(tarea del PLN conocida como Atribución de Autoría) (Stamatatos, 2009), identificar el estado de ánimo de quien escribe (Mishne, 2005). Además, mediante métodos automáticos que analizan textos también es posible incursionar en tareas como predecir las fluctuaciones en la bolsa de valores (Bollen, Mao, & Zeng, 2011), conocer el perfil del autor (tarea del PLN conocida como Perfilado de Autores), por ejemplo identificación de pedófilos (Escalante, Villatoro-Tello, Juárez, Montes-y-Gómez, & Villaseñor-Pineda, 2013; Villatoro-Tello, Juárez-González, Escalante, Montes-y-Gómez, & Villaseñor-Pineda, 2012) en salas de chat, identificación de usuarios influyentes en redes sociales (Villatoro-Tello, Ramírez-de-la-Rosa, Sánchez-Sánchez, Jiménez-Salazar, Luna-Ramírez, & Rodríguez-Lucatero, 2014; Ramírez-de-la-Rosa, Villatoro-Tello, Jiménez-Salazar, & Sánchez-Sánchez, 2014) e identificación de genero y edad (López-Monroy A. P., Montes-y-Gómez, Escalante, Villaseñor-Pineda, & Villatoro-Tello, 2013), entre otras.

Es importante resaltar que en años recientes, otra tarea relacionada al perfilado de autores que ha adquirido mucha importancia y que realiza análisis detallado de los usuarios de Internet, es la detección automática de la personalidad. La personalidad puede reflejarse, por un lado, mediante las palabras (*lenguaje escrito*) que se usan para describirse o que se emplean en la interacción con otras personas y, por otro lado, a través de la voz o imágenes producidas en una secuencia de video, es decir por medio del analisis del *lenguaje no-escrito* (Argamon, Dhawle, Koppel, & Pennebaker, 2005; Oberlander & Nowson, 2006; Mairesse, Walker, Mehl, & Moore, 2007; Pennebaker, The Secret Life of Pronouns: What Our Words Say About Us, 2011; Batrinca, Mana, Lepri, Pianesi, & Sebe, 2011).

1.1 Relevancia del tema de investigación para la comunidad especializada

Una de las razones por las cuales el interés de la comunidad científica en el estudio de la personalidad se ha incrementado es debido a que la detección de la personalidad en individuos, empleando técnicas tradicionales desarrolladas en el área de la psicología, ha probado ser eficaz en ayudar a identificar e incluso anticipar distitos aspectos en las personas. Así por ejemplo, conocer la personalidad de un individuao ayuda a predecir sus patrones de pensamientos, emociones y comportamiento (Funder, 2001). Así también ha probado ser útil en la detección de aspectos que incluyen la determinación del nivel de bienestar, salud física y mental, calidad de las relaciones interpersonales, involucramiento con la comunidad, nivel actividad criminal e ideología política (Ozer & Martínez, 2006).

Agregado a lo anterior, estimar la personalidad de los usuarios es una tarea relevante en varias áreas de las Ciencias Computacionales. Por ejemplo, en el campo de Interacción Humano-Computadora (HCI) donde por un lado, conocer la personalidad del individuo ayuda a mejorar su experiencia con la computadora; y por otro lado, dotar a la computadora de una personalidad compatible con la del humano produciría una interacción más natural entre las partes (Bickmore & Picard, 2005). En los juegos de computadoras, la noción de personalidad ha sido clave para mejorar la credibilidad de los personajes (André, Klesen, Gebhard, Allen, & Rist, 1999). En otros estudios se ha reportado que dotar de personalidad a robots asistentes ayuda a los humanos asistidos a sentirse en confianza, por ejemplo en (Tapus, Tapus, & Mataric, 2008) reportan que al dotar a un robot asistente de la personalidad del paciente que haya padecido un derrame

cerebral ha apoyado positivamente en su recuperación. En el campo de la educación, los tutores inteligentes podrían ser más efectivos si se adaptan a la personalidad del estudiante (Komarraju & Karau, 2005). Finalmente, conocer la personalidad del usuario es relevante en la tarea de recuperación de información basada en tendencias o gustos particulares (Ghorab, Zhou, O'Connor, & Wade, 2013).

Todo los anterior es posible ya que la personalidad es la combinación de todos los atributos que caracterizan a un individuo de forma única. Estos atributos pueden ser de comportamiento, temperamentales, emocionales y mentales. La personalidad, desde el punto de vista psicológico, captura las características estables de un individuo. Típicamente estas características pueden ser medidas en términos cuantitativos de forma que puedan explicar y predecir diferencias de comportamiento observables (Matthews, Deary, & Whiteman, 2009). Como se puede notar, desarrollar métodos automáticos para la detección de la personalidad tiene implicaciones no sólo en el área de la psicología, donde su impacto radica en proporcionar a los especialistas de mecanismos más económicos para detectar la personalidad; sino también tiene implicaciones positivas en diversas áreas de la computación con miras en el beneficio de los usuarios a través de la mejora tecnológica de los servicios y aplicaciones.

1.1.1 Impacto

En conclusión, podemos decir que conocer la personalidad de los usuarios tiene grandes implicaciones en el campo de la mercadotecnia, pues permitiría a grandes empresas tener publicidad dirigida u orientada a diferentes tipos de usuarios. Agregado a esto, la personalización de interfaces así como los sistemas de recomendación son ejemplos de aplicaciones que podrían verse beneficiados de herramientas que sean capaces de identificar la *personalidad* de los usuarios. En el campo de la salud, éste tipo de herramientas puede ser igualmente útil. El análisis efectivo de la personalidad, realizado por un psicólogo, le ayuda a identificar ciertas patologías en las personas. Así entonces, detectar estas desviaciones de forma no costosa permitiría focalizar la atención de los psicólogos a los sujetos que requieran atención más especializada

1.2 Los rasgos de personalidad y el Big Five

Desde el punto de vista psicológico, el reconocimiento de la personalidad asume que las características estables de los sujetos ocurren en patrones de comportamiento estable que estos sujetos muestran, independientemente de la situación en la que se vean inmersos (al menos hasta cierto punto) (Vinciarelli & Mohammadi, 2014). Los modelos psicológicos más eficientes para medir los aspectos en la vida de las personas están basados en rasgos (Goldberg, 1993). Estos rasgos de personalidad son disposiciones internas que se manifiestan, de cierta forma, en los procesos de pensar, sentir, o actuar frente a situaciones específicas con resultados esperados (Wrzus & Mehl, 2015). Uno de los modelos basados en rasgos es conocido como *Big Five* (BF) o FFM (*Five-Factor Model*). El *Big Five* es el paradigma dominante en la investigación de la personalidad, además es uno de los modelos qué más ha influenciado la psicología actualmente (McCrae R. R., 2002). El modelo *Big Five* consta de cinco rasgos con dos polaridades cada uno (positiva y negativa):

- *Extroversión* (Extroversion) es asociado con la energía, emociones positivas, asertividad, sociabilidad y expresividad; su polo negativo es *introversión*.
- Neuroticismo (*Neuroticism*) es la tendencia a experimentar emociones no placenteras como enojo, ansiedad, depresión o vulnerabilidad. A veces, es referido como su otro: *Estabilidad emocional* que está asociado con controlar el impulso.
- Amabilidad (*Agreeableness*) se refiere a la tendencia de ser compasivo y cooperativo. En su polo negativo hace referencias a la desconfianza y apatía hacia los demás.
- *Responsabilidad* (*Conscientiousness*) es la tendencia a mostrar autodisciplina, actuar de forma leal, alcanzar objetivos y a ser planificador, organizado y confiable. Su polo negativo refiere a comportamientos espontáneos.
- *Apertura* (*Openess to experience*) es asociado con la apreciación a ideas inusuales, personas imaginativas y curiosas. El polo negativo de este rasgo está asociado a seres inflexibles al cambio y poco imaginativos.

Usualmente la personalidad de un sujeto, definida por factores de estos cinco rasgos, se mide mediante la aplicación de baterías (cuestionarios) estandarizadas (Batería BFQ) (McCrae & Costa Jr., 1997). Estos cuestionarios son por lo regular costosos de obtener y en general no son fácilmente reutilizables pues son acotados a regiones geográficas e incluso a factores socio-culturales. Algunos de los cuestionarios más populares para esta tarea son el NEO-Personality-Inventory Revised (NEO-PI-R con 240 preguntas), el NEO Five Factor Inventory (NEO-FFI con 60 preguntas) y el Big-Five Inventory (BFI con 44 preguntas). Algunos enfoques más recientes proponen la realización de ejercicios de escritura en los cuales el sujeto describe algún evento muy particular (es decir, redactan un texto dirigido, el cual tiene un tema muy específico). A través de este escrito se intentan determinar algunos rasgos de personalidad (Mairesse, Walker, Mehl, & Moore, 2007). En ambos casos, es necesario que psicólogos expertos evalúen y analicen los instrumentos (*i.e.*, baterías y ensayos) para determinar los factores de cada rasgos de cada sujeto. Aunque hasta el momento, este proceso es el más efectivo para identificar la personalidad, el mismo representa un proceso lento y costoso, tanto para los psicólogos que evalúan como para los sujetos que son evaluados. Por si fuera poco, dichos mecanismos pueden estar sesgados de alguna manera debido a que un sujeto al saberse evaluado, puede no ser completamente honesto al momento de resolver los ejercicios, lo cual impacta directamente en la correcta determinación de su personalidad.

1.3 Investigación propuesta

El estudio sobre el procesamiento automático de la personalidad se ha dividido en tres problemas principales (Vinciarelli & Mohammadi, 2014): *i)* reconocimiento automático de la personalidad (RAP), que involucra el descubrir la personalidad manifiesta de un sujeto; *ii)* detección automática de la percepción de la personalidad, que tiene que ver con la personalidad que otros perciben de un sujeto (esta puede o no puede ser la personalidad real del sujeto); y *iii)* reproducción automática de la personalidad, que trata de dotar a

agentes o robots con una personalidad particular, usualmente esta característica de los sistemas computaciones (agentes o robots) se imprime mediante síntesis de voz.

Dentro de este proyecto de investigación nos enfocaremos en atacar el problema del *reconocimiento automático de la personalidad* (RAP), mismo que se encarga de identificar la personalidad real de un sujeto. De acuerdo al modelo cognitivo de Brunswik (*Brunswik Lens*) (Brunswik, 1956), la personalidad *real* de un sujeto (la cual es distinta a la personalidad *percibida* por otros) puede ser capturada mediante pistas observables. Estas pistas observables, una vez que son capturadas pueden usarse en el análisis y proceso del RAP. Por otro lado, en el problema de detección automática de la percepción de la personalidad, las pistas observables son primero recibidas y analizadas por terceras personas, y posteriormente las impresiones de estos sujetos son las que se deben analizar; es decir, se analizan las pistas proximales. Este último enfoque resulta costoso pues para determinar la percepción de la personalidad de un sujeto es necesario el juicio de varios jueces.

En general, la comunidad científica del área de Inteligencia Artificial ha abordado el problema de la identificación (o reconocimiento) automático de la personalidad desde dos grandes enfoques. El primer enfoque tiene que ver con la utilización de atributos unimodales para extraer un tipo de *pista observable*. En este primer enfoque se pueden identificar dos líneas, una línea utiliza características que tienen que ver con mediciones de proximidad entre usuarios, tiempo de intervención en conversaciones con grupos de trabajo, etc. Una segunda línea hace uso de la modalidad textual, analizando aspectos asociados al contenido de textos escritos o transcripciones de grabaciones. El segundo gran enfoque utiliza un conjunto de atributos multimodales que tienen como objetivo extraer información de diferentes pistas observables que apoyen de manera más eficaz en la detección de la personalidad. En este sentido, la investigación se ha dirigido principalmente a encontrar correlaciones entre atributos de comportamiento y atributos extraídos de la señal acústica de la voz. En general, dentro de estos dos enfoques (los que utilizan representaciones unimodales y aquellos que hacen uso de atributos multimodales) se han encontrado resultados alentadores; sin embargo, la tarea está lejos de estar resuelta, pues en promedio los resultados reportados varían entre 40% al 70% de exactitud (en experimentos realizados con sujetos de habla inglesa en algunos casos y de habla italiana en otros). Hasta el momento no hemos encontrado trabajos que combinen características de texto producido por los individuos de manera informal (es decir, texto producido en forma libre y no textos sobre ejecicios específicos orientados a la detección de la personalidad), con atributos extraidos de la señal de voz, ni con atributos asociados a las imágenes. Este hallazgo nos da la pauta para dirigir la investigación hacia un enfoque multimodal que incluya tanto el análisis del texto, atributos extraidos de la señales de voz e incluso del video, pues como se ha observado en trabajos previos, cada modalidad aporta información importante para resolver el problema.

1.3.1 Consideraciones sobre la originalidad de la porpuesta

Dado lo anterior, podemos afirmar que la relevancia científica del presente proyecto recae en el hecho de proponer nuevos enfoques que sean capaces de considerar atributos asociados al texto (e.g., estilo, palabras, frases, y/o temáticas), atributos asociados a la

señal de voz (e.g., timbre, frecuencias, turnos, etc.), e incluso atributos asociados a las imágenes (e.g., gestos, movimientos, posturas, etc.) que permitan hacer la asociación efectiva de la información multimodal de un usuario en particular a uno de los rasgos de personalidad concebidos en el *"Big Five"*.

2. Trabajo Relacionado

La tarea del reconocimiento automático de la personalidad identifica los rasgos de un individuo mediante un conjunto de pistas observables de su personalidad. Entre las *pistas observables* que han sido usados para el desarrollo de métodos automáticos se pueden mencionar tres grupos: *i)* texto, ya sea escrito o transcripciones de audio; *ii)* atributos capturados con dispositivos móviles y sensores; por ejemplo, la proximidad que un individuo guarda en un grupo, la cantidad de intervenciones que tiene en una reunión de trabajo, entre otros; y *iii)* comportamiento en redes sociales, por ejemplo en Facebook se puede determinar la cantidad de *likes* que se asignan a actualizaciones de estado o a páginas de eventos, personas públicas, deportes, etc. A continuación se hace una descripción de trabajos que caen en cada uno de estos tres grupos.

2.1 Análisis de Texto

El lenguaje escrito es un buen indicador de la personalidad pues a través de él se puede expresar la forma de pensar o sentir (Pennebaker, The Secret Life of Pronouns: What Our Words Say About Us, 2011). Los primeros trabajos que aparecieron sobre la identificación de la personalidad a través del texto se basaron enteramente en la identificación de ciertos tipos de palabras utilizadas por los individuos. Por ejemplo, en Argamon et al. (Argamon, Dhawle, Koppel, & Pennebaker, 2005) realizaron experimentos con ensayos escritos por estudiantes para determinar la identificación de dos rasgos: *Extraversión* y *Neurotisismo* y asociaron las palabras usadas en estos escritos en cuatro grandes categorías: palabras de función (como artículos y preposiciones), palabras de cohesión (como pronombres), palabras de valoración (términos que evalúan el contexto como aquellas palabras usadas para validar, desear, aprobar, etc.), y palabras de evaluación (términos que se usan para expresar la actitud del escritor sobre lo que escribe).

En el trabajo realizado por Mairesse et al. (Mairesse, Walker, Mehl, & Moore, 2007) se quiere identificar los cinco rasgos del *Big Five* utilizando, en lugar de cuatro, 88 categorías de palabras de la herramienta LIWC (*Linguistic Inquiry and Word Count*) (Pennebaker, Francis, & Booth, Linguistic Inquiry and Word Count, 2001), junto con un conjunto de 14 atributos tomados de la base de datos psicolongüística MRC que contiene estadísticas de uso sobre más de 150,000 palabras. Los resultados que obtuvieron no fueron favorecedores en todos los rasgos, los autores exploraron otros enfoques de clasificación automática además de la clasificación binaria y probaron que los tres modelos de clasificación usados tuvieron resultados estadísticamente significativos en la detección automática de la personalidad.

Existen otros trabajos que han seguido esta línea de utilizar diccionarios de palabras categorizadas en clases que, en cierta medida, puedan estar correlacionadas con todos o algún rasgo de personalidad. Sin embargo, la principal desventaja de este tipo de

investigaciones es precisamente que las palabras consideradas deben tener una entrada en los diccionarios consultados (*e.g.*, LIWC, MRC) y además deben estar clasificadas en alguna de las categorías que cada uno ha establecido. Adicionalmente, la naturaleza estática de estas herramientas, hace que los enfoques basados en ellos no soporten la evolución natural del lenguaje.

Otra línea de investigación que ha surgido se ha centrado en analizar la relación de unidades textuales, ya sea el conjunto de palabras independientes o las secuencias de éstas (*n*-gramas de palabras) a los distintos rasgos de personalidad. En (Oberlander & Nowson, 2006) los autores evitan el uso de recursos sociolinguisticas (diccionarios) y deciden hacer una representación del texto en forma de *n*-gramas de palabras de longitud *n=2* y *n=3*. A pesar de que el conjunto de individuos que formaron el conjunto experimental fue pequeño (de solamente 71 sujetos), la exactitud de su método es elevada (entre 70% y 90% de exactitud, dependiendo del rasgo detectado). Esto es un indicio de que este tipo de representación (no atada a diccionarios) puede ser empleada con buenos resultados. Como hemos mencionado antes, dentro de este proyecto nos interesa proponer métodos que sean flexibles a la evolución del lenguaje, no limitados a condiciones preestablecidas (*e.g.*, ejercicios ex profeso) ni recursos lingüísticos, es decir se propone emplear representaciones basadas en un vocabulario abierto.

2.2 Personalidad en Redes Sociales

Debido al creciente uso de redes sociales, la investigación del reconocimiento de la personalidad ha tomado ventaja de la enorme cantidad de información personal publicada en estas redes. El interés particular de la computación, es determinar si es posible, a través de la forma de usar (o de comportarse en) las redes sociales determinar la personalidad del usuario. Un ejemplo claro de esta tendencia es el trabajo reportando en (Adali & Golbeck, 2012) donde proponen un conjunto de atributos de comportamiento para determinar los diferentes grados de personalidad de usuarios en Twitter. Este conjunto de atributos tienen que ver, en su mayoría, con la forma en que los usuarios interactúan con otros usuarios en Twitter: ancho de su red, contenido de los mensajes, comportamiento de sus amigos, reciprocidad de acciones, el tipo del mensaje (informativo o particular a un usuario). Concluyen que la predicción de la personalidad mediante atributos de comportamiento es comparable con la predicción de usar solamente texto.

En Facebook, se ha tratado de determinar la personalidad de sus usuarios mediante la forma en la que interactúan y cómo se comunican en su red de amigos. Por ejemplo, Celli y Polonio (Celli & Polonio, 2013) realizan un estudio con 25 usuarios de Facebook y el análisis de 5,200 posts. Los autores utilizan 8 características de estilo que extraen de los posts de los usuarios. Además, por cada usuarios analizado extraen el conjunto de amigos para formar una red (o grafo) de donde calculan aspectos como el diámetro de la red, grado de centralidad, coeficiente de agrupamiento, etc. Los resultados que obtuvieron no fueron concluyentes, sin embargo el uso de otros atributos, como los obtenidos de la red, mostró tener ciertas correlaciones con algunos rasgos de personalidad.

Otro trabajo que ha realizado estudio en Facebook es el de Ortigosa et al. (Ortigosa, Carro, & Quiroga, 2014). En este trabajo, los autores intentan probar que los

usuarios con personalidad similar tienen patrones de interacción similares. Para realizar su estudio utilizaron solamente atributos que pudieran describir el comportamiento de los usuarios, por lo que por cada individuo recolectaron el número de amigos, número de posts, si algún amigo les ha escrito o no, etc. Sus resultados muestran que es posible capturar la personalidad a través del comportamiento de los usuarios de redes sociales con la utilización de un pequeño conjunto de atributos. En la misma línea, en (Youyou, Kosinski, & Stillwell, 2015) se analiza si un sistema de clasificación automático puede predecir la personalidad de los usuarios con el análisis de solamente los *likes* que los usuarios proporcionan a páginas de eventos, personajes públicos, música, películas, libros, etc. Los resultados obtenidos son comparables con los resultados obtenidos del consenso de familiares y amigos cercanos de los sujetos.

A pesar que las redes sociales proporcionan una gran cantidad de información, como la explorada por los trabajos descritos en esta sección, ninguno de ellos ha involucrado el texto que los usuarios producen en estas redes sociales. Dado que el uso del texto puede servir para determinar la personalidad de los individuos (como se ha mencionado en la Sección 2.1) nuestra intuición es que al combinarlo con otro tipo de atributos, puede ayudar a enriquecer los modelos automáticos para la detección de la personalidad.

2.3 Trabajos que hacen combinación de tipos de atributos

Las representaciones que utilizan un sólo tipo de atributos se les denomina representación *unimodal*, como es el caso de los trabajos descritos en las secciones anteriores. Los trabajos que se describirán en esta sección han utilizado al menos dos tipos de atributos (es decir, utilizan una representación *multimodal*) para la caracterización de la información producida por los sujetos, con el objetivo de determinar su personalidad.

En (Mairesse, Walker, Mehl, & Moore, 2007) se utilizó un corpus que corresponde a un conjunto de conversaciones de 96 participantes que está compuesto de más de 97,000 palabras y más de 15,000 expresiones. Los atributos extraídos se dividen en 3 tipos: palabras en LIWC, palabras en MRC, y atributos extraídos de la señal de voz (prosodia). Los resultados mostraron que existen correlaciones entre los tipos de atributos y algunos rasgos, es decir, no todos los atributos son útiles para identificar todos los rasgos de personalidad.

En el caso de (Batrinca, Mana, Lepri, Pianesi, & Sebe, 2011) se realizaron experimentos sobre 89 videos de auto-presentaciones hechas por Skype. Los autores analizaron atributos prosódicos tales como estadísticas del pitch, intensidad y duración de los segmentos de voz, pero además, utilizaron características extraídas de la imagen del video, como por ejemplo la dirección de la mirada, postura, movimientos de la cabeza, entre otros. Los resultados reportados varían entre el 65% y 75% de exactitud (dependiendo del rasgo en cuestión). Este mismo conjunto de atributos se ha utilizado en el reconocimiento de los rasgos de personalidad que se manifiestan en tareas colaborativas (Batrinca, Lepri, Mana, & Pianesi, 2012).

El trabajo propuesto en (Kalimeri, 2013) hace uso de atributos extraídos de dispositivos bluetooth que sirven como indicadores de proximidad entre individuos que

los porten; a este conjunto de atributos le agregan los obtenidos por sensores infrarrojos (que proporciona más detalle de la proximidad de los individuos). Además, hace uso de atributos obtenidos de la comunicación vía correo electrónico, por ejemplo, el número de correos que el sujeto envía y recibe, el número de contactos que tiene, la longitud del correo electrónico, entre otros. A pesar de combinar diferentes tipos de atributos, los resultados obtenidos no son significativamente mejores que los reportados por otros trabajos, es decir, no rebasan el 70% de exactitud dependiendo del rasgo detectado.

2.4 Discusión

En esta sección se han presentado trabajos que usan dos enfoques de representación. El primer enfoque es la *representación unimodal* que ha mostrado que una sola modalidad, en particular el texto, captura algunos rasgos de la personalidad de los sujetos, dependido del método de clasificación usado.

Por otro lado, la *representación multimodal* que utilizan los trabajos presentados en esta revisión del estado del arte, utilizan atributos de sensores y de la señal de voz. Bajo la combinación de dos tipos de atributos se ha mostrado que el uso de más de una modalidad puede ser benéfico para resolver esta tarea. Sin embargo, es posible que las modalidades usadas por este segundo enfoque (multimodal) no sean del todo adecuadas, sobre todo por aquellos trabajos que recurren al uso de sensores, lo cual resulta invasivo para los sujetos pues se requiere portar dispositivos especiales para poder obtener los atributos requeridos por estos métodos.

En general, hemos observado que el texto resulta útil en la tarea y que la multimodalidad es pertinente, por lo cual en este proyecto se propone investigar una representación que fusione modalidades no invasivas, tales como el texto, la señal acústica e incluso atributos del video (*i.e.*, atributos extraidos del lenguaje no-escrito), para la identificación automática de la personalidad (vea sección 1.3.1).

3. Hipótesis

"Los rasgos de personalidad son reflejados tanto a través del lenguaje escrito como del no-escrito, y en consecuencia pueden ser proyectados por medio de atributos multimodales"

Agregado a la hipótesis planteada anteriormente, se desea dar respuesta a las siguientes preguntas de investigación.

- ¿Hasta qué grado es posible determinar los rasgos de personalidad empleando lenguaje escrito y no-escrito?
- ¿Con qué efectividad es posible obtener una misma asignación de factor de personalidad empleando una representación multimodal, como por medio de ejercicios de escritura generados ex profeso y/o la aplicación de las baterías estandarizadas?
- ¿Qué tanta información es necesaria conocer respecto a un usuario para determinar efectivamente sus rasgos de personalidad?

4. Objetivos

El objetivo general de esta propuesta es el siguiente:

"Desarrollar un método automático para la categorización de rasgos de personalidad que integre información multimodal, propia de los contenidos digitales generados en formato libre, a través de técnicas de reconocimiento de patrones."

4.1 Objetivos particulares

- Construcción de un corpus multimodal en español (texto, audio y video) de individuos con perfiles de personalidad previamente conocidos.
- Proponer nuevas formas de análisis y representación multimodal de los contenidos digitales para la tarea de categorización automática de rasgos de personalidad.
- Evaluar la clasificación de rasgos de personalidad bajo diferentes modalidades, es decir, representando los documentos digitales sólo con atributos extraídos del lenguaje escrito y/o con atributos asociados lenguaje no-escrito.
- Realizar un análisis detenido de las características de personalidad, identificadas mediante la Batería BFQ "Big Five", así como una descripción de los factores externos relacionados.

5. Metas Científicas y de Formación de Recursos Humanos

El adecuado desarrollo del proyecto permitirá alcanzar metas científicas con importantes contribuciones en el área de Procesamiento del Lenguaje Natural, y en particular en la temática de perfilado de autores. Además se formará un grupo considerable de recursos humanos en diversos niveles.

5.1 Metas científicas

- Generación de un corpus de personalidad multimodal (texto, audio y video) en Español de México. La construcción de este corpus representará un recurso muy valioso para la comunidad científica de PLN en México, el cual podrá ser utilizado para futuros proyectos de investigación.
- Una representación basada en n-gramas de caracteres para el RAP, la cual integre información de contenido y estilo.
- Representaciones multimodales distribucionales para el RAP; primeramente considerando información textual de contenido y estilo, posteriormente, en una segunda etapa, incluyendo información del lenguaje no-escrito, por ejemplo de la señal acústica.
- Nuevos esquemas de pesado multimodales para el RPA, que capturen de mejor manera las pistas observables de los sujetos y valoren elementos de contenido de acuerdo con el estilo y el contexto de sus menciones.
- Un método de predicción de personalidad que sea capaz de determinar el grado de presencia de cada rasgo de acuerdo al modelo establecido en el Big Five.

5.2 Metas de formación de recursos

Se cubrirán tres metas principales bajo este rubro:

- ***Producción de tesis de posgrado:*** Dos estudiantes de doctorado estarán trabajando en temas asociados al proyecto. Se espera que al concluir el proyecto ambos estudiantes se gradúen o esté en proceso de graduación. Por un lado, actualmente la M.C. Gabriela Ramírez de la Rosa, adscrita al posgrado de Ciencias Naturales e Ingeniería en le línea de Computación de la UAM-C ya se encuentra trabajando en la propuesta de representación multimodal. Por otro lado, gracias a la colaboración que existe con el LIDIC de la UNSL en Argentina, recientemente se acaba de iniciar el proceso solicitud de beca doctoral de la C. María Paula Villegas, quien de ser aprobada su solicitud de ingreso al doctorado, trabajará en proponer formas de representación que incorpore atributos léxicos, estilo-métricos, e incluso sociolingüísticos (comportamiento) que permitan evidenciar el uso de lenguaje en tareas de perfilado de autor. Además, se espera incorporar a otros dos estudiantes de maestría en un periodo de tres años. En total, el proyecto producirá al menos 1 tesis de doctorado y 2 de maestría. Aunado a lo anterior otros estudiantes, provenientes de las instituciones participantes, estarían colaborando en el proyecto y se beneficiarían del mismo.
- ***Fortalecimiento de programas de posgrado.*** El fortalecimiento del programa de postgrado de Diseño, Información y Comunicación (MADIC) de la UAM-C al impartir cursos relacionados con el tratamiento automático del lenguaje, así como a través de seminarios impartidos por los colaboradores del grupo de Lenguaje y Razonamiento de la UAM-C, así como de los colaboradores externos (LIDIC e INAOE)**.** Con lo anterior, además de fortalecer al programa de posgrado se fortalecerá la línea de investigación de Procesamiento de Lenguaje Natural y de Reconocimiento de Patrones.
- ***Producción de tesis de Licenciatura***. Inclusión de alumnos de nivel licenciatura en el proyecto. Se considera la participación de al menos 4 alumnos de la Licenciatura en Tecnologías y Sistemas de la Información de la UAM Cuajimalpa. Actualmente la estudiante Janet V. Hernández García, adscrita a la LTSI de la UAM-C, ya se encuentra trabajando en la representación del texto.

6. Metodología

Las hipótesis planteadas en este proyecto serán confirmadas o rechazadas con base en pruebas empíricas. Se requiere de realizar experimentos los cuales se han distribuido entre subgrupos de participantes del presente proyecto. Además de la comunicación permanente entre los participantes del grupo de investigación, se tienen previstas tres reuniones plenarias al finalizar cada cuatrimestre de cada año. En estas reuniones se presentarán los avances, se interpretarán los resultados, y se discutirá la mejor manera de realizar los experimentos posteriores. En las etapas anuales, que a continuación se presentan, se describen las acciones que se llevaran a cabo durante la ejecución del proyecto.

Etapa 1. *Formación de un corpus multimodal etiquetado con rasgos de personalidad para el Español de México. El corpus tendrá como características principales que estará conformado de: i) textos en formato libre (twitter, correo electrónico, etc.); ii) grabaciones*

(video y voz) de auto-presentaciones de los sujetos que formen parte del estudio; iii) resultados de aplicar las baterías estandarizadas del "Big Five"; y iv) información de perfil del individuo (por ejemplo, edad, sexo, nivel socioeconómico, orden de nacimiento etc.).

1. Convocar a una población considerable de estudiantes de nivel Licenciatura (principalmente de la UPAEP y UAM-C) a participar en pruebas de personalidad que involucran atender tres diferentes pruebas.
2. Aplicar las baterías estandarizadas "Batería BFQ" al conjunto de estudiantes que participaran en los ejercicios de identificación de rasgos de personalidad.
3. Definir el medio de comunicación digital por el cual se hará la recolección de textos en formato libre. Inicialmente se piensa en convocar a estudiantes de nivel licenciatura que tengan cuenta de Twitter y que estén dispuestos a compartir la información de sus cuentas. Agregado a esto, se considera mantener como canal de comunicación principal el correo electrónico con la finalidad de obtener otro tipo de textos en formato libre.
4. Se les pedirá a los participantes realizar una auto-presentación ante una video-cámara. Por medio de esta grabación será posible contar con documentos de audio y video, los cuales podrán ser explotados en etapas posteriores.
5. Etiquetado manual de los datos. A pesar de que se emplearán métodos de transcripción automática para obtener textos de las auto-presentaciones, se considera realizar un etiquetado manual de los datos por expertos lingüistas.
6. Realizar una base de datos la cual deberá contener los resultados de todas las pruebas aplicadas al conjunto de estudiantes que completaron las pruebas. Esta base de datos representará en su conjunto al corpus etiquetado con rasgos de personalidad en Español de México, mismo que será puesto a disposición de la comunidad científica interesada en el problema.
7. En el caso de contar con la posibilidad de involucrar a otras instituciones de educación superior, la etapa 1 podría replicarse en otras poblaciones de estudiantes.

Etapa 2. *Definir una nueva forma de representación para las colecciones de documentos a través de atributos multimodales.*

1. Analizar el conjunto de características extraídas del texto que la literatura ha mostrado que son útiles en el reconocimiento automático de la personalidad. Algunos atributos que se tienen identificados atributos que capturen parte de la dependencia de las palabras (n-gramas de palabras de diversos tamaños), y atributos que pueda capturar parte del estilo del autor del escrito y por lo tanto puede ser indicador de algunos aspectos que sean reflejo de la personalidad (n-gramas de caracteres), entre otros.
2. Identificar el conjunto de características extraídas de la señal acústica que se ha utilizado en tareas para el reconocimiento automático de la personalidad.
3. Definir un conjunto de atributos textuales y acústicos que tengan una mejor correlación entre con los cinco rasgos de personalidad del Big Five.
4. Proponer formas de representación multimodal, que fusione atributos textuales y acústicos; y que además deberá considerar la relevancia (o correlación) de cada uno con los rasgos de personalidad del Big Five.

5. Proponer dos modelos de clasificación: uno que considere una clasificación binaria para cada rasgo y un modelo de regresión lineal para predecir los valores exactos de cada rasgo. Para ambos casos, dado que una persona no está limitada a un sólo rasgo de personalidad, sino que puede tener presente más de uno en diferente grado, se analizará la mejor configuración de combinar los 5 modelos resultantes (uno por rasgo) para presentar un resultado único.
6. Diseñar un conjunto de experimentos orientados a validar tanto la representación propuesta como el modelo generado para la detección de la personalidad de forma automática. La evaluación considerará un corpus etiquetado manualmente y estará sujeta no solo a la exactitud, como la mayoría de los trabajos descritos previamente, sino también se reportarán medidas de precisión, recuerdo, f-score y error cuadrado.
7. Se considera realizar evaluaciones en al menos dos corpora, uno en Inglés que haya sido usado por otros trabajos a fin de realizar comparaciones entre los métodos propuestos, y un segundo corpus que apoye la investigación nacional, que sea en español y con sujetos de nuestro entorno social.

Etapa 3. *Medir la correlación entre los atributos más discriminativos identificados en la etapa anterior y las características más importantes identificadas por medio de aplicar las baterías estandarizadas en combinación con la información de perfil de usuario, el lenguaje escrito y lenguaje no-escrito*.

1. Realizar un análisis cualitativo y cuantitativo de la correlación que la representación propuesta y el método de clasificación automático desarrollado tiene con los diferentes aspectos del perfil de los sujetos.
2. Evaluar el desempeño de la representación multimodal propuesta en otras tareas de interés en el área de psicología, por ejemplo determinar el nivel de bienestar de las personas.
3. Evaluar el aporte de representaciones multimodales distribucionales para la tarea de identificación de la personalidad; en primer lugar, una técnica que considere información textual de contenido y estilo, posteriormente, incluir información del lenguaje no-escrito, por ejemplo atributos extraídos de la señal acústica. Como punto de partida se está considerando el trabajo propuesto en (Cabrera, Escalante, & Montes, 2013).
4. Proponer nuevos esquemas de pesado multimodales para el reconocimiento automático de la personalidad. Interesa que dicho esquema de pesado permita capturar más efectivamente las pistas observables de los sujetos y valoren elementos de contenido de acuerdo con el estilo y el contexto de sus menciones (López-Monroy A. P., Montes-y-Gómez, Escalante, & Villaseñor-Pineda, 2014).
5. Evaluar y validar estas formas de representación construidas en la tarea de identificación de personalidad.

7. Grupo de Trabajo

El presente proyecto será realizado principalmente por los miembros del grupo de Investigación de Lenguaje y Razonamiento del Departamento de Tecnologías de la

Información (DTI) de la División de Ciencias de la Comunicación y Diseño (DCCD) de la Universidad Autónoma Metropolitana Unidad Cuajimalpa (UAM-C). Sin embargo, se contará con la valiosa colaboración de investigadores de grupos de reconocido prestigio. En lo que respecta al área de Tratamiento Automático de Documentos se contará con la participación del Laboratorio de Tecnologías del Lenguaje del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). En lo que respecta al área de Inteligencia Artificial e Inteligencia Colectiva se tendrá el apoyo del Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC) ubicado en la Universidad Nacional de San Luis, Argentina. En cuanto al área de Diagnóstico y Rehabilitación Neuropsicológica se contará con la participación del grupo de Psicología del Departamento de Ciencias y Humanidades de la Universidad Popular Autónoma del Estado de Puebla (UPAEP). Cabe señalar que la responsabilidad de la administración del proyecto recaerá completamente sobre el grupo de investigadores del DTI de la UAM-C, especialmente sobre el Dr. Esaú Villatoro Tello (responsable técnico de esta propuesta).

7.1 Integrantes

Del grupo de Investigación de Lenguaje y Razonamiento:

- Dr. Esaú Villatoro Tello (responsable de la coordinación del proyecto, y tendrá participación activa en todas las etapas del proyecto así como colaborará en la dirección de tesis de estudiantes asociados al proyecto). Profesor Titular C de la UAM-C, SNI C, especialista en la identificación de perfiles de usuario por medio de estrategias de aprendizaje que máquina a través de formas de representación que capturan etilo y contenido de los textos. Para más información ver: http://ccd.cua.uam.mx/~evillatoro
- Dr. Héctor Jiménez Salazar. Profesor Titular C de la UAM-C, SNI I, especialista en diversos temas de Procesamiento de Lenguaje Natural, Recuperación de Información y desambiguación del sentido de las palabras.

Del Laboratorio de Tecnologías del INAOE:

- Dr. Luis Villaseñor Pineda. Investigador Titular del INAOE, SNI II y miembro regular de la Academia Mexicana de Ciencias. Es especialista en recuperación de información en documentos multimodales (imagen y/o texto y/o audio), procesamiento de señales con especial énfasis en análisis y reconocimiento de fonemas. Para más información ver: https://ccc.inaoep.mx/~villasen

De la Facultad de Psicología de la UPAEP:

- Dra. Verónica Reyes Meza. Profesora/Investigadora de la UPAEP, SNI I. Es especialista en la línea de Neuropsicología y tiene amplia experiencia aplicando y evaluando la batería que utilizaremos en este proyecto ya que evaluó la personalidad de 500 sujetos durante su proyecto de postdoctorado.

Del Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC) de la Universidad Nacional de San Luis, Argentina:

- Dra. Leticia Cecilia Cagnina. Investigador asistente del LIDIC en la Universidad Nacional de San Luis, Argentina. Amplia experiencia en el área de la optimización

mono y multi-objetivo a través de heurísticas de inteligencia colectiva, además de tener múltiples publicaciones relevantes en el área de representación y clasificación de textos cortos. Para más información ver: https://sites.google.com/site/lcagnina/

7.2 Experiencia del equipo de trabajo

Como se ha mencionado antes, el proyecto de investigación aquí propuesto cae dentro del área de Procesamiento del Lenguaje Natural o Lingüística Computacional, y considera específicamente el problema de clasificación automática *no temática* de documentos[3], en particular al problema conocido como *perfilado del autor* (Author Profiling en Inglés). Cabe señalar que esta problemática ha sido motivo de diversos trabajos y proyectos dentro del Grupo de Lenguaje y Razonamiento de la UAM-C, entre los que destacan los siguientes:

- Towards Automatic Detection of User Influence in Twitter by means of Stylistic and Behavioral Features. Gabriela Ramírez-de-la-Rosa, Esaú Villatoro-Tello, Héctor Jimenez-Salzar, Christian Sánchez-Sánchez. In the 13th Mexican International Conference on Artificial Intelligence MICAI 2014. Tuxtla Gutierrez, Chiapas, November 2014. Lecture Notes in Artificial Intelligence, LNAI Vol. 8856, pp. 245-256, Springer 2014 (ISSN: 0302-9743)
- UAMCLyR at RepLab 2014: Author Profiling Task. Notebook for RepLab 2014 at CLEF 2014. Esaú Villatoro-Tello, Gabriela Ramírez-de-la-Rosa, Christian Sánchez-Sánchez, Héctor Jimenez-Salzar, Wulfrano A. Luna-Ramírez, and Carlos Rodríguez-Lucatero. In Proceedings of the Fifth International Conference on the CLEF Initiative, Sheffield, UK. September 2014.
- Sexual Predator Detection in Chats with Chained Classifiers. Hugo J. Escalante, Esaú Villatoro-Tello, Antonio Juárez-González, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, WASSA 2013, NAACL-HLT 2013, pp. 46-54. June 2013.
- UAMCLyR at RepLab 2013: Profiling Task. Esaú Villatoro-Tello, Carlos Rodríguez-Lucatero, Christian Sánchez-Sánchez, and A. Pastor López-Monroy. In Working Notes of the CLEF 2013 Conference and Labs of the Evaluation Forum, Valencia, Spain. September 2013.
- INAOE's participation at PAN'13: Author Profiling task. A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo J. Escalante, Luis Villaseñor-Pineda and Esaú Villatoro-Tello. In Working Notes of the CLEF 2013 Conference and Labs of the Evaluation Forum, Valencia, Spain. September 2013.
- A two-step Approach for Effective Detection of Misbehaving Users in Chats. Esaú Villatoro-Tello, Antonio Juárez-Gonzalez, Hugo J. Escalante, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda. In Working Notes of the CLEF 2012 Conference and Labs of the Evaluation Forum, Rome, Italy. September 2012.

[3] Entiéndase por *documentos,* documentos de texto, audio y/o video indistintamente.

Así mismo es importante mencionar que esta propuesta es una consecuencia natural de un proyecto previo conducido por el Dr. Esaú Villatoro Tello, responsable del grupo de Lenguaje y Razonamiento de la UAM-C. Dicho proyecto llevó por título "Identificación de Depredadores Sexuales en Conversaciones de Chat Incorporando Información Secuencial" y fue patrocinado por medio de la convocatoria de Apoyo a la incorporación de Nuevos Profesores de Tiempo Completo SEP-PROMEP 2013, con número de proyecto SEP-PROMEP UAM-PTC-380/48510349. Es importante mencionar que éste proyecto estuvo encaminado al desarrollo de herramientas y recursos lingüísticos básicos para el tratamiento del lenguaje escrito, en particular para el problema de perfilado de autores, específicamente orientado a la detección de perfiles pedófilos. De la misma manera el presente proyecto será antecedente de iniciativas futuras relacionadas con el acceso y organización de información producida por los usuarios en los actuales medios de comunicación multimedia.

8. Infraestructura Disponible

El presente trabajo se inscribe dentro de la investigación que realiza en el grupo de investigación de Lenguaje y Razonamiento del DTI de la UAM-C. Alrededor de este grupo de investigación trabajan seis investigadores con experiencia en las diversas líneas que abarca el proyecto. Además, el grupo de investigación cuenta con diferentes recursos propios de esta línea de investigación que se han adquirido y desarrollado a lo largo de otros proyectos. Es el caso de los recursos lingüísticos tanto para el tratamiento del lenguaje hablado como escrito, así como herramientas especializadas en el área. Por supuesto, el proyecto también busca ampliar esta infraestructura. Por otro lado, se cuenta con equipo de cómputo con capacidades y características especiales para estas tareas. Sin embargo, cabe mencionar, para este proyecto será necesario adquirir nuevas máquinas de escritorio y equipo de cómputo portátil las cuales se dedicarán exclusivamente al proyecto. Esto se debe principalmente, a la necesidad de integrar a nuevos estudiantes en el desarrollo de este proyecto y de desplazamiento de los miembros del mismo.

Por otro lado, es importante resaltar los servicios de apoyo a la investigación con los que cuenta el Departamento de Tecnologías de la Información de la UAM-C al cual está adscrito el grupo de Investigación de Lenguaje y Razonamiento. Entre ellas se cuenta con:

- *Servicios de biblioteca*. La UAM-C cuenta con servicios de préstamo interbibliotecario y de búsqueda bibliográfica. Un propósito importante del proyecto es acrecentar el acervo bibliográfico en temas de procesamiento del lenguaje natural y aprendizaje automático.
- *Servicios de cómputo*. Se cuenta con personal calificado en soporte e instalación de equipo de cómputo y paquetería. Además se cuenta con una red local para compartir los diferentes recursos del laboratorio (impresoras, servidores, etc.) y también se tiene acceso a la red institucional con salida a Internet, indispensable para los estudios que se desean realizar.
- *Servicios administrativos*. Se cuenta con servicio secretarial y de apoyo en las diversas tareas necesarias para la administración de proyectos de investigación.

9. Programa de Actividades

Se contempla que el proyecto puede concluirse satisfactoriamente en un periodo de a lo más tres años. Esta sección presenta el plan de actividades propuesto para alcanzar los objetivos planteados en la sección 4.

Para facilitar la asociación entre actividades y personas participantes en las diferentes etapas del proyecto se optó por utilizar una tabla, en la cual se usaron las siguientes abreviaturas para referirse a cada uno de los participantes: **EV** (Dr. Esaú Villatoro Tello, UAM-C, responsable de la coordinación del proyecto, y tendrá participación activa en todas las etapas del proyecto así como colaborará en la dirección/co-dirección de tesis de estudiantes asociados al proyecto); **HJ** (Dr. Héctor Jiménez Salazar, UAM-C); **GR** (Mtra. Adriana Gabriela Ramírez de la Rosa, estudiante de doctorado de la UAM-C); **JH** (Janet V. Hernández García, estudiante de Licenciatura en Tecnologías y Sistemas de Información (LTSI) de la UAM-C); **LV** (Dr. Luis Villaseñor Pineda, investigador titular del INAOE); **VR** (Dra. Verónica Reyes Meza, profesor/investigador de la UPAEP); **LC** (Dra. Leticia Cagnina, investigador del LIDIC en la UNSL Argentina); **MV** (María Paula Villegas, estudiante de doctorado de la UNSL, Argentina); **EL1**, **EL2**, **EL3** (estudiantes de licenciatura 1, 2 y 3 cuyos temas de tesis serán asociados a la carrera de LTSI); **EM1** y **EM2** (estudiantes de maestría por definir 1 y 2, uno de los temas de tesis estará asociado a la línea de psicología, mientras que el otro a líneas afines a las TIC); **PR** (Programador de tiempo completo); **ET1 y ET2** (lingüistas etiquetadores).

		Cuatrimestres									
	Actividades	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Participantes**
1ª Etapa	Puesta en marcha de la convocatoria a una población considerable de estudiantes de nivel Licenciatura (principalmente de la UPAEP y UAM-C) para participar en pruebas de personalidad	X									EV,VR,EM1
	Aplicar las baterías estandarizadas "Batería BFQ" al conjunto de estudiantes que participaran en los ejercicios de identificación de rasgos de personalidad	X	X								EV,VR,EM1
	Construcción de las herramientas necesarias para hacer la recolección de la información disponible en los medios de comunicación digital establecidos (Por ejemplo cuentas de Twitter, Facebook y/o de correo electrónico).		X	X							EV,VR,PR
	Grabación, edición y almacenaje de las auto-presentaciones de los sujetos involucrados en el estudio.		X	X							EV,VR,LV,EM1

	Etiquetado manual de los datos obtenidos de las grabaciones por medio de métodos automáticos (e.g. métodos de transcripción automática) y etiquetado manual con ayuda de expertos lingüistas.		X	X							EV,VR,LV,ET1, ET2
	Construcción de la base de datos con todos los elementos necesarios que el corpus construido deberá contener. Puesta a marcha del servidor de datos en el cual se podrá compartir la información con los participantes del proyecto y con la comunidad científica interesada en el área.		X	X							EV,LV,PR
	Replicar la metodología de recolección de datos de personalidad en una comunidad de estudiantes distinta a la UPAEP. En principio se considera replicar las técnicas empleadas con la comunidad de la UAM-C. Tener datos de estudiantes de otro sector socio-económico permitirá validar la robustez de los métodos propuestos. La descripción de la metodología seguida para la construcción del corpus, junto con un análisis detallado de varios aspectos contenidos en el mismo se considera será material suficiente para una publicación de revista indexada.		X	X	X						EV,VR,LV,PR,GR, EM1
2ª Etapa	Se realizará un análisis del conjunto de características extraídas del texto que la han demostrado ser útiles en el RAP. Se valorará la importancia de atributos que capturen parte de la dependencia de las palabras (n-gramas de palabras de diversos tamaños), y atributos que pueda capturar parte del estilo del autor el cual puede ser indicador de algunos aspectos que sean reflejo de la personalidad (n-gramas de caracteres). De este análisis detallado se considera poder tener material suficiente para una publicación en congreso.			X	X						EV,HJ,GR,MV,JH, GR,EM1
	Realizar experimentos con el conjunto de características extraídas de la señal acústica que se ha utilizado en tareas para el RPA. El análisis de la importancia de las características extraídas de la señal acústica en la solución del problema planteado se considera material suficiente para una publicación en congreso.			X	X						EV,LV,GR,EL1

	Identificar aquellos atributos (textuales y acústicos) con mejor correlación a los diferentes factores de personalidad definidos bajo el modelo del Big Five.				X						EV,LV,HJ,LC,GR
	Implementar una forma de representación multimodal, que fusione atributos textuales y acústicos; y que además deberá considerar la relevancia (o correlación) de cada uno con los rasgos de personalidad del Big Five				X	X					EV,HJ,LC,GR
	Construcción de dos modelos de clasificación: uno que considere una clasificación binaria para cada rasgo y un modelo de regresión lineal para predecir los valores exactos de cada rasgo. Para ambos casos, dado que una persona no está limitada a un sólo rasgo de personalidad, sino que puede tener presente más de uno en diferente grado, se analizará la mejor configuración de combinar los 5 modelos resultantes (uno por rasgo) para presentar un resultado único.					X					EV,LC,GR,EM2
	Diseñar un conjunto de experimentos orientados a validar tanto la representación propuesta como el modelo generado para la detección de la personalidad de forma automática. La evaluación considerará un corpus etiquetados manualmente y estará sujeta no solo a la exactitud, como la mayoría de los trabajos descritos previamente, sino también se reportarán medidas de precisión, recuerdo, f-score y el error cuadrado. El resultado de el trabajo realizado respecto a la representación multimodal se planea publicar en una revista indexada.					X	X				EV,HJ,LC,JH,GR, EM2
	Validar los métodos propuestos en al menos dos corpora, un corpus en Inglés que haya sido usado por otros trabajos a fin de realizar comparaciones entre los métodos propuestos (por ejemplo, el corpus liberado durante el PAN@CLEF 2015), y un segundo corpus que apoye la investigación nacional, que sea en español y con sujetos de nuestro entorno social. Este trabajo podrá ser presentado en un congreso.					X	X		X	X	EV,HJ,LV,LC,EM2 ,EL2

3ª Etapa	Realizar un análisis cualitativo y cuantitativo de la correlación que la representación propuesta y el método de clasificación automático desarrollado tiene con los diferentes aspectos del perfil de los sujetos, por ejemplo cual es el grado de asociación dependiendo el orden de nacimiento, el genero y/o edad. Dependiendo de los resultados, se considera esto dará material para una publicación de congreso.					X	X	X			EV,VR,HJ,LV,LC, GR,EM2
	Evaluar el desempeño de la representación multimodal propuesta en otras tareas de interés en el área de psicología, por ejemplo determinar el nivel de bienestar de las personas. Dependiendo de los resultados, se considera esto dará material para una publicación de congreso.						X	X			EV,VR,LC,MV,EL3
	Evaluar el aporte de representaciones multimodales distribucionales para la tarea de identificación de la personalidad; en primer lugar, una técnica que considere información textual de contenido y estilo, posteriormente, incluir información del lenguaje no-escrito, por ejemplo atributos extraídos de la señal acústica. Como punto de partida se está considerando el trabajo propuesto en (Cabrera, Escalante, & Montes, 2013). Dependiendo de los resultados, se considera esto dará material para una publicación de congreso.						X	X			EV,HJ,LV,MV
	Proponer nuevos esquemas de pesado multimodales para el reconocimiento automático de la personalidad. Interesa que dicho esquema de pesado permita capturar más efectivamente las pistas observables de los sujetos y valoren elementos de contenido de acuerdo con el estilo y el contexto de sus menciones. Para esto se está considerando partir de las técnicas de análisis semántico conciso (López-Monroy A. P., Montes-y-Gómez, Escalante, & Villaseñor-Pineda, 2014).							X	X		EV,LC,LV,MV
	Evaluar y validar estas formas de representación construidas en la tarea de identificación de personalidad. Con el resultado del análisis de estos experimentos se considera tener material suficiente para una revista indexada.								X	X	EV,LC,LV,HJ,MV

Publicación de Resultados			X		X	X		X	X	TODOS
Tesis de Licenciatura/Maestría/Doctorado		X	X			X	X	X	X	TODOS

10. Presupuesto

10.1 Gasto corriente:

Concepto	Cantidad	Justificación	1er año	2do año	3er año
Acervos bibliográficos	Alrededor de 20 libros	Libros relacionados a los temas del proyecto: procesamiento de lenguaje natural, minería de textos, aprendizaje computacional, etc. Se ha considerado un precio promedio de $1,500 pesos por libro, incluyendo gastos de envío.	$30,000	----	----
Formación de recursos humanos	Becas para alumnos de Licenciatura y una extensión de beca para un alumno doctoral	Se apoyará a los estudiantes de Licenciatura que estén realizando su tesis en temas afines al proyecto con una beca correspondiente a un salario mínimo general vigente en el DF por un periodo máximo de hasta 6 meses c/u. Extensión de beca para un alumno de doctorado (en caso de no terminar su tesis durante la vigencia de su beca) por hasta 6 meses: $12,000 por mes.	$12,618	$25,236	$84,618
Cuotas de inscripción	2 congresos por año	Las inscripciones a congresos internacionales se consideraron a $11,000. Esto debido a los costos actuales de congresos de interés como: COLING (540 EUROS); CIARP (400 USD); ACL (785 USD); CICLING (675 USD); MICAI(520 USD).	$22,000	$22,000	$22,000
Estancias participantes	2 estancias, de a lo más 15 días, de participante mexicano en Argentina	Durante el desarrollo del proyecto se tienen contempladas dos estancia de investigación con duración máxima de 15 días para participantes del proyecto, en la Universidad de San Luis Argentina, con la Dra. Leticia Cagnina. ($20,000 avión; $15,000 hospedaje; $8,000 comida; $2,000 transp. Local)	$45,000	$45,000	-----
Estancias visitantes	2 estancias, de 15 días cada una, de part. argentino en UAM-C e INAOE	Durante los tres años del proyecto se espera la visita del Dra. Leticia Cagnina y de la estudiante María Paula Villegas ($20,000 avión; $15,000 hospedaje; $8,000 comida; $2,000 transp. Local)	----	$45,000	$45,000
Pasajes	2 congresos internacionales por año	Para participación en 2 congresos europeos y 2 en USA. Boleto de avión a Europa se estima en $20,000 y para USA en $13,000. Se considera $3,000	$36,000	$36,000	$36,000

		por viaje de transporte local.			
Materiales	Varios	Consumibles, expansiones de memoria, fuentes de poder, accesorios para respaldo y transferencia de información.	$30,000	----	----
Honorarios por servicios profesionales	1 Programador por un año; 2 etiquetadores por 6 meses.	El programador se encargará del desarrollo de herramienta para facilitar el trabajo del etiquetado, configurará pondrá a disposición el servidor donde se almacenarán los datos del corpus (12 mil por mes). También se consideran 2 personas para etiquetar las transcripciones de los sujetos evaluados (6 mil por mes por persona).	$144,000	$72,000	----
Viáticos	2 congresos internacionales por año	Para los congresos internacionales se calculó una estancia de una semana (estancias de menos días provoca que el costo del boleto se incremente excesivamente). Considerando un costo por noche de 125 euros por 7 días y suponiendo el cambio a $20 pesos por euro para el hospedaje se necesitan $17,500 pesos. Para alimentos se consideraron 40 euros diarios, en total por los 7 días se estiman $5,600 pesos. En total el costo de los viáticos para un viaje internacional es de $23,100 pesos.	$46,200	$46,200	$46,200
			$365,818	$291,436	$233,818
	Total gasto corriente:		**$891,072**		

10.2 Gasto de inversión:

Concepto	Cantidad	Justificación	1er año	2do año	3er año
Equipo de cómputo	2 equipos portátiles tipo *Tablet*	Este equipo será utilizado para las presentaciones de resultados en las reuniones de trabajo. Además se usaran para aplicar los cuestionarios al momento de estar recolectando el corpus..	$16,800	----	----
	1 estación de trabajo de alto desempeño	Se consideran equipos con al menos procesador Corei7 y 64GB de memoria y 3TB de almacenamiento. Es imprescindible un equipo con estas características para poder aplicar y evaluar los métodos desarrollados en aplicaciones reales.	$68,000	----	----
			$84,800		
	Total gasto inversión:		**$84,800**		

11. Resultados

A continuación se listan los resultados concretos del proyecto por categoría:

- *Científicos*: 6 publicaciones en conferencias arbitradas internacionales (2 por año) y al menos 2 publicaciones en revistas indexadas. En total, al finalizar los tres años del proyecto se tendrán 8 publicaciones como mínimo.
- *Formación de recursos humanos*: 4 estudiantes de Licenciatura en el área de Tecnologías y Sistemas de Información, 2 alumnos de maestría en áreas afines a las TIC's, y dos estudiantes de doctorado en el área de Ciencias Computacionales.
- *Recursos:* Al finalizar el proyecto se tendrá un corpus multimodal (texto/audio/video) de personalidad en Español de México.
- *Divulgación:* A lo largo del proyecto se realizarán seminario de divulgación a través de los cuales se dará a conocer a la comunidad los logros alcanzados del proyecto. Con este tipo de seminarios se espera ayudar a fortalecer las líneas de conocimiento asociadas a las Tecnologías del Lenguaje y de Inteligencia Artificial en general.
- *Vinculación:* Se fortalecerá la relación de trabajo entre los miembros de laboratorios de Tecnologías del Lenguaje en el INAOE así como con investigadores y estudiantes del LIDIC en la UNSL en Argentina.

Bibliografía

Cabrera, J. M., Escalante, H. J., & Montes, M. (2013). Distributional Term Representations for Short Text Categorization. *Proceedings of CICLing 2013, Part II, LNCS , 7817*, 335--346.

Celli, F., & Polonio, L. (2013). Relationships between Personality and Interactions in Facebook. *Social Networking: Recent Trends, Emerging Issues and Future Outlook* (págs. 41--54). Nova Science Publishers, Inc.

Leon-Martagón, G., Villatoro-Tello, E., Jiménez-Salazar, H., & Sánchez-Sánchez, C. (2013). Análisis de Polaridad en Twitter. *In Journal of Research in Computing Science , 62*, 69-78.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H., & Villaseñor-Pineda, L. (2014). Using Intra-Profile Information For Author Profiling. *Proceedings of the PAN 2014.*

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H., Villaseñor-Pineda, L., & Villatoro-Tello, E. (2013). INAOE's participation at PAN'13: Author Profiling task. *Working Notes of the CLEF 2013 Conference and Labs Evaluation Forum.* Valencia, Spain.

Adali, S., & Golbeck, J. (2012). Predicting Personality with Social Behavior. *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, (págs. 302-309).

André, E., Klesen, M., Gebhard, P., Allen, S., & Rist, T. (1999). Integrating models of personality and emotions into lifelike characters.

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. (2005). Lexical Predictors Of Personality Type. *In Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America.*

Batrinca, L. M., Mana, N., Lepri, B., Pianesi, F., & Sebe, N. (2011). Please, Tell Me About Yourself: Automatic Personality Assessment Using Short Self-presentations. *Proceedings of the 13th International Conference on Multimodal Interfaces* (págs. 255--262). ACM.

Batrinca, L., Lepri, B., Mana, N., & Pianesi, F. (2012). Multimodal Recognition of Personality Traits in Human-computer Collaborative Tasks. *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (págs. 39--46). ACM.

Bickmore, T. W., & Picard, R. (2005). Establishing and Maintaining Long-term Human-computer Relationships. *ACM Trans. Comput.-Hum. Interact. , 12* (2), 293--327.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science , 2* (1), 1-8.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments.* University of California Press.

Escalante, H. J., Villatoro-Tello, E., Juárez, A., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2013). Sexual predator detection in chats with chained classifiers. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (págs. 46-54). Association for Computational Linguistics.

Funder, D. C. (2001). Personality. *Annual Review of Psychology , 52* (1), 197-221.

Ghorab, M., Zhou, D., O'Connor, A., & Wade, V. (2013). Personalised Information Retrieval: survey and classification. *User Modeling and User-Adapted Interaction , 23* (4), 381-443.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist , 48* (1), 26-34.

Kalimeri, K. (2013). Towards a dynamic view of personality: multimodal classification of personality states in everyday situations. *ICMI* (págs. 325-328). ACM.

Komarraju, M., & Karau, S. (2005). The relationship between the big five personality traits and academic motivation. *Personality and Individual Differences , 39* (3), 557-567.

McCrae, R. R. (2002). Cross-Cultural Research on the Five-Factor Model of Personality. *Online Readings in Psychology and Culture , 4* (4).

McCrae, R. R., & Costa Jr., P. (1997). Personality trait structure as a human universal. *American psychologist , 52* (5), 509.

Mairesse, F., Walker, M., Mehl, M., & Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)* , 457--500.

Matthews, G., Deary, I., & Whiteman, M. (2009). *Personality Traits.* USA: Cambridge University Press.

Mishne, G. (2005). Experiments with mood classification in blog posts. *1st Workshop on Stylistic Analysis Of Text For Information Access.*

Oberlander, J., & Nowson, S. (2006). Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (págs. 627--634). Association for Computational Linguistics.

Ortigosa, A., Carro, R., & Quiroga, J. (2014). Predicting User Personality by Mining Social Interactions in Facebook. *J. Comput. Syst. Sci. , 80* (1), 57--71.

Ozer, D. J., & Martínez, V. (2006). Personality and the Prediction of Consequential Outcomes. *Annual Review of Psychology* (1), 401--421.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr. , 2* (1-2), 1-135.

Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What Our Words Say About Us.* Bloomsbury Press.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count.* Lawerence Erlbaum Associates.

Ramírez-de-la-Rosa, G., Villatoro-Tello, E., Jiménez-Salazar, H., & Sánchez-Sánchez, C. (2014). Towards Automatic Detection of User Influence in Twitter by Means of Stylistic and Behavioral Features. *13th Mexican International Conference on Artificial Intelligence MICAI 2014. 8856*, págs. 245-156. Chiapas,Mexico: Human-Inspired Computing and Its Applications.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *J. Am. Soc. Inf. Sci. Technol.* , 538--556.

Tapus, A., Tapus, C., & Mataric, M. (2008). User-Robot Personality Matching and Robot Behavior Adaptation for Post-Stroke Rehabilitation Therapy. *Intelligent Service Robotics , 1* (2), 169-183.

Villatoro-Tello, E., Juárez-González, A., Escalante, H., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2012). A Two-step Approach for Effective Detection of Misbehaving Users in Chats. *Working Notes of the CLEF 2012 Conference and Labs of the Evaluation Forum* (págs. 1-12). Rome: CLEF (Online Working Notes/Labs/Workshop).

Villatoro-Tello, E., Ramírez-de-la-Rosa, G., Sánchez-Sánchez, C., Jiménez-Salazar, H., Luna-Ramírez, W., & Rodríguez-Lucatero, C. (2014). UAMCLyR at RepLab 2014: Author Profiling Task. *Working Notes for {CLEF} 2014 Conference* (págs. 1547-1558). Sheffield, UK: CLEF (Online Working Notes/Labs/Workshop).

Vinciarelli, A., & Mohammadi, G. (2014). A Survey of Personality Computing. *IEEE Transaction on Affective Computing* .

Wrzus, C., & Mehl, M. (2015). Lab and/or Field? Measuring Personality Processes and Their Social Consequences. *European Journal of Personality , 29* (2), 250--271.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences , 112* (4), 1036-1040.