UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Reconocimiento de la personalidad en Twitter

TRABAJO FIN DE MÁSTER

Máster en Big Data Analytics

Autor: Pilar Sáez Hernández

Tutor: Francisco Rangel

Curso 2017-2018

Resumen

????

Palabras clave: Aprendizaje Automático, Redes neuronales, vectores de soporte

Abstract

????

Key words: Machine Learning, Artificial Neural Networks, Support Vector Machine

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

1.1 Motivación

En la actualidad la aplicación de la tecnología en la mayoría de los aspectos de la sociedad humana ha generado un cambio profundo en el comportamiento de organizaciones y personas, afectado a los procesos organizativos, políticos, las estructuras socioeconómicas y comunicaciones interpersonales. Todo está en constante cambio, y la adaptación y el aprovechamiento de este es lo que se denomina *Transformación Digital*.

La *Transformación Digital* genera, para las organizaciones, un mundo de oportunidades y retos así como multitud de vulnerabilidades. Estas organizaciones y las personas que las dirigen se esfuerzan por comprender su entorno para desarrollar estrategias que mejoren su desempeño organizacional y mantengan su ventaja sobre la competencia. La creciente complejidad, el ritmo y la multitud de desafíos y oportunidades a los que se enfrentan las organizaciones crean necesidades cada vez mayores de innovación, automatización y la capacidad de obtener valor a partir de los datos con el fin de mantenerse ágiles y flexibles para poder asumir riesgos en el mundo cambiante en el que vivimos.

Autoritas Consulting es una de las organizaciones que diseña estrategias para aprovechar las ventajas que propone este nuevo paradigma de comunicación.
Su actividad se centra en el sector de la consultoría, mediante la generación de inteligencia a partir de datos sociales provenientes de fuentes abiertas. El objetivo es conocer el entorno en el cual opera cada cliente (su reputación, su competencia, la identificación de medios influyentes, la gestión de crisis,...) y así ayudar en la toma de decisiones con conocimiento de causa.

Una de las herramientas desarrolladas por *Autoritas Consulting* enmarcadas en este ámbito se denomina Cosmos. Cosmos es el conjunto de módulos de software para la captura, explotación y análisis de información en Internet. Son herramientas integradas en la lógica de escucha inteligente y diseñadas específicamente para dar soporte a cada una de las fases del ciclo de inteligencia, lo que finalmente permitirá extraer conocimiento y resultados que serán determinantes en la toma de decisiones.

En la figura (**??**) se puede ver el ciclo completo de la herramienta. Desde el momento de la planificación pasando por el análisis de la información hasta que se propone la acción o acciones a llevar a cabo.

La aportación de este trabajo profundiza en la técnica conocida como *Author Profiling*, que es un campo de investigación transversal a diferentes disciplinas como la lingüística (computacional), procesamiento del lenguaje natural, aprendizaje automático, recuperación de información, neurología, marketing,etc y que básicamente trata de averiguar la

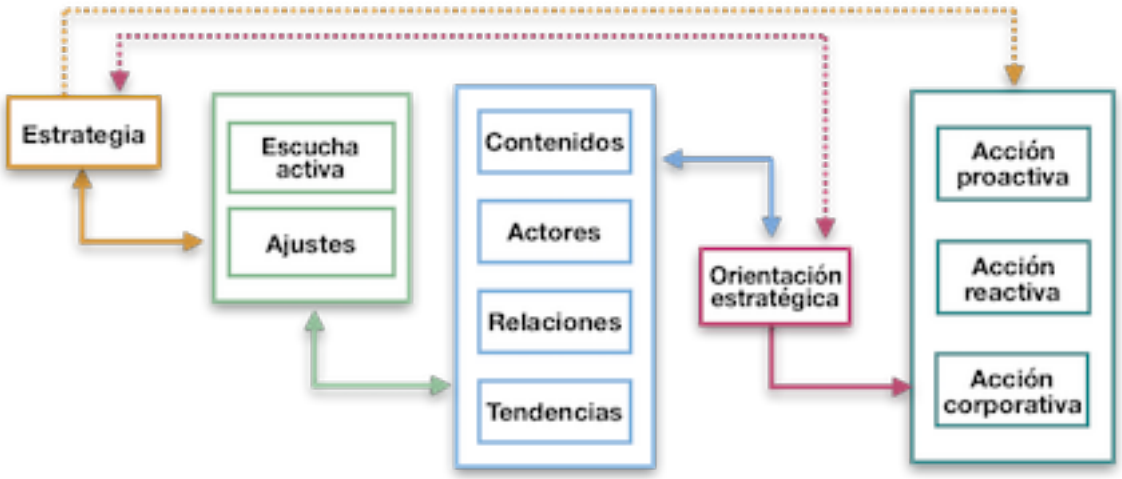


Figura 1.1: Ciclo completo Cosmos

máxima información personal posible de un autor o usuario a partir de lo que anónimamente escribe: edad, género, idioma nativo, perfil emocional, rasgos de personalidad.

1.2 Objetivos

El objetivo principal es desarrollar un ciclo completo de predicción para cada una de las características que definen la clasificación de personalidad *big five* (clasificación que se defnirá más adelante) y su integración en la herramienta *Cosmos*.

Como punto de partida del proyecto, se dispone de un conjunto de textos obtenidos de la red social *Twitter* cuyos usuarios han sido sometidos a un test de personalidad. Por lo que se dispone también de la clasificación necesaria para crear un modelo de predicción mediante técnicas de aprendizaje supervisado.

Si nos fijamos en el esquema de *Cosmos* definido en la figura **??**, este nuevo módulo para el reconocimiento de la personalidad se incluiría en la fase de análisis. Por lo tanto el módulo ha de estar preparado para la escucha activa de información proveniente de Twitter y ser capaz de analizar la misma en un tiempo prudencial, así como la muestra de resultados para que posteriormente se determinen las acciones necesarias.

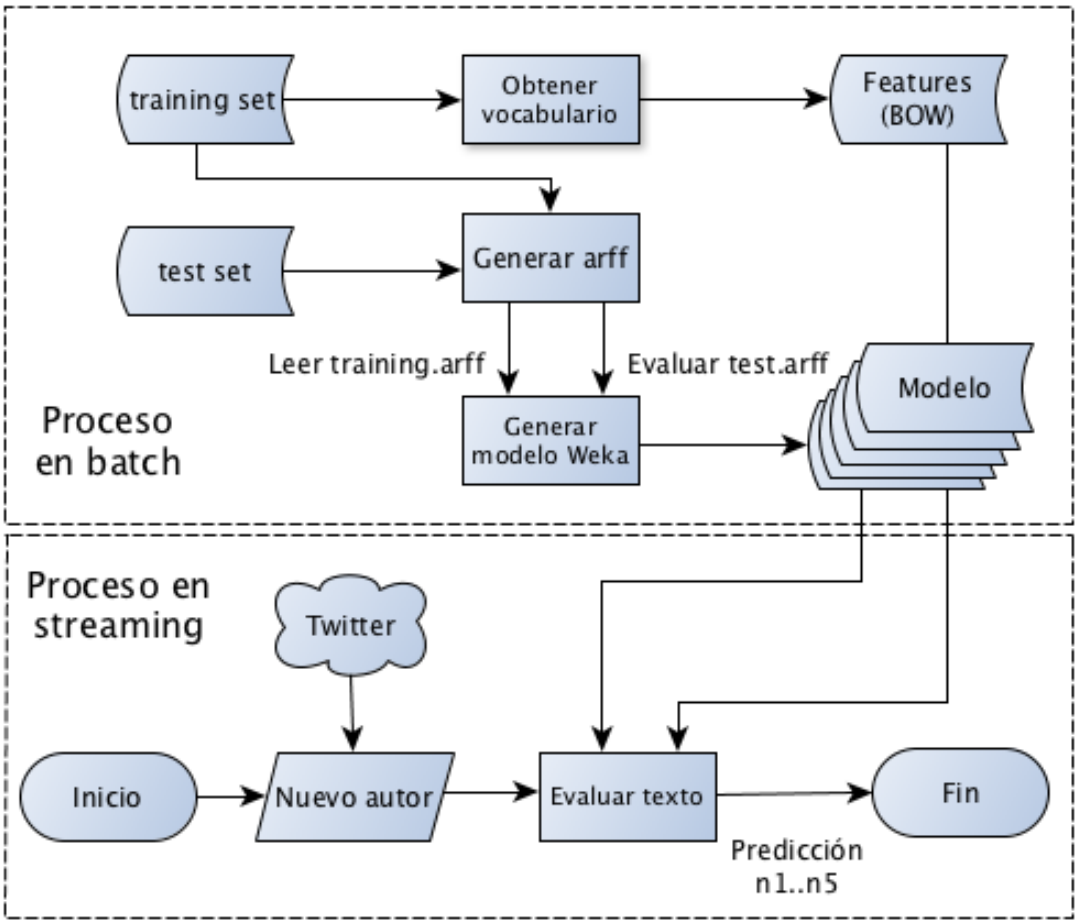


Figura 1.2: Diagrama de la implementación

En la figura **??** representa el diagrama con el ciclo completo desarrollado en este proyecto. Se puede dividir el desarrollo en dos grandes secciones, por un lado el proceso que compone la extracciónn, transformación y carga (ETL) y la generación de modelos y por otra la implementación del módulo que analiza y predice los rasgos de personalidad dado un texto de entrada.

1.3 Estructura de la memória

- Estado de la cuestión y fundamentación teórica
- Análisis de los datos y Metodología propuesta
- Explotación y visualización
- Conclusiones

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 2
Fundamentación teórica y Estado de la cuestión

En este capítulo se presenta un resumen sobre los estudios previos realizados o que podrían estar relacionados con el actual, así como la definición de los conceptos teóricos que sirven de fundamentación para el trabajo presentado.

2.1 Aprendizaje automático

Aprendizaje automático o *Machine Learning* es un subcampo dentro de Computer Science muy relacionado con la Inteligencia Artificial y el reconocimiento de patrones[WIKML]. También está muy relacionado con técnicas estadísticas para el cálculo de modelos de predicción y la optimización matemática.
Es la disciplina que explora la construcción y el estudio de algoritmos que aprenden a partir de los datos y permite hacer predicciones sobre los mismos.
Los algoritmos se utilizan para crear/estimar modelos a partir de muestras (de aprendizaje). Y estos modelos se utilizan para realizar predicciones o tomar decisiones.

2.1.1. Tipos de aprendizaje

Las tareas o problemas a los que pueden aplicarse las distintas técnicas de aprendizaje automático son:

- **Aprendizaje supervisado o *Supervised learning*** Se dispone del valor de la variable salida (etiqueta o valor).
- **Aprendizaje no supervisado o *Unsupervised learning*** No se dispone de la variable salida. Las muestras no están etiquetadas como pertenecientes a ninguna clase o no van acompañadas de un valor a predecir.
- **Aprendizaje por refuerzo o *Reinforcement learning*** Aquí el aprendizaje consiste en encontrar las acciones a realizar según la situación con el objetivo de maximizar una recompensa. El algoritmo no dispone de la variable de salida, debe descubrirla mediante un proceso de prueba y error. Existe una serie de estados posibles y unas acciones a realizar, el algoritmo interactúa con el entorno explorando posibilidades (distintas acciones) y rectifica o consolida sus reglas internas según la recompensa obtenida. La recompensa no siempre se recibe después de cada acción. Muchas veces viene tras una serie de acciones.

2.1.2. Aprendizaje supervisado

Los modelos se estiman aprendiendo reglas que permiten obtener la variable salida a partir de la(s) variable(s) de entrada. El aprendizaje supervisado se a aplica dos tareas:

- **Clasificación** Las muestras pertenecen a dos o más clases. Se utilizan muestras etiquetadas para el aprendizaje. El modelo se utiliza después para etiquetar muestras de las cuales no se conoce a qué clase pertenecen. Un buen ejemplo es la clasificación de dígitos manuscritos.

- **Regresión** La variable salida es uno o más valores reales (continuos). Un buen ejemplo es la predicción de la temperatura a diferentes horas del día.

2.1.3. Reconocimiento de patrones

El reconocimiento de patrones es la rama dentro del aprendizaje automático dedicada al desarrollo de algoritmos para descubrir, de manera automática, patrones de regularidad en los datos. Y aprovechar las regularidades detectadas para tareas como clasificación de los datos en diferentes categorías o clases [4].

2.1.4. Ejemplo de problema de clasificación mediante aprendizaje supervisado y reconocimiento de patrones

El objetivo en la clasificación de dígitos manuscritos es, dada una imagen, decir qué dígito es. La variable salida es un valor entero del 0 al 9.
El problema, nada trivial, es entrenar un sistema capaz de determinar a qué dígito corresponde una imagen ya preprocesada, es decir, normalizada y con el dígito centrado pero sin etiquetar.
Para entrenar el sistema necesitaremos de un conjunto de muestras etiquetadas (training set), es decir, un conjunto de imágenes (muestras d-dimensionales) y su correspondiente variable salida, un entero del 0 al 9 en este caso.

(a) Dígitos manuscritos

$X = \{x_1, x_2, ..., x_N\}$ Muestras con las variables de entrada: $x_i \in \mathbb{R}^d$
$t = \{t_1, t_2, ..., t_N\}$ *Target vector* $t_i \in \{0, 1, 2, ..., 9\}$
N es el tamaño del *training set*

El resultado de entrenar/aprender un algoritmo/modelo será disponer de una función que dada una muestra de entrada obtenga el valor ?correcto? de la variable de salida.

$$\hat{t} = y(\hat{x})$$

2.1.5. Posibles problemas del proceso de aprendizaje

- ***Over-fitting*** El objetivo de todo algoritmo de aprendizaje es la generalización, es decir, la capacidad de clasificar correctamente muestras que no se utilizaron durante el proceso de aprendizaje. El problema de *over-fitting* o sobre aprendizaje aparece cuando el algoritmo ha aprendido a predecir la clase o el valor de salida correspondiente a las muestras de aprendizaje de manera demasiado precisa, el error de clasificación o regresión para el training set es muy bajo, pero muy alto para las muestras del test set.

- **High-dimensionality** Cuando las muestras de aprendizaje son vectores de muchas dimensiones aparecen dos problemas:
  que los cálculos a realizar por muchos algoritmos de aprendizaje son caros en tiempo de cómputo y que muchas técnicas ven reducida su capacidad de aprendizaje.

  En muchos casos es necesario realizar un preproceso a los datos (feature extraction) con objeto de sólo trabajar con las variables de entrada que el experto humano considera servirán mejor para la tarea de clasificación, regresión, identificación o predicción.

2.2 Evaluación del error

2.2.1. Función de error a minimizar

La función de error debe ser siempre no negativa. Será igual a cero si el valor a predecir $\hat{f}_i$ coincide con el valor objetivo f_i.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{f}_i - f_i)^2}{n}}$$

Esta función es la raíz cuadrada de la suma de los cuadrados de la diferencia entre la estimación y la función a predecir dividida entre el total de muestras. Se conoce como *root-mean-square error*.

El problema consiste encontrar el valor de$\hat{f}_i$ para el cual el ***RMSE*** es lo más pequeño posible.

2.3 Reconocimiento de la personalidad utilizando la Teoría de los Cinco Grandes

La personalidad puede definirse por medio de cinco rasgos utilizando la Teoría de los Cinco Grandes [**??**] o *Big Five*, que es la más aceptada en psicología. Los cinco rasgos incluidos en esta teoría son:

- apertura a la experiencia (O) - Innovador vs conservador

- escrupolosidad (C) - Concienzudo vs descuidado
- extroversión (E) - Extrovertido vs tímido
- amabilidad (A) - Simpático vs serio
- estabilidad (S) - Estable emocionalmente vs neurótico

Dado un texto escrito de forma anónima se obtendrá una medida predictiva para cada uno de estos rasgos que definirá al autor en la proporción más ajustada posible.

2.4 Estado de la cuestión

El reconocimiento automático de la personalidad a partir del texto ha sido abordado por trabajos pioneros desde hace unos 10 años. Argamon [**??**] se centró en dos de los rasgos de los Cinco Grandes (Extroversión y Estabilidad Emocional), medidos por medio de autoinformes. En su estudio se usó el algoritmo de Support Vector Machines, complementado con categorías de palabras y frecuencia relativa de palabras funcionales, para reconocer estos dos rasgos.

De forma similar, Oberlander y Nowson [**??**] trabajaron en la clasificación de los tipos de personalidad de los bloggers mediante la extracción de patrones de forma ascendente.

Mairesse en [**??**], investigaron sistemáticamente la utilidad de diferentes conjuntos de características textuales que explotan los diccionarios psicolingüísticos (LIWC4 y MRC5). Extrajeron modelos de personalidad de autoinformes y datos observados, y obtuvieron que el rasgo de la apertura a la experiencia producía el mejor rendimiento.

En años más recientes, el interés de los investigadores se ha enfocado más en la predicción de la personalidad usando corpus de datos de redes sociales, como Twitter y Facebook. Explotando características lingüísticas en actualizaciones de estado, características sociales como conteo de amigos y actividad diaria [**??**, **??**, **??**].
Kosinski [**??**] realizó un análisis exhaustivo de las diferentes características, incluido el tamaño de la red de amistad, el recuento de fotos cargadas y los eventos asistidos, y encontró las correlaciones con los rasgos de personalidad de 180000 usuarios de Facebook. Obtuvieron muy buenos resultados en la predicción automática de Extroversión. Bachrach et al. realizó un análisis exhaustivo de los rasgos de la red (tamaño de la red de amistad, fotos cargadas, eventos atendidos, frecuencia con la que un usuario ha sido etiquetado en las fotos)

CAPÍTULO 3
Análisis y Metodología Propuesta

En el presente capítulo se analiza el problema planteado, se realiza un estudio de la información disponible y las posibles formas de abordarla para la consecución de los objetivos planteados.

3.1 Obtención de los datos

Los datos fueron obtenidos de Twitter por medio de una campaña publicitaria. El género y la edad fueron especificados por el usuario mientras que los rasgos de la personalidad fueron calculados mediante el test BFI-10 (Rammstedt & John, 2007) y normalizados en valores comprendidos entre -0.5 y +0.5. Asumimos que estos resultados definen con precisión los rasgos de personalidad.

Durante la campaña PAN CLEF 2015, son generadas estructuras XML para cada respuesta obtenida. La respuesta de género es codificada con un carácter binario y la edad se agrupa en cuatro bloques. Los rasgos de personalidad *Big Five* se codifican con valores comprendidos entre -0.5 y +0.5.

Para valorar la bondad del modelo de predicción se usará como medida el Root Mean Square Error (RMSE), cuya fórmula se definió en la sección **??**

3.2 Análisis del dataset

Se dispone de un dataset con información clasificada de 190 usuarios de Twitter. Para cada usuario tenemos, por un lado un XML con sus tuits y un archivo *es.txt* con su clasificación *Big Five*.

- Muestra-descripción del XML y txt

3.3 Metodología Propuesta

En este apartado se presenta la metodología seguida para la consecución de los objetivos planteados. El proceso de generación de modelos junto con las técnicas usadas en la predicción y su integración en la herramienta *social census*.

3.4 Generación de *Features*

Se llama *features* a las variables o características que se seleccionan para la generación de un modelo predictivo. Para obtener un buen modelo, es importante que las *features* contengan la máxima información relevante sobre la variable a predecir.

Este es un proceso de investigación en el que se pueden emplear multitud de técnicas matemáticas, estadísticas, etc y que requiere de un tiempo considerable para una reducción pequeña del *RMSE*. Por este motivo y dado que el proyecto se centra más en el proceso completo de predicción, se opta por un modelo sencillo de bolsa de palabras.

La bolsa de palabras se construye procesando la totalidad de los documentos contenidos en el conjunto de datos de training, creando una lista ordenada de los términos utilizados de más frecuente a menos frecuente y seleccionando los mil términos que más se repiten.

Además se contará la frecuencia de uso de los símbolos de puntuación: punto(**.**), coma(**,**) y dos puntos (**:**).

Por lo tanto el archivo resultante, *bow.txt*, será nuestro archivo de *features* y contendrá el listado de los mil términos y los tres símbolos de puntuación.

3.5 Generación de modelos en *Weka*

Weka es una plataforma de software para el aprendizaje automático y la minería de datos. Contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades.
Entre las cuales se encuentra la aplicación de multitud de técnicas de aprendizaje automático y la posibilidad de generar y almacenar los modelos de predicción resultantes de estas aplicaciones.

Weka posee el formato de entrada de datos definido como *Attribute-Relation File Format (ARFF)*. Es un archivo de texto que relaciona una lista de atributos con una lista de instancias y que nos servirá como punto de partida para la generación de la línea base en modelos de predicción.

Se genera un archivo de entrada para cada uno de los cinco rasgos de personalidad, donde la diferencia entre ellos, estriba en la variable a predecir.
Para construir estos archivos nos basaremos en la bolsa de palabras definida anteriormente y en los datos de training. El último atributo de cada archivo de entrada será el rasgo a predecir.

Tendremos una instancia por cada usuario de *Twitter*. Para dar un valor a cada atributo de la instancia contamos la frecuencia con la que el usuario usa cada palabra contenida en la bolsa y obtendremos un valor normalizado. Finalmente para completar el valor de la clase, lo obtenemos del training. Y haremos lo mismo para cada rasgo de personalidad que queremos predecir.

Una vez que tenemos los datos cargados en Weka probaremos las diferentes técnicas que nos ofrece el programa y seleccionaremos los modelos para los cuales se obtengan los mejores resultados.

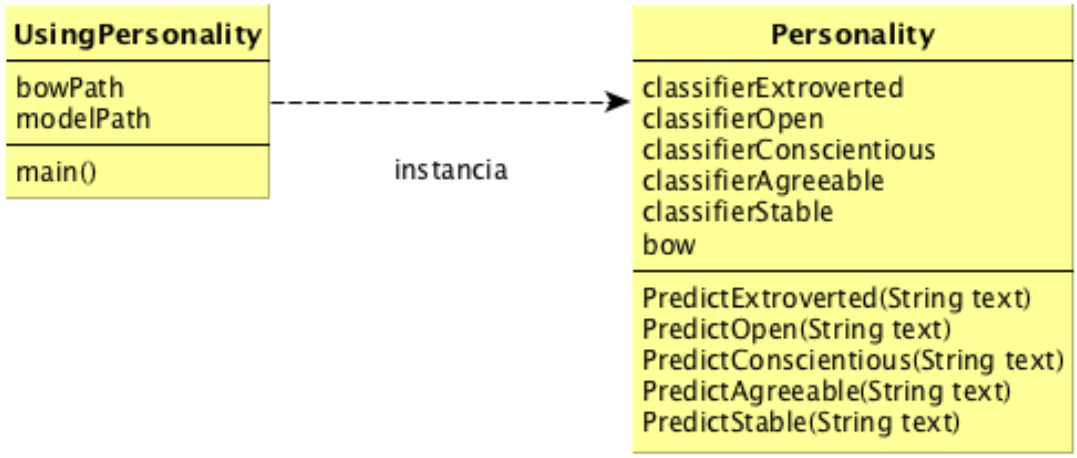
Para la evaluación de los modelos, construimos un fichero de test de la misma forma que para el training y obtenemos las siguientes medidas:

	Spanish		
	Bagging	RandomForest	RandomCommitte
Extroverted	0.1712	0.1749	0.1754
Open	0.1476	0.134	0.1343
Conscientious	0.1499	0.1457	0.1394
Agreeable	0.1694	0.1611	0.1662
Stable	0.2092	0.2008	0.1981

En la tabla **??** se muestran los resultados obtenidos correspondientes a los valores de *RMSE*. Hay que tener en cuenta que el problema que se plantea es de regresión, por lo que los algoritmos usados han de ser compatibles.
Por lo tanto construiremos el modelo para la predicción del rasgo de personalidad Extroverted basándonos en la técnica de Bagging. Random Forest para Open y Agreeable y Random Comitte para Conscientious y Stable porque son los que presentan el menor error.

3.6 Modulo de predicción en tiempo real

Una vez generados los modelos y la bolsa de palabras se implementará el módulo de predicción en tiempo real. El funcionamiento básico consiste en, dado un texto de entrada obtener cinco datos de salida, que corresponderán a una medida de predicción para cada rasgo de personalidad del autor.



(a) Diagrama UML

3.7 Integración

- Construcción de una librería

public double PredictOpen(String text) public double PredictConscientious(String text) public double PredictAgreeable(String text) public double PredictStable(String text)

3.8 Pruebas de carga

3.9 Mejora de resultados

CAPÍTULO 4

Casos de Uso

CAPÍTULO 5
Conclusions

????? ????????????? ????????????? ????????????? ????????????? ?????????????

Bibliografía

[1] Costa, P.T., McCrae, R.R. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment 2*, 179-198 (2008)

[2] Sushant, S.A., Argamon, S., Dhawle, S., Pennebaker, J.W. Lexical predictors of personality type *In Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, (2005)

[3] Oberlander, J., Nowson, S. Whose thumb is it anyway?: classifying author personality from weblog text. *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 627-634. Association for Computational Linguistics (2006)

[4] Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research 30(1)*, 457-5

[5] Golbeck, J., Robles, C., Turner, K. Predicting personality with social media. *CHI?11 Extended Abstracts on Human Factors in Computing Systems.*, pp. 253-262. ACM (2011)

[6] Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., Crowcroft, J. The personality of popular facebook users. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.*, p. 955-964. ACM (2012)

[7] Celli, F., Polonio, L. Relationships between personality and interactions in facebook *Social Networking: Recent Trends, Emerging Issues and Future Outlook*, pp. 41-54. Nova Science Publishers, Inc (2013)

[8] Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., Graepel, T. Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, pp. 1-24 (2013)

[9] Georges Ifrah. *Historia universal de las cifras*. Espasa Calpe, S.A., Madrid, sisena edició, 2008.

[10] Comunicat de premsa del Departament de la Guerra, emés el 16 de febrer de 1946. Consultat a http://americanhistory.si.edu/comphist/pr1.pdf.

APÉNDICE A
Configuració del sistema

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.1 Fase d'inicialització

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.2 Identificació de dispositius

????? ????????????? ????????????? ????????????? ????????????? ?????????????

APÉNDICE B
??? ????????????? ????

????? ????????????? ????????????? ????????????? ????????????? ?????????????