SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Subsecretaría de Educación Superior
Tecnológico Nacional de México
Coordinación Sectorial Académica
Dirección de Estudios de Posgrado e Investigación

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Centro Nacional de Investigación y Desarrollo Tecnológico

Subdirección académica
Departamento de ciencias computacionales

TESIS DE MAESTRÍA EN CIENCIAS

Servicio Web para la detección automática de personalidad a través del análisis lingüístico de textos

Presentada:
L. I. Carlos Acevedo Peña

Como requisito la obtención del grado de
Maestro en ciencias en ciencias de la computación

Directora de tesis:
Dra. Alicia Martínez Rebollar

Codirectora de tesis:
Dra. María Yasmín Hernández Pérez

Revisores:
Dr. Máximo López Sánchez

Dr. Joaquín Pérez Ortega

Cuernavaca, Morelos, México. Junio de 2018

Agradecimientos

Al llegar al final de este viaje que comenzó en circunstancias muy adversas es para mí, es un verdadero gusto dar las gracias a todos aquellos que me brindaron su apoyo de manera directa o indirecta aun sin conocerme y que con el tiempo se han convertido en verdaderos y estimados amigos.

En primer lugar, agradezco a Dios por ser mi guía, por haberme escuchado siempre, por poner muchas luces en mi camino y por estar en cada paso que doy, llevándome siempre de la mano por que sin Dios, nada y con él, todo.

A las instituciones que me apoyaron, al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico, que permitió el desarrollo de esta tesis.

Al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) por la oportunidad de realizar una Maestría en Ciencias de la Computación. Al personal administrativo y académico, por su atención y amabilidad, en especial a Lili, muchas gracias por tu apoyo.

Gracias Dra. Alicia y Dra. Yasmín por creer en mí, gracias por todo el tiempo que me han dado, por sus sugerencias e ideas que me han ayudado a crecer. No cabe duda que sin su apoyo este trabajo no hubiera sido posible, gracias por el respaldo y la amistad.

Y por supuesto, el agradecimiento más profundo y sentido es para mi familia. Este logro también es suyo. Sin su apoyo, colaboración e inspiración habría sido imposible llevar a cabo este duro viaje. A mi madre, quien me han enseñado siempre a poner mi fe en Dios y dar lo mejor de mí de corazón, a perseverar a través de sus sabios consejos, por hacer de mí una mejor persona, gracias por todo el cariño, comprensión y apoyo brindado a lo largo de mi vida, por guiar mi camino y estar junto a mí en todo momento. A mis queridos hermanos que son los mejores del mundo y no los cambiaría por nada. Por su apoyo incondicional, por comprenderme y hacer siempre las dificultades más ligeras y divertidas, aunque a veces parecieran muy complicadas, porque juntos siempre hemos salido adelante y siempre lo seguiremos haciendo.

A mis profesores, Lic. Patricia Armas León, Dr. Dante Mujica Vargas, Dr. Noé Castro Sánchez, Dra. Andrea Magadán Salazar, Dr. Máximo López Sánchez, Dr. Joaquín Pérez Ortega, Dr. Juan Gabriel González Serna, Dra. Azucena Montés Rendón, Dr. Gerardo Reyes Santiago, Dr. Gerardo Vela, Mtro. David Luviano Jiménez, Dra. Ana María Peña, Lic. Verónica Sotelo, por sus consejos, enseñanzas y porque todos ellos han aportado un granito de arena en mi formación profesional y sobre todo por su amistad.

A mis preciados amigos y compañeros: José de Jesús, Jaime, Marian, Ana Karen, Leopoldo, José Luis, Rubén, Miguel Ángel, Hermilo, Cristina, Claudia, Fortino, Melisa, David Ulises, David

Eduardo, Jeovanny, Samuel, Nelva, Andrea, German, Lupita, Roberto, Manuel, Arturo, José Carlo, Saraí, Nancy, Sergio, Víctor, Antonio, Abiud. Gracias también a Wendy, Pech, Alejandro, León Alberne, Fernando, Roberto Joshua, Edgar, Daniel, Luis y a todos mis hermanitos de tesis que me recibieron con los brazos abiertos.

A mi estimado amigo Eder Roberto Zavala Alarcón por ser un gran ser humano conmigo, por creer en mí y por apoyarme desde el primer momento muchas gracias.

Resumen

La personalidad es una combinación de características que definen el comportamiento de las personas. La personalidad puede influir en la elección de una persona en varias cosas, como páginas Web, ropa, libros, música o películas y afecta su relación con el ambiente que la rodea.

Actualmente los departamentos de recursos humanos en diversas instituciones o empresas, consultores, desarrollo del talento, psicólogos, entre otros tienen la necesidad de poder conocer la personalidad de sus candidatos, saber de qué manera se comportarán o se comportan esas personas por distintas razones, como por ejemplo, eficiencia, productividad, manejo de conflictos, seguridad. Actualmente para conocer la personalidad se realizan pruebas tradicionales que tienen distintos inconvenientes, por lo que se ha optado por utilizar métodos indirectos para poder calcular la personalidad como, el análisis de texto, análisis de comportamiento.

Este trabajo de investigación se centra en el desarrollo de un modelo para la detección automática de personalidad a través del análisis lingüístico de textos y la implementación de ese modelo mediante el desarrollo de un servicio Web utilizando el algoritmo de clasificación automática *SMO* y el peso de términos *TF-IDF* a partir de un conjunto de palabras que denotan personalidad en español creado en esta tesis.

Algunos trabajos de investigación llevan a cabo la predicción de personalidad pero la mayoría lo hacen trabajando con textos en idioma inglés haciendo uso de recursos léxicos en ese idioma que difícilmente se pueden encontrar disponibles para otros idiomas como el español. En este trabajo se lleva a cabo la creación de un conjunto de palabras que denotan personalidad y un corpus de personalidad en español basado en el modelo *DISC*.

En este trabajo de investigación se realizaron nueve experimentos para la clasificación de la personalidad utilizando el algoritmo *SMO* y los recursos léxicos generados, finalmente se obtuvo una precisión del 85.3% en el proceso de clasificación de personalidad y se desarrolló un servicio Web que implementa el algoritmo *SMO* que determina la personalidad de alguien mediante el análisis de su texto.

Abstract

Personality is a combination of characteristics that defines the behavior of people in different situations. The personality can influence the choice of a person in various things, such as Web pages, clothing, books, music or movies and affects their relationship with the environment that surrounds them.

Currently various institutions or companies, have the need to know the personality. For institutions it is important to know how the people that make up their work group will behave or behave for different reasons, such as efficiency, productivity, conflict management, security. Currently to know the personality traditional tests such as personality tests are carried out but these methods have different drawbacks, so different jobs have decided to use indirect methods to calculate personality such as text analysis, behavior analysis.

This research work focuses on the development of a method for the automatic detection of personality through the linguistic analysis of texts and the implementation of that method through the development of a Web service using the SMO automatic classification algorithm and the weight of terms TF-IDF from a set of words that denote personality in Spanish created in this thesis.

Some research works carry out personality prediction but most do it working with texts in English language using lexical resources in that language that can hardly be available for other languages such as Spanish. In this work the creation of a set of words that denote personality and a corpus of personality in Spanish based on the DISC model is carried out.

In this research work nine experiments were conducted for the classification of the personality using the SMO algorithm and the lexical resources generated, finally an accuracy of 85.3% was obtained in the personality classification process and a Web service was developed that implements the algorithm SMO that determines someone's personality by analyzing their text.

Tabla de contenido

Lista de figuras

Lista de tablas

Capítulo 1

Introducción

La personalidad es una combinación de características que definen el comportamiento de las personas que afecta a la interacción con otras personas y el medio ambiente.

A lo largo del tiempo se han propuesto varios modelos para identificar y etiquetar la personalidad en tipos o categorías, como el modelo *Big Five* (Christal, 1992), el modelo *PEN* (Eysenck, 1950) el modelo DISC (Marston, 1928), (Axiom, 2018), (Agung & Yunair, 2016)por ejemplo. Típicamente, para identificar la personalidad es necesario que el individuo se someta a una evaluación psicológica o una prueba de personalidad basada en un modelo de personalidad. Las pruebas de personalidad pueden ser relatos auto-descriptivos, entrevistas u observaciones realizadas por psicólogos. Estos métodos tradicionales son costosos y menos prácticos. Un estudio reciente muestra que los rasgos de personalidad pueden ser obtenidos automáticamente a partir del texto escrito (Mairesse, Walker, Mehl & Moore, 2007). La elección de las palabras más utilizadas puede describir la personalidad de una persona en particular (Pratama & Sarno, 2015). Es razonable esperar que diferentes individuos tengan diferentes maneras de expresarse a través de la palabra escrita, y estas diferencias corresponden a sus perfiles individuales de personalidad, así como a sus estados de ánimo(Sáez, Navarro, Mochón, & Isasi, 2014).

Se han realizado muchas investigaciones sobre la predicción de la personalidad, sin embargo, la mayoría de ellas es en el idioma inglés y se basan en el modelo de personalidad *Big Five.* En este trabajo de investigación se lleva a cabo el desarrollo de un servicio Web que permite determinar la personalidad de los usuarios mediante el análisis lingüístico de texto proporcionado por los mismos usuarios en el idioma español.

El modelo de personalidad que se utiliza en este trabajo de investigación es el modelo *DISC,* un modelo que consta de cuatro factores de personalidad del que se hablará en el capítulo II marco teórico.

Para el cálculo de la personalidad se llevó a cabo la construcción de un léxico y un corpus lingüístico en español anotado con pesos en los factores de personalidad del modelo *DISC* mediante la fórmula de ponderación de términos TF-IDF. Estos recursos resultan de gran importancia para futuras investigaciones sobre el cálculo de la personalidad y se describirán en el capítulo 4.

1.1 Planteamiento del problema

La personalidad es una combinación de características que definen el comportamiento de un individuo en diversas situaciones. La personalidad puede influir en la elección de una persona en varias cosas, tales como: páginas Web, ropa, libros, música y películas (Cantador et al., 2013). Además, la personalidad también afecta a la interacción con otras personas y el medio ambiente. La personalidad puede ser utilizada como un elemento de evaluación en la selección de personal en una empresa, en la orientación profesional, en terapias de pareja ó en el asesoramiento de la salud (Pratama & Sarno, 2015).

La personalidad ha sido analizada y evaluada desde diferentes puntos de vista. Actualmente, algunos modelos sicológicos de personalidad más utilizados para prededir la personalidad de un sujeto son el modelo *DISC* y el modelo *Big five* (Golbeck, Robles, Edmondson, & Turner, 2011).

El modelo de personalidad DISC contempla cuatro dimensiones o características únicas en una persona; es decir, este modelo nos indica las tendencias de una persona a ser dominante, influyente, estable o concienzudo (Bradberry, 2008).

En los últimos años, el test de personalidad *DISC* se ha posicionado como una de las herramientas favoritas entre los profesionales relacionados con la selección de personal y desarrollo del talento, por ser una solución sencilla y fiable que da respuesta a diversas necesidades. Sin embargo esta prueba consiste de 28 tablas con 4 grupos de palabras, de las cuales el sujeto a evaluar debe contestar de manera electrónica o escrita (Guisasola, 2016).

Actualmente, diferentes empresas e instituciones han descubierto la importancia de conocer la personalidad de quienes van a contratar o admitir en su organización (Alles, 2006) (Barret 1991), (Carbó, 2000). Esto debido a que la personalidad de los sujetos afecta directamente en el desempeño de sus actividades.

A pesar de existir diferentes pruebas psicológicas ampliamente reconocidas y utilizadas por las organizaciones, su aplicación y evaluación requiere de personal capacitado. Adicionalmente, la aplicación y evaluación de dichas pruebas requieren de un tiempo considerable. Uno de los métodos más utilizados son las llamadas pruebas psicométricas que están compuestas por un conjunto de preguntas que luego de ser contestadas por el sujeto bajo evaluación son analizadas, y de acuerdo con unos parámetros establecidos se genera el diagnóstico. Este procedimiento imprime subjetividad y potencia la aparición de errores (Martinez et al., 2018), lo que disminuye la confiabilidad de las técnicas psicométricas (Galindo & Aguilar, 2004).

También existen herramientas comerciales (PsicoSmart, 2018), (Prevue, 2018), (Apply magic sauce, 2018) que permiten la detección de personalidad. Sin embargo, estas herramientas han sido desarrolladas en diferentes idiomas excluyendo al idoma español; otras solo presentan un test para la determinación de la personalidad sin realizar un análisis automatico, que le evite al usuario el proceso de responder un largo cuestionario de manera manual.

Por lo anterior, se requiere de una herramienta automática para la detección de la personalidad con base en modelos de personalidad reconocidos y que represente una manera sencilla y confiable para detectar la personalidad.

1.2 Justificación

En este trabajo se propone un modelo para conocer la personalidad con base en el análisis lingüístico y en el modelo de personalidad DISC. El modelo incluye un corpus língüistico y conjunto de palabras que denotan personalidad.

El problema de conocer la personalidad es relevante ya que en muchas tareas se necesita conocer la personalidad de manera precisa y oportuna. Sobre todo, se requiere un método indirecto que evite información errónea.

Adicionalmente, este trabajo es importante ya que la generación de recursos lingüísticos se ha dado principalmente para el idioma inglés, por lo que es necesario desarrollar recursos para el idioma español.

Por el lado de la detección de la personalidad, existen varios enfoques orientados al análisis de comportamiento (Adali & Golbeck, 2012); otras investigaciones se han enfocado en el analisis de datos demográficos (Wald, Khoshgosftaar, & Summer, 2012) y de textos (Sáez et al., 2014), (Lima & de Castro, 2014), (Peng, Liou, Chang & Lee, 2015), (Celli, 2012). Este trabajo propone un modelo que integra un modelo psicológico de personalidad con el análisis lingüístico lo que representa una contribución importante, ya que el modelo DISC no ha sido muy utilizado en los modelos computacionales de personalidad.

Por el lado del procesamiento de lenguaje natural, y particularmente en el analisis de textos, consideramos que esta área resulta muy interesante y que cuenta con problemas que aun no se

han resuelto. Este trabajo de investigación contribuye a incrementar el conocimiento del área y a aportar nuevos enfoques y modelos.

1.3 Objetivo

Desarrollar un modelo para la detección automática de la personalidad de un sujeto a través del análisis lingüístico de texto escrito por el mismo sujeto utilizando el modelo de personalidad *DISC.* Este modelo será implementado a traves de un servicio Web.

1.3.1 Objetivos específicos

- Definir un conjunto de palabras que denoten personalidad a partir de textos obtenidos de un grupo de estudiantes de nivel universitario, para identificar las palabras que denotan en mayor cantidad cada uno de los rasgos de personalidad del modelo *DISC.*
- Desarrollar un corpus de personalidad *DISC* etiquetado con pesos en cada factor de personalidad del modelo *DISC.*
- Implementar un algoritmo de clasificación automática, para determinar la personalidad en textos.
- Desarrollar e implementar un servicio Web para determinar la personalidad.

1.4 Alcances y limitaciones

1.4.1 Alcances

Este trabajo de investigación tiene como alcance la determinación de la personalidad de un sujeto a través del análisis de un texto escrito. Para lograr esto, por se desarrolló un modelo que permite determinar la personalidad de un sujeto a través del análisis linguistico de un texto de manera automática, utilizando el modelo de personalidad DISC. El modelo se realizó tomando una muestra de sujetos de grado universitario, por lo que el modelo tiene una mayor precisión en un dominio similar.

Además se desarrolló un servicio Web que implementa el modelo creado. El servicio Web recibe como entrada un texto del usuario de manera escrita o a través de la carga de un archivo de texto plano y después de que el servicio realiza el análisis del mismo se obtiene como resultado la personalidad del usuario.

En este trabajo de investigación se trabaja únicamente con textos en español.

1.4.2 Limitaciones

En el desarrollo de este trabajo de investigación se presentaron las siguientes limitaciones:

- **Falta de recursos léxicos para determinar personalidad:** Durante el desarrollo del proyecto se realizó una búsqueda de recursos léxicos como: bases de datos, corpus, o lexicones que pudieran

servir para realizar el cálculo de personalidad a travez del análisis de textos, pero debido a la escasez de estos recursos en el idioma español fue necesario desarrollar recursos léxicos propios, un corpus de personalidad y un conjunto de palabras que denotan personalidad basados en el modelo de personalidad DISC.

- **Las características de la muestra:** La aplicación de las encuestas está limitada a la disponibilidad de las personas que requieran participar. En este trabajo de investigación nuestra muestra fue de 120 personas a las que se les aplicó una encuesta de personalidad basada en el modelo DISC. Las personas más disponibles para este caso, fueron estudiantes de nivel universitario en un rango de edad de 20 a 30 años, debido a que presentaron mayor interés en participar, y se facilitaba en este caso el acceso a ellos. Por lo que la determinación de la personalidad se realiza con mayor precisión en personas con las características mencionadas.

1.5 Estructura del documento

El contenido de esta tesis se encuentra organizado en los siguientes capítulos:

Capítulo 2. Marco Teórico: En este capítulo se definen los principales conceptos para el tema de tesis desarrollado.

Capítulo 3. Estado del Arte: En este capítulo se presentan los trabajos relacionados con análisis de personalidad más representativos en la revisión de la literatura.

Capítulo 4. Modelo para la detección de personalidad: se describe el modelo propuesto para la detección automática de personalidad en textos en español, mediante el desarrollo de un servicio Web utilizando el modelo de personalidad *DISC.*

Capítulo 5. Servicio Web para la detección automática de personalidad: En este capítulo se presenta la arquitectura y diseño del servicio Web.

Capítulo 6. Pruebas y Resultados: En este capítulo se presentan las pruebas realizadas al modelo de detección de personalidad de textos en español.

Capítulo 7. Conclusiones y Trabajo Futuro: En esta sección se presentan las conclusiones y contribuciones, los trabajos futuros que se pueden derivar de este proyecto de investigación y las actividades adicionales que se realizaron.

Capítulo 2

Marco teórico

En esta sección de presentan algunos conceptos utilizados en este trabajo de investigación para una óptima comprensión del resto del documento.

2.1 Servicio Web

Un servicio Web es un componente de software almacenado en una computadora, el cual se puede utilizar mediante llamadas a métodos desde una aplicación (u otro componente de software) en otra computadora, a través de una red. Los servicios Web se comunican mediante el uso de tecnologías como XML y HTTP. Los servicios Web son independientes de la plataforma y del lenguaje, por lo que no hay que preocuparse por la compatibilidad de sus tecnologías de hardware, software y comunicaciones (Deitel, 2008).

2.2 Red social

Se definen como sitios de redes sociales, a los servicios basados en Web que permiten a las personas: (1) construir un perfil público o semipúblico dentro de un sistema limitado, (2) articular una lista de otros usuarios con los que comparten una conexión y (3) ver y recorrer su lista de

conexiones y las realizadas por otros dentro del sistema. La naturaleza y la nomenclatura de estas conexiones pueden variar de un sitio a otro (Conole, Galley, & Culver, 2011).

2.3 Modelo de personalidad DISC

El modelo de personalidad DISC fue desarrollado por el doctor William Moulton Marston (1893-1947) psicólogo. En este modelo Marston observó que existen cuatro dimensiones o características únicas de la personalidad. Aunque estos rasgos representan necesidades que son importantes para cualquier ser humano en algún grado, saber cuál predomina en una persona es la clave para comprender su personalidad. Los cuatro rasgos de la personalidad de Marston se conocen con el nombre de modelo *DISC*, sigla en inglés que resume las tendencias de una persona a ser dominante, interpersonal, estable o meticulosa. En la figura 2 se muestra la representación del modelo DISC.

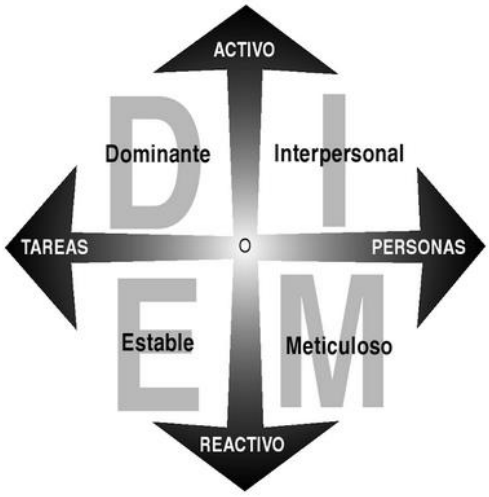


Figura 1. Representación del modelo DISC

El modelo *DISC* de Marston sintetiza lo que la gente suele pensar, sentir y hacer como producto de las tendencias inherentes a su personalidad. Este modelo nos brinda una imagen visual rápida para ilustrar las diferencias globales entre los cuatro rasgos de la personalidad (Bradberry, 2008).

El modelo *DISC* se basa en la elaboración de un test en el cual el individuo al contestar una serie de preguntas puede determinar su personalidad basada en los cuatro rasgos que conforman el modelo.

2.4 Recursos lingüísticos

La expresión recursos lingüísticos se refiere a un conjunto generalmente extenso de datos, así como a descripciones de una lengua en formato electrónico, empleados para mejorar y evaluar los sistemas de procesamiento del lenguaje natural. Ejemplos de recursos lingüísticos son los

corpus de textos y los lexicones, por ejemplo: bases de datos léxicas, tesauros, diccionarios electrónicos, etc. (Baca, 2014).

2.4.1 Corpus de textos

Un corpus es una colección de textos, representativos de una lengua, de un dialecto o un subconjunto de un lenguaje, que han sido elaborados con fines de investigación, y son aplicables a diversas tareas del Procesamiento del Lenguaje Natural y utilizados para el análisis lingüístico (Baca, 2014).

Un corpus también se puede definir como una colección de textos en formato electrónico, los cuales se convierten en repositorios de información a partir de los cuales se pueden encontrar, obtener, y por lo tanto, aprender los múltiples contextos en los que puede aparecer una determinada palabra, convirtiéndose así, en una fuente de información fundamental para los sistemas de desambiguación (Taulé, 2003).

Una tipología básica de los distintos corpus textuales se puede establecer dependiendo del propósito: corpus con fines generales, cuyo objetivo principal es el de constituir una fuente de información textual de una lengua para fines y aplicaciones diversas; y corpus con fines específicos, creados en respuesta a un propósito particular, como el estudio de aspectos concretos de la gramática o del léxico de la lengua, la extracción de datos estadísticos, el estudio del comportamiento lingüístico de una determinada población de hablantes, análisis comparativos de diversas variedades lingüísticas, o el desarrollo y evaluación de sistemas de Procesamiento de Lenguaje Natural (Baca, 2014).

2.5 Tokenización

Es el proceso de segmentar el texto en palabras y oraciones llamados tokens. El texto electrónico es una secuencia lineal de símbolos (caracteres o palabras o frases). Naturalmente, antes de que se realice cualquier procesamiento de texto real, el texto debe segmentarse en unidades lingüísticas tales como palabras, signos de puntuación, números, números alfanuméricos, etc. Este proceso se denomina tokenización. La tokenización es una especie de pre-procesamiento en cierto sentido; Una identificación de las unidades básicas a procesar (Trim, 2013).

2.6 Lematización

En el proceso conocido como *stemming* se busca la raíz (*stem*) de la palabra para utilizarla en aplicaciones relacionadas con la extracción de información.

*"*Stemming: proceso por el que se truncan las palabras de los documentos antes de indexarlos, con el objetivo de identificar palabras con la misma raíz" (Antonín, 2003).

2.7 *Weka (Waikato Environment for Knowledge Analysis)*

WEKA es una herramienta de minería de datos. *WEKA*, acrónimo de *Waikato Environment for Knowledge Analysis*, es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario. Para ello únicamente se requiere que los datos a analizar se almacenen con un cierto formato, conocido como ARFF (*Attribute-Relation File Format*).

WEKA se distribuye como software de libre distribución desarrollado en *Java*. Está constituido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación, y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. Estos paquetes pueden ser integrados en cualquier proyecto de análisis de datos, e incluso pueden extenderse con contribuciones de los usuarios que desarrollen nuevos algoritmos. Con objeto de facilitar su uso por un mayor número de usuarios, *WEKA* además incluye una interfaz gráfica de usuario para acceder y configurar las diferentes herramientas integradas (Corso, 2009).

2.8 Métodos de evaluación: *Ten-fold cross validation y Percentage Split*

2.8.1 Validación cruzada de 10 divisiones *(Ten-fold cross validation)*

Un aspecto a tener en cuenta a la hora de evaluar sistemas de clasificación supervisados es la distinción que hay que hacer entre el conjunto de entrenamiento (grupo de documentos a partir de los cuales el clasificador aprenderá) y el conjunto de evaluación (grupo de documentos que serán clasificados automáticamente por el sistema y sobre los que se calculan las métricas de evaluación).

Es necesario separar ambos conjuntos para obtener una evaluación correcta, ya que si no se realiza esta división los resultados del sistema podrían ser ligeramente mejores al estar el sistema sobre-adaptado al conjunto utilizado para el aprendizaje. Es por esto que, normalmente, se realizan k divisiones equitativas sobre el conjunto de documentos totales, donde k-1 partes son utilizadas para aprender y la restante para evaluar.

Este método, para realizar la evaluación de este tipo de sistemas, se denomina cross fold validation. Por regla general, se utilizan diez divisiones (10 fold-cross validation); es decir, se divide el conjunto total de documentos en 10 conjuntos. Cada subconjunto es utilizado como conjunto de evaluación, mientras que el resto es utilizado como conjunto de entrenamiento, evaluando el sistema diez veces y obteniendo como resultado final la media de las k evaluaciones parciale (Cuadrado, 2011).

2.8.2 Porcentaje dividido *(Percentage Split)*

Este método divide el conjunto de datos suministrado en dos subconjuntos, uno destinado al entrenamiento y otro al test, según un porcentaje especificado por el usuario. Weka calcula automáticamente las métricas más utilizadas para este tipo de tareas como son precision (precisión), recall (cobertura) y F-measure (medida-F) de cada clase(Cuadrado, 2011).

2.9 Algoritmo de aprendizaje automático SMO

El algoritmo SMO (*Sequential Minimal Optimization*) es un algoritmo simple que entrena al algoritmo SVM (*Support Vector Machine*). Las Maquinas de Vectores de Soporte o Support Vector Machines (SVM) son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik.

Entrenar un SVM requiere de la solución a un problema muy grande de QP (*quadratic programming*). SMO divide este problema en una serie de problemas QP más pequeños y resuelve rápidamente el problema SVM QP sin almacenar ninguna matriz extra y sin invocar una rutina iterativa numérica para cada sub problema (Baca, 2014).

El objetivo de los problemas de clasificación que aplican este tipo algoritmos de aprendizaje supervisado es el siguiente; dado un conjunto de entrenamiento con sus etiquetas de clase, entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra o conjunto de test.

Las SVM son una de las técnicas más poderosas del aprendizaje automático. Consiste en construir un hiperplano en un espacio de dimensionalidad muy alta (o incluso infinita) que separe las clases que tenemos. Una buena separación entre las clases permitirá una clasificación correcta de la nueva muestra, es decir, necesitamos encontrar la máxima separación a los puntos más cercanos a este hiperplano (Baca, 2014).

Capítulo 3

Estado del arte

En este capítulo se presentan los trabajos relacionados con análisis de personalidad más representativos en la revisión de la literatura:

3.1 Criterios de evaluación

Cada trabajo de investigación se describe basándose en ciertos criterios de evaluación, con el fin de realizar una comparación y evaluación objetiva, los criterios seleccionados son los siguientes:

- **Descripción General:** Se presenta una descripción general sobre los principales puntos de interés el trabajo seleccionado.

- **Métodos y técnicas:** Se realiza una descripción de los métodos y técnicas implementados en el trabajo presentado, es decir como realizaron la detección de la personalidad.

- **Ventajas:** Se presentan las ventajas observadas en el trabajo de investigación seleccionado.

- **Desventajas:** Se presentan las desventajas observadas en el trabajo de investigación seleccionado.

3.2 Prediciendo la personalidad con la conducta social

a) Descripción general

En este trabajo de investigación (Adali & Golbeck, 2012) se identifica el tipo de personalidad de un usuario por medio del análisis de su conducta social en la red social Twitter. El objetivo principal de este trabajo es saber si se puede determinar el tipo de personalidad de una persona analizando solo su comportamiento en las redes sociales. A demás compara los resultados con el de análisis de texto, mostrando un rendimiento estadísticamente equivalente.

b) Métodos y Técnicas

En el trabajo de (Adali & Golbeck, 2012) se introducen características que capturan información acerca del comportamiento social de un conjunto de usuarios de la red social Twitter. Se muestra correlación entre esas características y rasgos de personalidad, los valores esos rasgos se obtienen de una encuesta (versión de 44 preguntas del modelo de personalidad de los 5 grandes factores), aplicada a un grupo de 60 personas. Para encontrar la correlación entre las características de comportamiento y los rasgos de personalidad, se utiliza la regresión basada en selección de subconjuntos Forward (FSSreg).

Para predecir la puntuación de una característica de personalidad dada, se realiza un análisis de regresión en Weka. Para ello se usan dos algoritmos de regresión: Gaussian Process y ZeroR, cada uno con una validación cruzada de 10 veces con 10 iteraciones. Además de eso se lleva a cabo un análisis del texto de los tweets de los usuarios con el software LIWC (Linguistic Inquiry and Word Count) para la predicción de la personalidad.

c) Ventajas

- Este trabajo de investigación determina el tipo de personalidad mediante el análisis de la conducta social y del análisis de texto de los usuarios de la red social Twitter. A demás compara los resultados de estas dos maneras mostrando un rendimiento estadísticamente equivalente.

- Se muestra un alto grado de rendimiento de predicción al utilizar una combinación de algoritmos de regresión, características de conducta y la herramienta LIWC (Linguistic Inquiry and Word Count).

d) Desventajas

- Una de las desventajas es que en ciertas situaciones el usuario puede llevar un comportamiento muy similar al de cualquier grupo social, el llamado efecto de grupo, y de esa manera no mostrar su verdadera personalidad.
- La herramienta LIWC no se encuentra disponible en español de forma gratuita.

3.3 El reconocimiento de la personalidad no supervisado en redes sociales

a) Descripción general

En este trabajo de investigación (Celli, 2012) se presenta un sistema de identificación del tipo de personalidad de usuarios que utiliza características lingüísticas y el aprendizaje no supervisado. El sistema se ejecuta sobre un conjunto de 1065 publicaciones de 748 usuarios de la red social italiana FriendFeed. Se adoptan cinco clases a partir del modelo estándar conocido en psicología como el modelo de "los cinco grandes": extroversión, estabilidad emocional, amabilidad, responsabilidad y apertura a la experiencia.

b) Métodos y Técnicas

Para desarrollar el sistema de reconocimiento del tipo de personalidad sin supervisión, se convierten los coeficientes de correlación de los factores lingüísticos en características, que se pueden extraer automáticamente del texto. Como primer paso el sistema extrae una muestra aleatoria del conjunto de datos. A partir de esta muestra, el sistema extrae la media y la desviación estándar de cada característica. También se calcula la media de la frecuencia de palabras, con esta información el sistema genera sobre la marcha un modelo de personalidad para cada usuario. Entonces, el sistema evalúa los modelos mediante la comparación de todos los mensajes del mismo usuario. El sistema no supervisado toma todos los modelos construidos a partir de las publicaciones del usuario, compara sus valores y proporciona dos medidas, precisión y validez como medidas de evaluación.

c) Ventajas

- En este trabajo se utiliza una técnica de análisis no supervisado para el reconocimiento de los tipos de personalidad, esta técnica permite inferir el tipo de personalidad de usuarios de la red social FriendFeed sin necesidad de que se respondan algún tipo de prueba o presencia del usuario.

d) Desventajas

- Sólo se pueden evaluar modelos para los usuarios que tienen más de un mensaje en el conjunto de datos, y se desechan todos los otros usuarios.

- Este sistema se aplica a la red social FriendFeed que desapareció en abril 2015.

- El idioma utilizado es el italiano.

3.4 Máquina de predicción de la personalidad de perfiles de Facebook

a) Descripción general

En este estudio (Wald, Khoshgoftaar, & Sumner, 2012) se aplican técnicas de Minería de Datos y de aprendizaje automático para predecir los rasgos de personalidad de los usuarios (específicamente, los rasgos del modelo de personalidad *Big Five*) de la red social *Facebook*, usando atributos demográficos y textos extraídos de sus perfiles.

b) Métodos y Técnicas

En este trabajo de investigación se usan 31 atributos demográficos como: edad, sexo, localidad, citas, relaciones, el número de amigos, fotos, intereses y comentarios proporcionados por el usuario, etc. y 80 atributos basados en texto, procesados utilizando el paquete de software LIWC (Linguistic Inquiry and Word Count), los atributos son extraídos de los perfiles de 537 usuarios de la red social Facebook a quienes se les aplica una encuesta de 45 preguntas para categorizar su personalidad de acuerdo con el índice del modelo de los cinco factores de personalidad (Big five) y se emplean tres modelos de predicción numérica diferentes: regresión lineal (LinR), árboles de decisión (REPTree) y tablas de decisión (DTable) para predecir los rasgos de personalidad mediante la construcción de modelos haciendo uso de la herramienta *WEKA*.

c) Ventajas

- En este estudio, se determina el tipo de personalidad de los usuarios mediante el uso de técnicas de minería de datos, aprendizaje automático, atributos demográficos y también texto extraído de perfiles de Facebook.

d) Desventajas

- Al seleccionar el 10% de los individuos de la población, el algoritmo LinR presenta una baja efectividad al determinar los factores de personalidad.
- El algoritmo DTable es muy bueno sólo para predecir correctamente a los individuos más extremos en los grupos, pero no es tan capaz de seleccionar el grupo que está un poco más lejos del extremo.
- Los resultados del algoritmo REPTree son difíciles de clasificar.

3.5 Un Sistema de detección de la personalidad y de la felicidad

a) Descripción general

En este trabajo de investigación (Sáez et al., 2014) se presenta una aplicación móvil que se ejecuta sobre el sistema operativo Android, la aplicación permite la adquisición de información en forma de texto escrito de la aplicación WhatsApp y el servicio de SMS. Se presenta también un prototipo

para la clasificación de la información reunida, para la detección de los tipos de personalidad y de la felicidad. Todavía no se han obtenido resultados sobre la detección de los tipos de personalidad, ya que el prototipo se encuentra todavía en desarrollo.

b) Métodos y Técnicas

La aplicación implementa el modelo correspondiente a un sistema informático distribuido, este se compone de numerosos dispositivos. Con el objetivo de estimar los tipos de personalidad y la felicidad de los usuarios, esta aplicación permite la adquisición de información en forma de texto escrito y permite la comunicación con un servidor, la información se adquiere a partir de dos fuentes: la aplicación *WhatsApp* y el servicio de *SMS*. Un servidor recibe esta información de los clientes móviles y la almacena. Además, se menciona un prototipo, aun en desarrollo, para un módulo clasificador de la información recopilada, que busca marcadores que permitan clasificar al usuario según la teoría de la personalidad de Eysenck. Este módulo se comunica con la base de datos para obtener los datos de usuario a procesar y busca coincidencias de indicadores de personalidad. Debido a la falta información para analizar y probar el método completo con *Whatsapp* y mensajes *SMS*, se utilizan conjuntos de mensajes de otras fuentes como Twitter y blogs, con los que se realizan experimentos con algoritmos de clasificación en el software *WEKA*. La precisión de los algoritmos probados en el mejor de los casos es de un 57% (SVM).

c) Ventajas

- En este trabajo de investigación se construyó el primer corpus público de mensajes de la aplicación *WhatsApp* hasta la fecha.
- En este trabajo se realizan pruebas de análisis de textos en el idioma español de España para determinar el tipo de personalidad de los usuarios de *WhatsApp* y el servicio de SMS.

d) Desventajas

- Una desventaja de este trabajo de investigación es que se encuentra todavía en una fase inicial. Aún no está terminado y no se ha cumplido el objetivo de definir todos los factores de la personalidad de los usuarios de la aplicación *WhatsApp* y el servicio de SMS.

3.6 Un enfoque de clasificación multietiqueta semisupervisado, aplicado a la predicción de la personalidad en las redes sociales

a) Descripción general

En este trabajo de investigación (Lima & de Castro, 2014) se presenta un sistema de predicción de los tipos de personalidad de una persona a partir de grupos de mensajes de *Twitter* o '*tweets*'. Introduce un enfoque basado en el modelo de los cinco grandes factores de la personalidad para predecir la personalidad en los datos de medios sociales, más específicamente en grupos de mensajes de *Twitter* o *'Tweets*'. El sistema se denomina *PERSOMA* (Predicción de personalidad en medios sociales).

b) Métodos y Técnicas

El sistema presentado denomina *PERSOMA* (*PERsonality prediction in SOcial Media datA*) y trabaja con grupos de textos, en lugar de los textos individuales. *PERSOMA* se compone de tres módulos principales aplicados en cascada. En el primer módulo, se extrae un conjunto de meta atributos de los Tweets.

Cada una de las cinco dimensiones del modelo de los cinco factores (apertura a las nuevas experiencias, responsabilidad, extroversión, amabilidad y neuroticismo o inestabilidad emocional) se ve como un rasgo de personalidad, y, como cada tweet puede contener de cero a cinco rasgos, el problema se caracteriza como un problema de etiquetado múltiple.

En el segundo módulo, el problema de la etiqueta múltiple se transforma en un conjunto de cinco problemas de clasificación binaria (tiene o no tiene el rasgo de personalidad).

Por último, en el módulo de clasificación se aplican tres algoritmos diferentes: Naïve Bayes, una máquina de vectores soporte (SVM), y una red neuronal perceptrón multicapa para predecir el tipo de personalidad. En este módulo se toma un pequeño número de datos etiquetados y un enfoque de aprendizaje semi-supervisado para clasificar los datos.

c) Ventajas:

- En este estudio se trabaja con grupos de textos, en lugar de textos únicos, y no toma en cuenta la información de los perfiles de los usuarios de Twitter.
- *PERSOMA* hace uso de meta-atributos lo que lo hace menos dependiente de la lengua, lo que le permite extender fácilmente a otros idiomas.
- Alto grado de precisión al inferir los tipos de personalidad.

d) Desventajas:

- Se trabaja sólo con textos en inglés.
- Los mejores resultados solo se obtendrían en casos donde hay un pequeño número de datos etiquetados y un gran número de datos no etiquetados, esos serían los escenarios más adecuados ya que el método de clasificación utilizado es semi-supervisado.

3.7 Clasificación personalidad basada en textos de Twitter usando Naive Bayes, KNN y SVM

a) Descripción general

Este trabajo de investigación (Pratama & Sarno, 2015) se utiliza la clasificación de textos para predecir el tipo personalidad basada en texto escrito por los usuarios de *Twitter.* Los idiomas utilizados son el inglés y el indonesio. El Sistema desarrollado es una aplicación Web. El tipo de personalidad del usuario se predijo con un 60% de precisión.

b) Métodos y Técnicas

Este trabajo utiliza el conjunto de datos *MyPersonality* original, ligeramente modificado, está en forma de 250 documentos de 250 usuarios con los que se forma una sola cadena larga en un documento. El sistema toma los últimos 1.000 textos en forma de *tweets* (publicaciones hechas directamente por el usuario) y *re-tweets* (volver a publicar el texto de otra persona). La colección de *tweets* se convierte en un sola cadena larga en un documento, entonces es pre-procesado en datos vectoriales. El texto ya pre-procesado se clasifica en un conjunto de datos etiquetado. Los resultados son las predicciones para cada uno de los cinco grandes rasgos de personalidad. El lenguaje de programación utilizado es *Python* con la biblioteca *scikit-learn*. Se utilizan tres algoritmos para inferir el tipo de personalidad de los usuarios de la red social Twitter: *Naive Bayes*, K-vecinos más cercanos y máquinas de vectores soporte, se observa que *Naive Bayes* supera ligeramente los otros métodos.

c) Ventajas:

- De los tres métodos utilizados se muestra que el método *Naive Bayes* es más preciso.
- Este experimento utiliza la clasificación de texto para predecir los tipos de personalidad basada en el texto escrito por los usuarios de *Twitter.*
- El sistema tiene un 65% de precisión en comparación con la prueba basada en cuestionarios.

d) Desventajas:

- El sistema realiza detección solo con los idiomas inglés e indonesio.
- No se logra mejorar la precisión de la investigación previa que es de 61%.
- Los encuestados elegidos deben tener una cuenta de *Twitter* con un número mínimo de 1.000 tweets.

3.8 La predicción de rasgos de la personalidad de los usuarios chinos basada en publicaciones de Facebook

a) Descripción general

En este trabajo (Peng, Liou, Chang, & Lee, 2015) se intenta clasificar los rasgos de la personalidad a partir de textos chinos tomados de la red social *Facebook,* se aplican distintas pruebas, pero al final sólo llevan a cabo experimentos que logran determinar el factor de extraversión.

b) Métodos y Técnicas

En este trabajo, se recogen datos de personalidad de 222 usuarios de *Facebook* en Taiwán almacenándolos en documentos de texto sin formato, se leen los documentos y se utiliza la extracción de características y métodos de selección para construir la representación vectorial.

Se utiliza *Jieba,* una herramienta china de segmentación de texto, para la tarea de segmentación de texto, y finalmente se utiliza la máquina de vectores soporte (SVM) como el algoritmo de aprendizaje para la clasificación de la personalidad. Los resultados experimentales muestran que el rendimiento en la precisión y la recuperación se pueden mejorar de manera significativa con la ayuda de segmentación de texto. La precisión para clasificar extraversión que se muestra en este trabajo es del 73.5%.

c) Ventajas

- En base a los experimentos aplicados se observa que los usuarios clasificados como extravertidos escriben más oraciones y usan palabras más comunes que los introvertidos.
- Se muestra que el recuento de ocurrencias es un esquema básico pero fiable para la representación de un vector de términos en un documento.
- Se obtiene un porcentaje de precisión del 73.5% para el factor de extraversión.

d) Desventajas

- Este trabajo de investigación trabaja con textos chinos que son mucho más difíciles de delimitar que los textos en inglés.
- Solo se logra clasificar un factor de personalidad, la extraversión.
- Solo se toman en cuenta usuarios con más de 10 publicaciones no menos.

3.9 Tabla comparativa de trabajos relacionados

Trabajo	Técnicas	Red Social	Modelo de Personalidad	idioma	Recursos	Exactitud	Software
Prediciendo la personalidad con la conducta social	Regresión basada en selección de subconjuntos Forward (FSSreg) Algoritmos de regresión: Proceso de Gauss y ZeroR.	Twitter	Modelo de 5 factores (Big five)	Ingles	21,525 usuarios, Uso de Linguistic Inquiry and Word Count (LIWC).	85%	NO
El reconocimiento de la personalidad no supervisado en sitios de redes sociales	Media simple de la frecuencia de palabras, desviación estándar y correlación de Pearson.	FriendFeed	Modelo de los 5 factores (Big five)	Italiano	748 usuarios italianos de FriendFeed (1065 mensajes) y lista de coeficientes de correlación de Mairesse et al.	63.1%	NO
Máquina Predicción de la personalidad de los perfiles de Facebook	Regresión lineal (LinR), árboles de decisión (REPTree) y tablas de decisión (DTable)	Facebook	Modelo de los 5 factores (Big five)	Ingles	537 usuarios de Facebook, usando 31 atributos demográficos y 80 atributos basados en texto, Uso de Linguistic Inquiry and Word Count (LIWC) y WEKA	74.5%	NO
Un Sistema de detección de la personalidad y de la felicidad	Filtrado, Distancia euclidiana, Máquina de Vector de Soporte (SVM), LibLINEAR	Whatsapp y el servicio SMS	Modelo PEN	Español (España)	conjunto 1 = 1772 mensajes de Twitter, conjunto 2 = 314 mensajes de blogs, conjunto 3 = 200 mensajes de blogs, cuestionario EPQ-R, Freeling, WEKA	57%	Aplicación Movil sobre el sistema operativo Android
Un enfoque de clasificación multi-etiqueta, semi-supervisado aplicado a la predicción de la personalidad en los medios sociales.	Naïve Bayes Máquina de Vector de Soporte Red neuronal de Perceptron Multicapa	Twitter	Modelo de los 5 factores	Ingles	41 grupos de Tweets (18,435 Tweets), Java, WEKA, PersonalRecognizer (PRec), Uso de Linguistic Inquiry and Word Count (LIWC), base de datos psicolingüística MRC	83%	PERSOMA programa en Java
Clasificación personalidad basada en textos de Twitter usando Naive Bayes, KNN y SVM	Naive Bayes, K-vecinos más cercanos, máquinas de vectores soporte (SVM), Porter Stemmer para el idioma Inglés y Nazief-Andriani para el idioma indonesio	Twitter	Modelo de los 5 factores	Ingles, Indonesio	Version modificada del dataset MyPersonality (250 documentos de 250 usuarios), 1,000 twits del usuario, Phyton, biblioteca Scikit-Learn	60%	Aplicación Web
La predicción de rasgos de la personalidad de los usuarios chinos basada en publicaciones de muro de Facebook	Se utiliza Jieba, herramienta china de segmentación de texto, máquina de vectores soporte (SVM), kit de herramientas scikit-learn	Facebook	Modelo de los 5 factores	Chino	publicaciones de 222 usuarios de Facebook	73.5%.	NO

Tabla 1. Tabla comparativa de trabajos relacionados

Capítulo 4

Modelo para la detección de personalidad

En este capítulo se describe un modelo para la detección automática de personalidad en textos en español, mediante el desarrollo de un servicio Web utilizando el modelo de personalidad llamado *DISC*. En la primera sección se presenta una breve descripción del modelo y en las secciones siguientes se describen de manera detallada las fases que lo componen.

4.1 Descripción general del modelo para la detección de personalidad

En este proyecto de investigación se llevó a cabo el desarrollo de un servicio Web para la detección de personalidad en textos en español. En este modelo todo se propone utilizar un algoritmo de clasificación automática, un conjunto de palabras que denotan personalidad y un corpus de personalidad en español, ambos basados en el modelo de evaluación de personalidad *DISC*.

Una de las actividades realizadas fue la construcción de un conjunto de palabras que denotan personalidad en español basado en el modelo *DISC*, para extraer características del texto, con la finalidad de obtener información para identificar la personalidad de alguien analizando su texto. En este caso, se entienden por características; palabras, número de palabras y pesos TF-IDF de

esas palabras en cada uno de los factores del modelo de personalidad DISC. Con las características obtenidas se genera un vector, el que se utiliza en la clasificación automática.

En la clasificación automática es necesario contar con un corpus de entrenamiento. Por lo tanto, se llevó a cabo la construcción de un corpus de personalidad con las características del texto ya mencionadas. En la figura 3 se muestra el diagrama del modelo propuesto. Este modelo se divide en cinco fases principales:

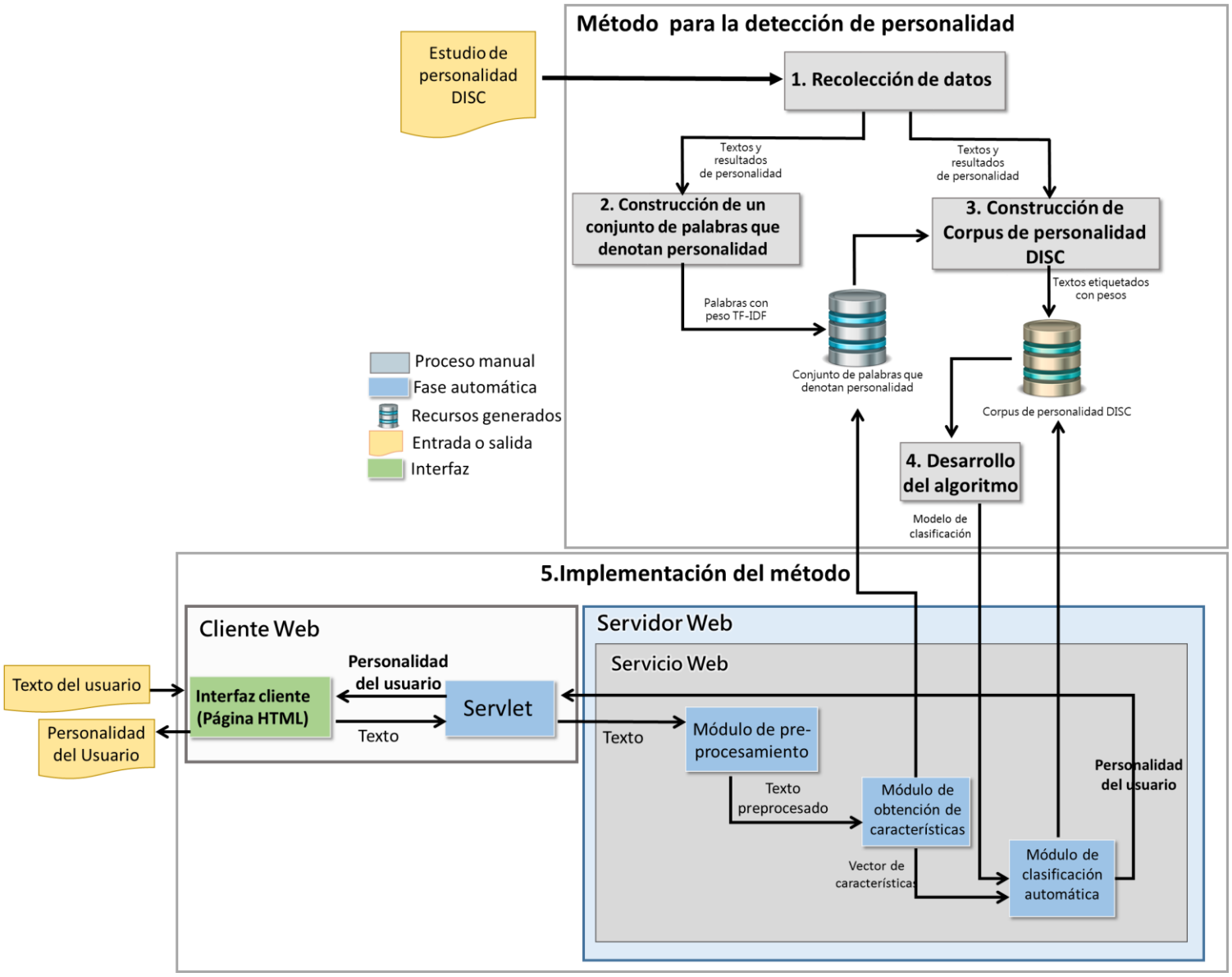


Figura 2. Modelo para la detección de personalidad.

4.2 Recolección de datos

En esta fase se llevó a cabo la aplicación de un estudio con base en el modelo *DISC* para conocer la personalidad de 120 personas, este estudio se anexo a un formato de encuesta que tenía como objetivo reunir datos demográficos y un texto de cada persona que tomaba la prueba para posteriormente, utilizarlos para la construcción de un conjunto de palabras que denoten personalidad y un corpus de personalidad. El formato de la encuesta se presenta en el anexo 1 *formato de encuesta DISC*. A continuación se describen los elementos que integran la fase de recolección de datos:

- **Perfil de los participantes:** el perfil de los participantes fue: estudiantes de grado universitario, de edades entre 20 a 30 años, 49 personas de sexo femenino y 71 personas de sexo masculino, formando un total de 120 personas, en su mayoría solteros.

- **Materiales:** El formato de encuesta que se aplicó estaba formado por una encuesta de datos demográficos, una encuesta con los reactivos del estudio de personalidad *DISC* y una sección para escribir un texto. Los materiales utilizados para la aplicación de la prueba de personalidad *DISC* son:

  Encuesta de datos demográficos: Esta parte de la prueba reunía información demográfica de la persona que respondía la encuesta, datos como; el rango de edad, red social que utiliza, número de amigos en redes sociales, sexo, ocupación, escolaridad, correo electrónico y estado civil.

  Encuesta de reactivos del estudio DISC: Esta parte de la encuesta está formada por los reactivos del estudio para determinar la personalidad basada en el modelo DISC que incluye un conjunto de 28 tablas con 4 grupos de palabras cada una. En este estudio la persona que responde tiene que seleccionar las palabras que considere la identifican más y las que consideren que la identifican menos.

  Sección de texto: En esta parte de la encuesta se presenta un espacio con líneas para que el participante escriba un texto del tema de su preferencia.

- **Procedimiento:**

  Se acudió a un Instituto Tecnológico para aplicar el estudio de personalidad. Ubicados en la entrada principal del Instituto se preguntó a las personas que pasaban si deseaban responder la encuesta para conocer su personalidad, también se acudió a algunas aulas para conseguir más participantes. A cada uno de los participantes se les explicó el procedimiento para responder la encuesta y se les solicitó un correo electrónico para informarles sus resultados después de evaluar sus encuestas.

Parte 1 (Encuesta de datos demográficos): Se solicita al participante información personal, como nombre, rango de edad, estado civil, etc.

Parte 2 (Encuesta de reactivos del estudio DISC) : Se presenta el cuestionario del estudio de personalidad DISC que consta de 28 grupos de palabras en donde hay que elegir solo un par de ellas en cada grupo (una palabra con la que más se identifique y una con la que menos se identifique la persona que está tomando la prueba).

Parte 3 (Sección de texto): Al final del estudio se solicitó a las personas que escribieran un texto del tema de su preferencia, para esta actividad se agregó una hoja con líneas a la prueba para que los participantes pudieran escribir su texto.

- **Resultados**

Los resultados principales obtenidos de la aplicación del estudio para conocer la personalidad basado en el modelo *DISC* se muestran en la tabla 2.

Genero	Estado civil	Rango de edad	Escolaridad	Resultado de la prueba	
49 Mujeres	2 casadas	De 20 a 30 años	Universidad (Carrera)	Dominante	8
				Influyente	10
	47 solteras			Estable	24
				Concienzudo	7
71 Hombres	2 casados	De 20 a 30 años	Universidad (Carrera)	Dominante	6
				Influyente	16
	69 solteros			Estable	38
				Concienzudo	11
Total = 120					

Tabla 2. Resultados de la aplicación del estudio de personalidad DISC.

4.2.1 Construcción de un conjunto de palabras que denotan personalidad

En esta fase se llevó a cabo la construcción de un conjunto de palabras que denotan personalidad basándose en los pesos obtenidos mediante la ponderación TF-IDF. Las actividades desarrolladas en esta fase son:

- Análisis de encuestas y textos
- Conteo y registro de adjetivos y verbos

- Cálculo de pesos TF-IDF
- Aplicación de la fórmula de pesos TF-IDF al conjunto de palabras que denotan personalidad

Análisis de encuestas y textos

En esta actividad todas las encuestas aplicadas fueron reunidas, contabilizadas, y evaluadas de acuerdo con el modelo de personalidad *DISC.* También se analizaron los textos, reunidos mediante la aplicación del estudio de personalidad en la fase de recolección de datos. Los textos reunidos ya se encontraban clasificados de acuerdo con el resultado obtenido en las encuestas de personalidad.

Para poder calcular la personalidad, de los textos se decidió tomar palabras para realizar el proceso de clasificación, estas palabras son adjetivos y verbos.

Se decidió tomar esas palabras porque se considera que son palabras representativas en los textos. El objetivo es formar una tabla de términos, en la cual se puedan ver las veces que esos términos aparecen en unos de los factores de personalidad *DISC* y en otros no. A modo de ejemplo se muestra a continuación la tabla 3.

Palabra	Dominante	Influyente	Estable	Concienzudo
Agradable	3	6	14	4
Comer	1	0	0	0
Productivo	3	5	20	3
Principal	3	6	5	3
Comprar	1	0	1	0
Gustar	4	15	24	9
Romántico	19	12	52	8
Ir	17	31	38	15
Pasear	1	1	3	0

Tabla 3. Ejemplo de tabla de registro de apariciones de palabras.

En esta tabla se pueden apreciar las veces que aparecen algunas palabras en cada factor de personalidad, esta información es útil para poder asignarle un peso a cada una. Este análisis sirvió para poder notar que algunas palabras son más o menos veces utilizadas en cada factor. Pero se desconoce qué valor asignar a cada palabra y como asignar ese valor, para ello se investigó una fórmula de cálculo de pesos, de la cual se hablará mas adelante.

Conteo y registro de adjetivos y verbos

En esta actividad se llevó a cabo la revisión de los textos reunidos en la fase de recolección de datos para identificar, contar y registrar manualmente los adjetivos y verbos de cada uno de los textos.

Para realizar esta actividad se utilizó la herramienta *Freeling* ("Freeling," 2018) para *lematizar* cada texto y para etiquetar las palabras con el objetivo de poder identificarlas correctamente.

Otros de los motivos para utilizar *Freeling* es que para la identificación de la personalidad el proceso de *lematización* del texto tiene que ser automatizado y esta herramienta facilita ese proceso. Además al identificar los lemas se facilita el conteo de palabras con el mismo lema que se encuentren conjugadas de distintas formas.

El procedimiento para realizar esta actividad fue el siguiente:

1. Se tomó cada uno de los textos y se procesaron con la herramienta *Freeling* para obtener las etiquetas y los lemas de las palabras de cada texto. En la figura 4 se puede observar el ejemplo del procesamiento de un texto en *Freeling.*

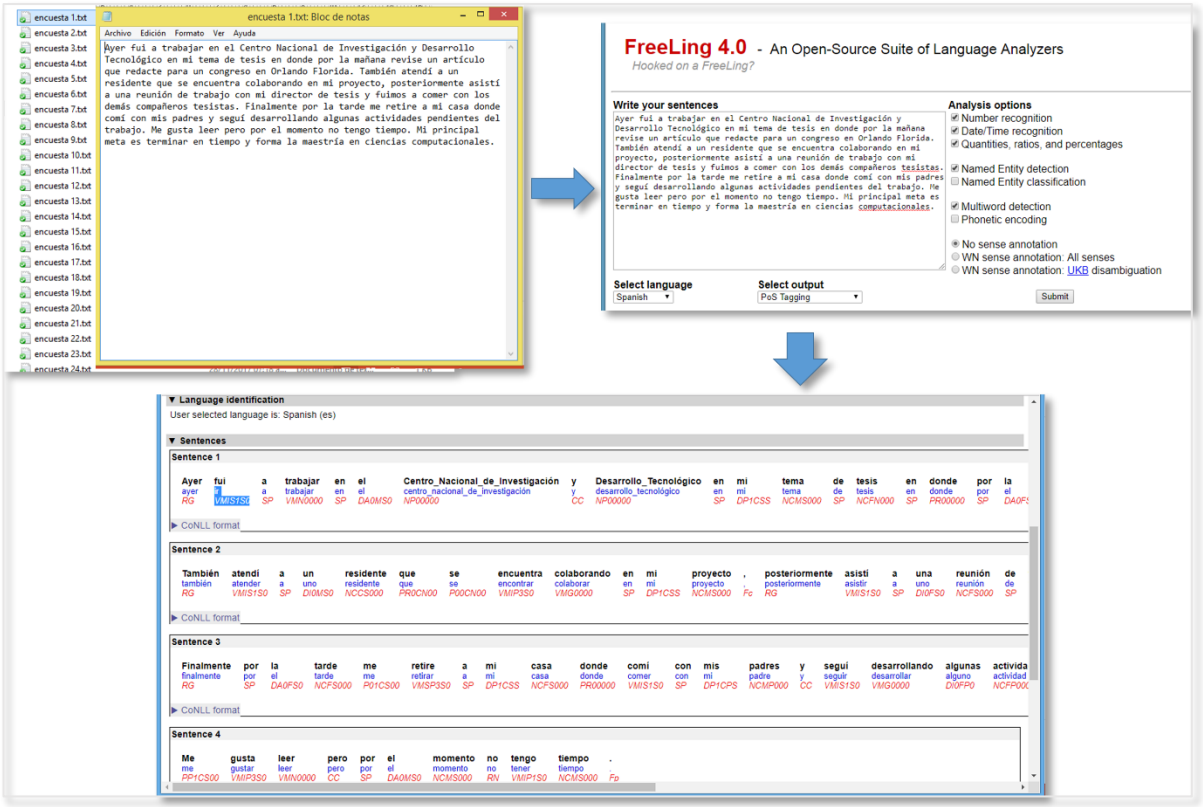
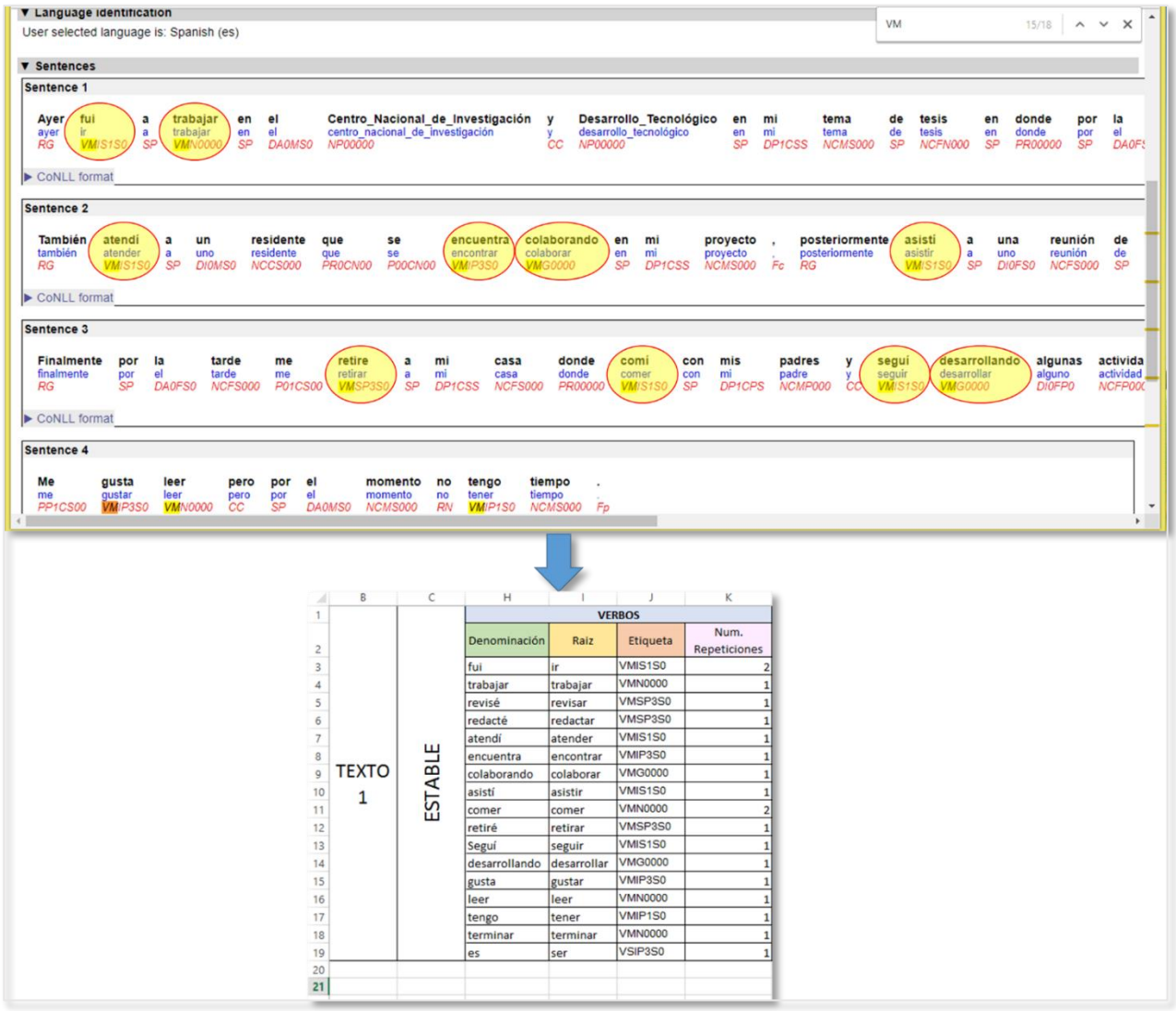


Figura 3. Ejemplo del procesamiento de un texto en la herramienta *Freeling*.

Los resultados que se obtuvieron de *Freeling* fueron revisados manualmente uno a uno para encontrar los adjetivos y verbos que contenían cada uno de los textos. La ocurrencia de estas palabras fue registrada en una tabla para su uso posterior. En la figura 5 se muestra un ejemplo del proceso de revisión y registro de cada verbo de un texto del corpus.



	B	C	H	I	J	K
1			VERBOS			
2			Denominación	Raiz	Etiqueta	Num. Repeticiones
3	TEXTO 1	ESTABLE	fui	ir	VMIS1S0	2
4			trabajar	trabajar	VMN0000	1
5			revisé	revisar	VMSP3S0	1
6			redacté	redactar	VMSP3S0	1
7			atendí	atender	VMIS1S0	1
8			encuentra	encontrar	VMIP3S0	1
9			colaborando	colaborar	VMG0000	1
10			asistí	asistir	VMIS1S0	1
11			comer	comer	VMN0000	2
12			retiré	retirar	VMSP3S0	1
13			Seguí	seguir	VMIS1S0	1
14			desarrollando	desarrollar	VMG0000	1
15			gusta	gustar	VMIP3S0	1
16			leer	leer	VMN0000	1
17			tengo	tener	VMIP1S0	1
18			terminar	terminar	VMN0000	1
19			es	ser	VSIP3S0	1
20						
21						

Figura 4. Ejemplo del proceso de revisión y registro de los verbos de un texto

2. Finalmente los resultados de los conteos se reunieron y se registraron en cuatro libros de Excel, cada uno correspondiente a uno de los cuatro tipos de personalidad del modelo DISC.

   En la figura 6 se muestran los registros de los adjetivos y verbos utilizados en uno de los textos que obtuvo un resultado de personalidad dominante.

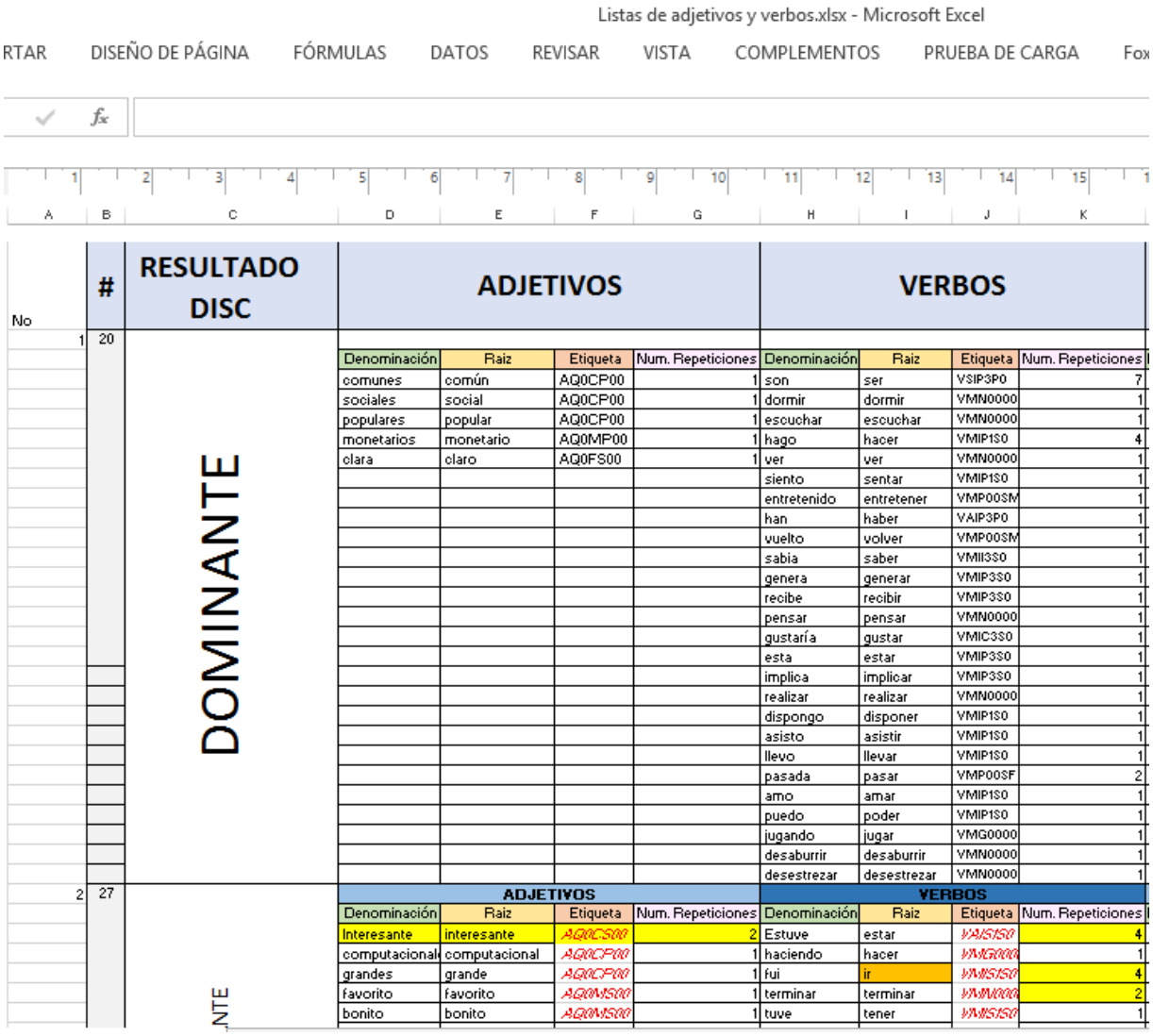


Figura 5. Registros de adjetivos y verbos

Cálculo de pesos TF-IDF

En esta actividad se analizó la fórmula para el cálculo de pesos o ponderación TF-IDF. Hasta este punto ya se cuenta con el registro de todos los adjetivos y verbos de todos los textos reunidos por las personas que tomaron el estudio de personalidad basado en el modelo *DISC*. Pero se desconoce qué valor asignar a cada palabra y como asignar ese valor.

Para saber cómo asignar el peso a cada término se hace uso de la fórmula de ponderación TF-IDF o fórmula para cálculo de pesos de términos.

La ponderación de los términos es el proceso que tiene como finalidad conocer la importancia de los términos para representar un documento y permitir su posterior recuperación. Esto implica que se debe determinar la capacidad de los términos para representar el contenido de los documentos en la colección, que permitan identificar cuáles son relevantes o no ante la consulta del usuario (Vera, 2017).

Al valor e índice que es capaz de determinar este extremo se le denomina "peso del término" o "ponderación del término" y su cálculo implica determinar la "Frecuencia de aparición del término TF" y la "Frecuencia inversa del documento para un término IDF".

El peso local se denomina ***Term Frequency (TF),*** y se calcula contando el número de veces que la palabra aparece en el documento dividido entre el número total de palabras contenidas en él mismo:

$$\mathrm{TF} = \frac{\text{El número de veces que la palabra aparece en el documento}}{\text{El número total de palabras contenidas en el documento}}$$

Contar el número de veces que una palabra aparece en un documento nos da el peso local de esa palabra en el documento.

Contar el número de documentos en los que aparece esa palabra, aunque sea una vez, nos da el peso global de ella con respecto a la colección de documentos que se está analizando.

El peso global se denomina Frecuencia Inversa del Documento (*Inverse Document Frequency*, *IDF*), y se calcula dividiendo el número total de documentos de la colección entre el número de documentos que contienen la palabra, y se calcula el logaritmo de ese valor. Al valor resultante se le suma 1 para corregir los valores para los términos con IDF muy bajos:

$$IDF = \mathrm{Log10}\frac{\text{Total de documentos}}{\text{Número de documentos donde aparece el termino}} + 1$$

Ahora, solamente se multiplican ambos pesos, y a esta técnica para asignar pesos a las palabras se le llama ***TF-IDF.*** El peso de un término en un documento es el producto de su frecuencia de aparición en dicho documento (TF) y su frecuencia inversa de documento (IDF) como se muestra en la siguiente fórmula:

$$TFIDF_{i,d} = TF_{i,d}\, x IDF_i$$

La fórmula TF-IDF permitirá saber los pesos de cada palabra en cada factor de personalidad *DISC*. Para ello se necesita saber el número de ocurrencias totales de las palabras en cada texto y en cada documento, así como el número total de palabras de cada documento y el número total de documentos.

De acuerdo con el análisis realizado se decidió reunir los textos correspondientes al mismo tipo de personalidad en un solo documento. De esta manera se crearon cuatro documentos que contenían a todos los demás textos con la misma personalidad.

Cada documento corresponde a cada uno de los factores del modelo *DISC* y se trabajó sobre esos cuatro documentos generados a partir de la unión de los textos.

En la figura 7 podemos ver de manera representativa como se agruparon los textos de los cuatro perfiles de personalidad *DISC.*

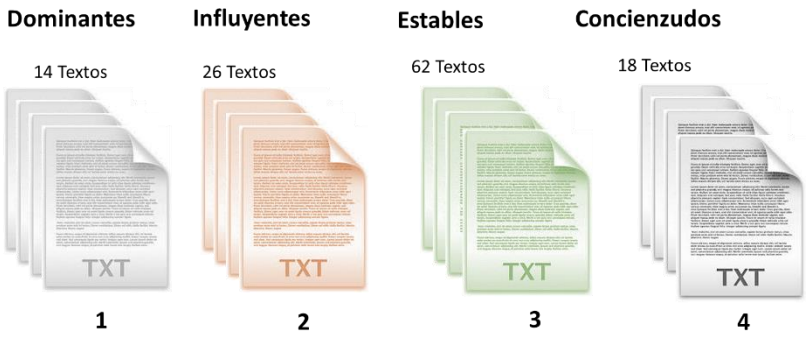


Figura 6. Agrupación de los textos de los cuatro perfiles de personalidad DISC

Esta decisión se tomó debido a que la fórmula TF-IDF nos indica el peso de cada término en un documento respecto a los otros documentos, pero no se toma en cuenta la clase del documento es decir la personalidad a la que pertenece.

Al aplicarse la fórmula sobre los 120 textos de manera independiente se obtienen diferentes pesos para las palabras que corresponden al mismo tipo de personalidad y no el peso de cada palabra en cada tipo de personalidad. Lo que se busca es la capacidad de cada término para representar la personalidad en los textos de la colección.

Para utilizar la fórmula TF-IDF requerimos de las variables que la conforman. Para poder obtener las variables y calcular primeramente la frecuencia de término TF, que es la primera parte de la fórmula, se llevó a cabo el conteo del número de veces que aparece cada adjetivo y verbo en cada documento y el número total de palabras de cada documento.

La primera parte del procedimiento para contar las apariciones de cada término en los documentos se llevó a cabo revisando uno por uno en las tablas de registros de adjetivos y verbos de cada uno de los 120 textos, generados en la actividad anterior. Las palabras y su cantidad de apariciones se marcaron con un color diferente para cada personalidad, esto para facilitar su identificación en el proceso de conteo manual.

En la figura 8 se muestra la primera parte del procedimiento que se siguió para el conteo y registro de las apariciones de los adjetivos y verbos de cada uno de los textos.

En la segunda parte del procedimiento se reunieron todas las palabras contabilizadas y marcadas con un color diferente y se fueron registrando una a una sus apariciones en cada uno de los factores del modelo *DISC* de manera manual en una tabla.

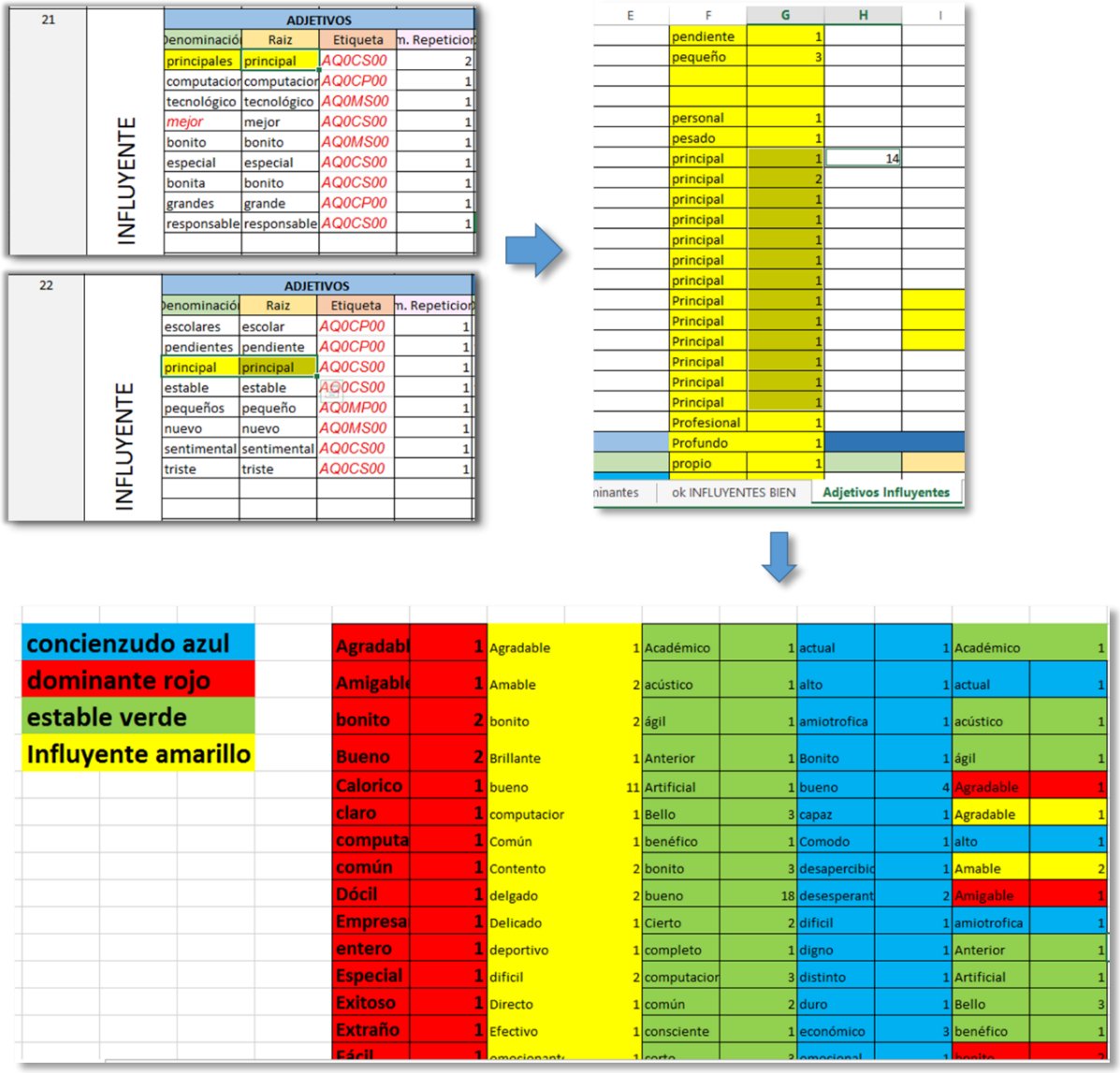


Figura 7. Primera parte del procedimiento de conteo de verbos y adjetivos

En la figura 9 se muestra la segunda parte del procedimiento que se siguió para el conteo y registro de las apariciones de los adjetivos y verbos de cada uno de los textos.

		Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece
callar	1	0	1	0	0	1
cambiar	5	0	0	5	2	2
caminar	1	1	0	3	0	2
Cansar	1	0	2	1	0	2
cantar	1	0	1	1	1	3
Capturar	1	0	0	1	0	1
Caracterizar	1	0	0	1	0	1
Casar	1	0	1	3	0	2
catalogar	1	0	0	1	0	1
cenar	1	0	1	4	0	2
cenar	4					
Centrar						
Cerrar						
Charlar	1					
Clasificar	1					

Figura 8. Segunda parte del procedimiento de conteo de adjetivos y verbos

Una vez que se obtuvieron las apariciones totales de cada palabra en cada uno de los textos y la cantidad de textos en los que aparecen, nos faltaba una variable para poder aplicar la fórmula y esa variable es la cantidad total de palabras de cada uno de los textos.

Para contar el número total de palabras en cada uno de los textos se utilizó la herramienta llamada *AntConc*. Esta herramienta permite contar el número de palabras de un documento, pero además nos permite eliminar las *stopwords* del documento (Laurence, 2018).

Las *stopwords* son palabras que carecen de un significado por si solas y que se repiten muchas veces en los textos ocasionando ruido. En el procedimiento automatizado estas palabras tienen que ser eliminadas para limpiar el texto.

Para contar el número total de palabras de los cuatro documentos con los que estamos trabajando se realizó el siguiente procedimiento:

1) Se creó una lista de *stopwords* para poder cargarlas en la herramienta *AntConc.*

2) Se cargaron uno a uno los cuatro documentos correspondientes a cada personalidad del modelo *DISC* y también se cargó la lista de *stopwords* a eliminar del texto en la herramienta *AntConc*.

3) Se utilizó la herramienta *AntConc* mediante su función de conteo de palabras y nos proporcionó el número total de palabras en cada uno de los textos. En la figura 10 se muestra el ejemplo de uso de la herramienta *AntConc* que consiste en cargar el archivo que contiene el texto que se va a procesar, en este caso el archivo es CONCIENZUDO.txt, cargar la lista de *stopwords* en el menú *tool preferences* y presionar el botón *Start* para obtener el número total de palabras sin contar las *stopwords.*

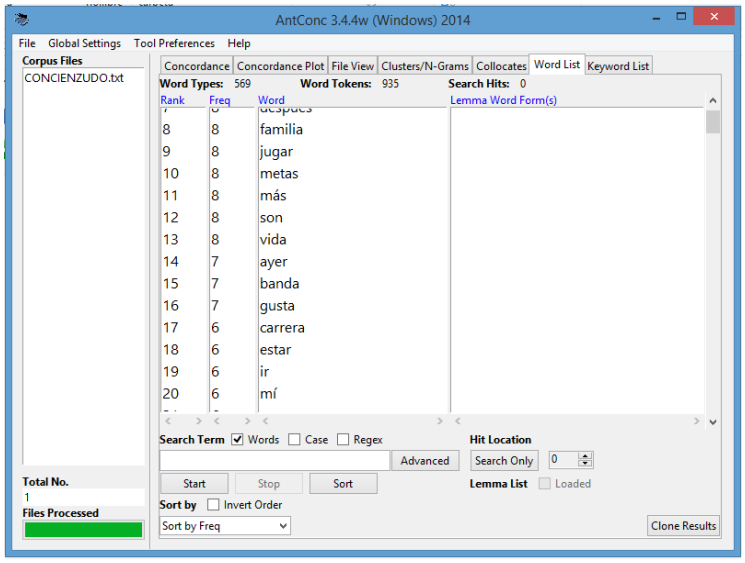


Figura 9. Ejemplo de uso de la herramienta AntConc, conteo de palabras del texto.

Aplicación de la fórmula de pesos TF-IDF al conjunto de palabras que denotan personalidad

En esta actividad se realizó el registro de cada una de las palabras identificadas (adjetivos y verbos) de los textos, con sus respectivos pesos TF-IDF.

Una vez que se contó el número de palabras en cada uno de los textos, ya se tenían las variables necesarias para poder aplicar la fórmula TD-IDF y formar el conjunto de palabras que denoten personalidad. La tabla 4 muestra la cantidad total de palabras en cada uno de los cuatro textos de cada personalidad una vez eliminadas las *stopwords*.

Factor DISC	Total de palabras
Dominante	528
Influyente	1182
Estable	2216
Concienzudo	728

Tabla 4. Cantidad total de palabras en cada uno de los cuatro textos de cada personalidad**.**

Para aplicar la fórmula a las palabras se agregaron cuatro columnas más a la tabla de registro como se muestran en la tabla 5. En estas columnas se introdujo la fórmula de ponderación TF-IDF sustituyendo los valores de las variables por los valores de aparición y número de palabras que ya se tenían registrados en la tabla. El conjunto completo de palabras se presenta en el anexo 3 *Conjunto de palabras que denotan personalidad basado en el modelo DISC.*

El propósito de la creación de este conjunto de palabras que denoten personalidad es emplearlo para identificar la personalidad de los textos. Con la ayuda de este conjunto de palabras se construyó el corpus mediante el etiquetado de cada uno de los textos, con sus correspondientes valores TF-IDF de las palabras (adjetivos y verbos) que lo integran como se explica en la siguiente fase del modelo de detección de personalidad presentado en este trabajo de investigación.

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Académico	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
actual	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
acústico	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
ágil	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Agradable	1	1	0	0	2	0.0025	0.0011	0.0000	0.0000	Adjetivo
alto	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Amable	0	2	0	0	1	0.0000	0.0027	0.0000	0.0000	Adjetivo
Amigable	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
amiotrofica	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo

Tabla 5. Columnas adicionales de ponderación de términos TF-IDF

4.3 Construcción de corpus de personalidad DISC

En esta actividad se hizo uso del conjunto de palabras que denotan personalidad para poder calcular el peso que le corresponde a cada uno de los textos en cada factor del modelo *DISC*.

Posteriormente con los pesos correspondientes a cada texto se complementó el corpus etiquetándolo con los valores obtenidos.

Se revisó manualmente cada uno de los textos para identificar los adjetivos y verbos que contenían y así asignarles un peso correspondiente a cada personalidad utilizando el conjunto de palabras que denotan personalidad.

El procedimiento para realizar esta actividad es el siguiente:

1. Los textos obtenidos por las encuestas se revisaron manualmente uno a uno, para poder identificar los adjetivos y verbos que los integraban. Después se buscaron uno a uno sus pesos correspondientes a cada factor *DISC* en el conjunto de palabras que denotan personalidad, como se muestra en la figura 11.

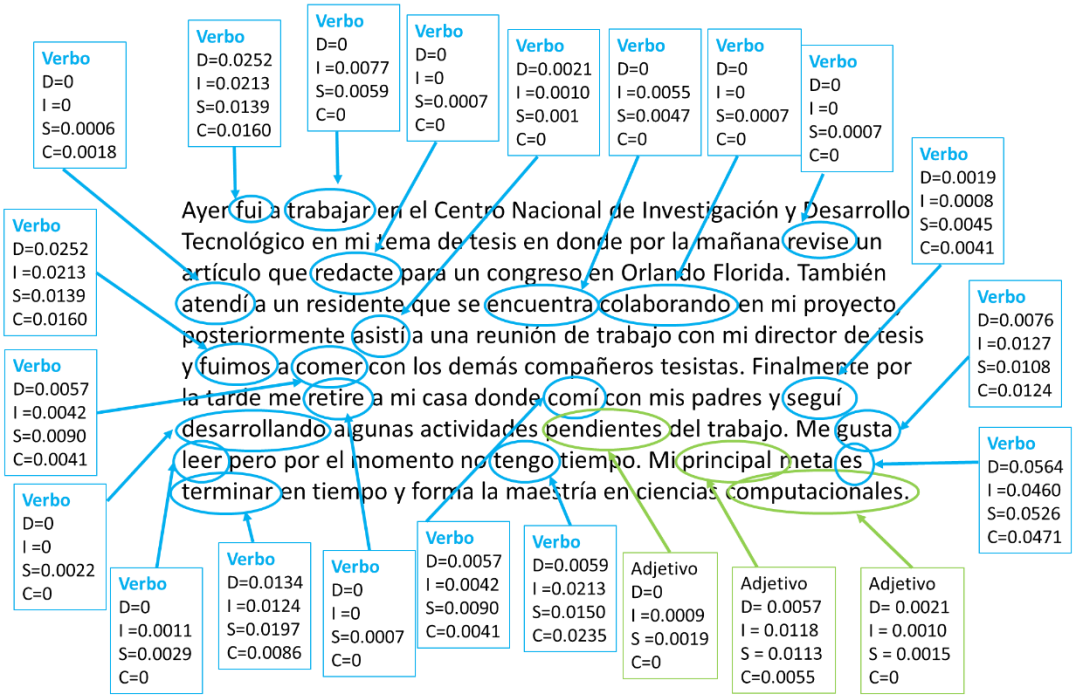


Figura 10. Ejemplo de etiquetado de un texto.

Todos los resultados se registraron en tablas separadas, una por cada texto, quedando así 120 tablas. La tablas 6 y 7 muestran un ejemplo del registro de resultados de dos textos del corpus.

	TEXTO NUMERO 2		NUMERO DE ENCUESTA:		27									
Palabra (Lema)	Apariciones DOMINANTES	Apariciones INFLUYENTES	Apariciones ESTABLES	Apariciones CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo	Repeticiones				
						Dominante	Influyente	Estable	Concienzudo						
						D	I	S	C			D	I	S	C
interesante	2	0	1	0	2	0.0049	0.0000	0.0006	0.0000	Adjetivo	2	0.0099	0.0000	0.0012	0.0000
computacional	1	1	3	0	3	0.0021	0.0010	0.0015	0.0000	Adjetivo	1	0.0021	0.0010	0.0015	0.0000
grande	1	1	1	3	4	0.0019	0.0008	0.0005	0.0041	Adjetivo	1	0.0019	0.0008	0.0005	0.0041
favorito	4	7	25	5	4	0.0076	0.0059	0.0113	0.0069	Adjetivo	1	0.0076	0.0059	0.0113	0.0069
bonito	2	2	3	1	4	0.0038	0.0017	0.0014	0.0014	Adjetivo	1	0.0038	0.0017	0.0014	0.0014
ir	17	31	38	15	4	0.0322	0.0262	0.0171	0.0206	Verbo	4	0.1288	0.1049	0.0686	0.0824
terminar	9	18	54	8	4	0.0170	0.0152	0.0244	0.0110	Verbo	2	0.0341	0.0305	0.0487	0.0220
finalizar	1	1	0	0	2	0.0025	0.0011	0.0000	0.0000	Verbo	1	0.0025	0.0011	0.0000	0.0000
visitar	1	1	1	3	4	0.0019	0.0008	0.0005	0.0041	Verbo	1	0.0019	0.0008	0.0005	0.0041
encantar	2	1	4	1	4	0.0038	0.0008	0.0018	0.0014	Verbo	1	0.0038	0.0008	0.0018	0.0014
salir	4	21	29	4	4	0.0076	0.0178	0.0131	0.0055	Verbo	1	0.0076	0.0178	0.0131	0.0055
caminar	1	0	3	0	2	0.0025	0.0000	0.0018	0.0000	Verbo	1	0.0025	0.0000	0.0018	0.0000
correr	1	2	2	0	3	0.0021	0.0019	0.0010	0.0000	Verbo	1	0.0021	0.0019	0.0010	0.0000
jugar	5	11	21	15	4	0.0095	0.0093	0.0095	0.0206	Verbo	1	0.0095	0.0093	0.0095	0.0206
gustar	4	15	24	9	4	0.0076	0.0127	0.0108	0.0124	Verbo	1	0.0076	0.0127	0.0108	0.0124
viajar	1	6	5	5	4	0.0019	0.0051	0.0023	0.0069	Verbo	1	0.0019	0.0051	0.0023	0.0069
pasar	7	8	22	5	4	0.0133	0.0068	0.0099	0.0069	Verbo	1	0.0133	0.0068	0.0099	0.0069
				TOTAL =		0.1221	0.1072	0.1073	0.1016			0.2406	0.2011	0.1837	0.1745

Tabla 6. Ejemplo del registro de resultados de un texto de personalidad estable

	TEXTO NUMERO 1		NUMERO DE ENCUESTA:		1										
Palabra (Lema)	Apariciones DOMINANTES	Apariciones INFLUYENTES	Apariciones ESTABLES	Apariciones CONCIENZUDO	Doc. en los que aparece	Pesos en cada Factor				Tipo	Repeticiones en este texto				
						Dominante	Influyente	Estable	Concienzudo						
						D	I	S	C			D	I	S	C
computacional	1	1	3	0	3	0.0021	0.0010	0.0015	0.0000	Adjetiv	1	0.0021	0.0010	0.0015	0.0000
ir	17	31	38	15	4	0.0322	0.0262	0.0171	0.0206	Verbo	2	0.0644	0.0525	0.0343	0.0412
trabajar	0	7	10	0	2	0.0000	0.0077	0.0059	0.0000	Verbo	1	0.0000	0.0077	0.0059	0.0000
revisar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo	1	0.0000	0.0000	0.0007	0.0000
redactar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo	1	0.0000	0.0000	0.0007	0.0000
atender	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Verbo	1	0.0000	0.0000	0.0006	0.0018
encontrar	0	5	8	0	2	0.0000	0.0055	0.0047	0.0000	Verbo	1	0.0000	0.0055	0.0047	0.0000
colaborar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo	1	0.0000	0.0000	0.0007	0.0000
asistir	1	1	3	0	3	0.0021	0.0010	0.0015	0.0000	Verbo	1	0.0021	0.0010	0.0015	0.0000
Comer	3	5	20	3	4	0.0057	0.0042	0.0090	0.0041	Verbo	2	0.0114	0.0085	0.0181	0.0082
retirar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo	1	0.0000	0.0000	0.0007	0.0000
Seguir	1	1	10	3	4	0.0019	0.0008	0.0045	0.0041	Verbo	1	0.0019	0.0008	0.0045	0.0041
desarrollar	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Verbo	1	0.0000	0.0000	0.0022	0.0000
gustar	4	15	24	9	4	0.0076	0.0127	0.0108	0.0124	Verbo	1	0.0076	0.0127	0.0108	0.0124
Leer	0	1	5	0	2	0.0000	0.0011	0.0029	0.0000	Verbo	1	0.0000	0.0011	0.0029	0.0000
terminar	9	18	54	8	4	0.0170	0.0152	0.0244	0.0110	Verbo	1	0.0170	0.0152	0.0244	0.0110
				TOTAL =		0.0687	0.0754	0.0881	0.0540			0.1065	0.1059	0.1143	0.0787

Tabla 7. Ejemplo del registro de resultados de un texto de personalidad dominante

2. Una vez que se obtuvieron los pesos totales de cada uno de los textos se realizó el etiquetado del corpus, que consistió en registrar esos pesos de cada uno de los textos contenidos en el corpus. La tabla 8 muestran algunos de los textos del corpus etiquetado con sus respectivos pesos TF-IDF en cada factor de personalidad *DISC*.

Num_encuesta	NumPalabras	Texto	DominanteTFIDF	InfluyenteTFIDF	StableTFIDF	ConcienzudoTFIDF	Personalidad
1	53	Ayer fui a trabajar en el Centro Nacional de Investigación y Desarrollo Tecnológico en mi tema de tesis en donde por la mañana revise un artículo que redacte para un congreso en Orlando Florida. También atendí a un residente que se encuentra colaborando en mi proyecto, posteriormente asistí a una reunión de trabajo con mi director de tesis y fuimos a comer con los demás compañeros tesistas. Finalmente por la tarde me retire a mi casa donde comí con mis padres y seguí desarrollando algunas actividades pendientes del trabajo. Me gusta leer pero por el momento no tengo tiempo. Mi principal meta es terminar en tiempo y forma la maestría en ciencias computacionales.	0.0308	0.0501	0.0669	0.0320	ESTABLE
2	33	Mi pasatiempo favorito es pasarlo con mi familia en casa, platicar en la mesa a la hora de la comida, y hablar de todo lo sucedido durante el día de cada uno de nosotros, pedir opinión acerca de los acontecimientos a mis padres y como poder resolver algún problema que se halla presentado en la semana en el día tanto puede ser en casa, como en la escuela, adoro hablar con ellos y por este motivo se convierte en mi	0.0612	0.0601	0.0810	0.0495	ESTABLE
3	41	En el día 14/11/16 Me levante temprano y no asistí al servicio social para poder arreglar los documentos necesarios para tramitar una beca, tuve que hacer fila en la escuela y me mantuve parado alrededor de 30 min. La fila era muy larga. Al fin tramité un documento en la escuela, después camine hacia el ayuntamiento a tramitar otro documento, fue menos labor puesto que la fila era muy corta. Cuando termine regrese a casa a terminar tarea pendiente y a esperar la hora para entrar a la escuela.	0.0261	0.0346	0.0494	0.0168	ESTABLE
4	32	Lo que deseo en la vida es, realizar todas mis metas, lo primordial es terminar mi carrera, para poder trabajar y ayudar a mis padres. Formar una familia hermosa y poder disfrutar de ella día con día. Ser una persona de bien para la sociedad y el país, deseo que en el futuro la violencia ya termine, y que solo haya buenas personas en el mundo, sé que suena imposible, pero eso es lo que yo deseo.	0.0291	0.0555	0.0604	0.0394	ESTABLE
5	38	Dentro de mis principales metas a corto plazo, es concluir mis estudios, titularme, poder encontrar un trabajo próximo con el que pueda sustentarte para tener una vida digna; ya que en la actualidad nos enfrentamos a muchos problemas desempleo, crisis económica y el daño a nuestro medio ambiente. Desarrollarme profesionalmente es un logro personal y satisfactorio para mí y para las personas que me han ayudado y apoyado durante mis años de estudio a lo largo de mi vida.	0.0284	0.0637	0.0788	0.0380	ESTABLE

Tabla 8. Textos etiquetados con sus respectivos pesos TF-IDF.

4.4 Desarrollo del algoritmo

Para llevar a cabo el entrenamiento del algoritmo *SMO*, se utilizó el corpus de personalidad *DISC,* que fue creado en la Fase 4 "*construcción de corpus de personalidad DISC*" del modelo para la detección de personalidad.

Una vez entrenado el algoritmo *SMO* se generó un modelo de clasificación, utilizando la herramienta *Weka*. El modelo muestra los atributos y el algoritmo utilizados, también se muestran los pesos asignados a cada uno de los atributos del texto que aparecen en el corpus de entrenamiento.

Los pesos se generan de dos en dos clases, por ejemplo: concienzudo y dominante, concienzudo y estable, concienzudo influyente y así sucesivamente con todas las combinaciones posibles. En este caso se presentan los pesos asignados a los atributos entre la clase concienzudo y dominante, concienzudo y estable y concienzudo e influyente. En la figura 13 se muestran dos partes del modelo de clasificación generado por *Weka*.

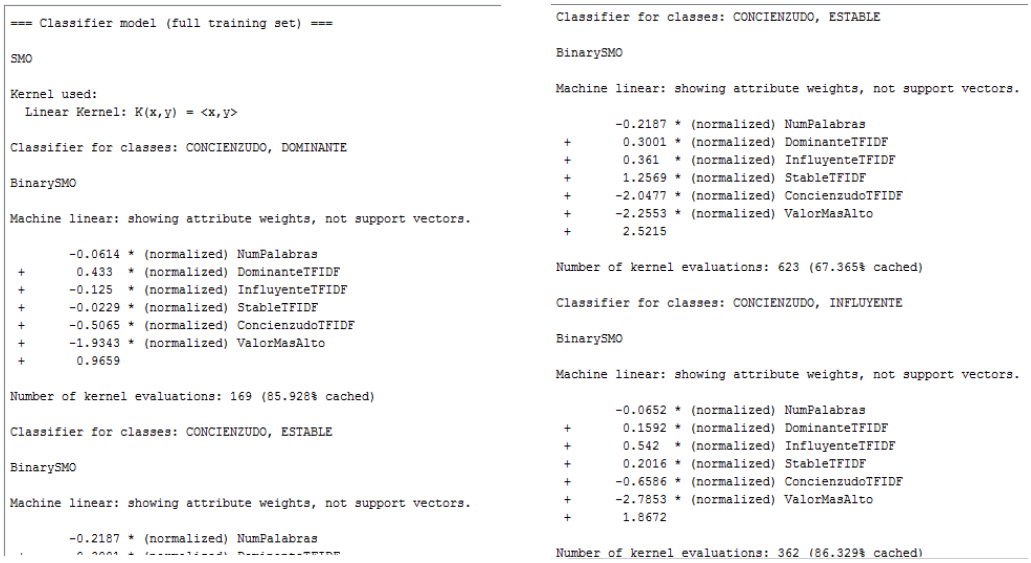


Figura 11. Modelo de clasificación generado por Weka.

El algoritmo de clasificación construye un modelo de clasificación a partir de las características obtenidas del corpus de entrenamiento. El propósito de generar un modelo de clasificación es para poder realizar la clasificación de los textos de los cuales no conocemos su clase, en este caso la personalidad a la que pertenece.

4.5 Implementación del modelo

En esta fase se desarrolló un servicio Web para la detección de personalidad, el cual consta de tres módulos, (I) Módulo de pre-procesamiento, (II) Módulo de obtención de características y (III)

Módulo de clasificación automática. El servicio Web ejecuta estos módulos siempre que se realiza el análisis de un texto. En la figura 12 se muestran los módulos que componen la aplicación de detección de personalidad.

El módulo de pre-procesamiento hace uso de la herramienta *freeling* para las tareas de lematización y etiquetado gramatical del texto introducido por el usuario.

FreeLing es una biblioteca en lenguaje de programación C ++ que proporciona funcionalidades de análisis de lenguaje (análisis morfológico, detección de entidades nombradas, etiquetado PoS, análisis sintáctico, desambiguación de sentido de palabras, etiquetado de funciones semánticas, etc.) para una variedad de idiomas (inglés, español, portugués, italiano, francés, Alemán, ruso, catalán, gallego, croata, esloveno, entre otros)(Padró, 2011).

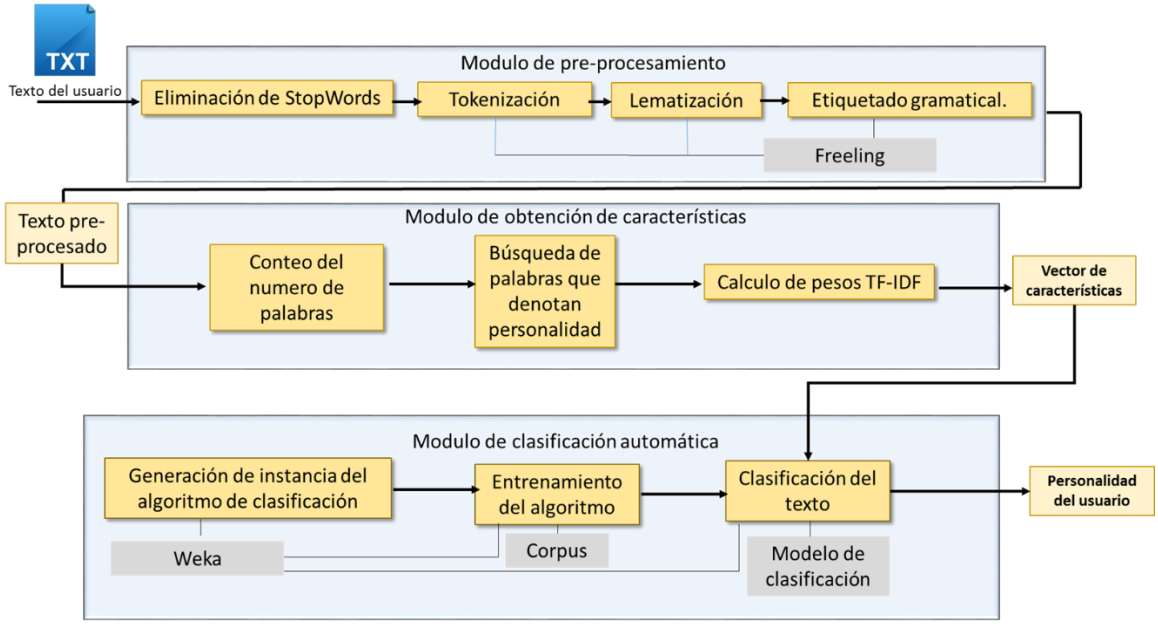


Figura 12. Módulos de la herramienta para la detección de personalidad.

En el segundo módulo se obtienen características del texto introducido. Estas características juntas conforman un vector llamado vector de características. El vector se genera a partir del texto introducido por el usuario, utilizando el conjunto de palabras que denotan personalidad en español. En este caso, se entiende como características el número de palabras en el texto, el peso TF-IDF general del texto en cada factor *DISC* y un identificador del factor que obtiene el valor TF-IDF más alto.

En el tercer módulo, el módulo de clasificación automática, se implementó el algoritmo SMO (*Sequential Minimal Optimization*) utilizando Weka. El algoritmo se entrenó con el corpus de personalidad que se formó anteriormente.

Una vez entrenado el algoritmo, se generó un modelo de clasificación. La detección de personalidad se realiza a partir del modelo de clasificación, se reciben las características del vector obtenido en el segundo módulo, obteniendo así la personalidad del texto.

4.5.1 Módulo de preprocesamiento

El primer módulo (pre-procesamiento), recibe como entrada un texto sin formato por parte de la interfaz cliente del servicio Web, el texto puede ser escrito por el usuario o cargado como un archivo en formato txt. En este módulo se realiza la *tokenizacion*, eliminación de *stopwords, lematizacion* y etiquetado gramatical del texto. A continuación, se describen brevemente cada una de estos procesos:

Tokenizacion: Primeramente el módulo toma el texto introducido y realiza una separación de cada palabra (*tokenización*).

Eliminación de *stopwords*: Consiste en eliminar las palabras que no aportan significado o información relevante, palabras como: artículos, pronombres, algunos verbos, conjunciones, etc. En esta tarea se carga una lista de *stopwords* automáticamente y se busca en esa lista cada una de las palabras que se obtuvieron del proceso de *tokenizacion,* y se eliminan las palabras que se encuentren en la lista. En el anexo 2 *lista de stopwords* se presenta el conjunto de *stopwords* utilizadas.

Lematización: la lematización consiste en convertir todas las palabras a su lema o forma raíz. Por ejemplo, *decir* es el lema de *dije*, pero también de *diré* o *dijéramos*, *guapo* es el lema de *guapas*, *mesa* es el lema de *mesas*. Para esta tarea se hace uso de la herramienta *freeling.*

El etiquetado gramatical: consiste en identificar la categoría gramatical a la que pertenece cada palabra que forma parte del texto del usuario. Por ejemplo: la palabra *"comer"* se etiqueta como verbo, la palabra *"el"* se etiqueta como artículo, etcétera. Para esta tarea se hace uso de la herramienta *freeling.*

4.5.2 Módulo de obtención de características

El módulo de obtención de características se implementó para extraer características del texto que aportan información para la detección de la personalidad. A continuación se describen brevemente las tareas que lleva a cabo este módulo:

Conteo de número de palabras: en esta tarea se obtiene como entrada el texto *tokenizado* y sin *stopwords* del módulo anterior y se obtiene como característica el número de palabras que contiene el texto.

Búsqueda de palabras de personalidad: como entrada de esta actividad el texto se recibe ya *tokenizado, lematizado* y sin *stopwords* del módulo anterior, se realiza una búsqueda de las palabras que integran el texto a analizar, la búsqueda se realiza en el conjunto de palabras que denotan personalidad construido, si la palabra se encuentra se le asigna un determinado peso correspondiente a cada uno de los factores de personalidad *DISC*, cada palabra tendrá asignado cuatro pesos diferentes (un peso por cada factor *DISC*).

Cálculo de pesos TF-IDF por frecuencia de palabras: enseguida se suman los pesos de cada una de las palabras encontradas para obtener un peso general del texto para cada uno de los factores de personalidad del modelo *DISC*.

Cálculo del valor TF-IDF más alto: Finalmente el módulo hace una comparación de los pesos TF-IDF del texto y, determina cual es el peso más alto obtenido y en que factor se encuentra. Estas características forman un vector que se envía al siguiente módulo para poder realizar la clasificación automática del texto y determinar la personalidad.

4.5.3 Módulo de clasificación automática

En este módulo se hace uso del algoritmo de clasificación SMO (*Sequential Minimal Optimization*), utilizando la librería para *Java* de la herramienta *Weka*. Se decidió utilizar este algoritmo debido a que mostró mejores resultados en comparación con otros en las pruebas de clasificación de la personalidad que se aplicaron y que se describen en el capítulo 6.

En este módulo se necesitan como entradas características que coincidan con las que se construyó el modelo de clasificación con el algoritmo *SMO*. Esas características son recibidas por este módulo a través del vector de características generado por el módulo anterior. En la figura 14 se muestra un ejemplo de la clasificación de un texto de un usuario.

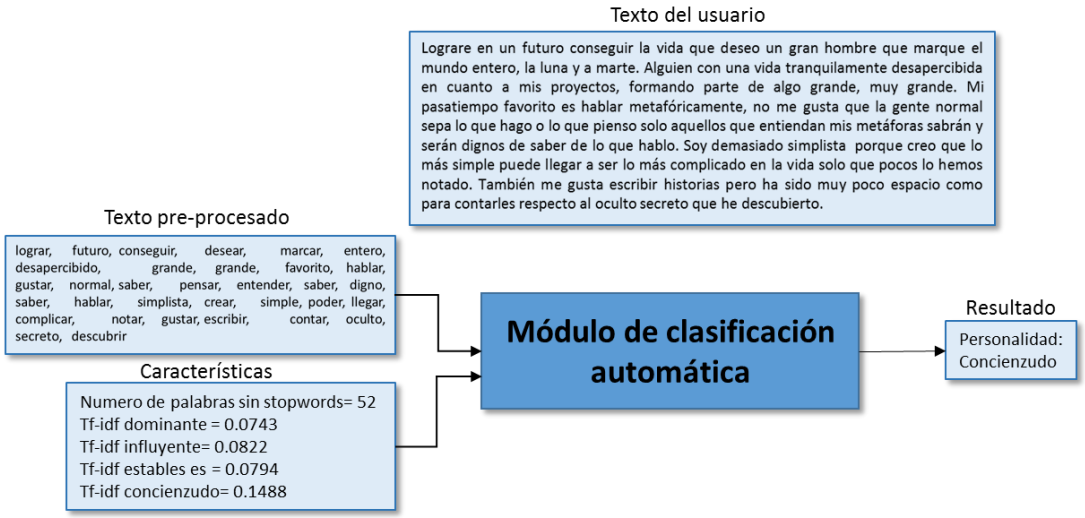


Figura 14. Ejemplo de clasificación.

Capítulo 5

Servicio Web para la detección automática de personalidad

En este capítulo se presenta la arquitectura del servicio Web para la detección automática de personalidad. El servicio Web está compuesto por tres módulos: pre-procesamiento, extracción de características y clasificación automática. La programación del servicio Web se llevó a cabo utilizando el lenguaje de programación *Java* y los recursos de la herramienta *Freeling* y *Weka.*

5.1 Arquitectura del servicio Web

En esta sección se muestra la arquitectura del servicio Web para la detección automática de personalidad. Como se observa en la figura 15 el usuario ingresa un texto en la aplicación cliente, que es una página Web escrita en lenguaje HTML. La aplicación Cliente, envía la llamada al servicio Web a través de una clase en *Java* llamada *Servlet,* la cual sirve como puente entre la aplicación cliente y el servicio Web se genera la conexión e intercambio de datos. Para que el *Servlet* pueda conectarse al servicio Web hace uso de un contrato WSDL (*Lenguaje de descripción de servicios Web o Web Services Description Language*) que es un archivo que contiene toda la información referente al servicio Web, métodos, nombre del servicio, etc.

La computadora en la que reside el servicio Web se identifica como servidor Web. Una vez que se realiza la conexión entre el servicio Web y la aplicación cliente, el servicio Web procesa el texto y devuelve al cliente la personalidad que corresponde al texto introducido.

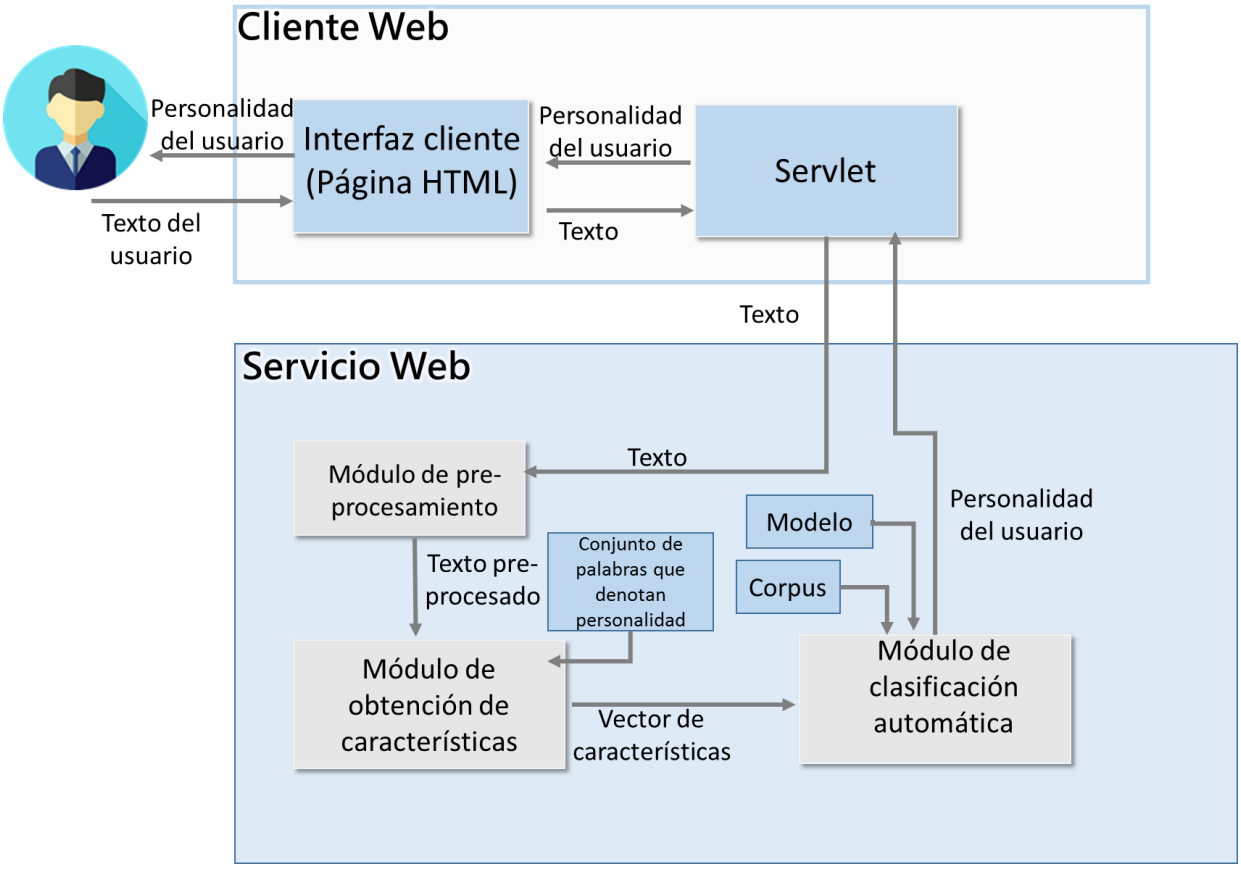


Figura 13. Arquitectura del servicio Web para la detección de personalidad

El servicio Web está compuesto por tres módulos:

I) *Módulo de pre-procesamiento.* Este módulo tiene como objetivo el pre-procesamiento del texto de entrada. Este módulo recibe como entrada un texto. En este módulo se lleva a cabo la *tokenización*, la eliminación de *stopwords*, la *lematización* y el etiquetado gramatical del texto. Este módulo da como resultado el texto pre-procesado, es decir un texto separado por palabras, sin stopwords y lematizado.

II) *Módulo de obtención de características.* Este módulo tiene como objetivo obtener características específicas del texto. Este módulo recibe como entrada el texto pre-procesado en el primer módulo y genera un vector de características que contiene el número de palabras en el texto, el peso TF-IDF del texto en cada factor DISC y un

identificador del factor de personalidad que obtiene el valor TF-IDF más alto, basándose en el conjunto de palabras que denotan personalidad. Este módulo da como salida un vector con las características del texto analizado.

III) *Módulo de clasificación automática.* Este módulo tiene como objetivo llevar a cabo la clasificación automática del texto. Este módulo recibe como entrada las características del vector que se genera en el módulo de extracción de características. Este módulo da como resultado la personalidad del texto introducido.

La clasificación se realizó a través de la herramienta *Weka* con el algoritmo SMO (*Sequential minimal optimization ó* algoritmo de optimización mínima secuencial). El algoritmo se entrena con el corpus de personalidad *DISC*, generado en este trabajo de tesis y hace uso del modelo de clasificación generado de igual manera con el corpus de personalidad *DISC*. Finalmente se obtiene la personalidad del comentario.

5.2 Interfaz gráfica del servicio Web para la detección de personalidad

En el desarrollo del servicio Web se generó una interfaz gráfica HTML, la que permite el uso del sistema a usuarios finales. En la figura 16 se muestra la interfaz gráfica del servicio Web, en donde el usuario puede escribir su nombre y un texto o cargarlo desde su computadora y al presionar el botón "Calcular" se comienza a analizar el texto mostrando una imagen de carga, cuando el análisis finaliza se muestra una página con el resultado del análisis del texto que contiene el nombre de la persona, y su resultado de personalidad, también se muestran las características que describen el tipo de personalidad obtenida como resultado.

Figura 14. Ingresar texto para obtener su personalidad

Para obtener la personalidad del texto que se desea analizar se introduce el nombre del usuario en el campo Nombre y el texto en el cuadro de texto y se presiona el botón “Calcular”. El servicio Web comenzará a hacer el análisis y se mostrara una imagen de carga. En la figura 17 se muestra un ejemplo.

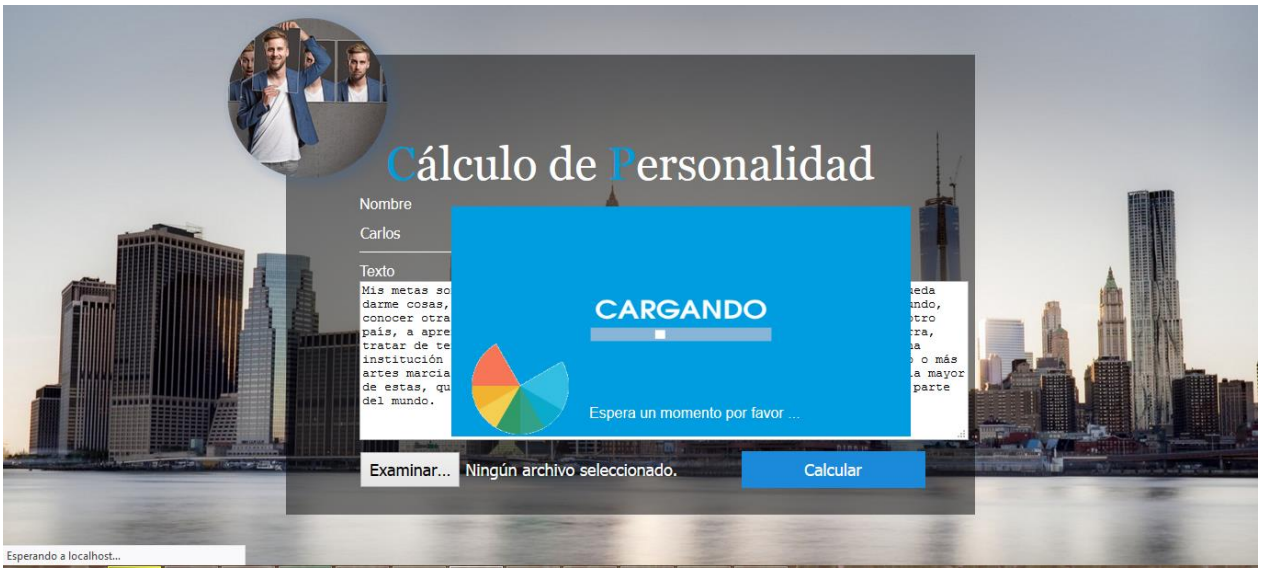


Figura 15. Calculando su personalidad

Finalmente, la página de interfaz del cliente establece una vía de comunicación con el *Servlet,* este envía los datos para establecer la comunicación con el servicio Web. Una vez establecida la comunicación, el servicio Web recibe el texto, lo analiza y devuelve la personalidad al *Servlet* y él envía el resultado mediante la generación de una página HTML. En la figura 18 se muestra un ejemplo de cómo se presenta el resultado al usuario.

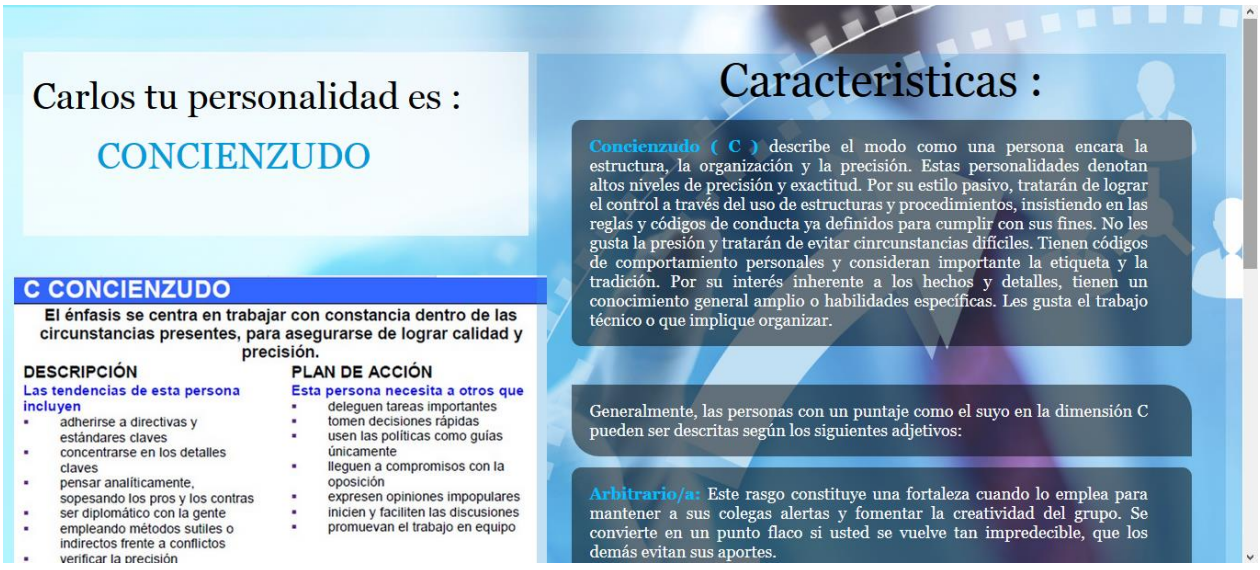


Figura 16. Resultado del análisis del texto

Capítulo 6

Pruebas y resultados

En este capítulo se describen las medidas de evaluación y el plan de pruebas establecido para llevar a cabo la evaluación del modelo de detección de personalidad.

6.1 Descripción de las pruebas

En esta sección se describen brevemente las pruebas realizadas para la evaluación del modelo de detección de personalidad presentado. Las medidas de evaluación aplicadas a las pruebas realizadas para la detección de personalidad fueron precisión, cobertura y medida-F. Las pruebas se realizaron en dos fases, utilizando la técnica *ten-fold cross validation y percentage split* con el algoritmo de clasificación automática *SMO* implementado con la herramienta *Weka* y entrenando al algoritmo con el corpus de personalidad *DISC*.

- En la primera fase se realizaron pruebas utilizando todos los adjetivos y verbos de los textos del corpus, tomándolos a ambos como características individuales del texto. El valor asignado a cada una de esas palabras en el corpus se calculó con la fórmula de pesos TF-

IDF de manera individual para cada palabra en cada uno de los 120 textos. El corpus que se obtuvo quedo formado por 540 características.

- En la segunda fase se utilizó el corpus con 120 textos obtenidos de la aplicación de encuestas de personalidad del modelo DISC. Esta versión del corpus contiene la información proporcionada por las personas que respondieron la prueba, y los pesos TF-IDF de cada uno de los textos calculados en base a el conjunto de palabras que denotan personalidad.

Las pruebas se llevaron a cabo con dos versiones del corpus con el objetivo de observar cual era el mejor resultado que se podía obtener en la clasificación de los textos. La prueba 1 se realizó con toda la información que se tenía hasta el momento sin ningún tipo de procesamiento con el propósito de tener un punto de partida e ir comparando que características pueden tomarse en cuenta para mejorar la precisión.

En la pruebas 2, 3, 4 y 5 se tomaron en cuenta diferentes combinaciones de características obtenidas del texto, para identificar cuales podían mejorar la precisión. En la prueba 6 de decidió realizar una igualación del número de registros por cada clase. Finalmente, en las pruebas 8 y 9 se obtuvo una mejora bastante significativa en el resultado de clasificación sobre las otras pruebas realizadas. A continuación se listan las pruebas que se ejecutaron:

Prueba 1: Sin procesamiento: No se realiza eliminación de *stopwords* ni ningún otro tipo de modificación en el texto, se utiliza toda la información reunida por las encuestas de personalidad.

Prueba 2: Convirtiendo palabras a características: Todos los adjetivos y verbos de los textos son utilizados en el corpus, cada uno como una característica independiente de cada registro, formando un corpus con 540 características. Para cada palabra se calcula su peso TF-IDF individual.

Prueba 3: Corpus de palabras en binario: Se sustituyen los pesos TF-IDF de cada palabra por un valor binario, 0 en caso de no aparecer en ese texto y 1 en caso de aparición.

Prueba 4: Agregando otras características del texto: Se agregan otras características al corpus de personalidad, estas características son: número de verbos, número de adjetivos, número de palabras, riqueza léxica y número de palabras diferentes.

Prueba 5: Binario agregando otras características del texto: Se agregan las características utilizadas en la prueba 5 al corpus en binario, estas características son: número de verbos, número de adjetivos, número de palabras, riqueza léxica y número de palabras diferentes.

Prueba 6: Igualar número de clases: Se iguala el número de registros para cada uno de los tipos de personalidad en el corpus con pesos TF-IDF y también en el corpus en binario, tomando como máximo de registros la clase con menos registros en este caso la clase Dominante con 14 registros.

Prueba 7: conjunto de palabras que denotan personalidad incluyendo verbos de *stopwords*: Se hace uso del conjunto de palabras que denotan personalidad *DISC* para calcular el peso TF-IDF de cada texto, no se eliminan los verbos que pertenecen a la lista de *stopwords* y el corpus se reduce al uso de ocho características. Cada texto es etiquetado con un peso para cada tipo de personalidad.

Prueba 8: conjunto de palabras que denotan personalidad eliminando verbos de *stopwords:* Se eliminan los verbos que pertenecen al grupo de *stopwords* del conjunto de palabras que denotan personalidad y se calculan nuevamente los pesos TF-IDF para cada texto en los que aparecen las palabras eliminadas. Cada texto es etiquetado con un peso para cada tipo de personalidad.

Prueba 9: Eliminando verbos repetidos: Se identifican y eliminan tres verbos que aparecieron de forma muy repetida en todos los textos registrados y causan un bajo porcentaje en la clasificación de personalidad. En la tabla 9 se muestra un resumen de las pruebas realizadas.

Prueba	Número de características	Eliminación de *Stopwords*	Uso de Léxico	Igualación del número de clases	Peso	Eliminación de verbos más repetitivos	Otras Características
Prueba 1	8	No	No	No	No	No	No
Prueba 2	540	Si, se incluyen los verbos de las *stopwords*	No	No	TF-IDF Asignado por palabra	No	No
Prueba 3	540	Si, se incluyen los verbos de las *stopwords*	No	No	Valor Binario	No	No
Prueba 4	545	Si, se incluyen los verbos de las *stopwords*	No	No	TF-IDF Asignado por palabra	No	Número de verbos, número de adjetivos, número de palabras, riqueza léxica y número de palabras diferentes.
Prueba 5	545	Si, se incluyen los verbos de las *stopwords*	No	No	Valor Binario	No	Número de verbos, número de adjetivos, número de palabras, riqueza léxica y número de palabras diferentes.
Prueba 6	545	Si, se incluyen los verbos de las *stopwords*	No	Si	Valor Binario y TF-IDF Asignado por palabra	No	Número de verbos, número de adjetivos, número de palabras, riqueza léxica y número de palabras diferentes.
Prueba 7	6	Si, se incluyen los verbos de las *stopwords*	Si	No	TF-IDF por factor DISC	No	Valor TF-IDF más alto.
Prueba 8	6	Eliminación de verbos pertenecientes a las *stopwords*	Si	No	TF-IDF por factor DISC	No	Valor TF-IDF más alto.
Prueba 9	6	Eliminación de verbos pertenecientes a las *stopwords*	Si	No	TF-IDF por factor DISC	Eliminación de verbos ir, terminar y jugar	Valor TF-IDF más alto.

Tabla 9. Resumen de pruebas.

6.2 Medidas de evaluación

La evaluación de los resultados de las pruebas se realizó con las siguientes medidas: precisión (*precision*), cobertura (*recall*) y medida-F (*F-Score*). Las variables utilizadas por estas medidas son TP (*true positive*), TN (*true negative*), FN (*false negative*) y FP (*false positive*).

Si un texto es clasificado automáticamente en la misma categoría que la clasificación manual, se trata de un verdadero positivo (TP, *true positive*) o un verdadero negativo (TN, *true negative*), pero si el texto se llega a clasificar en otra clase diferente se trata de un falso negativo (FN, *false negative*) o un falso positivo (FP, *false positive*) (Cuadrado, 2011).

En la tabla 10 se presentan las posibles combinaciones de clasificación en dos clases de un sistema.

	Automático	
Manual	Positivo	Positivo
Positivo	TP	FN
Negativo	FP	TN

Tabla 10. Posibles combinaciones de resultados en la clasificación en dos clases.

A continuación se describen las medidas de evaluación.

3.3.1 Precisión

La ***precisión*** mide el número de aciertos del clasificador para una clase entre el total de clasificados en esa clase. Es decir, la precisión mide la exactitud del clasificador para esa clase, una mayor precisión indica un menor número de *falsos positivos*, y por el contrario, una baja precisión indica muchos *falsos positivos*. La precisión se calcula según la siguiente ecuación.

$$precisiónp = \frac{\mathrm{TP}}{TP + FP}$$

3.3.2 Cobertura

La cobertura mide el número de aciertos para una clase en relación con el número de documentos que deberían haber sido clasificados en esa clase. Es decir, la cobertura se puede ver como la sensibilidad del clasificador para esa clase, ya que a mayor cobertura menor número de falsos negativos, mientras que una cobertura baja indica mayor número de falsos negativos (Cuadrado, 2011). La cobertura se calcula según la siguiente ecuación:

$$coberturap = \frac{\mathrm{TP}}{TP + FN}$$

3.3.3 Medida F

La ***medida-F*** *(F-Score)* es una combinación de las medidas anteriores, precisión y cobertura, que representa la media armónica de la precisión y la cobertura (Cuadrado, 2011). La medida-F se calcula según la siguiente ecuación para 2 clases:

$$medida - F = \frac{2\ x\ precisión\ x\ cobertura}{precisión + cobertura}$$

6.3 Selección del algoritmo de clasificación

En esta sección se llevó a cabo la selección del algoritmo de clasificación que se encarga de realizar la clasificación automática de textos para determinar la personalidad del usuario. El algoritmo seleccionado para realizar la tarea de clasificación fue el algoritmo *SMO* (*algoritmo de optimización mínima secuencial*). Para poder llegar a la conclusión de seleccionar el algoritmo de clasificación *SMO* de entre otros, se realizó una revisión de los algoritmos utilizados en el estado del arte.

En los trabajos de (Adali & Golbeck, 2012) se utiliza el algoritmo de regresión lineal *ZeroR,* en el trabajo de (Wald et al., 2012) se utilizan los algoritmos *DTable* y *REPTree*. En el trabajo de (Sáez et al., 2014) se utiliza *SVM* para llevar a cabo la clasificación, en el trabajo de (Lima & de Castro, 2014) se utiliza *Naive Bayes* , *SVM y Red neuronal de Perceptron Multicapa*, en el trabajo de (Pratama & Sarno, 2015) se utiliza *Naive Bayes, K-vecinos más cercanos y máquinas de vectores soporte (SVM),* y en el trabajo de (Peng et al., 2015) se utiliza también máquina de vectores soporte *(SVM*).

Se realizaron evaluaciones con siete algoritmos identificados en el estado del arte, los cuales se mencionan a continuación: *ZeroR, DTable, RepTree, SMO (SVM), Naive Bayes, Multilayer Perceptron y KNN.*

Para la evaluación se utilizó el corpus con los resultados de las evaluaciones de personalidad *DISC* que fueron aplicados a 120 personas.

6.4 Evaluación de algoritmos

La evaluación de algoritmos se llevó a cabo con la información sin procesar esto quiere decir que los datos se utilizan tal como fueron obtenidos de las encuestas de personalidad *DISC* aplicadas anteriormente. Los registros utilizados para esta prueba fueron: género, escolaridad, estado civil, ocupación, puesto, red social, número de amigos en redes sociales, texto del participante y resultado de la prueba. En la Tabla 11 se muestra un ejemplo de los registros utilizados para esta prueba.

Datos: Resultados de las encuestas sin ningún procesamiento (género, escolaridad, estado civil, ocupación, puesto, red social, número de amigos en redes sociales, texto del participante y resultado de la prueba).

Genero	Escolaridad	Estado_civil	Ocupacion	Puesto	Red_social	Num_Amigos	Texto	Resultado_prueba
Masculino	Posgrado	Soltero	Estudiante	Estudiante	Twitter	10	Ayer fui a trabajar en el Centro Nacional de Investigacion y Desarrollo Tecnologico en mi tema de tesis en donde por la mañana revise un articulo que redacte para un congreso en Orlando Florida. Tambien atendi a un residente que se encuentra colaborando en mi proyecto, posteriormente asisti a una reunion de trabajo con mi director de tesis y fuimos a comer con los demas compañeros tesistas. Finalmente por la tarde me retire a mi casa donde comi con mis padres y segui desarrollando algunas actividades pendientes del trabajo. Me gusta leer pero por el momento no tengo tiempo. Mi principal meta es terminar en tiempo y forma la maestria en ciencias computacionales.	ESTABLE
Femenino	Universidad	Soltero	Estudiante	Estudiante	Facebook	600	Mi pasatiempo favorito es pasarlo con mi familia en casa, platicar en la mesa a la hora de la comida, y hablar de todo lo sucedido durante el día de cada uno de nosotros, pedir opinion acerca de los acontecimientos a mis padres y como poder resolver algun problema que se halla presentado en la semana en el dia tanto puede ser en casa, como en la escuela, adoro hablar con ellos y por este motivo se convierte en mi	ESTABLE
Masculino	Preparatoria	Casado	Estudiante	Estudiante	Facebook	100	En el día 14/11/16 Me levante temprano y no asistí al servicio social para poder arreglar los documentos necesarios para tramitar una beca, tuve que hacer fila en la escuela y me mantuve parado alrededor de 30 min. La fila era muy larga. Al fin tramité un documento en la escuela, despues camine hacia el ayuntamiento a tramitar otro documento, fue menos labor puesto que la fila era muy corta. Cuando termine regrese a casa a terminar tarea pendiente y a esperar la hora para entrar a la escuela.	ESTABLE
Femenino	Preparatoria	Casado	Estudiante	Estudiante	Facebook	250	Lo que deseo en la vida es, realizar todas mis meta, lo primordial es terminar mi carrera, para poder trabajar y ayudar a mis padres. Formar una familia hermosa y poder disfrutar de ella día con día. Ser una persona de bien para la sociedad y el país, deseo que en el futuro la violencia ya termine, y que solo haya buenas personas en el mundo, se que suena imposible, pero eso es lo que yo deseo.	ESTABLE
Femenino	Universidad	Soltero	Estudiante	Estudiante	Facebook	200	Dentro de mis principales metas a corto plazo, es concluir mis estudios, titularme, poder encontrar un trabajo próximo con el que pueda sustentarte para tener una vida digna; ya que en la actualidad nos enfrentamos a muchos problemas desempleo, crisis económica y el daño a nuestro medio ambiente. Desarrollarme profesionalmente es un logro personal y satisfactorio para mí y para las personas que me an ayudado y apoyado durante mis años de estudio a lo largo de mi vida.	ESTABLE
							Mi pasatiempo favorito es tocar mi saxofón me gusta la música, también me gusta	

Tabla 11. Ejemplo de los registros utilizados en la evaluación de algoritmos.

Resultados: En la tabla 12 se muestran las características del corpus utilizado y los resultados obtenidos de la evaluación realizada a los algoritmos de aprendizaje automático.

Corpus	Características del corpus	Clasificador	Precisión	Tiempo de ejecución.	Método de evaluación
Corpus con los resultados de las evaluaciones de personalidad DISC	4 clases : 14 Dominantes 62 Estables 26 Influyentes 18 Concienzudos Total 120 textos	ZeroR	26%	0 seg	*Ten-fold cross validation*
		DTable	32%	0.03 seg	
		RepTree	26%	0 seg	
		SMO(SVM)	47%	0.15 seg	
		Naive Bayes	42%	0 seg	
		KNN	41%	0 seg	
		Multilayer Perceptron	47%	44.21 seg	
		ZeroR	44%	0 seg	*Percentage Split* 70% para entrenamiento y 30% para evaluación
		DTable	43.1%	0.06 seg	
		RepTree	44.4%	0.01 seg	
		SMO(SVM)	46.9%	0.39 seg	
		Naive Bayes	41.7%	0 seg	
		KNN	43%	0.01 seg	
		Multilayer Perceptron	46.9%	36.93 seg	

Tabla 12. Resultados de la evaluación de algoritmos.

Los resultados de evaluación de los algoritmos muestran que el algoritmo *SMO* y *Multilayer Perceptron* obtuvieron los porcentajes de exactitud más altos. Finalmente se seleccionó el algoritmo *SMO y Multilayer Perceptron* para utilizarlo en la primera prueba de clasificación.

6.5 Prueba 1: Sin procesamiento algoritmo *SMO* y *Multilayer Perceptron*

En esta prueba se realizó una comparación de los resultados obtenidos al utilizar el algoritmo *SMO* y el algoritmo *Multilayer Perceptron* sobre el conjunto de datos del corpus de personalidad, esta prueba se llevó a cabo con la información sin procesar esto quiere decir que los datos se utilizan tal como fueron obtenidos de las encuestas de personalidad *DISC* aplicadas anteriormente.

Datos: Resultados de las encuestas sin ningún procesamiento (género, escolaridad, estado civil, ocupación, puesto, red social, número de amigos en redes sociales, texto del participante y resultado de la prueba).

Genero	Escolaridad	Estado_civil	Ocupacion	Puesto	Red_social	Num_Amigos	Texto	Resultado_prueba
Masculino	Posgrado	Soltero	Estudiante	Estudiante	Twitter	10	Ayer fui a trabajar en el Centro Nacional de Investigacion y Desarrollo Tecnologico en mi tema de tesis en donde por la mañana revise un articulo que redacte para un congreso en Orlando Florida. Tambien atendi a un residente que se encuentra colaborando en mi proyecto, posteriormente asisti a una reunion de trabajo con mi director de tesis y fuimos a comer con los demas compañeros tesistas. Finalmente por la tarde me retire a mi casa donde comi con mis padres y segui desarrollando algunas actividades pendientes del trabajo. Me gusta leer pero por el momento no tengo tiempo. Mi principal meta es terminar en tiempo y forma la maestria en ciencias computacionales.	ESTABLE
Femenino	Universidad	Soltero	Estudiante	Estudiante	Facebook	600	Mi pasatiempo favorito es pasarlo con mi familia en casa, platicar en la mesa a la hora de la comida, y hablar de todo lo sucedido durante el día de cada uno de nosotros, pedir opinion acerca de los acontecimientos a mis padres y como poder resolver algun problema que se halla presentado en la semana en el dia tanto puede ser en casa, como en la escuela, adoro hablar con ellos y por este motivo se convierte en mi	ESTABLE
Masculino	Preparatoria	Casado	Estudiante	Estudiante	Facebook	100	En el día 14/11/16 Me levante temprano y no asistí al servicio social para poder arreglar los documentos necesarios para tramitar una beca, tuve que hacer fila en la escuela y me mantuve parado alrededor de 30 min. La fila era muy larga. Al fin tramité un documento en la escuela, despues camine hacia el ayuntamiento a tramitar otro documento, fue menos labor puesto que la fila era muy corta. Cuando termine regrese a casa a terminar tarea pendiente y a esperar la hora para entrar a la escuela.	ESTABLE
Femenino	Preparatoria	Casado	Estudiante	Estudiante	Facebook	250	Lo que deseo en la vida es, realizar todas mis meta, lo primordial es terminar mi carrera, para poder trabajar y ayudar a mis padres. Formar una familia hermosa y poder disfrutar de ella día con día. Ser una persona de bien para la sociedad y el país, deseo que en el futuro la violencia ya termine, y que solo haya buenas personas en el mundo, se que suena imposible, pero eso es lo que yo deseo.	ESTABLE
Femenino	Universidad	Soltero	Estudiante	Estudiante	Facebook	200	Dentro de mis principales metas a corto plazo, es concluir mis estudios, titularme, poder encontrar un trabajo próximo con el que pueda sustentarte para tener una vida digna; ya que en la actualidad nos enfrentamos a muchos problemas desempleo, crisis económica y el daño a nuestro medio ambiente. Desarrollarme profesionalmente es un logro personal y satisfactorio para mí y para las personas que me an ayudado y apoyado durante mis años de estudio a lo largo de mi vida.	ESTABLE
							Mi pasatiempo favorito es tocar mi saxofón me gusta la música, también me gusta	

Tabla 13. Ejemplo de los registros utilizados en la prueba 1.

Algoritmo	Precisión	Cobertura	Medida F	Método de evaluación	Instancias clasificadas correctamente
SMO	47%	52.5%	45.8%	*Ten-fold cross validation*	52.5%
Multilayer Perceptron	47%	51.7%	45.5%		51.6%
SMO	46.9%	55.6%	50.4%	*Percentage Split* 70% para entrenamiento y 30% para evaluación	55.5%
Multilayer Perceptron	46.9%	55.6%	50.4%		55.5%

Tabla 14. Resultados de la prueba 1

Resultados: Los resultados de esta prueba se muestran en la tabla 14. En ambos casos se obtienen resultados muy parecidos. En la prueba con el método de evaluación de validación cruzada el algoritmo *SMO* supera ligeramente al algoritmo *Multilayer Perceptron*, ambos obtienen el 47% de precisión, pero en cobertura, medida F e instancias clasificadas correctamente se obtienen mejores resultados con el algoritmo *SMO*. Es por eso que se decide utilizar el algoritmo *SMO* como clasificador en el resto de las pruebas.

6.6 Prueba 2: Convirtiendo palabras a características

La prueba 2 se llevó a cabo utilizando todos los adjetivos y verbos de los textos en el corpus, cada uno se tomó como una característica independiente de cada registro, formando un corpus con 540 características. Para cada palabra se calculó su peso TF-IDF individual. Las palabras utilizadas en el corpus fueron todos los adjetivos y verbos que aparecen en los textos obtenidos de las personas que respondieron la encuesta de personalidad *DISC*. En la tabla 15 se muestra un ejemplo de los registros utilizados para esta prueba. En el método de evaluación *percentages Split* se utilizó un 70% del corpus para entrenamiento del algoritmo y 30% para prueba.

Datos: 540 adjetivos y verbos con peso TF-IDF.

Único	Unir	Universitario	usar	útil	Utilizar	Valeroso	variar	Vender	venir	ver	viajar	visitar	Vivir	volver	PERSONALIDAD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0.02112587	0.0743482	0.06103106	0	0	CONCIENZUDO
0	0	0	0	0	0	0	0	0	0	0.0470805	0	0	0	0	INFLUYENTE
0.03393317	0	0	0	0	0	0	0	0	0	0	0	0	0	0	INFLUYENTE
0	0	0	0	0	0	0	0	0	0	0	0	0	0.02689796	0	CONCIENZUDO
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	CONCIENZUDO
0	0	0	0	0	0	0	0	0	0	0.0633776	0	0	0.0272428	0	ESTABLE
0.03753214	0	0	0	0	0	0	0	0	0	0	0.02928869	0	0.03219604	0	CONCIENZUDO
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0.06420068	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	INFLUYENTE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02797914	INFLUYENTE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0.02009534	0	0	0	0.03173244	DOMINANTE
0	0	0	0	0	0	0	0	0	0	0	0.02612234	0	0	0	INFLUYENTE
0	0	0	0	0	0	0	0	0	0	0.11133902	0	0	0	0	INFLUYENTE

Tabla 15. Ejemplo de los registros utilizados en la prueba 2

Resultados: En las Tabla 16 se muestran los resultados obtenidos en esta prueba. Se puede observar que el porcentaje en las medidas de evaluación no supera el 55.6 por ciento.

Algoritmo	**Precisión**	**Cobertura**	**Medida F**	**Método de evaluación**	**Instancias clasificadas correctamente**
SMO	42%	45%	41.1%	*Ten-fold cross validation*	45%
SMO	49.8%	55.6%	49.8%	*Percentage Split*	55.5%

Tabla 16. Resultados de la prueba 2

6.7 Prueba 3 Corpus de palabras en binario

En esta prueba se utiliza el corpus que contiene cada palabra como una característica independiente con sus respectivos pesos TF-IDF de cada palabra utilizado en la prueba 2, pero en este caso los pesos se cambian por un valor binario, se asignó 0 en caso de que la palabra no apareciera en el texto registrado en el corpus y 1 en caso de que si apareciera. Las palabras utilizadas en el corpus fueron todos los adjetivos y verbos que aparecen en los textos obtenidos por parte de las personas que respondieron la prueba de personalidad *DISC*.

El método de evaluación *percentage split* se utilizó con 70% del corpus para entrenamiento del algoritmo y 30% para prueba. En la tabla 17 se muestra un ejemplo de los registros utilizados para esta prueba.

Datos: 540 adjetivos y verbos con valores binarios

Único	Unir	Universitaric	usar	útil	Utilizar	Valeroso	variar	Vender	venir	ver	viajar	visitar	Vivir	volver	PERSONALIDAD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	ESTABLE
0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	CONCIENZUDO
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	INFLUYENTE
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	INFLUYENTE
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	CONCIENZUDO
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	CONCIENZUDO
0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	ESTABLE
1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	CONCIENZUDO

Tabla 17. Ejemplo de los registros utilizados en la prueba 3.

Resultados:
Los resultados obtenidos se muestran en la tabla 18.

Algoritmo	Precisión	Cobertura	Medida F	Método de evaluación	Instancias clasificadas correctamente
SMO	39.6%	45.8%	41.5%	*Ten-fold cross validation*	45.8%
SMO	26.6%	33.3%	29.6%	*Percentage Split*	33.3%

Tabla 18. Resultados de la prueba 3

6.8 Prueba 4: Agregando otras características del texto

En esta prueba se decide agregar otras características obtenidas al analizar los textos del corpus de personalidad *DISC*. Esto con el objetivo de mejorar los resultados de clasificación de la personalidad, se agregan como nuevas características; número de verbos, número de adjetivos, número de palabras, riqueza léxica y número de palabras diferentes. En la tabla 19 se muestra un ejemplo de los registros utilizados para esta prueba. El método de evaluación *percentage split* se utilizó con 70% del corpus para entrenamiento del algoritmo y 30% para prueba.

Datos: 540 adjetivos y verbos como características independientes con peso TF-IDF, número de verbos, número de adjetivos, número de palabras, riqueza léxica y número de palabras diferentes).

NumVerbos	NumAdjetivos	NumPalabras	RiquezaLexica	PalabrasDiferentes	apoyar	Arreglar	asistir	atender	Ayudar	Bueno	buscar	Vivir	volver	PERSONALIDAD
17	1	62	0.919354839	57	0	0	0.0383905	0.04480889	0	0	0	0	0	ESTABLE
12	1	38	0.868421053	33	0	0	0	0	0	0	0	0	0	ESTABLE
16	5	51	0.862745098	44	0	0.04511824	0.04667081	0	0	0	0	0	0	ESTABLE
12	3	38	0.842105263	32	0	0	0	0	0.04934372	0	0	0	0	ESTABLE
12	9	43	0.953488372	41	0.05535375	0	0	0	0.04360608	0	0	0	0	ESTABLE
10	4	37	0.945945946	35	0	0	0	0	0	0	0	0	0	ESTABLE
15	9	78	0.756410256	59	0	0	0	0	0	0.02092338	0	0	0	CONCIENZUDO
14	2	35	0.914285714	32	0	0	0	0	0	0	0	0	0	INFLUYENTE
21	6	73	0.726027397	53	0	0	0	0	0	0.06706945	0	0	0	INFLUYENTE
23	9	79	0.924050633	73	0	0	0	0.03516647	0	0.02065852	0.06025851	0.02689796	0	CONCIENZUDO
21	11	69	0.898550725	62	0	0	0	0	0	0	0	0	0	CONCIENZUDO
27	9	78	0.807692308	63	0.03051553	0	0	0	0	0.02092338	0	0.0272428	0	ESTABLE
15	7	66	0.818181818	54	0	0	0	0	0	0	0	0.03219604	0	CONCIENZUDO
16	2	72	0.680555556	49	0	0	0	0	0	0	0	0	0	ESTABLE
30	3	77	0.909090909	70	0	0	0	0	0	0	0	0	0	ESTABLE
25	9	84	0.94047619	79	0	0	0	0	0	0.01942885	0	0	0	INFLUYENTE
15	7	53	0.924528302	49	0	0	0	0	0	0	0	0	0	ESTABLE
31	7	93	0.806451613	75	0	0.02474226	0	0	0	0	0	0	0.02797914	INFLUYENTE
23	7	75	0.8	60	0	0	0	0	0	0	0	0	0	ESTABLE
26	5	82	0.853658537	70	0	0	0.02902697	0	0	0	0	0	0.03173244	DOMINANTE
19	8	74	0.810810811	60	0	0	0	0	0	0	0.06433003	0	0	INFLUYENTE
21	8	74	0.783783784	58	0	0	0	0	0	0	0	0	0	INFLUYENTE

Tabla 19. Ejemplo de los registros utilizados en la prueba 4

Resultados: en la tabla 20 se muestran los resultados obtenidos en esta prueba. Se puede observar que la precisión aumenta a 48.1% y 49.8% resultados que son muy bajos aun para poder realizar la clasificación.

Algoritmo	Precisión	Cobertura	Medida F	Método de evaluación	Instancias clasificadas correctamente
SMO	48.1%	45%	40.5%	*Ten-fold cross validation*	45%
SMO	49.8%	55.6%	49.8%	*Percentage Split*	55.5%

Tabla 20. Resultados de la prueba 4

6.9 Prueba 5: Binario agregando otras características del texto

En esta prueba se decide agregar las características de la prueba anterior pero esta vez se agregan al corpus con valores binarios en lugar de valores TF-IDF. Con el objetivo de mejorar los resultados de clasificación de la personalidad, se agregan las características utilizadas en la prueba 4; número de verbos, número de adjetivos, número de palabras, riqueza léxica y número de palabras diferentes. En la tabla 21 se muestra un ejemplo de los registros utilizados para esta prueba. El método de evaluación *percentage split* se utilizó con 70% del corpus para entrenamiento del algoritmo y 30% para prueba.

Datos: 540 adjetivos y verbos como características independientes con valores binarios, numero de verbos, numero de adjetivos, numero de palabras, riqueza léxica y numero de palabras diferentes).

NumVerbos	NumAdjetivos	NumPalabras	RiquezaLexica	PalabrasDiferentes	apoyar	Arreglar	asistir	atender	Ayudar	Bueno	buscar	Vivir	volver	PERSONALIDAD
17	1	62	0.919354839	57	0	0	1	1	0	0	0	0	0	ESTABLE
12	1	38	0.868421053	33	0	0	0	0	0	0	0	0	0	ESTABLE
16	5	51	0.862745098	44	0	1	1	0	0	0	0	0	0	ESTABLE
12	3	38	0.842105263	32	0	0	0	0	1	0	0	0	0	ESTABLE
12	9	43	0.953488372	41	1	0	0	0	1	0	0	0	0	ESTABLE
10	4	37	0.945945946	35	0	0	0	0	0	0	0	0	0	ESTABLE
15	9	78	0.756410256	59	0	0	0	0	0	1	0	0	0	CONCIENZUDO
14	2	35	0.914285714	32	0	0	0	0	0	0	0	0	0	INFLUYENTE
21	6	73	0.726027397	53	0	0	0	0	0	1	0	0	0	INFLUYENTE
23	9	79	0.924050633	73	0	0	0	1	0	1	1	1	0	CONCIENZUDO
21	11	69	0.898550725	62	0	0	0	0	0	0	0	0	0	CONCIENZUDO
27	9	78	0.807692308	63	1	0	0	0	0	1	0	1	0	ESTABLE
15	7	66	0.818181818	54	0	0	0	0	0	0	0	1	0	CONCIENZUDO

Tabla 21. Ejemplo de los registros utilizados en la prueba 5

Resultados: En la tabla 22 se muestran los resultados obtenidos en esta prueba.

Algoritmo	Precisión	Cobertura	Medida F	Método de evaluación	Instancias clasificadas correctamente
SMO	39.1%	46.7%	41.6%	*Ten-fold cross validation*	46.6%
SMO	29.7%	38.9%	33.5%	*Percentage Split*	38.8%

Tabla 22. Resultados de la prueba 5

6.10 Prueba 6: Igualar número de clases

En esta prueba, debido al bajo porcentaje de clasificación obtenido hasta el momento se decide igualar el número de clases ya que el corpus contiene registros de 120 textos de personalidades diferentes en diferente proporción. Se iguala el número de registros para cada uno de los tipos de personalidad en el corpus con pesos TF-IDF y también en el corpus en binario, tomando como máximo de registros la clase con menor cantidad, en este caso la clase dominante con 14 registros. El método de evaluación *percentage split* se utilizó con 70% del corpus para entrenamiento del algoritmo y 30% para prueba.

Datos: 14 registros de cada clase utilizando como características independientes adjetivos y verbos con peso TD-IDF y aparte con valores binarios, *numero de verbos, numero de adjetivos, numero de palabras, riqueza léxica y numero de palabras diferentes*). En la tabla 23 y 24 se muestra un ejemplo de los registros utilizados para esta prueba.

Verbos	Adjetivos	Num_De_Palabras	RiquezaLexica	PalabrasDiferentes	estar	gustar	haber	hacer	ir	ser	tener	terminar	Vivir	PERSONALIDAD	
15	9	78	0.756410256	59	0.093329	0	0	0	0.0524299	0.0410132	0.0339309	0	0	CONCIENZUDO	
23	9	79	0.924050633	73	0	0	0.0395106	0.0162005	0.0345108	0.0674901	0.0335014	0	0.026898	CONCIENZUDO	
21	11	69	0.898550725	62	0	0.0437965	0.0678551	0.0185484	0	0.0772713	0	0	0	CONCIENZUDO	
15	7	66	0.818181818	54	0	0	0	0	0.0206542	0.0323135	0.1203006	0	0.032196	CONCIENZUDO	
14	4	56	0.875	49	0.0259988	0	0	0.0228543	0	0	0.0945219	0.0231045	0	CONCIENZUDO	
25	1	85	0.8	68	0.0513858	0	0	0.0451708	0.0320748	0	0	0.0456653	0	CONCIENZUDO	
23	10	92	0.815217391	75	0.0316507	0	0.0508913	0	0	0.0115907	0	0	0	CONCIENZUDO	
10	3	31	0.870967742	27	0.1408966	0	0.0503441	0	0	0.1031946	0	0	0	CONCIENZUDO	
4	0	9	0.888888889	8	0	0	0	0	0	0.1184827	0	0	0	CONCIENZUDO	
8	3	30	0.933333333	28	0	0.050366	0	0	0	0.0710896	0.0882204	0	0	CONCIENZUDO	
15	5	60	0.85	51	0	0.025183	0	0	0.0454393	0.0710896	0.0220551	0.0215642	0	CONCIENZUDO	
20	3	65	0.907692308	59	0.0895958	0	0	0	0.020972	0.0492159	0.0610757	0	0	CONCIENZUDO	
18	4	62	0.887096774	55	0.0234828	0.0974826	0	0.0206426	0.0219867	0.0515973	0	0.0208686	0	CONCIENZUDO	
11	2	34	0.970588235	33	0	0	0	0	0	0.0627261	0	0	0	CONCIENZUDO	14
26	5	82	0.853658537	70	0.0177553	0.0184266	0.0190325	0.0624313	0	0.0910294	0	0	0	DOMINANTE	
17	5	67	0.865671642	58	0.0869213	0.0225519	0	0.0191021	0.0813838	0.0477467	0.0197508	0.0386224	0	DOMINANTE	
10	5	49	0.87755102	43	0	0	0	0.0261192	0	0.0435242	0	0	0	DOMINANTE	
9	1	32	0.78125	25	0	0	0.0975417	0.039995	0	0.0666465	0	0	0	DOMINANTE	
5	1	21	0.857142857	18	0.2773204	0	0	0	0	0	0	0	0	DOMINANTE	
22	3	100	0.74	74	0	0	0	0.0383952	0.1090542	0.0106634	0	0.0129385	0	DOMINANTE	
11	1	25	1	25	0.0582373	0	0	0.0511936	0.0545271	0	0	0.1035081	0	DOMINANTE	
7	5	29	0.896551724	26	0	0.0521027	0	0.0441324	0.0470061	0.1103115	0.0456313	0.0446156	0	DOMINANTE	
5	4	27	0.851851852	23	0	0	0	0.0474015	0	0.1579769	0.0490113	0	0	DOMINANTE	
27	4	80	0.8625	69	0.0363983	0	0.0195083	0.015998	0.0170397	0.0399879	0	0.0161731	0	DOMINANTE	
15	7	55	0.818181818	45	0	0	0.0283758	0.0232698	0	0.1163284	0	0.0235246	0.1159057	DOMINANTE	

Tabla 23. Ejemplo de los registros utilizados en la prueba 6, palabras con peso TF-IDF.

Verbos	Adjetivos	Num_De_Palabras	RiquezaLexica	PalabrasDiferentes	estar	gustar	haber	hacer	ir	ser	tener	terminar	Vivir	PERSONALIDAD
15	9	78	0.7564103	59	1	0	0	0	1	1	1	0	0	CONCIENZUDO
23	9	79	0.9240506	73	0	0	1	1	1	1	1	0	1	CONCIENZUDO
21	11	69	0.8985507	62	0	1	1	1	0	1	0	0	0	CONCIENZUDO
15	7	66	0.8181818	54	0	0	0	0	1	1	1	0	1	CONCIENZUDO
14	4	56	0.875	49	1	0	0	1	0	0	1	1	0	CONCIENZUDO
25	1	85	0.8	68	1	0	0	1	1	0	0	1	0	CONCIENZUDO
23	10	92	0.8152174	75	1	0	1	0	0	1	0	0	0	CONCIENZUDO
10	3	31	0.8709677	27	1	0	1	0	0	1	0	0	0	CONCIENZUDO
4	0	9	0.8888889	8	0	0	0	0	0	1	0	0	0	CONCIENZUDO
8	3	30	0.9333333	28	0	1	0	0	0	1	1	0	0	CONCIENZUDO
15	5	60	0.85	51	0	1	0	0	1	1	1	1	0	CONCIENZUDO
20	3	65	0.9076923	59	1	0	0	0	1	1	1	0	0	CONCIENZUDO
18	4	62	0.8870968	55	1	1	0	1	1	1	0	1	0	CONCIENZUDO
11	2	34	0.9705882	33	0	0	0	0	0	1	0	0	0	CONCIENZUDO
26	5	82	0.8536585	70	1	1	1	1	0	1	0	0	0	DOMINANTE
17	5	67	0.8656716	58	1	1	0	1	1	1	1	1	0	DOMINANTE

14

Tabla 24. Ejemplo de los registros utilizados en la prueba 6, palabras con valores binarios.

Resultados: En la tabla 25 se muestran los resultados de las pruebas utilizando ambos corpus.

Corpus	Algoritmo	Precisión	Cobertura	Medida F	Método de evaluación	Instancias clasificadas correctamente
Corpus con pesos TF-IDF por palabra.	SMO	29.1%	26.8%	27.3%	*Ten-fold cross validation*	26.7%
		17%	23.5%	19%	*Percentage Split*	23.5%
Corpus con valores binarios	SMO	27.8%	28.6%	28%	*Ten-fold cross validation*	28.5%
		25.5%	23.5%	23.5%	*Percentage Split*	23.5%

Tabla 25. Resultados de la prueba 6

Como se puede observar en la tabla 25 los resultados del clasificador al evaluar los dos corpus muestran muy bajos porcentajes en todas las medidas de evaluación y también en el porcentaje de instancias clasificadas correctamente. Debido a estos bajos resultados en esta prueba y las anteriores se decide crear un conjunto de palabras que denoten personalidad. El objetivo de crear un conjunto de palabras que denoten personalidad con los textos del corpus de personalidad es para poder calcular un peso de cada verbo y adjetivo que aparecen en cada uno de los textos. Si las palabras tienen un peso asignado se puede saber de esta manera cuánto pesa cada palabra

en cada factor de personalidad y así poder determinar qué peso tiene cada texto registrado en el corpus en cada personalidad.

6.11 Prueba 7: Conjunto de palabras que denotan personalidad incluyendo verbos de *stopwords*

En esta prueba se hace uso del conjunto de palabras que denotan personalidad para calcular el peso TF-IDF de cada texto en cada factor de personalidad, en esta prueba se toman en cuenta todos los verbos y adjetivos que aparecen en cada texto. No se eliminan los verbos que pertenecen a la lista de *stopwords* y el corpus se reduce a siete características de cada texto *(Número de palabras, Dominante_TF-IDF, Influyente_TF-IDF, Stable_TF-IDF, Concienzudo_TF-IDF, TF-IDF Mayor y Resultado de prueba*). Cada texto es etiquetado con un peso para cada tipo de personalidad. El método de evaluación *percentage split* se utilizó con 70% del corpus para entrenamiento del algoritmo y 30% para prueba.

Datos: *número de palabras, dominante TF-IDF, influyente TF-IDF, estable TF-IDF, concienzudo TF-IDF, TF-IDF mayor y resultado de prueba*. En la tabla 26 se muestra un ejemplo de los registros utilizados para esta prueba.

Num_palabras	Dominante_TF-IDF	Influyente_TF-IDF	Stable_TF-IDF	Concienzudo_TF-IDF	TF-IDF Mayor	Resultado_prueba
78	0.3978	0.4136	0.3468	0.4286	4	CONCIENZUDO
79	0.4203	0.3808	0.4039	0.4146	1	CONCIENZUDO
69	0.3957	0.3293	0.3791	0.398	4	CONCIENZUDO
66	0.2107	0.3028	0.269	0.3326	4	CONCIENZUDO
56	0.1395	0.1748	0.1751	0.2423	4	CONCIENZUDO
85	0.3383	0.2723	0.2841	0.3142	1	CONCIENZUDO
92	0.154	0.1519	0.1521	0.2268	4	CONCIENZUDO
31	0.2556	0.2123	0.2202	0.2557	4	CONCIENZUDO
9	0.0682	0.0659	0.0724	0.0695	3	CONCIENZUDO
30	0.1632	0.1791	0.1835	0.2186	4	CONCIENZUDO
60	0.3387	0.3076	0.3249	0.367	4	CONCIENZUDO
65	0.3264	0.348	0.3137	0.3914	4	CONCIENZUDO
62	0.3165	0.2981	0.3048	0.2822	1	CONCIENZUDO
34	0.1575	0.1181	0.134	0.1314	1	CONCIENZUDO
11	0.1261	0.1082	0.1266	0.1085	3	CONCIENZUDO
74	0.4468	0.4199	0.435	0.464	4	CONCIENZUDO
20	0.0475	0.0434	0.0364	0.0516	4	CONCIENZUDO
26	0.1736	0.1355	0.1545	0.1423	1	CONCIENZUDO
82	0.6384	0.4546	0.5388	0.4654	1	DOMINANTE
67	0.4719	0.4013	0.3787	0.3979	1	DOMINANTE
49	0.174	0.1095	0.1323	0.1095	1	DOMINANTE
32	0.1999	0.144	0.1845	0.1702	1	DOMINANTE

Tabla 26. Ejemplo de los registros utilizados en la prueba 7

Resultados: En la tabla 27 se pueden ver los resultados obtenidos en esta prueba.

Algoritmo	Precisión	Cobertura	Medida F	Método de evaluación	Instancias clasificadas correctamente
SMO	26.7%	51.7%	35.2%	*Ten-fold cross validation*	51.6%
SMO	44.4%	66.7%	53.3%	*Percentage Split*	66.6%

Tabla 27. Resultados de la prueba 7

Como se puede observar en la tabla 27 los resultados del clasificador al evaluar el corpus aún muestran muy bajos porcentajes, pero el porcentaje de instancias clasificadas correctamente obtenido con el método de evaluación *percentage split* es más alto que en las pruebas anteriores.

6.12 Prueba 8: Conjunto de palabras que denotan personalidad eliminando verbos de *stopwords*

En esta prueba se hace uso del conjunto de palabras que denotan personalidad nuevamente para calcular el peso TF-IDF de cada texto en cada factor de personalidad. En esta prueba se eliminan del conjunto de palabras todos los verbos que pertenecen al grupo de *stopwords* y se calculan nuevamente los pesos TF-IDF de los textos afectados, textos en los que aparecen las palabras eliminadas. Cada texto es etiquetado con un peso para cada tipo de personalidad en el corpus y se ejecuta nuevamente la prueba de clasificación. El método de evaluación *percentage split* se utilizó con 70% del corpus para entrenamiento del algoritmo y 30% para prueba.

Datos: *Número de palabras, dominante TF-IDF, influyente TF-IDF, estable TF-IDF, concienzudo TF-IDF, TF-IDF mayor y resultado de prueba*. En esta prueba los pesos TF-IDF son calculados después de eliminar verbos que pertenecen a la lista de *stopwords*. En la tabla 28 se muestra un ejemplo de los registros utilizados para esta prueba.

Num_palabras	Dominante_TF-IDF	Influyente_TF-IDF	Stable_TF-IDF	Concienzudo_TF-IDF	TF-IDF Mayor	Resultado_prueba
78	0.1499	0.1612	0.124	0.1576	2	CONCIENZUDO
79	0.0915	0.0955	0.0754	0.1204	4	CONCIENZUDO
69	0.0732	0.0831	0.0785	0.1485	4	CONCIENZUDO
66	0.0663	0.0835	0.0809	0.1117	4	CONCIENZUDO
56	0.0852	0.0765	0.0999	0.1399	4	CONCIENZUDO
85	0.25	0.2324	0.2416	0.2804	4	CONCIENZUDO
92	0.0413	0.0671	0.0637	0.1325	4	CONCIENZUDO
31	0.0176	0.0062	0.0159	0.0295	4	CONCIENZUDO
9	0.0152	0.0245	0.0244	0.0288	4	CONCIENZUDO
30	0.0473	0.0482	0.0569	0.0953	4	CONCIENZUDO
60	0.1331	0.1236	0.1157	0.1921	4	CONCIENZUDO
65	0.0871	0.1024	0.0954	0.1209	4	CONCIENZUDO
62	0.1104	0.1398	0.1258	0.1276	2	CONCIENZUDO
34	0.0419	0.0262	0.0283	0.0411	1	CONCIENZUDO
11	0.017	0.02	0.0264	0.0185	3	CONCIENZUDO
74	0.0608	0.0556	0.061	0.0863	4	CONCIENZUDO
20	0.0606	0.0535	0.0449	0.0663	4	CONCIENZUDO
26	0.0417	0.0435	0.0373	0.0531	4	CONCIENZUDO
82	0.1118	0.0715	0.0823	0.084	1	DOMINANTE
67	0.2406	0.2011	0.1837	0.1745	1	DOMINANTE
49	0.0382	0.007	0.0056	0.0075	1	DOMINANTE
32	0.0487	0.0362	0.0451	0.0565	4	DOMINANTE

Tabla 28. Ejemplo de los registros utilizados en la prueba 8

Resultados: En la tabla 29, se pueden ver los resultados obtenidos en esta prueba. Se puede ver que la precisión aumento a 41.6% y 67.9% y el porcentaje de instancias clasificadas correctamente aumento hasta un 75%.

Algoritmo	Precisión	Cobertura	Medida F	Método de evaluación	Instancias clasificadas correctamente
SMO	41.6%	53.3%	41.2%	*Ten-fold cross validation*	53.3%
SMO	67.9%	75%	67.8%	*Percentage Split*	75%

Tabla 29. Resultados de la prueba 8

6.13 Prueba 9: Eliminando verbos repetidos

En esta prueba analizan las pruebas realizadas anteriormente y se identifican tres verbos (ir, terminar y jugar) que aparecen muchas veces en la mayoría de los textos. Se decide eliminar estos tres verbos porque aparecen en casi todos los textos y no ayudan a diferenciarlos unos de otros, causando un bajo porcentaje en la clasificación de personalidad. El método de evaluación *percentage split* se utilizó con 70% del corpus para entrenamiento del algoritmo y 30% para prueba.

Datos: *Número de palabras, dominante TF-IDF, influyente TF-IDF, estable TF-IDF, concienzudo TF-IDF, TF-IDF mayor y resultado de prueba*. En esta prueba los pesos TF-IDF son calculados después de eliminar verbos que pertenecen a la lista de *stopwords* y los verbos ir, terminar y jugar. En la tabla 30 se muestra un ejemplo de los registros utilizados para esta prueba.

NumPalabras	DominanteTFIDF	InfluyenteTFIDF	StableTFIDF	ConcienzudoTFIDF	TF-IDF Mayor	Personalidad
58	0.0533	0.0825	0.0726	0.0958	4	CONCIENZUDO
59	0.0271	0.043	0.0411	0.0792	4	CONCIENZUDO
54	0.0732	0.0831	0.0785	0.1485	4	CONCIENZUDO
41	0.0341	0.0573	0.0637	0.0911	4	CONCIENZUDO
48	0.0303	0.0241	0.0376	0.0465	4	CONCIENZUDO
76	0.0966	0.097	0.0963	0.1238	4	CONCIENZUDO
67	0.0413	0.0671	0.0637	0.1325	4	CONCIENZUDO
17	0.0176	0.0062	0.0159	0.0295	4	CONCIENZUDO
7	0.0152	0.0245	0.0244	0.0288	4	CONCIENZUDO
25	0.0189	0.0203	0.0284	0.0335	4	CONCIENZUDO
48	0.0422	0.0466	0.0476	0.1192	4	CONCIENZUDO
53	0.036	0.0576	0.0593	0.059	3	CONCIENZUDO
51	0.0612	0.0984	0.0843	0.096	2	CONCIENZUDO
29	0.0419	0.0262	0.0283	0.0411	1	CONCIENZUDO
9	0	0.0048	0.002	0.0075	4	CONCIENZUDO
50	0.0116	0.0142	0.0195	0.0547	4	CONCIENZUDO
16	0.0189	0.018	0.0183	0.0251	4	CONCIENZUDO
20	0.0095	0.0172	0.0202	0.0325	4	CONCIENZUDO
59	0.1099	0.0681	0.0841	0.0702	1	DOMINANTE
55	0.0683	0.0564	0.0569	0.0495	1	DOMINANTE
41	0.0382	0.007	0.0056	0.0075	1	DOMINANTE

Tabla 30. Ejemplo de los registros utilizados en la prueba 9

Resultados: En la tabla 31 se muestran los resultados de esta prueba.

Algoritmo	Precisión	Cobertura	Medida F	Método de evaluación	Instancias clasificadas correctamente
SMO	80%	78.3%	77.6%	*Ten-fold cross validation*	78.3%
SMO	85.3%	83.3%	81.7%	*Percentage Split*	83.3%

Tabla 31. Resultados de la prueba 9

Se puede observar que en esta prueba se obtiene el porcentaje más alto que en cualquier otra de las pruebas realizadas llegando hasta un 85% de precisión del clasificador con el método de evaluación *percentage split*. Es por esto que se decide utilizar este corpus y este método de evaluación para generar el modelo de clasificación para el módulo de clasificación automática del servicio Web.

6.14 Resumen de resultados

En esta sección se presenta un resumen de los resultados de las pruebas realizadas. La primer prueba se llevó a cabo con la información sin procesar esto quiere decir que los datos se utilizan tal como fueron obtenidos de las encuestas de personalidad *DISC* aplicadas. Después de aplicar distintos algoritmos de clasificación se observó que dos de ellos *SMO* y *Multilayer Perceptron* mostraban los porcentajes de precisión más altos. Pero en otras medidas de evaluación y en el tiempo de procesamiento, el algoritmo *Multilayer Perceptron* presentó valores menos favorables. Es por esto que se decidió comenzar a realizar las otras pruebas con el algoritmo *SMO*.

En las pruebas 2, 3, 4, 5 y 6 se decidió trabajar con el corpus de personalidad como un vector de palabras, en primer lugar tomando cada verbo y adjetivo de cada texto y usándolos a todos como una característica diferente. Un peso TF-IDF fue calculado para cada palabra y registrado en el corpus. En segundo lugar, debido al bajo porcentaje presentado al realizar la prueba se decidió convertir los pesos de las palabras a valores binarios ya que algunos algoritmos trabajan mejor con datos binarios, pero igualmente se obtuvieron muy bajos resultados.

Se revisaron los resultados obtenidos en las pruebas anteriores y se decidió crear un conjunto de palabras que denotan personalidad. El conjunto de palabras se formó con todos los verbos y adjetivos de todos los textos obtenidos en la fase de recolección de datos del modelo para la detección de personalidad. Se revisaron las apariciones de esas palabras en cada texto y en cada factor de personalidad. Se calculó un peso para cada una de las palabras que permitiera diferenciar el valor que tenían unas de otras en cada personalidad. Los pesos de las palabras fueron calculados y se utilizaron para revisar cada texto y calcular su peso basándonos en el léxico en las pruebas 7, 8 y 9.

Se eliminaron *stopwords* y aun se presentaba un porcentaje de clasificación muy bajo. En la primera eliminación de *stopwords* no se eliminaron los verbos que pertenecían a esta lista, por lo que estaban causando el bajo porcentaje en la clasificación. Se decidió eliminar esos verbos en la siguiente prueba (prueba 8) y se observó que el porcentaje de clasificación aumentó. Entonces se decidió identificar palabras que se estuvieran repitiendo mucho en los textos del corpus. Palabras que estuvieran desempeñando un papel igual a las *stopwords* y es así que se identificaron tres verbos, los verbos; ir, terminar y jugar aparecían en la mayoría de los textos y se repetían muchas veces, es por esto que se llegó a la conclusión de eliminarlas y llevar a cabo la prueba número 9.

Los mejores resultados de las pruebas realizadas con todos los corpus se obtienen con la prueba 9, la cual implica el uso del conjunto de palabras que denotan personalidad para el cálculo y etiquetado de los textos con su correspondiente peso TF-IDF en cada personalidad, la eliminación de *stopwords* incluyendo todos los verbos que se encuentren en la lista de *stopwords* y también la eliminación de los verbos ir, terminar y jugar. En la tabla 32 se muestran los resultados obtenidos en las pruebas 1 – 9 después de llevar a cabo la evaluación de algoritmos.

No	Prueba	Precisión	Cobertura	Medida F	Evaluación
1	Sin procesamiento	47%	52.5%	45.8%	*Cross validation*
		46.9%	55.6%	50.4%	*Percentage Split*
2	Convirtiendo palabras a características	42%	45%	41.1%	*Cross validation*
		49.8%	55.6%	49.8%	*Percentage Split*
3	Corpus de palabras en binario	42%	45%	41.1%	*Cross validation*
		26.6%	33.3%	29.6%	*Percentage Split*
4	Agregando otras características del texto	39.6%	45.8%	41.5%	*Cross validation*
		49.8%	55.6%	49.8%	*Percentage Split*
5	Binario agregando otras características del texto	48.1%	45%	40.5%	*Cross validation*
		29.7%	38.9%	33.5%	*Percentage Split*
6	Igualar número de clases	29.1%	26.8%	27.3%	*Cross validation*
	Igualar número de clases Binario	27.8%	28.6%	28%	
	Igualar número de clases	17%	23.5%	19%	*Percentage Split*
	Igualar número de clases Binario	25.5%	23.5%	23.5%	
7	Conjunto de palabras que denotan personalidad incluyendo verbos de *stopwords*	26.7%	51.7%	35.2%	*Cross validation*
		44.4%	66.7%	53.3%	*Percentage Split*
8	Conjunto de palabras que denotan personalidad eliminando verbos de stopwords	41.6%	53.3%	41.2%	*Cross validation*
		67.9%	75%	67.8%	*Percentage Split*
9	Eliminando verbos repetidos	80%	78.3%	77.6%	*Cross validation*
		85.3%	83.3%	81.7%	*Percentage Split*

Tabla 322. Resultados de la prueba 9

En la tabla 33 se muestra la comparación de los resultados obtenidos en este trabajo de investigación con los resultados de los trabajos relacionados revisados en el estado del arte. Se puede observar que en los trabajos relacionados que utilizan el método de análisis de texto se obtiene un porcentaje de precisión de 83% en el mejor de los casos. Sólo un trabajo que utiliza el análisis del comportamiento en redes sociales obtiene 85% de precisión en sus resultados.

En este trabajo de investigación se obtiene 85.3% de precisión como resultado del cálculo de personalidad al analizar texto. También se puede observar que en ninguno de los trabajos relacionados se generan recursos para el cálculo de la personalidad en este trabajo de investigación se construyen dos recursos linguisticos útiles para el cálculo de personalidad uno de ellos es un conjunto de palabras que denotan personalidad basado en el modelo *DISC* y el otro recurso construido es un corpus de personalidad basado también en el modelo *DISC*. Además en los trabajos relacionados no se trabaja con textos en español a excepción de uno que trabaja con textos en español de España, en este trabajo se trabaja con textos en Español de México y se construye un servicio Web para el cálculo automatico de la personalidad.

Trabajo	Método	idioma	Generación de recursos	Precisión	Software
Prediciendo la personalidad con la conducta social	Análisis de Conducta social	Inglés	NO	85%	NO
El reconocimiento de la personalidad no supervisado en sitios de redes sociales	Análisis de Textos	Italiano	NO	63.1%	NO
Máquina Predicción de la personalidad de los perfiles de Facebook	Análisis de datos demográficos y Texto	Inglés	NO	74.5%	NO
Un Sistema de detección de la personalidad y de la felicidad	Análisis de Textos	Español (España)	NO	57%	Aplicación Móvil (sistema operativo Android)
Un enfoque de clasificación multi-etiqueta, semi-supervisado aplicado a la predicción de la personalidad en los medios sociales	Análisis de Textos	Inglés	NO	83%	PERSOMA programa en Java
Clasificación personalidad basada en textos de Twitter usando Naive Bayes, KNN y SVM	Análisis de Textos	Inglés, Indonesio	NO	60%	Aplicación Web
La predicción de rasgos de la personalidad de los usuarios chinos basada en publicaciones de muro de Facebook.	Análisis de Textos	Chino	NO	73.5%.	NO
Servicio Web para la detección automática de personalidad a través del análisis lingüístico de textos	Análisis de Textos	Español (México)	• Conjunto de palabras que denotan personalidad. • Corpus de personalidad DISC.	85.3%	Servicio Web

Tabla 333. Comparacion de resultados con trabajos relacionados

Capítulo 7

Conclusiones y trabajo futuro

En esta sección se presentan las conclusiones y contribuciones generadas durante el desarrollo de este proyecto de investigación. También se describen los trabajos futuros que se puede derivar a partir de este trabajo. Finalmente, se listan las actividades realizadas.

7.1 Conclusiones

La motivación principal para el desarrollo del presente trabajo de investigación surge de la necesidad de poder conocer mejor la conducta y características de comportamiento de una persona a través del cálculo de la personalidad, sin necesidad de aplicar una prueba, de manera rápida y saber con qué tipo de persona se está tratando. Esta información resulta de gran utilidad para distintas instituciones. Otra motivación importante para el desarrollo de este trabajo de investigación fue poder generar recursos lingüísticos y herramientas que permitan analizar y clasificar textos en español de manera automática, en este caso clasificar textos de acuerdo con su personalidad con base en el modelo de personalidad *DISC* que representa la personalidad en cuatro factores; Dominante, Influyente, Estable y Concienzudo.

En este trabajo de investigación se desarrollo un modelo para la detección automática de la personalidad de un sujeto a través del análisis linguistico de texto. Además se desarrolló un servicio Web que implementa el modelo creado. El servicio Web recibe como entrada un texto

del usuario de manera escrita o a través de la carga de un archivo de texto plano y después de que el servicio realiza el análisis del mismo se obtiene como resultado la personalidad del usuario.

De acuerdo a la investigación realizada, en general no existen muchos recursos en español para las tareas de análisis lingüísticos y particularmente para la detección de personalidad en el idioma español son escasos. Por lo que en este trabajo se creó un conjunto de palabras que denotan personalidad basado en el modelo de personalidad *DISC* y un corpus con textos etiquetados con pesos para a cada factor de personalidad del modelo *DISC*. Para poder asignar pesos a las palabras y a los textos se realizó la revisión, conteo y cálculo TF-IDF manualmente.

Al realizar las pruebas de clasificación sin el uso del conjunto de palabras que denotan personalidad se pudo observar que los resultados no mejoraban por lo que el conjunto de palabras que denotan personalidad resulta de gran utilidad, aporta el peso de cada palabra que ayudan en conjunto a determinar el peso de un texto completo en cada una de las personalidades es importante mencionar que la eliminación de *stopwords* es necesaria para poder disminuir el ruido en la información, al eliminarlas la mejora se ve reflejada en el incremento de los porcentajes de precisión.

Es importante resaltar que el modelo fue realizado tomando una muestra de sujetos universitarios, por lo que el modelo tendrá una mayor precisión en un dominio similar.

Para poder realizar la determinación de la personalidad a través del texto se llevo a cabo una clasificación de textos del corpus de personalidad que se evaluó utilizando el algoritmo *SMO*, las técnicas *ten-fold cross validation*, y *percentaje Split* y se entrenó con dos corpus con distintas características del texto obteniendo como porcentaje de precisión mas alto 85.3%. Los porcentajes más significativos de acuerdo con la revisión de los trabajos relacionados fueron de 83% (Lima & de Castro, 2014) y 85% (Adali & Golbeck, 2012). Aunque es importante mencionar que los modelos de personalidad y los métodos utilizados para su detección difieren unos de otros, además que la mayoría de trabajos están desarrollados para el idioma inglés, no desarrollan ninguna herramienta para la detección automática de la personalidad y se apoyan de diversos recursos lingüísticos con los que no se cuentan en el idioma español, lo que se hace más difícil llevar a cabo esta tarea.

7.2 Contribuciones

El objetivo general que se definió para este proyecto de investigación consistió en desarrollar un modelo para la detección automática de personalidad mediante el análisis lingüístico de textos y el desarrollo de un servicio Web que implementa ese modelo para determinar la personalidad de textos en español.

En el cumplimiento de este objetivo se lograron las siguientes contribuciones:

a) Se construyó un modelo para la detección automática de la personalidad mediante el análisis de textos en español.

b) Se desarrolló un servicio Web para el cálculo de personalidad mediante el análisis de texto que implementa el modelo construido, el servicio Web está formado por tres módulos:

I) Módulo de preprocesamiento: En el primer módulo se lleva a cabo el pre-procesamiento del texto; recibe como entrada un texto, se realiza la *tokenizacion*, eliminación de *stopwords*, *lematizacion* y etiquetado gramatical del texto.

II) Módulo de obtención de características: En el segundo módulo se lleva a cabo la extracción de características del texto; recibe como entrada el texto pre-procesado en el primer módulo y se genera un vector con el número de palabras en el texto, el peso TF-IDF general del texto en cada factor DISC y un identificador del factor de personalidad que obtiene el valor TF-IDF más alto, basándose en el léxico de personalidad DISC.

III) Módulo de clasificación automática: En el tercer módulo se lleva a cabo la clasificación automática; la entrada son los parámetros del vector que se genera en el módulo de extracción de características y da como resultado la personalidad del usuario.

c) Un conjunto de palabras en español que denotan personalidad basado en el modelo *DISC* a partir del cálculo de pesos TF-IDF. El conjunto de palabras quedó constituido por 346 verbos y 159 adjetivos diferentes y se presenta en el anexo 3 *Conjunto de palabras que denotan personalidad basado en el modelo DISC*.

d) Un corpus de textos etiquetados con sus pesos TF-IDF correspondientes a cada factor de personalidad del modelo *DISC* (dominante, influyente, estable y concienzudo), útil para poder llevar a cabo la detección de personalidad.

7.3 Trabajo futuro

Para ampliar el desarrollo de este proyecto de investigación, se propone la realización del siguiente trabajo futuro en el que este trabajo puede ser mejorado.

- Incrementar el tamaño del conjunto de palabras que denotan personalidad, al incrementar el tamaño de palabras en el conjunto el sistema será capaz de dar un peso más exacto a cada texto mejorando la precisión de la clasificación. También pueden incluirse más tipos de palabras al conjunto de palabras que denotan personalidad como los sustantivos.

- Identificar características representativas de los textos, las características que más identifiquen a un texto mejoran mucho el resultado de la clasificación se podrían buscar otras características que den más información sobre cada texto.

- Ampliar la variedad y cantidad de registros de resultados del corpus aplicando la prueba de personalidad a personas de otras edades, ocupaciones, estado civil etc. esto permitirá enriquecer la información del corpus con más palabras utilizadas por personas de distintas características.

- La creación o implementación de una base de datos con información lingüística sobre el texto en idioma español de México, en la revisión del estado del arte se encontró que muchos proyectos en idioma inglés cuentan con recursos como las bases de datos lingüísticas que cuentan con un gran número de diversas características de las palabras que proporcionan información en muchas categorías del texto y esta información resulta de gran utilidad para mejorar los resultados.

- La Implementación de otros modelos de personalidad; debido a que la personalidad es algo abstracto que no se puede medir, han surgido muchos modelos de personalidad que intentan medir este abstracto como el modelo Big Five (Christal, 1992) , modelo PEN (Eysenck, 1950) entre otros. Los modelos son variados, diferentes y parecidos en algunos aspectos (Bausela, 2005). Para poder enriquecer y proporcionar un resultado más completo del análisis sería recomendable comparar e implementar más modelos de personalidad.

7.4 Publicación realizada

Y. Hernández, C. Acevedo, A. Martínez (2017). ***Towards a linguistic corpus in Spanish with personality annotations.*** *In proceedings of 16th Mexican International Conference on Artificial Intelligence.* In press.Ensenada, Baja California, Mexico, 24 de Octubre 2017.

Referencias

Adali, S., y Golbeck, J. (2012). Predicting Personality with Social Behavior. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 302–309. doi: 10.1109/ASONAM.2012.58

Agung, A. A., y Yunair, I. (2016). Personality assessment website using DISC: A case study in information technology school. *International Conference on Information Management and Technology (ICIMTech).* 72-77. doi: 10.1109/ICIMTech.2016.7930305.

Antonín, M. M. A., (2003). *Tecnologías del lenguaje*. Catalunya, España: Editorial UOC.

Axiom, S. (2018). What is DISC?. discusonline. https://www.discusonline.com/es-es/disc/what-is-disc.php

Apply Magic Sauce (2018). *applymagicsauce*. Recuperado de http://applymagicsauce.com/

Baca, G. Y. R. (2014). Desarrollo de un Servicio Web para Determinar la Polaridad de Textos de Redes Sociales en Español (tesis de maestria). Centro Nacional de Investigacion y Desarrollo Tecnologico, Cuernavaca, Morelos, México.

Bausela, H. E. (2005). Modelo alternativos de evaluación de la personalidad: modelo de los cinco factores, modelo 16 pf y otros. *Avances En Salud Mental Relacional*, *4*(2), 29. Recuperado de http://www.bibliopsiquis.com/asmr/0402/adv.pdf

Bradberry, T. (2008). *El Codigo de La Personalidad*. Bogotá, Colombia: Editorial Norma. Recuperado de https://books.google.com.mx/books?id=O7de7ggKkO0C&printsec=frontcover#v=onepage&q&f=false

Cantador, I., Fernández-Tobías, I., & Bellogín, A. (2013). Relating personality types with user

preferences in multiple entertainment domains. *CEUR Workshop Proceedings, 997. ResearchGate*. Recuperado de https://www.researchgate.net/publication/283270671_Relating_Personality_Types_with_User_Preferences_Multiple_Entertainment_Domains

Celli, F. (2012). Unsupervised Personality Recognition for Social Network Sites. *The Sixth International Conference on Digital Society*, (c), 59–62. Recuperado de https://www.researchgate.net/publication/258045593_Unsupervised_Personality_Recognition_for_Social_Network_Sites

Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013). Workshop on Computational Personality Recognition : Shared Task. *Proceedings of the Workshop on Personality Recognition*, 2–5.

Christal, E. T. and R. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality*, *60*(2), 225–251.Recuperado de http://doi.org/10.1111/j.1467-6494.1992.tb00973.x

Conole, G., Galley, R., y Culver, J. (2011). Frameworks for understanding the nature of interactions, networking, and community in a social networking site for academic practice. *International Review of Research in Open and Distance Learning*, *12*(3), 119–138. Recuperado de http://doi.org/10.1111/j.1083-6101.2007.00393.x

Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando, 11. Recuperdado de http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf

Cuadrado, A. J. C. (2011). Un Modelo Lingüístico-Semántico Basado en Emociones para la Clasificación de Textos según su Polaridad e Intensidad. (tesis de doctoral). Universidad Complutense de Madrid, Madrid, España.

Deitel, P. J. y Deitel, H. M. (2008). *Cómo Programar en Java*. Monterrey, Mexico: Editorial: Pearson. Recuperado de http://doi.org/9702605318

Discprofiles4u (2018), *DISC behavioral styles.* https://www.discprofiles4u.com/pages/DiSC-Behavioral-Styles.html

Discprofile.(2008). https://www.discprofile.com/what-is-disc/research-reliability-and-validity/

Eysenck, H. J. (1950). *Dimensions of Personality*. Estados Unidos de America: Editorial Transaction Publishers.

Freeling (2018). Welcome FreeLing home page. http://nlp.lsi.upc.edu/freeling/node/1

Galindo,M. y Aguilar, J.(2004). Clasificación de la personalidad y sus trastornos, con la herramienta Lamda de inteligencia artificial en una muestra de personas de origen hispano que viven en Toulouse-Francia, *Revista de Estudios Sociales*, 18, 99-110.

Goh, J. X., Schlegel, K., Tignor, S. M., y Hall, J. A. (2016). Who is interested in personality? The Interest in Personality Scale and its correlates. *Personality and Individual Differences*, *101*, 185–191. Recuperado de http://doi.org/10.1016/j.paid.2016.05.366

Golbeck, J., Robles, C., Edmondson, M., y Turner, K. (2011). Predicting personality from twitter. *IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing*, 149–156. Recuperado de http://doi.org/10.1109/PASSAT/SocialCom.2011.33

Guisasola, V. (2016). Cómo saber si una Evaluación ? DISC es fiable. https://es.linkedin.com/pulse/cómo-saber-si-una-evaluación-disc-es-fiable-virginia-guisasola

John, O. P., y Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research,* 2, 102–138.

Laurence, A. (2018). AntConc. http://www.laurenceanthony.net/software/antconc/

Lima, E. S. A. C., y de Castro, L. N. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks*, *58*, 122–130. Recuperado de http://doi.org/10.1016/j.neunet.2014.05.020

Luyckx, K., Daelemans, W., & European Language Resources, A. (2008). Personae: a corpus for author and personality prediction from text. *Sixth International Conference on Language Resources and Evaluation*.

Mairesse, F., Walker, M. A., Mehl, M. R., y Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, *30*, 457–500.Recuperado de http://doi.org/10.1613/jair.2349

Marston, W. M. (1928). *Emotions of normal people*.Nueva York: Editorial harcourt, brace and company. Recuperado de http://doi.org/10.1037/13390-000

Martinez,M. J., Alvarado, M., Aldea, A. y Bañares, A. R. (2005). Modelling Human Behaviour at

Work using Fuzzy Logic: The Challenge of Work Teams Configuration. *Semanticscholar*, 1-24. Recuperado de https://pdfs.semanticscholar.org/6b3e/b86ceb7bba56426b2ebd99f4e b8448553a98.pdf

Padró, L. (2011). Analizadores Multilingues en FreeLing. *Linguamatica*. 3. 13-20. Recuperado de http://nlp.lsi.upc.edu/publications/papers/padro11.pdf

Peng, K.-H., Liou, L.-H., Chang, C.-S., y Lee, D.-S. (2015). Predicting personality traits of Chinese users based on Facebook wall posts. *24th Wireless and Optical Communication Conference*, 9–14. Recuperado de http://doi.org/10.1109/WOCC.2015.7346106

Pratama, B. Y., & Sarno, R. (2015). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. *International Conference on Data and Software Engineering* 170–174. Recuperado de http://doi.org/10.1109/ICODSE.2015.7436992

Prevue. prevuehr. Recuperado de https://www.prevuehr.com/

PsicoSmart.HumanSmart. Recuperado de http://humansmart.com.mx/sistema-de-psicometria-laboral-en-linea-software-online-pruebas-psicometricas#ul-id-17-2

Sáez, Y., Navarro, C., Mochón, A., y Isasi, P. (2014). A System for Personality and Happiness Detection. *Ijimai*, *2*(5), 8–16. Recuperado de http://doi.org/10.9781/ijimai.2014.251

Taulé, M. (2003). SENSEVAL, una aproximación computacional al significado, 1–13. Recuperado de http://www.uoc.edu/humfil/articles/esp/taule0303/taule0303.html

The DISC Insights Web Development Team (s.f.). https://www.discinsights.com/

Trim, C. (2018). The Art of Tokenization. IBM. https://www.ibm.com/developerworks/ community/blogs/nlp/entry/tokenization?lang=en

Vera, P. S. A. (2017). Análisis y Clasificación de Textos con Técnicas Semi Supervisadas Aplicado a Área Atención al Cliente (Tesis de pregrado) , 54. Recuperado de opac.pucv.cl/pucv_txt/txt-8500/UCD8533_01.pdf

Wald, R., Khoshgoftaar, T., y Sumner, C. (2012). Machine prediction of personality from Facebook profiles. *IEEE 13th International Conference on Information Reuse and Integration,* 109–115.Recuperado de http://doi.org/10.1109/IRI.2012.6302998

Anexos

Anexo 1

Formato de encuesta *DISC*

Un plan para autoconocerte

En el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) estamos realizando un estudio sobre la personalidad de los individuos, es decir, sobre aquellas características que definen la forma de ser de cada persona. Solicitamos su colaboración para participar en una encuesta que consta de 3 fases. El tiempo requerido para dicha encuesta es de 10-15 min aproximadamente. Es importante resaltar que toda la información recabada es estrictamente confidencial y no será utilizada para ningún otro fin distinto a esta investigación.

Fase 1. Se le solicita llene la siguiente información personal. En caso de no requerir una retroalimentación no es necesario colocar su nombre y correo electrónico.

Género: ◯ Masculino ◯ Femenino

Rango de edad: ◯ Entre 20 - 30 años ◯ Entre 31 - 40 años

◯ Entre 41 - 50 años ◯ Más de 50 años

Escolaridad: ◯ Preparatoria ◯ Universidad (Carrera) ◯ Posgrado

Estado civil: ◯ Soltero ◯ Casado

Ocupación o Profesión:_____

Puesto que ocupa: _____

Redes sociales que utiliza: ◯ Twitter ◯ Facebook

Numero de amigos en redes sociales: [_____]

Correo electrónico:_____

Fase 2. Consiste en contestar un cuestionario que recolecta información de algunas características que te identifican. **Instrucciones:** Elije una palabra que te identifique "más" y una palabra que te identifique "menos" en cada uno de los grupos.

1			8			15			22		
	MAS	MENOS		MAS	MENOS		MAS	MENOS		MAS	MENOS
Entusiasta			Extrovertido			Popular			Impulsivo		
Rápido			Precavido			Reflexivo			cuida los detalles		
Lógico			Constante			Tenaz			Enérgico		
Apacible			Impaciente			Calmado			Tranquilo		
2			9			16			23		
Cauteloso			Discreto			Analítico			Sociable		
Decidido			Complaciente			Audaz			Sistemático		
Receptivo			Encantador			Leal			Vigoroso		
Bondadoso			Insistente			Promotor			Tolerante		
3			10			17			24		
Amigable			Valeroso			Sociable			Cautivador		
Preciso			Anima a los demás			Paciente			Contento		
Franco			Pacífico			Autosuficiente			Exigente		
Tranquilo			Perfeccionista			Certero			Apegado a las normas		
4			11			18			25		
Elocuente			Reservado			Adaptable			Discute con frecuencia		
Controlado			Atento			Resuelto			Metódico		
Tolerante			Osado			Prevenido			Comedido		
Decisivo			Alegre			Vivaz			Desenvuelto		
5			12			19			26		
Atrevido			Estimulante			Agresivo			Jovial		
Concienzudo			Gentil			Impetuoso			Preciso		
comunicativo			Perceptivo			Amistoso			Directo		
Moderado			Independiente			Discerniente			Ecuánime		
6			13			20			27		
Ameno			competitivo			De trato Fácil			Inquieto		
Ingenioso			Considerado			Compasivo			Amable		
Investigador			Alegre			Cauto			Elocuente		
Acepta riesgos			Sagaz			Habla directo			Cuidadoso		
7			14			21			28		
Expresivo			Meticuloso			Evaluador			Prudente		
Cuidadoso			Obediente			Generoso			Pionero		
Dominante			Ideas firmes			Animado			Espontáneo		
Sensible			Alentador			Persistente			Colaborador		

Fase 3. Se le solicita escriba un texto mínimo XX líneas en esta página. Algunas ideas de lo que puede escribir es:

- Que hizo el día de ayer.
- Cuáles son sus principales metas
- Describa su pasatiempo favorito
- Escribir una historia
- Etc.
- También puede servirnos que nos envíe por correo (*personalidadproyecto@gmail.com*) algún correo que usted haya escrito o algún texto que tenga, que haya redactado.

Gracias por tu participación.

REDACCIÓN TEXTO:

Dudas y sugerencias puedes escribirnos al correo:
personalidadproyecto@gmail.com

Anexo 2

Lista de *stopwords*

a	cinco	dijo	estos	informó	ningunas
actualmente	comentó	dio	éstos	junto	ninguno
adelante	como	donde	estoy	la	ningunos
además	cómo	dos	estuvo	lado	no
afirmó	con	durante	ex	las	nos
agregó	conocer	e	existe	le	nosotras
ahí	considera	ejemplo	existen	les	nosotros
ahora	consideró	el	explicó	llegó	nuestra
al	contra	él	expresó	lleva	nuestras
algo	cosas	ella	fin	llevar	nuestro
algún	creo	ellas	fue	lo	nuestros
alguna	cual	ello	fuera	los	nueva
algunas	cuales	ellos	fueron	luego	nuevas
alguno	cualquier	embargo	gran	lugar	nuevo
algunos	cuando	en	grandes	manera	nuevos
alrededor	cuanto	encuentra	ha	manifestó	nunca
ambos	cuatro	entonces	haber	más	o
ante	cuenta	entre	había	mayor	ocho
anterior	da	era	habían	me	otra
antes	dado	eran	habrá	mediante	otras
añadió	dan	es	hace	mejor	otro
apenas	dar	esa	hacen	mencionó	otros
aproximadamente	de	esas	hacer	menos	para
aquí	debe	ese	hacerlo	mi	parece
aseguró	deben	eso	hacia	mientras	parte
así	debido	esos	haciendo	misma	partir
aún	decir	esta	han	mismas	pasada
aunque	dejó	está	hasta	mismo	pasado
ayer	del	ésta	hay	mismos	pero
bajo	demás	estaba	haya	momento	pesar
bien	dentro	estaban	he	mucha	poca
buen	desde	estamos	hecho	muchas	pocas
buena	después	están	hemos	mucho	poco
buenas	dice	estar	hicieron	muchos	pocos
bueno	dicen	estará	hizo	muy	podemos
buenos	dicho	estas	hoy	nada	podrá
cada	dieron	éstas	hubo	nadie	podrán
casi	diferente	este	igual	ni	podría
cerca	diferentes	éste	incluso	ningún	podrían
cierto	dijeron	esto	indicó	ninguna	poner

por	será	todas
porque	serán	todavía

posible
primer
primera
primero
primeros
principalmente
propia
propias
propio
propios
próximo
próximos
pudo
pueda
puede
pueden
pues
que
qué
quedó
queremos
quien
quién
quienes
quiere
realizado
realizar
realizó
respecto
se
sea
sean
según
segunda
segundo
seis
señaló
ser
sería
si
sí
sido
siempre
siendo
siete
sigue
siguiente
sin
sino
sobre
sola
solamente
solas
solo
sólo
solos
son
su
sus
tal
también
tampoco
tan
tanto
tendrá
tendrán
tenemos
tener
tenga
tengo
tenía
tenido
tercera
tiene
tienen
toda
todo
todos
total
tras
trata
través
tres
tuvo
última
últimas
último
últimos
un
una
unas
uno
unos
usted
va
vamos
van
varias
varios
veces
ver
vez
y
ya
yo

Anexo 3

Conjunto de palabras que denotan personalidad basado en el modelo *DISC*

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Académico	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Actual	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Acústico	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Ágil	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Agradable	1	1	0	0	2	0.0025	0.0011	0.0000	0.0000	Adjetivo
Alto	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Amable	0	2	0	0	1	0.0000	0.0027	0.0000	0.0000	Adjetivo
Amigable	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Amiotrofica	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Artificial	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Bello	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Adjetivo
Benéfico	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Bonito	2	2	3	1	4	0.0038	0.0017	0.0014	0.0014	Adjetivo
Brillante	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Calórico	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Capaz	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Claro	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Cómodo	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Completo	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Computacional	1	1	3	0	3	0.0021	0.0010	0.0015	0.0000	Adjetivo
Común	1	1	2	0	3	0.0021	0.0010	0.0010	0.0000	Adjetivo
Consciente	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Contento	0	2	0	0	1	0.0000	0.0027	0.0000	0.0000	Adjetivo
Corto	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Adjetivo
Critico	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Delgado	0	2	0	0	1	0.0000	0.0027	0.0000	0.0000	Adjetivo
Delicado	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Deportivo	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Desapercibido	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Desesperante	0	0	0	2	1	0.0000	0.0000	0.0000	0.0044	Adjetivo
Diario	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Adjetivo
Difícil	0	2	1	1	3	0.0000	0.0019	0.0005	0.0015	Adjetivo
Digno	0	0	2	1	2	0.0000	0.0000	0.0012	0.0018	Adjetivo
Directo	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Distinto	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Dócil	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Domiciliario	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Duro	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Económico	0	0	3	3	2	0.0000	0.0000	0.0018	0.0054	Adjetivo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Efectivo	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Emocional	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Emocionante	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Empresarial	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Enorme	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Entero	1	0	0	1	2	0.0025	0.0000	0.0000	0.0018	Adjetivo
Escolar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Especial	1	1	1	0	3	0.0021	0.0010	0.0005	0.0000	Adjetivo
Estable	0	1	2	3	3	0.0000	0.0010	0.0010	0.0046	Adjetivo
Estatal	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Estratégico	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Estresante	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Exitoso	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Exquisito	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Extenso	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Exterior	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Extraño	1	1	0	0	2	0.0025	0.0011	0.0000	0.0000	Adjetivo
Fácil	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Adjetivo
Familiar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Favorito	4	7	25	5	4	0.0076	0.0059	0.0113	0.0069	Adjetivo
Feliz	2	1	4	1	4	0.0038	0.0008	0.0018	0.0014	Adjetivo
Feo	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Adjetivo
Festivo	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Flojo	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Foráneo	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Fresco	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Frío	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Fuerte	0	1	0	1	2	0.0000	0.0011	0.0000	0.0018	Adjetivo
Futuro	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Gemelo	2	0	0	0	1	0.0061	0.0000	0.0000	0.0000	Adjetivo
Genial	1	1	0	0	2	0.0025	0.0011	0.0000	0.0000	Adjetivo
Grande	1	1	1	3	4	0.0019	0.0008	0.0005	0.0041	Adjetivo
Grupal	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Gubernamental	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Hábil	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Hermoso	0	1	1	0	2	0.0000	0.0011	0.0006	0.0000	Adjetivo
Higiénico	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Importante	0	0	2	1	2	0.0000	0.0000	0.0012	0.0018	Adjetivo
Imposible	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Adjetivo
Incondicional	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Increíble	0	3	1	0	2	0.0000	0.0033	0.0006	0.0000	Adjetivo
Independiente	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Indispensable	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Informático	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Inolvidable	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Interesante	2	0	1	0	2	0.0049	0.0000	0.0006	0.0000	Adjetivo
Internacional	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Laboral	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Largo	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Adjetivo
Lateral	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Libre	1	0	4	3	3	0.0021	0.0000	0.0020	0.0046	Adjetivo
Lindo	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Lleno	1	0	1	1	3	0.0021	0.0000	0.0005	0.0015	Adjetivo
Local	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Mal	0	1	1	0	2	0.0000	0.0011	0.0006	0.0000	Adjetivo
Malo	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Adjetivo
Maravilloso	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Marcial	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Máximo	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Mediano	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Médico	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Medio	0	1	5	0	2	0.0000	0.0011	0.0029	0.0000	Adjetivo
Menor	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Militar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Mío	0	0	2	1	2	0.0000	0.0000	0.0012	0.0018	Adjetivo
Monetario	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Moral	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Nacional	0	1	0	2	2	0.0000	0.0011	0.0000	0.0036	Adjetivo
Natal	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Navideño	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Necesario	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Adjetivo
Normal	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Obvio	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Oculto	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Olvidadizo	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Orgulloso	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Original	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Pendiente	0	1	4	0	2	0.0000	0.0011	0.0023	0.0000	Adjetivo
Pequeño	0	3	5	1	3	0.0000	0.0029	0.0025	0.0015	Adjetivo
Perfección	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Personal	1	1	1	1	4	0.0019	0.0008	0.0005	0.0014	Adjetivo
Pesado	0	1	1	0	2	0.0000	0.0011	0.0006	0.0000	Adjetivo
Popular	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Positivo	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Primario	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Primordial	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Adjetivo
Principal	3	14	25	4	4	0.0057	0.0118	0.0113	0.0055	Adjetivo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Productivo	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Adjetivo
Profesional	0	1	1	1	3	0.0000	0.0010	0.0005	0.0015	Adjetivo
Profundo	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Programable	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Adjetivo
Rápido	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Raro	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Real	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Responsable	0	2	1	0	2	0.0000	0.0022	0.0006	0.0000	Adjetivo
Revelador	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Rico	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Rock	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Romántico	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Adjetivo
Sabroso	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Sano	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Adjetivo
Satisfactorio	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Secreto	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Secundario	1	0	0	1	2	0.0025	0.0000	0.0000	0.0018	Adjetivo
Sentimental	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Simple	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Adjetivo
Simplista	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Adjetivo
Social	1	0	2	1	3	0.0021	0.0000	0.0010	0.0015	Adjetivo
Suave	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Adjetivo
Suficiente	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Súper	0	2	0	0	1	0.0000	0.0027	0.0000	0.0000	Adjetivo
Tecnológico	0	1	2	0	2	0.0000	0.0011	0.0012	0.0000	Adjetivo
Terrorista	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Adjetivo
Tímido	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Tranquilo	1	1	2	1	4	0.0019	0.0008	0.0009	0.0014	Adjetivo
Triste	0	1	2	0	2	0.0000	0.0011	0.0012	0.0000	Adjetivo
Único	0	1	1	2	3	0.0000	0.0010	0.0005	0.0031	Adjetivo
Universitario	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Útil	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Valeroso	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Adjetivo
Abrazar	0	2	0	0	1	0.0000	0.0027	0.0000	0.0000	Verbo
Abrir	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Aburrir	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Verbo
Acabar	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Verbo
Acarrear	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Acompañar	0	0	2	1	2	0.0000	0.0000	0.0012	0.0018	Verbo
Acordar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Actualizar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Actuar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Admirar	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Verbo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Adorar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Afectar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Agradar	0	1	1	0	2	0.0000	0.0011	0.0006	0.0000	Verbo
Agradecer	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Alcanzar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Alegrar	0	1	0	1	2	0.0000	0.0011	0.0000	0.0018	Verbo
Alistar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Almorzar	0	2	2	0	2	0.0000	0.0022	0.0012	0.0000	Verbo
Alzar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Amar	1	1	4	0	3	0.0021	0.0010	0.0020	0.0000	Verbo
Andar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Animar	0	1	0	1	2	0.0000	0.0011	0.0000	0.0018	Verbo
Apasionar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Apoyar	0	0	5	0	1	0.0000	0.0000	0.0036	0.0000	Verbo
Aprender	1	1	3	6	4	0.0019	0.0008	0.0014	0.0082	Verbo
Arreglar	1	3	2	0	3	0.0021	0.0029	0.0010	0.0000	Verbo
Arriesgar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Asegurar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Asignar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Asistir	1	1	3	0	3	0.0021	0.0010	0.0015	0.0000	Verbo
Aspirar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Atacar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Atender	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Verbo
Atentar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Atesorar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Atrasar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Avanzar	0	1	1	1	3	0.0000	0.0010	0.0005	0.0015	Verbo
Ayudar	1	2	15	1	4	0.0019	0.0017	0.0068	0.0014	Verbo
Bailar	0	4	2	0	2	0.0000	0.0044	0.0012	0.0000	Verbo
Bajar	0	0	6	1	2	0.0000	0.0000	0.0035	0.0018	Verbo
Bañar	1	2	4	3	4	0.0019	0.0017	0.0018	0.0041	Verbo
Basar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Besar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Brindar	0	3	1	0	2	0.0000	0.0033	0.0006	0.0000	Verbo
Buscar	0	3	4	2	3	0.0000	0.0029	0.0020	0.0031	Verbo
Caer	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Verbo
Callar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Cambiar	0	0	5	2	2	0.0000	0.0000	0.0029	0.0036	Verbo
Caminar	1	0	3	0	2	0.0025	0.0000	0.0018	0.0000	Verbo
Cansar	0	2	1	0	2	0.0000	0.0022	0.0006	0.0000	Verbo
Cantar	0	1	1	1	3	0.0000	0.0010	0.0005	0.0015	Verbo
Capturar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Caracterizar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Casar	0	1	3	0	2	0.0000	0.0011	0.0018	0.0000	Verbo
Catalogar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Cenar	0	1	4	0	2	0.0000	0.0011	0.0023	0.0000	Verbo
Centrar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Cerrar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Charlar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Clasificar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Cocinar	0	0	0	2	1	0.0000	0.0000	0.0000	0.0044	Verbo
Colaborar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Colocar	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Verbo
Combatir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Combinar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Comentar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Comenzar	0	3	5	1	3	0.0000	0.0029	0.0025	0.0015	Verbo
Comer	3	5	20	3	4	0.0057	0.0042	0.0090	0.0041	Verbo
Compartir	1	2	1	0	3	0.0021	0.0019	0.0005	0.0000	Verbo
Compensar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Complicar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Componer	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Comprar	3	6	5	3	4	0.0057	0.0051	0.0023	0.0041	Verbo
Concluir	0	4	5	1	3	0.0000	0.0038	0.0025	0.0015	Verbo
Concursar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Conducir	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Conectar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Confiar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Conseguir	1	2	7	1	4	0.0019	0.0017	0.0032	0.0014	Verbo
Consentir	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Considerar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Consistir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Construir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Contar	1	4	1	2	4	0.0019	0.0034	0.0005	0.0027	Verbo
Contentar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Continuar	0	2	4	0	2	0.0000	0.0022	0.0023	0.0000	Verbo
Controlar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Conversar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Convertir	0	0	4	0	1	0.0000	0.0000	0.0029	0.0000	Verbo
Convivir	1	3	3	2	4	0.0019	0.0025	0.0014	0.0027	Verbo
Correr	1	2	2	0	3	0.0021	0.0019	0.0010	0.0000	Verbo
Corresponder	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Cortar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Costar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Crear	0	1	0	1	2	0.0000	0.0011	0.0000	0.0018	Verbo
Crecer	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Verbo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Creer	1	2	0	1	3	0.0021	0.0019	0.0000	0.0015	Verbo
Cuidar	0	2	4	1	3	0.0000	0.0019	0.0020	0.0015	Verbo
Cumplir	2	1	2	1	4	0.0038	0.0008	0.0009	0.0014	Verbo
Curar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Cursar	0	0	3	1	2	0.0000	0.0000	0.0018	0.0018	Verbo
Deber	0	1	0	1	2	0.0000	0.0011	0.0000	0.0018	Verbo
Decidir	1	2	1	0	3	0.0021	0.0019	0.0005	0.0000	Verbo
Declamar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Decorar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Dedicar	1	0	1	1	3	0.0021	0.0000	0.0005	0.0015	Verbo
Defender	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Verbo
Dejar	0	3	7	1	3	0.0000	0.0029	0.0036	0.0015	Verbo
Demostrar	1	0	0	1	2	0.0025	0.0000	0.0000	0.0018	Verbo
Denominar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Depender	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Derivar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Desaburrir	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Desarrollar	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Verbo
Descansar	0	4	2	0	2	0.0000	0.0044	0.0012	0.0000	Verbo
Descender	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Descombrar	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Verbo
Describir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Descubrir	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Descuidar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Desear	1	0	5	1	3	0.0021	0.0000	0.0025	0.0015	Verbo
Desempeñar	0	1	1	0	2	0.0000	0.0011	0.0006	0.0000	Verbo
Desenvolver	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Desestresarse	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Despedir	0	2	1	0	2	0.0000	0.0022	0.0006	0.0000	Verbo
Despertar	0	4	6	1	3	0.0000	0.0038	0.0030	0.0015	Verbo
Destruir	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Verbo
Determinar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Dirigir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Disfrutar	0	4	2	0	2	0.0000	0.0044	0.0012	0.0000	Verbo
Disponer	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Distraer	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Divertir	0	4	2	3	3	0.0000	0.0038	0.0010	0.0046	Verbo
Dormir	3	4	12	4	4	0.0057	0.0034	0.0054	0.0055	Verbo
Echar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Ejercer	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Elaborar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Elegir	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Verbo
Emocionar	0	2	0	0	1	0.0000	0.0027	0.0000	0.0000	Verbo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Empezar	0	2	1	4	3	0.0000	0.0019	0.0005	0.0062	Verbo
Emplear	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Enamorar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Encajar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Encantar	2	1	4	1	4	0.0038	0.0008	0.0018	0.0014	Verbo
Encargar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Encender	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Verbo
Encontrar	0	5	8	0	2	0.0000	0.0055	0.0047	0.0000	Verbo
Endurar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Enfrentar	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Verbo
Engargolar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Enojar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Enorgullecer	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Enrollar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Ensayar	0	1	1	1	3	0.0000	0.0010	0.0005	0.0015	Verbo
Enseñar	0	2	0	1	2	0.0000	0.0022	0.0000	0.0018	Verbo
Entender	1	0	3	1	3	0.0021	0.0000	0.0015	0.0015	Verbo
Entrar	0	2	3	0	2	0.0000	0.0022	0.0018	0.0000	Verbo
Entregar	0	0	5	0	1	0.0000	0.0000	0.0036	0.0000	Verbo
Entrenar	1	1	0	0	2	0.0025	0.0011	0.0000	0.0000	Verbo
Entretener	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Verbo
Entrometer	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Escoger	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Esconder	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Escribir	2	1	3	4	4	0.0038	0.0008	0.0014	0.0055	Verbo
Escuchar	2	2	14	4	4	0.0038	0.0017	0.0063	0.0055	Verbo
Esperar	0	2	2	1	3	0.0000	0.0019	0.0010	0.0015	Verbo
Estrenar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Estructurar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Estudiar	1	4	10	2	4	0.0019	0.0034	0.0045	0.0027	Verbo
Existir	0	2	4	1	3	0.0000	0.0019	0.0020	0.0015	Verbo
Expresar	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Verbo
Extrañar	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Verbo
Fallecer	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Faltar	0	0	2	1	2	0.0000	0.0000	0.0012	0.0018	Verbo
Fascinar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Festejar	0	1	0	1	2	0.0000	0.0011	0.0000	0.0018	Verbo
Fijar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Finalizar	1	1	0	0	2	0.0025	0.0011	0.0000	0.0000	Verbo
Firmar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Fortalecer	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Funcionar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Ganar	0	2	4	2	3	0.0000	0.0019	0.0020	0.0031	Verbo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Generar	1	0	2	0	2	0.0025	0.0000	0.0012	0.0000	Verbo
Gobernar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Guardar	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Verbo
Guiar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Gustar	4	15	24	9	4	0.0076	0.0127	0.0108	0.0124	Verbo
Hablar	1	1	2	3	4	0.0019	0.0008	0.0009	0.0041	Verbo
Hallar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Hibridar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Hundir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Implicar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Importar	0	2	0	1	2	0.0000	0.0022	0.0000	0.0018	Verbo
Imprimir	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Incorporar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Identificar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Infectar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Influir	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Ingresar	0	0	0	2	1	0.0000	0.0000	0.0000	0.0044	Verbo
Inspirar	1	0	0	1	2	0.0025	0.0000	0.0000	0.0018	Verbo
Intentar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Interactuar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Invadir	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Invitar	1	1	0	2	3	0.0021	0.0010	0.0000	0.0031	Verbo
Jubilar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Lastimar	0	1	0	1	2	0.0000	0.0011	0.0000	0.0018	Verbo
Lavar	0	2	2	0	2	0.0000	0.0022	0.0012	0.0000	Verbo
Leer	0	1	5	0	2	0.0000	0.0011	0.0029	0.0000	Verbo
Levantar	1	2	9	1	4	0.0019	0.0017	0.0041	0.0014	Verbo
Liberar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Librar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Llamar	1	2	1	1	4	0.0019	0.0017	0.0005	0.0014	Verbo
Llegar	4	12	16	8	4	0.0076	0.0102	0.0072	0.0110	Verbo
Lograr	1	3	2	3	4	0.0019	0.0025	0.0009	0.0041	Verbo
Luchar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Mandar	0	1	1	0	2	0.0000	0.0011	0.0006	0.0000	Verbo
Mantener	1	2	4	0	3	0.0021	0.0019	0.0020	0.0000	Verbo
Marcar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Medir	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Mejorar	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Verbo
Merecer	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Verbo
Merendar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Meter	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Verbo
Mirar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Mojar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Montar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Morir	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Verbo
Navegar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Necesitar	1	2	3	0	3	0.0021	0.0019	0.0015	0.0000	Verbo
Notar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Nutrir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Obligar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Observar	0	1	2	0	2	0.0000	0.0011	0.0012	0.0000	Verbo
Obtener	0	2	1	2	3	0.0000	0.0019	0.0005	0.0031	Verbo
Ocupar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Ocurrir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Ofender	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Ofrecer	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Oír	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Ojala	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Oler	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Olvidar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Optar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Ordenar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Otorgar	0	1	1	0	2	0.0000	0.0011	0.0006	0.0000	Verbo
Padecer	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Pagar	0	1	2	1	3	0.0000	0.0010	0.0010	0.0015	Verbo
Parar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Parecer	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Participar	0	1	0	1	2	0.0000	0.0011	0.0000	0.0018	Verbo
Pasar	7	8	22	5	4	0.0133	0.0068	0.0099	0.0069	Verbo
Pasear	1	1	2	0	3	0.0021	0.0010	0.0010	0.0000	Verbo
Pedir	0	0	4	0	1	0.0000	0.0000	0.0029	0.0000	Verbo
Pelear	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Pensar	1	4	6	1	4	0.0019	0.0034	0.0027	0.0014	Verbo
Perder	0	0	3	1	2	0.0000	0.0000	0.0018	0.0018	Verbo
Permitir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Pescar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Pintar	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Verbo
Planchar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Planear	1	0	3	0	2	0.0025	0.0000	0.0018	0.0000	Verbo
Platicar	4	1	7	1	4	0.0076	0.0008	0.0032	0.0014	Verbo
Poder	5	23	38	7	4	0.0095	0.0195	0.0171	0.0096	Verbo
Practicar	0	5	4	2	3	0.0000	0.0048	0.0020	0.0031	Verbo
Preferir	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Preocupar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Preparar	1	0	0	1	2	0.0025	0.0000	0.0000	0.0018	Verbo
Presentar	1	0	1	0	2	0.0025	0.0000	0.0006	0.0000	Verbo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Prestar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Prestigiar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Pretender	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Probar	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Producir	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Programar	0	0	0	4	1	0.0000	0.0000	0.0000	0.0088	Verbo
Progresar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Proponer	0	1	1	0	2	0.0000	0.0011	0.0006	0.0000	Verbo
Publicar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Quedar	0	0	7	2	2	0.0000	0.0000	0.0041	0.0036	Verbo
Querer	1	2	16	2	4	0.0019	0.0017	0.0072	0.0027	Verbo
Recapacitar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Recibir	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Reciclar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Reconocer	0	1	1	2	3	0.0000	0.0010	0.0005	0.0031	Verbo
Recordar	0	1	2	0	2	0.0000	0.0011	0.0012	0.0000	Verbo
Recorrer	0	1	2	0	2	0.0000	0.0011	0.0012	0.0000	Verbo
Redactar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Reflexionar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Regresar	2	3	6	2	4	0.0038	0.0025	0.0027	0.0027	Verbo
Reír	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Relacionar	0	0	1	1	2	0.0000	0.0000	0.0006	0.0018	Verbo
Relajar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Relatar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Remunerar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Rendir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Reprimir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Resolver	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Resultar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Retirar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Reunir	0	1	2	0	2	0.0000	0.0011	0.0012	0.0000	Verbo
Revisar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Rodear	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Saber	1	2	3	9	4	0.0019	0.0017	0.0014	0.0124	Verbo
Sacar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Salir	4	21	29	4	4	0.0076	0.0178	0.0131	0.0055	Verbo
Saltar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Saludar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Satisfacer	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Seguir	1	1	10	3	4	0.0019	0.0008	0.0045	0.0041	Verbo
Sentar	3	3	3	1	4	0.0057	0.0025	0.0014	0.0014	Verbo
Sentir	3	3	3	1	4	0.0057	0.0025	0.0014	0.0014	Verbo
Servir	0	1	1	0	2	0.0000	0.0011	0.0006	0.0000	Verbo

Palabra (Lema)	Suma de Apariciones en el texto (frecuencia) DOMINANTES	Suma de Apariciones en el texto (frecuencia) INFLUYENTES	Suma de Apariciones en el texto (frecuencia) ESTABLES	Suma de Apariciones en el texto (frecuencia) CONCIENZUDO	Numero de doc. en los que aparece	Pesos en cada Factor				Tipo
						Dominante	Influyente	Estable	Concienzudo	
						D	I	S	C	
Sobrevivir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Solar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Soler	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Solucionar	0	0	2	0	1	0.0000	0.0000	0.0014	0.0000	Verbo
Solventar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Sonar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Sonreír	2	0	0	0	1	0.0061	0.0000	0.0000	0.0000	Verbo
Sostener	1	0	0	1	2	0.0025	0.0000	0.0000	0.0018	Verbo
Subir	0	1	2	2	3	0.0000	0.0010	0.0010	0.0031	Verbo
Suceder	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Sufrir	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Superar	1	1	0	1	3	0.0021	0.0010	0.0000	0.0015	Verbo
Surgir	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Suspender	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Sustentar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Tardar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Titular	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Tocar	0	1	7	2	3	0.0000	0.0010	0.0036	0.0031	Verbo
Tomar	2	3	6	0	3	0.0043	0.0029	0.0030	0.0000	Verbo
Trabajar	0	7	10	0	2	0.0000	0.0077	0.0059	0.0000	Verbo
Traer	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Tramitar	0	0	3	0	1	0.0000	0.0000	0.0022	0.0000	Verbo
Trascender	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Trasladar	0	2	1	0	2	0.0000	0.0022	0.0006	0.0000	Verbo
Tratar	0	1	3	3	3	0.0000	0.0010	0.0015	0.0046	Verbo
Unir	1	0	0	0	1	0.0030	0.0000	0.0000	0.0000	Verbo
Usar	0	1	0	0	1	0.0000	0.0014	0.0000	0.0000	Verbo
Utilizar	0	0	1	0	1	0.0000	0.0000	0.0007	0.0000	Verbo
Variar	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Vender	0	0	0	1	1	0.0000	0.0000	0.0000	0.0022	Verbo
Venir	0	1	4	0	2	0.0000	0.0011	0.0023	0.0000	Verbo
Viajar	1	6	5	5	4	0.0019	0.0051	0.0023	0.0069	Verbo
Visitar	1	1	1	3	4	0.0019	0.0008	0.0005	0.0041	Verbo
Vivir	3	2	8	2	4	0.0057	0.0017	0.0036	0.0027	Verbo
Volver	1	1	1	0	3	0.0021	0.0010	0.0005	0.0000	Verbo