra escrita, el usuario debe presionar los pictogramas dentro del tablero formando la frase.
     b) Borrar toda la frase.

4. Cuarta tarea:

   - Frase a escribir: *"papá comprar leche"*
   - Acciones a ejecutar:
     a) El usuario debe presionar los pictogramas dentro del tablero formando la frase.
     b) Borrar toda la frase

5. Quinta tarea:

   - Frase a escribir: *"yo llorar dolor pie"*
   - Acciones a ejecutar:
     a) El usuario debe presionar los pictogramas dentro del tablero formando la frase.
     b) Borrar toda la frase

6. Sexta tarea:

   - Frase a escribir: *"yo dormir cama mamá"*
   - Acciones a ejecutar:
     a) El usuario debe presionar los pictogramas dentro del tablero formando la frase.
     b) Borrar toda la frase

Durante el desarrollo de la evaluación, el evaluador midió las métricas de Wilson y Wixon [8], las cuales son:

- Tiempo en completar una tarea
- Número y tipo de errores por tarea
- Número de errores por unidad de tiempo
- Número de ayudas necesarias
- Número de usuarios que completan una tarea con éxito
- Número de recomendaciones usadas

La última variable no está presente en el documento desarrollado por Wilson y Wixon, sin embargo se agregó con el fin de poder medir la cantidad de veces que los usuarios usaban las recomendaciones entregadas por la aplicación en cada tarea. Para tomar registro de estas variables, se utilizó la plantilla presente en el anexo (para más detalle ver documento B.3).

Habiendo realizado todas las tareas, se le pidió al evaluado que conteste una evaluación basada en NASA-TLX (presente en el anexo, ver documento B.1), con el fin de medir las sensaciones de carga y dificultad percibidas por los usuarios durante la evaluación. Esta herramienta se utilizó para calificar la carga de trabajo percibida por el usuario durante la realización de las tareas. Las dimensiones que evalúa este instrumento son:

- Exigencia mental
- Exigencia física
- Exigencia temporal
- Rendimiento
- Esfuerzo
- Nivel de frustración

Adicionalmente se le pidió al usuario que contestase un cuestionario diseñada por el equipo de investigadores, validada por expertos de dominio con experiencia en HCI, que tiene como fin medir la utilidad percibida de las recomendaciones. Este documento entregó una medida de comparación entre los tres modelos de predicción integrados a la aplicación. En el cuestionario se le entregaron cinco aseveraciones al usuario, para que este evalúe en una escala de likert de cinco puntos qué tan de acuerdo esta con ellas. Las aseveraciones que incluía el cuestionario son:

- Me fue fácil identificar las recomendaciones
- Me parecieron apropiadas las recomendaciones
- Me parecieron valiosas las recomendaciones
- Me parecieron útiles las recomendaciones
- Es probable que vuelva a utilizar la aplicación

Para cada una de estas aseveraciones el usuario tenía como opción las siguientes respuestas.

- Muy en desacuerdo
- En desacuerdo
- Ni de acuerdo, ni en desacuerdo
- De acuerdo
- Muy de acuerdo

Adicionalmente se presentaba una pregunta abierta final en donde se le preguntaba al usuario *"¿Tiene comentarios o sugerencias adicionales?"*. Esto con el fin de recolectar respuestas cualitativas sobre la utilidad de las recomendaciones. Este cuestionario se presenta en el anexo (para más detalle ver documento B.2).

5.2.4. Instrumentos de recolección de datos

Tipo de estudio

Estudio exploratorio, porque mide cómo se comporta el sistema y los modelos predictivos en un escenario de prototipo (prueba de concepto). Todos los datos fueron recolectados en papel, con el fin de que la evaluación fuese más sencilla para los usuarios. Estas encuestas fueron impresas para que el usuario pudiese responder con el lápiz de su preferencia. El investigador tenía lápices y encuestas de sobra en caso de que alguno fallará. Con respecto a la medición de las métricas, el investigador estuvo atento a todo el proceso evaluativo, tomando notas y realizando observación activa del proceso para medir todos los datos referentes a las métricas de Wilson y Wixon [8].

¿Cómo se recolectó la información?

Para medir los tiempos que le tomó al usuario completar cada tarea, se utilizó el cronómetro de un celular. El tiempo comienza en el momento de terminar las instrucciones al usuario y termina cuando logra completar la tarea. Se utilizó una tabla en papel para recolectar los datos que permiten medir las métricas mencionadas en el apartado de *Proceso de experimentación* (para más detalles ver 5.2.3 y documento B.3).

Calibración de los instrumentos

Al usar el cronómetro integrado en un teléfono celular, había que fijarse que este se encontraba en 0 y que el dispositivo móvil tuviera batería suficiente para realizar toda la evaluación. Las encuestas entregadas a los usuarios y las tablas de recolección debían encontrarse impresas previo a realizar el proceso de investigación. Todos los cuestionarios usados pueden ser encontrados en el capitulo B del anexo.

Validez y confianza de los instrumentos

NASA-TLX[14] y Wilson y Wixon [8] son métodos validados. El diseño de la aplicación y el protocolo experimental están validados mediante una evaluación heurística con expertos del dominio de HCI y estudios de usabilidad.

Limitaciones de los instrumentos

Se utilizó una tablet de 8", descrita en detalle en la sección 5.2.2. Lo que implica una limitación física-espacial del instrumento evaluativo. Cabe mencionar que Bahamonde [5] obtuvo como resultado de su evaluación que no existe diferencia significativa entre usar el tablero digital integrado en una tablet de 8" y una de 12.4".

5.2.5. Procedimiento de recolección de datos

¿Cómo se recolectó la información?

[14] Método NASA-TLX, Ministerio del Trabajo y Previsión Social: https://ergomedia.isl.gob.cl/app_ergo/nasatlx/

- Pre evaluación: Previo a la investigación se le pidió al usuario que consienta verbalmente a que los datos usados sean utilizados en la presente investigación. Además se le especificó que el proceso tenia como objetivo evaluar el desempeño de la aplicación y no al usuario. Luego se le asignó un grupo de estudio según los utilizados en los dos usuarios anteriores, entregándole la aplicación con un modelo de predicción integrado tal que sea distinto al usado por las dos personas que lo precedieron.

- Durante la evaluación: Se tomaron los tiempos utilizados por los usuarios para completar las tareas con un cronómetro, además se relleno el documento de recolección de datos para registrar las otras métricas de usabilidad (para más detalle ver el documento B.3 en el anexo).

- Post evaluación: Finalizada la intervención, se le entregaron al usuario dos cuestionarios, uno basado en NASA-TLX para medir la carga de trabajo percibida durante la evaluación (para más detalle ver el documento B.1 en el anexo) y uno con preguntas que permiten medir la utilidad percibida por el usuario con respecto a las recomendaciones entregadas (para más detalle ver el documento B.2 en el anexo).

La información obtenida durante y después de la evaluación fue traspasada a un formato digital para realizar el análisis estadístico que se presentará más adelante.

Condiciones de observación y/o experimentación

El usuario y el facilitador debían encontrarse en el mismo lugar físico. Al iniciar la evaluación, el facilitador debía encontrarse un poco más alejado para no incidir en el uso de la aplicación. Durante todo el proceso evaluativo, el usuario no podía conversar con personas externas al proceso, con el fin de poder captar una representación fidedigna de como el usuario percibió y utilizó la aplicación.

Pilotaje del proceso experimental

Para pilotar el proceso, se realizó una evaluación a baja escala, con un grupo de 5 personas, las cuales debieron ejecutar 4 de las tareas propuestas. Esto se realizó con al finalidad de validar la plausibilidad del protocolo experimental, en particular verificar la entendibilidad de las tareas y cuestionarios a aplicar. Las tareas incluidas en el pilotaje son:

1. Primera tarea:
    - Frase a escribir *"yo quiero agua"*
    - Acciones a ejecutar:
        a) El usuario debe presionar los pictogramas dentro del tablero formando la frase.
        b) Borrar toda la frase
2. Segunda tarea:
    - Frase a escribir *"yo quiero jugar"*
    - Acciones a ejecutar:
        a) El usuario debe presionar los pictogramas dentro del tablero formando la frase.

b) Borrar toda la frase

3. Tercera tarea:

   - Frase a escribir *"yo llorar dolor pie"*
   - Acciones a ejecutar:
     a) El usuario debe presionar los pictogramas dentro del tablero formando la frase.
     b) Borrar toda la frase

4. Cuarta tarea:

   - Frase a escribir *"yo dormir cama mamá"*
   - Acciones a ejecutar:
     a) El usuario debe presionar los pictogramas dentro del tablero formando la frase.
     b) Borrar toda la frase

Es importante incluir estas tareas en el proceso de pilotaje, debido a que fueron diseñadas teniendo en cuenta la forma en que cada modelo realiza las predicciones. Luego de la realización de las 4 tareas presentadas, se le pidió al usuario piloto que responda las dos encuestas post-evaluatorias, NASA-TLX y el cuestionario que mide la utilidad percibida de las recomendaciones (para más detalles ver sección 5.2.3 y el documento B.1 en el anexo). El tamaño de muestra utilizado para el pilotaje fue de 5 personas, para verificar que el proceso sea entendible tanto para el investigador como para el usuario.

Verificabilidad del proceso

El proceso de pilotaje fue realizado con un experto de dominio con conocimiento en HCI y usabilidad, por lo que se encuentra verificado y validado. El protocolo se adhiere a las normas éticas en HCI [30]. Estas normas deben ser aplicadas durante todas las etapas del estudio presentadas en la sección 5.2.3.

Replicabilidad del proceso

El proceso es replicable utilizando todos los instrumentos presentados en el anexo de este documento y la tablet presentada en la sección 5.2.2.

Limitaciones del proceso

Con respecto a la validez interna, se encuentra que solamente se utilizó una tablet de 8", sin probar como se adaptaba la aplicación a otros tamaños. Cabe señalar que Bahamonde [5] logró como resultado que no existe distinción significante entre una tablet de 8" y una de 12.4", sin embargo esta evaluación no se hizo con el sistema de recomendación integrado a la aplicación, por ende, estos resultados podrían diferir utilizando el protocolo experimental descrito en esta sección.

Con respecto a la muestra utilizada, se realizó la evaluación con un perfil de cuidador de niño con PC. Por ende los resultados pueden variar al aplicar esta evaluación en la población

objetiva de niños con PC.

Considerando el procedimiento, se tiene como limitación los instrumentos de recolección de datos, NASA-TLX que mide sensaciones del usuario cualitativamente y las métricas de Wilson y Wixon [8] que miden la usabilidad de la aplicación cuantitativamente, en conjunto con el teléfono usado para medir el tiempo.

Capítulo 6

Resultados

Se presentan a continuación los resultados de las evaluaciones explicadas en el capitulo 5.

6.1. Hipótesis

Se plantean las siguientes hipótesis de trabajo con respecto a los resultados de las evaluaciones:

- Intrínsecamente el modelo con mejor desempeño será el basado en transformadores.
- Extrínsecamente el modelo con mejor desempeño será el que utiliza cadenas de Markov.

6.2. Resultados evaluación intrínseca

En esta sección se presentan los resultados de cada modelo para las métricas intrínsecas presentadas en la sección 5.1.

6.2.1. Resultados Top-N

Con respecto a la métrica de *Top-N*, se consiguieron los siguientes resultados para los valores de 1,2,3,4.

Tabla 6.1: Tabla con los valores de Top-N para los tres modelos

Modelos	**Top-1**	**Top-2**	**Top-3**	**Top-4**
Frecuencia	41.50 %	46.23 %	47.16 %	49.06 %
Markov	53.85 %	64.98 %	72.08 %	75.53 %
BERT	49.67 %	62.41 %	68.12 %	71.93 %

Esta tabla muestra que Markov posee el mejor desempeño, dentro de los tres modelos desarrollados, en predecir la siguiente palabra para todos los valores de N. En particular, estos resultados muestran que cada vez que este modelo entregue 4 palabras como recomendaciones, la palabra correcta estará en este grupo un 75.53 % de las veces.

Se puede destacar también que ambos modelos de Markov y BERT, logran mejores resultados que el modelo de frecuencias, con respecto a la precisión de predecir la siguiente palabra.

6.2.2. Resultados Perplexity

Con respecto a la métrica de *Perplexity*, se consiguieron los siguientes resultados para los tres modelos integrados en la aplicación.

Tabla 6.2: Tabla con los valores de perplexity para los tres modelos

Modelos	**Perplexity**
Frecuencia	76.65
Markov	5.14
BERT	1.24

De esta tabla, podemos desprender que BERT es el modelo que logra una mejor representación del lenguaje de este problema.

De la misma forma que para la métrica de *Top-N*, tanto el modelo de BERT como el de Markov logran una mejor representación del lenguaje que el modelo de frecuencias. Esto se debe a que le asigna probabilidad muy baja a las palabras que no siguen la estructura sintáctica predeterminada (sujeto-verbo-sustantivo-...-sustantivo ; revisar 2.1.1 para más detalle). Por ejemplo, estima que la probabilidad de que dos verbos se encuentren seguidos es muy baja, lo cual contrasta con la alta ocurrencia de esta combinación en el dataset creado por la doctora Hidalgo.

Luego de ver los resultados se determina que esta métrica no es categórica para discernir sobre qué modelo tiene un mejor desempeño intrínseco. Esto, porque existe un modelo que consigue una *perplexity* casi perfecta (su cota inferior es 1 por definición). Se hipotetiza que esto ocurre debido a la similaridad entre el corpus de testeo y el ocupado para entrenamiento, lo que viene dado por la poca cantidad de oraciones creadas. Por ende, los resultados entregados no representan fidedignamente qué tan bien cada modelo representa el lenguaje.

Considerando estos resultados, se determina que el modelo con un mejor desempeño intrínseco es el que utiliza las cadenas de Markov.

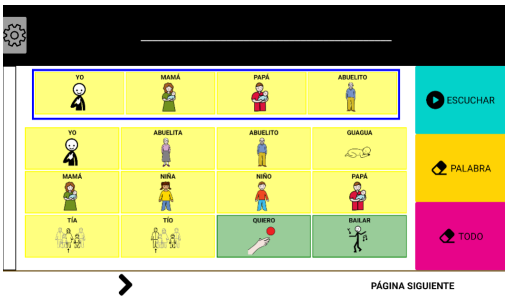
6.3. Resultados evaluación heurística

El nuevo diseño de la aplicación fue validado en una inspección de usabilidad con expertos de dominio de trata de pacientes, pertenecientes a las fundaciones Coaniquem y Teleton, en donde se incluían a la doctora Gabriela Hidalgo (Fisiatra), Matiás Orellana (Fisiatra), Daniel Durán (Terapeuta ocupacional) y Claudia Sepúlveda (Fonoaudióloga). Las cuatro personas catalogadas como especialistas trabajan diariamente con sistemas aumentativos y alternativos de comunicación, teniendo el conocimiento requerido para poder discernir si la interfaz diseñada mantiene la usabilidad del tablero de comunicación previo a su modificación.

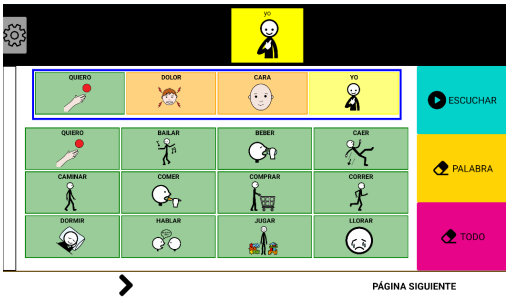
Para empezar la reunión se les mostró la secuencia de imágenes mostradas en la figura 2.2.a. De esta forma los expertos conocerían la situación que se tenía previa a la realización de este trabajo de título. Posteriormente, se realizó una demo en vivo de la aplicación con el sistema de predicciones integrado. Se decidió mostrar el modelo de Markov. Sin embargo,

lo que se quería evaluar no era el desempeño del sistema de recomendaciones, sino que se deseaba validar que el hecho de agregar recomendaciones constituía a una posible mejora a la aplicación y además verificar que la interfaz aplicada para integrar las sugerencias efectivamente cumple con los principios de accesibilidad y usabilidad alcanzados como resultado del trabajo de título de Bahamonde.

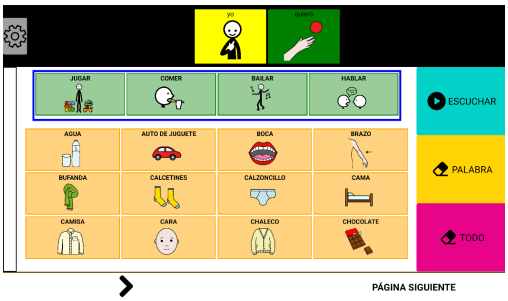
Se presenta a continuación una secuencia de imágenes que ejemplifica la demostración realizada durante la inspección de usabilidad con expertos de dominio:



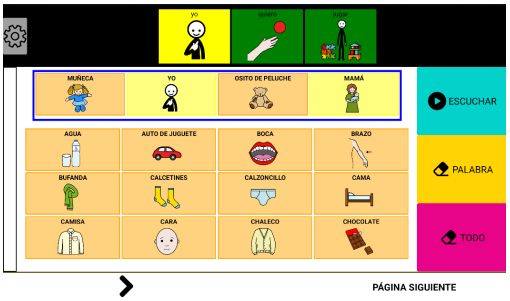
(a) Input vacío



(b) Palabra “yo” escrita



(a) Palabras “yo quiero” escritas



(b) Palabras “yo quiero jugar” escritas

Figura 6.5: Modelo de Markov generando recomendaciones

Los expertos realizaron comentarios con respecto a la aplicación, y a su vez, entregaron muchas oportunidades de mejora para el sistema de recomendación.

Con respecto a la interfaz que integra los modelos de predicción al tablero de comunicación, se comentó inicialmente que el tamaño permite discriminar claramente la imagen de las recomendaciones del teclado completo. Siguiendo el mismo punto, el azul aporta a la visibilidad de la sección, debido a que contrasta con el resto de los colores de la aplicación. Esto confirma que el nuevo diseño permite destacar las predicciones, cumpliendo así el primer objetivo propuesto.

El que este bajo la construcción de la frase permite al usuario verlo de manera inmediata al ingresar a la aplicación y utilizar el teclado.

En línea con el segundo eje bajo el cual se diseño la interfaz, los expertos opinaron que la inclusión de recomendaciones parece súper intuitivo. A medida que se van escribiendo los pictogramas, van cambiando los presentados dentro del recuadro azul, por lo que se logra entender que son una propuesta por parte de la aplicación. Es por esto que se determina que

la curva de aprendizaje al utilizar la aplicación por primera vez no será grande.

Con respecto a la utilidad percibida por los expertos, se dijo que las respuestas sugeridas permiten al usuario ver de inmediato posibles opciones con las cuales puede continuar la frase. Lo cual le permitirá escribir oraciones de forma más intuitiva y rápida, logrando así una comunicación más efectiva y por ende, agrega un gran valor a la aplicación.

En general, todos estuvieron de acuerdo con que lo presentado es una evolución lógica para la aplicación y que agregar un predictor constituye una muy buena idea.

Como oportunidades de mejora, se propusieron algunas sobre el funcionamiento de la aplicación, mientras que otras iban más enfocadas a las predicciones como tal.

Por ejemplo, una modificación sugerida es la de incluir un botón a la derecha de la nueva sección de recomendaciones. Este botón, al momento de ser presionado, debe generar predicciones nuevas, distintas a las presentadas con anterioridad. Esto permitiría entregar al usuario un número mayor de recomendaciones al refrescar la fila superior y quizás aumentar aún más los efectos positivos de entregar predicciones relevantes. Sin embargo, esta decisión compromete la usabilidad de la aplicación, ya que es un mecanismo no muy utilizado dentro del diseño común de sistemas de predicciones. Es por esta razón que no se incluyó dentro de este trabajo de título, no obstante, es algo que vale la pena explorar y evaluar si constituye una mejoría a la aplicación.

Los expertos hicieron notar sobre la importancia de que el sistema de recomendaciones reconozca la individualidad de cada usuario, sobretodo si esta enfocado a niños que tienen gustos cambiantes a medida que van creciendo. Un niño de 5 años querrá jugar con su auto de juguete, mientras que uno de 12 probablemente querrá ir de shopping o escuchar música.

Con respecto al diagnóstico clínico, se hizo notar que cada uno de los niños que utilicen este tablero digital tendrán un grado de movilidad distinto y por ende, tendrán preferencias distintas con respecto a las frases que querrán utilizar. Un paciente con compromiso motor severo, necesitará comunicar actividades que para él son cotidianas, que no lo son para niños con mayor grado de movilidad. Como por ejemplo, acciones frecuentemente utilizadas por integrantes de este grupo incluyen: aspirar secreciones respiratorias, cambiar de posición o dar analgésico. Las cuales no constituyen a frases que actualmente se puedan escribir en la aplicación, debido a que se utilizó un vocabulario estandarizado para niños de 2 a 4 años y no especialmente adecuado para personas con parálisis cerebral.

Siguiendo con la linea de la personalización, el sistema predictivo no considera experiencias previas del usuario. Algo que para los expertos era de suma importancia. Para ellos, mientras un usuario utilice la tableta digital, las predicciones deberían utilizar las preferencias personales de cada usuario para ponderar las probabilidades de la palabra escrita a continuación. Esto aportará a generar predicciones más adecuadas a cada usuario, por ende, mejorando la comunicación.

Durante la reunión se levantó la necesidad de que el sistema de predicción realice distinción entre los ambientes en los que se puede encontrar el usuario. Es distinto usar la tablet

en el colegio que en la casa y es necesario que la aplicación haga tal distinción. Para incluir esto, se puede recuperar la geolocalización del usuario y a partir de ella, presentar palabras distintas como recomendaciones. Sin embargo, se dijo también que las palabras incluidas en el vocabulario no dan abasto para cubrir las situaciones de comunicación en la sala de clases o en la clínica.

Es parte de una siguiente investigación, realizar un nuevo vocabulario adaptado a niños con parálisis cerebral con palabras que permitan la comunicación en distintos ambientes, como lo es el colegio, la casa o la clínica.

Muchas de estas oportunidades de mejora se encuentran con el mismo cuello de botella, el dataset de entrenamiento debe de incluir el ambiente en donde se escribió la frase. Esto con el fin de que el modelo de predicción lo ocupe como evidencia previa para generar las recomendaciones. Como se vio al inicio de la sección 3, esto se intentó hacer e incluir, sin embargo no se llego al volumen necesario de frases para que el corpus tuviera significancia. Se presenta como una oportunidad de mejora en los siguientes capitulos.

6.4. Resultados evaluación extrínseca

Para cada tarea, lo que se analizará es la variable tiempo en los tres modelos, incluyendo también en la comparación los datos obtenidos durante la evaluación realizada por Bahamonde. Esto permite realizar una comparación directa entre la eficiencia de cada uno de los modelos predictivos y la eficiencia de la aplicación sin ellos.

Para asegurar verificabilidad de los resultados presentados en esta sección, se incluyen como anexo las tablas para cada tarea que tienen todas las observaciones realizadas durante los experimentos (para más detalle ver capitulo A en el anexo). En el caso de los datos usados como control, estos se encuentran disponibles en la sección de anexo de la memoria de Bahamonde [5].

A continuación se presentan los resultados para las métricas registradas durante y posterior a la evaluación (para más detalles revisar las sección 5.2.3)

Debido a que el objetivo de este análisis es comparar los modelos bajo el mismo contexto, es que se hizo un análisis del tiempo que tomó cada usuario en completar todas las tareas y además un análisis separado en cada tarea. Se permite así entregar une métrica que evalúa la eficiencia de cada modelo entregando las recomendaciones.

Se comparan los resultados en la métrica de Tiempo, la cual hace referencia a la cantidad de tiempo, en segundos, que le toma a un usuario completar la tareas. Los datos se agruparán según el modelo integrado en la aplicación al momento de la evaluación, para así poder analizar de forma comparativa la eficiencia de cada grupo experimental. Se realiza también una agrupación por género, con el fin de determinar si existe diferencia entre la usabilidad percibida por los usuarios de distinto género.

Se realiza una comparación entre tres (tarea 4,5 y 6) o cuatro (tarea 1,2 y 3) medias de grupos independientes. Para poder hacer este análisis es necesario primero determinar

si los datos se distribuyen de forma normal, o no. Esta verificación se hace con el método de Shapiro-Wilk. En el caso de que la hipótesis nula no se pueda rechazar con un grado de confianza de 0.05, se asumirá normalidad de los datos y se utilizará un análisis de la varianza (ANOVA) y pruebas T para determinar si existe diferencia determinante entre dos o más grupos experimentales independientes, en caso de que se cumpla la hipótesis de homocedasticidad. En caso contrario, se usará la correspondencia no paramétrica de Kruskal-Wallis y pruebas U (de Mann-Whitney) como análisis post-hoc.

6.4.1. Promedio agregado

Se presentan los resultados de la variable Tiempo, que constituye al tiempo que le demoró al usuario completar, en promedio, cada tarea.

Agrupado por Modelo

Considerando el tiempo, en promedio, en que los usuarios completaron cada tarea, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR:

Tabla 6.3: Valores estadísticos para el tiempo que les tomó a los usuarios completar, en promedio, cada tarea según modelo usado

Modelo	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
BERT	20.4	6.72	18.3	14.3	44.2	3.75
Frecuencia	23.2	4.14	21.8	18.2	33.3	4.83
Markov	19.7	5.50	19	11.8	34	4.92

El diagrama de caja que muestra los resultados del tiempo requerido, en promedio para cada tarea, según cada modelo, se presenta a continuación:

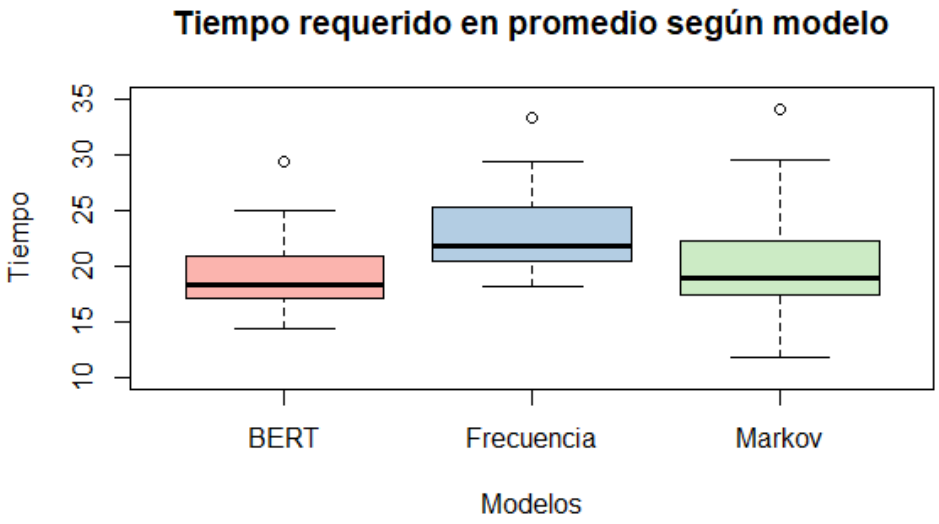


Figura 6.6: Tiempo requerido en completar todas las tareas según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 4 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value = 0.008$, lo cual al ser menor a 0.05 permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba pareada de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.4: Prueba pareada de Mann-Whitney para el tiempo total requerido en tareas según modelo

	BERT	**Frecuencia**
Frecuencia	0.014	-
Markov	0.942	0.021

Considerando los datos presentados en 6.4, se puede determinar que existe una diferencia significativa entre los grupos de BERT y Frecuencia (magnitud de efecto grande, delta de Cliff $= -0.54$) y Markov y Frecuencia (magnitud de efecto grande, delta de Cliff $= 0.47$). Mostrando en todos estos pares un valor U menor a 0.05.

Agrupado por Género

Considerando el tiempo, en promedio, en que los usuarios completaron cada tarea, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.5: Valores estadísticos para el tiempo que les tomó a los usuarios completar todas las tareas según su género

Género	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
Femenino	21.0	6.34	20.5	11.8	44.2	5.83
Masculino	21.2	4.99	19.3	13.5	34	4.42

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un valor U de 0.64, el cual al ser mayor a 0.005, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no hay diferencia significativa entre ambos grupos.

6.4.2. Tarea 1

Se presentan los resultados de la variable Tiempo, que constituye al tiempo que le demoró al usuario completar la Tarea 1.

6.4.2.1. Agrupado por Modelo

Considerando el tiempo en que los usuarios completaron la tarea 1, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.6: Valores estadísticos para el tiempo requerido en completar tarea 1 según modelo usado

	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
Control	29.2	21.0	21	10	76	23
BERT	13.2	4.31	13	7	23	7
Frecuencia	14.1	5.18	13	8	27	4.5
Markov	15.1	10.4	12	4	50	7

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 4 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p-value < 0.005$, lo cual nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba de Mann-Whitney, la cual entrega los siguientes

Tabla 6.7: Prueba de Mann-Whitney para tiempo utilizado en completar tarea 1 según modelo

	BERT	**Control**	**Frecuencia**
Control	0.0027	-	-
Frecuencia	0.8490	0.0029	-
Markov	0.8490	0.0029	0.8490

Considerando los datos presentados en 6.7, se puede determinar que existe una diferencia significativa entre los grupos de Markov y Control (magnitud de efecto grande, delta de Cliff = 0.54), BERT y Control (magnitud de efecto grande, delta de Cliff = -0.60), Frecuencia y Control (magnitud de efecto grande, delta de Cliff = 0.56). Mostrando en todos estos pares un valor U menor a 0.005.

6.4.2.2. Agrupado por Género

Considerando el tiempo en que los usuarios completaron la tarea 1, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.8: Valores estadísticos para el tiempo requerido en completar tarea 1 según género del usuario

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	13.9	6.02	13	4	30	5
Masculino	14.4	8.23	13	7	50	7

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un valor U de 0.94, el cual al ser mayor a 0.005, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no hay diferencia significativa entre ambos grupos.

6.4.3. Tarea 2

Se presentan los resultados de la variable Tiempo, que constituye al tiempo que le demoró al usuario completar la Tarea 2.

6.4.3.1. Agrupado por Modelo

Considerando el tiempo en que los usuarios completaron la tarea 2, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.9: Valores estadísticos para el tiempo requerido en completar tarea 2 según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	11.6	8.46	8	4	38	6
Control	69.7	24.2	73	21	110	37
Frecuencia	30.8	5.04	31.5	22	40	5.25
Markov	15.8	11.9	11	5	41	13.5

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 4 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value < 0.05$, lo cual nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.10: Prueba de Mann-Whitney para tiempo utilizado en tarea 2 según modelo

	BERT	Control	Frecuencia
Control	7.9e-08	-	-
Frecuencia	5.6e-06	2.6e-06	-
Markov	0.26571	1.3e-07	0.00079

Considerando los datos presentados en 6.10, se puede determinar que existe una diferencia significativa entre los grupos de BERT y Control (magnitud de efecto grande, delta de Cliff $= -0.98$), BERT y Frecuencia (magnitud de efecto grande, delta de Cliff $= -0.89$), Frecuencia y Control (magnitud de efecto grande, delta de Cliff $= 0.84$), Markov y Control (magintud de efecto grande, delta de Cliff $= 0.94$) y Markov y Frecuencia (magnitud de efecto grande, delta de Cliff 0.65). Mostrando en todos estos pares un valor U menor a 0.05.

6.4.3.2. Agrupado por género

Considerando el tiempo en que los usuarios completaron la tarea 2, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.11: Valores estadísticos para el tiempo requerido en completar tarea 2 según género del usuario

Género	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
Femenino	18.6	11.1	20	4	38	21
Masculino	19.8	13.2	12	5	41	24.5

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un valor U de 0.61, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no hay diferencia significativa entre ambos grupos.

6.4.4. Tarea 3

Se presentan los resultados de la variable Tiempo, que constituye al tiempo que le demoró al usuario completar la Tarea 3.

6.4.4.1. Agrupado por Modelo

Considerando el tiempo en que los usuarios completaron la tarea 3, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.12: Valores estadísticos para el tiempo requerido en completar tarea 3 según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	31.6	9.29	31	24	65	8
Control	43.6	20.8	36	15	85	38
Frecuencia	22.2	5.07	21.5	11	30	5
Markov	34.4	15.1	30	15	87	10

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 4 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value < 0.05$, lo cual nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.13: Prueba de Mann-Whitney para tiempo utilizado en tarea 3 según modelo

	BERT	**Control**	**Frecuencia**
Control	0.08209	-	-
Frecuencia	0.00024	0.00017	-
Markov	0.60826	0.28951	0.00024

Considerando los datos presentados en 6.13, se puede determinar que existe una diferencia significativa entre los grupos de BERT y Frecuencia (magnitud de efecto grande, delta de Cliff = 0.75), Frecuencia y Control (magnitud de efecto grande, delta de Cliff = 0.73) y Markov y Frecuencia (magnitud de efecto grande, delta de Cliff -0.74). Mostrando en todos estos pares un valor U menor a 0.05.

6.4.4.2. Agrupado por Género

Considerando el tiempo en que los usuarios completaron la tarea 3, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.14: Valores estadísticos para el tiempo requerido en completar tarea 3 según género del usuario

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	31.3	15.0	28	15	87	9
Masculino	27.6	6.54	27	11	40	7.5

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un valor U de 0.69, el cual al ser mayor a 0.005, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no hay diferencia significativa entre ambos grupos.

6.4.5. Tarea 4

Se presentan los resultados de la variable Tiempo, que constituye al tiempo que le demoró al usuario completar la Tarea 4.

A diferencia de las tres tareas anteriores, no se incluyó el grupo de Control, correspondiente a la evaluación hecha por Bahamonde [5]. Esto debido a que en su evaluación existía una tarea previa que cambiaba el funcionamiento del tablero y los resultados obtenidos con respecto a la escritura de la frase “papa comprar leche” fueron condicionados por como este cambio

afectaba la usabilidad de la aplicación [15]

6.4.5.1. Agrupado por Modelo

Considerando el tiempo en que los usuarios completaron la tarea 4, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.15: Valores estadísticos para el tiempo requerido en completar tarea 4 según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	10.8	7.31	9	5	38	3.5
Frecuencia	10.8	4.53	9	5	19	7
Markov	15.3	7.44	15	6	35	9

El diagrama de caja que muestra los resultados del tiempo requerido para cada modelo en la tarea 4 se presenta a continuación:

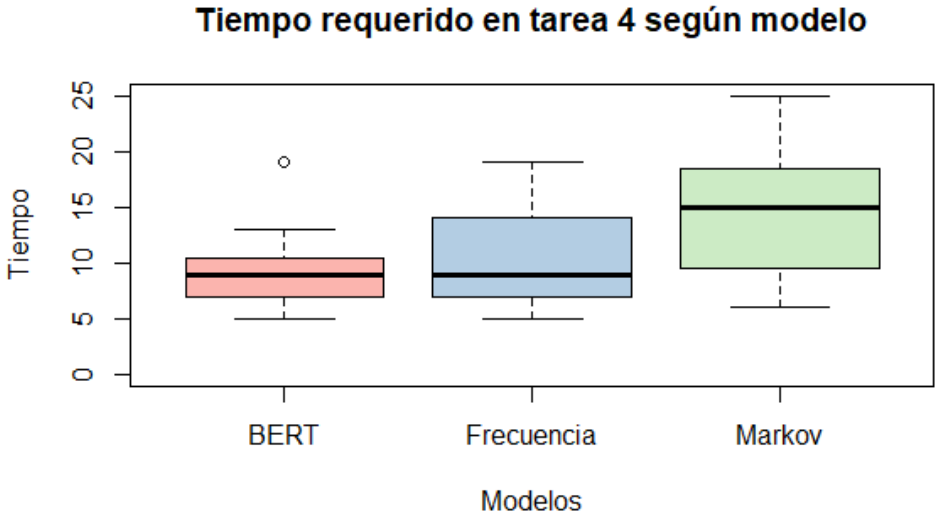


Figura 6.7: Tiempo requerido en tarea 4 según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p-value < 0.005$, lo cual nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo

[15] Para más detalles, revisar la memoria de Bahamonde, disponible en https://repositorio.uchile.cl/handle/2250/193929

menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.16: Prueba de Mann-Whitney para tiempo utilizado tarea 4 según modelo

	BERT	**Frecuencia**
Frecuencia	0.657	-
Markov	0.047	0.077

Considerando los datos presentados en 6.16, se puede determinar que existe una diferencia significativa entre los grupos de BERT y Markov (magnitud de efecto mediana, delta de Cliff $= -0.45$) y Frecuencia y Markov (magnitud de efecto mediana, delta de Cliff $= -0.37$). Mostrando en todos estos pares un valor U menor a 0.005.

6.4.5.2. Agrupado por Género

Considerando el tiempo en que los usuarios completaron la tarea 4, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.17: Valores estadísticos para el tiempo requerido en completar tarea 6 según género del usuario

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	13.0	8.30	10	5	38	7
Masculino	11.7	4.84	10	5	24	7.5

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un valor U de 0.99, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no hay diferencia significativa entre ambos grupos.

6.4.6. Tarea 5

Se presentan los resultados de la variable Tiempo, que constituye al tiempo que le demoró al usuario completar la Tarea 5.

6.4.6.1. Agrupado por Modelo

Considerando el tiempo en que los usuarios completaron la tarea 5, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.18: Valores estadísticos para el tiempo requerido en completar tarea 5 según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	45.8	17.3	40	27	105	14
Frecuencia	34.4	11.6	34	19	60	16.5
Markov	22	14.1	20	8	71	12

El diagrama de caja que muestra los resultados del tiempo requerido para cada modelo en la tarea 5 se presenta a continuación:

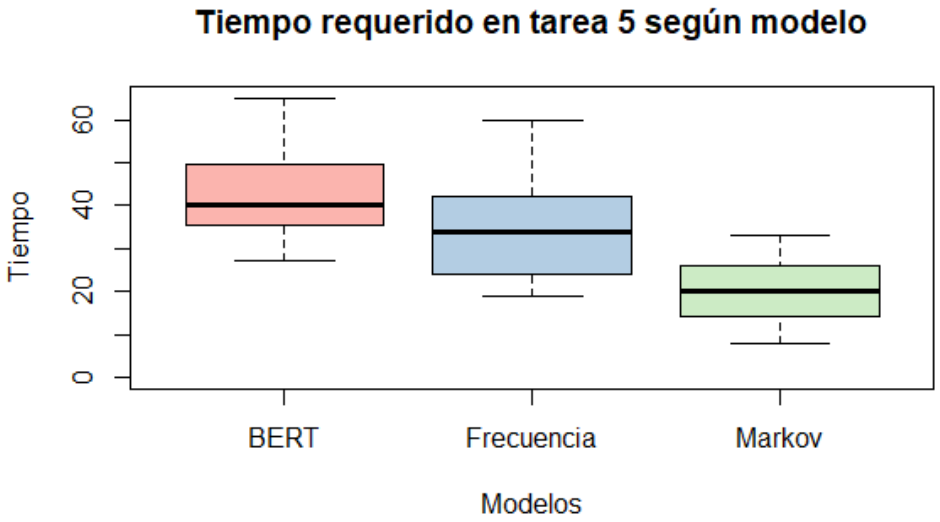


Figura 6.8: Tiempo requerido en tarea 5 según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value < 0.05$, lo cual nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.19: Prueba de Mann-Whitney para tiempo utilizado en tarea 5 según modelo

	BERT	**Frecuencia**
Frecuencia	0.0191	-
Markov	1e-05	0.0013

Considerando los datos presentados en 6.19, se puede determinar que existe una diferencia significativa entre los grupos de BERT y Frecuencia (magnitud de efecto mediana, delta de Cliff = 0.45), BERT y Markov (magnitud de efecto grande, delta de Cliff = 0.88) y Markov y Frecuencia (magnitud de efecto grande, delta de Cliff 0.64). Mostrando en todos estos pares un valor U menor a 0.05.

6.4.6.2. Agrupado por Género

Considerando el tiempo en que los usuarios completaron la tarea 5, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.20: Valores estadísticos para el tiempo requerido en completar tarea 5 según género del usuario

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	33.2	19.1	27	8	105	15
Masculino	35	15.6	34	8	71	20

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un valor U de 0.37, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no hay diferencia significativa entre ambos grupos.

6.4.7. Tarea 6

Se presentan los resultados de la variable Tiempo, que constituye al tiempo que le demoró al usuario completar la Tarea 6.

6.4.7.1. Agrupado por Modelo

Considerando el tiempo en que los usuarios completaron la tarea 6, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.21: Valores estadísticos para el tiempo requerido en completar tarea 6 según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	9.63	2.95	10	5	18	2.5
Frecuencia	26.7	5.34	26.5	20	41	4.75
Markov	15.6	9.93	12	5	48	5

El diagrama de caja que muestra los resultados del tiempo requerido para cada modelo en la tarea 6 se presenta a continuación:

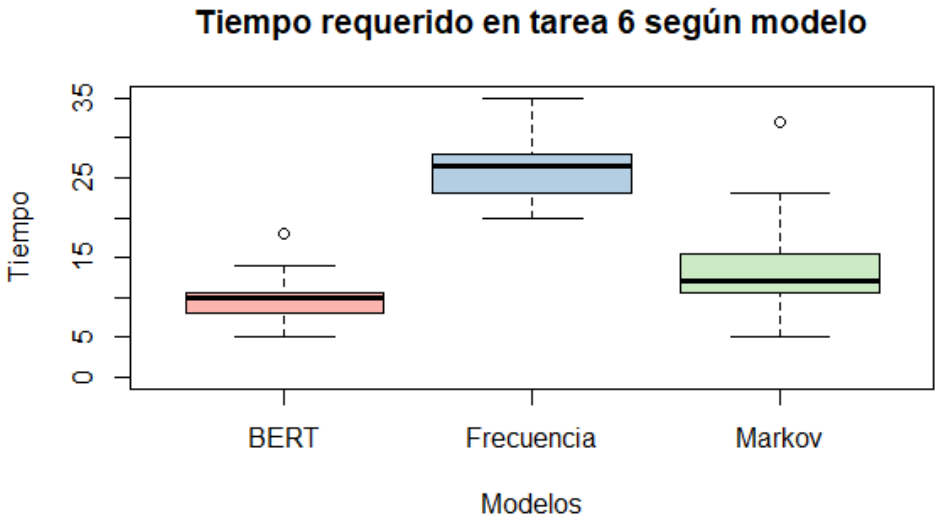


Figura 6.9: Tiempo requerido en tarea 6 según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value < 0.05$, lo cual nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.22: Prueba de Mann-Whitney para tiempo utilizado en tarea 6 según modelo

	BERT	Frecuencia
Frecuencia	6.2e-07	-
Markov	0.00246	0.00014

Considerando los datos presentados en 6.22, se puede determinar que existe una diferencia significativa entre los grupos de BERT y Frecuencia (magnitud de efecto grande, delta de Cliff = 0.51), BERT y Markov (magnitud de efecto grande, delta de Cliff = −0.57) y Markov y Frecuencia (magnitud de efecto grande, delta de Cliff 0.75). Mostrando en todos estos pares un valor U menor a 0.05.

6.4.7.2. Agrupado por Género

Considerando el tiempo en que los usuarios completaron la tarea 6, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.23: Valores estadísticos para el tiempo requerido en completar tarea 6 según género del usuario

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	15.8	9.51	12	5	48	13
Masculino	18.6	9.91	14	7	41	16.5

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un valor U de 0.23, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no hay diferencia significativa entre ambos grupos.

6.4.8. NASA-TLX

Se presentan los resultados del cuestionario NASA-TLX, aplicado a los usuarios posterior a la evaluación. Este instrumento tiene como fin medir la exigencia percibida durante el experimento. Se incluyen los datos recolectados durante este trabajo de título, como también los recolectados por Bahamonde [5] como grupo de control.

Cada una de las variables medidas corresponde a una pregunta en el cuestionario, donde el usuario debía categorizar su experiencia en un valor entre 1 y 20.

6.4.8.1. Exigencia mental

Se presentan los resultados de la variable Exigencia mental, la cual representa qué tan demandante mentalmente fue para el usuario escribir las frases.

Agrupado por modelo

Considerando la exigencia mental percibida por los usuarios durante la evaluación, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores

para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.24: Valores estadísticos para la variable de exigencia mental según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	7.95	5.34	8	1	15	9.5
Control	7.24	4.61	7	1	16	7
Frecuencia	6.44	4.84	5	1	20	6.5
Markov	6.26	4.16	5	1	15	6

El diagrama de caja que muestra los resultados de la exigencia mental percibida, agrupado según cada modelo se presenta a continuación:

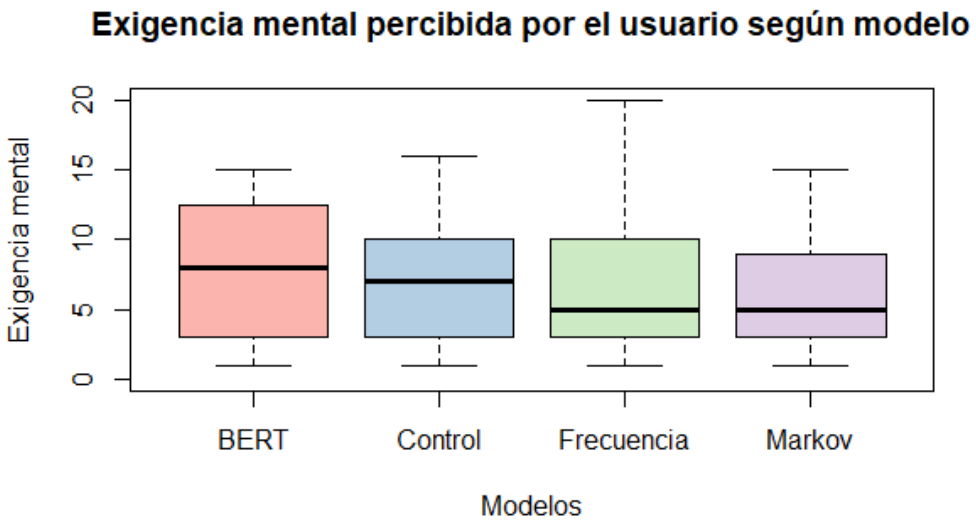


Figura 6.10: Exigencia mental según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p-value = 0.70$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 4 grupos.

Agrupado por género

Considerando la exigencia mental percibida por los usuarios durante la evaluación, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.25: Valores estadísticos para la variable de exigencia mental según género del usuario

Género	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
Femenino	6.55	4.59	7	1	15	8
Masculino	7.26	5.03	5	1	20	7.5

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un $p - value = 0.59$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 2 grupos.

6.4.8.2. Exigencia física

Se presentan los resultados de la variable Exigencia física, la cual representa qué tan demandante físicamente fue para el usuario escribir las frases.

Agrupado por modelo

Considerando la exigencia física percibida por los usuarios durante la evaluación, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.26: Valores estadísticos para la variable de exigencia física según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	3.53	3.75	2	1	13	3.5
Control	3.97	3.67	3	1	16	4
Frecuencia	2.78	4.54	1	1	20	1
Markov	2.42	2.19	1	1	9	2

El diagrama de caja que muestra los resultados de la exigencia física percibida, agrupado según cada modelo se presenta a continuación:

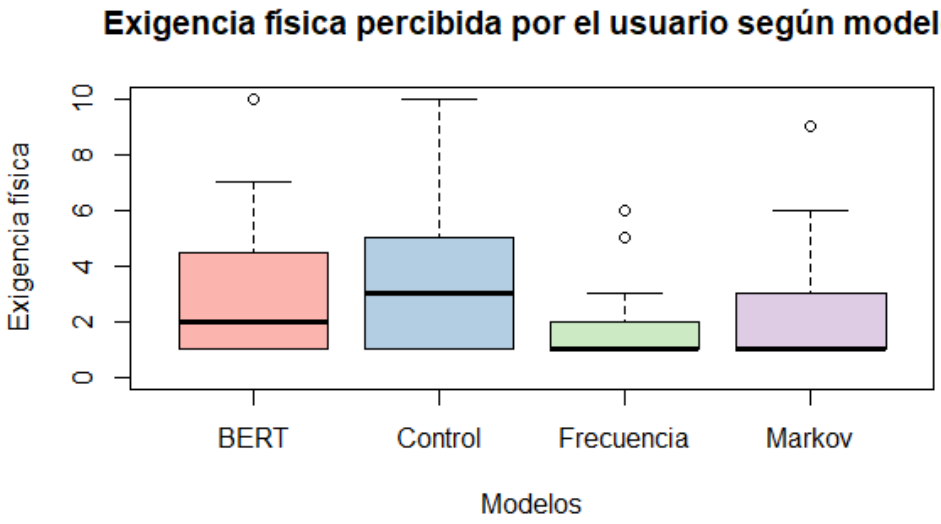


Figura 6.11: Exigencia física según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p-value = 0.15$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 4 grupos.

Agrupado por género

Considerando la exigencia física percibida por los usuarios durante la evaluación, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.27: Valores estadísticos para la variable de exigencia física según género usado

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	2.24	1.83	1	1	7	2
Masculino	3.63	4.73	1	1	20	2

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un $p-value = 0.53$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 2 grupos.

6.4.8.3. Exigencia temporal

Se presentan los resultados de la variable Exigencia temporal, la cual representa qué tan fuerte o rápido fue el ritmo impuesto para que el usuario escriba las frases.

Agrupado por modelo

Considerando la exigencia temporal percibida por los usuarios durante la evaluación, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.28: Valores estadísticos para la variable de exigencia temporal según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	7.89	3.68	8	1	15	5.5
Control	6.38	5.05	5	1	17	8
Frecuencia	5.78	3.37	5.5	1	11	5.25
Markov	7.32	4.87	6	1	18	5.5

El diagrama de caja que muestra los resultados de la exigencia temporal percibida, agrupado según cada modelo se presenta a continuación:

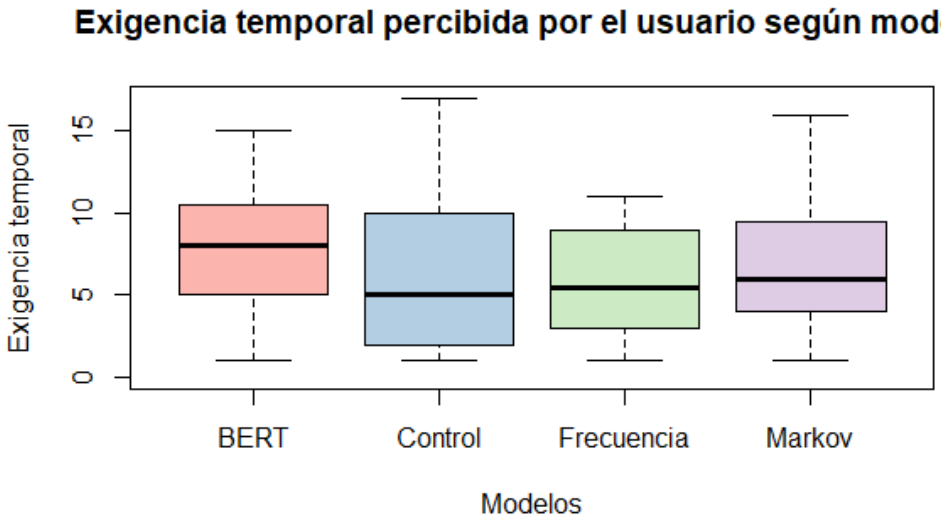


Figura 6.12: Exigencia temporal según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p-value = 0.33$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 4 grupos.

Agrupado por género

Considerando la exigencia temporal percibida por los usuarios durante la evaluación, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.29: Valores estadísticos para la variable de exigencia temporal según género usado

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	7.48	4.69	8	1	18	7
Masculino	6.52	3.27	6	1	15	3

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un $p-value = 0.48$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 2 grupos.

6.4.8.4. Rendimiento

Se presentan los resultados de la variable Rendimiento, la cual representa qué tan exitoso fue el usuario escribiendo las frases.

Agrupado por modelo

Considerando el rendimiento percibido por los usuarios durante la evaluación, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.30: Valores estadísticos para la variable de rendimiento según modelo usado

Modelo	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
BERT	15.7	4.00	17	6	20	3
Control	12.8	5.43	12	2	20	8
Frecuencia	14.8	6.16	17	1	20	9.5
Markov	18.2	2.30	19	13	20	2.5

El diagrama de caja que muestra los resultados del rendimiento percibido agrupado según cada modelo se presenta a continuación:

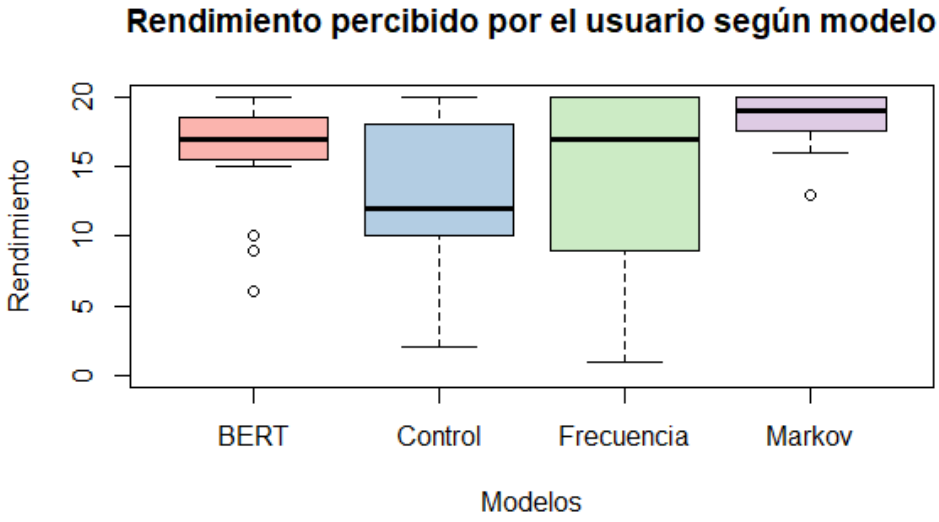


Figura 6.13: Rendimiento percibido según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value = 0.0029$, el cual al ser menor a 0.05, nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba pareada de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.31: Prueba pareada de Mann-Whitney para Rendimiento percibido agrupado por modelo

	BERT	**Control**	**Frecuencia**
Control	0.2065	-	-
Frecuencia	0.8539	0.2118	-
Markov	0.0446	0.0012	0.2005

Considerando los datos presentados en 6.31, se puede determinar que existe una diferencia significativa entre los grupos de Markov y Bert (magnitud de efecto mediano, delta de Cliff $= -0.45$) y Markov y Control (magnitud de efecto grande, delta de Cliff $= -0.63$). Mostrando en todos estos pares un valor U menor a 0.05.

Agrupado por género

Considerando el rendimiento percibido por los usuarios durante la evaluación, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.32: Valores estadísticos para la variable de rendimiento según género del usuario

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	17	3.67	18	6	20	4
Masculino	15.4	5.32	17	1	20	5.5

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un $p - value = 0.35$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 2 grupos.

6.4.8.5. Esfuerzo

Se presentan los resultados de la variable Esfuerzo, la cual representa qué tan duro tuvo que trabajar el usuario para escribir las frases.

Agrupado por modelo

Considerando el esfuerzo percibido por los usuarios durante la evaluación, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.33: Valores estadísticos para la variable de esfuerzo según modelo usado

Modelo	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
BERT	8.47	5.43	6	1	16 11	3
Control	6.86	4.41	7	1	16	7
Frecuencia	6.61	4.72	6.5	1	20	5.75
Markov	7.42	5.81	5	1	19	7.5

El diagrama de caja que muestra los resultados del esfuerzo percibido agrupado según cada modelo se presenta a continuación:

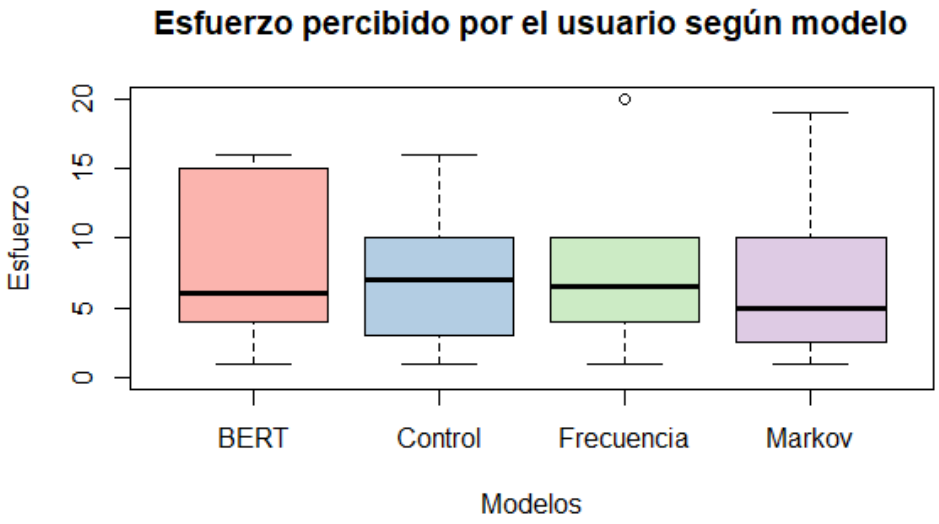


Figura 6.14: Exigencia mental según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p-value = 0.76$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 4 grupos.

Agrupado por género

Considerando el esfuerzo percibido por los usuarios durante la evaluación, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.34: Valores estadísticos para la variable de esfuerzo según género del usuario

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	6.86	5.39	6	1	20	7
Masculino	8.22	5.23	7	1	19	8

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un $p-value = 0.26$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 2 grupos.

6.4.8.6. Nivel de esfuerzo

Se presentan los resultados de la variable Nivel de esfuerzo, la cual representa qué tan inseguro, irritado y molesto se sintió el usuario al escribir las frases.

Agrupado por modelo

Considerando el nivel de esfuerzo percibido por los usuarios durante la evaluación, agrupados según el modelo utilizado para generar predicciones, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.35: Valores estadísticos para la variable de nivel de esfuerzo según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	4	3.70	3	1	13	4
Control	5.48	5.12	4	1	18	6
Frecuencia	3.11	3.55	2	1	16	2
Markov	2.84	3.08	1	1	11	2

El diagrama de caja que muestra los resultados del nivel de esfuerzo percibido agrupado según cada modelo se presenta a continuación:

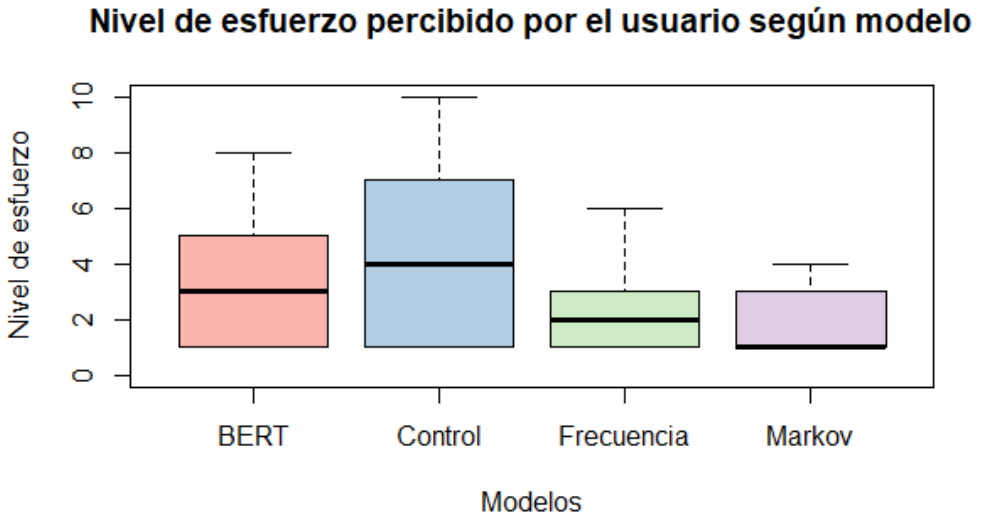


Figura 6.15: Nivel de esfuerzo según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p-value = 0.15$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 4 grupos.

Agrupado por género

Considerando el nivel de esfuerzo percibido por los usuarios durante la evaluación, agrupados según el género del usuario, se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR.

Tabla 6.36: Valores estadísticos para la variable de nivel de esfuerzo según género

Género	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
Femenino	3.10	3.26	2	1	13	3
Masculino	3.56	3.64	3	1	16	3

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Mann-Whitney para comparar las medias de los 2 grupos independientes.

Realizando el test de Mann-Whitney, se consigue un $p-value = 0.42$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 2 grupos.

6.4.9. Cuestionario post-evaluatorio

Se presentan los resultados del cuestionario de utilidad percibida, aplicado a los usuarios posterior a la evaluación. Este instrumento tiene como fin medir la utilidad percibida de las recomendaciones entregadas por el modelo durante el experimento. Debido a la naturaleza del cuestionario, solamente se incluyen la comparación entre los tres modelos integrados a la aplicación, sin distinguir entre género o edad del usuario.

Cada una de las variables medidas corresponde a una aseveración en el cuestionario, donde el usuario debía calificar su concordancia con ella con una categoría en la escala likert, pudiendo elegir entre: “Muy en desacuerdo”, “En desacuerdo”, “Ni de acuerdo, ni en desacuerdo”, “De acuerdo”, “Muy de acuerdo”, donde cada una de estas opciones se representa numéricamente con un valor del 1 al 5, respectivamente.

6.4.9.1. Me fue fácil identificar las recomendaciones

Se presentan los resultados de la concordancia de los usuarios con la aseveración “Me fue fácil identificar las recomendaciones”. Es decir, este apartado muestra que tan fácil les fue a los usuarios identificar las recomendaciones según el modelo que se les fue asignado.

Se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR:

Tabla 6.37: Valores estadísticos para la variable de facilidad de identificación según modelo usado

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	3.68	1.16	4	1	5	0.5
Frecuencia	3.11	1.41	3	1	5	2
Markov	3.68	1.34	4	1	5	2

El diagrama de caja que muestra los resultados de la facilidad percibida para identificar las recomendaciones, agrupado según cada modelo, se presenta a continuación:

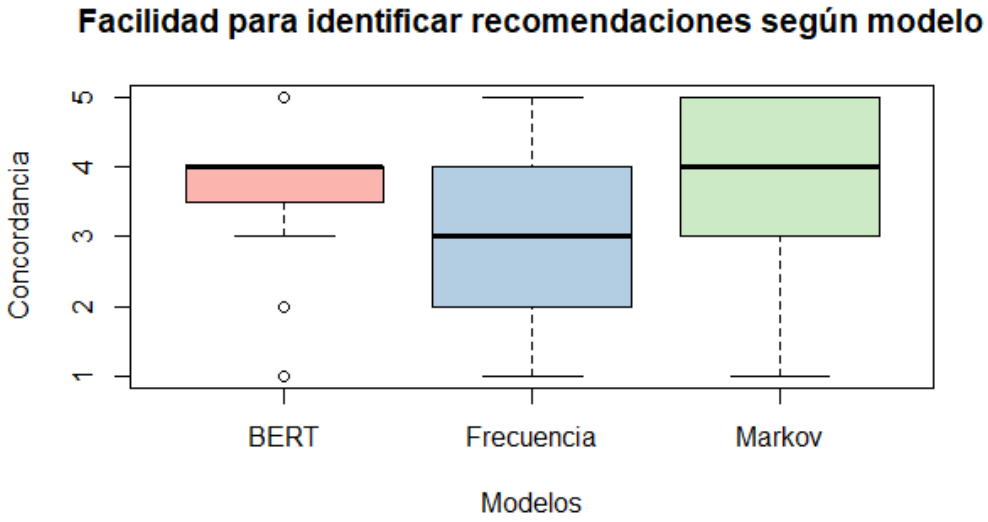


Figura 6.16: Facilidad de identificar las recomendaciones según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p-value = 0.35$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 3 grupos.

6.4.9.2. Me parecieron apropiadas las recomendaciones

Se presentan los resultados de la concordancia de los usuarios con la aseveración "Me parecieron apropiadas las recomendaciones". Es decir, este apartado muestra que tan apropiadas

les parecieron a los usuarios las recomendaciones según el modelo que se les fue asignado.

Se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR:

Tabla 6.38: Valores estadísticos para lo apropiadas que le parecieron las recomendaciones según modelo

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	4.47	0.513	4	4	5	1
Frecuencia	3.56	1.10	4	1	5	1
Markov	4.26	0.653	4	3	5	1

El diagrama de caja que muestra los resultados de la apropiadas que le parecieron a los usuarios las recomendaciones según cada modelo asignado, se presenta a continuación:

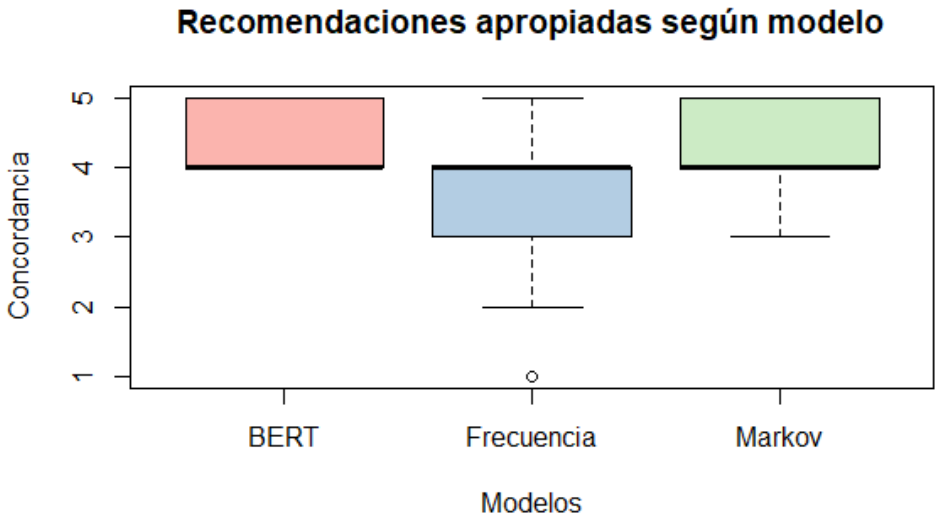


Figura 6.17: Recomendaciones apropiadas según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value = 0.0095$, el cual al ser menor a 0.05, nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba pareada de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.39: Prueba de Mann-Whitney para lo apropiadas que son las recomendaciones según modelo

	B	**F**
F	0.012	-
M	0.346	0.054

Considerando los datos presentados en 6.39, se puede determinar que existe una diferencia significativa entre los grupos de BERT y Frecuencia (magnitud de efecto grande, delta de Cliff = 0.51). Mostrando en este par un valor U menor a 0.05.

6.4.9.3. Me parecieron valiosas las recomendaciones

Se presentan los resultados de la concordancia de los usuarios con la aseveración "Me parecieron valiosas las recomendaciones". Es decir, este apartado muestra que tan valiosas les parecieron a los usuarios las recomendaciones según el modelo que se les fue asignado.

Se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR:

Tabla 6.40: Valores estadísticos para lo valiosas que le parecieron las recomendaciones según modelo

Modelo	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
BERT	4.26	0.653	4	3	5	1
Frecuencia	3.56	0.984	4	2	5	1
Markov	4.47	0.513	4	4	5	1

El diagrama de caja que muestra los resultados de lo valiosas que le parecieron a los usuarios las recomendaciones, agrupados según cada modelo, se presenta a continuación:

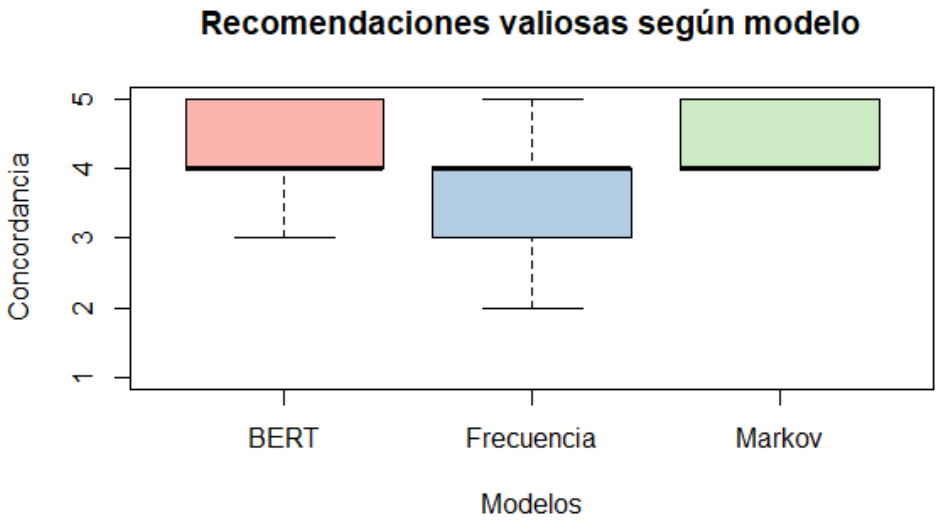


Figura 6.18: Recomendaciones valiosas según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value = 0.0055$, el cual al ser menor a 0.05, nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba pareada de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.41: Prueba de Mann-Whitney sobre lo valiosas que son las recomendaciones según modelo

	BERT	**Frecuencia**
Frecuencia	0.035	-
Markov	0.346	0.008

Considerando los datos presentados en 6.41, se puede determinar que existe una diferencia significativa entre los grupos de BERT y Frecuencia (magnitud de efecto mediano, delta de Cliff = 0.41) y Markov y Frecuencia (magnitud de efecto grande, delta de Cliff = -0.54). Mostrando en cada uno de estos pares un valor U menor a 0.05.

6.4.9.4. Me parecieron útiles las recomendaciones

Se presentan los resultados de la concordancia de los usuarios con la aseveración “Me parecieron útiles las recomendaciones”. Es decir, este apartado muestra que tan útiles les parecieron a los usuarios las recomendaciones según el modelo que se les fue asignado.

Se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR:

Tabla 6.42: Valores estadísticos para lo útiles que le parecieron las recomendaciones según modelo

Modelo	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo	IQR
BERT	4.26	0.562	4	3	5	1
Frecuencia	3.67	1.03	4	2	5	1
Markov	4.53	0.612	5	3	5	1

El diagrama de caja que muestra los resultados de la apropiadas que le parecieron a los usuarios las recomendaciones según cada modelo asignado, se presenta a continuación:

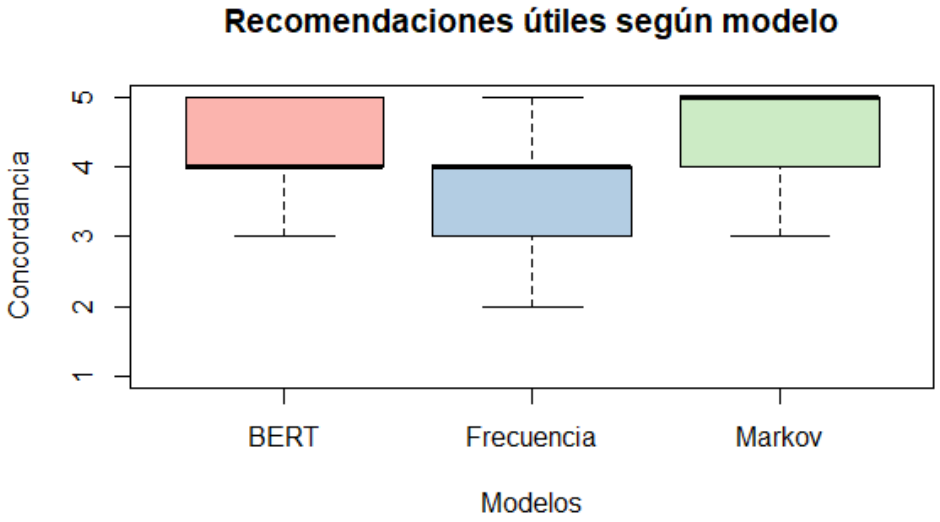


Figura 6.19: Recomendaciones útiles según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value = 0.012$, el cual al ser menor a 0.05, nos permite determinar que existe un efecto conjunto dentro de los datos. Lo que significa que hay por lo menos dos variables distintas entre sí. Para realizar una comparativa pareada entre los grupos experimentales aplicamos una prueba pareada de Mann-Whitney, la cual entrega los siguientes resultados:

Tabla 6.43: Prueba de Mann-Whitney sobre lo útiles que son las recomendaciones según modelo

	BERT	**Frecuencia**
Frecuencia	0.099	-
Markov	0.143	0.021

Considerando los datos presentados en 6.43, se puede determinar que existe una diferencia significativa entre los grupos de BERT y Frecuencia (magnitud de efecto pequeña, delta de Cliff = 0.32) y Markov y Frecuencia (magnitud de efecto grande, delta de Cliff = -0.48). Mostrando en cada uno de estos pares un valor U menor a 0.05.

6.4.9.5. Es probable que vuelva a usar la aplicación

Se presentan los resultados de la concordancia de los usuarios con la aseveración "Es probable que vuelva a usar la aplicación". Es decir, este apartado muestra que tan probable es que los usuarios vuelvan a usar la aplicación, agrupando sus respuestas según el modelo que se les fue asignado.

Se consiguen los siguientes valores para promedio, mediana, desviación estándar, mínimo, máximo e IQR:

Tabla 6.44: Valores estadísticos para la probabilidad de que vuelva a usar la aplicación según modelo

Modelo	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**	**IQR**
BERT	4.47	0.697	5	3	5	1
Frecuencia	4.17	0.857	4	2	5	1
Markov	4.32	0.820	4	2	5	1

El diagrama de caja que muestra los resultados de la probabilidad de que el usuario vuelva a usar la aplicación según cada modelo asignado, se presenta a continuación:

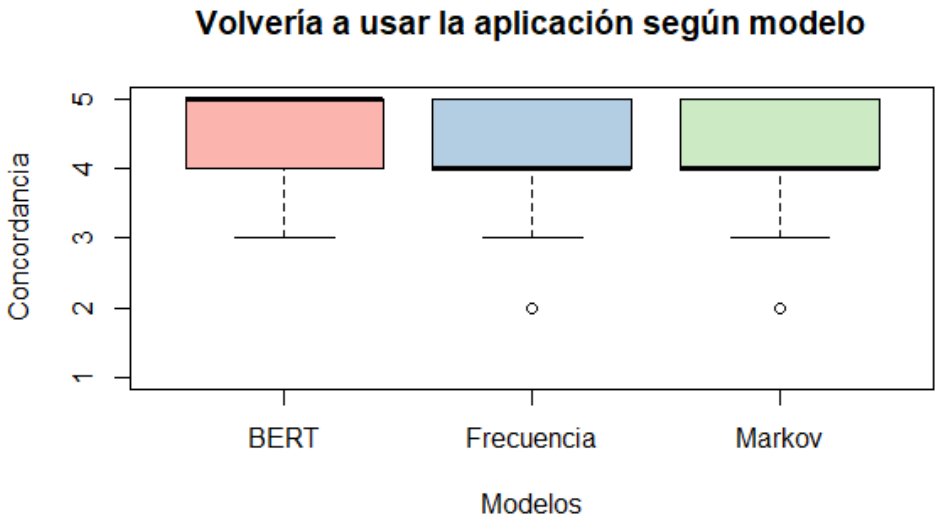


Figura 6.20: Volvería a usar la aplicación, según modelo asignado

Al realizar el test de Shapiro-Wilk, se consigue un valor menor a 0.05, lo que permite rechazar la hipótesis nula de normalidad de los datos. Por ende, se utilizará el test de Kruskal-Wallis para comparar las medias de los 3 grupos independientes.

Realizando el test de Kruskal-Wallis, se consigue un $p - value = 0.50$, el cual al ser mayor a 0.05, nos permite determinar que no existe un efecto conjunto dentro de los datos. Es decir, que no existe diferencia significativa entre los 3 grupos.

6.4.10. Comentarios cualitativos

Se le pidió a cada usuario que durante la evaluación diera comentarios sobre cómo se sentía mientras usaba la aplicación y realizaba las tareas especificadas (para más detalles sobre el proceso experimental, revisar la sección5.2.3). Adicionalmente, el cuestionario post-evaluatorio incluye también una pregunta abierta en donde el usuario podía entregar comentarios adicionales con respecto al tablero digital y el sistema de recomendaciones que tiene integrado (para ver el cuestionario y su objetivo, revisar la sección 5.2.5). Los comentarios entregados se incluyen en el presente apartado.

Con respecto a la aplicación como tal y el objetivo que tiene de mejorar la comunicación en una población de niños con PC, los usuarios opinaron que *"Ayuda a que sea mas sencilla la comunicación, especialmente entre cuidadora y niño"* y que *"Buena aplicación, muy útil para los niños"*. Estos comentarios muestran que para los usuarios que cumplen el perfil de cuidador de niños con PC (para más detalle ver sección 5.2.1), el objetivo de mejorar la comunicación se siente cumplido.

Los usuarios también concordaron en que la aplicación se siente sencilla de usar, incluso para usuarios con nula experiencia en soluciones de este tipo. Con respecto a esto, los usuarios opinaron que: *"La aplicación tiene una baja curva de aprendizaje, siendo fácil de usar la*

primera vez", *"Muy novedoso, bueno y amigable para todos"* y *"Súper bueno, me parece súper útil al acostumbrarse uno"*. Lo que muestra que se consiguió uno de los objetivos planteados en 6.3, mantener la accesibilidad y usabilidad conseguida como resultado del trabajo de Bahamonde [5], en particular, mantener la aplicación lo más sencilla de utilizar para usuarios nuevos.

El orden con el cual se presentan los pictogramas al usuario fue objeto de comentarios, en donde se dijo que: *"Confunde que no sean solo sujetos en primera plana, quizás se podrían poner más como amigo/amiga o mascota"*, haciendo alusión a que la primera página contiene sujetos y verbos,*"No logré ver ni entender el orden"*, *"¿No está ordenado por abecedario? Debería estarlo"*. Habiendo concordancia entre algunos usuarios con respecto a que la forma en la que se distribuyen los pictogramas no es la idónea, debido no se entiende en el primer uso. Como oportunidad de mejora, los usuarios entregaron las siguientes opciones a modo de comentario: *"Categorizaría los pictogramas por tipo de palabra"*, *"Deberían haber divisiones según categoría de significado de la palabra"* y *"Deberían estar por página y categoría"*, donde todos estos concordaron que debería haber una categorización adicional a la existente que utiliza la categoría sintáctica de la palabra. Esto sugiere que el layout actual es responsable de provocar errores de paginación innecesaria en la búsqueda de pictogramas, en donde los usuarios cambian de página sin darse cuenta que el pictograma que buscaban estaba en realidad en la página actual. Se puede concluir así, que el teclado actual provoca un uso ineficiente de la aplicación.

Con respecto a los materiales usados durante la evaluación, algunos usuarios consideraron que la tablet era muy pequeña para utilizar correctamente la aplicación, entregando como comentarios que: *"Las letras y dibujos deben ser más grandes, y estar distribuidos de mejor forma"*, *"El tamaño de letra muy chico"*, *"La tablet es muy pequeña para poder ver bien cada pictograma"* y *"La lectura es primordial para que la aplicación funcione bien en una tableta tan chica, me enfoque más en la palabra que en el pictograma"*. Lo anterior evidencia que el tamaño del material usado podría ser muy pequeño para la evaluación; sin embargo, Bahamonde consiguió como resultado de su experimentación que no existe diferencia significativa entre el uso de esta aplicación en una tablet de 8" y una de 12.4" [5], contrastando así con los comentarios entregados por los usuarios durante esta evaluación.

Durante el desarrollo de la solución, la doctora Hidalgo indicó la relevancia del pictograma *"dolor"* para que las necesidades de comunicación de un niño con PC quedarán mejor cubiertas (para más detalle ver la sección 3). Sin embargo, este no fue bien recibido por los usuarios, entregando comentarios que denotan el poco entendimiento del mismo, tales como: *"Me parece muy poco significativo y representativo el dibujo de dolor"* , *"¿Esto es dolor? No se nota"* y *"Me parece muy poco significativo el icono de dolor"*. Evidenciando la dificultad que trae aumentar el vocabulario de la aplicación.

Los usuarios tuvieron comentarios con respecto a las recomendaciones entregadas durante y posterior a la evaluación.

Durante la evaluación algunos usuarios que identificaron las recomendaciones, tuvieron los siguientes comentarios sobre su utilidad: *"Me di cuenta en la última frase donde estaba todo y fue súper útil"*, *"Algunas te iban saliendo al tiro y eso era súper útil para escribir mas*

rápido", *"Las recomendaciones eran buenas, inconscientemente me iba a ellas"* y *"Buenísima idea las recomendaciones, algunas veces servían mucho"* complementandosé así con lo dicho por los expertos de dominio, los cuales mencionaron que este avance de la aplicación será de mucha utilidad al momento de utilizar el tablero (para más detalle ver la sección 6.3).

Con respecto a la forma en la que se presentaron las recomendaciones, se opinó que: *"A medida que avancé fui buscando arriba primero y luego abajo si no estaba la opción"* y *"Atención se me desviaba al recuadro azul al tiro, tiendes a ver eso primero y luego voy secuencialmente"*. Esto pone en evidencia que para algunos usuarios la forma en la que se presentaron las recomendaciones es adecuada, validando las decisiones de diseño tomadas durante la integración de las predicciones al tablero de comunicación (para más detalle ver 4). Esto concuerda también con la experiencia que demostraron algunos usuarios al usar la aplicación, los cuales no veían las recomendaciones hasta pasar de página y realizar un barrido secuencial del tablero, poniendo en evidencia que la decisión de colocar las predicciones en la parte superior del tablero fue la correcta para maximizar la visibilidad de esta sección.

Las opiniones con respecto a las recomendaciones contrastó entre los usuarios, debido a que muchos no las identificaron durante la evaluación. Por ejemplo, algunos usuarios opinaron que: *"Ni me fije en el recuadro de arriba, miraba siempre al medio"*, *"Las recomendaciones se mimetizan mucho con las normales, el azul no es suficientemente visible, falta más separación"* y *"No las use porque no me di cuenta que eran. Pensé que eran más de lo mismo"*, mostrando que la forma de presentar las recomendaciones puede no haber sido adecuada para algunos usuarios. Sin embargo, estos mismos usuarios opinaron que si bien no identificaron las recomendaciones, sí les parecían de utilidad luego de haberse explicado su existencia: *"Las recomendaciones me podrían haber ayudado a comunicarme más rápido si las hubiera visto"*, *"Me parecen súper útiles porque todas las aplicaciones hoy en día te adivinan lo que viene a continuación"*.

Un punto levantado por los usuarios en forma de comentario fue que el hecho de que las predicciones cambien con cada palabra escrita puede provocar confusión pudiendo generar errores. Por ejemplo, si las palabras $p1, p2, p3, p4$ están como recomendaciones actuales, y el usuario presiona p1, el modelo volverá a generar predicciones, las que son $p4, p1', p2', p3'$, teniendo p4 presente en ambos conjuntos, pero cambiando el lugar donde se presenta.

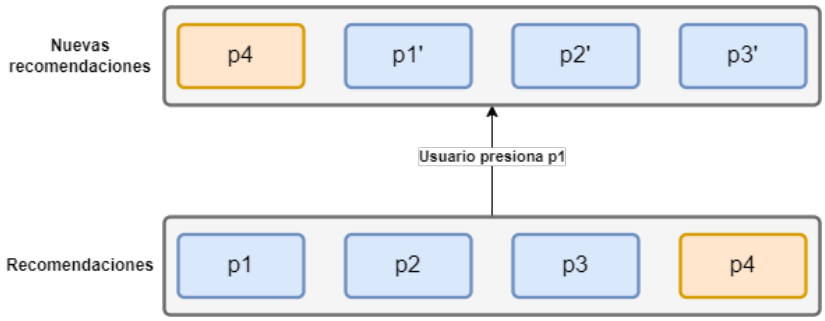


Figura 6.21: Cambio de recomendaciones luego de que usuario presione p1

Esto presentó problemas para los usuarios, ya que algunos presionaban la posición antigua de p4 antes que el modelo logre generar las predicciones. Esto ocurre porque existe un tiempo entre el momento en cuando se presiona un pictograma y el momento que el modelo entrega

las predicciones al usuario. Este comportamiento se ve reflejado en algunos comentarios entregados por los usuarios, tales como: “No me gustaba que cambiaran de posición cuando lo apretaba”, “Me confunde que cambien de lugar tan seguido” y “Fui a apretar donde estaba *“quiero”* antes y cambió de lugar el tablero”. Es importante modificar la solución para que este problema no influya en la experiencia de uso del tablero.

Capítulo 7

Discusión y análisis

En este capitulo se presenta un análisis de los resultados obtenidos de todas las evaluaciones realizadas sobre la solución (para más detalle ver los capítulos 5 y 6).

7.1. Comparación con tablero de control

Uno de los objetivos principales de la integración del modelo de predicción al tablero digital de comunicación creado por Bahamonde [5] era mantener la accesibilidad y usabilidad que él había alcanzado como resultado de su trabajo de título. Para evaluar si esto se cumplió se utilizan los resultados obtenidos en el focus group de expertos de dominio (ver sección 6.3), en conjunto con los resultados del cuestionario NASA-TLX (ver sección 6.4.8).

Considerando la inspección de usabilidad realizada, siguiendo una evaluación heurística con expertos de dominio, tanto en manejos de pacientes como con diseño de usabilidad y diseño de interacción (para más detalle ver 6.3). Todos los presentes estuvieron de acuerdo que la usabilidad de la aplicación no se degradó al integrar la recomendaciones, en particular consideraron que el sistema de predicciones le otorgaría una utilidad extra al usuario y en ningún caso empeoraría la experiencia de uso.

Esta evaluación se complementa con los resultados obtenidos mediante NASA-TLX (para más detalle sobre la evaluación ver sección 5.2.3). Debido a que tenemos los datos de control registrados por Bahamonde [5], podemos comparar los usuarios con algún sistema de recomendación integrado a la aplicación y el grupo de control, es decir el grupo con la aplicación desarrollada sin predicciones. Teniendo consideración las siguientes variables del cuestionario: Exigencia mental, Exigencia física, Exigencia temporal, Esfuerzo y Nivel de esfuerzo, se registran datos que no muestran diferencia significativa entre los grupos experimentales (para más detalle sobre esto ver 6.4.8). Esto nos permite decir que:

- La exigencia mental percibida por los usuarios al utilizar la aplicación sin recomendaciones es la misma que la percibida utilizando la aplicación con cada uno de los modelos de predicción integrados.
- La exigencia física percibida por los usuarios al utilizar la aplicación sin recomendaciones es la misma que la percibida utilizando la aplicación con cada uno de los modelos de predicción integrados.

- La exigencia temporal percibida por los usuarios al utilizar la aplicación sin recomendaciones es la misma que la percibida utilizando la aplicación con cada uno de los modelos de predicción integrados.
- El esfuerzo que tuvieron que aplicar los usuarios al utilizar la aplicación sin recomendaciones es la misma que el que tuvieron que aplicar los usuarios utilizando la aplicación con cada uno de los modelos de predicción integrados.
- Lo inseguro, irritado y molesto que se siente el usuario al utilizar la aplicación sin recomendaciones es la misma que las sensaciones percibidas por cada usuario al utilizar la aplicación con cada uno de los modelos de predicción integrados.

Considerando estos puntos, podemos determinar que la integración del modelo de predicciones a la aplicación no sacrifica la accesibilidad y usabilidad alcanzada como resultado del trabajo de Bahamonde [5], en particular, no aumenta la carga de trabajo percibida por el usuario al usar la aplicación.

El objetivo principal que guió este trabajo de título es que el sistema de predicción diseñado debe mejorar la utilidad percibida de la aplicación (para más detalle ver 1.1.1). Para analizar cómo las predicciones lograron mejorar la comunicación se comparan los resultados obtenidos en la variable Rendimiento de NASA-TLX (ver sección 6.4.8), en conjunto con los tiempos que le toma a cada usuario resolver las primeras tres tareas (para más detalle ver sección 6.4). En el análisis se usan los resultados obtenidos por el grupo de Control, en contraste con los que se obtuvieron integrando cada uno de los modelos a la aplicación.

La variable de Rendimiento del cuestionario NASA-TLX mide en una escala del 1 al 20 qué tan exitoso se sintió el usuario escribiendo las frases presentadas por el investigado durante el experimento (para más detalle ver 5.2.3). Considerando esta métrica, los usuarios que no tenían un modelo de recomendación integrado percibieron un nivel de rendimiento de 12.8, mientras que los usuarios que usaron los modelos de BERT, Frecuencia y Markov percibieron un rendimiento de 15.7,14.8 y 18.2 respectivamente. En particular se encontró una diferencia significativa, con una magnitud de efecto grande, entre el Modelo de Markov, donde ningún usuario reportó un rendimiento menor a 13 y un 50 % percibió un rendimiento mayor o igual a 19, y el de Control, donde un 50 % de los usuarios percibieron un rendimiento menor a 12 y solamente un 13 % de los usuarios reportó un rendimiento mayor o igual a 19 (para más detalle ver la tabla 6.30 y la figura 6.13)). Estos datos nos permiten decir que los usuarios que utilizaron la aplicación con el modelo de Markov integrado, reportaron un rendimiento **significativamente mejor** que los usuarios que no tenían ningún modelo de predicción integrado. Esto impacta positivamente en el valor percibido por el usuario que utiliza el modelo de predicción de Markov.

Comparando el tiempo que les tomó realizar las primeras tres tareas a los usuarios sin recomendaciones, con el tiempo utilizado por usuarios con algún modelo de predicción integrado, podemos ver una diferencia significativa, entre el grupo de Control y uno o más modelos de predicción. Se desglosa a continuación el análisis según cada tarea.

En la tarea 1 (ver 6.4.2 para más detalle), al agrupar los datos según el modelo utilizado, se consigue que los usuarios del grupo de Control utilizaron en promedio 29.2 segundos para

completar la tarea, mientras que para los modelos de BERT, Frecuencia y Markov se usaron en promedio 13.2, 14.1 y 15.1 segundos respectivamente (para más detalle ver tabla 6.6). Se encontró una diferencia significativa entre el grupo de Control y cada uno de los modelos integrados a la aplicación, con una magnitud de efecto grande (para más detalle ver tabla 6.7). Por ende, se puede determinar que los modelos de BERT, Frecuencia y Markov permiten un uso **mucho más eficiente** de la aplicación que un tablero que no tenga integrado un sistema de recomendaciones al escribir la frase *"yo quiero agua"*. Esto impacta positivamente en la utilidad percibida por el usuario que utiliza los modelos de predicciones, ya que le permite comunicarse de manera más eficiente usando la aplicación.

En la tarea 2 (ver 6.4.3 para más detalle), al agrupar los datos según el modelo utilizado, se consigue que los usuarios del grupo de Control utilizaron en promedio 69.7 segundos para completar la tarea, mientras que para los modelos de BERT, Frecuencia y Markov se usaron en promedio 11.6, 30.8 y 15.8 segundos respectivamente (para más detalle ver tabla 6.9). Se encontró una diferencia significativa entre el grupo de Control y cada uno de los modelos integrados a la aplicación, con una magnitud de efecto grande (para más detalle ver tabla 6.10). Por ende, se puede determinar que los modelos de BERT, Frecuencia y Markov permiten un uso **más eficiente** de la aplicación que un tablero que no tenga integrado un sistema de recomendaciones al escribir la frase *"yo quiero jugar"*. Esto impacta positivamente en la utilidad percibida por el usuario que utiliza los modelos de predicciones, ya que le permite comunicarse de manera más eficiente usando la aplicación.

En la tarea 3 (ver 6.4.4 para más detalle), al agrupar los datos según el modelo utilizado, se consigue que los usuarios del grupo de Control utilizaron en promedio 43.6 segundos para completar la tarea, mientras que para los modelos de BERT, Frecuencia y Markov se usaron en promedio 31.6, 22.2 y 34.4 segundos respectivamente (para más detalle ver tabla 6.9). En particular, se encontró una diferencia significativa entre el grupo de Control y el modelo de Frecuencia, con una magnitud de efecto grande (para más detalle ver tabla 6.10). Por ende, se puede determinar que el modelo de Frecuencia permite un uso **más eficiente** de la aplicación que un tablero que no tenga integrado un sistema de recomendaciones al escribir la frase *"mamá quiero galletas jugo"*. Esto impacta positivamente en la utilidad agregada por el usuario que utiliza el modelo de predicción de Frecuencia, ya que le permite comunicarse de manera más eficiente usando la aplicación.

Utilizando este análisis, podemos determinar que el modelo de predicción diseñado, al ser integrado a la aplicación produce un aumento en la percepción de rendimiento si se utiliza un modelo de Markov y también produce un uso **mas eficiente** si utilizamos un modelo de BERT, Markov, y Frecuencia, encontrando una mayor diferencia entre el grupo de control y grupo que uso el modelo de Frecuencia en las primeras tres tareas. **Esto nos permite decir que se cumplió el objetivo principal, al presentar una solución que es más eficiente y produce mayor rendimiento según los usuarios.**

7.2. Comparación intra-modelo

Veremos a continuación una comparación entre cada modelo diseñado, para así determinar cuál es el que mejor resuelve el problema planteado en este trabajo de título. En

particular, ver qué modelo logra mejores resultados intrínsecamente y extrínsecamente, para posteriormente evaluar las concesiones existentes entre el desempeño de un sistema de NLP y la utilidad percibida por el usuario al aplicarlo al tablero de comunicación digital.

7.2.1. Comparación intrínseca

Considerando los datos presentados como resultados de la evaluación intrínseca 5.1 de la solución, se presenta el siguiente análisis.

El modelo de Markov posee los valores más altos para la métrica Top-N, ya que las predicciones que realiza son correctas un 53.85 % de las veces si entrega solamente un resultado como opción, 64.98 % de las veces si entrega dos resultados como opciones, 72.08 % de las veces si entrega tres resultados como opciones y 75.53 % de las veces si entrega cuatro resultados como opciones (para más detalles ver tabla 6.1). Este último valor (representando la métrica de Top-4), es lo que se buscaba maximizar durante el desarrollo de la solución, debido a que es la cantidad de predicciones que se necesitan entregar al integrar el modelo al tablero de comunicación digital. Es por esto que se puede determinar que el modelo de Markov genera predicciones **más precisas** que los modelos de BERT y Frecuencias.

El valor de Perplexity no se consideró dentro de la determinación del mejor modelo intrínseco (para más detalles ver 6.2.2).

Utilizando estas métricas, se determina que el modelo de Markov posee el mejor desempeño de forma intrínseca.

Este resultado contrasta con el primer punto de la hipótesis de trabajo presentada en la sección 6.1, donde se pensó que el mejor modelo iba a ser BERT. Este resultado se puede deber a que el problema a modelar no era tan complejo, provocado por: el limitado vocabulario utilizado por la aplicación, la falta de conectores en las oraciones que se pueden usar y por el reducido corpus utilizado para el entrenamiento de los modelos.

Contrastando con las hipótesis de trabajo planteadas en la sección 6.1, se determina que el modelo con mejor desempeño intrínseco es el que utiliza cadenas de Markov y no el que utiliza BERT.

7.2.2. Comparación extrínseca

Considerando los datos presentados como resultados de la evaluación extrínseca 5.2 de la solución. Se desprende el siguiente análisis a partir de estos resultados.

Al analizar los tiempos que utilizaron los usuarios para completar en promedio cada una de las tareas, se pueden sacar conclusiones sobre qué modelo es en promedio más eficiente, independiente de la frase que se escriba (ver 6.4.1 para más detalle). Al agrupar los datos según el modelo utilizado, se consigue que el grupo que utilizó el modelo de Markov para generar recomendaciones requirió en promedio 19.7 segundos para completar cada tarea, este valor es de 20.4 para el modelo de BERT y 23.2 para el grupo que utilizó el modelo de

Frecuencias (para más detalle ver tabla 6.3 y 6.6). Se encontró una diferencia significativa entre el grupo que utilizó el modelo de Frecuencia y los otros dos grupos que usaron BERT y Markov para generar recomendaciones, la magnitud de efecto para ambos pares se determinó como grande (para más detalle ver tabla 6.7). Considerando esto, se puede determinar que al usar los modelos de BERT y Markov se permite un uso **mucho más eficiente** de la aplicación, en promedio, que un tablero que utilice un modelo de Frecuencias como sistema de recomendaciones. Esto impacta positivamente en la utilidad percibida al permitir comunicarse de manera más eficiente usando la aplicación.

Para poder discernir entre la diferencia de eficiencia que produce usar un modelo de Markov o un modelo de BERT, se analizarán las tareas 4, 5 y 6.

En la tarea 4 (ver 6.4.5 para más detalle), al agrupar los datos según el modelo utilizado, se consigue que el grupo que utilizó el modelo de BERT para generar recomendaciones requirió en promedio 10.8 segundos para completar esta tarea; este valor es de 10.8 para el modelo de Frecuencia y 15.3 para el grupo que utilizó el modelo de Markov (para más detalle ver tabla 6.15 y 6.7). Se encontró una diferencia significativa entre el grupo que utilizó el modelo de Markov y los otros dos grupos que usaron BERT y Frecuencia para generar recomendaciones, la magnitud de efecto para ambos pares se determinó como mediana (para más detalle ver tabla 6.16). Considerando esto, se puede determinar que al usar los modelos de BERT y Frecuencia se permite un uso **más eficiente** de la aplicación, que un tablero que utilice un modelo de Markov como sistema de recomendaciones al escribir la frase *"papá comprar leche"*.

En la tarea 5 (ver 6.4.6 para más detalle), al agrupar los datos según el modelo utilizado, se consigue que el grupo que utilizó el modelo de BERT para generar recomendaciones requirió en promedio 45.8 segundos para completar esta tarea, este valor es de 34.4 para el modelo de Frecuencia y 22 para el grupo que utilizó el modelo de Markov (para más detalle ver tabla 6.18 y figura 6.8). Se encontró una diferencia significativa entre el grupo que utilizó el modelo de Markov y los otros dos grupos que usaron BERT y Frecuencia para generar recomendaciones, la magnitud de efecto para ambos pares se determinó como grande. También se encontró una diferencia significativa, con una magnitud de efecto mediana, entre los grupos que usaron como modelo de predicción BERT y los que usaron el modelo de Frecuencia (para más detalle ver tabla 6.19). Considerando esto, se puede determinar que el usar el modelo de Markov permite un uso **mucho más eficiente** de la aplicación, que si se utilice un modelo de BERT o uno de Frecuencia como sistema de recomendaciones al escribir la frase *"yo llorar dolor pie"*. Se desprende de esto también que el usar el modelo de Frecuencia permite un uso **más eficiente** de la aplicación, que si se utilice un modelo de BERT como sistema de recomendaciones al escribir la frase *"yo llorar dolor pie"*.

En la tarea 6 (ver 6.4.7 para más detalle), al agrupar los datos según el modelo utilizado, se consigue que el grupo que utilizó el modelo de BERT para generar recomendaciones requirió en promedio 9.63 segundos para completar esta tarea, este valor es de 26.7 para el modelo de Frecuencia y 15.6 para el grupo que utilizó el modelo de Markov (para más detalle ver tabla 6.21 y 6.9). Se encontró una diferencia significativa entre el grupo que utilizó el modelo de BERT y los otros dos grupos que usaron Markov y Frecuencia para generar recomendaciones, la magnitud de efecto para ambos pares se determinó como grande. También se encontró una diferencia significativa, con una magnitud de efecto grande, entre los grupos que usaron como

modelo de predicción Markov y los que usaron el modelo de Frecuencia (para más detalle ver tabla 6.22). Considerando esto, se puede determinar que el usar el modelo de BERT permite un uso **mucho más eficiente** de la aplicación, que si se utilice un modelo de Markov o uno de Frecuencia como sistema de recomendaciones al escribir la frase *"yo dormir cama mamá"*. Se desprende de esto también que el usar el modelo de Markov permite un uso **más eficiente** de la aplicación, que si se utilice un modelo de Frecuencia como sistema de recomendaciones al escribir la frase *"yo dormir cama mamá"*.

Es importante considerar la naturaleza de las frases al momento de analizar estos resultados. La tarea 5 fue diseñada con el objetivo tal que su estructura sintáctica se asemeje lo más posible a las diseñadas por la doctora Hidalgo para el entrenamiento de los modelos, mientras que la tarea 6 fue diseñada para evaluar que tan bien podían los modelos capturar la información semántica dentro de una oración. Cabe destacar que ninguna de estas estaba presente en el corpus diseñado como evidencia. Considerando esto, se puede decir que el modelo de Markov es **mucho más eficiente** que los otros dos al momento de realizar predicciones sobre oraciones con una estructura sintácticamente correcta, según lo definido como base por la doctora experta de dominio. Por otro lado, el modelo de BERT es **mucho más eficiente** que los otros dos al generar recomendaciones usando como evidencia frases con mucha relación semántica entre las palabras que la conforman.

Para analizar la percepción que tuvieron los usuarios con respecto a las recomendaciones entregadas por cada modelo de NLP, se comparan los resultados obtenidos en el cuestionario post-evaluatorio (para más detalle ver 5.2.5). Solamente se encontró diferencia significativa entre los grupos experimentales sobre la concordancia en relación a las aseveraciones: "Me parecieron *apropiadas* las recomendaciones", "Me parecieron *valiosas* las recomendaciones" y "Me parecieron *útiles* las recomendaciones".

Con respecto a lo apropiadas que les parecieron las recomendaciones a los usuarios, los que utilizaron un modelo de Frecuencia poseen un promedio de concordancia menor que el grupo que usó un Modelo de Markov y el que usó uno de BERT (para más detalle ver la tabla 6.38 y la figura 6.17). Esta diferencia se determinó como significativa, con un efecto de magnitud grande para el par de modelos BERT y Frecuencia (ver 6.39). Por lo tanto, se puede determinar que los usuarios que utilizaron el modelo de BERT perciben una **utilidad mayor** en el modelo al considerar las recomendaciones como más apropiadas.

Con respecto a lo valiosas que les parecieron las recomendaciones a los usuarios, los que utilizaron un modelo de Frecuencia poseen un promedio de concordancia menor que el grupo que usó un Modelo de Markov y el que usó uno de BERT (para más detalle ver la tabla 6.40 y la figura 6.18). Esta diferencia se determinó como significativa, con un efecto de magnitud grande para ambos pares de modelos (ver 6.41). Por lo tanto, se puede determinar que tanto los usuarios que utilizaron el modelo de BERT como los que usaron el de Markov, perciben un **valor mayor** en el modelo al considerar las recomendaciones como más valiosas.

Con respecto a lo útiles que les parecieron las recomendaciones a los usuarios, los que utilizaron un modelo de Frecuencia poseen un promedio de concordancia menor que el grupo que usó un Modelo de Markov y el que usó uno de BERT (para más detalle ver la tabla 6.42

y la figura 6.19). Esta diferencia se determinó como significativa, con un efecto de magnitud grande para el par de modelos Markov y Frecuencia y pequeña para el par de modelos BERT y Frecuencia (ver 6.43). Por lo tanto, se puede determinar que los usuarios que utilizaron el modelo de Markov perciben una utilidad mayor en el modelo al considerar las recomendaciones como más útiles. Percibiendo una **utilidad agregada muy grande** en el caso de el grupo que utilizó un modelo de Markov y **utilidad agregada pequeña** para el grupo que usó BERT.

Estos resultados permiten demostrar que el modelo de Markov produce recomendaciones percibidas como mucho más valiosas y mucho más útiles que las generadas por el modelo de Frecuencia. Por otro lado, demuestran también que el modelo de BERT produce recomendaciones percibidas como **mucho más valiosas** y **un poco más útiles** que las generadas por el modelo de Frecuencia. Mostrando así para ambos casos un aumento del valor y de la utilidad percibida por los usuarios al usar estos modelos de recomendación.

Para discernir qué modelo es mejor extrínsecamente, es necesario incluir en el análisis el hecho de que las frases que se pueden crear con este tablero digital de comunicación utilizan un vocabulario de 79 pictogramas, sin conectores entre cada palabra. Estas limitaciones obligan a que las relaciones semánticas presentes entre cada palabra se pierdan en contraste con la importancia sintáctica que tiene cada palabra en la estructura de la oración. Es decir que, para la aplicación diseñada, el modelo que mejor se adapte a la solución será uno que se desempeñe mejor realizando predicciones sobre oraciones que sintácticamente sean similares a las que un niño con PC pudiese escribir. **Considerando esto, el modelo que tiene un mejor desempeño extrínseco es el que utiliza un modelo de Markov para realizar las predicciones.**

Con respecto a las hipótesis de trabajo planteadas en la sección 6.1, se determina que el segundo punto era correcto, debido a que el modelo con mejor desempeño extrínseco es efectivamente el que utiliza cadenas de Markov. Entregandolé una mayor utilidad y valor a la aplicación al integrarle un modelo de predicción de Markov.

7.3. Diseño de la aplicación

En este apartado se presentan los resultados encontrados sobre el efecto que tuvo el diseño de la aplicación en la percepción de los usuarios sobre las recomendaciones.

Hubo opiniones contrastadas con respecto a la forma de representar las recomendaciones. Independiente del modelo utilizado, hubo usuarios que entregaron opiniones negativas y positivas sobre la forma en la que se entregan las predicciones. Algunos dijeron que el color azul lograba resaltarlas y permitía verlas de forma inmediata al usar la aplicación, mientras que hubo gente que no las identificó nunca y por ende no las utilizó (para más detalle ver sección 6.4.10). Si complementamos esto con los resultados cuantitativos presentados en 6.37 y 6.16, podemos ver que el promedio de los usuarios es inferior a 4 en todos los modelos, mostrando así que el nivel de concordancia con respecto a la facilidad de identificación de las recomendaciones es, en promedio, menor a “De acuerdo”. Esto nos permite decir que para un grupo no menor de usuarios fue un problema identificar las recomendaciones. En particular,

un 50 % de los usuarios que utilizaron el modelo de Frecuencia se encontró "Ni de acuerdo, ni en desacuerdo" con que las recomendaciones eran fáciles de identificar. Considerando que se determinó que no existe diferencia significativa entre estos los tres grupos experimentales, se puede decir que los problemas de identificación de las recomendaciones ocurren de forma transversal en la muestra, y en particular, independiente del modelo.

Con el fin de mitigar este problema es que se propone la creación de un tutorial que incluya un apartado que hable sobre las recomendaciones, el cual explique cómo usarlas y en particular, donde se encuentran. Esto aseguraría que todos los usuarios logren identificar las recomendaciones como tales y así aprovechar la utilidad que estas entregan.

Uno de los puntos más problemáticos durante la evaluación fue el diseño del teclado, específicamente el orden con el cual se presentan los pictogramas. Viendo el orden por defecto que posee la aplicación, se pueden apreciar inconsistencias entre cada categoría sintáctica, por ejemplo sujetos se tiene *"yo"* como primer pictograma y luego se sigue un orden alfabético, asimismo en los verbos se tiene *"quiero"* como primer pictograma y luego también se sigue un orden alfabético, pero por otro lado, en los sustantivos se sigue un orden alfabético a lo largo de toda la categoría (ver figura 2.2 para un ejemplo gráfico). Estas inconsistencias provocaron confusión en los usuarios y por ende se debe definir un orden estándar que se debe seguir en todas las categorías por igual.

Con respecto al cambio de posición dentro de las recomendaciones presentado en la figura 6.21, fue un problema que indujo gran cantidad de errores en los usuarios y fue también objeto de comentarios que mostraban confusión. Para mitigar este problema se determina que la aplicación debe seguir la siguiente regla: si es que la palabra esta dentro de las recomendaciones actuales y también se incluye en las próximas recomendaciones, se debe colocar en el mismo lugar que estaba anteriormente. De esta forma se pueden evitar los errores provocados por el cambio de lugar de los pictogramas.

7.4. Limitaciones del trabajo presentado

Es importante recordar que tanto los resultados como el análisis correspondiente fueron obtenidos mediante una experimentación sobre una muestra que no representa al usuario final. Los participantes cumplían con el perfil de cuidador de niños con PC (para más detalle ver 5.2.1). Si se quiere utilizar esta aplicación con niños que padezcan parálisis cerebral se debe realizar otra experimentación que valide los resultados obtenidos se mantienen al usar el tablero digital de comunicación con usuarios finales.

Sin embargo, con lo obtenido en la presente memoria, se puede determinar que el modelo que posee un mejor desempeño, considerando el problema abordado en este trabajo de título y todas sus limitaciones, es el modelo basado en cadenas de Markov.

Existen dos escenarios que vale la pena comentar. ¿Qué pasa si agregamos vocabulario al problema? y ¿Qué pasa si agregamos información al entrenamiento? ¿Se mantienen los resultados?

El primer escenario fue levantado como idea de mejora por un integrante del focus group de expertos de dominio (para más detalle ver sección 6.3), en donde se preguntaba que tan factible era añadir pictogramas personalizados a la aplicación, es decir que cada usuario pudiese agregar los iconos y palabras que este desease. Esto trae consigo dos problemas principales.

Primero, el sistema de recomendaciones de Markov, utiliza la frecuencia de combinaciones de palabras para realizar las recomendaciones y por ende, entrega una probabilidad nula a cualquier combinación que incluya alguna palabra no presente en el vocabulario. Es decir que agregar un pictograma a la aplicación significaría tener que adaptar el corpus de entrenamiento de donde el modelo recolecta evidencia para generar las predicciones, para que luego aprenda las relaciones sintácticas y semánticas que tiene esa palabra con el resto, lo cual no es un problema trivial.

El segundo problema con añadir pictogramas, es que deben ser validados semántica y semíoticamente para que el usuario pueda usarlos de forma correcta, esto es muy complejo. En este trabajo de título se agregó un pictograma al vocabulario, luego de una sugerencia realizada por un experto de dominio del área de trata de pacientes. Sin embargo, no fue bien recibido por los usuarios, debido a que muchos comentaron sobre la poca representabilidad del icono, teniendo problemas para identificar el pictograma. Los comentarios realizados al respecto del pictograma dolor muestran lo complicado semántica y semioticamente que es incluir pictogramas a la aplicación.

El segundo escenario responde a otra de las limitaciones de este trabajo de título, el tamaño del corpus de entrenamiento. Este influye de forma directa en los resultados de la evaluación, en particular en la parte intrínseca de la misma, donde se tuvo que invalidar el resultado obtenido en *perplexity* por la poca magnitud del corpus. Si aumentamos el tamaño del corpus de entrenamiento, es posible y probable que los resultados evaluación intrínseca cambien, en particular a medida que aumentemos la evidencia, el modelo basado en transformadores podrá usar más información para identificar relaciones semánticas entre los componentes de las oraciones y por ende es probable que supere al modelo de Markov en la precisión de las recomendaciones generadas. Por otro lado, debido a la forma en que cada modelo extrae información de las palabras, es poco probable que los resultados de la evaluación extrínseca varíen al agregar oraciones al corpus de entrenamiento. En particular, en frases que siguen una estructura sintáctica similar a las generadas por la doctora Hidalgo, el modelo de Markov seguirá teniendo un mejor desempeño. Por estas razones, es que se cree que el modelo de Markov seguirá siendo mejor que los anteriores aún en este nuevo escenario planteado.

El hecho de que la muestra se haya compuesto de adultos que cumplen con el perfil de cuidador introduce una serie de sesgos a los resultados y percepciones de la aplicación. Por ejemplo muchos comentaron que se necesitan subdivisiones semánticas para poder organizar mejor los pictogramas en el tablero, sin embargo, al realizar esta nueva categorización, es necesario tener en consideración que los niños que se consideran como usuario objetivo no tienen el mismo conocimiento sobre el significado de las palabras. Es decir, que si realizamos la subdivisión semántica que los usuarios mencionaron como posible mejora para que la aplicación sea más fácil de navegar, es posible que los niños no la entiendan y por ende, no sirva. Es por esta razón que se decide que la categorización sintáctica decidida en el trabajo de Bahamonde [5] funciona mejor para soluciones que ocupan tableros de comunicación.

Además concuerda con la importancia que tiene esta estructura para las oraciones creadas por la doctora Hidalgo, fisiatra experta en el dominio de trata de pacientes.

Otro punto relevante que se debe considerar al momento de adaptar esta solución para ser usada con niños con parálisis cerebral es el tamaño del tablet. Si bien Bahamonde [5] logró como resultado que no existía diferencia significativa en el uso de la aplicación en un tablet de 8" y uno de 12.4", este resultado también se encuentra limitado por la muestra utilizada para la experimentación, en donde los usuarios cumplen el mismo perfil explicitado en 5.2.1. Durante la evaluación hubo un número no menor de participantes que mencionó que para buscar y seleccionar los pictogramas utilizaba solamente la palabra asociada y no el icono. Las razones que daban para esto variaban entre la mala representación percibida de algunos dibujos y que el tamaño de los mismos era muy pequeño para la dimensión de la tablet utilizada en la evaluación (para más detalles revisar sección 5.2.2). Es por estas razones que al usar esta aplicación con niños, se debe validar de nuevo el tamaño. Es posible que al usar la misma tablet utilizada en este trabajo de título, los usuarios no podrán identificar los iconos de forma correcta, y al no saber leer, no podrán usar bien la aplicación, entorpeciendo con el objetivo final de la misma, el cual es facilitar la comunicación.

Capítulo 8

Conclusión y Trabajo Futuro

En este trabajo de título se buscaba integrar un sistema de predicción de la palabra siguiente a un tablero digital de comunicación basado en pictogramas. Esto con el fin de aumentar la utilidad y el valor que percibe el usuario al utilizar la aplicación.

Se desarrollaron tres modelos de predicción distinto, los cuales al ser integrados en la aplicación permitieron aumentar la eficiencia, permitiendo a los usuarios comunicarse en menos tiempo, en promedio, usando cualquiera de los tres modelos diseñados.

Al realizar una comparativa entre los modelos de predicción se llegó a la conclusión que tanto el modelo de Markov como el de BERT permiten, de forma transversal, un uso más eficiente de la aplicación. En particular al analizar el tipo de oraciones para las cuales cada uno de estos modelos tenía el mejor desempeño, se obtuvo que Markov entregaba una mayor utilidad y valor al usuario si es que se quiere realizar recomendaciones sobre oraciones que se benefician de una estructura sintáctica (tarea 4 y 5) , mientras que BERT aprovechaba mejor las relaciones semánticas entre las palabras de una oración para realizar las predicciones (tarea 6).

En el caso particular de esta aplicación, la relación sintáctica entre las palabras de una oración es más relevante para los usuarios que la información semántica que se tenga. Es por esto, que se determinó que el modelo que entrega una utilidad y valor agregado mayor a la aplicación es el basado en las cadenas de Markov.

Como trabajo futuro se plantean las siguientes implementaciones, las cuales surgen de los comentarios entregados en la evaluación heurística con expertos del dominio del trata de pacientes (para más detalle ver sección 6.3).

Implementar un sistema que tome registro de las frases que ha escrito el usuario con la aplicación y entregue predicciones ponderando las preferencias personales del sujeto con el corpus general de oraciones. De esta forma se pueden generar recomendaciones que reconozcan la individualidad del sujeto. Es interesante y se plantea como trabajo futuro evaluar como esta personalización influye en la precisión, utilidad y valor percibido de las recomendaciones generadas.

Evaluar el desempeño de los modelos de Markov y BERT al incluir datos adicionales como evidencia, como vendrían siendo la localización del individuo al escribir la frase (im-

plementado por [23]); haciendo que palabras relacionadas con la casa se predigan con mayor probabilidad al estar el usuario en esta localización física, el timestamp de el momento en que se registró la frase; permitiendo que las palabras predichas dependan y varíen dependiendo de la hora del día; y predicciones pasadas elegidas. Es interesante evaluar como cada una de estas variables influye en la precisión, utilidad y valor percibido de las recomendaciones generadas.

Además se plantea utilizar los comentarios entregados por los usuarios durante la evaluación extrínseca (para más detalles ver subsección 6.4.10) para diseñar mejoras a la aplicación. Los comentarios plasman problemas sufridos por los usuarios que participaron en este primer experimento, se especula que modificar la aplicación teniendo como guía estas respuesta resultará en un aumento de la utilidad del sistema de recomendación y también de la la usabilidad de la aplicación en general.

Bibliografía

[1] Vela, C. C. V., & Ruiz, C. A. V. (2014). **Parálisis cerebral infantil: definición y clasificación a través de la historia.** Revista mexicana de Ortopedia pediátrica, 16(1), 6-10.

[2] Goldstein, H., & Cameron, H. (1952). **New method of communication for the aphasic patient.** Arizona Medicine, 8, 17–21.

[3] Goldberg, H. R., & Fenton, J. (1960). **Aphonic communication for those with cerebral palsy: Guide for the development and use of a communication board.** New York: United Cerebral Palsy of New York State.

[4] Glennen, S. L., & DeCoste, D. C. (1997). **Augmentative and alternative communication systems.** The Handbook of Augmentative and Alternative Communication, 59-69.

[5] Bahamonde Santander, M. (2023). **Tablero digital de comunicación para niños con parálisis cerebral.** Disponible en https://repositorio.uchile.cl/handle/2250/193929

[6] Newell, A., Langer, S., & Hickey, M. (1998). **The rôle of natural language processing in alternative and augmentative communication.** Natural Language Engineering, 4(1), 1-16.

[7] Martin, J. H. (2009). **Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.** Pearson/Prentice Hall.

[8] Chauncey Wilson Dennis Wixon. **Handbook of Human-Computer Interaction**. Chapter 27 - The Usability Engineering Framework for Product Design and Evaluation, pages 653–688. North Holland, 2 edition, 1997.

[9] Dudy, S., & Bedrick, S. (2018, July). **Compositional language modeling for icon-based augmentative and alternative communication.** In Proceedings of the conference. Association for Computational Linguistics. Meeting (Vol. 2018, p. 25). NIH Public Access.

[10] Pereira, J. A., Macêdo, D., Zanchettin, C., de Oliveira, A. L. I., & do Nascimento Fidalgo, R. (2022).**Pictobert: Transformers for next pictogram prediction.** Expert Systems with Applications, 202, 117231.

[11] Sennott, S. C., Akagi, L., Lee, M., & Rhodes, A. (2019). **AAC and artificial intelligence (AI).** Topics in language disorders, 39(4), 389.

[12] Schadle, I. (2004). **Sibyl: AAC system using NLP techniques.** In Computers Helping People with Special Needs: 9th International Conference, ICCHP 2004, Paris, France, July 7-9, 2004. Proceedings 9 (pp. 1009-1015). Springer Berlin Heidelberg.

[13] Copestake, A. (1997). **Augmented and alternative NLP techniques for augmen-**

tative and alternative communication. In Natural Language Processing for Communication Aids.

[14] Trnka, K., Yarrington, D., McCaw, J., McCoy, K. F., & Pennington, C. (2007, April). **The effects of word prediction on communication rate for AAC.** In Human language technologies 2007: The conference of the north american chapter of the association for computational linguistics; companion volume, short papers (pp. 173-176).

[15] McCoy, K. F., Pennington, C. A., & Badman, A. L. (1998). **Compansion: From research prototype to practical integration.** Natural Language Engineering, 4(1), 73-95.

[16] Shakhovska, K., Dumyn, I., Kryvinska, N., & Kagita, M. K. (2021). **An Approach for a Next-Word Prediction for Ukrainian Language.** Wireless Communications and Mobile Computing, 2021, 1-9.

[17] Hamarashid, H. K., Saeed, S. A., & Rashid, T. A. (2021). **Next word prediction based on the N-gram model for Kurdish Sorani and Kurmanji.** Neural Computing and Applications, 33, 4547-4566.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). **Attention is all you need.** Advances in neural information processing systems, 30.

[19] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). **Bert: Pre-training of deep bidirectional transformers for language understanding.** arXiv preprint arXiv:1810.04805.

[20] Pereira, J., Franco, N., & Fidalgo, R. (2020, September). **A semantic grammar for augmentative and alternative communication systems.** In International Conference on Text, Speech, and Dialogue (pp. 257-264). Cham: Springer International Publishing.

[21] Martínez-Santiago, F., García-Cumbreras, M. Á., Montejo-Ráez, A., & Díaz-Galiano, M. C. (2016, June). **Pictogrammar: an AAC device based on a semantic grammar.** In Proceedings of the 11th workshop on innovative use of NLP for building educational applications (pp. 142-150).

[22] Morgan & Claypool Publishers. Hervás, R., Bautista, S., Méndez, G., Galván, P., & Gervás, P. (2020). **Predictive composition of pictogram messages for users with autism.** Journal of Ambient Intelligence and Humanized Computing, 11(11), 5649–5664.

[23] Garcia, L. F., de Oliveira, L. C., & de Matos, D. M. (2016). **Evaluating pictogram prediction in a location-aware augmentative and alternative communication system.** Assistive Technology, 28(2), 83–92.

[24] Miller, G. A. (1995). **WordNet: a lexical database for English.** Communications of the ACM, 38(11), 39–41. http://dx.doi.org/10.1145/219717.219748.

[25] Russell, S. J., & Norvig, P. (2010). **Artificial intelligence a modern approach.** London.

[26] Cheng, J., Dong, L., & Lapata, M. (2016). **Long short-term memory-networks for machine reading.** arXiv preprint arXiv:1601.06733.

[27] Lisa Torrey and Jude Shavlik. **Transfer learning.** In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pages 242–264. IGI global, 2010.

[28] Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2023). **Spanish pre-trained bert model and evaluation data.** arXiv preprint arXiv:2308.02976.

[29] Holzinger, A. (2005). **Usability engineering methods for software developers.** Communications of the ACM, 48(1), 71-74.

[30] Creswell, J. W., & Creswell, J. D. (2017). **Research design: Qualitative, quantitative, and mixed methods approaches.** Sage publications.

Anexo A

Tablas

Tabla A.1: Pictogramas según categoría sintáctica

Sujetos	**Verbos**	**Sustantivos**	
yo	quiero	agua	guata
abuelita	bailar	auto de juguete	helado
abuelito	beber	boca	huevo
guagua	caer	brazo	jugo
mamá	caminar	bufanda	lámpara
niña	comer	calcetines	leche
niño	comprar	calzoncillo	mano
papá	correr	cama	mesa
tía	dormir	camisa	muñeca
tío	hablar	cara	nariz
	jugar	chaleco	ojo
	llorar	chocolate	oreja
	subir	cojín	osito de peluche
		cuchara	pan
		cuchillo	pantalón
		dedo	pantalones cortos
		dulce	pelo
		galletas	pelota
		globo	pie
		gorro	pierna
		guante	plato
		tenedor	polera
		torta	robot
		tren	silla
		vaso	sillón
		vestido	sopa
		zapato	tambor
		dolor	taza

Tabla A.2: Caracterización de los usuarios de la evaluación

Caracterización Usuario				
ID	**Genero**	**Perfil**	**Edad**	**Modelo**
1	M	Cuidador/a	56	Bert
2	F	Cuidador/a	26	Bert
3	F	Cuidador/a	55	Bert
4	M	Cuidador/a	38	Bert
5	F	Cuidador/a	46	Bert
6	M	Cuidador/a	27	Bert
7	F	Cuidador/a	26	Bert
8	M	Cuidador/a	28	Bert
9	F	Cuidador/a	36	Bert
10	M	Cuidador/a	46	Bert
11	M	Cuidador/a	43	Bert
12	F	Cuidador/a	43	Bert
13	F	Personal salud	23	Bert
14	M	Cuidador/a	53	Bert
15	M	Cuidador/a	39	Bert
16	F	Cuidador/a	39	Bert
17	M	Cuidador/a	47	Bert
18	M	Cuidador/a	43	Bert
19	F	Cuidador/a	44	Bert
20	F	Cuidador/a	58	Frecuencia
21	M	Cuidador/a	49	Frecuencia
22	F	Cuidador/a	42	Frecuencia
23	F	Cuidador/a	43	Frecuencia
24	M	Cuidador/a	67	Frecuencia
25	F	Cuidador/a	40	Frecuencia
26	M	Cuidador/a	40	Frecuencia
27	F	Cuidador/a	43	Frecuencia
28	M	Cuidador/a	42	Frecuencia
29	M	Cuidador/a	44	Frecuencia
30	F	Cuidador/a	43	Frecuencia
31	M	Cuidador/a	50	Frecuencia
32	F	Cuidador/a	37	Frecuencia
33	M	Cuidador/a	45	Frecuencia
34	M	Cuidador/a	40	Frecuencia
35	F	Cuidador/a	39	Frecuencia
36	M	Cuidador/a	51	Frecuencia

37	M	Cuidador/a	54	Frecuencia
38	F	Cuidador/a	50	Markov
39	M	Cuidador/a	37	Markov
40	M	Cuidador/a	54	Markov
41	M	Cuidador/a	41	Markov
42	F	Cuidador/a	34	Markov
43	F	Cuidador/a	55	Markov
44	F	Personal salud	26	Markov
45	M	Cuidador/a	55	Markov
46	F	Cuidador/a	51	Markov
47	F	Cuidador/a	26	Markov
48	M	Cuidador/a	29	Markov
49	F	Cuidador/a	49	Markov
50	F	Cuidador/a	44	Markov
51	F	Cuidador/a	46	Markov
52	F	Cuidador/a	40	Markov
53	M	Cuidador/a	58	Markov
54	M	Cuidador/a	43	Markov
55	F	Cuidador/a	48	Markov
56	F	Cuidador/a	46	Markov

Tabla A.3: Evaluación de la tarea 1

Tarea 1						
User ID	**Tiempo (s)**	**Errores**	**Errores/ Tiempo**	**Ayudas**	**# Recom.**	**Completa 1:Si 0:No**
1	18	0	0	0	1	1
2	14	0	0	0	2	1
3	13	0	0	0	1	1
4	7	0	0	0	1	1
5	23	0	0	0	2	1
6	17	0	0	0	0	1
7	8	0	0	0	1	1
8	17	1	0,0588	0	1	1
9	11	0	0	0	1	1
10	9	0	0	0	2	1
11	9	0	0	0	2	1
12	16	0	0	0	1	1
13	7	0	0	0	1	1
14	16	1	0,0625	0	0	1
15	12	0	0	0	1	1
16	13	0	0	0	2	1
17	17	0	0	0	0	1
18	10	0	0	0	1	1
19	13	0	0	0	1	1
20	27	0	0	0	1	1
21	13	0	0	0	1	1
22	14	0	0	0	2	1
23	22	0	0	0	2	1
24	18	0	0	0	1	1
25	13	0	0	0	0	1
26	13	0	0	0	0	1
27	14	0	0	0	0	1
28	8	0	0	0	0	1
29	12	0	0	0	2	1
30	11	0	0	0	0	1
31	8	0	0	0	2	1
32	22	1	0,0454	0	1	1
33	13	1	0,0769	0	1	1
34	10	0	0	0	2	1
35	9	0	0	0	1	1

36	16	0	0	0	1	1
37	11	0	0	0	0	1
38	11	0	0	0	1	1
39	24	0	0	0	0	1
40	15	1	0,0666	0	2	1
41	9	0	0	0	2	1
42	15	0	0	0	2	1
43	9	0	0	0	2	1
44	8	0	0	0	0	1
45	13	0	0	0	0	1
46	17	0	0	0	1	1
47	12	0	0	0	2	1
48	50	2	0,04	0	1	1
49	16	0	0	0	2	1
50	30	1	0,0333	0	2	1
51	12	0	0	0	0	1
52	7	0	0	0	0	1
53	7	0	0	0	1	1
54	16	0	0	0	2	1
55	12	0	0	0	2	1
56	4	0	0	0	1	1

Tabla A.4: Evaluación de la tarea 2

Tarea 2						
User ID	**Tiempo (s)**	**Errores**	**Errores/ Tiempo**	**Ayudas**	**# Recom.**	**Completa 1:Si 0:No**
1	7	0	0	0	2	1
2	6	0	0	0	2	1
3	18	1	0,0556	0	3	1
4	7	0	0	0	1	1
5	20	0	0	0	3	1
6	7	0	0	0	1	1
7	25	0	0	0	0	1
8	7	0	0	0	3	1
9	8	0	0	0	2	1
10	5	0	0	0	3	1
11	14	0	0	0	2	1
12	38	0	0	1	2	1
13	8	0	0	0	2	1
14	12	0	0	0	2	1
15	9	0	0	0	2	1
16	6	0	0	0	3	1
17	8	0	0	0	3	1
18	11	0	0	0	1	1
19	4	0	0	0	3	1
20	31	0	0	1	1	1
21	32	0	0	1	1	1
22	38	0	0	0	2	1
23	32	0	0	1	2	1
24	39	0	0	1	1	1
25	26	0	0	1	2	1
26	40	0	0	1	0	1
27	30	0	0	1	0	1
28	32	0	0	1	0	1
29	33	0	0	1	2	1
30	29	0	0	1	0	1
31	30	0	0	1	1	1
32	22	0	0	1	2	1
33	27	0	0	1	1	1
34	24	0	0	1	1	1
35	24	0	0	1	1	1

36	32	0	0	1	1	1
37	33	0	0	1	0	1
38	22	0	0	0	2	1
39	12	0	0	0	1	1
40	10	0	0	0	3	1
41	11	0	0	0	1	1
42	5	0	0	0	3	1
43	7	0	0	0	3	1
44	9	0	0	0	1	1
45	41	0	0	1	0	1
46	33	0	0	1	2	1
47	31	1	0,0323	0	2	1
48	5	0	0	0	3	1
49	8	0	0	0	3	1
50	13	0	0	1	3	1
51	13	0	0	0	1	1
52	8	0	0	0	1	1
53	40	2	0,05	0	3	1
54	6	0	0	0	3	1
55	6	0	0	0	3	1
56	20	1	0,05	0	3	1

Tabla A.5: Evaluacuón de la tarea 3

Tarea 3						
User ID	**Tiempo (s)**	**Errores**	**Errores/ Tiempo**	**Ayudas**	**# Recom.**	**Completa 1:Si 0:No**
1	40	0	0	1	0	1
2	31	0	0	1	1	1
3	27	0	0	1	1	1
4	32	0	0	1	1	1
5	65	1	0,0153	1	0	1
6	25	1	0,04	1	1	1
7	24	0	0	0	0	1
8	32	0	0	1	1	1
9	33	0	0	1	1	1
10	24	0	0	1	1	1
11	24	0	0	0	1	1
12	36	0	0	0	0	1
13	28	0	0	1	1	1
14	33	0	0	1	1	1
15	33	0	0	1	1	1
16	33	0	0	1	1	1
17	25	0	0	0	1	1
18	25	0	0	0	1	1
19	30	0	0	1	1	1
20	20	0	0	0	1	1
21	11	0	0	0	2	1
22	18	0	0	0	3	1
23	23	0	0	0	2	1
24	30	0	0	0	1	1
25	15	0	0	0	3	1
26	29	0	0	0	0	1
27	28	0	0	0	2	1
28	24	0	0	0	1	1
29	18	0	0	0	3	1
30	22	0	0	0	1	1
31	25	0	0	0	2	1
32	25	0	0	0	3	1
33	21	0	0	0	1	1
34	20	0	0	0	2	1
35	20	0	0	0	2	1

36	29	0	0	0	2	1
37	21	0	0	0	0	1
38	28	0	0	1	1	1
39	40	0	0	1	1	1
40	30	0	0	1	2	1
41	27	0	0	1	1	1
42	37	0	0	1	1	1
43	87	2	0,0230	1	1	1
44	29	0	0	1	1	1
45	31	0	0	0	0	1
46	25	0	0	0	2	1
47	15	0	0	1	1	1
48	27	0	0	1	1	1
49	27	0	0	1	1	1
50	39	0	0	1	1	1
51	30	0	0	0	0	1
52	27	0	0	1	0	1
53	31	0	0	0	1	1
54	37	0	0	1	0	1
55	55	1	0,0182	1	1	1
56	32	0	0	1	0	1

Tabla A.6: Evaluación de la tarea 4

Tarea 4						
User ID	**Tiempo (s)**	**Errores**	**Errores/ Tiempo**	**Ayudas**	**# Recom.**	**Completa 1:Si 0:No**
1	10	0	0	0	2	1
2	8	0	0	0	2	1
3	9	0	0	0	2	1
4	7	0	0	0	1	1
5	38	0	0	0	2	1
6	8	0	0	0	2	1
7	11	0	0	0	0	1
8	13	0	0	0	2	1
9	8	0	0	0	2	1
10	6	0	0	0	2	1
11	7	0	0	0	2	1
12	13	0	0	0	2	1
13	10	0	0	0	1	1
14	9	0	0	0	2	1
15	19	0	0	0	2	1
16	7	0	0	0	2	1
17	5	0	0	0	2	1
18	10	0	0	0	2	1
19	7	0	0	0	2	1
20	11	0	0	0	1	1
21	10	0	0	0	2	1
22	8	0	0	0	2	1
23	19	0	0	0	1	1
24	18	0	0	0	1	1
25	7	0	0	0	2	1
26	13	0	0	0	0	1
27	8	0	0	0	2	1
28	8	0	0	0	1	1
29	7	0	0	0	2	1
30	16	0	0	0	1	1
31	7	0	0	0	2	1
32	5	0	0	0	2	1
33	14	0	0	0	1	1
34	7	0	0	0	2	1
35	6	0	0	0	2	1

36	17	0	0	0	2	1
37	14	0	0	0	2	1
38	15	0	0	0	1	1
39	12	0	0	0	1	1
40	18	0	0	0	1	1
41	8	0	0	0	1	1
42	9	0	0	0	1	1
43	35	1	0,0286	0	0	1
44	7	0	0	0	1	1
45	16	0	0	0	0	1
46	23	0	0	0	0	1
47	8	0	0	0	1	1
48	12	0	0	0	1	1
49	10	0	0	0	1	1
50	25	0	0	0	1	1
51	15	0	0	0	1	1
52	6	0	0	0	1	1
53	16	0	0	0	1	1
54	24	0	0	0	1	1
55	19	0	0	0	1	1
56	13	0	0	0	1	1

Tabla A.7: Evaluación de la tarea 5

Tarea 5						
User ID	**Tiempo (s)**	**Errores**	**Errores/ Tiempo**	**Ayudas**	**# Recom.**	**Completa 1:Si 0:No**
1	65	0	0	0	0	1
2	40	0	0	0	1	1
3	50	0	0	0	1	1
4	45	0	0	0	0	1
5	105	1	0,0095	0	1	1
6	36	0	0	0	1	1
7	27	0	0	0	0	1
8	40	0	0	0	1	1
9	35	0	0	0	1	1
10	34	0	0	0	1	1
11	40	0	0	0	1	1
12	62	0	0	0	0	1
13	33	0	0	0	1	1
14	43	0	0	0	0	1
15	49	0	0	0	1	1
16	36	0	0	0	1	1
17	34	0	0	0	1	1
18	47	0	0	0	1	1
19	50	0	0	0	1	1
20	60	0	0	0	1	1
21	29	0	0	0	2	1
22	19	0	0	0	2	1
23	36	0	0	0	1	1
24	54	0	0	0	0	1
25	43	0	0	0	1	1
26	34	0	0	0	0	1
27	37	0	0	0	1	1
28	24	0	0	0	1	1
29	33	0	0	0	2	1
30	22	0	0	0	1	1
31	20	0	0	0	2	1
32	34	0	0	0	0	1
33	37	0	0	0	1	1
34	24	0	0	0	2	1
35	25	0	0	0	2	1

36	47	0	0	0	0	1
37	42	0	0	0	0	1
38	26	1	0,0384	0	1	0
39	33	0	0	0	2	1
40	18	0	0	0	3	1
41	15	0	0	0	2	1
42	24	0	0	0	3	1
43	27	1	0,0370	0	1	0
44	8	0	0	0	2	1
45	71	1	0,0141	0	0	1
46	20	0	0	0	1	0
47	32	0	0	0	1	1
48	8	0	0	0	3	1
49	11	0	0	0	3	1
50	26	0	0	0	3	1
51	21	0	0	0	2	1
52	23	0	0	0	2	1
53	10	0	0	0	3	1
54	13	0	0	0	3	1
55	17	0	0	0	3	1
56	15	0	0	0	3	1

Tabla A.8: Evaluación de la tarea 6

Tarea 6						
User ID	**Tiempo (s)**	**Errores**	**Errores/ Tiempo**	**Ayudas**	**# Recom.**	**Completa 1:Si 0:No**
1	10	0	0	0	3	1
2	10	0	0	0	4	1
3	9	0	0	0	4	1
4	18	0	0	0	4	1
5	14	0	0	0	4	1
6	10	0	0	0	3	1
7	9	0	0	0	1	1
8	7	0	0	0	2	1
9	5	0	0	0	4	1
10	8	0	0	0	3	1
11	11	0	0	0	4	1
12	11	0	0	0	2	1
13	6	0	0	0	4	1
14	12	0	0	0	2	1
15	10	0	0	0	4	1
16	8	0	0	0	3	1
17	10	0	0	0	2	1
18	7	0	0	0	3	1
19	8	0	0	0	3	1
20	24	0	0	0	1	1
21	28	0	0	0	1	1
22	26	0	0	0	1	1
23	22	0	0	0	0	1
24	41	0	0	0	1	1
25	21	0	0	0	1	1
26	23	0	0	0	0	1
27	28	0	0	0	1	1
28	20	0	0	0	0	1
29	30	0	0	0	1	1
30	28	0	0	0	0	1
31	28	0	0	0	1	1
32	20	0	0	0	1	1
33	25	0	0	0	1	1
34	24	0	0	0	1	1
35	27	0	0	0	1	1

36	35	0	0	0	1	1
37	31	0	0	0	0	1
38	5	0	0	0	2	0
39	14	0	0	0	2	1
40	23	1	0,0434	0	2	1
41	11	0	0	0	2	1
42	13	0	0	0	2	1
43	12	0	0	0	1	0
44	10	0	0	0	2	1
45	32	0	0	0	0	1
46	15	0	0	0	2	0
47	10	0	0	0	2	1
48	12	0	0	0	2	1
49	9	0	0	0	2	1
50	12	0	0	0	3	1
51	48	0	0	0	1	1
52	11	0	0	0	1	1
53	11	0	0	0	2	1
54	10	0	0	0	2	1
55	16	0	0	0	2	1
56	22	1	0,0454	0	2	1

Tabla A.9: Resultados NASA-TLX

NASA-TLX						
User ID	**Mental**	**Física**	**Temp.**	**Rendim.**	**Esfuerzo**	**Nvl Esfu.**
1	15	10	5	20	16	5
2	2	1	3	17	4	5
3	11	1	8	17	8	1
4	5	1	5	17	4	3
5	1	1	1	19	1	1
6	3	1	9	10	15	1
7	15	4	5	18	6	1
8	3	2	7	6	6	5
9	1	1	5	9	4	2
10	15	10	15	15	14	4
11	12	1	12	16	16	3
12	10	5	12	16	15	13
13	10	1	10	20	6	2
14	15	13	11	17	15	8
15	13	3	10	10	15	13
16	4	2	12	16	3	3
17	8	1	4	17	4	1
18	1	2	7	19	4	1
19	7	7	9	19	5	4
20	5	2	3	17	4	4
21	5	1	7	19	7	1
22	1	1	10	20	1	1
23	10	1	1	19	10	1
24	5	1	5	20	5	2
25	8	1	9	20	10	1
26	10	6	11	5	9	6
27	12	5	10	9	8	3
28	11	3	7	8	10	2
29	7	2	5	19	7	3
30	2	1	8	14	6	1
31	3	1	1	17	1	3
32	7	1	9	6	20	5
33	4	1	6	16	4	2
34	2	1	5	1	1	1
35	1	1	1	20	1	1
36	3	1	5	16	5	3

37	20	20	1	20	10	16
38	10	1	10	18	2	1
39	1	1	6	20	10	1
40	3	2	2	18	19	2
41	9	2	4	13	4	1
42	14	1	11	17	10	11
43	8	6	8	13	7	11
44	9	3	15	20	18	3
45	5	9	9	18	5	3
46	1	1	4	16	1	3
47	5	3	3	19	4	4
48	5	1	6	20	2	1
49	3	1	18	18	3	1
50	10	4	5	20	10	1
51	2	2	2	20	2	1
52	5	1	8	20	13	1
53	6	1	6	20	10	1
54	7	1	5	20	4	4
55	1	1	1	20	1	1
56	15	5	16	16	16	3

Tabla A.10: Resultados cuestionario post evaluación

Cuestionario Post-Evaluación					
User ID	**P1**	**P2**	**P3**	**P4**	**P5**
1	2	4	3	3	4
2	5	5	4	5	5
3	4	5	5	4	5
4	4	5	5	4	5
5	2	5	5	4	5
6	5	5	5	5	5
7	1	4	4	4	5
8	4	4	4	4	4
9	5	5	5	5	5
10	4	4	4	4	4
11	5	5	4	5	5
12	2	5	5	5	4
13	4	5	5	5	5
14	4	4	4	4	5
15	4	4	4	4	3
16	3	4	3	4	4
17	4	4	4	4	4
18	4	4	4	4	5
19	4	4	4	4	3
20	1	5	3	3	5
21	5	4	4	5	3
22	5	4	5	5	3
23	1	1	4	4	4
24	5	4	4	4	5
25	5	5	5	5	5
26	3	3	3	2	4
27	2	3	2	3	4
28	2	3	3	4	4
29	4	4	4	4	5
30	2	2	2	2	4
31	4	4	4	4	5
32	3	4	3	3	2
33	2	2	2	2	4
34	4	4	4	4	4
35	2	3	3	3	5
36	4	4	4	4	4

37	2	5	5	5	5
38	5	5	5	5	4
39	3	4	4	4	5
40	4	4	4	3	4
41	5	4	4	4	5
42	3	4	4	4	4
43	1	5	5	5	5
44	4	3	5	5	2
45	2	3	4	5	4
46	5	4	4	4	4
47	3	4	4	4	4
48	3	4	5	5	3
49	4	4	4	5	5
50	5	5	5	5	5
51	5	5	5	5	5
52	1	5	5	5	4
53	5	5	5	5	5
54	4	4	4	4	5
55	5	5	5	5	5
56	3	4	4	4	4

Anexo B

Formularios

Cuestionario de Experiencia de Uso

Nombre: _____ Edad: _____

Le pedimos que por favor conteste a las siguientes preguntas respecto a su experiencia interactuando con el tablero digital que le fue facilitado. No hay respuestas correctas o incorrectas. Recuerde que no lo estamos evaluando a usted, sino que es usted quien está evaluando la aplicación.

1. Respecto a la tarea de componer una oración con la ayuda de los pictogramas:

Exigencia mental: ¿Qué tan demandante mentalmente fue escribir las frases?

Baja Alta

Exigencia física: ¿Qué tan demandante físicamente fue escribir las frases?

Baja Alta

Exigencia temporal: ¿Qué tan fuerte o rápido es el ritmo impuesto para escribir frases?

Baja Alta

Rendimiento: ¿Qué tan exitoso ha sido en lograr escribir las frases?

Malo Bueno

Esfuerzo: ¿Qué tan duro estima que tuvo que trabajar para lograr un buen nivel de rendimiento?

Bajo Alto

Nivel de esfuerzo: ¿Qué tan inseguro, irritado y molesto se sintió al escribir las frases?

Bajo Alto

Figura B.1: Formulario NASA-TLX post evaluatorio

2. Indique qué tan de acuerdo está con las siguientes afirmaciones:

Me fue fácil identificar las recomendaciones

O Muy en desacuerdo	O En desacuerdo	O Ni de acuerdo ni en desacuerdo	O De acuerdo	O Muy de acuerdo

Me parecieron apropiadas las recomendaciones

O Muy en desacuerdo	O En desacuerdo	O Ni de acuerdo ni en desacuerdo	O De acuerdo	O Muy de acuerdo

Me parecieron valiosas las recomendaciones

O Muy en desacuerdo	O En desacuerdo	O Ni de acuerdo ni en desacuerdo	O De acuerdo	O Muy de acuerdo

Me parecieron útiles las recomendaciones

O Muy en desacuerdo	O En desacuerdo	O Ni de acuerdo ni en desacuerdo	O De acuerdo	O Muy de acuerdo

Es probable que vuelva a utilizar la aplicación

O Muy en desacuerdo	O En desacuerdo	O Ni de acuerdo ni en desacuerdo	O De acuerdo	O Muy de acuerdo

¿Tiene comentarios o sugerencias adicionales?

_____

_____

_____

_____

_____

_____

_____

Figura B.2: Formulario cualitativo post evaluatorio

Recolección de datos

Nombre: _____ Edad: _____

Género: _____ Modelo utilizado: _____ Perfil: _____

Tarea1: Escribir “yo quiero agua”

Métrica	**Valor**
Tiempo en completar	
Número y tipo de errores	
Número de errores por unidad de tiempo	
Número de ayudas necesarias	
Tarea completada	
Uso de recomendaciones	

Errores cometidos: _____

Tarea2: Escribir “yo quiero jugar”

Métrica	**Valor**
Tiempo en completar	
Número y tipo de errores	
Número de errores por unidad de tiempo	
Número de ayudas necesarias	
Tarea completada	
Uso de recomendaciones	

Errores cometidos: _____

Tarea3: Escribir “mamá quiero galletas jugo”

Métrica	**Valor**
Tiempo en completar	
Número y tipo de errores	
Número de errores por unidad de tiempo	
Número de ayudas necesarias	
Tarea completada	
Uso de recomendaciones	

Errores cometidos: _____

Figura B.3: Formulario para recolectar datos sobre las primeras 3

Tarea4: Escribir “papá comprar leche”

Métrica	**Valor**
Tiempo en completar	
Número y tipo de errores	
Número de errores por unidad de tiempo	
Número de ayudas necesarias	
Tarea completada	
Uso de recomendaciones	

Errores cometidos: _____

Tarea5: Escribir “yo llorar dolor pie”

Métrica	**Valor**
Tiempo en completar	
Número y tipo de errores	
Número de errores por unidad de tiempo	
Número de ayudas necesarias	
Tarea completada	
Uso de recomendaciones	

Errores cometidos: _____

Tarea6: Escribir “yo dormir cama papá”

Métrica	**Valor**
Tiempo en completar	
Número y tipo de errores	
Número de errores por unidad de tiempo	
Número de ayudas necesarias	
Tarea completada	
Uso de recomendaciones	

Errores cometidos: _____

Figura B.4: Formulario para recolectar datos sobre las primeras 3

Anexo C

Información Experimental Adicional

C.1. Caracterización de la muestra

En esta sección se presentan los resultados obtenidos con respecto a la información de los usuarios evaluados, tales como edad y género.

Tamaño de la muestra

El proceso de construcción de la muestra se realizó utilziando la técnica de muestreo no probabilístico, "muestreo deliberado", explicado en la sección 5.2.1. Usando este método de captación se obtuvo una muestra de tamaño 60, de las cuales 4 debieron ser descartadas por errores durante la toma de resultados. Por lo que el tamaño final fue de 56.

Proporción de cada perfil de usuario

De los 56 usuarios, 2 corresponden a personal considerado como experto de dominio (para más detalles ver sección 5.2.1) y el resto 54 corresponden al perfil de cuidador de niños con PC.

Género de la muestra reclutada

Dentro de la muestra 29 personas respondieron que se identifican con el género femenino, esto corresponde a un 51.8 %, mientras que 27 personas se identifican con el género masculino, correspondiente con un 48.2 % de la muestra.

Información etaria de la muestra

Con respecto a la muestra general, se tiene que el promedio de la edades es 48.25, con una mediana de 47, un mínimo de 23 y un máximo de 81. Estas variables se distribuyen de forma normal con una desviación estándar de 12.32.

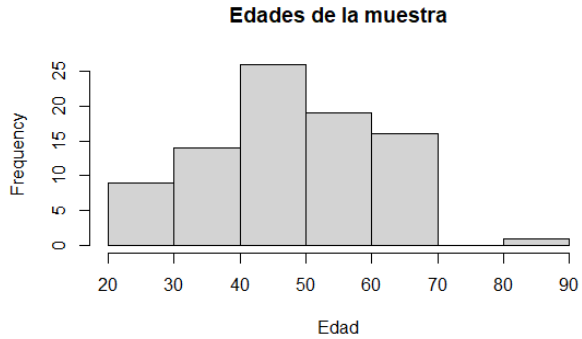


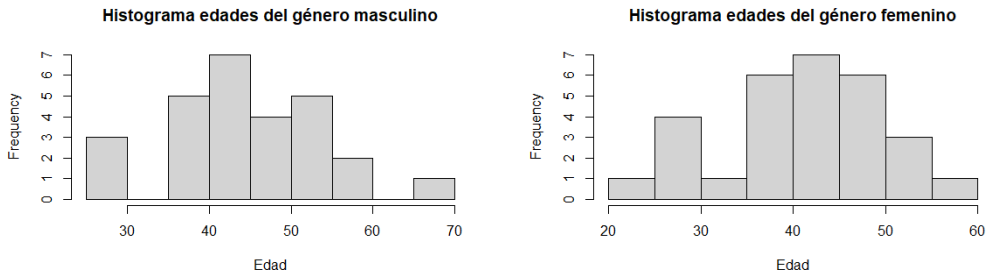
Figura C.1: Histograma de las edades de toda la muestra

Al agrupar la muestra por género se consiguieron los siguientes valores para promedio, desviación estándar, mediana, mínimo y máximo.

Tabla C.1: Estadísticas para las edades en cada género

Género	Promedio	Desviación Estándar	Mediana	Mínimo	Máximo
Femenino	41.3	9.27	43	23	58
M	45.1	9.38	44	27	67

Los datos etarios de la muestra agrupados según su género se distribuyen de la siguiente forma:



(a) Edades género masculino (b) Edades género femenino

Figura C.2: Histogramas de edades en cada género

Asignación de entornos de prueba

Con respecto a los modelos ocupados por cada usuario, se tiene que 18 usuarios usaron la aplicación con el modelo de frecuencias integrado, 19 usaron el modelo de Markov y 19 usaron el modelo de BERT.

La distribución de género según modelo se presenta a continuación:

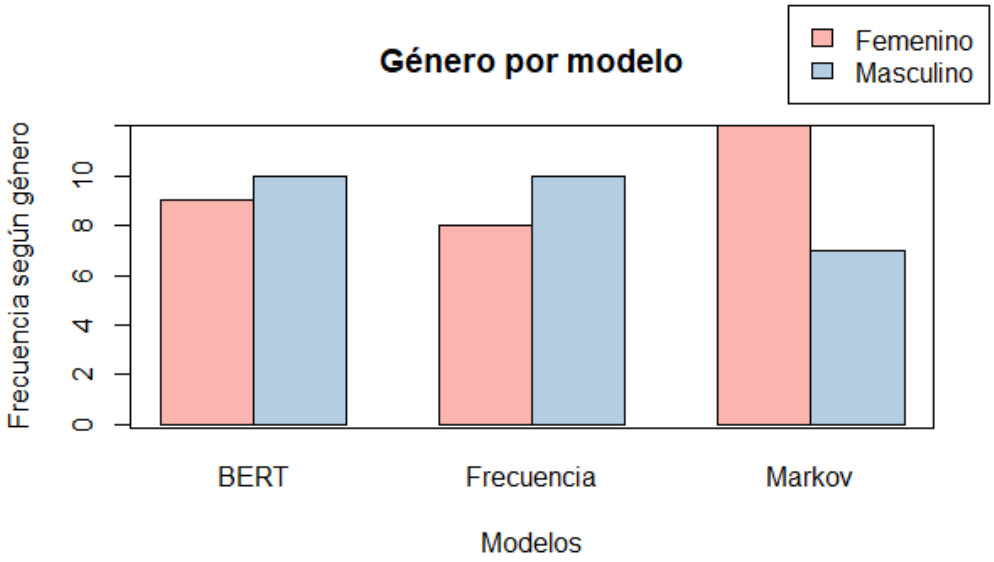


Figura C.3: Distribución de género según modelo

Mostrando una diferencia de 1 entre el género masculino y femenino para el modelo de BERT, una diferencia de 2 para el modelo de Frecuencias y una diferencia de 5 para el modelo de Markov.

Considerando la distribución etaria de la muestra entre cada modelo, se consiguieron los siguientes valores para promedio, desviación estándar, mediana, mínimo y máximo.

Tabla C.2: Distribución etaria según modelo de predicción asignado

Modelo	**Promedio**	**Desviación Estándar**	**Mediana**	**Mínimo**	**Máximo**
BERT	39.9	10.1	43	23	56
Frecuencias	45.9	7.67	43	37	67
Markov	43.8	9.77	46	26	58