Universidad Autónoma de Querétaro
Facultad de Ingeniería
Maestría en Ciencias en Inteligencia Artificial

Procesamiento de lenguaje natura para la búsqueda de patrones en demencia

Tesis

Que como parte de los requisitos para obtener el Grado de
Maestría en Ciencias en Inteligencia Artificial

Presenta:
Damián Solís Rosas
Dirigido por:
Dr. Saúl Tovar Arriaga

Dr. Saúl Tovar Arriaga
Presidente
Dr. Arturo González Gutiérrez
Secretario
Dr. Wilfrido Jacobo Paredes García
Vocal
Dr. Humberto Güendulain Arenas
Suplente
Dr. Marco Antonio Aceves Fernández
Suplente

Centro Universitario, Querétaro, Qro.
Enero 2020
México

A mi esposa, padres y hermanos

AGRADECIMIENTOS

Agradezco al Consejo Nacional de Ciencia y Tecnología y la Universidad Autónoma de Querétaro por el apoyo económico para el estudio de mi maestría y la realización de esta investigación.

Agradezco especialmente al Dr. Saúl Tovar Arriaga por toda la paciencia, apoyo y comprehensión que me brindó durante toda la maestría.

Expreso mi gran agradecimiento a la Dra. Davis Boyd por su apoyo valioso para obtener acceso a la base de datos *Carolina Conversation Collection*.

A mi esposa y a mis padres les agradezco todo el apoyo brindado para cumplir todas mis metas.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Los efectos en la capacidad lingüística de las personas con algún tipo de demencia se reflejan en el léxico (sus diccionarios mentales y su habilidad para entender palabras complejas) más que en su habilidad para formular enunciados completos y fluentes.

Análisis indican que la riqueza del léxico y la fluencia para hablar no son buenas cualidades en personas que padecen algún tipo de demencia (Bucks et al 2000).

Existen estudios previos de las patologías del habla, las cuales incluyen el uso de pausas, palabras de relleno, palabras inventadas, reinicios, repeticiones, enunciados incompletos y difluencias (Guinn and Singer, 2014). Todos los factores anteriores se pueden presentar en individuos con algún tipo de demencia.

A través de un análisis estadístico de métricas lingüísticas se buscan patrones para clasificar conversaciones de personas con demencia y personas sin demencia. Se usan dos algoritmos de aprendizaje automático para la clasificación binaria de la presencia o ausencia de demencia. El primero es una red neuronal de 3 capas la cual obtuvo una exactitud del 78.01 % en la clasificación de las conversaciones, y el segundo una máquina de soporte de vectores el cual obtuvo un 86.42% de exactitud.

(**Palabras clave**: procesamiento de lenguaje natural, patrones de demencia, clasificación de texto, máquina de soporte de vectores, redes neuronales)

SUMMARY

The effects on the linguistic capacity of the people with some type of dementia are reflected in the lexicon (his mental dictionaries and his ability to understand complex words) rather than his ability to formulate complete and fluent enunciations.

Analysis indicate that the richness of the lexicon and the fluency to speak are not good qualities in people who suffer dementia (Bucks et al 2000).

There are previous studies of the pathologies of speech, which include the use of pauses, words of filling, words invented, restarts, repetitions, incomplete statements and diffluent speech (Guinn and Singer, 2014). All previous factors may occur in individuals with some type of dementia.

Through discriminatory analysis of conversation and metrics analysis, we search patterns to classify conversations of people who suffer dementia and people who don't suffer it. Additionally, we use two machine learning algorithms to automatically classify presence or absence of dementia. The first one, a 3-layer neural network reaching a binary classification accuracy of 78.3%, and the second a support vector machine reaching a binary classification accuracy of 86.42%.

(**Keywords**: natural language processing, dementia patterns, text classification, support vector machine, neural networks)

1. INTRODUCCIÓN

1.1 Inteligencia Artificial

El término de inteligencia artificial surge en 1956 en una reunión de investigadores organizada por John McCarthy, estos investigadores fueron los pioneros en este campo de la computación. En esta reunión se propuso la que puede ser considerada como la primera definición de inteligencia artificial (IEEE, 2002). Como parte de la propuesta se incluyó que el problema de la inteligencia artificial es el de construir una máquina que se comporte de manera que, si el mismo comportamiento lo realizara un ser humano, el comportamiento sería llamado inteligente.

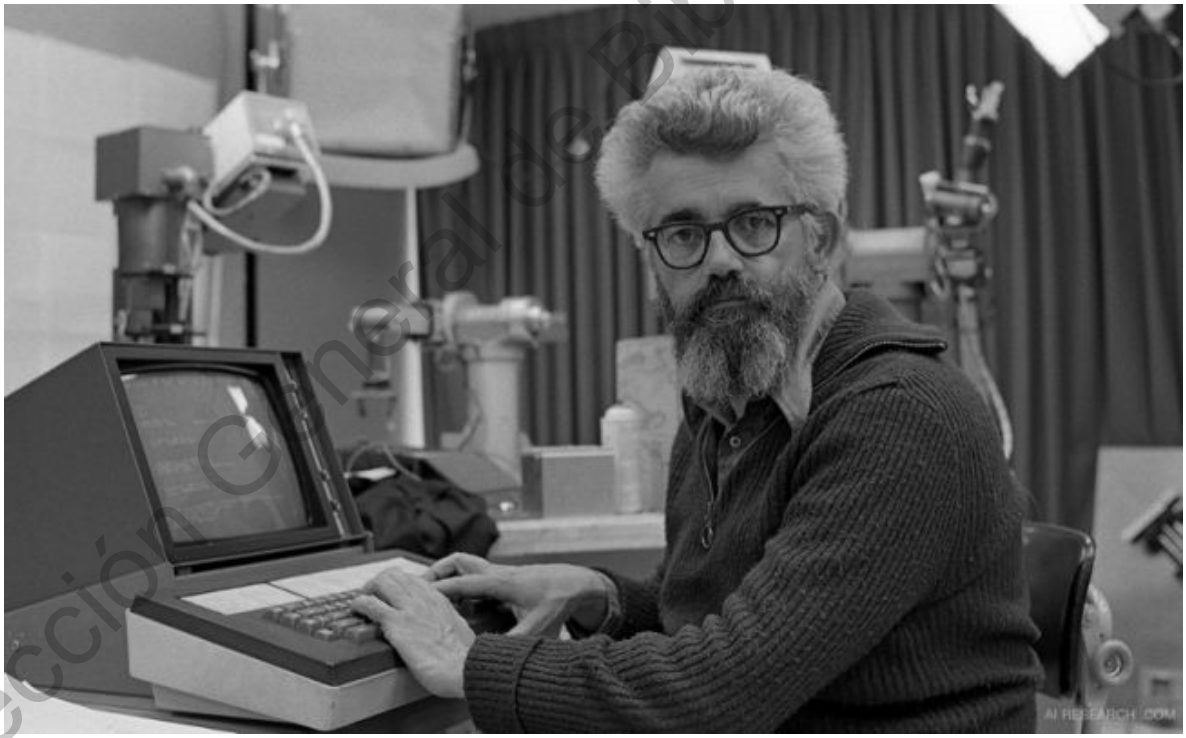

Figura 1: John McCarthy, él introdujo el término de Inteligencia Artificial. Fuente: (Chuck Painter, 1974).

Cabe mencionar que el primer término que sugirió McCarthy para nombrar al estudio de las máquinas pensantes fue el de “estudio de autómatas”, pero al preparar la

propuesta para la reunión antes mencionada pensó que sería mejor un nombre con más marketing. Años más tarde McCarthy mencionó que también considero el término de inteligencia computacional. McCarthy además de ser considerado el padre de la inteligencia artificial creo el lenguaje de programación LISP (List-Processing) el primer lenguaje de programación funcional.

El concepto de inteligencia artificial cada vez es más usado en la vida cotidiana. El mundo de la tecnología se ha encargado de proveernos de sistemas inteligentes. En la industria se han implementado sistemas inteligentes que mejoran los procesos, en la vida cotidiana usamos dispositivos inteligentes que nos facilitan tareas, nos brindan nuevas formas entretenimiento y facilitan la comunicación.

El término “inteligente” empleado en la tecnología se refiere a una rama de la computación, la inteligencia artificial, esta busca simular procesos de la inteligencia humana con técnicas de computación. La inteligencia artificial busca simular características de la inteligencia humana tales como la capacidad de aprendizaje, autocorrección y razonamiento entre otras.

La inteligencia artificial se puede dividir en los siguientes campos de estudio:

Figura 2: Campos de estudio de la inteligencia artificial. Fuente: (Medium, 2018).

1.1.1 Procesamiento de lenguaje natural

El lenguaje es un sistema de signos, orales, escritos o gestuales, que a través de su significado y la relación permiten que las personas puedan expresarse para lograr el entendimiento con el resto.

El lenguaje natural es el lenguaje que usa el ser humano para comunicarse todos los días, el cual fue construido con reglas y convenciones lingüísticas con el fin de lograr la comunicación. La comunicación es un proceso que implica millones de conexiones neuronales y procesos complejos de comprensión y captación.

La riqueza de los componentes semánticos da a los lenguajes naturales un gran poder expresivo, la sintaxis de un lenguaje natural puede ser modelada por un lenguaje formal.

El lenguaje natural hace referencia a la comunicación oral, escrita y señas. En la siguiente figura se muestra la estructura con ejemplos del lenguaje natural.

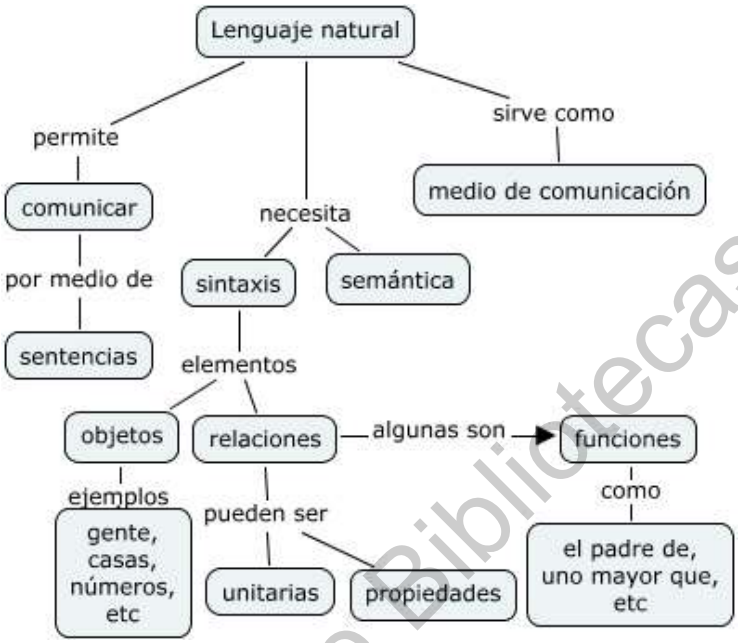

Figura 3: Estructura del lenguaje natural. Fuente: (Cursa, 2018).

El procesamiento de lenguaje natural o NLP por sus siglas en inglés es un campo de estudio interdisciplinario entre la inteligencia artificial y la lingüística el cual se ocupa de estudiar e investigar la manera de comunicar las máquinas con los seres humanos mediante el uso de lenguajes naturales. El objetivo de estudio del NLP es entender, interpretar y manipular el leguaje natural.

En los años 50´s Alan Turing propuso un test con el fin de determinar la habilidad de una máquina para mostrar su inteligencia (Turing, 1950). La forma de evaluar dicho test es emitiendo un juicio de una conversación entre la máquina a evaluar y un humano usado lenguaje natural. A lo largo de la historia de la computación se ha buscado que una máquina tenga la capacidad de entender y expresarse tal y como lo haría un humano (Hugo Banda, 2014).

El NLP procesa los lenguajes naturales para diseñar mecanismos que simulen la comunicación, para lograr esto el NLP hace diferentes análisis de lenguaje natural

basados en la gramática. A continuación, se muestra un esquema del estudio de la gramática:

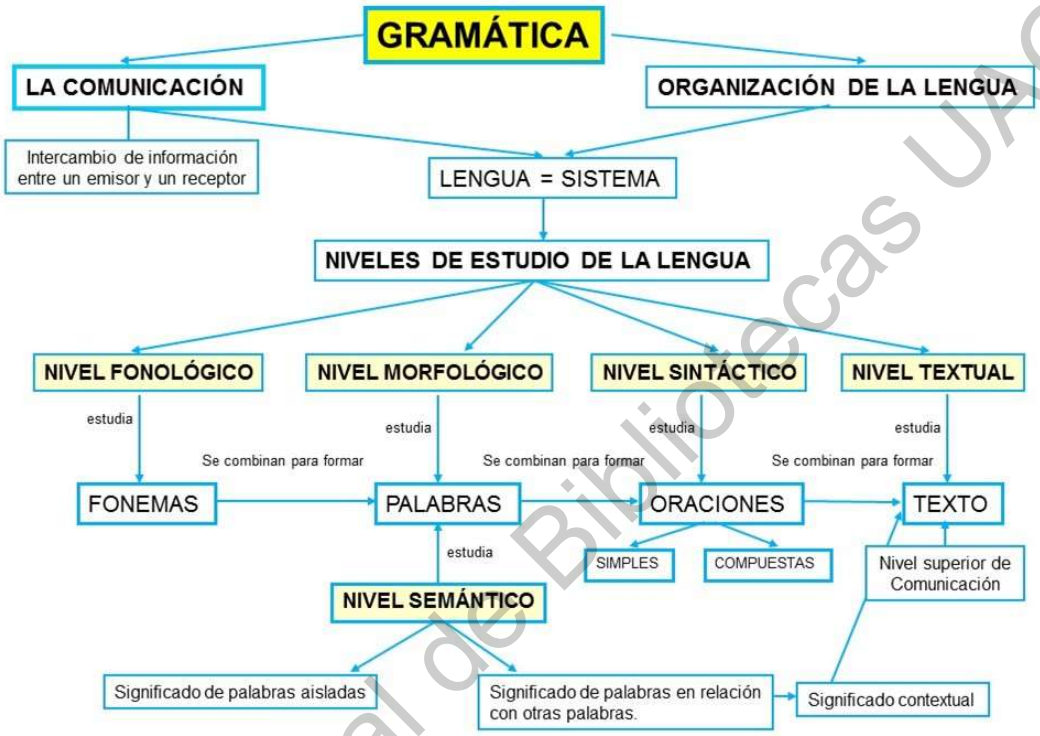

Figura 4: Niveles de estudio de la gramática. Fuente: (MDM, 2015).

El análisis gramatical del lenguaje natural que realiza el NLP, se realiza en los diversos niveles de estudio de la lengua, estos niveles están ejemplificados en el diagrama anterior. Este análisis gramatical se divide en análisis para cada nivel los cuales son:

- Análisis morfológico - El análisis de las palabras para extraer sus raíces, rasgos reflexivos y unidades léxicas.
- Análisis sintáctico - Análisis de la estructura sintáctica de las frases mediante gramáticas.
- Análisis semántico – Extracción del significado de las frases, y la resolución de ambigüedades estructurales y léxicas.

- Análisis pragmático – Análisis del contexto de uso a la interpretación final.
- Análisis textual - Extracción de oraciones y descomposición de las mismas.
- Análisis fonológico – Extracción y procesamiento de fonemas, de esta manera se logran diferenciar palabras con pronunciación similar.

En la siguiente figura se muestran aplicaciones del NLP, algunas de estas son usadas en este estudio:

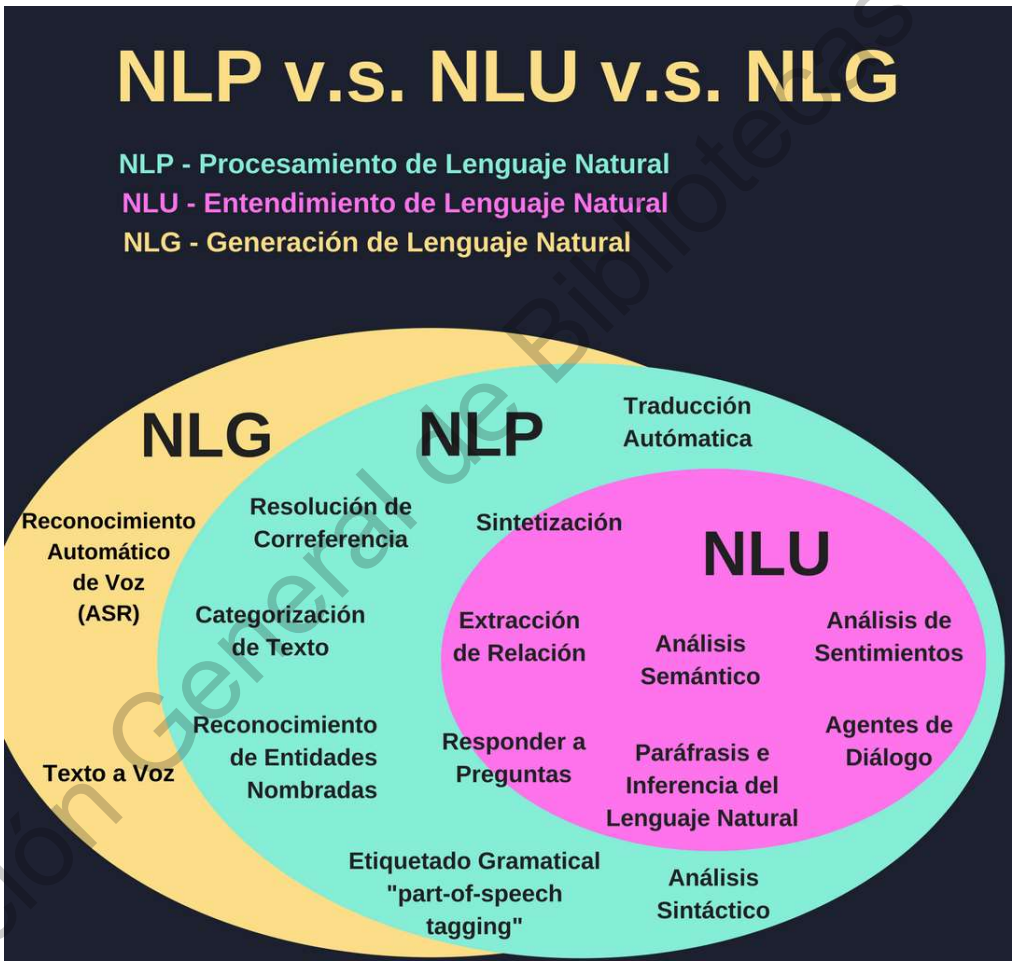

Figura 5: Aplicaciones del NLP. Fuente: (Bill MacCarteny, 2014).

1.2 Demencia y el Lenguaje

La demencia es un término general que describe un deterioro de la capacidad mental que se presenta en varios síntomas. Impacta a la sociedad de manera negativa ya que la gente padece de deterioro en la memoria y deterioro en sus capacidades de razonamiento, impidiendo o afectando la realización de tareas cotidianas (Alzheimer´s Association, 2014).

Una gran mayoría de los pacientes con demencia se caracterizan por la degradación de su lenguaje y su funcionalidad cognitiva resultando en complicaciones significativas en la comunicación vocal. Los efectos en la capacidad lingüística de las personas con demencia se reflejan en el léxico (sus diccionarios mentales y su habilidad para entender palabras complejas) más que en su habilidad para formular enunciados completos y fluentes (Guinn, 2012).

Análisis indican que la riqueza del léxico y la fluencia para hablar no son buenas cualidades en personas que padecen demencia (Bucks et al 2000). Existen estudios previos de las patologías del habla, las cuales incluyen el uso de pausas, palabras de relleno, palabras inventadas, reinicios, repeticiones, enunciados incompletos y difluencias. Todos los factores anteriores se presentan en individuos con demencia.

Las cualidades gramaticales de las personas con demencia (Janeth, 2006) en estado inicial se muestran en la Tabla 1 y 2:

Tabla 1: Cualidades de la comprensión en estado inicial.

NIVEL GRAMATICAL	CUALIDADES
SEMÁNTICA	Dificultad para comprender oraciones de contenido complejo (analogías, doble sentido, humor). Comprensión de ideas simples conservada.

SINTÁCTICA	Dificultad leve para comprender oraciones de organización extensa y/o compleja.
FONOLÓGICA	Recuperación conservada de forma léxica fonológica.
PRAGMÁTICA	Olvido ocasional de su interlocutor. Pérdida de preguntas y referencias del narrador.
LECTOESCRITURA	Comprensión lectora conservada para contenidos simples.

Tabla 2: Cualidades de la expresión en estado inicial.

NIVEL GRAMATICAL	CUALIDADES
SEMÁNTICA	Dificultad para recuperar palabras en conversación espontánea. Lenguaje fluido, pero poco concreto. Sustitución de palabras (parafasias semánticas, uso de circunloquios).
SINTÁCTICA	Estructura sintáctica conservada. Repetición afectada para oraciones largas.
FONOLÓGICA	Sistema fonológico conservado.
PRAGMÁTICA	Disminución en tiempo y contenido del discurso. Divagación y tópico difuso. Olvido, reiteración e ideas incompletas en la conversación. Auto monitoreo y autocorrección.
LECTOESCRITURA	Escritura alterada en la forma (ej., pérdida de acentuaciones, disortografía).

Las cualidades gramaticales de las personas con demencia en estado moderado se muestran en la Tabla 3 y 4:

Tabla 3: Cualidades de la comprensión en estado moderado.

NIVEL GRAMATICAL	CUALIDADES
SEMÁNTICA	Dificultad moderada para recuperar palabras en conversación espontánea y en tareas específicas. Comprensión escasa de frases de contenido complejo.
SINTÁCTICA	Comprensión de oraciones sintácticamente simples. Disminución en la comprensión de secuencias y series.
FONOLÓGICA	Escaso recobro de la representación auditiva de la palabra. Dificultades en el procesamiento del lenguaje oral.
PRAGMÁTICA	Olvido ocasional de su interlocutor. Incomprensión de rasgos intencionales de la conversación.
LECTOESCRITURA	Lectura en voz alta conservada, pero sin atribución de sentido. Comprensión lectora alterada para mensajes de alta complejidad gramatical.

Tabla 4: Cualidades de la expresión en estado moderado.

NIVEL GRAMATICAL	CUALIDADES
SEMÁNTICA	Dificultad para tareas de nominación y categorización (aunque mejora con claves). Aumento y predominio de parafasias semánticas. Reducción del vocabulario expresivo.
SINTÁCTICA	Producción verbal con formas sintácticas reconocibles, pero cortas. Se mantienen morfemas sintácticos. Repetición afectada para oraciones

	simples. Omisión de conectores y palabras funcionales de la oración. Conservación de estereotipos verbales sociales.
FONOLÓGICA	Confusión ocasional de patrones de pronunciación. Parafasias fonológicas (cambios de sonidos dentro de una palabra).
PRAGMÁTICA	Frases inacabadas. Repetición de ideas en la conversación. Olvido de su interlocutor. Pérdida del tópico y abandono de la conversación. Auto monitoreo y autocorrección poco frecuente. Ecolalia. Limitación en la toma de turnos en la conversación.
LECTOESCRITURA	Escritura de palabras y frases cortas. Reducción de la variedad de elementos en la redacción. Dificultad para iniciar de forma espontánea la escritura.

Las cualidades gramaticales de las personas con demencia en estado avanzado se muestran en la Tabla 5 y 6:

Tabla 5: Cualidades de la comprensión en estado avanzado.

NIVEL GRAMATICAL	CUALIDADES
SEMÁNTICA	Comprensión únicamente de elementos significativos. Limitaciones importantes para nominar y recuperar las palabras (incluso con claves). Comprensión solamente de palabras u oraciones muy cortas o con elementos familiares.
SINTÁCTICA	Comprensión severamente alterada para frases o limitada a palabras cortas y familiares.
FONOLÓGICA	Imprecisiones en la conversión fonológica.

PRAGMÁTICA	Indiferencia casi total hacia su entorno e interlocutor. Pérdida de ideas claves del discurso del otro. No reconoce patrones de intencionalidad.
LECTOESCRITURA	Afectación severa para la comprensión lectora.

Tabla 6: Cualidades de la expresión en estado avanzado.

NIVEL GRAMATICAL	CUALIDADES
SEMÁNTICA	Reducción del vocabulario. Utilización únicamente de elementos significativos. Predominio de parafasias semánticas (cambio de una palabra por otra).
SINTÁCTICA	Limitación en el uso de lenguaje automático. Repetición casi extinta, incluso para palabras monosílabas. Omisión frecuente de palabras funcionales (artículos, conjunciones, adverbios).
FONOLÓGICA	Parafasias fonológicas. Ecolalia (repetición parcial o total de forma incontrolada de frases expresadas por el interlocutor). Palilalia (repetición de una palabra de forma incontrolada). Logoclonias (repetición de sílabas) de forma incontrolada.
PRAGMÁTICA	Conversación casi ausente, limitada o imprecisa. Circunloquios (designar de forma indirecta un concepto a través de sus características). Ausencia de auto monitoreo y autocorrección. Mutismo. Imposibilidad para mantener el tópico.
LECTOESCRITURA	Afectación casi total de la escritura por componente apráxico. Puede conservarse la escritura de letras y/o palabras monosílabas.

1.3 Descripción del problema

La demencia como se pueden observar en las Tablas 1,2,3,4,5 y 6 se puede clasificar en 3 estados o fases, cada una con ciertas características.

En el estado inicial, las personas pueden olvidar rápidamente lo que han escuchado, visto o pensado, es común que pierdan el tópico y repiten una idea muchas veces, en consecuencia, se les dificulta seguir una conversación.

En el estado intermedio, las personas suelen estar desorientadas en tiempo y espacio, también pasar momentos en los que tienen gran déficit de memoria, por lo cual no pueden recordar hechos recientes, estas características generan una producción verbal pobre, sin sentido y poco elaborada.

En el estado final (el más grave), las personas tienen severos problemas para prestar atención, para codificar, recuperar información, tienen problemas de percepción y sus funciones ejecutivas están limitadas o carecen de estas. La memoria semántica y la capacidad para recordar conceptos están muy deteriorados o parecen ausentes en la persona.

En todas las etapas de la demencia, las características y capacidades lingüísticas están deterioradas, lo cual puede servir como objeto de estudio y análisis para su detección temprana o modelación de la evolución del trastorno.

Una gran mayoría de los pacientes con demencia se caracterizan por la degradación de su lenguaje y su funcionalidad cognitiva resultando en complicaciones significativas en la comunicación vocal.

Las características antes mencionadas se reflejan en el lenguaje cotidiano usado por las personas que padecen demencia. Las conversaciones que ellos mantienen contienen estas características. Para poder analizar y estudiar estas características se necesita hacerlo sobre las conversaciones, siendo fundamental tener una

colección de conversaciones de personas que padecen demencia y personas que no.

Obtener una base de datos basado en forma de texto de las conversaciones es un problema que se puede afrontar usando software, actualmente hay muchas herramientas para convertir conversaciones a texto.

Una vez que se tiene la base de datos en forma de texto, descomponer el texto en forma de oraciones, analizar sintáctica y semánticamente cada oración se puede lograr mediante herramientas basados en algoritmos creados por lenguajes de programación.

Existen herramientas tales como Natural Language Tool Kit (NLTK) la cual ha implementado muchas funciones básicas para el análisis de texto, usando las funciones básicas y desarrollando funciones más complejas, se puede lograr un análisis muy completo en cuestión de sintáctica y semántica con el fin de analizar las características léxicas de las conversaciones, en este caso de las conversaciones de personas con algún tipo de demencia.

También existen varias librerías para Python con las cuales se pueden implementar redes neuronales y máquina de soporte de vectores. Tales como:

- TensorFlow (1.13.1), (TensorFlow, 2019).
- Keras, (2.2.4-tf), (Keras, 2019).
- Scikit-Learn, (0.21.3), (Scikit-Learn, 2019).

1.4 Justificación

El aumento en la esperanza de vida en los últimos años ha traído con ello un aumento en el índice de algunas enfermedades, ya que a mayor edad mayores son las complicaciones en la salud a las que se enfrenta el ser humano.

En México la esperanza de vida para las mujeres es de 78 años y para los hombres de 73 años (INEGI, 2018). Estas cifras aumentan con el paso de los años, así como las enfermedades asociadas con la edad, estas enfermedades afectan directamente la calidad de vida de las personas.

Actualmente en México hay poco más de 12 millones de adultos mayores, lo que equivale aproximadamente al 12% de la población. El porcentaje de la población de adultos mayores en México se estima que crezca año con año, este aumento significa que habrá más personas que atender en temas de salubridad (INEGI, 2017).

En México los reportes clínicos indican que hay más de 750 mil personas que padecen algún tipo de demencia, siendo el Alzheimer la más frecuente. Se prevé que para 2050 la cifra alcanzará más de 3.5 millones de personas con algún tipo de demencia (INEGI, 2017).

Para el diagnóstico de demencia, los médicos realizan pruebas para evaluar el deterioro de la memoria y otras habilidades de razonamiento, determinar las capacidades funcionales e identificar cambios en la conducta. También llevan a cabo una serie de pruebas para descartar otras posibles causas de deterioro. Este diagnóstico puede ser tardado, caro y tedioso para el paciente.

El diagnóstico que es basado en la memoria analiza las capacidades de la persona, si el paciente tiene demencia esto se refleja en deterioros en la capacidad de aprender, memorizar y recordar. Todas esas características son reflejadas en el habla, ya que, al tener dificultad para poder memorizar, aprender y recordar, el léxico del paciente tiene características específicas.

En base a las características léxicas de una persona que padece algún tipo de demencia y el uso de técnicas de inteligencia artificial se puede lograr un algoritmo para sustraer las características del habla de una persona que tiene demencia y una que no.

2. ANTECEDENTES

Los efectos en la capacidad lingüística de las personas con demencia se reflejan en el léxico (sus diccionarios mentales y su habilidad para entender palabras complejas) más que en su habilidad para formular enunciados completos y fluentes (Singh, 2000).

Una de las áreas más importantes afectadas por la enfermedad es la capacidad de comunicación funcional, a medida que las habilidades lingüísticas se deterioran. Resultados indican que las soluciones obtenidas con métodos computacionales ofrecen una alternativa viable a las evaluaciones tradicionales para diagnosticar el nivel de discapacidad en los pacientes con algún tipo de demencia. Estos resultados son un avance significativo hacia medios automáticos y objetivos para identificar los primeros síntomas de demencia en adultos mayores (Khodabakhsh, 2014).

Se han desarrollado algoritmos para el diagnóstico automático enfocado en las características derivadas de las conversaciones con personas. Oponiéndose al diagnóstico tradicional basado en la memoria, esta prueba se basa en conversaciones informales con las personas.

Según Pope & Davis (2011) la riqueza del léxico y la fluencia para hablar no son buenas cualidades en personas que padecen demencia. Existen estudios previos de las patologías del habla, las cuales incluyen el uso de pausas, palabras de relleno, palabras formuladas, reinicios, repeticiones, enunciados incompletos y difluencias. Todos los factores anteriores se presentan en individuos con demencia (Davis, 2009).

Se han desarrollado algoritmos de detección automática de demencia (Roark, 2011) utilizando análisis de métricas en transcripciones de bases de datos.

Guinn (2014) realizó análisis discriminativos de conversaciones entre personas que sufren demencia y personas que no. Estos análisis fueron hechos a métricas de las transcripciones de una base de datos para determinar si hay diferencias estadísticas significativas entre personas con y sin demencia. En este estudio se identificó una diferencia lingüística medible entre los exámenes realizados a personas que padecen demencia y personas que no.

A continuación, se muestra una tabla del estado del arte, en la cual se incluyen los autores del estudio, el año de publicación, el formato de entrada de la base de datos usada, el nombre de la base de datos usada, el clasificador o análisis implementado y la exactitud obtenida:

Tabla 7: Estado del arte.

Autor	Año	Formato de entrada	Técnica	Base de datos	Resultados
C. Guinn Ben Singer A. Habash [ ]	2015	Conversaciones Transcritas	Árbol de decisión Clasif. Bayesiano	Carolina Corpus Conversations	80%
A. Khodabakhsh S. Kuscuoglu C. Demiroglu [ ]	2014	Conversaciones Transcritas	Árbol de decisión SVM	Propia	76.34%
B. Roark M. Mitchell J. Hosom K. Hollingshead [ ]	2011	Audio	SVM	Oregon Health and Science University	86%
B. Roark M. Mitchell J. Hosom J. A. Kaye [ ]	2011	Audio	Dif. en el habla Marcadores del lenguaje	Layton Aging and Alzheimer's	84.4%
C. Thomas M. Mitchell Vlado Keselj Kenneth Rockwood []	2005	Conversaciones Transcritas	Clasif. Bayesiano	ACADI Dataset	90%
R. S. Bucks S. Singh J. M. Cuerden G. K.Wilcock []	2010	Conversaciones Transcritas	Análisis de métricas lingüísticas	Propia	87.5%
R. S. Bucks S. Singh J. M. Cuerden G. K.Wilcock []	2011	Audio	Análisis fonético	Propia	-

El mejor resultado en la clasificación de conversaciones lo obtuvieron C. Thomas y M. Mitchel usando un clasificador Bayesiano en el 2005 usando la base de datos ACADI Dataset.

3. HIPÓTESIS

Es posible encontrar patrones en las conversaciones de personas con demencia y sin demencia usando técnicas de procesamiento del lenguaje natural e inteligencia artificial.

4. OBJETIVOS

4.1 Objetivo general

Clasificar por medio de la búsqueda de patrones en conversaciones transcritas la presencia o ausencia de algún tipo de demencia.

4.2 Objetivos Específicos

- Obtención de la base de datos de las conversaciones de personas con demencia y sin demencia
- Análisis estadístico de los datos
- Estudio y aplicación del Procesamiento de Lenguaje Natural
- Implementación de clasificadores automáticos

5. METODOLOGÍA

5.1 Descripción general

El procedimiento para la obtención de resultados consta de:

- Una base de datos: La cual consta de conversaciones transcritas de pacientes con y sin demencia.
- NLP: Procesamiento de lenguaje natural de la base de datos.
- Clasificador automático: Clasificación de las conversaciones usando datos de la base y del NLP.

El procedimiento antes mencionado se simboliza en la siguiente figura.

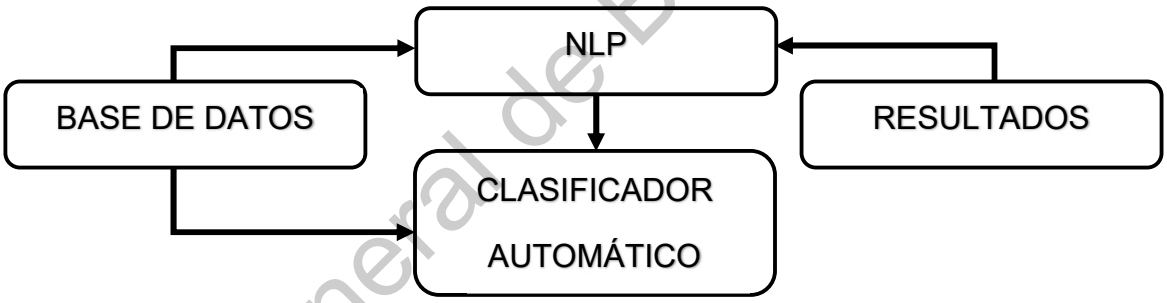

Figura 6: Metodología para la obtención de resultados.

Las conversaciones transcritas forman una base de datos, la cuál es procesada usando técnicas de NLP. El clasificador automático hace dos clasificaciones basadas en diferentes tipos de datos; para la primera clasificación los datos se toman directamente de la base de datos, para la segunda clasificación los datos se toman del procesamiento del lenguaje natural de nuestra base de datos. Los resultados obtenidos del NLP y del clasificador automático son analizados.

5.2 La base de datos

Se buscó una base de datos que tuviera conversaciones es español de personas Mexicanas con algún tipo de demencia, lamentablemente no existe ninguna con estas características en el idioma español, por este motivo se optó por buscar una base de datos en otro lenguaje.

Una base de datos que cumple con las características necesarias para este estudio es Carolina Conversation Collection (UNCC, 2008), la cual es administrada por la Universidad de Coralina del Norte de Charlotte. Esta consta de conversaciones transcritas del idioma inglés entre un entrevistador y un paciente. Los pacientes presentes en esta base de datos pueden tener algún tipo de demencia.

Los pasos para la obtención de la base de datos fueron los siguientes:

- Justificar el por qué es necesario el uso de estos datos y para qué fin serán usados.
- Mostar aprobación oficial del comité de ética de la institución a la que pertenece.
- Cursar y aprobar el Collaborative Institutional Training Initiative (CITI), el cual es un curso para el manejo de datos personales.

La justificación del uso de la base de datos se basó en la hipótesis de este trabajo: encontrar patrones en las conversaciones de personas con demencia y sin demencia usando técnicas de procesamiento del lenguaje natural e inteligencia artificial.

La aprobación oficial por para de la institución de pertenencia se logró justificando el uso de la base datos para la hipótesis antes mencionada y manifestando que los resultados son a partir de datos en el idioma inglés y la aplicación de los resultados a la población Mexicana está sujeta a una investigación adicional ya que no existe

una base de datos en México con las características necesarias para probar la hipótesis.

El curso consta de diferentes módulos, los cuales tratan temas tales como: principios éticos, regulaciones federales, protección de datos personales, consentimiento, regulaciones y procesos, investigación en entornos educativos, investigación en prisioneros, investigación genética, personas que requieren consideraciones y protecciones adicionales, evaluación de riesgos, investigaciones que involucren niños, privacidad, confidencialidad y protección del uso de datos personales. Cada módulo contiene un examen, los resultados de cada examen son reflejados en un puntaje. Para la aprobación del curso CITI se requería un mínimo de 90 puntos en la escala del 1 al 100, el puntaje obtenido fue de 96 (ver Anexo 9.1).

La base de datos está formada por conversaciones las cuales pueden ser encontradas mediante un buscador, los criterios de búsqueda pueden ser: el nombre del archivo, la fecha, el género del entrevistador, edad del paciente, lugar de nacimiento del paciente, lugar de residencia del paciente y padecimiento.

5.3 Las conversaciones

Las conversaciones transcritas en la base de datos se encuentran separadas en archivos de texto. Cada archivo tiene un encabezado el cual incluye: el nombre del archivo, un número de identificación del archivo, una lista de participantes en la conversación, los participantes con su número de identificación y un asterisco que diferencian el nombre del paciente y el del entrevistador. A continuación, se muestra un ejemplo del encabezado con datos ficticios:

```
TRANSCRIPT:     PersonX_Interviewer_01.txt
ID:     1000800

SPEAKERS
ID      NAME
1       Interviewer
2*      PersonX
```

Figura 7: Encabezado de las conversaciones en el Carolina Conversation Collection.

Después del encabezado los archivos contienen los diálogos entre el paciente y el entrevistador, cada diálogo comienza con el nombre del paciente y con el nombre del entrevistador. A continuación, se muestra un fragmento de una conversación con datos ficticios:

```
Interviewer: How are you ?

PersonX:  Fine.

Interviewer:  Did you take your breakfast ?

PersonX:  Yes.

Interviewer: What did you eat ?

PersonX:  An egg, milk and fruit.
```

Figura 8: Ejemplo de un fragmento de conversación.

En los diálogos las pausas cortas, las pausas largas y las muestras de aturdimiento están reflejadas por diferentes cadenas de signos, por ejemplo: el símbolo '[.]' reflejan una pausa corta, los símbolos '[…]' una pausa larga y los símbolos '[?]' una muestra de aturdimiento o confusión. En la siguiente figura se muestra un ejemplo de una conversación transcrita ficticia con pausas cortas, pausas largas y muestras de aturdimiento o confusión:

```
Mr. X: I am [.] fine .

Mr. X: I went to [...] my house.

Mr. X: I don´t know [?].
```

Figura 9: Ejemplo de pausa corta, pausa larga y confusión.

5.2 Procesamiento de las conversaciones

Para poder aplicar las técnicas del NLP, así como el análisis de métricas lingüísticas se realizó un preprocesamiento a cada conversación el cual se describe a continuación.

- Se eliminó el encabezado de cada archivo.
- Se eliminaron todos los nombres de los participantes en los diálogos.
- Se extrajeron solo los diálogos del paciente.

A continuación, se muestra un ejemplo de un archivo sin pre-procesar a la izquierda y a la derecha se muestra el mismo archivo pre-procesado:

```
TRANSCRIPT:     PersonX_Interviewer_01.txt
ID:     1000800

SPEAKERS
ID      NAME
1       Interviewer
2*      PersonX


Interviewer: How are you ?

PersonX:  Fine.

Interviewer:  Did you take your breakfast ?

PersonX:  Yes.

Interviewer: What did you eat ?

PersonX:  An egg, milk and fruit.
```

```
Fine. Yes. An egg, milk and fruit.
```

Figura 10: A la izquierda un archivo sin pre-procesar, a la derecha ya pre-procesado.

Como se muestra en la figura anterior, los nombres se eliminaron, solo se mantuvieron los diálogos del paciente, esto con el fin de no usar datos personales y evitar que el nombre del paciente afecte las métricas lingüísticas.

Una vez que se extrajeron solo los diálogos del paciente, todo el texto se cambió a minúsculas, se eliminaron todos los signos de puntuación y se codificaron todas las pausas y muestras de aturdimiento o confusión, a continuación, se muestra un ejemplo del proceso mencionado anteriormente:

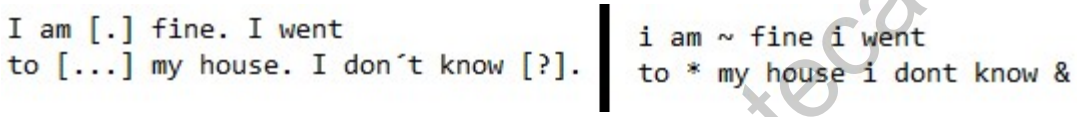

Figura 11: Ejemplo del procesamiento de los diálogos.

5.2 Creación de la base de datos

Todos los archivos procesados fueron separados en carpetas de acuerdo a si el paciente padecía algún tipo de demencia o no, de acuerdo a esta separación quedaron dos grupos, el grupo de pacientes con demencia el cual cuenta con 62 conversaciones y el grupo de pacientes sin demencia con 160 conversaciones.

Para poder identificar al grupo que pertenecen las conversaciones en los algoritmos, cada conversación fue etiquetada siguiendo las siguientes reglas:

- Si el paciente tiene demencia la conversación se etiqueta con el número 1.
- Si el paciente no tiene demencia la conversación se etiqueta con el número 0.

Siguiendo las reglas anteriores se creó una base de datos, la cual incluye el diálogo de cada paciente y la etiqueta, a continuación, se muestra un ejemplo con datos ficticios:

text	label
i am ~ fine i went to * my house i dont know &	1
yeah i went to visit my family was great we ate pizza	0

Figura 12: Ejemplo de la base de datos creada.

La base de datos creada cuenta con todas las características necesarias para poder procesar, analizar y clasificar las conversaciones.

5.3 Procesamiento de Lenguaje Natural

Para poder analizar y clasificar las conversaciones es necesario aplicar el procesamiento de lenguaje natural a cada diálogo. Para esto se usó el Natural Language ToolKit (NLTK, 2019) el cual es un kit de herramientas desarrolladas en Python.

El kit incluye bibliotecas y programas para el procesamiento de lenguaje natural simbólico y estadístico, también incluye bases de datos de diferentes tipos (cuentos, artículos, novelas), tutoriales y programas de ejemplo.

Para procesar las conversaciones se usaron las siguientes técnicas en cada conversación:

- Tokenización: Dividir un diálogo en oraciones y cada oración en palabras.

```
In[0]   : nltk.word_tokenize(text)
Out[0] : ['doing', 'what', 'i', 'can', 'not', 'read', 'it', 'i', 'see']
```

Figura 13: Tokenización..

- Partes de la oración: Cada oración fue dividida en partes de acuerdo a la clasificación de cada parte (Sujeto, Verbo, Adverbio, etc.).

```
In[1]  : nltk.pos_tag(text_tokens)
Out[1] : [('doing', 'Verb'), ('what', 'Pronoun'), ('i', 'Noun')]
```

Figura 14: Partes de la oración.

- Conteo de palabras: Se contó el número de palabras.

```
In[2]  : len(text)
Out[2] : 2885
In[3]  : len(text_tokens)
Out[3] : 687
```

Figura 15: Conteo de palabras.

- Conteo de pausas cortas, pausas largas y muestras de aturdimiento o confusión.

```
In[4]  : len(pauses(text))
Out[4] : 2885
In[5]  : len(long_pauses(text))
Out[5] : 687
In[6]  : len(sconf_stuns(text))
Out[6] : 128
```

Figura 16: Conteo de pausas cortas, pausas largas y muestras de aturdimiento o confusión

- Conteo de palabras con longitud menor o igual a 4.

```
In[7]  : len(words_len_less(text, 4))
Out[7] : 43
In[8]  : words_len_less(text, 4)
Out[8] : [and, 'what', 'i', 'can', 'not', 'read', 'it']
```

Figura 17: Conteo de palabras con longitud menor o igual a 4

- Conteo de palabras con longitud mayor o igual a 5.

```
In[9]    : len(words_len_greater(text, 5))
Out[9]  : 31
In[10]   : (words_len_greater(text, 5)
Out[10]: ['children', 'walked', 'forgot', 'soldiers', 'joined']
```

Figura 18: Conteo de palabras con longitud mayor o igual a 5.

- Extracción y frecuencia de las palabras más usadas con longitud menor o igual a 4.

```
In[11]   : most_freq_words_4(text)
Out[11] : [('and', 49), ('i'. 76), ('my', 23), ('the', 34)]
```

Figura 19: Palabras más usadas con longitud menor o igual a 4.

- Extracción y frecuencia de las palabras más usadas con longitud mayor o igual a 5.

```
In[12]   : most_freq_words_5(text)
Out[12] : [('children', 12), ('walked'. 11), ('forgot', 7)]
```

Figura 20: Palabras más usadas con longitud mayor o igual a 5.

- Conteo y extracción de palabras inusuales.

```
In[13]   : unusual_words(text)
Out[13] : ['evangilistic', 'soldiers', 'preached', 'belonged']
```

Figura 21: Conteo y extracción de palabras inusuales.

- Conteo y extracción de interjecciones.

```
In[14]   : interjections_words(text)
Out[14] : ['uh', 'ah', 'mmm', 'ammm']
```

Figura 22: Conteo y extracción de interjecciones.

- Conteo y extracción de las partes de la oración.

```
In[15]  : freq_pos(text)
Out[15] : [('Noun', 23), ('Verb', 11), ('Article', 7)]
```

Figura 23: Conteo y extracción de las partes de la oración.

Una vez que se procesaron las conversaciones, estas pueden ser usadas en los algoritmos de análisis y clasificación.

5.4 Análisis Estadístico

Una vez procesadas las conversaciones se realizó un análisis estadístico de diferentes métricas lingüísticas de acuerdo al estudio previo de las características en la comunicación de las personas con demencia.

Las métricas analizadas son:

- Palabras de acuerdo a su tipo: En cada grupo de conversaciones se contabilizó el número de palabras de acuerdo a su tipo y se calculó el porcentaje de uso de cada tipo (PTP) dividiendo el número de palabras de un tipo (UTP) entre el número de palabras usadas (TL) tal como lo muestra la fórmula 1:

$$PTP = \left(\frac{UTP}{TL}\right) * 100\% \quad (1)$$

- Pausas cortas, pausas largas y muestras de aturdimiento o confusión: En cada grupo de conversaciones se contabilizó el número de pausas cortas, pausas largas y muestras de aturdimiento o confusión y se calculó el porcentaje de presencia de cada una (PPA) dividiendo el número de pausa corta, pausa larga o aturdimiento (CLA) entre el número de palabras usadas (TL) tal como lo muestra la fórmula 2:

$$PPA = \left(\frac{CLA}{TL}\right) * 100\% \quad (2)$$

- Palabras inusuales: En cada grupo de conversaciones se contabilizó el número de palabras inusuales y se calculó el porcentaje de presencia de estas (PPI) dividiendo el número de palabras inusuales (PI) entre el número de palabras usadas (TL) tal como lo muestra la fórmula 3:

$$PPI = \left(\frac{PI}{TL}\right) * 100\% \quad (3)$$

- Diversidad Léxica: Este índice nos dice que tanto vocabulario fue usado, entre mayor sea el valor significa que más palabras diferentes fueron usadas. En cada grupo de conversaciones se calculó la diversidad léxica (DL) dividiendo la longitud del vocabulario (VL) entre el número de palabras usadas (TL) tal como lo muestra la fórmula 4:

$$\mathrm{DL} = \mathrm{VL}/\mathrm{TL} \quad (4)$$

- Palabras de acuerdo a su longitud: De acuerdo a las características del idioma inglés las palabras de longitud grande suelen ser más significativas ya que generalmente son sustantivos comunes o verbos que pueden representar los temas principales de la conversación y las palabras de longitud pequeña suelen ser artículos, pronombres, etc.

- Interjecciones: Las interjecciones suelen estar muy presentes en diálogos de personas con demencia de acuerdo a las características lingüísticas de estas personas. En cada grupo de conversaciones se calculó el índice de interjecciones usadas (PIN) dividiendo el número de interjecciones (IN) entre el número de palabras usadas (TL) tal como lo muestra la siguiente fórmula:

$$\mathrm{PIN} = \left(\frac{\mathrm{IN}}{\mathrm{TL}}\right) * 100\% \quad (5)$$

5.5 Clasificadores automáticos

Para poder clasificar las conversaciones de personas con demencia y sin demencia se implementaron dos clasificadores automáticos, una red neuronal y una máquina de soporte de vectores.

5.5.1 Red Neuronal

Se implementó una red neuronal para la clasificación de las conversaciones. La red neuronal hace una clasificación binaria ya que se quieren clasificar las conversaciones en: conversaciones de pacientes con demencia y sin demencia.

Las redes neuronales están inspiradas en las redes neuronales biológicas del cerebro humano. Estas están formadas por elementos procesadores que es la unidad análoga a la neurona biológica, estos elementos tienen varias entradas y las combina, la combinación es modificada por una función de transferencia y el valor resultante de esta función se pasa directamente a la salida del elemento procesador.

En concreto una red neuronal consiste en un conjunto de unidades elementales conectadas de una forma concreta. Estas unidades elementales están organizadas en grupos llamados capas.

La capa de entrada y de salida son las únicas con conexiones exteriores a la red. En la capa de entrada se presentan los datos y en la capa de salida se presenta la respuesta de la red a la entrada, las capas restantes son llamadas “capas ocultas”.

Para poder entrenar la red neuronal cada conversación fue codificada ya que el texto no puede ser la entrada en la red neuronal.

Para poder codificar las conversaciones se realizó una normalización del texto para la cual se extrae la palabra raíz de cada palabra, a continuación, se muestra un ejemplo de la normalización de las palabras 'liked', 'likes', 'likely' y 'liking':

```
'liked'  -> 'like'
'likes'  -> 'like'
'likely' -> 'like'
'liking' -> 'like'
```

Figura 24: Ejemplo de normalización de texto

Todas las palabras ejemplificadas en la figura anterior provienen de la palabra raíz 'like', con la normalización del texto se obtienen solo las palabras raíz reduciendo significativamente el vocabulario a analizar.

La codificación del texto se hizo basándose en un diccionario del idioma inglés incluido en el NLTK, en este diccionario están todas las palabras raíz del idioma inglés, a cada palabra se le tiene asignado un número único, de esta manera se pudieron codificar todas las palabras a números y así poder usarlos como entrada de la red neuronal.

A continuación, se muestra como la oración ‘the neural networks can be used to classify text’ es codificada:

```
In[16]  : word_to_id('the neural networks can be used to classify text')
Out[16] : [4, 73517, 8060, 70, 30, 343, 8, 12371, 3004]
```

Figura 25: Ejemplo de una oración codificada

Una vez que se tienen los datos de entrada listos se entrenó la red neuronal.

La red neuronal implementada es un modelo secuencial, es una pila de capas lineares. Esta cuenta con 3 capas, la primera y la segunda capa tienen 16 nodos e

implementan una función de activación rectificadora lineal, la tercera capa tiene solo 1 nodo e implementa la función de activación Sigmoide.

La función de activación rectificadora lineal, transforma los valores introducidos, todos los valores negativos los transforma a cero y los valores positivos los acepta sin modificarlos, la fórmula de esta función es:

$$f(x) = \max(0, x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \quad (6)$$

A continuación, se muestra una gráfica de la función de activación rectificadora lineal (ReLu):

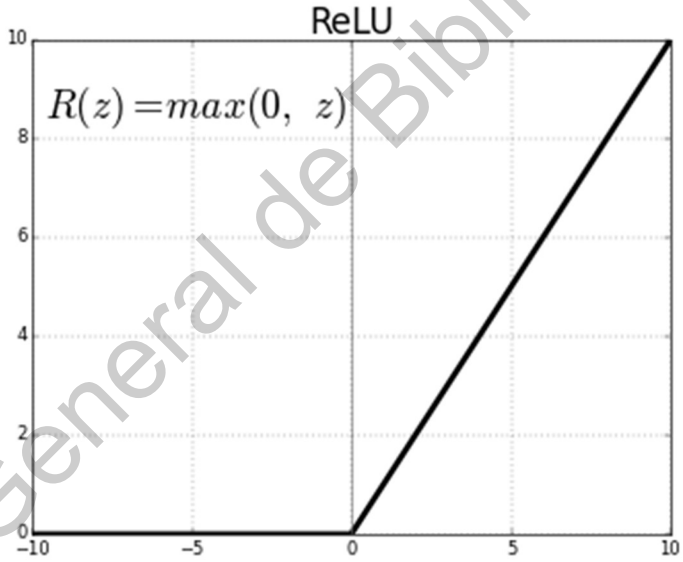

Figura 26: Gráfica de la función ReLu. Fuente: (Towardsdatascience, 2017).

Como se puede ver en la gráfica anterior, la función ReLu los valores negativos los transforma a 0 y los valores positivos no sufren ningún cambio.

La función de activación Sigmoide transforma los valores introducidos a una escala (0, 1), los valores altos tienden de manera asintótica a 1 y los valores muy bajos tienden de manera asintótica a 0, la fórmula de esta función es:

$$f(x) = \frac{1}{1 - e^{-x}} \tag{7}$$

Esta función tiene un buen rendimiento en la última capa o capa de salida, debido a que acota los valores entre 1 y 0 es perfecta para una clasificación binaria.

A continuación, se muestra una gráfica de la función de activación Sigmoide (sigmoid):

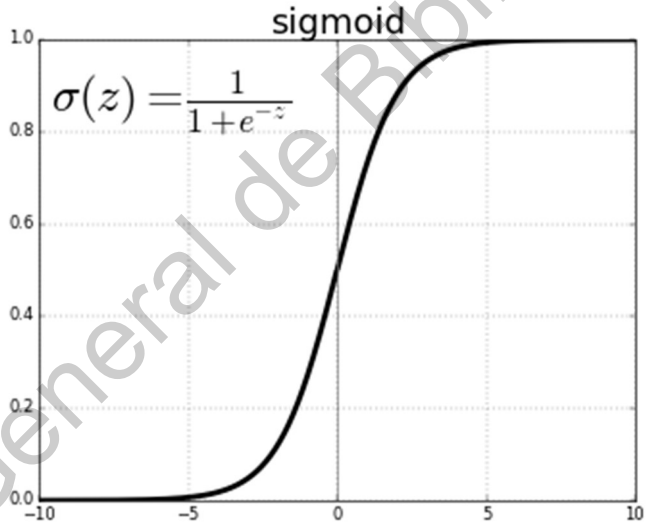

Figura 27: Gráfica de la función de activación Sigmoide. Fuente: (Towardsdatascience, 2017).

Como se puede ver en la gráfica anterior, la función Sigmoide acota todos los valores a 1 y 0.

El optimizador Adam es una extensión del descenso de gradiente estocástico. Este optimizador produce una actualización iterativa de los pesos de la red con los datos de entrenamiento. El nombre del algoritmo se deriva de: *adaptative moment estimation.*

Este algoritmo fue presentado por Diederik Kingma y Jimmy Ba en un artículo titulado "Adam: A Method for Stochastic Optimization" publicado en 2015 (Kingma and Ba, 2015).

Este algoritmo combina las ventajas de dos extensiones del descenso de gradiente estocástico, las cuales son:

- Algoritmo de gradiente adaptativo: Este mantiene una tasa de aprendizaje por parámetro que mejora el rendimiento en problemas con gradientes dispersos (por ejemplo, problemas de lenguaje natural y visión computarizada).
- Propagación del valor cuadrático medio: Este algoritmo mantiene tasas de aprendizaje por parámetro que se adaptan en función de la media de las magnitudes de los gradientes para el peso. La tasa de aprendizaje se adapta a cada uno de los parámetros. La idea es dividir la tasa de aprendizaje de un peso por un promedio de ejecución de las magnitudes de los gradientes recientes para ese peso.

Para evaluar la implementación de la red neuronal se usó una función de pérdida. Debido a que la clasificación es binaria, la función de pérdida usada es la de entropía cruzada binaria, la cual se calcula con la fórmula 8:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \quad (8)$$

El software usado para la implementación de la red neuronal fueron las siguientes librerías usando Python:

- TensorFlow (1.13.1), (TensorFlow, 2019).
- Keras, (2.2.4-tf), (Keras, 2019).

5.5.2 Máquina de Soporte de Vectores

Se implementó una máquina de soporte de vectores (SVM) para la clasificación de las conversaciones. La SVM realiza una clasificación binaria ya que se quieren clasificar las conversaciones en: conversaciones de pacientes con demencia y sin demencia.

Las máquinas de soporte de vectores (SVM) son algoritmos de aprendizaje supervisado, estas se pueden usar para regresión o clasificación. Esta técnica fue inventada por Vapnik y Chervonenkis en 1964 (Vapnik, 1964).

Las SVM tienen como objetivo encontrar una separación entre dos clases de objetos, entre más grande sea la separación más confiable será la clasificación. En el caso más simple, el caso lineal y separable, el algoritmo encuentra un hiperplano que separa el conjunto de observaciones en dos clases distintas de tal manera que maximiza la distancia entre el hiperplano y la observación más cercana del conjunto de entrenamiento.

Los vectores de soporte son un pequeño subconjunto de observaciones de entrenamiento que se utilizan como soporte para la ubicación óptima del hiperplano. Este pequeño subconjunto de observaciones son los puntos que definen el margen máximo de separación del hiperplano. Estos puntos tienen tantos elementos como dimensiones tenga el espacio de entrada, en pocas palabras son puntos multidimensionales que se representan con un vector de *n* dimensiones.

Para entrenar una SVM básicamente se requiere de dos pasos:

- Transformación de predictores: Los predictores son los datos de entrada, estos se transforman a un espacio de características altamente dimensional.
- Resolución de optimización cuadrática: Resolver el problema de optimización cuadrática que se ajuste al hiperplano óptimo para clasificar las características transformadas previamente en dos clases.

Para poder usar las conversaciones como datos de entrada para la SVM, estas fueron codificadas siguiendo los siguientes pasos:

- Tokenización: Dividir un diálogo en oraciones y cada oración en palabras.
- Eliminación de palabras vacías: Eliminar las palabras sin significado como artículos, pronombres, preposiciones, etc.
- Partes de la oración: Cada oración fue dividida en partes de acuerdo a la clasificación de cada parte en sujeto, adjetivo, verbo y adverbio. Cada parte fue etiquetada de acuerdo a las siguientes reglas:
    - Sujetos: 'N'
    - Adjetivos: 'J'
    - Verbos: 'V'
    - Adverbios: 'R'

A continuación, se muestra una oración codificada para poder ser usada como entrada en la SVM:

```
In[17]  : codif_text_to_svm('the brown dog is running fast')
Out[17] : ['J', 'N', 'V', 'R']
```

Figura 28: Ejemplo de codificación de una frase para ser usada en la SVM

Las conversaciones codificadas son los predictores, los cuales fueron transformados a vectores para la búsqueda del hiperplano óptimo.

- El software usado para la implementación de la SVM fue: Scikit-Learn, (0.21.3), (Scikit-Learn, 2019).

6. RESULTADOS

6.1 Análisis estadístico

Se realizó un análisis estadístico en las conversaciones de las siguientes métricas lingüísticas:

- Palabras de acuerdo a su tipo.
- Pausas cortas, pausas largas y muestras de aturdimiento o confusión.
- Palabras inusuales.
- Diversidad Léxica.
- Palabras de acuerdo a su longitud.
- Interjecciones.

A continuación, se muestran los resultados del análisis estadístico de las conversaciones de 20 personas, 10 de ellas padecen algún tipo de demencia y las otras 10 no padecen algún tipo de demencia.

En la tabla 8 se muestran los resultados de la diversidad léxica, palabras con longitud menor o igual a 4, palabras con longitud mayor o igual a 5 e interjecciones usadas en las conversaciones. La etiqueta “CD” hace referencia a una conversación de una persona que padece demencia y la etiqueta “SD” hace referencia a una conversación de una persona que no padece demencia.

Tabla 8: Diversidad léxica, palabras con longitud menor o igual a 4, palabras con longitud mayor o igual a 5 e interjecciones.

Sujeto	Diversidad Léxica	Palabras l <=4 (%)	Palabras l >=5 (%)	Interjecciones (%)
CD	2.03	64.14	31.81	4.05
CD	2.08	62.99	34.24	2.77
CD	1.79	68.95	28.15	2.9
CD	1.99	67.54	29.97	2.52
CD	2.47	74.38	22.92	2.7
CD	2.31	69.11	28.77	2.12
CD	2.04	71.32	26.67	2.01
CD	2.21	65.87	31.10	3.03
CD	2.48	62.43	34.83	2.74
CD	1.85	59.42	35.61	4.97
Promedio	2.12	66.61	33.39	2.98
SD	2.51	66.54	32.29	1.17
SD	1.87	70.27	27.04	2.69
SD	2.37	67.31	30.29	2.4
SD	2.28	69.81	27.48	2.71
SD	2.53	68.42	29.31	2.27
SD	1.98	65.41	32.91	1.68
SD	2.51	72.32	26.64	1.04
SD	2.34	71.29	25.62	3.09
SD	2.41	69.31	29.81	0.88
SD	2.07	70.51	28.86	0.63
Promedio	2.28	69.11	30.89	1.85

Los porcentajes de las métricas indican que tan usadas fueron en las conversaciones. En cada grupo de conversaciones se muestra un renglón con el promedio de cada métrica.

Los promedios de la diversidad léxica apenas muestran diferencia entre los grupos, de acuerdo a estudios previos (Guinn, 2014), la diversidad léxica no es una característica en las personas que padecen algún tipo de demencia.

Los promedios de palabras usadas de acuerdo a su longitud no muestran diferencia significativa.

En el caso de las interjecciones los promedios si muestran una diferencia significativa, de acuerdo a estudios previos, el uso de interjecciones de personas que padecen demencia es mucho mayor que en personas que no la padecen. La fluencia para hablar, el manejo de un vocabulario basto, la memorización de palabras no son buenas cualidades en las personas que padecen demencia esto se refleja en un uso mayor de interjecciones (Janeth, 2006).

En la tabla 9 se muestran los resultados de aturdimiento o confusión, pausas cortas, pausas largas y palabras inusuales usadas en las conversaciones.

Tabla 9: Aturdimiento o confusión, pausas cortas, pausas largas y palabras inusuales

Sujeto	Aturdimiento/ Confusión	Pausas Cortas (%)	Pausas Largas (%)	Palabras Inusuales (%)
CD	2.78	3.18	1.11	6.01
CD	1.42	1.17	0.76	8.14
CD	1.2	2.87	1.05	8.05
CD	0.98	2.34	0.97	7.3
CD	1.13	1.83	0.81	9.76
CD	0.87	2.94	1.21	8.47
CD	2.3	4.12	2.03	10.51
CD	1.92	3.35	1.52	10.26
CD	0.76	2.1	0.84	9.11
CD	0.81	1.87	0.73	7.63
Promedio	1.41	2.57	1.1	8.52
SD	0	0.56	0	1.97
SD	0	0.75	0	6.19
SD	0	1.9	0	5.77
SD	0.87	0.46	0.3	4.21
SD	0.76	1.37	0.41	7.88
SD	0	1.13	0.51	2.04
SD	0	0.87	0.21	3.1
SD	0.3	1.26	0	2.91
SD	0	1.1	0	4.51
SD	0	1.35	0	3.77
Promedio	0.19	1.07	0.14	4.23

Los promedios de todas las métricas de la tabla anterior muestran una diferencia significativa, de acuerdo a estudios previos, la presencia de pausas cortas y pausas largas, las muestras de aturdimiento o confusión y el uso de palabras inusuales son características de las personas que padecen demencia (Janeth, 2006) debido a que su capacidad para formular y entender oraciones está deteriorada de acuerdo a la etapa de demencia.

Un análisis de las partes de la oración fue hecho usando técnicas del NLP y técnicas de estadística. En la figura 29 se muestran las principales partes de la oración las cuáles son: sujeto, adverbio, adjetivo, verbo, palabras vacías y otras. Éstas están representadas en una gráfica de cajas. Las gráficas en azul corresponden a las conversaciones de personas que padecen demencia y las gráficas en gris corresponden a las personas que no padecen demencia.

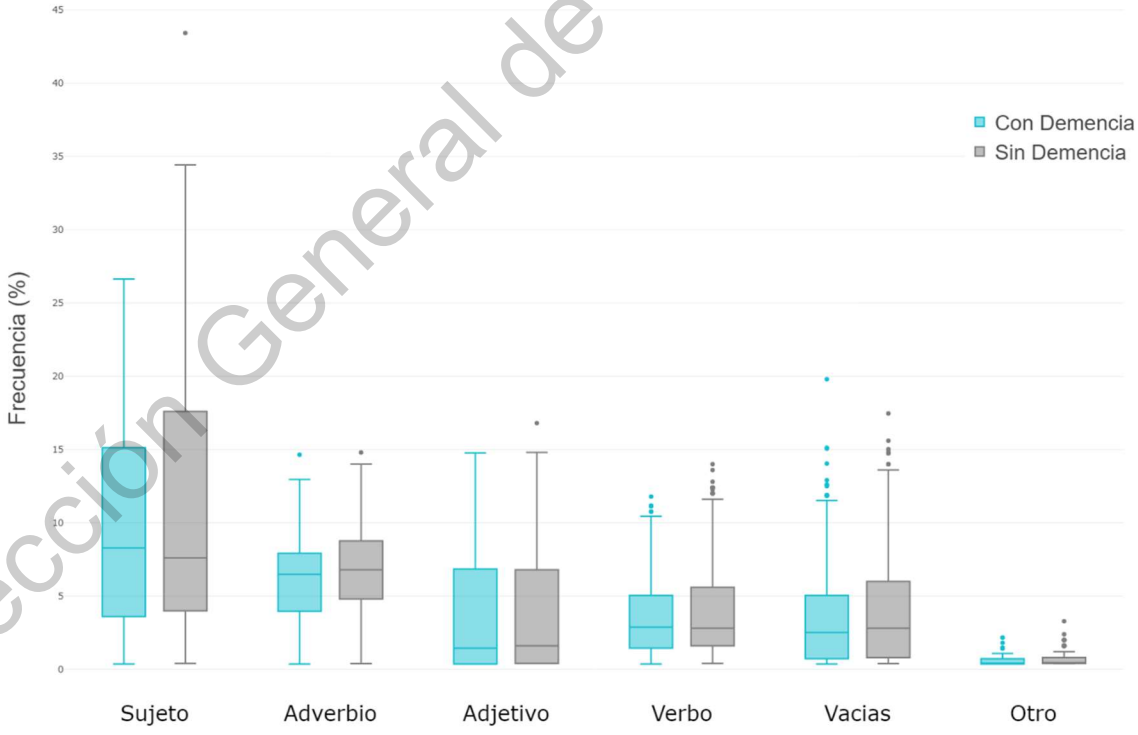

Figura 29: Gráfica de cajas de las partes de la oración en los dos grupos de conversaciones

En la de parte sujeto se pueden apreciar diferencias visuales, los cuartiles inferiores son casi iguales, en la gráfica gris se observa que el cuartil superior es relativamente mayor que el cuartil superior de la gráfica azul, pero las medianas son muy similares.

En la parte de adverbio se pueden apreciar diferencias visuales, el cuartil inferior de la gráfica gris es mayor que el de la gráfica azul, el cuartil inferior de la gráfica gris es mayor que el de la gráfica azul y las medianas son muy similares.

Para las partes de adjetivo y verbo, prácticamente no hay diferencia visual entre las gráficas de ambos grupos.

La parte de palabras vacías muestra diferencia solo en el cuartil superior, siendo el de la gráfica gris mayor.

No existe diferencia apreciable en otros.

A partir de la gráfica de caja no se pueden encontrar diferencias significativas entre las partes analizadas de cada grupo de conversaciones. De acuerdo a (Bayles, 2000) la riqueza del léxico y el uso de oraciones complejas no son características de la comunicación de personas que padecen demencia. Con el análisis estadístico de las partes de la oración no se pudo encontrar diferencias significativas de las características antes mencionadas entre personas que padecen demencia y personas que no.

6.2 Red Neuronal

Se implementó una red neuronal de 3 capas. La primera y la segunda capa tienen 16 nodos e implementan una función de activación rectificadora lineal, la tercera capa tiene solo 1 nodo e implementa la función de activación Sigmoide. La representación de la red neuronal implementada se muestra en la siguiente figura:

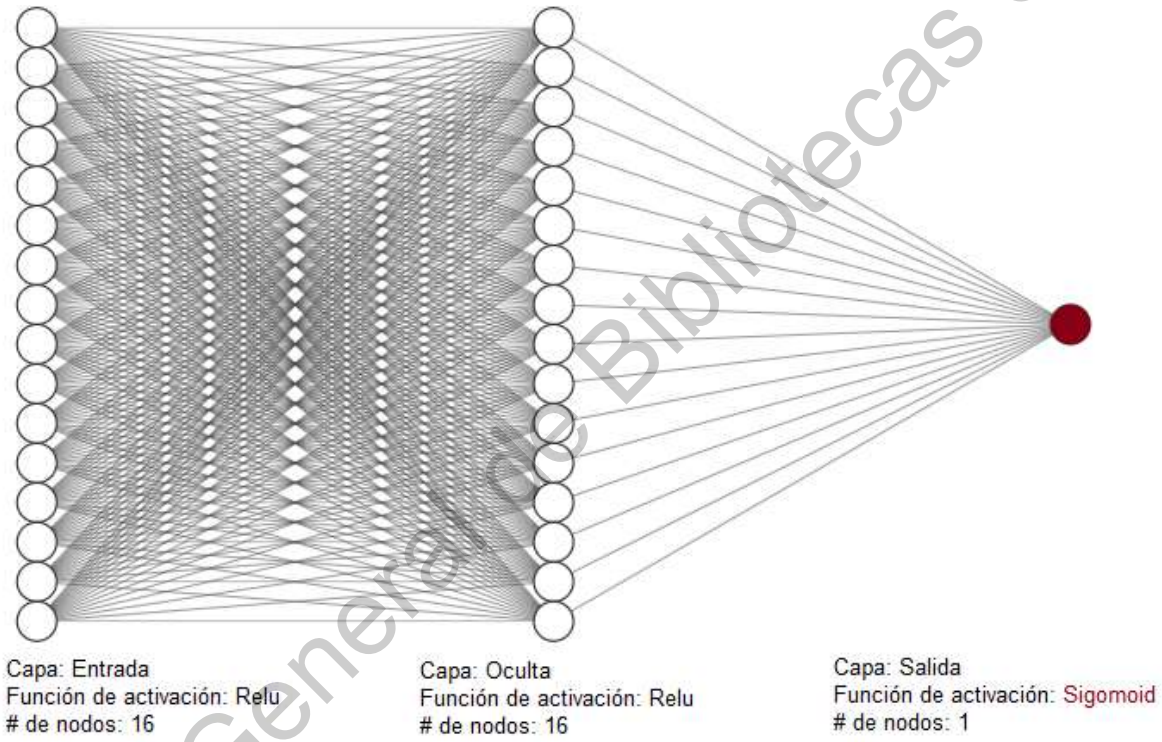

Figura 30: Red neuronal implementada

La red neuronal fue entrenada en 25 épocas. La red neuronal sobre aprende a partir de la época 25.

Para entrenar la red neuronal se usó el 70 % de los datos y el 30 % restante fue usado para validar. La relación de datos para entrenamiento y validación fue variada a través de las pruebas, debido a que la cantidad de datos que se tienen son muy pocos, aunado a la cantidad de datos, la cantidad de datos de los dos grupos diferentes (con y sin demencia) no es equitativa. El porcentaje de entrenamiento tiene que ser mucho mayor al de validación para procurar que la red neuronal logre

extraer las características de cada grupo para realizar una buena clasificación. En la figura 31 se muestra la exactitud del entrenamiento y la exactitud de la validación a través de las épocas:

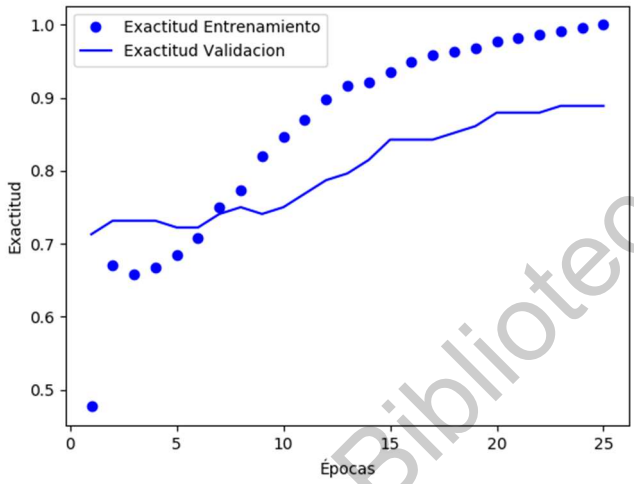

Figura 31: Exactitud de entrenamiento y validación

Tanto la exactitud del entrenamiento como el de la validación mejoran en cada época. La red neuronal obtuvo un 78.01 % de exactitud en la clasificación de las conversaciones. También obtuvo una entropía cruzada binaria de 0.52 tal como se muestra en la figura 32:

```
 - 0s - loss: 0.0859 - acc: 1.0000 - binary_crossentropy: 0.0859 - val_loss: 0.3150 - val_acc: 0.8889 -
val_binary_crossentropy: 0.3150
 32/141 [=====>........................] - ETA: 0s - loss: 0.5018 - acc: 0.7812 - binary_crossentropy:
0.5018
141/141 [==============================]        - 0s 285us/sample - loss: 0.5276 - acc: 0.7801 -
binary_crossentropy: 0.5276
Resultados:
Perdida, Exactitud, Binary Cross Entropy
[0.5275806421083762, 0.78014183, 0.5275806]
```

Figura 32: Resultados obtenidos por la red neuronal

6.2 Máquina de Soporte de Vectores

El entrenamiento de la máquina de soporte de vectores se realizó con el 70% de los datos al igual que la red neuronal debido a las características de los datos antes mencionadas.

La exactitud lograda con la máquina de soporte de vectores implementada para clasificar las conversaciones fue del 86.42 %.

```
=== RESTART: C:/Users/damia/Desktop/NLP2/SVM Text Classif/SVM text classif.py ==
SVM Accuracy Score -> 86.42
>>>
```

Figura 33: Resultados obtenidos por la máquina de soporte de vectores.

7. CONCLUSIONES

De acuerdo a las métricas de la diversidad léxica y las palabras usadas de acuerdo a su longitud no se pudo encontrar una diferencia significativa o un patrón determinante para clasificar las conversaciones. Los estudios previos sugieren que estas dos métricas tienen un índice bajo en personas que padecen algún tipo de demencia. El análisis estadístico realizado no toma en cuenta el grado de estudio de la persona, ni la etapa de la demencia, y tampoco la localidad donde la persona ha aprendido la lengua, todas estas características influyen en estas dos métricas y al no ser tomadas en cuenta no se pueden tomar en cuenta los resultados estadísticos de estas.

Las interjecciones son usadas con mayor frecuencia en las personas que padecen demencia tal como lo se pudo observar en el análisis estadístico debido a que algunas características de las personas con demencia son un vocabulario pobre y poca fluencia para hablar.

Los resultados de pausas cortas, pausas largas y las muestras de aturdimiento o confusión demuestran una diferencia significativa, en las personas que padecen demencia el uso de interjecciones es mayor que en las personas que no la padecen. Estos resultados eran esperados debido a que la capacidad para formular y entender oraciones está deteriorada en las personas que padecen demencia.

A partir del análisis estadístico de las partes de la oración de las conversaciones no se pudo encontrar diferencias significativas, la formulación de oraciones complejas y la riqueza del vocabulario no son cualidades de las personas que padecen demencia, pero estas cualidades dependen de otros factores ya mencionados como, el grado de estudios, etapa de la demencia, etc.

El grado de estudio de la persona, la etapa de la demencia, y la localidad donde la persona ha aprendido la lengua son factores que influyen directamente en las capacidades de comunicación de las personas que padecen demencia, estos factores podrían mejorar la clasificación o mostrar alguna diferencia estadística significativa, lamentablemente no están representados en la base de datos.

Los clasificadores automáticos obtuvieron índices altos en la clasificación, la red neuronal obtuvo el 78.01 % y la máquina de soporte de vectores obtuvo 86.42 % ambos clasificadores tienen un buen desempeño en la clasificación de texto, la exactitud alcanzada fue alta (considerando la naturaleza de la tarea). La exactitud de estos clasificadores se podría mejorar agregando más datos a la base de datos, ya que la relación de conversaciones de personas que padecen demencia es de 1 a 3 en relación a las conversaciones de personas que no la padecen.

8. REFERENCIAS

R. Bucks, S. Singh, J. Cuerden, and G. Wilcock, “Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance”, 2000.

Curry Guinn and Ben Singer, A Comparison of Syntax, Semantics, and Pragmatics in Spoken Language among Residents with Alzheimer’s Disease in Managed-Care Facilities., 2014.

S. L. Andresen, "John McCarthy: father of AI," in *IEEE Intelligent Systems*, vol. 17, no. 5, pp. 84-85, Sept.-Oct. 2002.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. 2010.

Zettlemoyer, Luke S. and Collins, Michael. Learning to map sentences to logical form: structured classification with probabilistic categorial grammars. 2005.

Alzheimer’s Association. 2014 Alzheimer’s Disease Facts and Figures. Downloaded on 07/05/14 from www.alz.org, 2014.

Jaramillo, Janeth & Márquez, Andrea & Rodríguez Riaño, Leidy Johanna. Demencia tipo Alzheimer y lenguaje. 2006.

Khodabakhsh, Ali & Kusxuoglu, Serhan & Demiroglu, Cenk. Natural language features for detection of Alzheimer's disease in conversational speech. 2014.

Kingma, Diederik & Ba, Jimmy. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations. 2014.APNIK, V. N., and A. Ya. CHERVONENKIS, Theory of pattern recognition: Statistical problems of learning]. 1974.

C. Pope and B. Davis, “Finding a balance: The Carolinas Conversation Collection”, Corpus Linguistics and Linguistic Theory 7(1), 143—161, 2011.

B. Davis, B. and M. Maclagan,”Examining pauses in Alzheimer's discourse”, American journal of Alzheimer's Disease and other dementias 24(2), 141—154, 2009.

M. Snover, B. Dorr, B. and R. Schwartz, “A lexically-driven algorithm for disfluency detection”, in Proceedings of HLT-NAACL 2004: Short Papers, pp. 157—160, 2004.

H. Bortfeld, S. Leon, S.; J. Bloom, M. Schober, and S. Brennan, “Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender”, Language and Speech 44(2), 123-147, 2001.

S. Singh, Neural Networks In Spontaneous Speech Assessment Of Dysphasic Patients, Ph.D. Dissertation, University of the West of England, Bristol, UK, 1996.

S. Singh, S. and T. Bookless, “Analysing spontaneous speech in dysphasic adults”, International Journal of Applied Linguistics 7(2), 165—181, 1997.

P. Ten Have, Doing conversation analysis, Sage Publications Ltd., 1999.

C. Guinn and A. Habash, “Language Analysis of Speakers with Dementia of the Alzheimer's Type”, AAAI Fall Symposium: Artificial Intelligence for Gerontechnology, 2012.

A. Habash, C. Guinn, D. Kline, and L. Patterson. “Language Analysis of Speakers with Dementia of the Alzheimer's Type”. Annals of the Master of Science in Computer Science and Information Systems at UNC Wilmington, 6(1) paper 11, 2012.

S. Bird, E. Klein, E. and E. Loper, E., Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'Reilly, Beijing, 2009.

Lai Yi-Hsiu, Language Processing of Seniors with Alzheimer’s Disease: From the Perspective of Temporal Parameters., 2017.

B. Roark, M. Mitchell, J. Hosom, K. Hollingshead and J. Kaye, Spoken Language Derived Measures for Detecting Mild Cognitive Impairment., 2011.

Calvin Thomas, Vlado Keselj, Nick Cercone and Kenneth Rockwood, Automatic Detection and Rating of Dementia of Alzheimer Type through Lexical Analysis of Spontaneous Speech., 2005.

Clare L., Wilson B., Carter G., Breen K., Gosses A., Hodges J., Intervening with everyday memory problems in dementia of the Alzheimer type: An errorless learning approach. Journal of Clinical and Experimental, 2000.

Cuetos F, Rodríguez J, Menéndez M. Semantic markers in the diagnosis of neurodegenerative dementias, 2009.

Bélanger S, Belleville S. Semantic inhibition impairment in mild cognitive impairment: a distinctive feature of upcoming cognitive decline, 2009.

S. Ray, M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L. Friedman, D. Galasko, M. Jutel, A. Karydas, J. Kaye, J. Leszek, B. Miller, L. Minthon, J. Quinn, G. Rabinovici, W. Robinson, M. Sabbagh, Y. So, D. Sparks, M. Tabaton, J. Tinklenberg, J. Yesavage, R. Tibshirani, R. and T Wyss-Coray, “Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins”, 2007.

Banda, H., Inteligencia Artificial: Principios y Aplicaciones. 2016.

9. ANEXOS

9.1 Certificación del Collaborative Institutional Training Initiative

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)
COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** damian solis (ID: 7326085)
- **Institution Affiliation:** University of North Carolina at Charlotte (ID: 1418)
- **Institution Email:** dsolis06@alumnos.uaq.mx
- **Institution Unit:** Engineering
- **Phone:** +5214423248862

- **Curriculum Group:** IRB Members - Basic/Refresher
- **Course Learner Group:** IRB Members (IRB) - Basic/Refresher
- **Stage:** Stage 2 - Refresher Course

- **Record ID:** 29374639
- **Completion Date:** 02-Dec-2018
- **Expiration Date:** 01-Dec-2022
- **Minimum Passing:** 80
- **Reported Score*:** 96

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Biomed Refresher 1 – History and Ethical Principles (ID: 975)	14-Nov-2018	2/2 (100%)
Biomed Refresher 1 – Regulations and Process (ID: 981)	22-Nov-2018	3/3 (100%)
Biomed Refresher 1 – Informed Consent (ID: 980)	28-Nov-2018	2/2 (100%)
Biomed Refresher 1 – SBR Methodologies in Biomedical Research (ID: 982)	28-Nov-2018	3/3 (100%)
Biomed Refresher 1 – Records-Based Research (ID: 983)	30-Nov-2018	1/2 (50%)
Biomed Refresher 1 – Genetics Research (ID: 984)	02-Dec-2018	4/4 (100%)
Biomed Refresher 1 - Populations in Research Requiring Additional Considerations and/or Protections (ID: 985)	02-Dec-2018	2/2 (100%)
Biomed Refresher 1 – Research Involving Prisoners (ID: 973)	02-Dec-2018	2/2 (100%)
Biomed Refresher 1 – Research Involving Children (ID: 974)	02-Dec-2018	4/4 (100%)
Biomed Refresher 1 – Research Involving Pregnant Women, Fetuses, and Neonates (ID: 986)	02-Dec-2018	3/3 (100%)
Biomed Refresher 1 – FDA-Regulated Research (ID: 987)	02-Dec-2018	2/3 (67%)
Biomed Refresher 1 - Research and HIPAA Privacy Protections (ID: 17261)	02-Dec-2018	4/4 (100%)
SBE Refresher 1 – History and Ethical Principles (ID: 936)	02-Dec-2018	2/2 (100%)
SBE Refresher 1 – Federal Regulations for Protecting Research Subjects (ID: 937)	02-Dec-2018	2/2 (100%)
SBE Refresher 1 – Defining Research with Human Subjects (ID: 15029)	02-Dec-2018	2/2 (100%)
SBE Refresher 1 – Informed Consent (ID: 938)	02-Dec-2018	2/2 (100%)
SBE Refresher 1 – Assessing Risk (ID: 15034)	02-Dec-2018	2/2 (100%)
SBE Refresher 1 – Privacy and Confidentiality (ID: 15035)	02-Dec-2018	4/4 (100%)
SBE Refresher 1 – Research with Prisoners (ID: 939)	02-Dec-2018	2/2 (100%)
SBE Refresher 1 – Research with Children (ID: 15036)	02-Dec-2018	2/2 (100%)
SBE Refresher 1 – Research in Educational Settings (ID: 940)	02-Dec-2018	2/2 (100%)
SBE Refresher 1 – International Research (ID: 15028)	02-Dec-2018	2/2 (100%)
Biomed Refresher 1 - Instructions (ID: 960)	02-Dec-2018	No Quiz
SBE Refresher 1 - Instructions (ID: 943)	02-Dec-2018	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?kd4217fee-af88-485b-af35-41339212d406-29374639

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

9.2 Aprobación para el uso de Carolinas Conversation Collection

UNC CHARLOTTE

The University of North Carolina at Charlotte
9201 University City Boulevard
Charlotte, NC 28223-0001

December 7, 2018

To: C. Damián Solis Rosas and Comité de Ética

Fr: Boyd H. Davis, PhD

This letter confirms that you have submitted all required documents and statements of ethics in order to use data from the Carolinas Conversation Collection, hosted at Medical University of South Carolina (MUSC), originally sponsored by MUSC and University of North Carolina-Charlotte for funding from the National Library of Medicine.

This data is for English-speakers only.

Kind regards,

Boyd Davis

Boyd H. Davis, PhD. | Bonnie E. Cone Professor of Teaching
Professor, Applied Linguistics/English | Affiliate: Gerontology;
Communications, Cognitive Science, Health Services Research
Fretwell: UNC Charlotte | 9201 University City Blvd | Charlotte NC 28223
Adjunct Research Professor, Nursing, Medical U of South Carolina
Researcher, COIN/HEROIC Ralph H Johnson VA, Charleston SC
http://english.uncc.edu/people/davis-boyd-h-phd
My NCBI:
https://www.ncbi.nlm.nih.gov/sites/myncbi/boyd.davis.1/bibliography/48091808/public/?sort=date&direction=descending
bdavis@uncc.edu | Skype boydhilldavis | ph + 866 966 391 420

2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE) Mexico City, Mexico. September 11-13, 2019

Search for Dementia Patterns in Transcribed Conversations using Natural Language Processing

Damián Solís Rosas
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México
damian.solis@hotmail.com

Saúl Tovar Arriaga
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México
saul.tovar@uaq.mx

Marco Antonio Aceves Fernández
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México
marco.aceves@gmail.com

***Abstract*—The effects on the linguistic capacity of the people with some type of dementia are reflected in the lexicon (his mental dictionaries and his ability to understand complex words) rather than his ability to formulate complete and fluent enunciations. Analysis indicate that the richness of the lexicon and the fluency to speak are not good qualities in people who suffer Alzheimer. There are previous studies of the pathologies of speech, which include the use of pauses, words of filling, words formulated, restarts, repetitions, incomplete statements and diffluent speech. All previous factors may occur in individuals with some type of dementia. Through discriminatory analysis of conversation and metrics analysis, we found slight statistical differences between people with and without dementia. Additionally, we use two machine learning algorithms to automatically classify presence/absence of dementia. The first one, a 3-layer neural network reaching a binary classification accuracy of 78.3%, and the second a support vector machine reaching a binary classification accuracy of 86.42%.**

***Index Terms*—Natural Language Processing, Dementia Patterns, Text Classification, Support Vector Machine, Neural Networks**

I. Introduction

Dementia is a general term for a decline in mental ability, describes a set of symptoms that may include memory loss and difficulties with thinking, problem-solving or language. A person with dementia may also experience changes in their mood or behaviour to such an extent that it interferes with a person's daily life and activities.

Most patients with some type of dementia are characterized by the degradation of their language and cognitive functions resulting in significant communication difficulties.

The detection of dementia can be an expensive and exhausting process for the person to be diagnosed [1].

Based in how far a person's dementia has progressed, dementia can be divided in stages. Defining a person's disease stage helps physicians determine the best treatment approach and aids communication between health providers and caregivers. The stages of Dementia can be grouped in Early-Stage (Moderate Cognitive Decline), Mid-Stage (Severe Cognitive Decline) and Late-Stage (Very Severe Cognitive Decline).

Each Stage has signs and symptoms. In the Early-Stage people can quickly forget what they have heard, seen or thought, it is common for them to lose the topic and repeat an idea many times, consequently, it is difficult to follow a conversation.

In the Mid-Stage people are often disoriented in time and space, usually have large memory deficits, so they can not remember recent events, these characteristics generate a poor verbal production, meaningless and simple.

In the Late-Stage people have serious problems to pay attention, to codify, retrieve information, have perception problems and their executive functions are limited or lack them. Semantic memory and the ability to remember concepts are very deteriorated or seem absent in the person.

In all the stages of dementia, the characteristics and linguistic capacities are deteriorated, these characteristics can serve as object of study and analysis to get an early diagnosis.

It is possible to find markers that give signs of early dementia by analyzing and processing the natural language obtained from patients through medical tests, questionnaires or simple conversations [2].For the analysis and search of markers in conversations, it is necessary to know the characteristics of communication in people who suffer dementia. We can found these characteristics from the Early-Stage.

The communication characteristics of people with dementia vary in each stage. As dementia progresses, communication problems increase, causing a deterioration in people's communication skills. The deterioration in communication impacts in daily life of patients and deteriorates one of the main tools and qualities of the human being: communication. Losing the ability to communicate can be one of the most frustrating and difficult problems for people with dementia, their families and carers. They find it more and more difficult to express themselves clearly and to understand what others say.

The main characteristics of the communication of people according to the stage of dementia are [3]:

- Early-Stage: Difficulty to understand sentences with complex content. Conserved syntactic structure and problems

978-1-7281-4840-3/19/$31.00 © 2019 IEEE

to repeat long sentences. Phonological system conserved. Incomplete Ideas.

- Mid-Stage: Difficulty to nominate and categorize. Reduced expressive vocabulary. Difficulty to repeat simple sentences. Omission of connectors and functional words in the sentences. Incomplete Phrases.
- Late-Stage: Reduction in Vocabulary. Frequent omission of functional words. Uncontrolled repetition of phrases said by the speaker. Uncontrolled repetition of the same word.

The conversations can be classified applying an statistical analysis and computational techniques such as support vector machine, decision tree and Bayes classifier[1][2][3][4].

In the next table are shown the previous studies with their authors, objective and results.

TABLE I
STATE OF THE ART

Author	Data Format	Technique	Accuracy
C. Guinn Ben Singer A. Habash [1]	Transcribed Conversations	Decision Tree Bayesian	67% 80%
A. Khodabakhsh S. Kuscuoglu C. Demiroglu [2]	Transcribed Conversations	Decision Tree SVM	90% 80%
B. Roark M. Mitchell J. Hosom K. Hollingshead [3]	Audio	SVM	86%
C. Thomas M. Mitchell Vlado Keselj Kenneth Rockwood [4]	Transcribed Conversations	Bayesian	90%

A neural network can be used to classify text, the accuracy obtained by the neural network can be compared with other classification methods to evaluate its performance.

II. Materials and Methods

A. Statistical Analysis

To process and analyze the text, the NLP (Natural Language Processing) use techniques as Lemmatization, Morphological segmentation, Word segmentation, Stemming, etc.

Based on the lexical and grammatical characteristics of people who have some type of dementia and those who do not have it and with the help of computational techniques, a computer system can be developed to find markers in the conversations of these people. For the analysis of the conversations and the application of techniques, algorithms were implemented with the help of the NLTK (Natural Language Toolkit) [4] library.

In this study, to get markers in the conversations, conversations of people who have some type of dementia and people who do not have it were used. These conversations are in the English language and are part of the Carolina Corpus Conversation database[5].

The Carolina Corpus Conversation database has hundreds of conversations of people with some pathological condition, the conversations are transcribed and include features such as: Comments from the recorder; Sounds recorded during conversations; Feelings reflected by the people involved in the conversations. Pauses, reflected feelings and signs of bewilderment or confusion are represented in conversations by different signs or chains of signs.

The techniques that were used to search characteristics in the conversations were:

- Text tokenization.

```
In[0]  : nltk.word_tokenize(text)
Out[0] : ['doing', 'what', 'i', 'can', 'not', 'read', 'it', 'i', 'see']
```

- Part of Speech

```
In[1]  : nltk.pos_tag(text_tokens)
Out[1] : ['doing', 'Verb'), ('what', 'Pronoun'), ('i', 'Noun')]
```

- Word Count.

```
In[2]  : len(text)
Out[2] : 2885
In[3]  : len(text_tokens)
Out[3] : 687
```

- Count pauses, long pauses and stuns or confusions recorded in the conversation.

```
In[4]  : len(pauses(text))
Out[4] : 2885
In[5]  : len(long_pauses(text))
Out[5] : 687
In[6]  : len(sconf_stuns(text))
Out[6] : 128
```

- Word count of length less than or equal to 4

```
In[7]  : len(words_len_less(text, 4))
Out[7] : 43
In[8]  : words_len_less(text, 4)
Out[8] : [and, 'what', 'i', 'can', 'not', 'read', 'it']
```

- Word count of length greater than or equal to 5

```
In[9]  : len(words_len_greater(text, 5))
Out[9] : 31
In[10] : (words_len_greater(text, 5)
Out[10]: ['children', 'walked', 'forgot', 'soldiers', 'joined']
```

- Extraction of the most used words of length less than or equal to 4 and frequency

```
In[11]  : most_freq_words_4(text)
Out[11] : [('and', 49), ('i'. 76), ('my', 23), ('the', 34)]
```

- Extraction of the most used words of length greater than or equal to 5 and frequency

```
In[12]  : most_freq_words_5(text)
Out[12] : [('children', 12), ('walked'. 11), ('forgot', 7)]
```

- Counting and extraction of unusual words

```
In[13]  : unusual_words(text)
Out[13] : ['evangilistic', 'soldiers', 'preached', 'belonged']
```

- Counting and extraction of interjections

```
In[14]  : interjections_words(text)
Out[14] : ['uh', 'ah', 'mmm', 'ammm']
```

TABLE II
CHARACTERISTICS OF COMMUNICATION ACCORDING TO THE STAGE OF DEMENTIA

	Semantic	**Sintactic**	**Phonological**	**Pragmatic**	**Literacy**
Early-Stage	Difficulty to understand sentences with complex content **N.A.**	Conserved syntactic structure and problems to repeat long sentences **Possible**.	Phonological system conserved **N.A.**	Ramble **Possible (A.I.)** Incomplete Ideas **Possible (A.I.)**	Disortography **N.A.**
Mid-Stage	Difficulty to nominate and categorize **Possible**. Reduced expressive vocabulary **Possible**.	Difficulty to repeat simple sentences **Possible**. Omission of connectors and functional words in the sentences **Possible**.	Occasional confusion of pronunciation patterns **N.A.**	Incomplete Phrases **Possible (A.I.)** Repetition of ideas in the conversation **Possible (A.I.)** Loss of topic and leave the conversation **Possible (A.I.)**	Writing of words and short phrases **N.A.**
Late-Stage	Reduction in Vocabulary **Possible**. Use only significant elements **Possible (A.I.)** Semantic paraphasias **Possible (A.I.)**	Limited automatic language **Possible (A.I.)** Almost zero repetition, even for monosyllabic words **Possible**. Frequent omission of functional words **Possible**.	Uncontrolled repetition of phrases said by the speaker **Possible**. Uncontrolled repetition of the same word **Possible**. Uncontrolled repetition of syllables **Possible**.	Conversation almost absent, limited or inaccurate **Possible**. Impossibility to maintain the topic **Possible (A.I.)**	Almost total impairment of writing **N.A.**

- Counting, extraction and analysis of Part of Speech

```
In[15]   : freq_pos(text)
Out[15] : [('Noun', 23), ('Verb', 11), ('Article', 7)]
```

To use the techniques mentioned previously, the conversations were preprocessed; punctuation marks were removed; the pauses and recorded stuns were coded; all text was converted to lowercase; The names of the people involved were eliminated. All the above was done for the counting and extraction of characteristic data that were later used to obtain statistical results and use artificial intelligence techniques.

Table II shows the communication characteristics in each stage of dementia. After an analysis of the database, taking into consideration the nature of the questions and answers, we labeled each of the characteristics found in each dementia stages. The first category is the N.A.(Not Available), which means that taking into consideration the available information in the database, to our knowledge, it is not possible to get the pattern. E.g., we can not find if a person have difficulties to understand sentences. Possible means that these characteristics can be analyzed using statistical techniques, and Possible A.I. means that this characteristics may be analyzed using Artificial Intelligence techniques.

The lexical diversity (LD) tells us how much vocabulary was deployed. This measure is important because the vocabulary of people with dementia in Mid-Stage and Late-Stage is very poor (Table II). The greater value of lexical diversity means that more different words were used in the conversation. The lexical diversity is obtained by dividing the number of words in the text Text Length (TL) by the number of words of the vocabulary used Vocabulary Length (VL). The formula is:

$$LD = TL/VL \tag{1}$$

The size of the words is a characteristic with which it can be inferred that a word has greater importance in the text, since the words of small length are usually articles, pronouns, conjunctions, etc. Persons with dementia tend to decrease using long words and replace them by simple ones easier to remember. Analyzing the frequency of long length words, we can infer what is the main topic or the topics that were discussed in the conversation.

Unusual words are those that are not present in informal language and even some of them are not in dictionaries. Some of them are scientific jargon and others are even invented by people. Persons with dementia tend to use these 'invented' words without noticing that they do not belong to the language. The percentage of unusual words in the vocabulary were counted and calculated. To obtain the percentage of unusual words (PPL), the number of unusual words (NPL) is divided by the number of words of the vocabulary used Vocabulary Length (VL) and multiplied by 100 %. The formula is:

$$PPL = (NPL/VL) * 100\% \tag{2}$$

The number of interjections, pauses and stun or confusion were counted. A high percentage of those words are a sign of Pragmatic issues (Table II).

An analysis of the words used was made based in the Part of Speech. The amount of articles, pronouns, conjunctions and adverbs used with respect to nouns and verbs give us a statistic of the complexity of the sentences formed in conversations [1].

B. Automatic Classifiers

An additional way to do an automatic classification of persons with and without dementia is using a neural network (NN). The application of the neural network in this case is to make a binary classification. To train the neural network, the conversations were encoded because we can not use raw text in the networks. Each conversation was divided in blocks of 256 words, a number was assigned to each word according to a dictionary. The data used in the neural networks are arrays of numbers with length of 256.
For example, the sentence *'The neural networks can be used to classify text'* is encoded:

```
In[16]  : word_to_id('the neural networks can be used to classify text')
Out[16] : [4, 73517, 8060, 70, 30, 343, 8, 12371, 3004]
```

The NN is a sequential model. The Sequential model is a linear stack of layers. This NN has 3 layers, the first and the second have 16 nodes and every node implements a rectified linear activation unit, the third has 1 node and this one implements a Sigmoid function.

The NN use the Adam optimizer, this is an optimization algorithm that can used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data. The loss function is measured with the Binary Cross-Entropy, wich is a loss function used on problems involving binary decisions.

The software used to make the automatic classifier is TensorFlow[6] combined with Keras[7].

To train the neural network the data was obtained from 62 conversations from people who have dementia and 160 from people who do not have dementia. Each conversation was tagged with 1 if the person have dementia and 0 if the person do not have dementia.

Another way to do an automatic classification of persons with and without dementia is using a support vector machine (support vector machine). Like the neuronal network the SVM is applied to make a binary classification. To apply the SVM the stop words were removed; the stem of the words were subtracted; the words obtained after the first steps were divided in four groups: nouns, adjectives, verbs and adverbs.

After processing the words, the conversations were transformed in vectors then the SVM were applied.

The software used to apply the SVM is Scikit-learn[8].

III. Results

The results showed below are from conversations of 20 persons, 10 of them have dementia and 10 do not have dementia. The lexical diversity, the percentage of words with length less than or equal to 4 and the percentage of words with a length greater than or equal to 5 are shown in the Table III. The persons who have dementia are represented with the tag "CD" and who do not have dementia with the tag "SD".

TABLE III
LEXICAL DIVERSITY, USED WORDS AND INTERJECTIONS IN THE CONVERSATIONS

Person	Lexical Diversity	Words l <=4 (%)	Words l >=5 (%)	Interjections (%)
CD	2.03	64.14	31.81	4.05
CD	2.08	62.99	34.24	2.77
CD	1.79	68.95	28.15	2.9
CD	1.99	67.54	29.97	2.52
CD	2.47	74.38	22.92	2.7
CD	2.31	69.11	28.77	2.12
CD	2.04	71.32	26.67	2.01
CD	2.21	65.87	31.10	3.03
CD	2.48	62.43	34.83	2.74
CD	1.85	59.42	35.61	4.97
Average	2.12	66.61	33.39	2.98
SD	2.51	66.54	32.29	1.17
SD	1.87	70.27	27.04	2.69
SD	2.37	67.31	30.29	2.4
SD	2.28	69.81	27.48	2.71
SD	2.53	68.42	29.31	2.27
SD	1.98	65.41	32.91	1.68
SD	2.51	72.32	26.64	1.04
SD	2.34	71.29	25.62	3.09
SD	2.41	69.31	29.81	0.88
SD	2.07	70.51	28.86	0.63
Average	2.28	69.11	30.89	1.85

The following table shows the percentage of daze or confusions, short pauses, long pauses and unusual words recorded in conversations.

TABLE IV
DAZE OR CONFUSION, SHORT PAUSES, LONG PAUSES AND UNUSUAL WORDS RECORDED IN THE CONVERSATIONS

Person	Daze/ Confusion	Short Pauses (%)	Long Pauses (%)	Unusual Words (%)
CD	2.78	3.18	1.11	6.01
CD	1.42	1.17	0.76	8.14
CD	1.2	2.87	1.05	8.05
CD	0.98	2.34	0.97	7.3
CD	1.13	1.83	0.81	9.76
CD	0.87	2.94	1.21	8.47
CD	2.3	4.12	2.03	10.51
CD	1.92	3.35	1.52	10.26
CD	0.76	2.1	0.84	9.11
CD	0.81	1.87	0.73	7.63
Average	1.41	2.57	1.1	8.52
SD	0	0.56	0	1.97
SD	0	0.75	0	6.19
SD	0	1.9	0	5.77
SD	0.87	0.46	0.3	4.21
SD	0.76	1.37	0.41	7.88
SD	0	1.13	0.51	2.04
SD	0	0.87	0.21	3.1
SD	0.3	1.26	0	2.91
SD	0	1.1	0	4.51
SD	0	1.35	0	3.77
Average	0.19	1.07	0.14	4.23

The percentages indicate how much the words were used throughout the conversation. The row of average shows the averages of each metric of the words according to the classification of the conversations.

An analysis of the Part of Speech was made using techniques of the Natural Language Toolkit and statistical techniques. In Figure 1 are shown the main types of word which are conjunctions, adjectives, determiners, and some others in a Box Chart. The blue chart represent the Part of Speech of people who have dementia and the gray one represent the Part of Speech of people who do not have dementia.

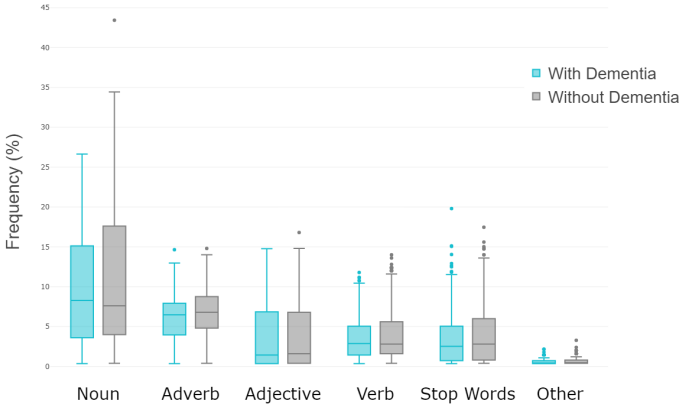

Fig. 1 - Types of words used in the conversations

There are a visual difference in the groups of noun, adverb, determiner, conjunction and preposition in the people who have dementia with the people who do not have it.

A 3-layer neural network was used, with 25 epochs, after the twenty-five epoch the neural network suffer an overfitting. To train the neural network 70% of the data was used and the remaining 30% was used to validate. The training and validation accuracy of the neural network are shown in the Figure 2.

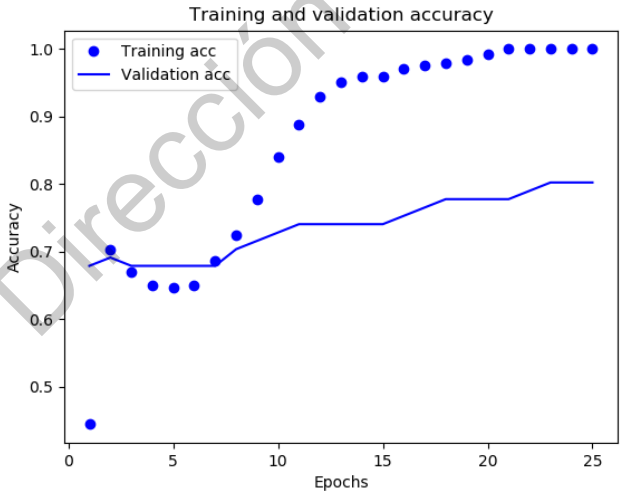

Fig. 2 - Training and validation accuracy of the NN

As we can see in Figure 2 the accuracy is improved in each epoch.

To apply the support vector machine classifier 70% of the data was used to training and the remaining 30% was used to validate.The accuracy of the Support Vector Machine to classify the conversations was 86.42%.

IV. Conclusion

According to our measures of lexical diversity, used words and interjections we can not find a significant or conclusive pattern to determine if a person has dementia. With statistical analysis we can not found a conclusive pattern in the conversations. There are a lot of factors that are not represented in the data such as: the stage of dementia and the education level [9]. According to previous studies the stage of dementia and the education level are directly related with the communication.

In contrast, our results in percentage of daze, pauses and unusual words show that it is useful to take this metric into consideration. The results indicate that these metrics are greater in the people who have dementia, because the capacities to understand, reason, learn and enunciate words are deteriorated in people with dementia according to the stage in which they are.

As we can see in the figure 1, the groups of noun, adverb and stop words have a visual difference, in people who have dementia are fewer used, as we said previously. The language in people who have dementia is simple, the sentences that they use have a lack of connectors, descriptors and nouns they want to say something in the easiest way.

The neural network used obtained a score of 78.3% in the classification of conversations of people with dementia and without dementia. Different networks were trained and the best score was obtained with a 3-layer neural network in 25 epochs. This accuracy may be improved using data of more subjects or adding data such as stage of dementia and the education level.

The support vector machined used obtained a score of 86.42% in the classification of conversations of people with dementia and without dementia. This automatic classifier have a better performance than the neural network.

Although we reached a good classification accuracy (according to the nature of the task), we believe that this classification depends a lot on patterns like pauses, daze, confusions and word type. We are not finding patterns related to other mental abilities like understanding of complex content or ability to nominate and categorize. These characteristics are important and should be taken into consideration if we want to improve our analysis.

Acknowledgment

We would like to express our very great appreciation to Dr. Davis Boyd for her valuable support to get the access to the Carolina Corpus Conversations.

REFERENCES

[1] Curry Guinn and Ben Singer, *A Comparison of Syntax, Semantics, and Pragmatics in Spoken Language among Residents with Alzheimer's Disease in Managed-Care Facilities.*, 2014.

[2] Ali Khodabakshsh and Cenk Demiroglu, *Natural language features for detection of Alzheimer's disease in conversational speech.*, 2014.

[3] B. Roark, M. Mitchell, J. Hosom, K. Hollingshead and J. Kaye, *Spoken Language Derived Measures for Detecting Mild Cognitive Impairment.*, 2011.

[4] Calvin Thomas, Vlado Keselj, Nick Cercone and Kenneth Rockwood, *Automatic Detection and Rating of Dementia of Alzheimer Type through Lexical Analysis of Spontaneous Speech.*, 2005.

[5] Carolinas Conversation Collection, *Carolinas Conversation Collection*, 2019. [Online]. Available:https://carolinaconversations.musc.edu/about/collection. [Accessed: March 11, 2018].

[6] TensorFlow, *TensorFlow*, 2019. [Online]. Available:https://www.tensorflow.org/. [Accessed March 6, 2019].

[7] Keras, *Keras*, 2019. [Online]. Available:https://keras.io/. [Accessed March 6, 2019.]

[8] Scikit-learn, *Scikit-learn*, 2019. [Online]. Available:https://scikit-learn.org/stable/. [Accessed July 15, 2019].

[9] Luis Miguel Gutiérrez-Robledo e Isabel Arrieta-Cruz, *Demencias en México: la necesidad de un Plan de Acción.*, 2015. [Online]. Available:http://www.medigraphic.com/pdfs/gaceta/gm-2015/gm155p.pdf. [Accessed: Feb 15, 2018].

[10] NLTK, *Natural Language Toolkit*, 2019. [Online]. Available:https://www.nltk.org/. [Accessed Feb. 3, 2018].

[11] NLTK, *Natural Language Toolkit*, 2019. [Online]. Available:https://www.nltk.org/. [Accessed Feb. 3, 2018].

[12] Janeth Hernández Jaramillo, *Demencias: los problemas de lenguaje como hallazgos tempranos.*, 2010.

[13] Lai Yi-Hsiu, *Language Processing of Seniors with Alzheimer's Disease: From the Perspective of Temporal Parameters.*, 2017.

[14] S. Bird, E. Klein and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.*, 2009.

[15] INEGI, *Población Esperanza de vida.*, 2017. [Online]. Available:http://cuentame.inegi.org.mx/poblacion/esperanza.aspx?tema=P. [Accessed: Feb. 3, 2018].