Licenciatura en Tecnologías y Sistemas de Información
División de Ciencias de la Comunicación y Diseño
Departamento de Tecnologías de la Información

Identificación de rasgos de personalidad a través del Análisis Automatizado de Textos.

Alumno:
Janet Viridiana Hernández García

Asesores:

Dr. Esaú Villatoro Tello

MC. A. Gabriela Ramírez de la Rosa

Contenido

1. Introducción 1

1.1. Planteamiento del problema 2
1.2. Motivación 2
1.3. Objetivos 3
1.3.1 Objetivo General 3
1.3.2. Objetivos Específicos 3
1.4. Organización del documento 3

2. Marco teórico 4

2.1. Aprendizaje Automático 4
2.1.1. Clasificación Automática de textos 5
2.1.2. Perfilado de Autor 6
2.1.3. Reconocimiento de la Personalidad a través del texto, PRT 7
2.2. Modelos de personalidad 7

3. Trabajo relacionado 10

3.1. Trabajos basados en análisis de texto 10
3.2. Trabajos basados en comportamiento 12
3.3. Análisis multi-modal 13
3.4. Herramientas 16
3.4.1. AnalyzeWords 16
3.4.2. Calcula el Language Style Matching 18
3.4.3. Apply Magic Sauce PredictionAPI 20
3.4.4. MyPersonality 25
3.4.5. Test de personalidad TP2010 26
3.4.6. Herramienta propuesta 28

4. Método 30

4.1. Grafo de n-gramas de caracteres 30
4.2. Medidas de similitud 32
4.3. Construcción de los modelos 33
4.4. Pruebas 36
4.4.1. Conjunto de datos 36
4.4.2. Algoritmos de aprendizaje 37
4.4.3. Evaluación 38
4.4.4. Experimentos y resultados 38

5. Desarrollo del sistema 40

5.1. Esquema general del sistema 40
5.2. Módulo de Extracción de tuits 41
5.3. Módulo de Representación del documento 42
5.4. Módulo de Extracción de atributos 43
5.5. Módulo de identificación personalidad 44
5.6. Módulo de evaluación 45
5.7. Sistema web 46
5.8. Vistas del sistema 46

6. Conclusiones y trabajo futuro 51

7. Bibliografía 52

Apéndices.....54

Apéndice diagramas casos de uso 54
Apéndice diagramas de secuencia 57

Índice de Figuras

Figura 1.Ejemplo de perfilado de autor con las tareas más comunes a clasificar, como conocer el género, edad y personalidad de un autor. 6
Figura 2 .Pantalla principal de Analyze Words 17
Figura 3. Captura de pantalla de los datos arrojados por AnalyzeWords 18
Figura 4. Interfaz de la herramienta LSM 18
Figura 5. Información requerida antes de hacer el análisis. 19
Figura 6. Resultados de LSM 19
Figura 7. Interfaz de Apply Magic Sauce 20
Figura 8. Inicio de sesión por medio cuenta de Facebook y autorización de privacidad 21
Figura 9. Predicción de edad y género. 21
Figura 10. Predicción de preferencias sexuales y estatus de relación. 22
Figura 11. Predicción de Personalidad y de educación 22
Figura 12. Predicción de Orientación Política y Religiosa 23
Figura 13. Test de personalidad en Línea 24
Figura 14. Resultados del test en línea. 24
Figura 15.Interfaz de myPersonality. 25
Figura 16. Captura de pantalla de TP2010 que muestra los datos de personalidad inferida. 26
Figura 17. Comparación entre la personalidad de dos amigos en TP201 27
Figura 18.Recomendador de amigo basado en compatibilidad en TP201 27
Figura 21. Representación en bi-gramas de las palabras "wiki" y "kiwi" respectivamente. 30
Figura 22. Grafo de tri-gama para la cadena de texto "home_phone" 31
Figura 23. Clasificación de un documento usando el modelo de grafos de n-gramas 33
Figura 24. Tipos de ventanas de n–gramas (No simétrica, simétrica y Gauss simétrica normalizada). 34
Figura 25. Diagrama de construcción de los grafos representativos de cada clase 35
Figura 26. Atributos del corpus PAN 2015 36
Figura 27. Esquema general del sistema 40
Figura 28.Módulo de extracción de tuits 41
Figura 29.Módulo de representación del documento. 42
Figura 30.Módulo de extracción de atributos 43
Figura 31.Módulo de identificación de personalidad 44
Figura 32.Módulo de evaluación 45
Figura 33.Página de inicio. 46
Figura 34.Vista cuenta de Twitter del usuario 47
Figura 35.Tuits más recientes del usuario. 47
Figura 36.Resultados de la identificación de personalidad. 48
Figura 37.Vista cuestionario Big Five 48
Figura 38.Validación del cuestionario 49
Figura 39.Resultados del cuestionario 49
Figura 40.Caso de uso 1: Identificar personalidad 54
Figura 41.Caso de uso 2: Evaluación 55
Figura 42.Diagrama de secuencia caso de uso 1 57
Figura 43.Diagrama de secuencia caso de uso 2 57

Índice de Tablas

Tabla 1. Los cinco rasgos de personalidad con sus características lingüísticas y de comportamiento. *9*
Tabla 2.Tabla comparativa de trabajos relacionados *15*
Tabla 3. Tabla comparativa de herramientas mencionadas. *29*
Tabla 4. Resultados obtenidos del experimento 1, edad. *38*
Tabla 5. Resultados del experimento 2, género *39*
Tabla 6. Resultados del experimento 3, personalidad. *39*
Tabla 7. Caso de uso 1: Identificar Personalidad *55*
Tabla 8. Caso de uso 2: Evaluación *56*

1. Introducción

En la actualidad, las Ciencias Sociales se están involucrando cada vez más en lo que se conoce como la "Era de la información". Este fenómeno se ha dado gracias a que el Internet representa una fuente de información muy valiosa, en la cual es posible obtener grandes cantidades de textos escritos en lenguaje natural.

Uno de los medios que generan más contenidos son las redes sociales, las cuales son empleadas por una cantidad considerable de la población a nivel mundial. En un estudio de tendencias de uso de redes sociales del 2014, realizado por la agencia especializada en contenidos, marketing y medios sociales, Kamber [1], se reporta el número de usuarios activos en el mundo de las nueve redes sociales principales, siendo Facebook la red social con la mayor cifra, contando con 1.19 billones de usuarios activos; YouTube 1 billón de usuarios activos; Google+ con 300 millones de usuarios y Twitter con 232 millones de usuarios activos.

Todos estos usuarios a su vez generan grandes cantidades de diversos tipos de contenidos como son videos, fotografías, posts, entre otros. La compañía de análisis de datos, DOMO [2] muestra en una infografía [3] cuánto contenido se genera cada minuto en los medios sociales. Revelando que en tan solo sesenta segundos se generan un aproximado de 72 horas de video en YouTube, se comparten más de 2.4 millones de piezas de contenido en Facebook, se envían más de 277 mil tuits. Además en las plataformas Facebook, Instagram, Pinterest, Flick y Twitter se comparten más de 500 millones de fotos al día [4].

A través de esta información disponible, es que se ha tratado de identificar la personalidad, así como el estado de ánimo de las personas. Algunas áreas que se podrían beneficiar con esto, sería el desarrollo de marketing personalizado, conocer la reputación de marcas y/o personajes públicos, sistemas de recomendación que sean más precisos, identificación de pedófilos, generación de perfiles de usuario, servicios de asistente personal inteligente, entre otros más.

En particular la identificación de rasgos de personalidad tiene una gran importancia desde el punto de vista psicológico [5] y social [6], ya sea en la presentación, tratamiento y principalmente el diagnóstico de los trastornos psiquiátricos, como trastornos de personalidad, e incluso de enfermedades orgánicas, como es el caso de los trastornos psicosomáticos. Además, es importante en la predicción de conductas relacionadas, como son en las relaciones sociales, desempeño escolar y laboral, elección vocacional, entre otras cosas.

Para lograr esto, es necesario construir herramientas que permitan analizar y hacer conclusiones a partir de grandes cantidades de información. Para esto es necesaria la colaboración entre diferentes campos del conocimiento; en particular entre las ciencias sociales y la lingüística computacional.

1.1. Planteamiento del problema

Uno de los problemas que recientemente ha llamado la atención de áreas del conocimiento, tales como las ciencias sociales y la lingüística computacional, debido a su gran variedad de posibles aplicaciones, es la identificación de rasgos de personalidad por medio del texto escrito en lenguaje natural.

En el campo de la psicología se han propuesto mecanismos que permiten conocer el factor de personalidad de un individuo a través de la aplicación de cuestionarios estandarizados. Agregado a esto, algunos ejercicios de escritura que implican la descripción de algún evento muy particular ayudan a determinar con cierta efectividad algunos rasgos de personalidad.

Sin embargo, en los medios que involucran formas de comunicación más impersonales (por ejemplo, comunicación mediada por la computadora), tales como las redes sociales y/o el correo electrónico, el determinar estos rasgos de personalidad se vuelve una tarea más complicada debido a que las señales no verbales que se dan de manera natural en las interacciones cara-a-cara no suelen estar presentes en este tipo de comunicación. De manera similar, el texto producido en estos medios digitales (redes sociales, correo electrónico, entre otros) carece de los atributos que se evalúan tanto en las baterías estandarizadas como en los textos dirigidos, complicando la tarea de identificar el factor de personalidad de un determinado individuo.

1.2. Motivación

En este proyecto se busca identificar hasta qué punto es posible hacer la clasificación automática de rasgos de personalidad empleando técnicas de minería de textos. Se pretende aplicar enfoques que han sido exitosos, que sean capaces de considerar atributos de estilo, palabras, frases, y/o temáticas, extraídas de los mismos textos; que permitan hacer la asociación efectiva del texto de un usuario a uno de los rasgos de personalidad concebidos en el modelo psicológico de personalidad “Big Five”.

El estudio de la personalidad y el lenguaje tiene una gran variedad de aplicaciones potencialmente importantes en el campo de las tecnologías de la información. 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.

1.3. Objetivos

1.3.1 Objetivo General

Desarrollo de una herramienta para la identificación de rasgos de personalidad en textos espontáneos basada en técnicas de clasificación de textos.

1.3.2. Objetivos Específicos

a) Utilizar una representación basada en atributos estilográficos para el problema de identificación de rasgos de personalidad.

b) Evaluar el desempeño de la representación propuesta en datos reales.

c) Construcción de un sistema Web que sea capaz de predecir el o los rasgos de personalidad dado un conjunto de textos de entrada.

1.4. Organización del documento

El contenido restante de este documento se encuentra organizado de la siguiente manera:

En el capítulo 2, Marco teórico, se explican algunos de los conceptos y elementos teóricos que fundamentan este proyecto terminal, como lo son los modelos de personalidad, aprendizaje automático, clasificación automática de textos, perfilado de autor, entre otros, que fueron de gran utilidad para el desarrollo de este sistema.

En el capítulo 3, Trabajo relacionado, se hizo una revisión de trabajos previos donde se habla acerca de los marcadores de personalidad. Aquí se describen los tres enfoques más importantes de los trabajos de investigación que se han realizado en relación a este trabajo y se presenta una tabla comparativa para poder analizar cuáles fueron sus resultados, así como observaciones independientemente del enfoque que se empleó. También se describen algunos de los sistemas y herramientas que se basan en análisis de texto y comportamiento. Por último se describe la herramienta propuesta.

En el capítulo 4, Método, se explica el modelo de grafos de n-gramas de caracteres. Así como las medidas de similitud que se utilizaron en este proyecto. También se explica cómo se llevó a cabo la construcción de los grafos modelos y por último se muestran algunas de las pruebas que se realizaron para clasificar edad, género y personalidad utilizando diferentes modelos de clasificación.

El capítulo 5, Desarrollo del sistema de identificación de personalidad, consta del esquema general del sistema. Se describen cuáles son los componentes del sistema que se creó y se explica el funcionamiento de cada uno de los módulos que se desarrollaron para poder implementar el sistema. Y por último, se muestran las vistas del sistema Web desarrollado y se explican sus componentes. Finalmente, en las conclusiones se habla sobre los resultados obtenidos, las metas logradas y el trabajo a futuro del sistema.

2. Marco teórico

Esta sección tiene como objetivo introducir brevemente al lector con algunos conceptos y elementos teóricos que fundamentan este proyecto terminal. Primeramente, se habla sobre el aprendizaje automático. Posteriormente, se pretende introducir al lector a la tarea de clasificación automática de textos. Luego, se describe la tarea de perfilado de autor, y se describe el reconocimiento de la personalidad a través del texto, que es uno de los puntos clave dentro de este trabajo.

Por último se describe qué son los modelos de personalidad, así como la descripción del que se usará para este trabajo.

2.1. Aprendizaje Automático

El Aprendizaje Automático (AA) o ML[1] es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender como lo hacen los humanos.

Una de las definiciones más usadas [9] de aprendizaje automático:

Se dice que una computadora es capaz de ***aprender*** de una experiencia ***E*** con respecto a una tarea ***T*** y una medida de desempeño ***P***, si su desempeño en la tarea ***T***, medida por medio de ***P***, mejora con la experiencia ***E***.

Por lo tanto, un problema de aprendizaje bien definido requiere que ***T, E*** y ***P*** estén bien especificados.

Hay tres tipos de algoritmos de aprendizaje automático:

1. Aprendizaje supervisado: algún mecanismo externo (como retroalimentación humana) proporciona datos de entrenamiento etiquetados, es decir, se conoce a que clase pertenece cada instancia del conjunto de entrenamiento.

2. Aprendizaje no supervisado: se debe realizar la clasificación totalmente sin hacer referencia a la información externa, donde no se cuenta con datos etiquetados.

3. Aprendizaje semi-supervisado: sólo una parte de los documentos están etiquetados por un mecanismo externo.

El aprendizaje automático tiene una amplia gama de aplicaciones, entre ellas se encuentra la clasificación automática de textos.

[1] Por sus siglas en inglés, Machine Learning.

Para el problema de clasificación se requiere de la disponibilidad inicial de una colección de ejemplos (instancias) donde se conoce a qué clase pertenece cada una. A esta colección se le conoce como conjunto de entrenamiento. A este proceso se le conoce como aprendizaje supervisado, ya que se cuenta con un conjunto de datos etiquetados.

2.1.1. Clasificación Automática de textos

Otro concepto importante en este trabajo, es la clasificación de textos, TC, (también conocida como categorización de textos o ubicación de temas). La cual se define como la actividad de etiquetar un conjunto de textos en lenguaje natural, en categorías temáticas dentro de un conjunto predefinido.

Esta clasificación se puede hacer manual o bien utilizando un algoritmo. Esta última es conocida también como clasificación automática de documentos (o textos), ADC[2] [10].

En la Inteligencia Artificial[3] se han desarrollado varios métodos que se pueden aplicar a diversos tipos de datos, por ejemplo, los métodos de aprendizaje automático. El uso del aprendizaje automático para la clasificación de textos se ha convertido en un enfoque muy utilizado para construir sistemas de categorización de textos [11].

Algunas formas de representación de documentos [12] en el ADC son:

- **Vector de términos:** Es una de las formas de representación más comunes, utilizada para representar cada documento, consiste en un vector con términos ponderados como entradas. Un texto d_j es representado como un vector, donde al conjunto de términos que ocurre al menos una vez en el documento tiene un peso *w*, que representa la importancia del término t_k dentro del contenido del documento d_j.

  El peso *w* se puede calcular de diferentes maneras, estas son las más usadas:

  - Ponderado Booleano: Consiste en asignar el peso de 1 si la palabra ocurre en el documento y 0 en otro caso.

  - Ponderado por frecuencia de término (TF)[4]: En este caso el valor asignado es el número de veces que el término t_k ocurre en el documento d*j*.

  - Ponderado por frecuencia relativa: Este es una variación del anterior, en el cual se combina la TF, con la medida de la frecuencia inversa de documento, IDF[5]. Esta es

[2] Por sus siglas en inglés, Automatic Document Classification
[3] De manera muy general la IA es la ciencia que tiene como propósito modelar formalmente la inteligencia humana.
[4] Por sus siglas en inglés Term Frequency

una manera de medir la "rareza" del término t_k, es decir, si un término se encuentra en todos los documentos dj, entonces este término es incapaz de distinguir entre los documentos, por otro lado, si este término sólo se encuentra en un sólo documento, es un término útil.

- **Bolsa de Palabras (BOW)[6]:** Dentro del área de TC, la BOW es la forma tradicionalmente utilizada para representar los documentos. Este método de representación utiliza a las palabras simples como los elementos del vector de términos.

- **n-Gramas:** Este tipo de representación, utilizado también dentro del área de TC, ha demostrado ser una buena forma de representación, pues compite con la representación por medio de BOW. Hay dos tipos de n-gramas, n-gramas de palabras y n-gramas de caracteres. Una de las ventajas de los primeros es que permite capturar información de contexto, mientras que los últimos capturan información de estilo de escritura. Este método de representación utiliza n palabras, o caracteres consecutivos como los elementos del vector de términos.

2.1.2. Perfilado de Autor

Más allá de la identificación de autor y las tareas de verificación de autor donde el estilo de autores es examinado, la tarea de Perfilado de Autor (AP) consiste en conocer lo más posible acerca de un autor no conocido, extrayendo información de aspectos de perfiles como el género, la edad, la lengua materna, o el tipo de personalidad, por medio del análisis de sus textos publicados [11].

En la Figura 1 se muestra un ejemplo de las tareas más comunes del perfilado de autor. Donde a partir de un texto de entrada de algún autor, se pretende determinar cuál es su género, rango de edad y sus rasgos de personalidad.

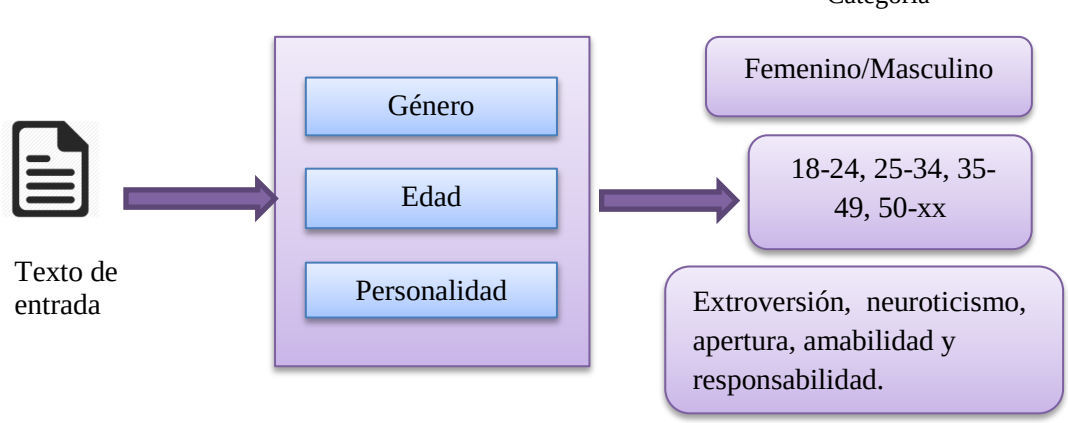

Figura 1.Ejemplo de perfilado de autor con las tareas más comunes a clasificar, como conocer el género, edad y personalidad de un autor.

[5] Por sus siglas en inglés Inverse Document Frequency
[6] Por sus siglas en inglés Bag Of Words.

2.1.3. Reconocimiento de la Personalidad a través del texto, PRT

PRT, por sus siglas en inglés *Personality Recognition from Text*, consiste en la clasificación automática de los rasgos de personalidad de autores, a través de fragmentos de texto que hayan escrito. Esta tarea, que está parcialmente conectada con atribución de autoría, requiere habilidades y técnicas de la Lingüística, Psicología, Minería de Datos y otras áreas más, principalmente de las Ciencias Sociales y la Lingüística Computacional.

Por ejemplo PRT requiere correlaciones entre características del lenguaje y los rasgos de personalidad (proporcionados por los psicólogos), una sólida formación en minería de datos para clasificación, un buen conocimiento de prácticas de comunicación para el análisis social y, lo más importante, un esquema formalizado con el fin de definir clases [13].

El PRT en textos espontáneos, como el que se genera en redes sociales es una tarea que representa todo un reto para las herramientas normales del PLN, ya que las publicaciones en estos medios son frecuentemente muy cortas y con mucho ruido.

2.2. Modelos de personalidad

En psicología, la estructura de la personalidad, se describe en términos de los rasgos (características) las cuales influyen en el comportamiento de una persona, y además capturan las cualidades fundamentales de ésta.

Existen diversos tipos de modelos psicológicos de personalidad [7], los cuales se listan a continuación:

- *Procesuales:* centrados en el estudio de los mecanismos y procesos afectivos y/o cognitivos que determinan la conducta.

- *Estructurales o de rasgos*: centrados en la identificación de los aspectos personales estables y generales que constituyen la estructura básica de la personalidad, cuya identificación permitiría describir, predecir la conducta de los individuos.

- *Factorial biológico*: las bases de las diferencias individuales en personalidad se encuentran en los mecanismos biológicos que sustentan los procesos de aprendizaje, emoción y motivación. Estas dimensiones causan la estabilidad y consistencia de la conducta y los modelos tienen pretensiones explicativas.

- *Factorial léxico:* consideran que el lenguaje recoge todos los términos relativos a diferencias individuales en personalidad, y esos términos son los que deben factorializarse para encontrar las dimensiones básicas. Aún no tienen un desarrollo teórico y explicativo consistente.

El modelo Big Five, es un ejemplo de tipo Factorial Léxico. Se usará el Big Five para este proyecto debido a que es uno de los más reconocidos.

Además, Big Five se adecúa mucho para este tipo de problemáticas de identificar rasgos de personalidad en texto. Esto debido a que se basa en el análisis de los descriptores de la personalidad que aparecen en el lenguaje natural. Estos términos lingüísticos recogen los aspectos descriptivos más importantes de diferencias entre individuos, así como los atributos esenciales de la personalidad.

Las cinco dimensiones de personalidad de este modelo cuentan con dos polos cada una (Neuroticismo, Extraversión, Apertura, Amabilidad y Responsabilidad).

Éstas se suelen denominar tradicionalmente con el acrónimo OCEAN, las cuales se describen en la Tabla 1, donde se presentan algunas de las características del lenguaje y comportamiento para cada uno de los rasgos [8].

Así, se muestran las características lingüísticas y en comportamiento de los polos alto y bajo, de cada uno de los cinco rasgos de personalidad antes mencionados, que se han encontrado en estudios previos [8] de personalidad y lenguaje.

Rasgo	Tipo	Polo alto	Polo bajo
Neuroticismo	Comportamiento	Inestabilidad emocional, ansiedad, hostilidad, depresión, ansiedad social, impulsividad, vulnerabilidad.	Estabilidad emocional, calma, menos fácil de molestar.
	Lingüístico	Uso de la primera persona del singular y palabras de emoción negativas (en ensayos); hablar de discrepancias, trabajos, y estados físicos (en blogs); conectores exclusivos e inclusivos, uso de expresiones múltiples de puntuación (en emails)	Uso de referencias a otras personas (en blogs); uso de más sustantivos y adverbios (email).
Extraversión	Comportamiento	Extrovertidos, cálidos, asertivo, orientados hacia la acción, búsqueda de emociones fuertes.	Introvertidos, bajo perfil, deliberados, fácilmente estimulados.
	Lingüístico	Uso de palabras sociales; palabras de emoción positivas; mayor seguridad (en emails y ensayos); mayor complejidad, conjunciones y adjetivos (en ensayos y email); verbos en tiempo presente.	Uso de negaciones y expresiones de emoción negativa, palabras inclusivas, exclusivas, artículos (en ensayos); mayor tentatividad (en email).
Apertura	Comportamiento	Aprecian las ideas y el arte, imaginativos, consciente de los sentimientos.	Intereses sencillos, conservadores, se resisten al cambio.
	Lingüístico	Uso de artículos, palabras largas y palabras intuitivas.(en ensayos); uso de palabras largas, expresión de sentimientos positivos , y palabras inclusivas (en blogs).	Uso de la primera persona del singular, tiempo presente, y palabras de causa (en ensayos), negaciones, referencias a la escuela (en blogs).
Amabilidad	Comportamiento	Compasivo; cooperativo; considerado; amistoso.	Sospechoso; antipático; cautelosos; antagónico; Poco colaborador.
	Lingüístico	El uso de la primera persona del singular, palabras positivas de emoción (en ensayos).	Uso de artículos, palabras negativas de emoción (en ensayos), discrepancias.
Responsabilidad	Comportamiento	Disciplinado; cumplidor; persistente; compulsivo; perfeccionista.	Espontáneo; impulsivo; logros menos importantes
	Lingüístico	Uso de palabras positivas de emoción (en ensayos)	Uso de negaciones, palabras negativas, palabras exclusivas, discrepancias (en ensayos), temas relacionados con la muerte (en blogs).

Tabla 1. Los cinco rasgos de personalidad con sus características lingüísticas y de comportamiento.

3. Trabajo relacionado

En esta sección se describe una revisión de trabajos de investigación que se han llevado a cabo en relación a nuestra propuesta.

El uso de Internet ha motivado muchas investigaciones en el medio social, el cual se ha convertido en un espacio donde los usuarios generan y comparten contenidos. Los estudios que se revisaron son los que se relacionan con el PRT.

En [8], [19] y [20] se estudian las características lingüísticas de la personalidad. También se vio que hay muchos estudios psicológicos sociales que se relacionan con la detección de rasgos de personalidad basados en la conducta social en línea, donde se pretende demostrar que la personalidad y el comportamiento en línea se relacionan [15]. Algunos de estos estudios se basan en contenido textual, así como meta-información acerca de una persona a través de las redes sociales y otros medios [22], [17], [16].

Asimismo se hallaron estudios donde se investigan diferentes modos de comunicación como son la voz, el habla, los gestos, los movimientos, para explorar relaciones con la personalidad, [18], [21].

3.1. Trabajos basados en análisis de texto

Gracias a la creciente popularidad del uso de las redes sociales han surgido muchos estudios tanto psicológicos, como sociales. Estos se han basado principalmente en Facebook y Twitter, ya que son las redes sociales más populares.

Estos trabajos se han enfocado en determinar el grado con el que es posible correlacionar la información contenida en las redes sociales con algunas características de los rasgos de personalidad.

Uno de los enfoques más utilizados en estos trabajos es el del estudio de las características lingüísticas de la personalidad. Un ejemplo de esto es el trabajo del psicólogo social James W. Pennebaker[7] en su libro *The Secret Life of Pronouns* [19]. La idea principal de éste es que las palabras que se usan para comunicar un mensaje, como lo son las publicaciones en Twitter y Facebook. Éstas revelan no sólo pistas de la personalidad, sino también del estilo de pensar, estado emocional y la conexión con otras personas.

Pennebaker, en colaboración con otros investigadores, ha desarrollado diversas herramientas que permiten el análisis del uso de las palabras en nuestros textos, una de

[7] James W. Pennebaker es un psicólogo social estadounidense. Su investigación se centra en la relación entre el uso del lenguaje natural, la salud y el comportamiento social. Autor o editor de 10 libros y más de 250 artículos, Pennebaker ha recibido numerosos premios y honores. [28]

estas herramientas es AnalyzeWords.com. Este es un sitio experimental que permite hacer un análisis rápido basado en publicaciones de Twitter. En este se hace un análisis de los datos utilizando el programa de análisis de texto LIWC[8]. La idea detrás de LIWC es que las palabras que usamos en un discurso cotidiano, pueden reflejar nuestras emociones. A partir del simple conteo de esas palabras es posible obtener una idea del estado emocional de quien las escribió.

LIWC, es una herramienta de análisis de texto muy utilizada para estudios con enfoques similares [16], [14], [20], [21]. Ésta se centra en las palabras funcionales que, aunque tienen poco significado léxico o tienen significado ambiguo, revelan más información de la personalidad. Las palabras funcionales pueden ser preposiciones, pronombres, verbos auxiliares, conjunciones, artículos, entre otros.

Otro trabajo con una idea similar es el de [8], el cual hace uso de un gran corpus de blogs para explorar la selección de características que permiten identificar información del lenguaje relacionado a la personalidad. Para esta investigación se usó una colección de textos aproximada de 3000 blogueros que escriben regularmente. A cada autor se le aplicó un test de personalidad basado en Big Five. La comparación del conjunto de características de texto se obtuvo a partir de uni-gramas y bi-gramas, así como el uso de TAWC[9], solo que en lugar de marcar la frecuencia de uso, se empleó el booleano, la presencia o ausencia de características que se observaron. Este estudio también tiene el enfoque de qué tanto la estructura del texto y la presencia de palabras comunes son importantes, ya que el mejor desempeño se obtuvo con el uso de bi-gramas.

También, se ha hecho una exploración del lenguaje [20] en textos asociados con depresión y otros más describiendo similitudes y diferencias entre los grupos de personas en función de su uso del lenguaje. Se aprovecha lo escrito por las personas en sus redes sociales para encontrar palabras distintivas, frases y temas, y usarlas como funciones de atributos como el género, la edad, la ubicación o características psicológicas. Tal es el caso del estudio presentado en [15], se analizaron un total de 14.3 millones de mensajes de Facebook obtenidos de 75,000 voluntarios, a los cuales se les pidió que realizaran un cuestionario de personalidad, así como reportaran su género y edad.

Con lo anterior se pudieron generar 452 millones de instancias de n-gramas y tópicos, los cuales se visualizan en grupos separados, presentando los n-gramas para distinguir tanto género, edad y los rasgos personalidad.

[8] LIWC (Linguistic Inquiry and Word Count): Herramienta de análisis de texto diseñado por James W. Pennebaker, Roger J. Booth, y Martha E. Francis. LIWC calcula el grado en que las personas utilizan diferentes categorías de palabras a través de una amplia variedad de textos, incluyendo correos electrónicos, discursos, poemas, etc.

[9] TAWC (Text Analysis and Word Count): Programa de análisis de texto y contador de palabras.

3.2. Trabajos basados en comportamiento

Por otro lado, muchos de los trabajos que existen relacionados a la detección de rasgos de personalidad tienen un enfoque diferente a los anteriores. Éstos usan diversas estadísticas de uso, como el número de seguidores, *likes*, número de amigos, entre otros.

En [15] se hace exploración del lenguaje usando como función la edad, género, y personalidad de un conjunto de datos de publicaciones de Facebook de 75,000 personas, que también realizaron test de personalidad.

Un ejemplo de aplicación en Facebook es *myPersonality* [22], proyecto cuyo enfoque es el de recolectar información de los datos de los usuarios que dieron su autorización para acceder a ésta, así como del llenado de cuestionarios de personalidad. Esta aplicación ofrecía al usuario una explicación detallada de su personalidad.

Actualmente existe el sitio en Internet, el cual permite tener acceso a una base de datos con la información que recolectaron a través de la aplicación *myPersonality*, con el propósito de permitir el intercambio de información con los investigadores.

En [17] el autor analiza la relación entre características de la personalidad, basado en el modelo Big Five y tres estatus públicos en Twitter. Se extrajeron tres datos de los perfiles de twitter: siguiendo, seguidores y listas.

La medición de las características de personalidad fueron realizadas indirectamente, a través de la aplicación de Facebook *myPersonality*. El autor consideró sólo a los usuarios que especificaron sus cuentas de Twitter en sus perfiles de Facebook, contando con un total de 335 usuarios. Su investigación consistió en analizar la relación entre los cinco rasgos de personalidad de Big Five y los cinco diferentes tipos de usuarios de Twitter identificados, que son: oyentes; populares; altamente leídos; y dos tipos de personas influyentes.

Uno de los resultados que encontraron fue que *Apertura* es el rasgo más sencillo de predecir, mientras que *Extraversión* es el más difícil.

Un estudio que también busca encontrar la relación entre características y rasgos de personalidad se presenta en [16]. Este, tiene como objetivo mejorar la predicción de los cinco rasgos de personalidad, reduciendo el error mediante la incorporación de datos reunidos de una combinación de múltiples OSN[10].

En ese estudio se usaron dos OSN: Facebook y LinkedIn. Recolectando datos de 31 voluntarios. Muchas características fueron extraídas de estas OSN. De Facebook se

[10] OSN (Online social Network) : Red social

extrajeron 9 características numéricas de: actualizaciones de estatus, *post* de amigos, amigos, likes en páginas, fotos, *tags*, eventos, grupos, y juegos.

De LinkedIn sólo se recolectó texto, esto se debe a que el texto de esta red social refleja el nivel de experiencia, logros académicos, entre otros, teniendo en cuenta cinco características como son: (1) longitud del texto, (2) número de conectores, (3) número de habilidades, (4) número de palabras positivas y (5) negativas.

A partir de eso se obtuvo que la exactitud de los resultados de la evaluación de la personalidad para apertura, responsabilidad, amabilidad y neuroticismo mejora cuando ambos sets de características tanto de Facebook como de LinkedIn son usadas juntas.

3.3. Análisis multi-modal

Como hemos estado viendo en los enfoques antes mencionados, en la psicología social se ha usado por mucho tiempo el análisis de texto para estudiar las propiedades psicométricas del uso de palabras y la exploración de relaciones entre ciertas dimensiones del lenguaje y personalidad. Sin embargo, casi no se han investigado el resto de las modalidades de este tipo de comunicación, que son la voz, el habla, los gestos, los movimientos, entre otros.

Este es un campo en el cual se está investigando, y una de las mayores dificultades que se presenta es en cómo representar los datos e información para que una computadora los pueda interpretar.

Las señales no verbales transmiten información útil para la predicción de la personalidad humana [15], [16], pero estas aún no han sido exploradas lo suficiente, y esto se puede deber en parte al costo de transcribir la información.

Como ya se ha comentado, Internet ha motivado muchas investigaciones en el medio social. Uno de estos medios que más han sido estudiados son los blogs, debido a su gran popularidad.

Los blogs, son la forma de expresión por excelencia. Sus autores, más conocidos como bloggers o blogueros, expresan sus opiniones, gustos o emociones sobre un tema concreto a través de sus páginas, reflejo de su personalidad y forma de vida. Pero el formato blog da un paso más y abre el camino al videoblog o *vblog*, galería de vídeos que se publican cronológicamente, una o dos veces al mes, por uno o más usuarios. Este formato hace el blog más humano que nunca, dado que sustituye el texto por un rostro. Los servicios de video online más populares son YouTube, Vimeo o Dailymotion.

Uno de los estudios que sostiene que los videos en las redes sociales transmiten enriquecedora información de la personalidad humana se presentan en [18]. En este se menciona que la información de personalidad puede ser codificada en otras dimensiones de comportamiento como el canal verbal, que como se ha dicho, ha sido menos estudiado en trabajos de interacción multimodal.

Este estudio investiga la viabilidad de usar contenido verbal para la predicción de huellas de personalidad de vloggers usando transcripciones manuales y reconocimiento automático de voz, ASR[11].

Se analiza en otros trabajos de enfoques similares, el hecho de que ciertos rasgos de personalidad se predicen más fácilmente que otros. Por ejemplo, con los dos enfoques anteriormente comentados, donde se hace uso de señales verbales, los rasgos de extraversión y neuroticismo se pueden predecir más fácilmente que los demás rasgos.

Y con la combinación de señales verbales y no verbales se puede incrementar el rendimiento de la predicción de personalidad.

En otra investigación [21] se encontró la identificación de la personalidad en conversaciones y texto. Se obtuvieron dos fuentes de información. El primero fueron 2479 ensayos (1.9 millones de palabras, 15,269 expresiones) de 96 estudiantes de psicología, se les pidió que escribieran cualquier cosa que se les viniera a la mente durante 20 minutos.

La segunda fuente de información que se empleó fueron extractos de conversaciones grabadas usando EAR[12] (97,468 palabras, 96 temas, 1,015.3 palabras por tema). Esta es una herramienta para el muestreo de los datos de comportamiento en ambientes naturales. Este corpus contiene tanto extracto de sonido como transcripciones. Permitiendo con esto construir modelos de reconocimiento de la personalidad por medio del habla. En este estudio se hizo una revisión de trabajos previos donde se habla acerca de los marcadores de personalidad.

Según estas revisiones, los psicólogos han documentado la existencia de dichas señales mediante el descubrimiento de correlaciones entre un rango de variables lingüísticas y rasgos de personalidad, a través de una amplia variedad de niveles lingüísticos, incluyendo parámetros acústicos.

Por ejemplo, los hallazgos incluyen que existe una mayor correlación entre extroversión y el lenguaje oral, especialmente cuando el estudio implica una tarea compleja. Los extrovertidos hablan más, más fuerte y más repetitiva, con menos pausas y vacilaciones, tienen tasas más altas de voz, silencios cortos, una producción verbal más alta, un lenguaje menos formal, mientras que los introvertidos usan un vocabulario más amplio.

En la Tabla 2 se hace una comparación entre algunos de los trabajos que se revisaron.

[11] ASR , Automatic Speech Recognition , Reconocimiento automático de voz.
[12] Electronically Activated Recorder (EAR) (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001) http://dingo.sbs.arizona.edu/~mehl/EAR.htm

Trabajo	Descripción	OSN / Fuentes	Enfoques	Observación
[15]	Exploración del lenguaje usando como función la edad, género, y personalidad de un conjunto de datos de publicaciones de Facebook de 75,000 personas.	Facebook	Análisis de texto	- Los voluntarios realizaron un cuestionario de personalidad y reportaron su género y edad. - 452 millones de instancias de n-gramas y tópicos y tópicos LDA[13] como función de género, personalidad y edad.
[8]	Analizar cómo las características con mejor desempeño pueden proporcionar una comprensión más profunda del comportamiento lingüístico de personalidad en línea.	Blogs	Análisis de texto	- Se usó una colección aproximada de 3000 blogueros que escribieron durante varios meses. - Se compara conjuntos de características derivadas de uni-gramas y bi-gramas. Obteniendo mejor desempeño en el uso de bi-gramas.
[16]	Encontrar la relación entre características y rasgos de personalidad, y como este se puede mejorar con la combinación de múltiples OSN	Facebook LinkedIn	Comportamiento	- Extracción de muchos atributos de Facebook como el número de amigos, número de publicaciones, número de Likes, etc. - De LinkedIn sólo se recolectó texto.
[17]	Se analiza la relación entre los cinco rasgos de personalidad de Big Five y los cinco diferentes tipos de usuarios de Twitter identificados, que son: Oyentes; Populares; Altamente leídos; y dos tipos de personas influyentes.	Twitter	Comportamiento	- Se recolectaron los datos de myPersonality [23], considerando solamente a los usuarios que especificaron sus cuentas de twitter, contando con un total de 335 usuarios de Twitter. - Se extrajeron tres datos de los perfiles de twitter: siguiendo, seguidores y listas. - Encontraron que *Apertura* es el rasgo más sencillo de predecir, mientras que *Extraversión* es el más difícil.
[21]	Reconocimiento de rasgos de personalidad tanto en texto como en conversaciones.	Ensayos EAR	Multi-modal	- Se obtuvieron dos fuentes de información: 2479 ensayos de cualquier cosa que se les ocurriera a estudiantes de psicología(1.9 millones de palabras) - Extractos de conversaciones grabadas usando EAR
[18]	Se investiga la viabilidad de usar contenido verbal para la predicción de rasgos de personalidad de vloggers	YouTube	Multi-modal	- Se usaron 442 YouTube vlogs - Transcripciones manuales de voz - Reconocimiento automático de voz.

Tabla 2.Tabla comparativa de trabajos relacionados

[13] LDA (Latent Dirichlet Allocation) es un proceso generativo en cuyos documentos son definidos como una distribución de tópicos, y cada tópico en turno es una distribución de tokens.

Estos trabajos, independientemente del enfoque que manejan, se enfrentan a problemas específicos de la APC. Uno de ellos es que la medición de personalidad de los participantes se vuelve una tarea complicada porque requiere de su cooperación en gran medida.
Además de que se depende de la honestidad de los participantes al momento de responder cuestionarios de personalidad.

Asimismo, la mayoría de los trabajos cuentan con fuentes de información muy pequeñas. Y aún no se ha logrado un buen desempeño en la predicción de los cinco rasgos de personalidad.

3.4. Herramientas

A continuación se describen, brevemente, algunas de las herramientas encontradas que se basan en análisis de texto y comportamiento. También, se mostrará cómo se tiene pensada desarrollar la herramienta propuesta.

La mayoría de estas herramientas están disponibles, actualmente, en Internet, a excepción de *myPersonality* [22], que fue cerrado en 2012 y el Test de personalidad TP2010.

Las primeras dos herramientas que se presentan, no se basan exactamente en predecir la personalidad del usuario pero vale la pena revisarlas ya que buscan pistas de estilo de lenguaje en el texto escrito para revelar información acerca de ciertas cualidades del individuo.

3.4.1. AnalyzeWords[14]

Este sitio experimental analiza los tuits de cualquier cuenta que se proporcione. Donde, según investigaciones en las que se basa [24], las palabras que se usan para comunicar un mensaje, como lo son las publicaciones en redes sociales, en este caso Twitter, revelan no sólo pistas de la personalidad, sino del estilo de pensar, estado emocional y la conexión con otras personas.

En la Figura 2 .Pantalla principal de Analyze Words, se muestra la pantalla principal de Analyze Words, donde se puede ingresar la cuenta de Twitter que los tuits puedan ser analizados.

[14] http://www.analyzewords.com

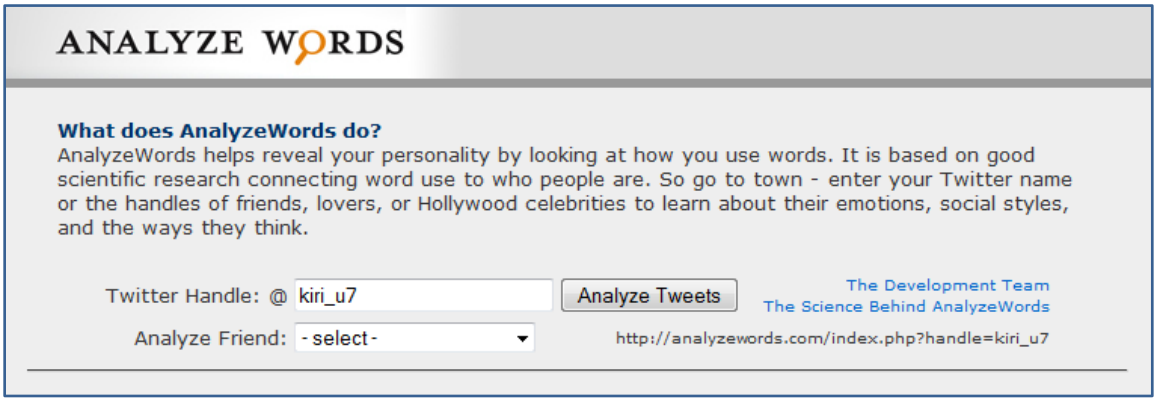

Figura 2 .Pantalla principal de Analyze Words

El sitio nos permite introducir una cuenta de Twitter y poder realizar un análisis de las 748 palabras más recientes hasta la fecha. Los resultados se muestran en la Figura 3Figura 3. Captura de pantalla de los datos arrojados por AnalyzeWords, donde estos resultados se despliegan en tres categorías.

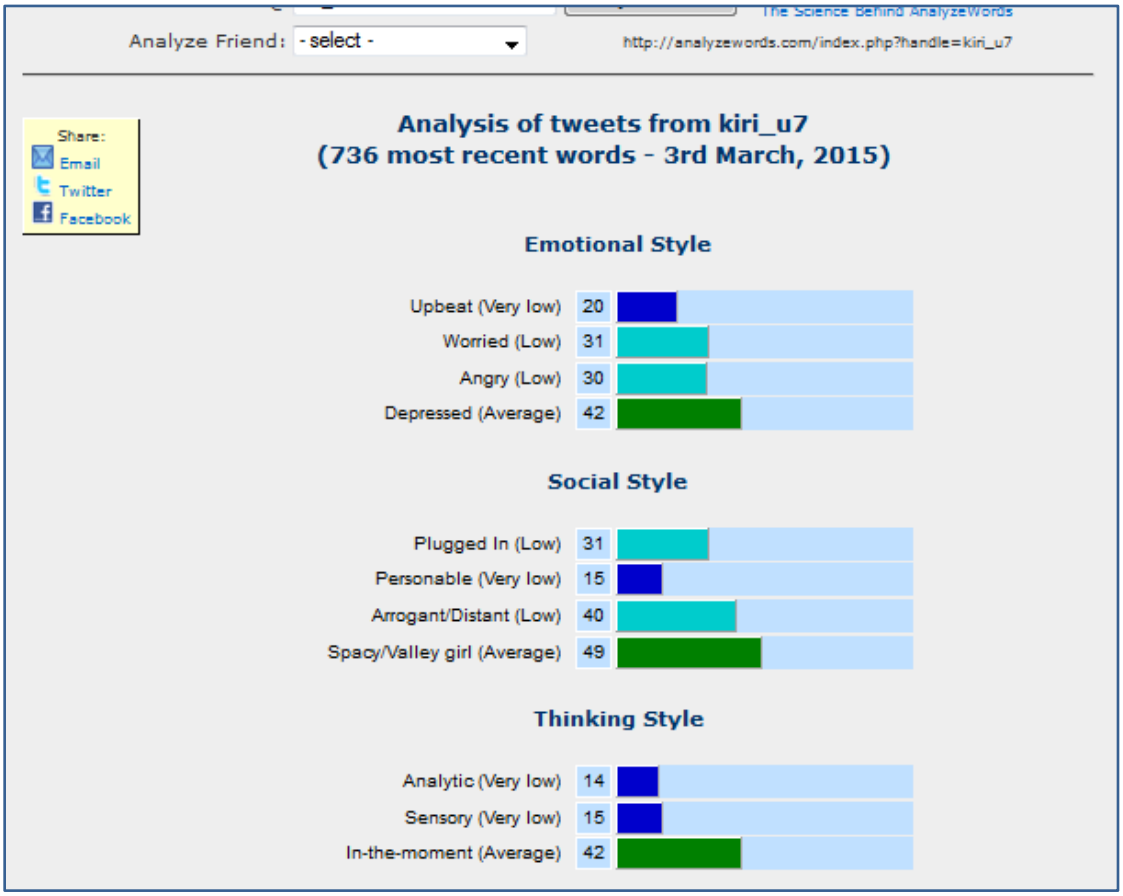

Figura 3. Captura de pantalla de los datos arrojados por AnalyzeWords

En la Figura 3 se muestra como se despliega el análisis que se realizó de la cuenta de Twitter especificada. Se presentan distintas características de personalidad con su puntaje, agrupadas en tres categorías: estilo emocional, estilo social y estilo de pensamiento.

En este caso, en estilo emocional, se obtuvo mayor puntuación para "Deprimido" y la más baja para "Optimista". En estilo social, el tipo "Valley Girl" (popular) tuvo la mayor puntuación seguido de un estilo social "Arrogante/Distante". Y por último, en estilo de pensar el resultado fue "En el momento".

3.4.2. Calcula el Language Style Matching

Language Style Matching (LSM), se refiere a la coincidencia de estilo de lenguaje en conversaciones de texto, email, y otras formas de comunicación interactiva. Calcula la tendencia de los participantes de usar un vocabulario común y con una estructura de sentencia similar. El termino LSM, también llamado coincidencia de estilo, fue introducido por Kate G. Niederhoffer y James W. Pennebaker en su artículo [24] .

En el sitio en Internet *http://www.utpsyc.org/*, está disponible una herramienta que permite comparar dos textos de diferentes personas.

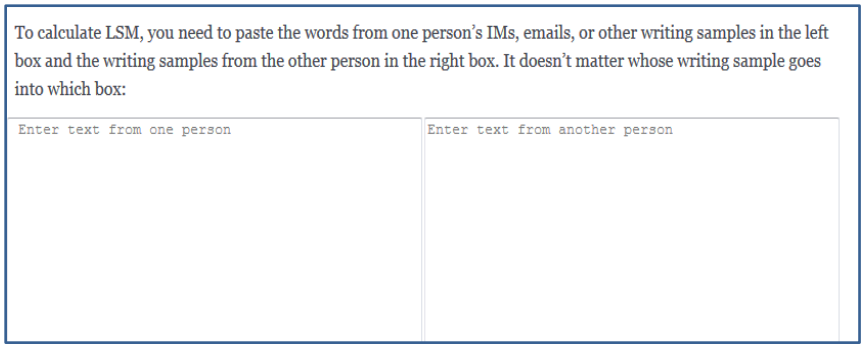

Figura 4. Interfaz de la herramienta LSM

Lo único que se tiene que hacer es copiar dos textos de dos diferentes personas en cada uno de los campos como se muestra en la Figura 4. Estos textos pueden ser extraídos ya sea de mensajería instantánea, emails, conversaciones, etc.

Antes de realizar el análisis se pedirá algunos datos como qué tipo de muestras son, mensajes instantáneos, emails, chats en línea, conversaciones transcritas o ejemplos generales de escritura (ver Figura 5). La relación entre las dos personas, que pueden ser

extraños, amigos, intereses potenciales en amor, novios, enemigos, las mismas personas, compañeros de trabajo, familia, entre otras. El usuario debe elegir qué tan sincronizados cree que están, eligiendo desde completamente conectados, hasta su opuesto de totalmente no conectados. Y por último la edad de las dos personas.

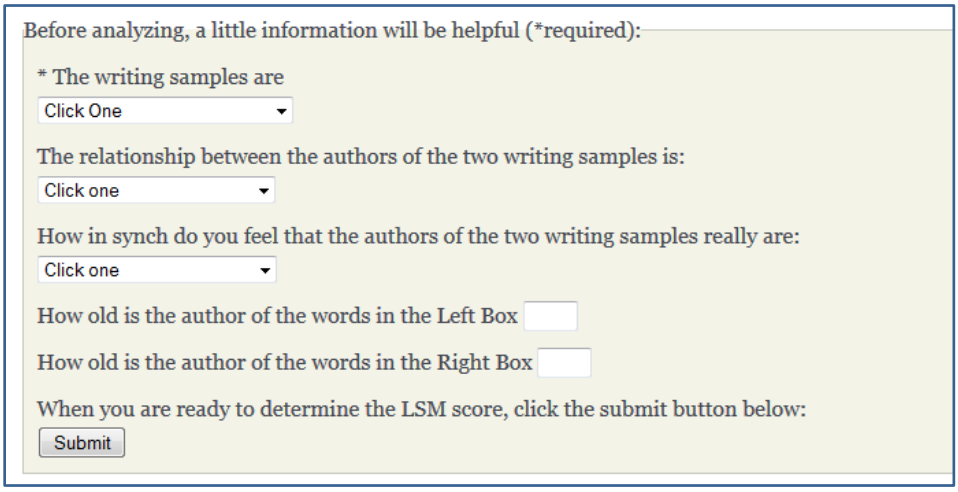

Figura 5. Información requerida antes de hacer el análisis.

Los resultados se muestran mediante una puntuación de similitud del uso de lenguaje de dos personas (ver Figura 6). Dónde se obtuvo un puntaje de LSM de 0.77. Este puntaje indica que las muestras que se dieron para comprar es de dos personas que piensan de manera similar. Entre mayor sea esta similitud mayor es el puntaje de LSM.

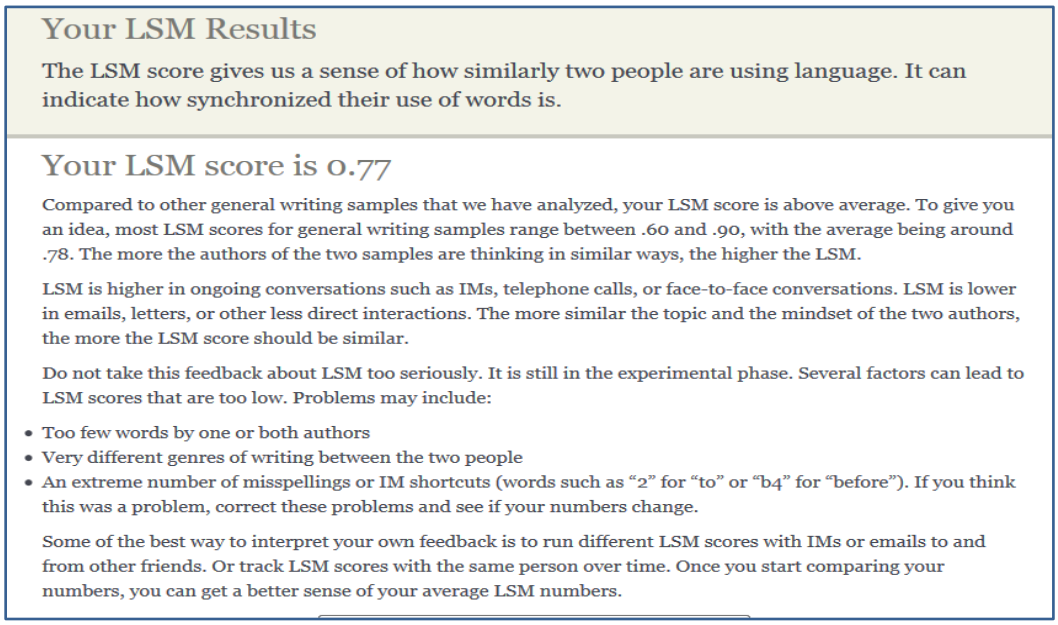

Figura 6. Resultados de LSM

3.4.3. Apply Magic Sauce[15] PredictionAPI

Apply Magic Sauce (AMS) [29], es un servicio que predice características demográficas basado en métodos basados en fingerprint. Esta aplicación permite conocer cuál es la personalidad en redes sociales prediciéndola en base a en los Likes en Facebook.
El motor de predicción fue montado por los investigadores del Centro de Psicometría de la Universidad de Cambridge[16].

Este modelo está basado en el conjunto de datos de *myPersonality* de más de 6 millones de usuarios.

Figura 7. Interfaz de Apply Magic Sauce

Como se muestra en la Figura 7 y la Figura 8, la aplicación pide que se inicie sesión con una cuenta de Facebook y se le dé autorización a la herramienta para a tener acceso al perfil público, lista de amigos intereses y *Me gusta*. Al iniciar, se envían los Likes de Facebook al motor de predicción de AMS.

http://tests.e-psychometrics.com (Test de personalidad tradicional de 10 minutos, para ver si coincide con su auto-percepción)

[16] http://www.psychometrics.cam.ac.uk/

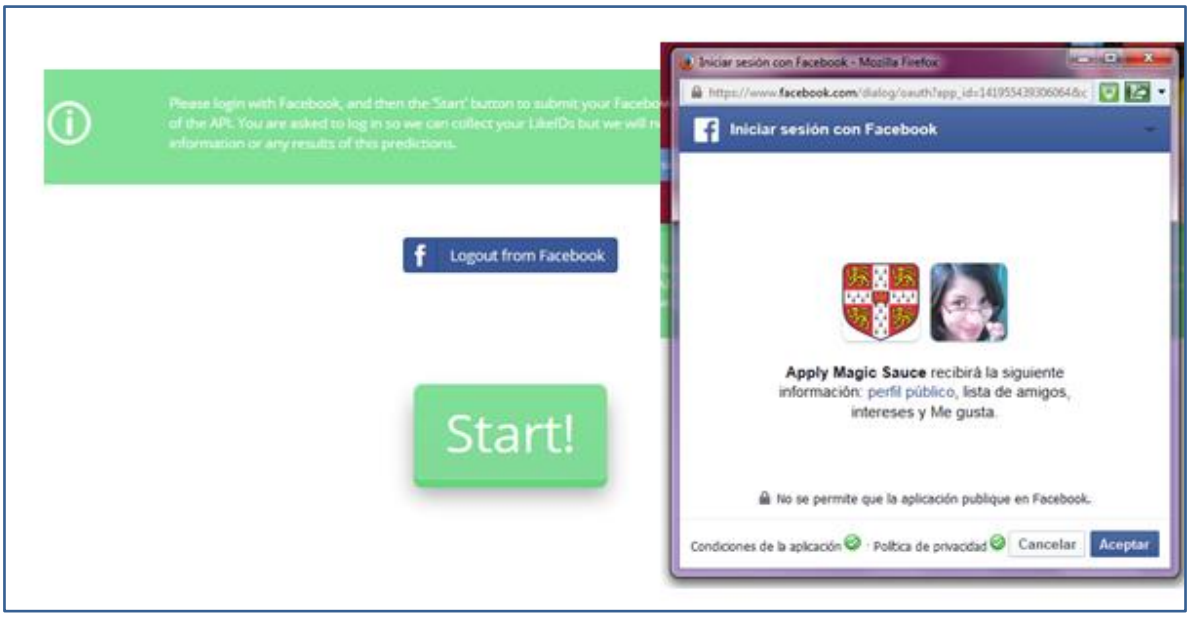

Figura 8. Inicio de sesión por medio cuenta de Facebook y autorización de privacidad

El resultado que arroja son una serie de predicciones basada en los Likes de la cuenta de Facebook. Usa esto para visualizar cómo perciben al usuario los demás en línea. Estás predicciones van desde la edad, genero, personalidad, hasta lo que es la orientación religiosa y política. En las Figuras 9 y 10 se muestran los resultados que se obtienen para cada uno de ellos.

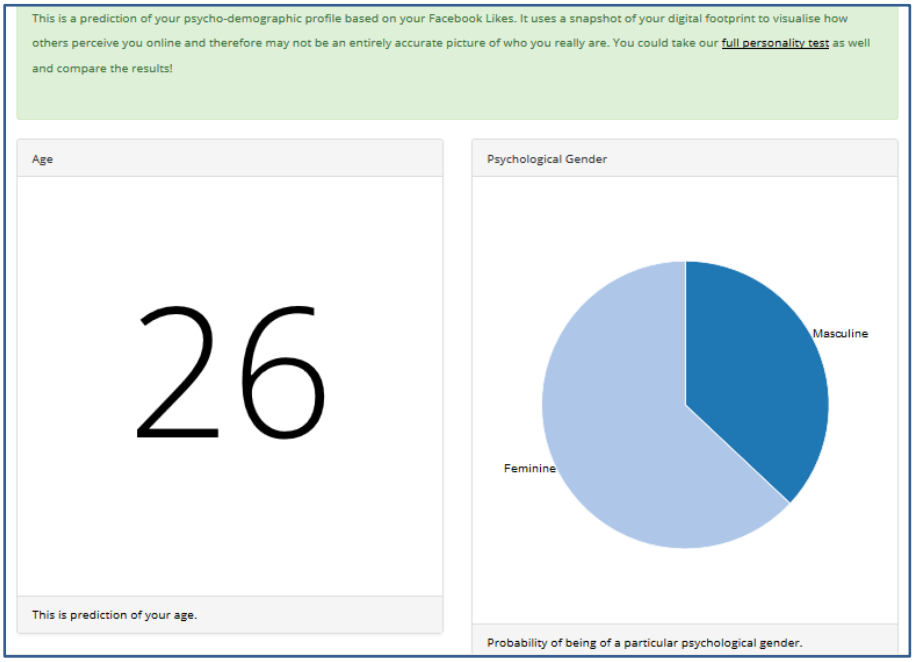

Figura 9. Predicción de edad y género.

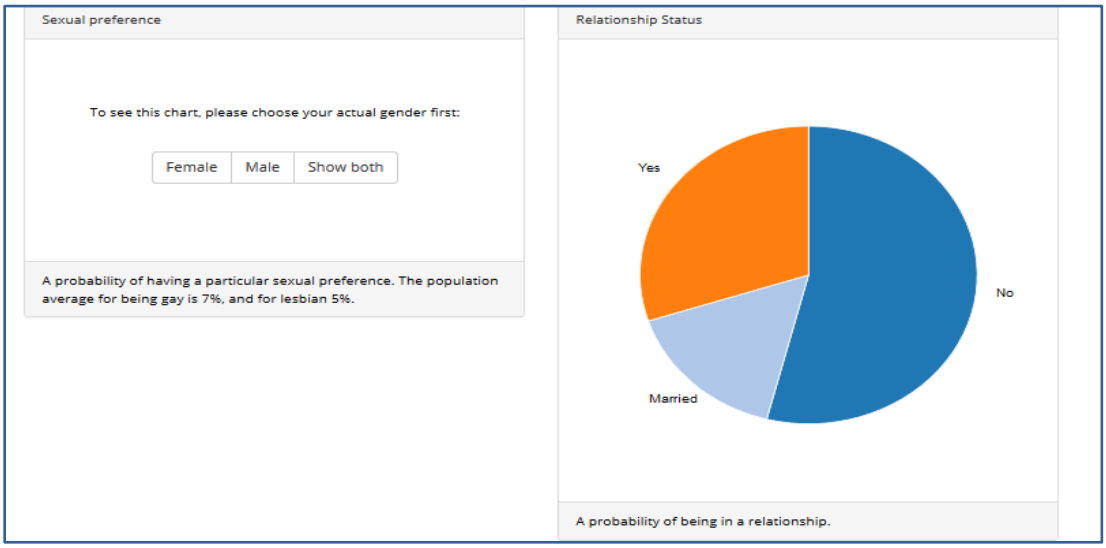

Figura 10. Predicción de preferencias sexuales y estatus de relación.

En la Figura 11, se visualiza información más detallada de la personalidad, pasando el cursor sobre la imagen.

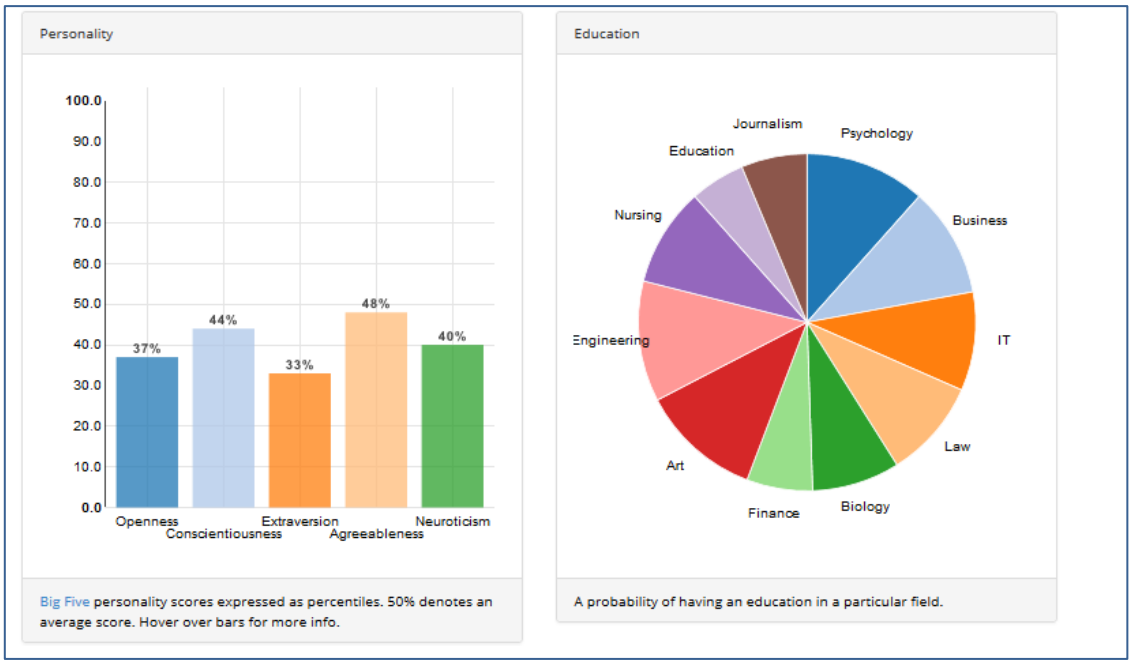

Figura 11. Predicción de Personalidad y de educación

Obteniendo en este caso para cada uno de los rasgos los siguientes resultados:

- Apertura (Openess) 37%: Conservador, tradicional
- Responsabilidad (Conscientousness) 44%: Espontáneo y flexible.
- Extraversión (Extraversion) 33%: Tímido y reservado
- Amabilidad (Agreableness) 48%: Asertivo y competitivo
- Neuroticismo (Neuroticism) 40%: Calmado y relajado

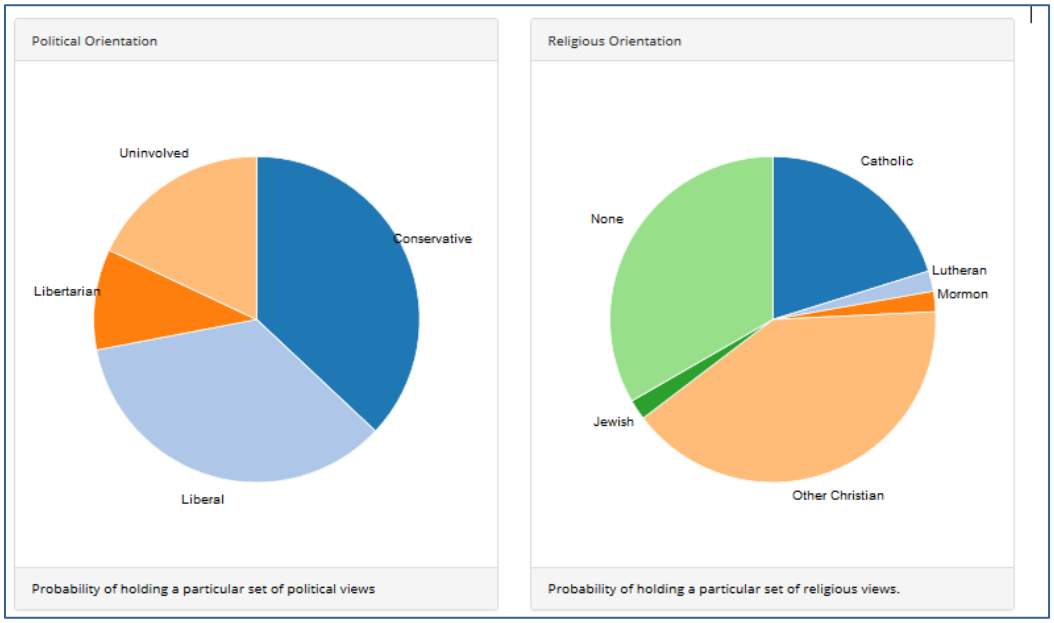

Figura 12. Predicción de Orientación Política y Religiosa

También sugiere realizar un cuestionario de personalidad en línea para así comparar los resultados.

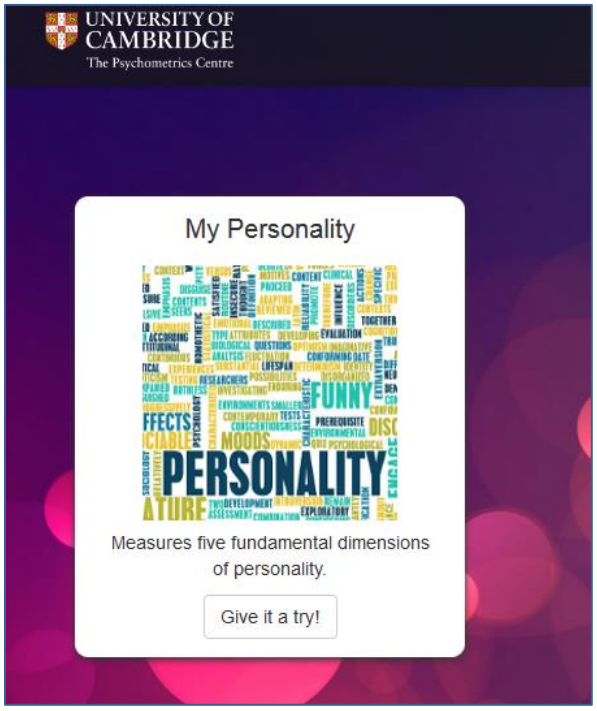

Figura 13. Test de personalidad en Línea

El cuestionario basado en el Big Five al que nos redirige es el que está disponible en el Centro de Psicometría de la Universidad de Cambridge. En la Figura 13 y Figura 14 se muestra la captura del cuestionario y los resultados.

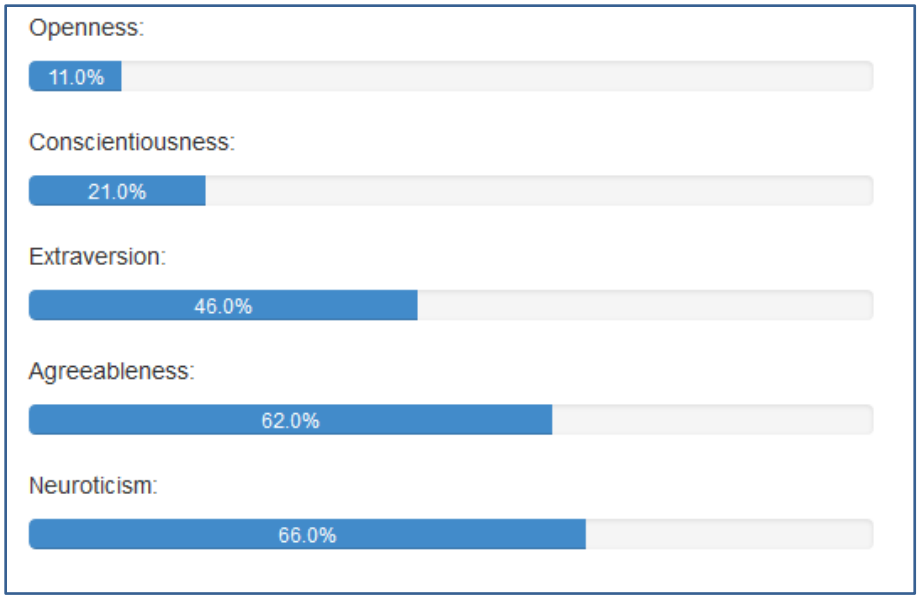

Figura 14. Resultados del test en línea.

3.4.4. MyPersonality[17]

MyPersonality fue una aplicación popular de Facebook que permitió a los usuarios realizar pruebas psicométricas reales, y almacenar, con su consentimiento, su perfil psicológico y perfil de Facebook. En la Figura 15 se muestra la interfaz de la página.

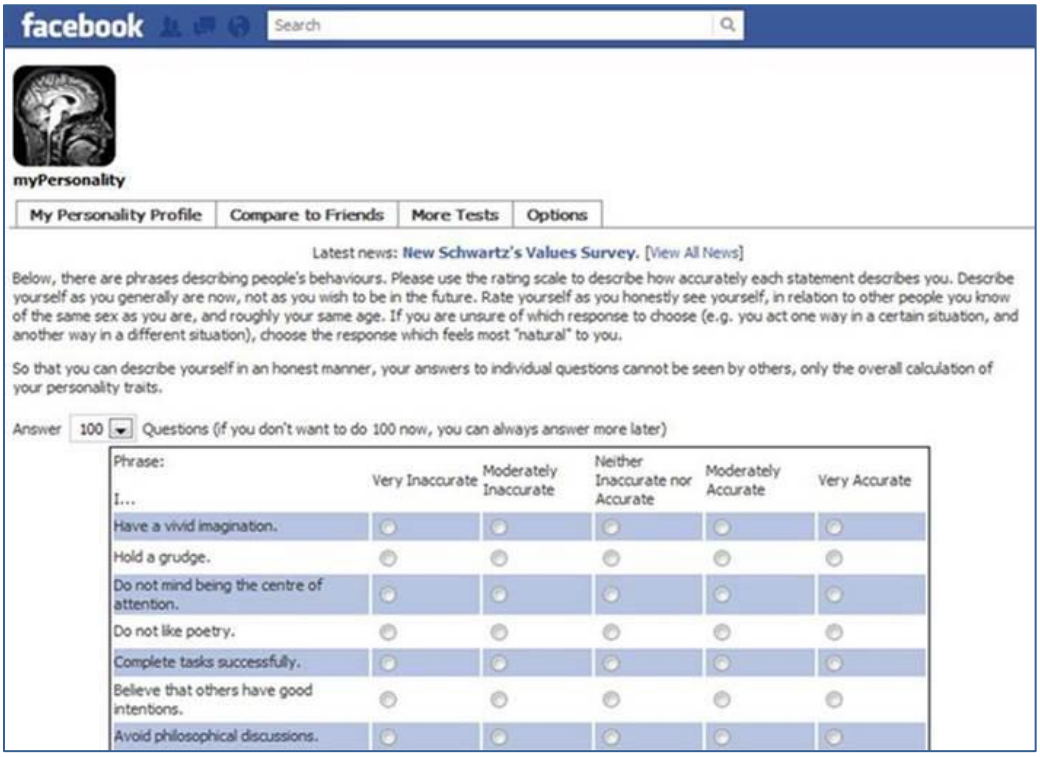

Figura 15.Interfaz de myPersonality.

En la actualidad, la base de datos de myPersonality contiene más de 6.000.000 resultados de las pruebas junto con más de 4.000.000 de perfiles individuales de Facebook. Los encuestados provienen de distintos grupos de edad, orígenes y culturas.

Actualmente cuenta con un sitio en Internet[18], que permite tener acceso a esta base de datos con la información que recolectó a través de la aplicación myPersonality, para permitir el intercambio de información con los investigadores.

[17] http://www.nbcnews.com/health/health-news/can-facebook-assess-your-personality-n284736
[18] http://mypersonality.org/

La información que está disponible para colaboradores registrados, incluye:

- Puntajes de pruebas psicométricas,
- Registros de perfiles de Facebook de los usuarios,
- Los datos de nivel de elemento de prueba,
- Información adicional, por ejemplo, registros de Likes de los usuarios.

Los datos recopilados provienen de más de 86,000 personas que voluntariamente realizaron un test de personalidad usando la aplicación mencionada ofrecida en Facebook desde 2007 hasta 2012. La prueba consistía en 100 preguntas basadas en el modelo Big Five que se ha mencionado con anterioridad. También analizaron los Likes de cada usuario.

3.4.5. Test de personalidad TP2010

TP2010 era una aplicación de Facebook que obtenía información acerca de los usuarios a través de un test de personalidad, así como también de la recolección de datos disponibles de las interacciones de este con la red social. En la Figura 16 se muestra una captura de pantalla del TP2010.

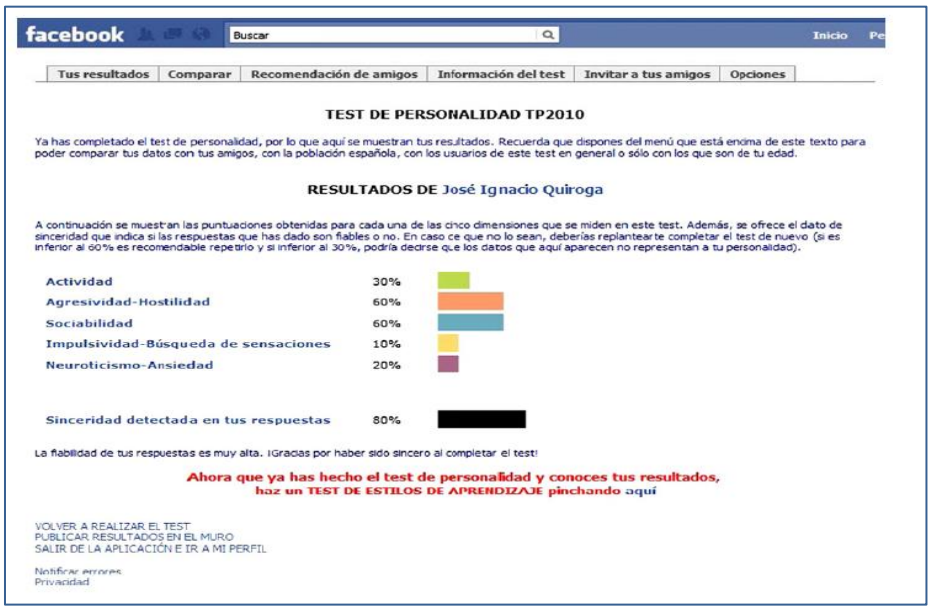

Figura 16. Captura de pantalla de TP2010 que muestra los datos de personalidad inferida.

La meta de TP2010 era descubrir la relación entre los resultados del usuario en la prueba de personalidad y todos aquellos atributos que describen la interacción con Facebook.

Esta aplicación también permitía realizar una comparación de personalidad entre dos usuarios de Facebook, como se muestra en la Figura 17.

Figura 17. Comparación entre la personalidad de dos amigos en TP201

Otra funcionalidad que poseía es la de realizar una recomendación de amigos basados en compatibilidad. Se mostraba a las personas, que también habían realizado el test, y cuyos resultados eran más similares a los de algún perfil. En la Figura 18 se muestra un ejemplo de esto.

Fig. 5. Compatibility-based friend recommender in TP2010.

Figura 18.Recomendador de amigo basado en compatibilidad en TP201

3.4.6. Herramienta propuesta

La herramienta que se propone permitirá que un usuario ingrese su cuenta de Twitter para extraer los tuits más recientes escritos por el usuario. El sistema hará un análisis de su texto, dando como resultado una visualización del porcentaje que obtuvo en cada uno de los cinco rasgos de personalidad. Esta visualización tentativamente se hará por medio de gráficas.

Además, la página permitirá la opción de realizar un test de personalidad en línea. Para poder comparar ambos resultados.

En la siguiente Tabla 3, se presenta una comparación entre las herramientas antes mencionadas. Estas herramientas aún no son muy exactas en sus aproximaciones de predicción de todos los rasgos de personalidad. Ya que aún se sigue investigando cuales son los atributos y características del lenguaje que se relacionan mejor con cada uno de los cinco rasgos de personalidad.

Nombre	Enfoque	Descripción	OSN[19]	Entrada	Salida
AnalyzeWords	Análisis de texto	Análisis de tuits para revelar la personalidad por medio de como usamos las palabras	Twitter	Últimos tuits de un usuario	Estilo emocional, social y forma de pensar del autor de los tuits.
Calculate LSM	Análisis de texto	Permite comparar dos textos de diferentes personas.	Mensajería instantánea	Mensajes instantáneos, emails, u otras muestras de escritura.	Puntuación de similitud del uso de lenguaje de dos personas.
Apply Magic Sauce PredictionAPI	Comporta miento	Predice la personalidad basado en los Likes en Facebook.	Facebook	Likes de alguna cuenta de Facebook	Una visualización de cómo te perciben los demás en línea.
MyPersonality	Comporta miento	Permitía a los usuarios realizar pruebas psicométricas, almacena su perfil psicológico y de Facebook.	Facebook	Cuestionario de 100 preguntas; Likes de Facebook; otros metadatos del perfil de Facebook	Análisis de sus rasgos de personalidad; comparación con otros usuarios.
Test de personalidad TP2010	Comporta miento	Descubrir la relación entre los resultados del usuario en test de personalidad y atributos que describen la interacción con Facebook.	Facebook	Cuestionario de personalidad; Metadatos del perfil de Facebook	Análisis de los cinco rasgos de personalidad; Comparaciones de personalidad entre usuarios; recomendaciones de usuarios basado en compatibilidad
Herramienta propuesta	Análisis de texto	Se extraen los tuits de algún usuario y a partir de este se hará un análisis de la personalidad de la persona, mostrando los resultados	Texto escrito de usuarios	Últimos 200 tuits dela cuenta que el usuario proporcione.	Un análisis de sus rasgos de personalidad, visualizado en gráficas con tu porcentaje obtenido en cada rasgo de personalidad.

Tabla 3. Tabla comparativa de herramientas mencionadas.

[19] OSN (Online social Network) : Red social

4. Método

El presente capítulo tiene el propósito de describir el método empleado para la construcción del sistema de predicción y la representación de los documentos para poder hacer la identificación de personalidad.

En la primera parte de este capítulo se describe en qué consisten los Grafos de n-gramas de caracteres, que es en lo que se basa nuestro sistema.

4.1. Grafo de n-gramas de caracteres

En general el problema de las representaciones basadas en bolsas de palabras es que no incluyen información secuencial, la cual en tareas de clasificación NO-TEMÁTICA puede resultar relevante, pues ayudaría a capturar ciertas características asociadas al estilo de escritura. La idea principal del método que se utilizó propone una representación que aprovecha estas dos grandes ventajas.

Ya que el modelo de la bolsa de caracteres de n-gramas no toma en cuenta el orden de aparición de los caracteres en el texto original, se pierde información valiosa. Como resultado, palabras de documentos con secuencia de caracteres diferentes terminan teniendo la misma o muy similar representación. Por ejemplo la palabra “wiki” y “kiwi” tienen la misma representación en bi-gramas, aunque su significado es totalmente diferente. En la Figura 21 se ejemplifica lo anterior.

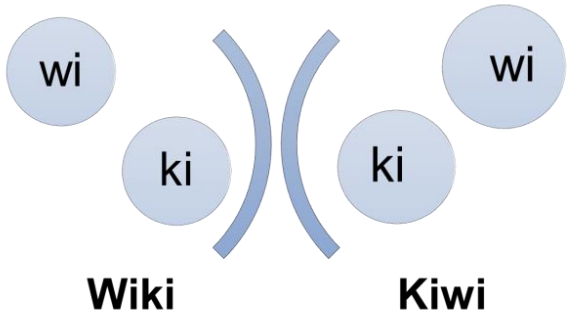

Figura 19. Representación en bi-gramas de las palabras "wiki" y "kiwi" respectivamente.

Se representa cada documento como un grafo, donde cada nodo corresponde a un n-grama específico, y con un peso en sus aristas .y de esta forma se agrega información contextual al

modelo de n-gramas. En Figura 20 se muestra un ejemplo de un grafo de tri-grama para el texto “home phone”

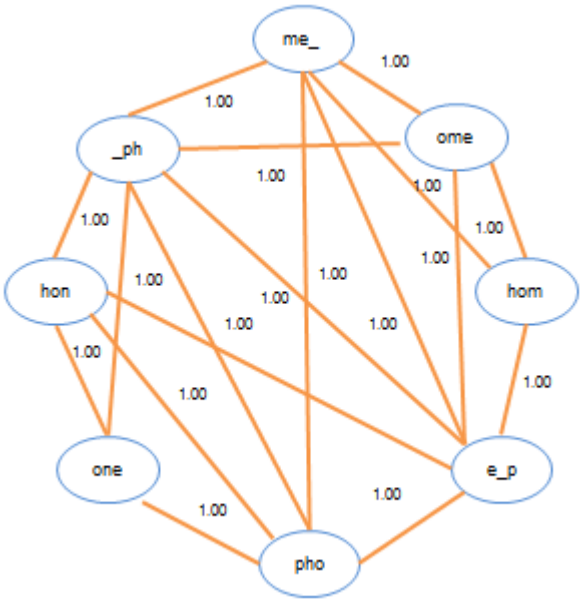

Figura 20. Grafo de tri-gama para la cadena de texto "home_phone"

Grafo de N-grama: *Un grafo de n-grama es un grafo unidireccional*

$G = \{V^G, E^G, W\}$, donde V^G es el conjunto de nodos que son etiquetados por el correspondiente n-grama, E^G es el conjunto de aristas, y W es una función que asigna un peso a cada arista. [14]

El grafo representativo de cada tema se combina en un solo grafo de la clase a través de la función de *Update*. Dada una colección de documentos *D*, un grafo G_D vacío se construye inicialmente, el *i-th* documento $d_i \epsilon D$ posteriormente se transforma en el grafo $G\ d_i$, que se combina con G_D para formar un nuevo grafo G^u_D con las siguientes propiedades: sus nodos incluyen la unión de los nodos de los grafos individuales, y sus pesos se ajustan de manera que converjan con el valor medio de los respectivos pesos. El grafo de clase resultante captura patrones comunes en el contenido del tema entero, como de recurrencia y secuencias de caracteres vecinos.

4.2. Medidas de similitud

La similitud entre documentos y temas es calculada a través de la cercanía de su representación de grafos [26]. A continuación se describen las medidas de similitud que fueron usadas.

1. Containment Similarity (CS), expresa la porción de nodos de un grafo G^i, que son compartidos con un segundo grafo G^j. Asumiendo que G es un grafo de n-gramas, e es un nodo de un grafo de n-grama, su función $u(e, G) = 1$, si y sólo si $e \in G$, y de otro modo equivale a 0.

$$\mathrm{CS}(G^i, G^j) = \frac{\sum_{e \in G^1} \mu(e, G^j)}{\min(|G^i|, |G^j|)}$$

Donde |G| indica el número de nodos del grafo G (es decir, el tamaño del grafo de n-grama).

2. Size Similarity (SS), Indica la relación de tamaños de dos grafos:

$$SS(G^i, G^j) = \frac{\min(|G^i|, |G^j|)}{\max(|G^i|, |G^j|)}$$

3. Value Similarity (VS), la cual indica cuantos de los nodos contenidos en el grafo G^i son contenidos en el grafo G^j, considerando también los pesos de los nodos que coinciden. En esta medida, cada nodo que coincide e teniendo un peso $W^i(e)$ en el grafo G^i contribuye a la suma VR(e)/max($|G^i|$, $|G^j|$) , donde VR(e) (es decir, Value ratio, valor de relación) es un factor escalable simétrico que es definido como VR(e) $VR(e) = \frac{\min(w_e^i, w_e^j)}{max(w_e^i, w_e^j)}$, con el fin de obtener valores en el intervalo de [0,1]. Los nodos que no coinciden no contribuyen a VS: $w_e^i = 0$ por un nodo $e \notin G^i$. Con todas estas medidas juntas se obtiene:

$$VS\left(G^i, G^j\right) = \frac{\sum_{e \in g^1} \frac{\min(w_e^i, w_e^j)}{\max(w_e^i, w_e^i)}}{\max(|G^i|, |G^j|)}$$

4. Normalized Value Similarity (NVS), este aumenta el valor de similitud VS, haciendo caso omiso del tamaño relativo de los grafos comparados.
   En esencia, la similidad entre dos grafos implica coocurrencias de subcadenas similares en los textos correspondientes.
   CS considera la coocurrencia de pares de n-gramas, en lugar de las coocurrencias de grafos individuales. Por el contrario, el valor normalizado de similitud toma en

cuenta la frecuencia de coocurrencias de n-gramas. NVS funciona en pares de n-gramas, en lugar de n-gramas individuales.

$$NVS(G^i, G^j) = \frac{VS(G^i, G^j)}{SS(G^i, G^j)}$$

4.3. Construcción de los modelos

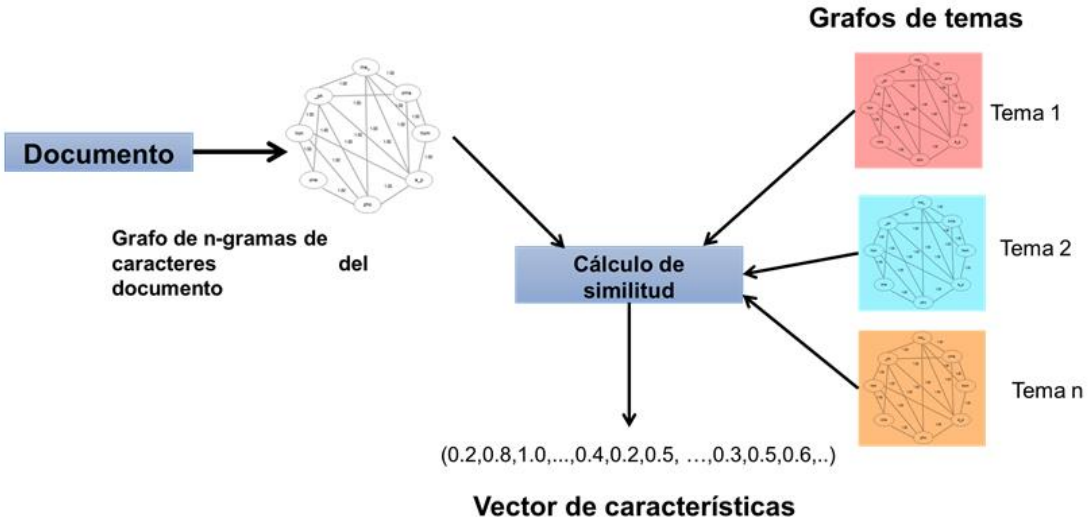

Figura 21. Clasificación de un documento usando el modelo de grafos de n-gramas

Para la construcción de los grafos representativos de cada una de las clases de nuestro sistema se utiliza una serie de cinco módulos para extraer el texto, hacer la representación en n-gramas de caracteres, construcción de los grafos y la función *Update*.

La construcción de los grafos representativos consta de cuatro etapas que a continuación se describen.

1. Limpieza del corpus: la primera etapa consiste en la extracción del texto del corpus del PAN[20] 2015. Se obtiene la clase de cada uno de los archivos para organizarlos por carpetas de acuerdo a sus clases.

[20] http://pan.webis.de/

2. Pre-procesamiento: esta segunda etapa consiste en realizar un pre-procesamiento del texto, donde se eliminan urls y se reemplazan símbolos y caracteres especiales además de números, con una letra específica para poder identificar cada uno de estos casos. Posteriormente, estos documentos se representan en tri-gramas de caracteres.

3. Representación de documentos a grafos: la tercera etapa consiste en la construcción de los grafos. Para representar un documento d, se crea un grafo G, corriendo una ventana de tamaño D_{win} sobre el contenido textual con el fin de analizar en el traslape de n-gramas de caracteres.

Cualquier par de n-gramas que se encuentran dentro de la misma ventana son conectados con una arista con un pesado [15], el cual denotará su frecuencia de co-ocurrencia en el documento.

Figura 22. Tipos de ventanas de n–gramas (No simétrica, simétrica y Gauss simétrica normalizada).

El tamaño de la ventana D_{win} que se utilizó fue de tamaño 4. Existen tres enfoques para tipos de ventanas: No simétrica, Simétrica y el enfoque simétrico normalizado-Gauss [25]. Para nuestro sistema se usó el enfoque de ventana simétrica. En la Figura 22 se ejemplifica estos tres enfoques.

4. Construcción de grafos modelos: por último, por medio de la función *Update*, se unen todos los grafos por sus respectivas clases creando así los grafos modelos que servirán para posteriormente poder extraer el vector de atributos. Como se muestra en la Figura 25.

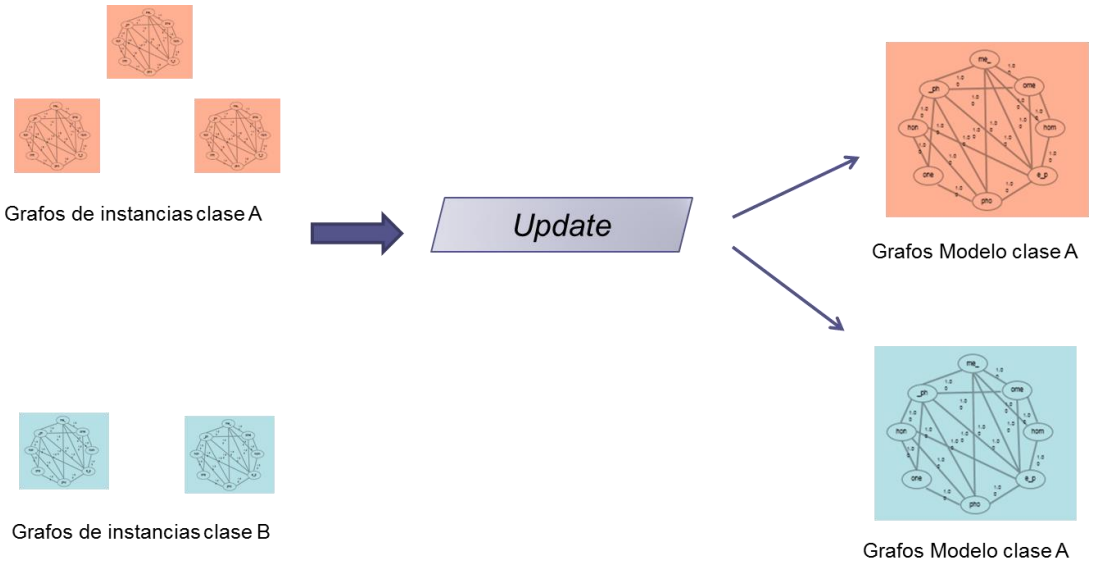

Figura 23. Diagrama de construcción de los grafos representativos de cada clase

Esta función se describe como $update(G_1, G_2) \equiv G^u = (E^u, V^u, L, W^u)$, donde $E^u = E_1^G \cup E_2^G$, donde E_1^G, E_2^G son el conjunto de aristas de G_1, G_2 respectivamente. Se considera que dos aristas son equivalentes $e_1 = e_2$ cuando tienen la misma etiqueta, por ejemplo L(e_1) = L(e_2).

Los pesos de los grafos resultantes se calculan de la siguiente manera: $W^i(e) = W^1(e) + \left(W^2(e) - W^1(e)\right) \times l$.

El factor de aprendizaje $l \in [0,1]$, cuanto mayor sea el valor de factor de aprendizaje, mayor es el impacto del segundo grafo en el primero. Cuando $l = 0$ indica que nuevo grafo ignorara al segundo grafo. Y cuando el valor es $l = 1$ indica que los pesos de los nodos del primer grafo serán asignados a los pesos del grafo resultante.

Para mantener el peso promedio de todos los grafos individuales que contribuyen a este modelo, el i-ésimo grafo que actualiza el grafo de la clase (modelo) usa un factor de aprendizaje $l = \left(1 - \frac{i-1}{i}\right), i > 1$.

4.4. Pruebas

En este capítulo se describe el conjunto de datos utilizado para las pruebas que se realizaron, así como una breve descripción de los algoritmos de aprendizaje y las medidas de evaluación que se emplearon. Y por último se muestran los resultados obtenidos en las pruebas.

Se realizaron una serie de pruebas para poder evaluar que algoritmo de aprendizaje obtiene mejores resultados, así como ver su eficacia en cada una de las tareas.

4.4.1. Conjunto de datos

Se hizo uso del corpus del PAN[21] 2015, que consiste en un conjunto de datos de entrenamiento de tuits en inglés, español italiano y alemán. En específico se usó el corpus en español para la construcción de los modelos.

- Para la edad se consideran las siguientes clases: 18-24, 25-34, 35-49, 50-xx.

- Género: male, female.

- Rasgos de personalidad, para cada rasgo se proporcionan puntajes (entre -0.5 y 0.5). Para nuestras pruebas se tomó como un problema de clasificación binaria, se asigna 1 para el polo positivo del rasgo y 0 para el polo negativo. Por ejemplo 1 si es extrovertido y 0 si es introvertido.

```
<author id="{author-id}"
        type="twitter"
        lang="en|es|it|nl"
        age_group="18-24|25-34|35-49|50-xx"
        gender="male|female"
        extroverted="-0.5 to +0.5"
        stable="-0.5 to +0.5"
        agreeable="-0.5 to +0.5"
        conscientious="-0.5 to +0.5"
        open="-0.5 to +0.5"
/>
```

Figura 24. Atributos del corpus PAN 2015

[21] PAN Promueve la investigación forense de texto digital mediante la organización de eventos de informática donde se invita a investigadores y profesionales para trabajar en alguna tarea compartida específica de su interés. En la categoría de *Authorship (*Autoría) se encuentran diez tareas, entre las cuales se encuentra Identificación de Autor y Perfilado de Autor. [59]

En la Figura 26 se muestran los atributos del corpus. Se cuenta con el rango de edad del usuario, así como el género y la puntuación de cada uno de los cinco rasgos de personalidad.

Para estas pruebas se usó el conjunto de datos en español, que tiene 100 archivos de texto de tuits de usuarios diferentes.

4.4.2. Algoritmos de aprendizaje

Para los experimentos realizados se seleccionaron cuatro algoritmos de aprendizaje, de los más representativos y utilizados en el campo del aprendizaje automático.

- **Naive Bayes (NB):** Método probabilístico, de los más utilizados por su simplicidad y rapidez, que asume la independencia de los atributos entre las diferentes clases del conjunto de entrenamiento.

- **J48**: Un algoritmo que permite generar un árbol de decisión, el cual selecciona los atributos más discriminativos basándose en su medida. La característica fundamental de este algoritmo es que incorpora una poda del árbol de clasificación una vez que éste ha sido inducido, es decir, una vez construido el árbol de decisión, se podan aquellas ramas del árbol con menor capacidad predictiva. Este algoritmo es una mejora de ID3, también basado en árboles, donde el criterio escogido para seleccionar la variable más informativa está basado en el concepto de cantidad de información mutua entre dicha variable y la variable clase [18] .

- **Optimización mínima secuencial (SMO):** Este algoritmo está basado en redes neuronales (funcionamiento inspirado en el cerebro humano, de ahí su nombre) cuya característica más importante es su capacidad de aprender a partir de ejemplos, lo cual les permite generalizar sin tener que formalizar el conocimiento adquirido.

- **Redes de función de base radial (RBF):** Este algoritmo, al igual que el anterior, está basado en redes neuronales. Este tipo de redes se caracteriza por tener un aprendizaje o entrenamiento híbrido. La arquitectura de estas redes se caracteriza por la presencia de tres capas: una de entrada, una única capa oculta y una capa de salida. [19]

4.4.3. Evaluación

Para evaluar el método propuesto se utilizaron las medidas tradicionales para la evaluación de sistemas de clasificación, tales como precisión, recuerdo y la medida F.

La precisión (*P*) es la proporción de instancias clasificadas correctamente en una clase c_i con respecto a la cantidad de instancias clasificadas en esa misma clase. El recuerdo (*R*), es la proporción de instancias clasificadas correctamente en una clase c_i con respecto a la cantidad de instancias que realmente pertenecen a esa clase. Así, la precisión se puede ver como una medida de la corrección del sistema, mientras que el recuerdo da una medida de cobertura o completitud.

Normalmente se emplea la medida F para describir el comportamiento de la clasificación, la cual se define:

$$F = \frac{(1+\beta^2)P*R}{\beta^2(P+R)}$$

Donde β representa la media armónica entre la precisión y el recuerdo. La función de β es la de controlar la importancia relativa entre las medidas de precisión y recuerdo. Es común asignar un valor de 1 indicando igual importancia a ambas medidas.

4.4.4. Experimentos y resultados

Como parte de las pruebas que se hicieron, se realizaron tres experimentos. El primero fue de edad con cuatro clases (18-24, 25-34, 35-49, 50-XX), el segundo experimento género con dos clases (Masculino y Femenino) y por último el experimento de personalidad con cinco clases, definido como un problema de clasificación binario (Introverted / Extroverted, Unstable /Stable, Disagreeable/Agreeable, unconscientious / Conscientious, y Closed / Open).

Se realizaron dos rondas, cambiando el conjunto de entrenamiento y de test. Las siguientes tablas muestran el promedio de la medida f de ambas rondas.

Clase	NB	J48	SMO	RBF
18-24	0.743	0.7	0.8335	0.643
25-34	0.5665	0.4665	0.5665	0.8455
35-49	0.5545	0.5	0.554	0.7
50-XX	0	0.598	0.3335	0

Tabla 4. Resultados obtenidos del experimento 1, edad.

En la Tabla 4 se muestran los resultados del experimento 1. En esta se puede observar el desempeño de los algoritmos empleados. En general tuvieron resultados similares, pero los algoritmos SMO y J48 pudieron clasificar todas las clases. Se puede ver, en base a estos resultados, que es más fácil clasificar los textos de personas que están entre los 18 y 24 años de edad. Mientras que los de 50-XX se obtiene un desempeño menor.

Clase	NB	J48	SMO	RBF
F	0.807	0.807	0.807	0.784
M	0.4705	0.4705	0.4705	0.4705

Tabla 5. Resultados del experimento 2, género

Los resultados del experimento 2 se muestran en la Tabla 5. En este experimento los cuatro algoritmos tuvieron un rendimiento muy similar para clasificar el género. Pese a que se contaba con un conjunto de datos con el 50% de usuarios femeninos y el restante 50% masculinos, se observa que se tiene un mejor desempeño para clasificar el género femenino.

Clase	NB	J48	SMO	RBF
Introverted	0.914	0.8655	0.86	0.8675
Extroverted	0.9705	0.9105	0.9125	0.92
Unstable	0.9255	0.96	0.9175	0.937
Stable	0.926	0.967	0.906	0.9445
Disagreeable	0.923	0.959	0.9115	0.937
Agreeable	0.981	0.988	0.9775	0.9815
Unconscientious	0.896	0.987	0.896	0.9685
Conscientious	0.978	0.9965	0.978	0.9915
Closed	0.875	0.875	0.875	0.875
Open	0.992	0.992	0.992	0.992

Tabla 6. Resultados del experimento 3, personalidad.

Por último, en la Tabla 6 se pueden ver los resultados del experimento de personalidad. Se observa que el desempeño general de todos los algoritmos que se usaron fue similar, además que el polo positivo de los rasgos de personalidad tiene ligeramente un mejor desempeño.

5. Desarrollo del sistema

En esta sección se muestra cómo es el funcionamiento del sistema de identificación de rasgos de personalidad y la estructura de los módulos que lo componen. El sistema web desarrollado IYP - Identifying Your Personality consta de cuatro módulos generales: Modulo de extracción de tuits, Módulo de representación del documento, Módulo de extracción de atributos, y Módulo de identificación de personalidad.

A continuación se muestra el esquema general del sistema así como cada uno de los módulos que lo integran.

5.1. Esquema general del sistema

El esquema general del sistema contempla los cuatro módulos que se desarrollaron en la implementación. Cada módulo realiza una tarea específica, como se muestra en la Figura 25.

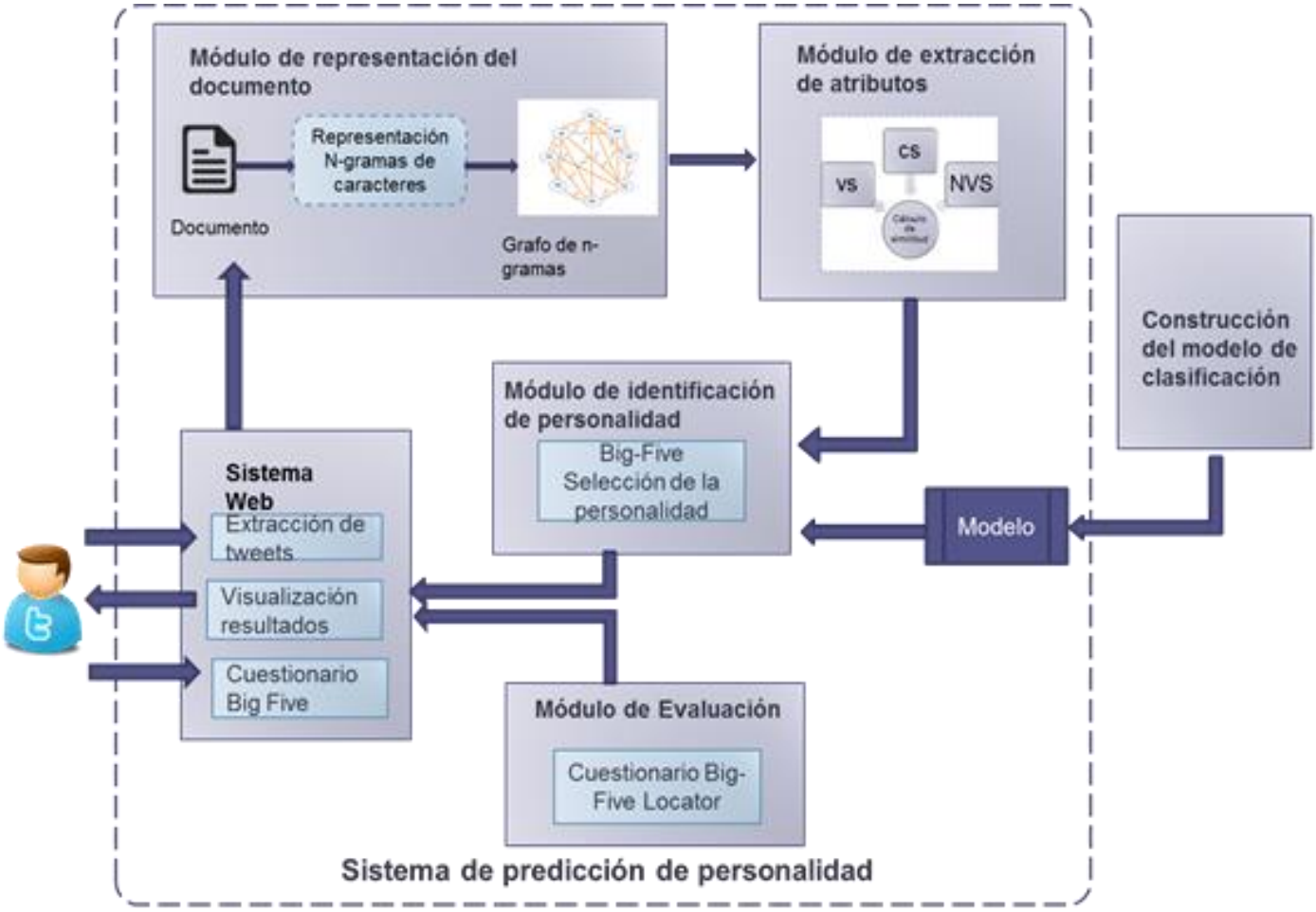

Figura 25. Esquema general del sistema

5.2. Módulo de Extracción de tuits

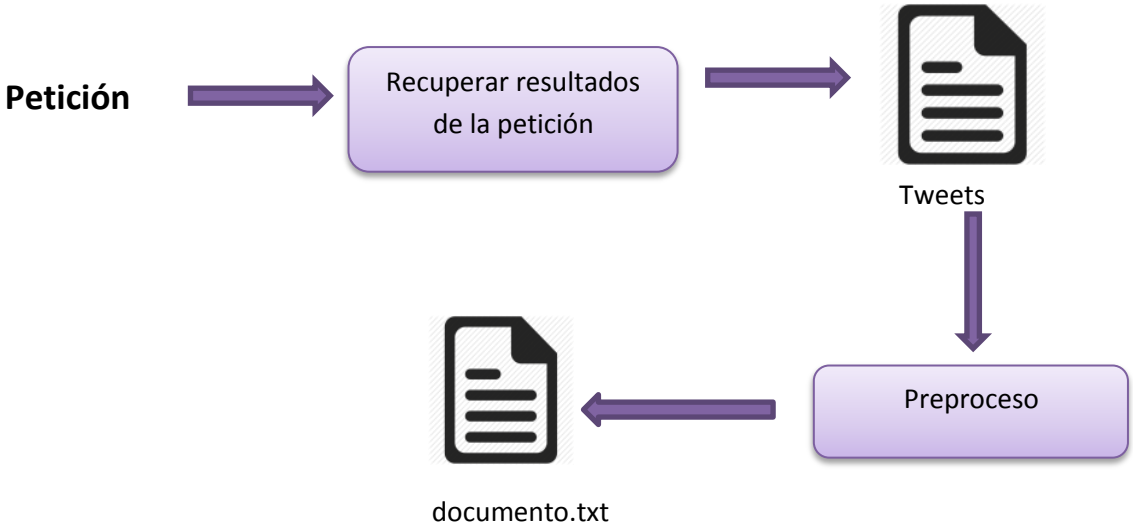

Figura 26.Módulo de extracción de tuits

La recuperación de los tuits comienza con una petición, como se muestra en la Figura 26, se ingresa el nombre de usuario a buscar. Para poder realizar esta tarea se utilizó el API-PHP de Twitter, el cual permite la conexión con Twitter, esto, previo registro en la página de desarrolladores de Twitter, en donde se obtienen las llaves de acceso que permiten hacer uso de la información que se genera en esa red social.

Si la conexión a Twitter es exitosa, se recuperan los 200 tuits más recientes del usuario y se muestran en el sistema.

Se realiza un pre-procesamiento de estos documentos, el cual considera las siguientes tareas:

- Reemplazar símbolos por el carácter ‘S’
- Reemplazar números por el carácter ‘N’
- Eliminar saltos de línea
- Eliminar url
- Reemplazar vocales acentuadas por vocales sin acento.

Una vez recuperados los tuits se guardan en un archivo *txt*, únicamente los tuits publicados por el usuario, ignorando todos los que sean citas o re-tuit .Dentro de la carpeta “Text” que se encuentra en el módulo “OCEAN” del sitio web.

5.3. Módulo de Representación del documento

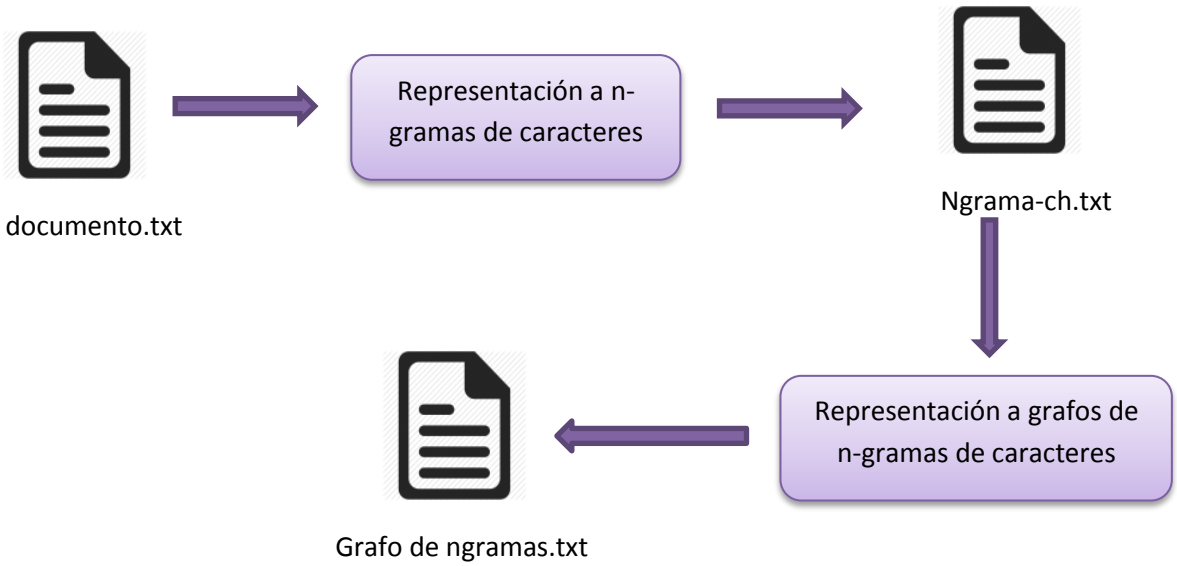

Figura 27.Módulo de representación del documento.

La Figura 27 muestra el proceso de representación del documento, primero en n-gramas de caracteres y luego en grafos.

Para la representación en n-gramas se hace uso del programa *"ngrams_CH",* que hace uso de la librería NLTK, la cual es el kit de herramientas de lenguaje natural. Es un conjunto de bibliotecas y programas para el procesamiento del lenguaje natural (PLN) para el lenguaje de programación Python. Para nuestro sistema se eligió usar tri-gramas de caracteres.

También se utilizó el módulo que se desarrolló en Python "*graph*", que contiene diversas funciones para la creación de grafos. '*ngram2GP.py'* es un programa en Python que lee un archivo de caracteres de n-gramas y crea a partir de este un grafo, recibiendo el tamaño de Dwin para calcular la coocurrencia.

5.4. Módulo de Extracción de atributos

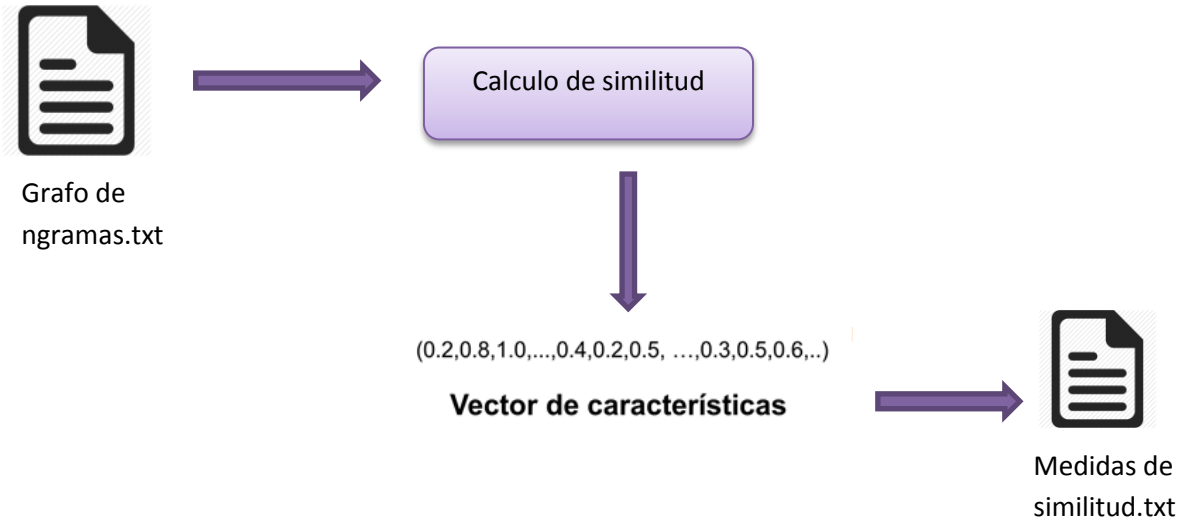

Figura 28.Módulo de extracción de atributos

La extracción de atributos de hace por medio de los cálculos de similitud, tal como se muestra en la Figura 28.

Estas son las medidas de similitud que son calculadas:

- Containment Similarity (CS)
- Size Similarity (SS)
- Value Similarity (VS)
- Normalized Value Similarity (NVS)

En el capítulo anterior se describieron más ampliamente estas medidas.

5.5. Módulo de identificación personalidad

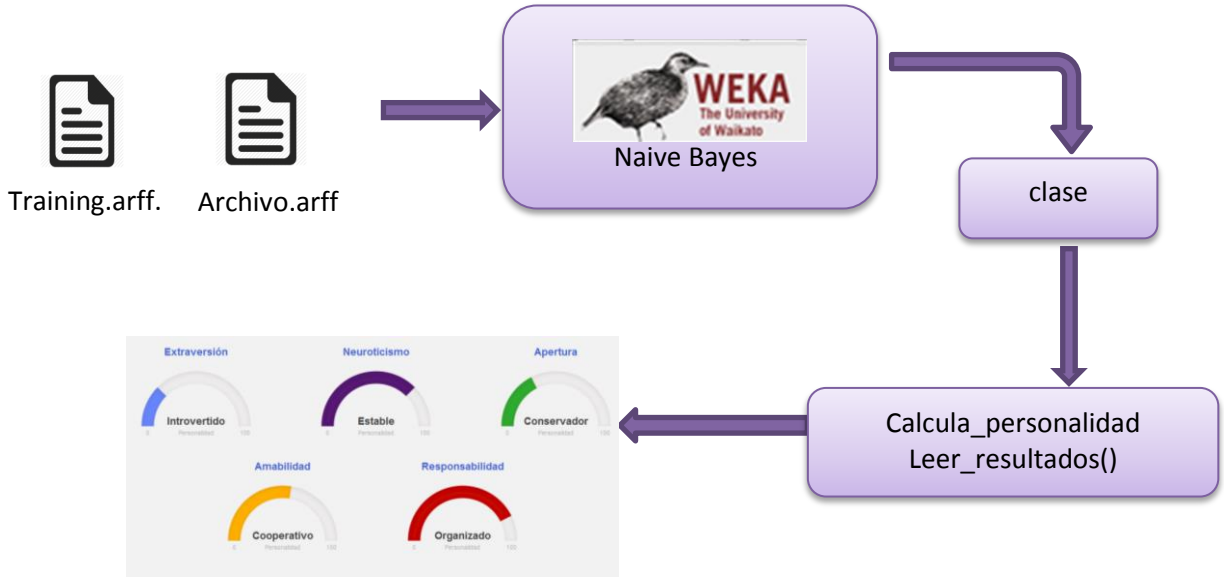

Figura 29.Módulo de identificación de personalidad

Para la identificación de personalidad se creó un módulo que hace uso de Weka[22] , que es es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java.

Se eligió el algoritmo de aprendizaje, Naive Bayes por su rapidez, su fácil implementación y porque que en las pruebas que se realizaron se obtuvieron buenos resultados de clasificación con este algoritmo.

Dentro del sistema Web se llama a weka por línea de comando y se le pasan los archivos arff que se crearon previamente, tanto el arff de entrenamiento, como el de la nueva instancia que queremos clasificar. El resultado de la clase obtenida para la nueva instancia es almacenado en un archivo de text. Por medio de la función *leer_resultados()* , se lee el archivo y se grafica el resultado y es mostrado en el sistema web. En la Figura 29 se muestra este proceso de identificación de personalidad.

[22] Por sus siglas en inglés *Waikato Environment for Knowledge Analysis*

5.6. Módulo de evaluación

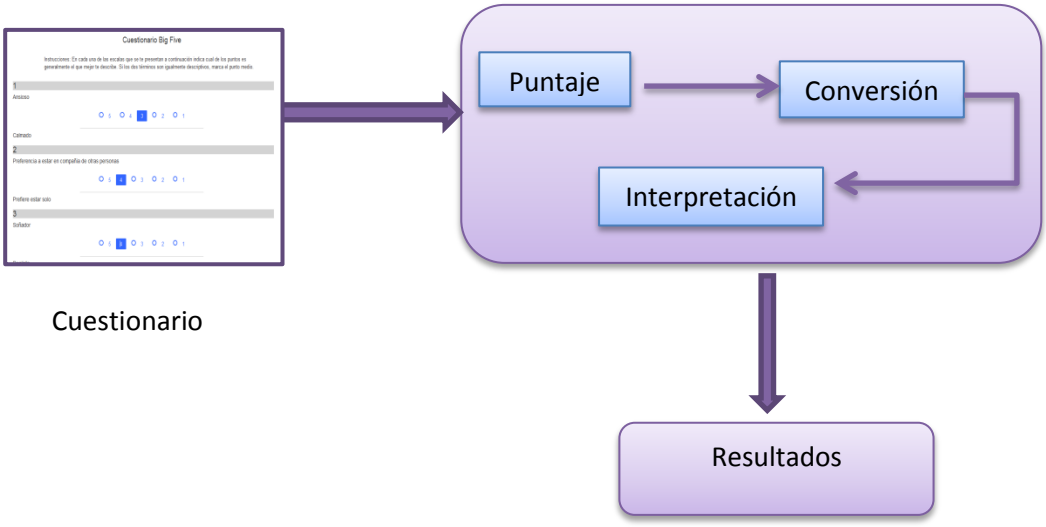

Figura 30.Módulo de evaluación

Se creó el módulo *big_five_locator* para poder evaluar la aplicación, y permitir tener una retroalimentación con el usuario.

Este módulo consta de tres funciones: puntaje, conversión e interpretación. A continuación se describe la función de cada una de ellas.

La evaluación consiste en un cuestionario de personalidad "Big Five Locator" que consta de 25 preguntas, y cinco opciones de respuesta. Una vez que el usuario contestó todas las preguntas (se hace una validación al formulario del cuestionario para asegurar esto), por medio de *puntaje.php* se obtiene el puntaje de cada pregunta, y con esta información se procede a calcular el puntaje de cada reactivo. Teniendo cinco factores: Emotividad negativa, Extroversión, Apertura, Adaptabilidad y Enfoque a metas.

El puntaje de cada uno de los reactivos es pasado a la siguiente función *conversión.php*, la cual realiza la conversión de estos puntajes en la tabla de conversión *BIG FIVE LOCATOR SCORE*.

La conversión nos da como resultado el score normal de cada reactivo. Este score normal es pasado a la siguiente función *interpretación.php* que es donde se determina en qué polo, positivo o negativo se encuentra en cada uno de los rasgos de personalidad. En la imagen anterior se describe el proceso de la evaluación que se realiza en el sistema web.

Y por último se guardan los resultados en un archivo de texto, para poder compararlos con los resultados que el módulo de predicción de personalidad arrojo, y de esta manera poder realizar una evaluación de la eficiencia de nuestro sistema de identificación de personalidad.

5.7. Sistema web

El sitio Web Identifying Your Personality tiene las siguientes características:

- Acerca de
- Cuestionario de personalidad
- Contacto
- Ingreso de cuenta de Twitter
    - Extracción de los tuits más recientes
    - Predicción de personalidad
        - Resultados

5.8. Vistas del sistema

A continuación se muestran las diferentes vistas que componen al sistema web Identifiyig Your Personality-IYP

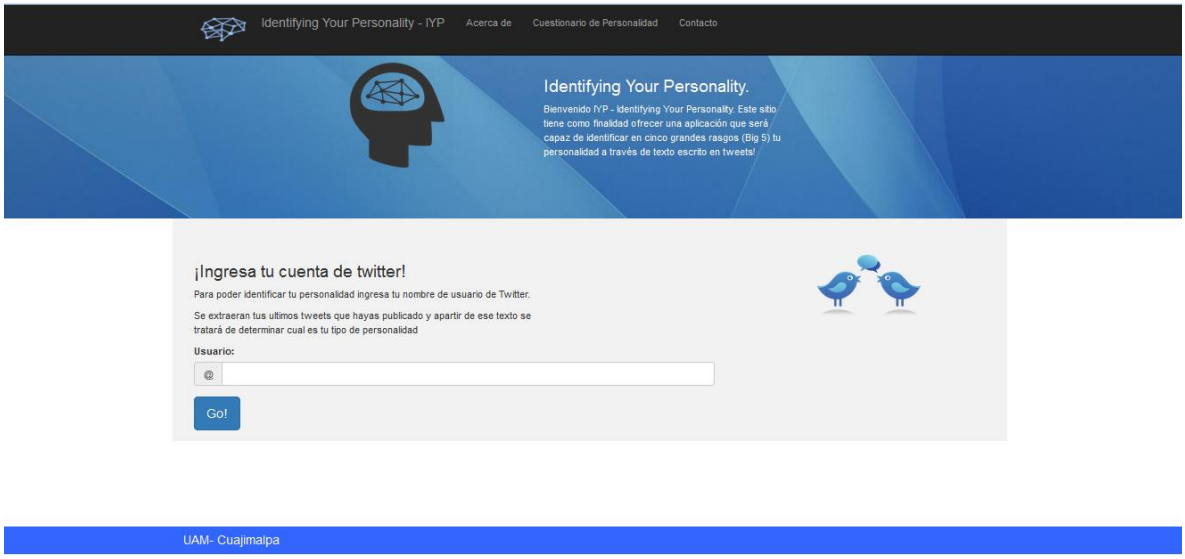

Figura 31.Página de inicio

En la Figura 31 se muestra la página principal, que es la página de inicio que despliega el sistema al ingresar.

En la parte superior se encuentra el menú de la página que contiene enlaces a la página Acerca de, Cuestionario de personalidad y por último la página de contacto.

En el contenido de la página se encuentra una sección donde hay un campo de texto para que el usuario pueda ingresar su cuenta de Twitter con la que se llevará a cabo el proceso de identificación de la personalidad.

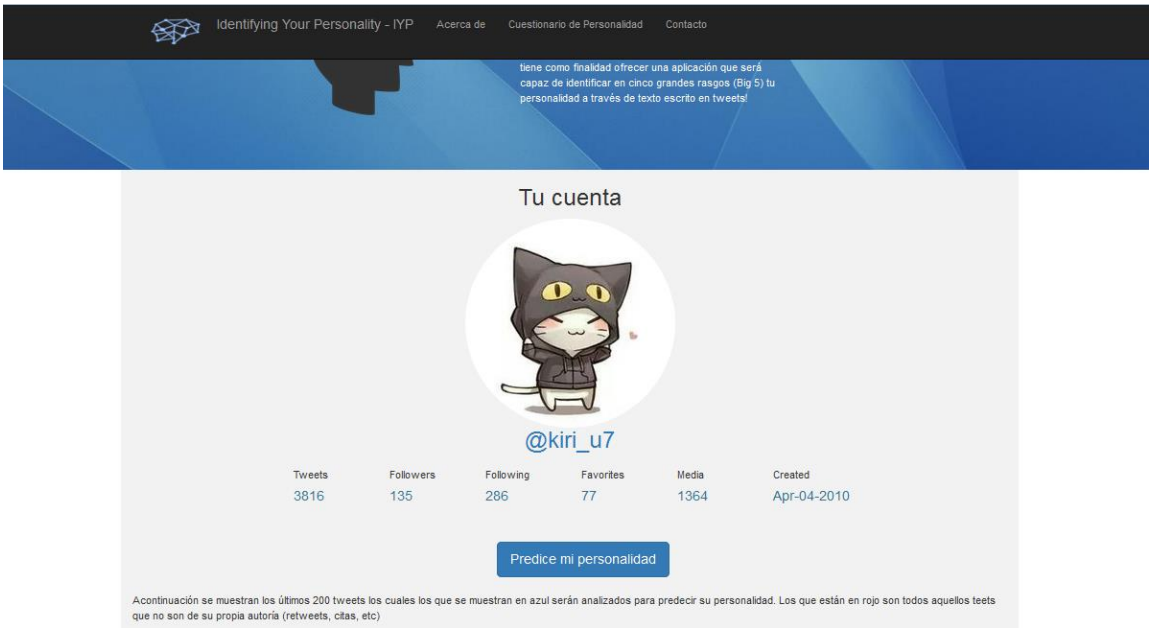

Figura 32.Vista cuenta de Twitter del usuario

El la Figura 32 se muestra la vista una vez que el usuario ingresó a su cuenta de Twitter, en esta se muestra la imagen de perfil del twitter, así como otros datos obtenidos de la API, como son número de seguidores, tuits, favoritos, etc.

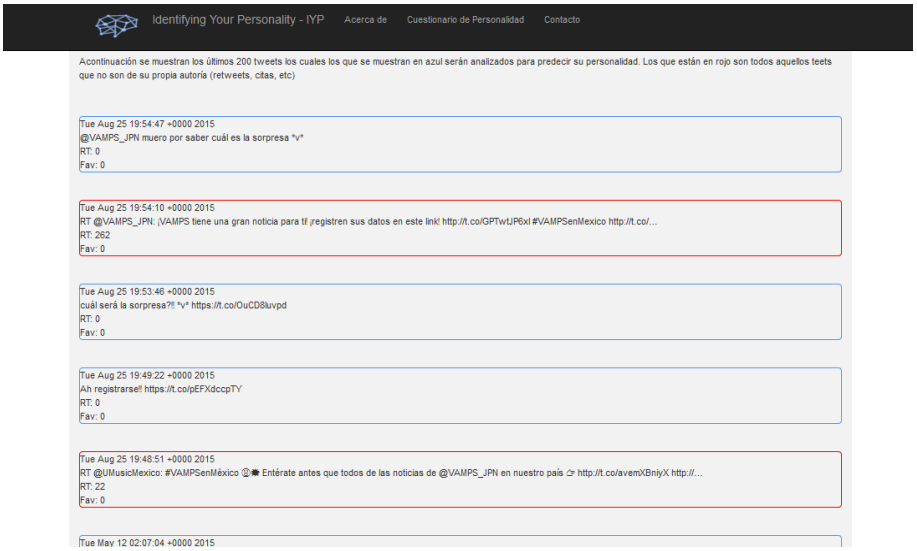

Figura 33.Tuits más recientes del usuario.

También se muestran los últimos 200 tuits más recientes del usuario. En color rojo se muestran todos aquellos que no fueron escritos por el usuario cómo re-tuits, menciones, etc. Y en color azul se visualizan los tuits que el usuario escribió los cuales sistema usará para la predicción de personalidad. Como se observa en la Figura 33.

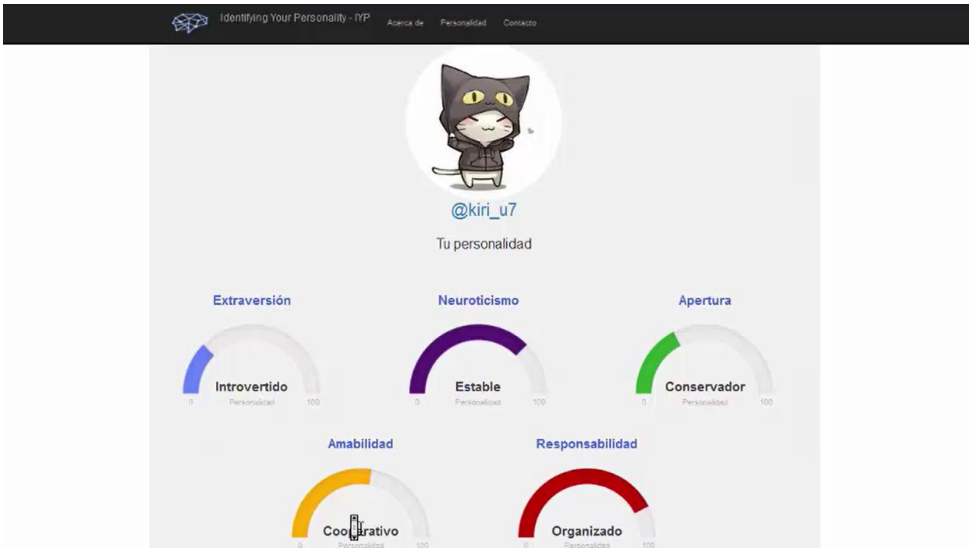

Figura 34.Resultados de la identificación de personalidad.

En la Figura 34 se muestra la pantalla de resultados, donde una vez realizado todo el proceso de identificación internamente, se obtienen los resultados para cada uno de los cinco rasgos de personalidad. Estos resultados se muestran en intervalos de 0-100 y se visualizan como gráficas de colores.

En la Figura 35 se muestra el cuestionario de personalidad que servirá para evaluar al sistema.

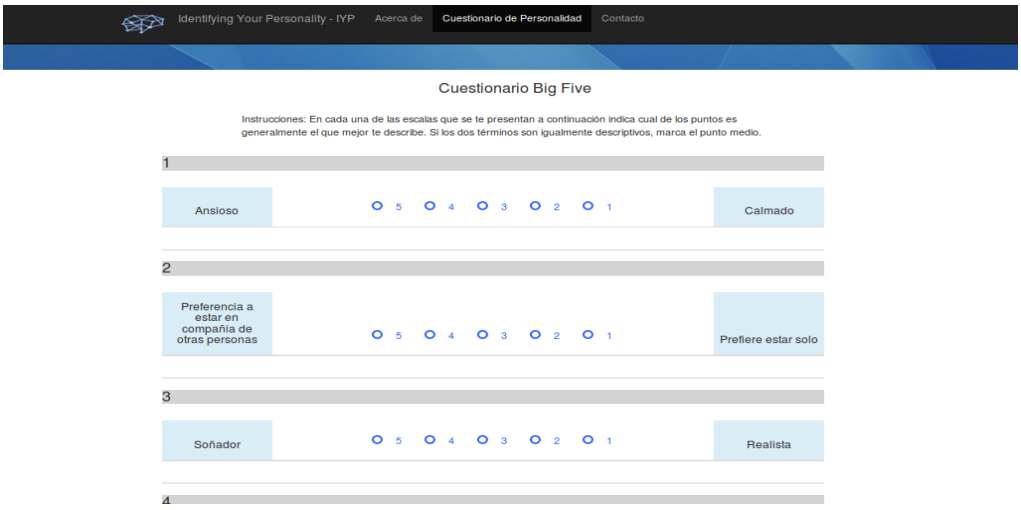

Figura 35.Vista cuestionario Big Five

El cuestionario consta de 25 preguntas, con cinco opciones de respuesta. Ninguna pregunta debe de quedar sin ser contestada, es por eso que se hace una validación del cuestionario para asegurar esto. El la Figura 36 se muestra esta validación, donde se despliega un mensaje de error cuando una pregunta queda sin contestar y el contorno se resalta en color rojo para marcar las preguntas que faltan responder y el verde las que ya fueron respondidas.

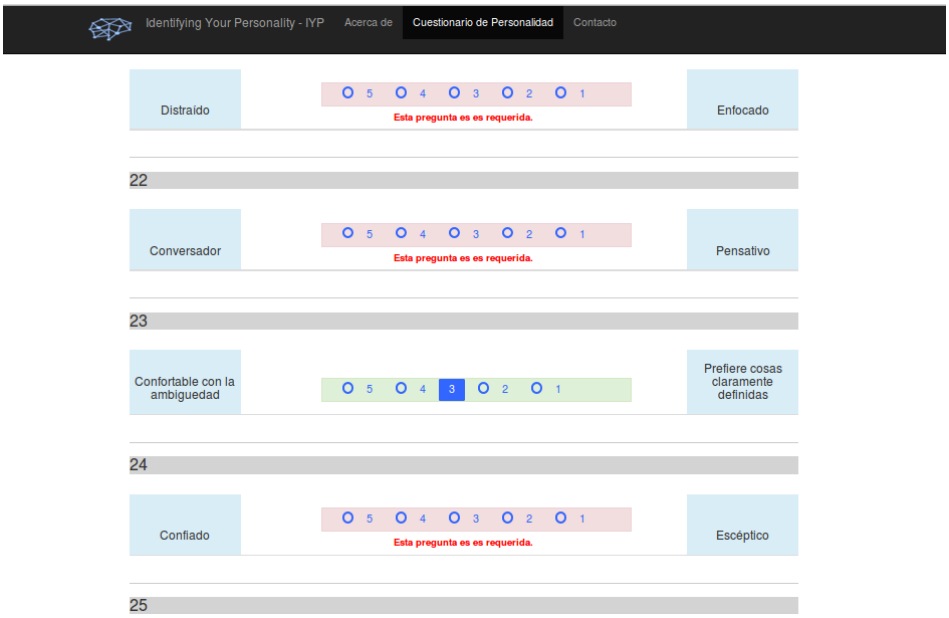

Figura 36.Validación del cuestionario

La

Figura 37 muestra la vista de la página de los resultados del cuestionario, en estos re muestra en qué polo se encuentra (polo positivo o polo negativo) en cada uno de los cinco rasgos de personalidad.

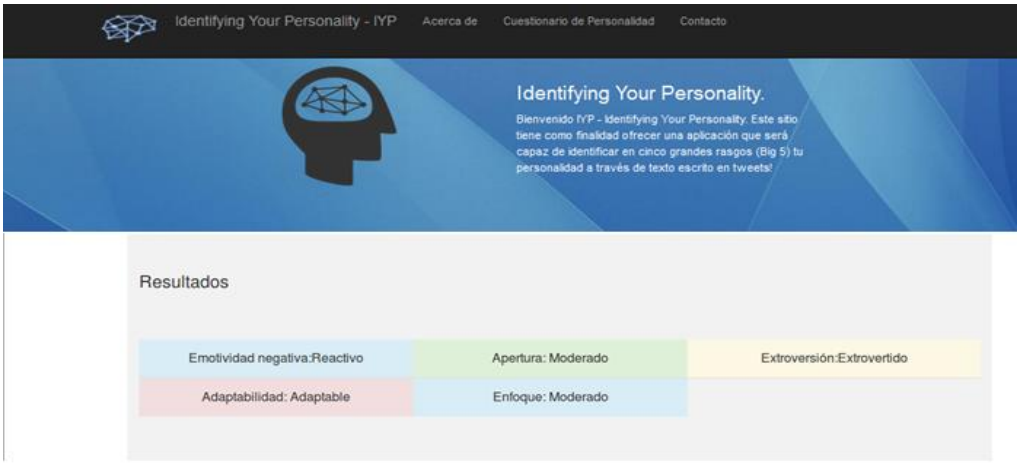

Figura 37.Resultados del cuestionario

6. Conclusiones y trabajo futuro

En este apartado se presentan los objetivos que se lograron cumplir en la realización del sistema Web, así como las conclusiones a las que se llegaron con el trabajo. Y por último se detalla el trabajo pensado a realizarse en el futuro.

Los objetivos que se plantearon en un principio se lograron satisfactoriamente. Se empleó una representación basada en grafos de n-gramas de caracteres y se realizó la clasificación de edad, género, así como de los diferentes rasgos de personalidad.

Otro de los objetivos principales que se logró fue el de implementar el método de grafos de n-gramas de caracteres en un sistema Web, *Identifiying Your Personality*. En este sistema Web el usuario es capaz de ingresar su cuenta de Twitter y el sistema predice los rasgos de personalidad de este usuario.

Otro de los objetivos que se plantearon fue el de poder evaluar el desempeño de la predicción de personalidad. Para esto se incorporó al sitio Web un cuestionario de personalidad Big Five Locator, donde el usuario sólo debe contestar una serie de preguntas para calcular los polos de personalidad en cada uno de los rasgos. Estos resultados se almacenan, permitiendo al sistema poder comparar los resultados que arrojó el cuestionario con los propios.

Las conclusiones que se llegó con este trabajo fue que los grafos de caracteres de n-gramas si pueden ayudar a mejorar los resultados de ciertas tareas. Pero aún se necesitan de más pruebas para poder evaluar la eficacia del sistema.

Como parte del trabajo futuro se tiene pensado ampliar el sistema Web para que también sea capaz de detectar género y edad del usuario, para que de este modo haga una identificación más completa del perfil del usuario.

Para mejorar la eficacia del sistema se tienen contempladas muchas ideas, entre las que están: aprovechar y explorar otras métricas de los grafos, así como explorar más opciones de búsqueda de patrones en un grafo para una mejor interpretación de los datos; se requiere conseguir un corpus más grande y mejor etiquetado para poder entrenar a los modelos de clasificación; experimentar con otras formas de representación del documento, entre otros.

7. Bibliografía

[1] Kamber, «Kamber,» 2014. [En línea]. Disponible: http://kamber.com.au/social-media-trends-2014-part-five/. [Último acceso: Marzo 2015].

[2] «Domo,» 2015. [En línea]. Disponible: Los contenidos que se generan en las redes sociales.. [Último acceso: Marzo 2015].

[3] «luismaram,» Marzo 2014. [En línea]. Disponible: http://www.luismaram.com/2014/03/24/cuanto-contenido-se-genera-por-minuto-en-twitter-y-facebook/. [Último acceso: Marzo 2015].

[4] I. Fried, «All Things D,» 29 Mayo 2013. [En línea]. Disponible: http://allthingsd.com/20130529/meeker-500-million-photos-shared-per-day-and-thats-on-track-to-double-in-12-months/. [Último acceso: Marzo 2015].

[5] E. J. PEDRERO PÉREZ, «Evaluación de la personalidad de sujetos drogodependientes que solicitan tratamiento mediante el Big-Five Questionnaire,» Madrid.

[6] J. t. Laak, «La cinco grandes dimensiones de la Personalidad,» vol. XIV, nº 2, 1996.

[7] R. S. C. Lerena, «Escalas para Evaluar la Personalidad y Trastornos de la Personalidad,» Cochabamba-Bolivia, 2010.

[8] F. Iacobelli, A. J. Gill, S. Nowson y J. Oberlander, «Large Scale Personality Classification of Bloggers,» de *Affective Computing and Intelligent Interaction*, Springer Berlin Heidelberg, 2011.

[9] T. M. Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math; (March 1, 1997), p. 432.

[10] Wikipedia, «Wikipedia,» [En línea]. Disponible: http://en.wikipedia.org/wiki/Document_classification.

[11] F. Sebastiani, «Machine Learning in Automated Text Categorization,» Italy.

[12] G. Sidorov, Construcción no lineal de n-gramas en la Linguistica Computacional, México, 2013.

[13] F. Celli y L. Polonio, «RELATIONSHIPS BETWEEN PERSONALITY AND,» 2013.

[14] A. Kartelj, V. Filipović y V. Milutinović, «Novel Approaches to Automated Personality,» Serbia.

[15] A. H. Schwartz, J. . C. Eichstaedt, L. Dziurzynski, M. . L. Kern, M. . E. P. Seligman y L. H. Ungar, «Toward Personality Insights from Language Exploration in Social Media,» AAAI Spring Symposium, 2013.

[16] S. Bhardwaj, «Personality Assessment Using,» Ottawa, 2014.

[17] D. Quercia, M. Kosinskiy, D. Stillwell y J. Crowcroft, «Our Twitter Profiles, Our Selves: Predicting Personality with Twitter,» UK.

[18] J. I. Biel, V. Tsiminaki, J. Dines y D. G. Perez, «Hi YouTube! Personality Impressions and,» Switzerland.

[19] J. W. Pennebaker, The Secret Life of Pronouns, Bloomsbury Press, 2011.

[20] N. Ramirez Esparza, C. . K. Chung, E. Kacewicz y J. . W. Pennebake, «The Psychology of Word Use in Depression Forums».

[21] F. Mairesse, M. . A. Walker, M. . R. Mehl y R. K. Moore, «Using Linguistic Cues for the Automatic Recognition of,» 2007.

[22] M. S. D. G. T. Kosinski, «Private traits and attributes are predictable from digital records of human behavior.Proceedings of the National Academy of Sciences (PNAS).,» 2013.

[23] «myPersonality,» [En línea]. Disponible: http://mypersonality.org/. [Último acceso: Marzo 2015].

[24] J. W. Pennebaker y K. G. Niederhoffer, «Linguistic Style Matching in Social Interaction,» de *Language and Social Psychology*, 2002.

[25] Search Engine Journal , «SEJ,» 15 Noviembre 2013. [En línea]. Disponible: http://www.searchenginejournal.com/growth-social-media-2-0-infographic/77055/. [Último acceso: Marzo 2015].

[26] F. Sebastiani, «Machine Learning in Automated Text Categorization,» Italy.

[27] E. J. K. M. D. L. R. Schwartz HA, «Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach,» nº doi:10.1371, 2013.

[28] «Department of Psychology,» [En línea]. Disponible: http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/Home2000/Background.htm. [Último acceso: 2015].

Apéndices

- **Apéndice diagramas casos de uso**

El sistema contará con un actor principal: Usuario.

Actor Usuario: es la persona que usará la el sistema. Podrá acceder al sitio web y acceder a todas las secciones de este, así como realizar las siguientes actividades:

- Ingresar su cuenta de usuario de twitter
- Ingresar sus tuits a la opción "Predice mi personalidad"
- Visualizar los resultados obtenidos
- Realizar el cuestionario de personalidad y visualizar los resultados de este cuestionario

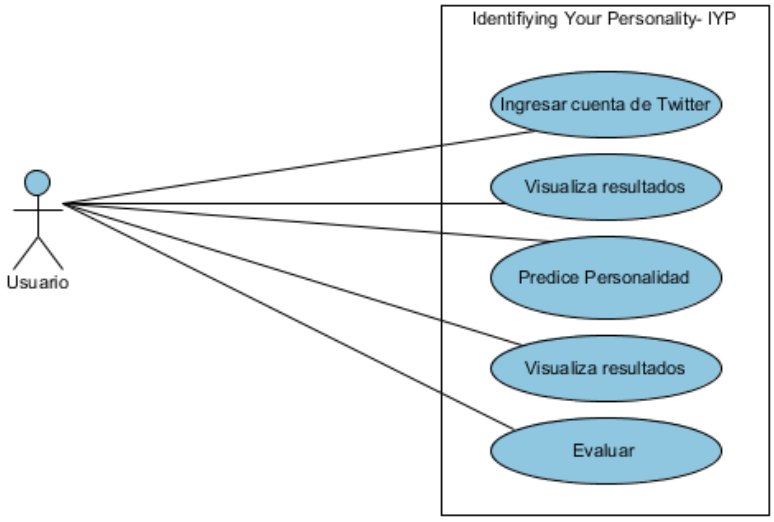

Figura 38.Caso de uso 1: Identificar personalidad

Caso de Uso No. 1	**Identificar Personalidad**	
Descripción	Permitirá al usuario conocer su personalidad	
Actor(es)	Usuario	
Precondición	El usuario ingrese su cuenta de Twitter desde la página principal y seleccione la opción *"Go"*	
Propósito	La necesidad de identificar la personalidad del usuario.	
Secuencia normal	Paso	Acción
	1	El usuario Visualiza los datos de su cuenta de Twitter
	2	Visualiza los 200 tuits más recientes.
	3	El usuario elije la opción 'Predice mi personalidad'
	4	Visualiza los resultados de su personalidad en los cinco rasgos Extraversión, Neuroticismo, Apertura, Amabilidad y Responsabilidad.
	5	Se visualizará la opción de *"Evaluar"*
Post condición	Se almacena en un archivo de texto el usuario de Twitter.	
Excepciones	Que el usuario no ingrese una cuenta válida de Twitter.	

Tabla 7. Caso de uso 1: Identificar Personalidad

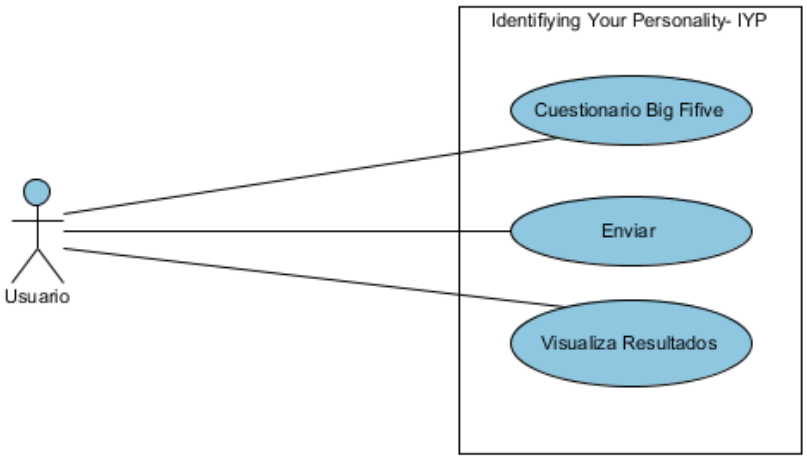

Figura 39.Caso de uso 2: Evaluación

Caso de Uso No. 2	**Evaluación**	
Descripción	Permitirá al usuario evaluar los resultados que dio el sistema	
Actor(es)	Usuario	
Precondición	El usuario y haya visualizado los resultados de personalidad de la sección “Predecir mi personalidad”	
Propósito	La necesidad de acceder al perfil de usuario.	
Secuencia normal	Paso	Acción
	1	El usuario entra a la opción *“Evaluar”*
	2	Se visualiza el cuestionario Big Five
	3	Si el usuario contesta todas las preguntas del cuestionario de personalidad
	4	Selecciona la opción *“Enviar”*
	5	Si el usuario no contesto todas las preguntas, se resaltarán todas las preguntas que faltan por responder.
	6	Visualiza los resultados de personalidad del cuestionario.
Flujo secundario	1	El usuario acceda desde el menú principal la opción “Cuestionario de personalidad”
	2	Repita los pasos de la secuencia normal a partir del paso 2.
Post condición	Se guardan en un archivo de texto los resultados del usuario.	
Excepciones	Que el usuario no responda el cuestionario	

Tabla 8. Caso de uso 2: Evaluación

- **Apéndice diagramas de secuencia**

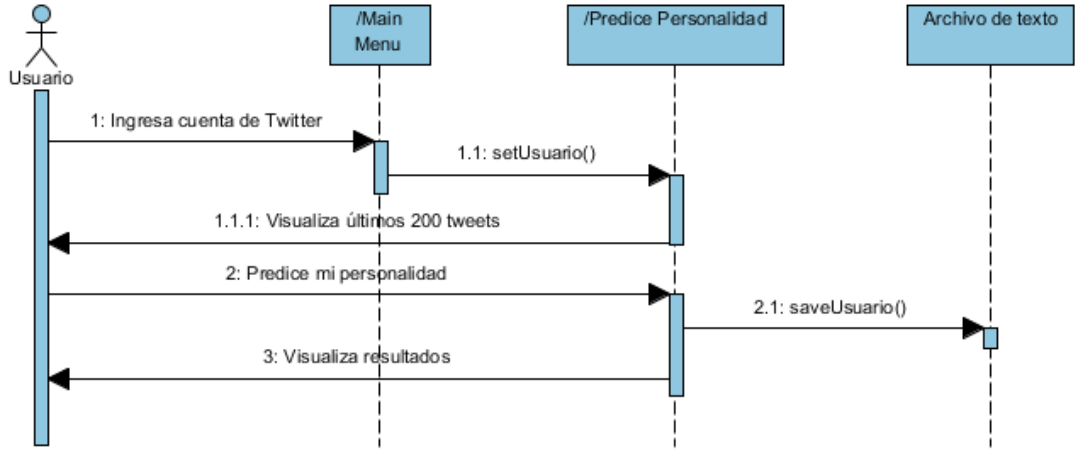

Figura 40.Diagrama de secuencia caso de uso 1

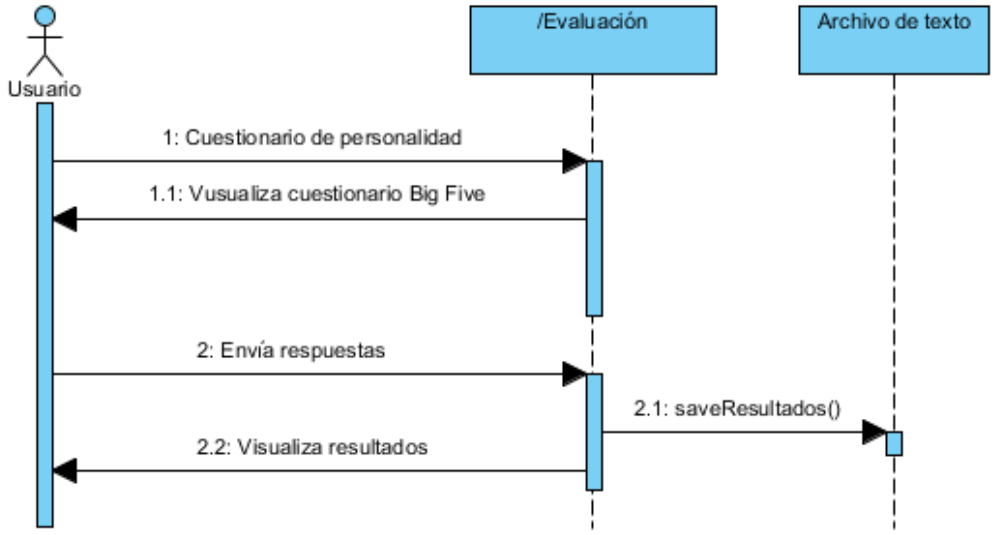

Figura 41.Diagrama de secuencia caso de uso 2