Universidad Autónoma de Querétaro
Facultad de Informática
Maestría en Ciencias Computacionales

Desarrollo de una herramienta inteligente con técnicas de aprendizaje automático para la aplicación de entrevistas de trabajo

Opción de titulación
Tesis

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

Presenta:
Ing. Adriana Mansilla Hermosillo

Dirigida por:
Dr. Fausto Abraham Jacques Garcia

Dr. Fausto Abraham Jacques Garcia
Presidente — Firma

Dra. Ana Marcela Herrera Navarro
Secretario — Firma

Dra. Diana Margarita Córdova Esparza
Vocal — Firma

M.C. Ricardo Chaparro Sánchez
Suplente — Firma

Dr. Alberto Lara Guevara
Suplente — Firma

M.I.S.D Juan Salvador Hernández Valerio

Dra. Ma. Guadalupe Flavia Loarca Piña
Directora de Investigación y Posgrado

Centro Universitario
Querétaro, Qro.
Marzo 2019

1

RESUMEN

Hoy en día el Internet contiene una cantidad impresionante de material útil para explotar por medio de la minería de personalidad, sin embargo, muchas compañías no aprovechan la información que los candidatos tienen de sí mismos en línea y utilizan pruebas psicométricas tradicionales, que, si bien pueden determinar ciertos rasgos de la personalidad de los candidatos, los resultados de estas tienden a estar sesgados por la impresión que los candidatos quieren dar al empleador. El objetivo de esta investigación consiste en desarrollar una herramienta de aprendizaje automático que pueda ser utilizada para predecir la personalidad de un candidato basándose en el contenido escrito que se encuentra en sus redes sociales. Para evaluar la precisión de la herramienta se les pidió a los candidatos que participaron en esta investigación realizar la prueba de personalidad de Myers-Briggs de manera tradicional y después se compararon los resultados de los participantes y se compararon contra los resultados obtenidos por el clasificador automático. Para evaluar el desempeño de la herramienta se realizaron una serie de pruebas con las métricas más importantes como exactitud, precisión, y exhaustividad. Como resultado se obtuvo un mayor grado de exactitud con el clasificador basado en el algoritmo de bosque aleatorio, ya que, este logro un 69.0% de exactitud promedio además de obtener las mejores métricas.

(**Palabras clave:** Aprendizaje Automático, Clasificación automática de personalidad, Indicador Myers-Briggs, Clasificación de texto en español)

SUMMARY

Nowadays the Internet has an astonishing amount of useful material for personality mining, nevertheless many companies fail to exploit the information and screen job candidates using personality tests that fail to grasp the very information they are trying to gather. The aim of this investigation is to develop a machine learning classifier that can be used to predict the personality of a Spanish speaking job applicant based on the written content posted on their social networks. To evaluate the performance of the classifiers a test harness with the most important performance measures such as accuracy, precision and recall was made. The results show that the random forest classifier outperforms other classifiers in accuracy and most performance metrics with 69.0% of average accuracy.

(**Keywords**: Machine learning, Automatic personality classification, Myers-Briggs, Spanish text classification)

DEDICATORIA

Patricia Hermosillo y Enrique Mansilla mis padres:

Por ser mi mayor ejemplo de trabajo y lucha, por guiarme en lograr mis objetivos, educarme con amor y proporcionarme los medios para continuar mis estudios.

Laura y Enrique, mis hermanos:

Gracias por compartirme todo su apoyo y amor.

Claudia, Selene, Miguel, Gabriela, y Aldo, mis compañeros:

Por compartir su conocimiento conmigo, compartir experiencias, aventuras y mucho esfuerzo para la realización de este trabajo.

AGRADECIMIENTOS

Se agradece al Consejo Nacional de Ciencia y Tecnología (CONACYT) por solventar los recursos económicos necesarios para el desarrollo de este trabajo, de igual manera a la Dirección de Posgrado de la Facultad de Informática por permitirnos utilizar la infraestructura y materiales de los laboratorios.

A mi director de tesis Doctor Fausto Abraham Jacques García por su orientación en el desarrollo de esta investigación, sobre todo por ayudarme en mis momentos de necesidad y guiarme para completar este proyecto.

A la Universidad Autónoma de Querétaro por educarme en la verdad y en el honor.

1. TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

INDICE DE TABLAS

2. INTRODUCCIÓN

Los encargados de reclutar personal para las empresas generalmente tienen dos formas principales de determinar qué tipo de personalidad tiene una persona: por medio de tests psicométricos o por medio de entrevistas.

La entrevista es un método de gran importancia para analizar las características más importantes de una persona; sin embargo, también es uno de los métodos que requieren más entrenamiento por parte del aplicador. Este hecho puede representar una desventaja ya que no todos los reclutadores cuentan con el entrenamiento necesario para aplicar las entrevistas y menos al ser necesario el evaluar una gran cantidad de personal.

Por otro lado, los tests psicológicos se pueden generar y aplicar de manera sencilla, sin embargo, éstos suelen presentar el problema de arrojar resultados con un sesgo en la veracidad de las respuestas de los participantes ya que éstos responden en base a lo que piensan deben de contestar.

Una alternativa a estos métodos tradicionales para determinar la personalidad del usuario es el análisis lingüístico de textos generados por las personas sin previo conocimiento de que éstos serán analizados. La forma que se consideró más adecuada para realizar este análisis lingüístico es mediante la extracción de patrones lingüísticos característicos de cada uno de los tipos de personalidad, para que posteriormente se pueda calcular la cantidad en que estos patrones se encuentran en un texto determinado y de esta forma poder clasificar al individuo dentro de un tipo de personalidad.

La lingüística computacional se presenta como una de las alternativas más importantes para buscar y medir estos patrones. Gracias al poder de procesamiento de los algoritmos de aprendizaje automático, se puede realizar estudios más cuantitativos. Esto nos da una ventaja contra los métodos tradicionales que son más tardados por ser realizados por humanos

El método que se utilizará para abordar el problema será el modelo Bolsa de palabras. Con él, se generará una vectorización del vocabulario utilizado en las publicaciones de los usuarios. Una vez que se vectorice el vocabulario de las publicaciones, se buscarán correlaciones características para cada uno de los cuatro tipos de preferencias existentes de acuerdo con el modelo Myers-Briggs; dicha búsqueda se realizará mediante el uso del algoritmo de bosque aleatorio.

En caso de que se encuentren los patrones lingüísticos, éstos podrían servir como fundamento para realizar una clasificación automática de la personalidad del autor. De hecho, esta hipotética técnica de clasificación tendría la ventaja de no necesitar que el individuo sea consciente de que está siendo evaluado, ya que esta consciencia lleva el riesgo de que se presente un sesgo en sus respuestas; por ejemplo, mediante el deseo de complacer al experimentador.

Una mayor confiabilidad en los resultados de las formas de calificar la personalidad tiene repercusiones buenas en campos como el laboral y en este caso en específico para los reclutadores de personal.

Además, esta investigación es una manera de combinar distintas áreas del conocimiento. Entre estas se puede encontrar la psicología social, la lingüística computacional y la inteligencia artificial. Es así como se logra una interdisciplinariedad, uno de los objetivos que se plantean de forma recurrente en las tendencias académicas actuales.

2.1 Estado del arte

En las dos últimas décadas el aprendizaje automático y los sistemas de reclutamiento han florecido y muchos no se concentran en la personalidad de los candidatos, pero la cantidad de información disponible sobre éstos es una buena oportunidad para ampliar las fronteras de estos sistemas de reclutamiento, además de que, una herramienta como esta puede ser útil para centros de reclutamiento que no pueden probar a todos los candidatos con una entrevista personal.

En Faliagka et al (2012) se creó un sistema de reclutamiento en línea que incluía minería de personalidad automática para los candidatos a una posición de trabajo, ellos medían el grado de introversión/extroversión a través de la polaridad de las palabras que un candidato utilizaba en su blog personal, y después los ordenaban de acuerdo con las necesidades del reclutador. El nivel de extroversión lo calculan usando el modelo desarrollado por Pennebaker (2007) llamado LIWC (por sus siglas en inglés Linguistic Inquiry and Word Count, Buscador Lingüístico y Contador de Palabras).

En otra investigación de Tandera, et al (2017) se desarrolló un clasificador de personalidad para usuarios de Facebook utilizando técnicas de Deep Learning y el modelo de los cinco grandes y obtuvieron un 74.14% de exactitud.

Por último, un estudio llevado a cabo por Ortigosa, et al (2014) tuvo éxito al encontrar patrones mediante interacciones en Facebook de personas con personalidades similares basándose en el modelo de los cinco grandes alternativo.

Estos estudios enfocados a la clasificación automática de la personalidad utilizando redes sociales y algún modelo de personalidad no se desarrollaron como una herramienta orientada al reclutamiento de personal, sin embargo, estas investigaciones previas sugieren un precedente exitoso en donde aún queda mucho por ahondar en especial en un contexto hispano donde el procesamiento de lenguaje es más escaso que en lenguajes que cuentan con más conjuntos de datos como el inglés.

La psicología se encuentra ligada al reconocimiento de patrones desde su inicio como ciencia en 1879. La rama que se enfoca en encontrar estos patrones: la psicología cognitiva, inició al final de los años 50 cuando los psicólogos ingleses y americanos rechazaron la corriente del conductismo y adoptaron un modelo basado en una computadora.

Esta nueva corriente fue inspirada en tres acontecimientos externos: la teoría de la información; el acercamiento al modelado por computadora; y, por último, el trabajo de lingüística generativa de Chomsky.

Actualmente, el reconocimiento de patrones sigue estando estrechamente relacionado con las diversas ramas de la psicología. Por ejemplo, la psicofisiología busca patrones dentro de las ondas arrojadas por electroencefalogramas y busca asociarlos a diferentes estados como vigilia, sueño y comer, así como con estados patológicos como la epilepsia. (Carlson, 2006). Por otro lado, la Psicología de la personalidad estudia la personalidad y su variación entre los individuos (Blei, Ng, y Jordan, 2003)

El indicador de Myers-Briggs es un examen que se aplica comúnmente para identificar las características más importantes de la personalidad de un individuo.

Este indicador está basado en la teoría de C.G. Jung de los tipos psicológicos, la cual dicta que, aunque a simple vista algunos comportamientos parecen aleatorios en realidad son patrones y tienen que ver con la manera en que los individuos usan su percepción y su juicio.

Los exámenes de personalidad se refieren a probar las habilidades, personalidad, o aptitud de un individuo en relación con un escenario específico. Los exámenes de personalidad han adquirido más popularidad en organizaciones donde se busca seleccionar y reclutar a los candidatos más adecuados para distintos roles, y también para ascender o promover a otros empleados.

La confianza en las pruebas de personalidad ha ido en aumento a nivel mundial, incentivada por el crecimiento en mercados nuevos. En 2016 la empresa Global Assessment Barometer hizo un estudio a nivel mundial a 2,776 firmas de reclutamiento y recursos humanos. El estudio encontró que el 94% de las organizaciones aplicó estos exámenes durante el proceso de contratación.

Por su parte, el análisis de personalidad mediante algoritmos de aprendizaje automático se encuentra presente en trabajos como los siguientes:

En Plank et al. (2016) encontraron la correlación entre las características de la personalidad y el comportamiento lingüístico de los usuarios. Utilizan técnicas de procesamiento de lenguaje natural aplicadas a una base de datos de 1.2 millones de tweets en inglés. En este estudio se demostró que los datos en redes sociales pueden proveer suficiente evidencia para predecir de manera correcta dos de las 4 dimensiones de la personalidad.

En Ma (2017) entrenaron una red neuronal recurrente para predecir el tipo de personalidad MBTI utilizando como entrada extractos de novelas. Utilizando solo cinco oraciones de texto el modelo era capaz de predecir el tipo de personalidad de un escritor. El modelo tuvo un 37% de exactitud.

En otro trabajo, (Peng et al.,2015) utilizaron el reconocimiento automático de personalidad basados en las publicaciones de redes sociales de usuarios en chino. Colectaron un conjunto de datos con publicaciones y resultados de la prueba de personalidad de 222 usuarios de Facebook que utilizaban chino como su principal lenguaje. Después, utilizaron Jieba, una herramienta para segmentar chino, y Máquinas de Soporte Vectorial como el algoritmo de aprendizaje para realizar la clasificación. Sus resultados experimentales muestran que el desempeño en la precisión y índice de recuperación se podrían mejorar con la ayuda de la segmentación de texto.

En Xue et al (2018) también proponen el reconocimiento de la personalidad por medio de las redes sociales, sin embargo, proponen hacer uso del análisis de la semántica en los textos de los usuarios. De manera que implementan una red neuronal jerárquica con una estructura AttRCNN que aprende las características de la semántica para las publicaciones de cada usuario. Los resultados experimentales muestran que estos vectores semánticos aprenden de manera más efectiva que otros 4 métodos probados.

Varios sistemas han sido propuestos para la automatización de la pre-selección de candidatos. El sistema E-Gen de Kessler et al (2008) lleva a cabo la categorización de ofertas de trabajo, así como un análisis de relevancia de los candidatos. Para

hacer esto, el sistema usa un filtrado automático de curriculums, que son representados como vectores, mientras que el análisis de relevancia de candidatos se basa en el algoritmo de Vector de Máquinas de Soporte (VMS).

El sistema integral CommOn (Radevski y Trichet, 2006) utiliza tecnologías web semánticas en el campo de recursos humanos. En este sistema el candidato tiene que llenar un cuestionario en línea que después determina las características de su personalidad, sin embargo, el proceso de aplicar para una posición consume mucho tiempo, por lo tanto, no es un sistema amigable para el usuario.

Los sistemas de reclutamiento en línea deben de tener como objetivo el asistir a los reclutadores humanos en su toma de decisiones, así como incrementar la eficiencia del proceso. El sistema propuesto extrae criterios objetivos sobre los perfiles de los candidatos, los cuales se comparan contra los requerimientos del puesto para estimar la relevancia de cada candidato. Además, la presencia social de los candidatos es examinada para encontrar características que reflejen su personalidad.

A pesar de que, como se ha observado, ha habido algunas aproximaciones entre las técnicas de aprendizaje automático y los tipos de personalidad, no hemos encontrado en la literatura la aplicación de ambas al idioma español y al proceso de reclutamiento. Es por esto que se considera importante explorar este terreno con ayuda del modelo de bolsa de palabras y las máquinas de soporte vectorial debido a los buenos resultados que ha dado en las investigaciones mencionadas de este apartado. Además, investigaciones como las de Plank et al. (2016) refuerzan la idea de que se pueden predecir los tipos de personalidad a partir de patrones lingüísticos.

2.2 Marco Teórico

Los bosques aleatorios fueron introducidos en (Cutler, Cutler & Stevens, 2011) y ellos a su vez fueron influenciados por el trabajo de (Amit & Geman, 1997). Los árboles aleatorios son una extensión de la idea de Breiman de embolsado (Breiman, 1996) y fueron desarrollados como una alternativa al boosting (impulso). El

algoritmo del bosque aleatorio se puede usar tanto para problemas de clasificación como de regresión. De manera similar, las predicciones de este clasificador pueden ser categóricas o continuas. Desde un punto de vista computacional el bosque aleatorio es un algoritmo que tiene las siguientes ventajas:

- Se entrena y prueba de manera rápida.
- Sólo depende de uno o dos parámetros.
- Se puede utilizar para problemas de alta dimensionalidad.

El bosque aleatorio es un algoritmo de clasificación con una estructura simple. Para entender cómo funciona un bosque aleatorio primero se observa que saber cómo funciona un árbol de decisión.

Dado un conjunto de datos, un árbol hace niveles o particiona los datos con reglas (if-else). Es decir, un árbol crea reglas y estas se determinan a partir de la contribución de esa variable a la pureza de los nodos hijos.

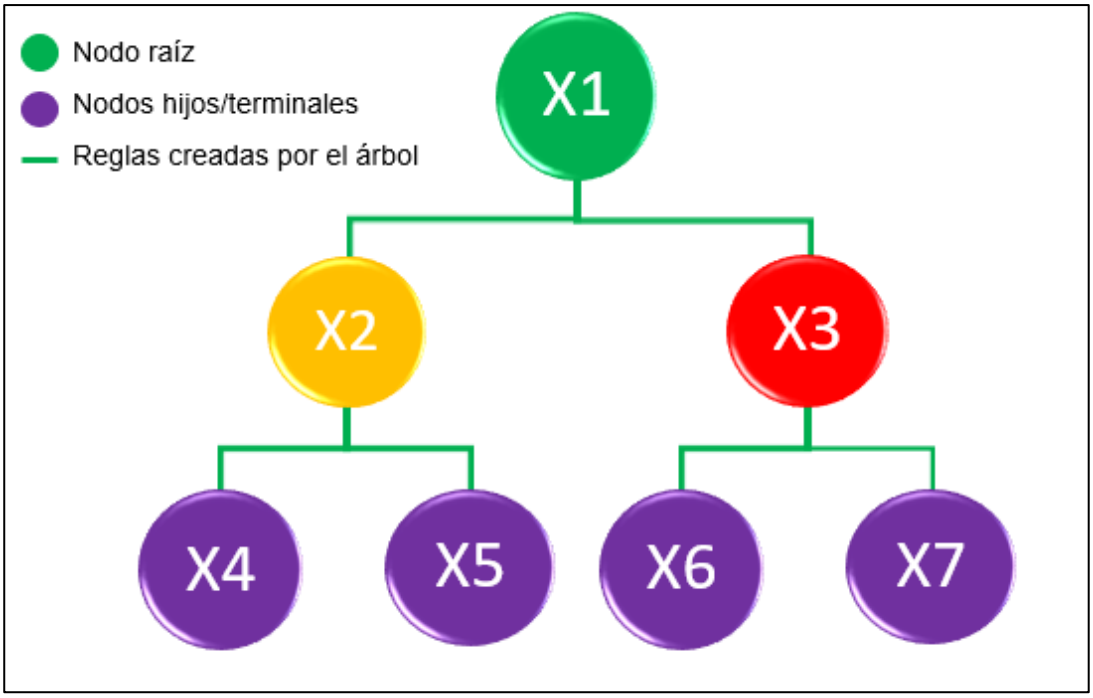


Figura 1 Árbol de decisión.

En la , la variable X1 da como resultado los nodos hijos con mayor pureza, por lo tanto, esa variable se convierte en un nodo raíz. Un nodo raíz es de las variables más importantes en un conjunto de datos.

Para determinar dónde se separa el árbol se usa la entropía, la cual es una medida de la impureza del nodo.

De manera que un nodo se va a dividir si su impureza es mayor a la del umbral establecido, de otra manera solo es un nodo terminal.

Una vez que se sabe cómo funciona un árbol de decisión se forma un conjunto de árboles de decisión que se transforman en un bosque, y este funciona de la siguiente manera:

- Se forma un subconjunto S2 () de entrenamiento con muestras de tamaño n y muestreo de reemplazo aleatorio a partir del conjunto original de muestras (S1).

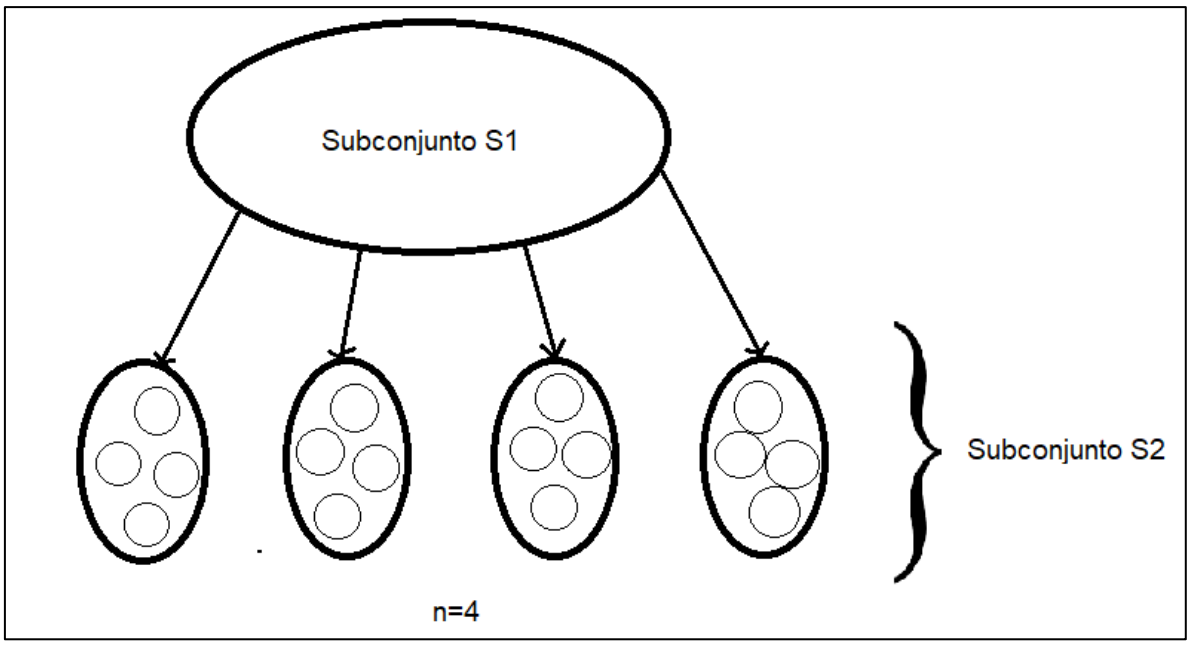


Figura 2 Subconjunto de entrenamiento.

- Si existen M variables de entrada, un número m<M se debe de especificar de manera que se seleccionen m variables de M. Aquí se emplea el método de división por nivel de impureza para separar el nodo y crear a los nodos hijos con mayor pureza posible. El valor de m es constante a lo largo del crecimiento del bosque.
- Cada árbol crece sin límite de profundidad.
- Finalmente, varios árboles han crecido y se obtiene una predicción final por medio de votación o promedio.

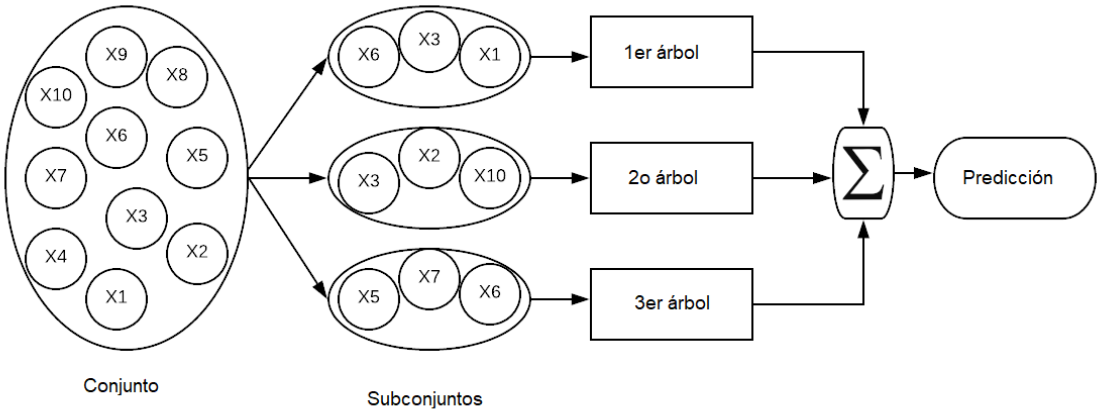


Figura 3 Formación del bosque aleatorio

2.3 Indicador Myers-Briggs

Este indicador nació para hacer la teoría de los tipos psicológicos descrita por C.G. Jung entendible y útil para la vida de las personas. La esencia de la teoría es que el comportamiento es más consistente de lo que se cree, en todos los seres se observan diferencias básicas en la manera en que preferimos usar nuestra percepción y juicio. Si la gente tiene distintas formas de percibir nueva información y de llegar a conclusiones entonces es lógico que tengan diferentes intereses, reacciones, valores, motivaciones y habilidades. Para dividir los tipos de personalidad diferentes se establecieron cuatro ejes distintos, según The Myers Briggs Foundation (2018), como se muestra en la .

Tabla 1 Ejes principales del Indicador Myers-Briggs

Preferencia	Dicotomía
Mundo favorito: ¿Exterior o Interior?	Extroversión(E)/Introversión(I)
Información: ¿Prefieres concentrarte en la información básica que te llega o prefieres agregar significado?	Sensorial (S) / Intuitivo (N)
Decisiones: ¿Al tomar decisiones prefieres primero ver la lógica y después las circunstancias?	Pensar (T) / Sentir (F)
Estructura: ¿Prefieres tomar un lado o permanecer abierto a nueva información?	Juzgar (J) / Percibir (P)

Una vez que se decide la categoría de una persona para cada uno de los 4 ejes se ejemplifica su personalidad con un código de 4 letras. Los 16 tipos posibles de personalidad se muestran en la a continuación:

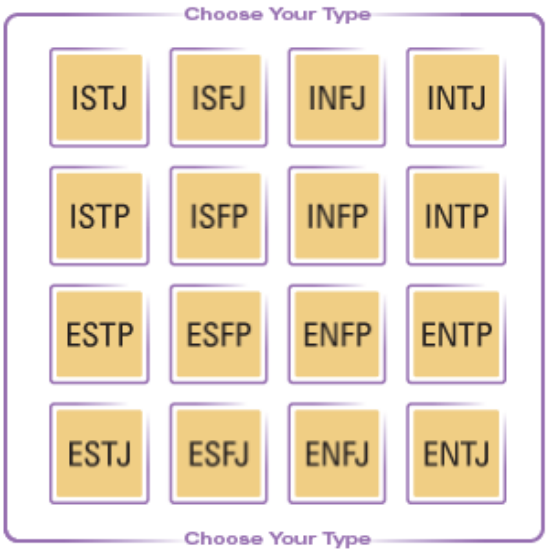


Figura 4 16 tipos de personalidad según el indicador Myers-Briggs.

De manera tradicional alguien puede tomar esta prueba de personalidad contestando preguntas como las que se muestran en la :

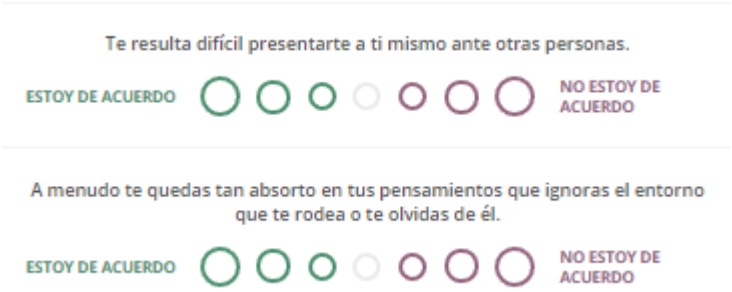


Figura 5 Ejemplos de preguntas y escala de respuestas del indicador Myers-Briggs.

3. OBJETIVOS

3.1 Objetivo General:

El objetivo principal de esta investigación es buscar patrones lingüísticos característicos de cada uno de los tipos de personalidad definidos en el modelo Myers-Briggs a partir de textos obtenidos de las redes sociales de los usuarios mediante el uso de lingüística computacional y técnicas de aprendizaje automático.

3.2 Objetivos Específicos:

- Establecer criterios de credibilidad para clasificar a los candidatos basados en el modelo de jerarquía analítica.
- Utilizar un método de vectorización para generar los tópicos principales que componen a todos los textos obtenidos de las redes sociales.
- Utilizar un algoritmo de aprendizaje automático para aprender directamente de un modelo de espacio de palabras las características para clasificar la personalidad de cada usuario, así como también aprender de los temas que componen los textos de las redes sociales.
- Argumentar las posibles causas de los patrones encontrados o, en su defecto, del porqué no se encontraron dichos patrones.

4. METODOLOGÍA

La Figura 6 muestra el proceso que se siguió para realizar esta investigación. Ya que se utilizó el modelo más conocido para probar algoritmos de aprendizaje automático se puede resumir en los siguientes pasos: recolección de datos, limpieza, vectorización, entrenamiento y pruebas de clasificación.

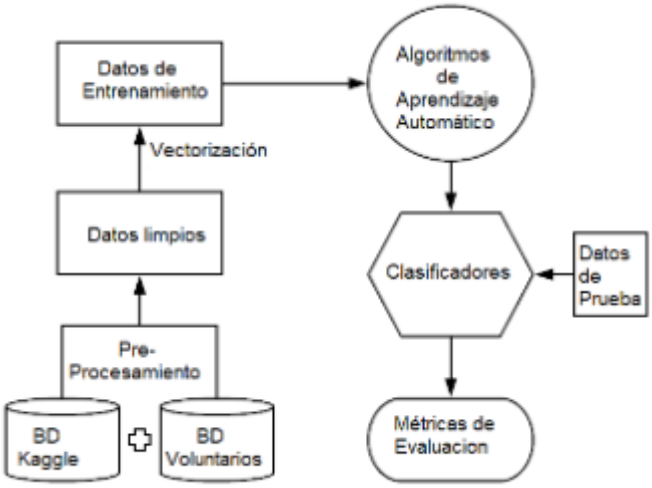


Figura 6 Metodología para probar algoritmos de aprendizaje automático.

4.1 Conjunto de Datos

Los datos para el entrenamiento se obtuvieron por medio de la descarga en el sitio kaggle donde se encuentra una base de datos con los posts de más de 8 mil usuarios y su resultado del indicador Myers-Briggs. De éstos más de 8 mil registros se tradujeron en total 852 al español y se usaron como datos de entrenamiento para los 3 clasificadores. Para los datos de prueba se consiguieron 44 estudiantes voluntarios de la facultad de informática de la Universidad Autónoma de Querétaro quienes aceptaron tomar la prueba y compartir el contenido de su perfil de Twitter para el experimento.

En la Figura 7 y Figura 8 se muestra un fragmento de la base de datos de entrenamiento y la de prueba:

4	ENTP	"Creí en Dios toda mi vida hasta hace un año. Mi madre creía mucho en Dios y nos alentó a mí y a mis hermanos a ser de la misma manera. Ella no nos forzó en nuestras gargantas o ... ||| Me acecho en todas partes para ser honesto. ||| Tuve un cone...
5	INTP	'Estoy rebotando dentro y fuera de este hilo como lo permite mi temperamento, lo admitiré. Estoy usando mis reservas de temperamento para responder esto. La letra cursiva no es razonable. Eso no va a argumentar en contra. En ... ||| Solo diré que...
6	ISFJ	'Vencí eso buscando a un compañero de casa que trabajaba meses a la vez, y cuando estaba en casa, visitó a sus compañeros después de que terminaron el trabajo, así que estaba solo el 80% del tiempo: D oh, yo no ... | || No hay pegajoso para que s...

Figura 7 Muestra de conjunto de entrenamiento

32	ENTP	Cada vez que tomo decisiones, trato de reducir los riesgos y problemas que puedan existir.???????? Si una persona no tiene sueños no tiene razón de vivir, soñar es necesario aún cuando el sueño va más allá de la re… https://t.co/QBdO2iU68X
33	ENTP	RT @_Cinthhya: Exactamente un mes para mi cumple ! ???? #PalPendiente ???? Pensando en ti!! ???????????? https://t.co/6MflY7nXtP @lapizito123 Carnal yo solo veía acábatelo por ti! Haaa eres quien hacia reír!!! #Uvimeza @_Cinthhya vamos!...
34	INTP	RT @MegBucher: Heading to bed, hopefully we hit 700 RTs and I can give you guys this #ArcadeMF code in the morning! https://t.co/PdMq7YV0UG RT @cdollarsc: Need 40 more followers! RT @cdollarsc: 34 more followers! RT @freddyomana94: Follo...

Figura 8 Muestra de conjunto de prueba.

Una vez recabados los datos se hizo un estudio para determinar la distribución de los datos tanto de la muestra obtenida de internet del sitio kaggle como de los datos que se obtuvieron de los participantes voluntarios, en la tabla 1 se muestra el promedio de edad entre los estudiantes voluntarios que tomaron el indicador Myers-Briggs. En la Tabla 2 se muestra como están divididos sus resultados por eje y por último en la Tabla 3 y Tabla 4 se muestra la distribución de la base de datos de kaggle y en la Tabla 5 a la muestra en total:

Tabla 2 Características del conjunto de datos de prueba.

Característica	Valor
Tamaño de la muestra	44
Edad promedio	22.72
Distribución de géneros	35 masculino / 9 femenino
Edad mínima-máxima	19-36

Tabla 3 Distribución de tipos de personalidad del conjunto de datos de prueba.

Característica	Valor
Introversión/Extroversión	33/11
Sensorial/Intuitivo	6/38
Pensar/Sentir	17/27
Juzgar/Percibir	27/17

Tabla 4 Distribución de tipos de personalidad del conjunto de datos de entrenamiento.

Característica	Valor
Introversión/Extroversión	191/661
Sensorial/Intuitivo	165/687
Pensar/Sentir	426/426
Juzgar/Percibir	523/329

Tabla 5 Distribución de tipos de personalidad del conjunto de datos entero.

Característica	Valor
Introversión/Extroversión	224/672
Sensorial/Intuitivo	171/725
Pensar/Sentir	443/453
Juzgar/Percibir	550/346

4.2 Pre-procesamiento

El preprocesamiento incluye eliminar signos de puntuación, HTML, emojis, números, urls, y demás caracteres, poner todo en minúsculas y dejar solamente la raíz de las palabras. Este proceso se aplicó en ambas bases de datos, tanto en la de entrenamiento como en la de prueba y después se procedió a entrenar y probar a los clasificadores. Sin embargo, primero se vectorizaron los datos, esto se llevó a cabo por medio de un proceso conocido como extracción de características. En la Figura 9, se muestra un ejemplo del tratamiento que recibió la base de datos para prepararla para la vectorización:

```
¿Cuál ha sido la experiencia que más cambió tu vida? || http: //
www.youtube.com/watch?v=vXZeYwwRDw8
http://www.youtube.com/watch?v=u8ejam5DP3E En repetición durante
la mayor parte del día. ||| Que la experiencia PerC te sumerja.
||| Lo último que mi El amigo de INFJ publicó en su facebook antes
de suicidarse al día siguiente. Descansa en paz ~
http://vimeo.com/22842206|||...
```

Figura 9 Texto muestra sin pre-procesamiento.

En la Figura 10 y Figura 11 se pasa todo a minúsculas, se eliminan símbolos (.,/*-+) y palabras comunes, asi como dígitos.

```
cuál sido experiencia cambió vida http
wwwyoutubecomwatchvvxzeywwrdw8 httpwwwyoutubecomwatchvu8ejam5dp3e
repetición mayor parte día experiencia perc sumerja último amigo
infj publicó facebook suicidarse día siguiente descansa paz
httpvimeocom22842206...
```

Figura 10 Texto muestra sin símbolos y palabras comunes.

```
cuál sido experiencia cambió vida repetición mayor parte día
experiencia perc sumerja último amigo publicó facebook suicidarse
día siguiente descansa paz
```

Figura 11 Texto sin urls ni tipos de personalidad.

En la Figura 12 se visualiza las 15 palabras que aparecen con menos frecuencia y se eliminan.

```
shadowlands    1
plagada        1
trabajara      1
cacique        1
lind           1
riften         1
primordial     1
xnfps          1
ynsieztmks4    1
dactilar       1
comportamos    1
demi           1
t7yzourfbno    1
sorbo          1
chevron        1
```

Figura 12 Lista de 15 palabras menos frecuentes en la muestra.

Enseguida, como se muestra en la Figura 13 se procede a lematizar todas las palabras, es decir, obtener la raíz de todas las palabras para así contabilizarlas como una sola.

```
cual sid experient camb vid repeticion mayor part dia experient
perc sumerj ultim amig public facebook suicid dia siguient descans
paz
```

Figura 13 Texto muestra con todas las palabras lematizadas.

Una vez lematizadas todas las palabras de ambas bases de datos se pueden pasar a vectores numéricos, los cuales después serán analizados por el algoritmo

clasificador para encontrar patrones que definan cada dicotomía de los tipos de personalidad.

4.3 Vectorización de los datos

El algoritmo de vectorización consiste principalmente en construir un vocabulario a partir de las palabras que se extrajeron en la sección previa. Primero se llama a una clase vectorizadora que “aprende” el vocabulario de todas las publicaciones pre-procesadas anteriormente. Después, se codifica o transforma ese vocabulario a un vector numérico que contiene el tamaño del vocabulario y un conteo del número de veces que aparece esa palabra en cada muestra de las publicaciones de los usuarios.

En la Figura 14 se muestra una parte del vocabulario:

```
{'moment': 265, 'mejor': 250, 'jug': 221, 'brom': 43, 'cual': 86,
'sid': 376, 'experient': 151, 'vid': 438, 'mayor': 247, 'part':
305, 'dia': 107, 'perc': 311, 'ultim': 417, 'amig': 20, 'public':
33...}
```

Figura 14 Vector de vocabulario de tamaño 449. En este caso se puede ver el índice de las palabras, es decir, la palabra 'moment' ocupa el lugar 265 en el vector de tamaño 449.

El vector resultante contiene 852 vectores de tamaño 449. En la Figura 15 se muestra un fragmento del vector para el primer registro.

```
  (0, 214)  1

  (0, 47)   3

  (0, 403)  5

  (0, 313)  1

  (0, 356)  1

  (0, 141)  2

    : :
```

Figura 15 En este ejemplo se ve como la palabra con el índice 403 aparece 5 veces, la palabra 356-1, y la 141-2 veces, todo esto para el primer usuario.

4.4 Normalización de los datos

Una vez que se contaron las veces que aparece cada palabra del vocabulario en cada registro se puede proceder a normalizar este vector con el algoritmo Tf-Idf(Term frequency - Inverse document frequency).

Este algoritmo como su nombre indica: primero, calcula la frecuencia con la que aparece una palabra, después, se encarga de escalar la importancia de las palabras que se repiten mucho a través de la base de datos, de manera que el algoritmo se encarga de resaltar las palabras más interesantes de cada registro de la base de datos y no las palabras más repetidas en toda la base de datos.

En la Figura 16 aparece una muestra de las frecuencias inversas del vocabulario. Se asigna el puntaje más bajo de 1.0 a las palabras que aparecen con más frecuencia.

```
[3.0125611  2.96963605 2.67358573 2.34608217 3.11403056 2.77202581
 3.1536397  2.86595762 2.14664073 2.52301287 2.85091975 3.27142273
 1.60672214...]
```

Figura 16 Vector de frecuencias inversas del vocabulario.

Finalmente, se calculan los puntajes normalizados de cada registro en la base de datos. En la Figura 17 se muestra una parte del vector resultante para el primer registro.

```
[0.         0.         0.         0.         0.         0.
 0.         0.07904235 0.         0.         0.         0.
 0.         0.         0.04225639 0.04751695 0.         0....]
```

Figura 17 Vector con frecuencias normalizadas entre 0 y 1.

En la Figura 18 se muestran los 15 términos con mayor frecuencia para demostrar los resultados del proceso.

```
 'si',
 'sol',
 'hac',
 'cre',
 'ser',
 'person',
 'tip',
 'gust',
 'buen',
 'com',
 'vez',
 'cos',
 'realment',
 'asi',
 'tiemp'
```

Figura 18 Lista con las 15 palabras con mayor frecuencia en la muestra.

Una vez que el vector normalizado está listo se puede usar directamente para calcular los resultados.

4.5 Transformar tipo de personalidad MBTI a vector binario

Por último, antes de concluir la etapa de procesamiento de la base de datos. Se observa el vector normalizado X, sin embargo, falta binarizar el vector con el tipo de personalidad para que los datos queden en forma X-Y.

- X: Publicaciones de los usuarios en representación tf-idf.
- Y: Vector de tipo de personalidad MBTI binarizado.

La primera columna de nuestra base de datos llamada 'tipo' se compone de un indicador de 4 letras, las cuales pueden ser:

- IE: Introversión (I) / Extroversión (E)
- NS: Intuición (N) / Sensorial (S)
- FT: Sentir (F) / Pensar (T)
- JP: Juzgar (J) / Percibir (P)

Para binarizar este indicador se tomarán las letras como 1 o 0, como lo muestra la Figura 19:

```
'I':0, 'E':1
'N':0, 'S':1
'F':0, 'T':1
'J':0, 'P':1
```

Figura 19 Binarización del tipo de personalidad

De manera que el indicador binarizado de un usuario del tipo INFJ quedaría como lo muestra la Figura 20:

```
I N F J

[0 0 0 0]
```

Figura 20 Ejemplo de binarización del tipo de personalidad INFJ.

Una vez que el vector ha sido binarizado se concluye con el tratamiento de la base de datos y se procede a entrenar el clasificador.

5. RESULTADOS Y DISCUSIÓN

5.1 Primera Fase

En la primera fase de la experimentación se consideró por separado a cada una de las cuatro dicotomías de tipo de personalidad de Myers-Briggs. Se entrenó al modelo con el algoritmo de Bosque aleatorio y se realizó un entrenamiento individual para cada uno de los tipos de personalidad: **Introversión/Extroversión, Intuitivo/Sensorial, Pensador/Emocional, y Calificador/Perceptivo**.

Además, se separa la base de datos de manera que los registros obtenidos en línea se utilizan como datos de entrenamiento para el clasificador, y los restantes 44 de alumnos voluntarios se utilizaron para las pruebas. En seguida se muestran los cuadros con el resultado de exactitud por cada rubro.

Es así como para el entrenamiento del tipo 'Extrovertido' se tomaron como valores positivos las publicaciones de participantes con este tipo de personalidad, mientras que las publicaciones de los participantes con el tipo 'Introvertido' se tomaron como negativas. Esto sucedió para cada uno de los cuatro tipos.

A partir de este entrenamiento se pidió al algoritmo de bosque aleatorio intentar reconocer en el conjunto prueba si las publicaciones pertenecían o no al tipo de personalidad analizado. Hay que recordar que el conjunto de prueba está

conformado por las 44 publicaciones de usuarios restantes; es decir, de las 852 publicaciones, 852 se asignaron al conjunto de entrenamiento y 44 al conjunto de prueba. La Tabla 6 muestra los resultados sintetizados de esta prueba.

Tabla 6 Resultados con conjunto de prueba y entrenamiento separados.

Tipo de personalidad	I/E	S/N	T/F	J/P	Promedio
Exactitud	25.00%	86.36%	45.45%	63.64%	55.11%
Precisión	6.00%	75.00%	77.00%	77.00%	58.75%
Ind. Recup.	25.00%	86.00%	45.00%	64.00%	55.00%
F1	10.00%	80.00%	35.00%	52.00%	44.25%

La exactitud de los resultados es la métrica más común para medir el desempeño de un modelo, ya que, es igual a la proporción de las observaciones predichas de manera correcta entre las observaciones totales.

Aunque esta métrica es buena para medir los resultados de manera general en este caso se emplearon más métricas ya que nuestra muestra de datos no está balanceada, es decir, los valores de falsos positivos y falsos negativos no vienen en una cantidad similar. Por lo tanto, agregamos otros parámetros para evaluar el desempeño del modelo de bosque aleatorio. Para el modelo se obtuvo una exactitud promedio del 55.11%

La precisión es la proporción de resultados predichos de manera correcta entre el total de resultados correctos. Para este rubro se obtuvo un promedio total del 58.75%.

El índice de recuperación es la proporción de observaciones predichas correctamente entre el número total de observaciones en cada clase.

F1 es el promedio de la precisión y el índice de recuperación. Por lo tanto, esta métrica toma los falsos positivos y los falsos negativos en cuenta. Esta métrica no es tan fácil de entender como la exactitud, pero es más útil.

En resumen, los mejores resultados del clasificador creado con el algoritmo de Bosque aleatorio obtuvieron un promedio de exactitud de 55.11%.

5.2 Segunda Fase

Para esta prueba se mezclaron los datos de prueba y entrenamiento de manera aleatoria para promover un índice más alto de exactitud. Los datos se separaron con el 80% como datos de entrenamiento y el resto de prueba. El resultado aparece en la Tabla 7.

Tabla 7 Resultados con conjunto de prueba.

Tipo de personalidad	I/E	S/N	T/F	J/P	Promedio
Exactitud	77.23%	77.23%	64.73%	57.14%	69.08%
Precisión	82.00%	60.00%	65.00%	54.00%	65.25%
Ind. Recup.	77.00%	77.00%	65.00%	57.00%	69.00%
F1	68.00%	67.00%	65.00%	45.00%	61.25%

Los mejores resultados obtenidos para la segunda fase en cuanto a clasificaciones correctas (exactitud) fueron de 77.23% para Introversión/Extroversión y 77.23% para Sensorial/Intuitivo.

Se puede notar como los resultados del experimento B son mejores que los del experimento A. Sin embargo, en los dos experimentos el clasificador con mejor exactitud y métricas resultó ser el del Bosque aleatorio con un porcentaje de 55.11 en el experimento A y de 69.08 en el experimento B. Estos resultados pueden ser debido a que el algoritmo del bosque aleatorio hace conjuntos y subconjuntos de los datos de entrenamiento lo que reduce el efecto que pueden tener los datos más

alejados de una clase, sin embargo, como se afirma en el estudio (Wolpert, 1995) siempre van a haber conjuntos de datos en los que un clasificador es mejor que otro. En este estudio se podrían pulir todavía más los datos al igual que recolectar una muestra mayor de las clases con menos ejemplos, al igual que, también se podría intentar reducir en mayor grado el vector de vocabulario para mejorar los resultados de los tres algoritmos. Otra manera de mejorar los resultados sería utilizando técnicas lingüísticas como la asociación de las palabras con sinónimos y antónimos para hacer el vocabulario más robusto con respecto a las variaciones de vocabulario y parafraseo. Además, se podría probar una combinación de clasificadores como en (Segrera et al., 2006).

6. CONCLUSIÓN

En este trabajo se utilizaron herramientas de aprendizaje automático, específicamente de lingüística computacional, para intentar encontrar una relación entre las palabras empleadas en las redes sociales por candidatos y su tipo de personalidad de acuerdo con el indicador Myers-Briggs. Para esta hipótesis el estudio se basó la afirmación de que se pueden predecir los tipos de personalidad a partir de un análisis lingüístico, en conjunto con el reciente florecimiento de técnicas de análisis de datos cada vez más potentes como lo es el agrupamiento automático de palabras relacionadas entre sí. Para buscar la posible relación antes mencionada se realizaron experimentos en dos fases:

En la primera fase se clasificaron los individuos de acuerdo con uno de los 16 tipos de personalidad del indicador Myers Briggs. Esta clasificación se hizo en base a las palabras que los usuarios utilizaron en sus textos de Twitter. En este intento se lematizaron las palabras, para que palabras como "niño" y "niña" contaran como una sola. Como resultado se tuvo un buen aprendizaje, al grado que se clasificó adecuadamente un promedio de 55.11% de los datos en la primera fase y 69.08% en la segunda.

Además, dentro de las dos fases se revisó la capacidad del algoritmo de Bosque aleatorio, famoso por sus altos niveles de exactitud y bajo nivel de complejidad, para clasificar de manera automática la personalidad de usuarios de habla hispana. Utilizando el clasificador entrenado con el algoritmo bosque aleatorio y una mezcla diferente de los datos de entrenamiento y prueba. Los mejores resultados se obtuvieron mezclando los datos de entrenamiento y los datos de prueba. Con el tratamiento que se le dio al conjunto de datos se pudo clasificar correctamente el 69.08% de los casos. Por lo tanto, los resultados son positivos y con algo de mejoras en la cantidad de datos para la muestra y mejor preprocesamiento, este clasificador podría mejorar y obtener un mejor porcentaje para clasificar la personalidad de los usuarios. La aplicación que puede tener esta herramienta en el campo de reclutamiento es grande, ya que las empresas hoy más

que nunca buscan una fuerza laboral que se amolde a su filosofía, por lo tanto, un clasificador de personalidad que funcione con tal rapidez y exactitud supondría una buena referencia para encontrar los rasgos que se buscan en una nueva contratación. Para continuar con esta investigación se plantea un nuevo experimento donde, después de incrementar la exactitud del clasificador, este pueda ser utilizado en un proceso de reclutamiento en un escenario concreto y delimitado donde los candidatos obtengan sus resultados y retroalimentación del sistema, así como una entrevista personal para cotejar los resultados que se obtengan mediante estas dos técnicas y así ver su grado de exactitud también.

6.1 Aportaciones del trabajo final

Como parte de las aportaciones del presente trabajo terminal se puede mencionar que este es uno de los primeros trabajos en los que con técnicas de Aprendizaje Automático e inteligencia artificial se intenta hallar una correlación entre la escritura y la clasificación que de este arroja un instrumento psicométrico como lo es el indicador Myers-Briggs. De igual forma, se realizaron sugerencias y ajustes de parámetros tanto de Bosque aleatorio como de bolsa de palabras para realizar esta tarea en particular.

Adicional, se implementó la conversión de un espacio de palabras a un espacio numérico de 449 palabras. Respecto al impacto en la Lingüística Computacional, se encontró que el uso de un preprocesamiento mejora el aprendizaje del bosque aleatorio, comparado con el aprendizaje que se tiene sobre el conjunto solo de palabras, es decir, el modelo sencillo de palabras.

Por otro lado, para las técnicas empleadas en el trabajo se observó que la normalización de los datos juega un papel importante para su clasificación en el algoritmo del bosque aleatorio. Esto indica que la cantidad de palabras utilizadas en el momento de escribir un tweet es importante para su clasificación.

6.2 Trabajo futuro

Como trabajo a futuro, se propone aplicar los métodos presentados en este trabajo terminal a otros conceptos psicológicos y otros instrumentos psicométricos, para determinar en qué grado se puede encontrar una correlación ente la escritura de un sujeto y los valores asignados por dicho instrumento.

Como posible mejora para las técnicas empleadas en este trabajo, se propone complementar las características utilizadas con otros datos que pudieran resultar relevantes, por ejemplo, sinónimos extraídos automática o manualmente, categorías gramaticales de las palabras, el uso de artículos y de diccionarios de clasificación de palabras, como podría ser el uso de ontologías, etc. Es posible que de esta forma mejore el nivel de aprendizaje del clasificador de bosque aleatorio y haya una mayor exactitud en las clasificaciones de los sujetos respecto a su tipo de personalidad.

También como trabajo futuro se sugiere determinar el grado en que la utilización de preprocesamiento mejora el aprendizaje del bosque aleatorio con respecto a otras técnicas.

Finalmente, podría ser importante para trabajos a futuro el considerar que en este trabajo se intentó evitar el uso de recursos adicionales como diccionarios, analizadores sintácticos y ontologías, con el objetivo de depender lo menos posible de dichos recursos y permitir que lo hecho en el trabajo sea fácilmente generalizable a otros idiomas.

7. REFERENCIAS

Amit, Y.,y Geman, D. (1997). Shape Quantization And Recognition With Randomized Trees.. Neural Computation. 9. 1545-1588. 10.1162/neco.1997.9.7.1545.

Blei, D., Ng, A., y Jordan, M. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: 993-1022

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2),123–140.

Carlson, N. R. (2006) Fisiología de la Conducta. México: Pearson-Addison.

Cutler, A, Cutler, D y Stevens, J. (2011). Random Forests. 10.1007/978-1-4419-9326-7_5.

Faliagka, E., Tsakalidis, A., y Tzimas, G. (2012). An Integrated E-Recruitment System for Automated Personality Mining and Applicant Ranking. *Internet Research*, 551-568.

Kessler, R., Torres-Moreno, J. M., y El-Bèze, M. (2008). E-Gen: Profilage automatique de candidatures. TALN 2008, Avignon, France, 370-379.

Ma, A. V. (2017). Neural networks in predicting myers brigg personality type from writing style.

Ortigosa, A., Carro, R., y Quiroga, J. (2014). Predicting user personality by mining social interactions in Facebook. *Journal of Computer and System Sciences*, 57-71.

Peng, K., Liou, L., Chang, C., y Lee, D. (2015). Predicting personality traits of chinese users based on facebook wall posts. doi: 10.1109/WOCC.2015.7346106

Pennebaker, J. W. (2007). The development and psychometric properties of LIWC2007. *LIWC.net*.

Plank, B., Sogaard, A., y Goldberg, Y. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss.

Radevski, V y Trichet, F. (2006). Ontology-Based Systems Dedicated to Human Resources Management: An Application in e-Recruitment. 4278. 1068-1077. 10.1007/11915072_9.

Segrera, S., y Moreno, M. N. (2006). An experimental comparative study of web mining methods for recommender systems. *Proceedings of the Sixth WSEAS International Conference on Distance Learning and WebEngineering*, 56-61..

Tandera, T., Suhartono, H., Suhartono, D., Wongso, R., y Prasetio, Y. (2017). Personality Prediction System from Facebook Users. *Procedia Computer Science*, 604 - 611.

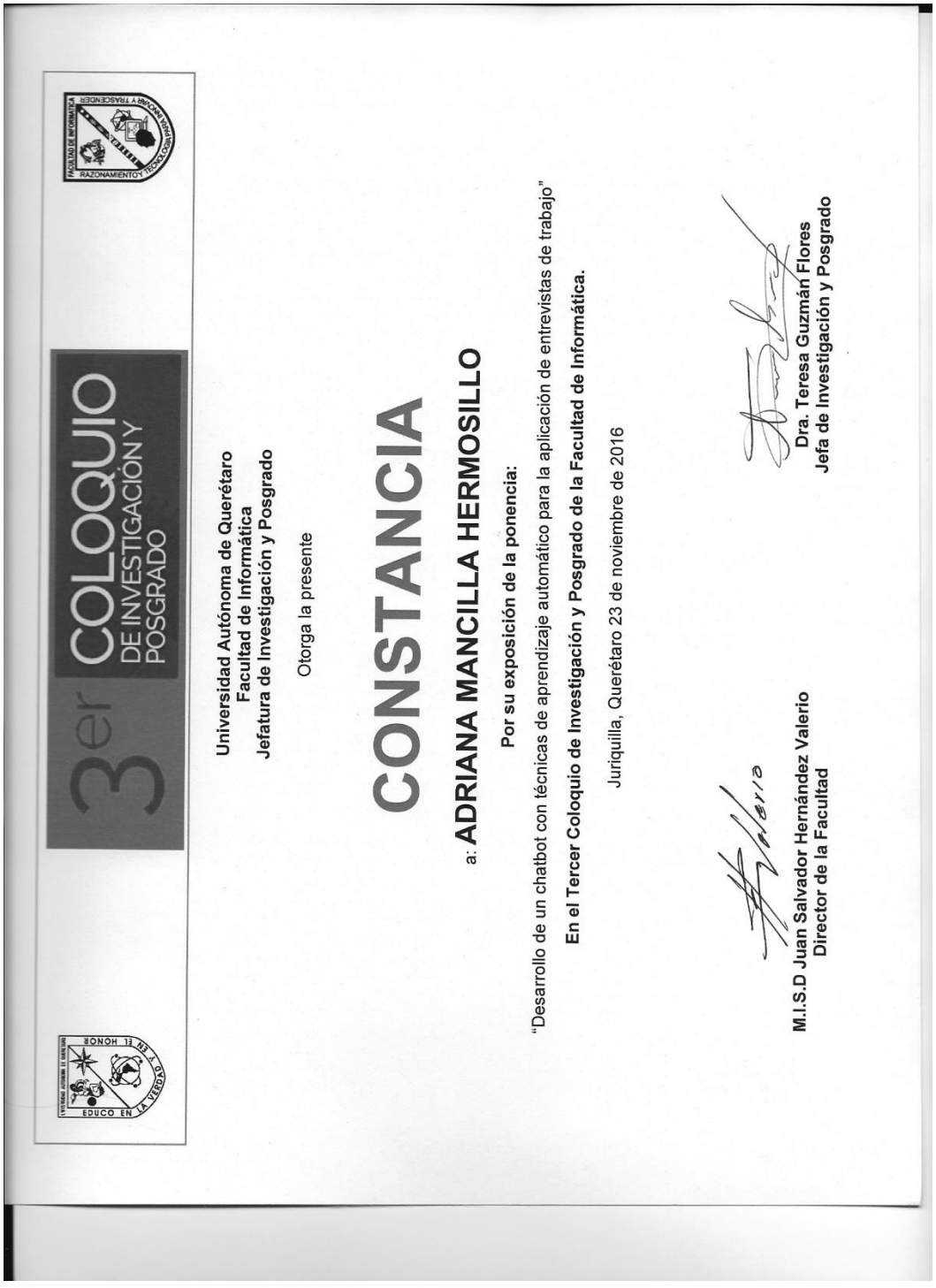
The Myers y Briggs Foundation. (2018). Retrieved from https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm?bhcp=1

Wolpert, D. M. (1995). No Free Lunch Theorems for Search. *Technical Report SFI-TR-95-02-010*.

Xue, D., Wu, L., Hong, Z., Guo, S., Gao, L., Wu, Z., Zhong, X., Sun, J. (2018). Deep learning-based personality recognition from text posts of online social networks. Applied Intelligence.

8. APÉNDICE

8.1 Coloquio de investigación y Posgrado



3er COLOQUIO DE INVESTIGACIÓN Y POSGRADO

Universidad Autónoma de Querétaro
Facultad de Informática
Jefatura de Investigación y Posgrado

Otorga la presente

CONSTANCIA

a: **ADRIANA MANCILLA HERMOSILLO**

Por su exposición de la ponencia:

"Desarrollo de un chatbot con técnicas de aprendizaje automático para la aplicación de entrevistas de trabajo"

En el Tercer Coloquio de Investigación y Posgrado de la Facultad de Informática.

Juriquilla, Querétaro 23 de noviembre de 2016

M.I.S.D Juan Salvador Hernández Valerio
Director de la Facultad

Dra. Teresa Guzmán Flores
Jefa de Investigación y Posgrado

8.2 Congreso Nacional en Computación y Tecnología Educativa

1er Congreso Nacional en Computación y Tecnología Educativa

Universidad Autónoma de Querétaro
Facultad de Informática
Jefatura de Investigación y Posgrado

Otorga la presente

CONSTANCIA

a: **Mansilla Hermosillo Adriana**

Por presentar la ponencia "Análisis de sentimiento de críticas de películas con el algoritmo de bosque aleatorio"
Facultad de Informática, Juriquilla, Querétaro 27 de noviembre de 2017

M.I.S.D Juan Salvador Hernández Valerio
Director de la Facultad

Dra. Ana Marcela Herrera Navarro
Jefa de Investigación y Posgrado

8.3 Taller Mexicano de Detección de Plagio y Análisis de Autoría

El Instituto Nacional de Astrofísica,
Óptica y Electrónica

otorga la presente constancia a

Adriana Mansilla Hermosillo

Por su participación con la ponencia

"Desarrollo de un chatbot inteligente para selección de personal"

en el 2do Taller Mexicano de Detección
de Plagio y Análisis de Autoría, llevado a cabo
el 17 de julio de 2017 en Cholula, Puebla.

Luis Villaseñor Pineda
Organizador, INAOE

Manuel Montes y Gómez
Organizador, INAOE

8.4 Publicación de artículo en congreso internacional.

IEEE Xplore® Digital Library

> Institutional Sign In

Browse ⌄ My Settings ⌄ Get Help ⌄ Subscribe

All ⌄ Enter keywords or phrases (Note: Searches metadata only by default. A search for 'smart grid' = 'smart AND grid')

Conferences > 2017 Twelfth Latin American C...

Implementation of an chatbot in a serious game associated with the acquisition of social skills and the promotion of collaborative tasks in children

6 Author(s) Adriana Mansilla ; Alberto Ochoa ; Julio Ponce ; Marcela Herrera ; Alberto Hernández ; Edgar Cossio View All Authors

273 Full Text Views

Abstract

Document Sections

I. Introduction

II. Background

III. Prototype Proposed

IV. Discussion and Future Work

Authors

Figures

References

Abstract:
Just a few decades ago the society has been interested in the field of children's rights. For this reason, the society as a whole has been developing strategies to achieve that commitment. As a result, it has given more attention to this society segment that is vulnerable. With this the society does not pretend to expose the individual in the early stages of his life. The way in which children live is critical in his formation when he reaches adulthood. Considering this, every field of study must be present with a childhood approach to the early stages of life. Hoping to contribute to build better generations. In the area of serious games we have better opportunities, because the games are a major part in the early stages of life and are present regardless of race, culture, religion or social status of the children. The multiple factors involved in shaping the character of children are out of their control. They are so many and varied that it becomes difficult for a child gets out unharmed. Although it is part of the natural formation of each individual processes it is different and therefore, is affected in a special way. This paper presents a serious game prototype used as an aid in the treatment of children with emotional disorder, particularly in children with some sort of depressive behavior.

Published in: 2017 Twelfth Latin American Conference on Learning Technologies (LACLO)

Date of Conference: 9-13 Oct. 2017 INSPEC Accession Number: 17396209

8.5 Publicación de artículo en revista arbitrada

8.6 Glosario de términos que se utilizan en la tesis

Aprendizaje automático
Programa o sistema que desarrolla (entrena) un modelo predictivo a partir de datos de entrada. El sistema usa el modelo aprendido para realizar predicciones útiles a partir de datos nuevos (nunca vistos) obtenidos de la misma distribución que la que se usó para entrenar el modelo. El aprendizaje automático también se conoce como el campo de estudio relacionado con estos programas o sistemas.

Clase
Valor de un conjunto de valores de segmentación enumerados para una etiqueta. Por ejemplo, en un modelo de clasificación binaria que detecta spam, las dos clases son es spam y no es spam. En un modelo de clasificación de clases múltiples que identifica razas de perros, las clases serían *poodle*, *beagle*, *pug*, etc.

Clasificación binaria
Tipo de tarea de predicción que da como resultado una de dos clases mutuamente exclusivas. Por ejemplo, un modelo de aprendizaje

	automático que evalúa mensajes de correo electrónico y da como resultado "es spam" o "no es spam" es un clasificador binario.
Conjunto de datos	Colección de ejemplos.
Exactitud	Fracción de predicciones que se realizaron correctamente en un modelo de clasificación
Entrenamiento de modelos	Proceso mediante el que se determina el mejor modelo.
Métrica	Número de gran interés. Puede optimizarse directamente o no en un sistema de aprendizaje automático. Una métrica que el sistema intenta optimizar se denomina un objetivo.
Modelo de clasificación	Tipo de modelo de aprendizaje automático para distinguir entre dos o más clases discretas. Por ejemplo, un modelo de clasificación de procesamiento de lenguaje natural podría determinar si una oración de entrada está en francés, español o italiano
Normalización	Proceso de convertir un rango real de valores en un rango estándar de valores, generalmente -1 a +1 o 0 a 1. Por ejemplo, imagina que el rango natural de un atributo específico es 800 a 6,000. A través de resta y división, puedes normalizar esos valores en el rango -1 a +1.
Precisión	Métrica para los modelos de clasificación. La precisión identifica la frecuencia con la que un modelo predijo correctamente la clase positiva
Recuperación	Métrica para los modelos de clasificación que responde a la siguiente pregunta: de todas las etiquetas positivas posibles, ¿cuántas identificó correctamente el modelo?