Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Estimación automática del afecto en textos en español

Tesis presentada para optar al título de
Licenciado en Ciencias de la Computación

Matías Gabriel Dell' Amerlina Ríos

Director: Agustín Gravano

RESUMEN

El análisis de sentimientos y afectos en textos (*sentiment analysis*) ha sido estudiado en el idioma inglés durante los últimos 30 años. Uno de los trabajos pioneros, y que sigue evolucionando, fue el de Whissell con su Diccionario de Afectos (*Dictionary of Affect in Language*, DAL), que permite puntuar palabras de acuerdo a tres dimensiones: *agrado, activación e imaginabilidad*. Dado que en el idioma español son escasas las herramientas disponibles para hacer estos análisis, en esta tesis se realizó una réplica en el idioma español del trabajo de Whissell. Este trabajo comenta de qué manera fue creado el sistema, que contó con más de **2500** palabras, donde cada vocablo fue puntuado en las mismas tres dimensiones. Además describe cómo el sistema realiza las estimaciones y qué índices arroja como resultado. Finalmente, se evaluó el sistema comparando sus resultados con los de un grupo de personas y, por otro lado, se evaluó la utilidad del sistema para clasificar opiniones de productos.

Palabras claves: Procesamiento del Lenguaje Natural, Afecto, Emociones, español.

ABSTRACT

The topic of sentiment and affections analysis in text has been studied in English for the last 30 years. One of the first works, and still useful, is the one made by Whissell, the Dictionary of Affect in Language (DAL), that allows to rate words along three dimensions: pleasantness, activation and imagery. Since there is a lack of such tools for the Spanish language, we decided to replicate Whissell's work in Spanish. This thesis describes the development of a system with more than 2500 words rated along the same three dimensions. Additionally, it describes how the system estimates affect and the indexes it outputs. Finally, the system was tested comparing its results against those of a group of human annotators. Moreover, the utility of the system was tested for the task of classifying products opinions into positive or negative.

Keywords: Natural Language Processing, Affect, Emotions, Spanish.

Índice general

1. INTRODUCCIÓN

En un intento por cuantificar las emociones que se pueden desprender de un texto en el idioma inglés, Cynthia Whissell desarrolló una herramienta llamada Dictionary of Affect in Language (*DAL*)[12]. *DAL* tiene como propósito puntuar un texto en base a tres dimensiones: agrado (*pleasantness*), activación (*activation*) e imaginabilidad (*imagery*). Ese resultado es obtenido a través de una base de conocimiento creada con la puntuación realizada por 200 voluntarios.

El trabajo de Whissell se realizó en varias etapas. Inicialmente seleccionó un cuerpo de datos conformado por palabras. Luego, realizó la puntuación de las palabras en las 3 dimensiones, que fue hecha por un grupo de voluntarios. Finalmente se construyó la herramienta, que dado un texto, realiza el cálculo de los valores para cada dimensión. Adicionalmente, se evaluó *DAL* en términos relativos a la confiabilidad y validez de los resultados obtenidos.

Para la selección del cuerpo de datos, Whissell tomó un conjunto compuesto de 1.000.000 de palabras elaborado por Kucera y Francis[6] . El cuerpo de datos fue procesado sobre un grupo de textos y se filtró aquellas palabras que tenían una frecuencia menor a 10 apariciones, evitando de esa manera palabras raras o específicas. Las palabras que pasaban el filtrado eran comparadas con nuevos textos y en aquellos casos en que se encontraba una nueva palabra, ésta era agregada a la lista de palabras seleccionadas. La finalización de la etapa de selección dejó un total de 8742 palabras.

Para la puntuación de las palabras se contó con 200 personas entre las que se incluían tanto hombres como mujeres. La tarea que tenían asignada era decidir qué puntuación le asignaban a cada palabra según las siguientes escalas de valores:

Agrado	**Activación**	**Imaginabilidad**
Desagradable (*Unpleasant*)	Pasivo (*Passive*)	Difícil de imaginar (*Hard to imagine*)
Ni agradable ni desagradable (*In between*)	Ni activo ni pasivo (*In between*)	Ni difícil ni fácil de imaginar (*In between*)
Agradable (*Pleasant*)	Activo (*Active*)	Fácil de imaginar (*Easy to imagine*)

Tab. 1.1: *Valores posibles para cada una de las distintas dimensiones.*

La última versión disponible de *DAL* ejecutada sobre un texto, arroja entre sus resultados las medias de *agrado, activación e imaginabilidad* que se corresponden con:

Mean Pleasantness, Mean Activation y Mean Imagery

En esta tesis se replicó la herramienta desarrollada por Whissell, para el idioma español. Dados los pocos detalles de la investigación original y a que fue hecha en 1989, se ha innovado en algunos aspectos. Entre los aspectos que se innovaron se destaca la automatización de la etapa de filtrado de palabras, la puntuación del cuerpo de datos con voluntarios y su posterior evaluación.

Las implicancias de este trabajo son amplias y serviría de piedra angular para distintos tipos de investigaciones llevadas a cabo a posteriori. Entre las disciplinas que pueden beneficiarse de este trabajo figuran Neurociencias, Psicología y Procesamiento del Lenguaje Natural, entre otras.

En el caso del Procesamiento del Lenguaje Natural el beneficio podría ser aplicado al análisis de sentimiento en opiniones (Sentiment Analysis). Por ejemplo, día a día en los sitios de ventas la gente genera opiniones relativas a productos comprados. Si estas opiniones pudieran ser automáticamente procesadas y enviadas a los fabricantes para hacer mejoras en los productos, se daría la oportunidad de hacer una mejor experiencia a los usuarios.

En Neurociencias algunos experimentos requieren la utilización de un conjunto de palabras que disparen ciertos sentimientos, como podría ser el agrado. Por

ejemplo, en el estudio de Gray[7] se analiza cómo un grupo de personas pueden ser influenciadas en el proceso cognitivo según el estado emocional. En dicho estudio se seleccionan videos y palabras para los experimentos. Para esa clase de trabajos la utilización del sistema desarrollado en esta tesis permitirá evitar realizar dicha selección manual, de modo de obtener un conjunto imparcial de palabras.

En el capítulo 2, comenzaremos describiendo en detalle cómo fueron seleccionadas las palabras. Luego continuaremos en el capítulo 3 detallando el sistema de puntuación que se creó para que los voluntarios hagan las puntuaciones de cada palabra. Para terminar, en el capítulo 4 se comenta la creación del sistema de estimación y se lo evalúa contra dos cuerpos de datos.

2. SELECCIÓN DE PALABRAS

En este capítulo se describe en detalle el proceso con el cual se obtuvo el listado de palabras que luego fueron puntuadas. Este proceso contó con varias etapas, entre las cuales figuran: la selección de distintos cuerpos de datos, la construcción y la aplicación de filtros sobre los cuerpos. Finalmente mostramos un análisis estadístico sobre las palabras seleccionadas.

2.1. Cuerpo de datos a procesar

El primer desafío para la construcción del sistema fue seleccionar un conjunto de palabras que sean representativas del idioma español. Este cuerpo de datos debía ser diverso, numeroso y representativo de la lengua española. Tras estudiar distintas alternativas, se optó por *Wikipedia*[1], en español, que cumplía con las premisas propuestas.

En la búsqueda de diversificar y sumar nuevas palabras al conjunto se decidió procesar otro cuerpo de datos: *Los Cuentos*[2], un sitio que pertenece a una comunidad literaria dedicada a cuentos en español. Los cuentos se encuentran publicados online, de forma tal que cualquiera tiene acceso a los mismos. Los autores de estas narrativas varían desde autores conocidos y de renombre, como puede ser Julio Cortázar, hasta autores no tan conocidos y con intenciones de iniciarse en el mundo literario. De esta manera, *Los Cuentos* aportó palabras más vinculadas al genero literario, lo cual dio mayor diversidad al cuerpo de datos final.

2.2. Extracción de palabras de Wikipedia y Los Cuentos

El procesamiento de las palabras de *Wikipedia* constó de varios pasos. Como primera medida se utilizó un *parser XML* para tomar sólo el texto de los artículos

[1] http://es.wikipedia.org
[2] http://www.loscuentos.net

publicados, eliminando de esa manera tags propios de *Wikipedia*. El *parser* generó varios archivos de tamaño más pequeño (20 MB), los cuales se fueron leyendo de a líneas. Esas líneas fueron tratadas para eliminar aquellos símbolos que no aportan significado, como por ejemplo los símbolos de puntuación (ver el listado completo de símbolos en el apéndice A.1). Las palabras resultantes fueron transformadas a minúsculas y agregadas a una estructura de datos del tipo diccionario, con la palabra como clave y su cantidad de apariciones, como valor. Los cuentos se bajaron de internet mediante un script. Del sitio online se obtuvieron más de 200 mil cuentos, los cuales dieron un total de 45.486 palabras.

2.2.1. Filtrado de palabras

Para eliminar aquellas palabras que no aportaban valor a los fines del estudio aplicamos un conjunto de reglas de filtrado. Cuando decimos que no aportaban valor, nos referimos a que no tenía sentido en una etapa posterior puntuar esas palabras por voluntarios. En su mayoría, las palabras dejadas fuera del cuerpo de datos eran difíciles de calificar en las 3 dimensiones que se requiere para el experimento. A continuación describimos estas reglas de filtrado.

Al haber en *Wikipedia* un vocabulario tan diverso, varias palabras tienden a tener pocas apariciones, dando como resultado un cuerpo de datos demasiado grande. Ese cuerpo estaba comprendido por cerca de 1.200.000 palabras. Con el fin de reducir la cantidad de palabras, decidimos eliminar aquellas que tenían dos caracteres o menos dado que no aportan datos significativos para el estudio.

Otra regla que aplicamos fue, que, luego de procesar cada uno de los archivos de 20 MB, se eliminaron aquellas palabras que tienen menos de 10 apariciones. Con esta estrategia, evitamos aquellas palabras poco frecuentes y demasiado específicas, quedándonos con las palabras más frecuentes.

También fue necesario filtrar nombres de ciudades, países y regiones. Para distin-

guir esa clase, utilizamos una base de datos con información geográfica[3]. Para este filtrado se procesó cada palabra buscando si pertenecía a un país, ciudad o región de las que estaba disponible en la base de datos. Por ejemplo, la palabra *Ezeiza* formaba parte de las ciudades en la base, por lo que fue eliminada del listado.

A continuación, filtramos los nombres propios. Para ello creamos manualmente una base de datos que constó de 762 nombres propios, obtenidos de distintos sitios de Internet. Corresponde aclarar que algunos nombres no fueron tenidos en cuenta dado que tenían otras acepciones además de ser un nombre propio. Por ejemplo, *Victoria* puede ser tanto un nombre como sustantivo, y en este último caso nos interesa que quede en el listado de palabras.

Luego se trabajó manualmente sobre cierto grupo de palabras. Del total de palabras se revisaron unas 5 mil, donde se eliminaron gentilicios, palabras en inglés, número en su forma textual, números romanos y otras palabras más (ver en el apéndice A.2 las palabras excluidas). Algunos ejemplos de este tipo de palabras son: *estadounidense* como gentilicio, la palabra en inglés *the*, el número romano *XIX* y el número de orden *sexto*, entre otros.

Otro conjunto de palabras que fueron excluidas del cuerpo de datos fueron algunos adverbios. Se excluyeron aquellos que son adverbios de lugar, tiempo, modo, duda, negación, afirmación y de cantidad. La lista completa de adverbios eliminados se puede ver en el apéndice A.3. Esta medida fue consecuencia del hecho que no aportan valores significativos para el estudio.

En una etapa más avanzada, luego de que se hizo el análisis morfológico, se detectó cierto grupo de palabras que en Lingüística se denominan conectores (ejemplos: *en lugar de*, *por lo cual*, *porque*). Estos conectores no aportaban valor al estudio, por lo que se decidió eliminar algunos de ellos. Para ver los conectores eliminados ver apéndice A.4.

[3] http://www.geobytes.com/FreeServices.htm

Al finalizar la etapa de filtrado se distinguieron, en *Wikipedia*, **163.071** palabras y para el cuerpo de datos de *Los Cuentos* se obtuvo un total de **30.544** palabras.

2.2.2. Análisis morfológico de las palabras

Al finalizar la ejecución de las reglas de filtrado, se generó un archivo con el listado de palabras. A ese listado luego se lo analizó con el analizador del lenguaje *Freeling*[8], que es de código abierto y de licencia GNU. *Freeling* permitió hacer análisis morfológico, reconocimiento de días, monedas, etc. sobre un texto. El análisis morfológico sirvió para determinar la forma, y la categoría gramatical de cada palabra en una oración.

Freeling funciona procesando texto y según el tipo de salida que se configure en el servidor arroja un resultado con cierto formato. Para procesar el listado de palabras de *Wikipedia* y *Los Cuentos*, se configuró la salida con la opción *morfo*. Esa opción significa que se analiza el texto de manera morfológica, es decir, se determina la forma, y la categoría gramatical de cada palabra. La figura 2.1 muestra un ejemplo del resultado con la opción *morfo* para la frase *El gato come pescado y bebe agua*. Nótese que la palabra *pescado* tiene dos acepciones posibles: sustantivo o verbo participio. Asimismo, *agua* puede ser sustantivo o verbo.

El	gato	come	pescado	y	bebe	agua	.
el	*gato*	*comer*	*pescado*	*y*	*beber*	*agua*	*.*
DA0MS0	NCMS000	VMIP3S0	NCMS000	CC	VMIP3S0	NCCS000	Fp
1	1	0.916667	0.954545	0.999962	0.994868	0.99177	1
		comer	*pescar*	*y*	*beber*	*aguar*	
		VMM02S0	VMP00SM	NCFS000	VMM02S0	VMIP3S0	
		0.0833333	0.0454545	3.76761e-05	0.00513196	0.00411523	
						aguar	
						VMM02S0	
						0.00411523	

Fig. 2.1: *Resultado de procesar la frase: "El gato come pescado y bebe agua". con la opción morfo. En color negro (1^{ra} fila) aparece el texto original. En color azul (2^{da} fila) se muestran los distintos lemas que son parte de la palabra que está en la misma columna. En color rojo (3^{ra} fila) se ven los tags que especifican más detalle de la palabra. Finalmente en color negro (4^{ta} fila) se ve la probabilidad que se le asigna a cada lema.*

Al procesar un texto con *Freeling* con la opción *morfo* se obtiene el siguiente formato para cada palabra:

palabra (lema tag probabilidad)+

donde *palabra* corresponde a la palabra original que se procesó; *lema* es el lema que se obtuvo de *palabra*; *tag* es una etiqueta alfanumérica que aporta más detalles de la palabra[4]; *probabilidad* es la probabilidad de que la palabra tenga el *tag* especificado.

El uso del analizador morfológico permitió filtrar algunas palabras, como por ejemplo, las preposiciones. La identificación de las características de cada palabra fue obtenida por medio de la etiqueta *tag*, eliminando así: determinantes posesivos, interjecciones, preposiciones, conjunciones, cifras, números, fechas y horas.

Una de las principales tareas del analizador morfológico fue lematizar los verbos llevándolos a su forma en infinitivo. Eso permitió colapsar todos los verbos en distintos tiempos verbales en uno solo. Por ejemplo, en el caso del verbo *surgir* fue colapsado de las siguientes conjugaciones: *surgió, surge, surgir, surgieron, surgido, surgen, surgiendo, surgida, surgidos, surgidas y surgían.*

En el caso de los sustantivos también fueron lematizados. Sin embargo, se hizo una excepción con el grado aumentativo y diminutivo. Es decir, si el sustantivo era aumentativo o diminutivo, se lo dejaba igual a la palabra original. La finalidad de considerar estos casos de manera distinta fue la suposición, de que su evaluación subjetiva sería distinta a la de su lema. Por ejemplo, entre las palabras *gato* y *gatito*, probablemente *gatito* tendrá un nivel de agrado más alto que *gato.*

Para la utilización de *Freeling* se probaron 2 variantes. Por un lado, se intentó usarlo desde *Java* como una biblioteca externa. Por otro lado, se utilizó directamente la opción por línea de comando que es provista con la instalación. La ejecución vía biblioteca externa con *Java* no funcionaba de manera correcta para la

[4] Los distintos tipos de tags se pueden ver en la documentación de *Freeling* http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

cantidad de palabras que era necesario procesar. Por esa razón, se decidió utilizar el comando *analyzer_server* junto con el comando *analyzer_client*. Para tal fin, se partió el archivo original con las palabras extraídas, en archivos más pequeños de 5 mil palabras. Estos archivos de 5 mil palabras se procesaron en un servidor desde distintas máquinas utilizando el comando *analyzer_client*. La ventaja de realizarlo de esta manera era que permitía escalar la cantidad de clientes y de servidores.

La mejor alternativa para el análisis hubiera sido configurar el analizador morfológico con la opción *parsed*, para hacer un *full parsing* del texto, y así obtener un detalle amplio sobre la estructura sintáctica que tiene la sentencia que se analiza.

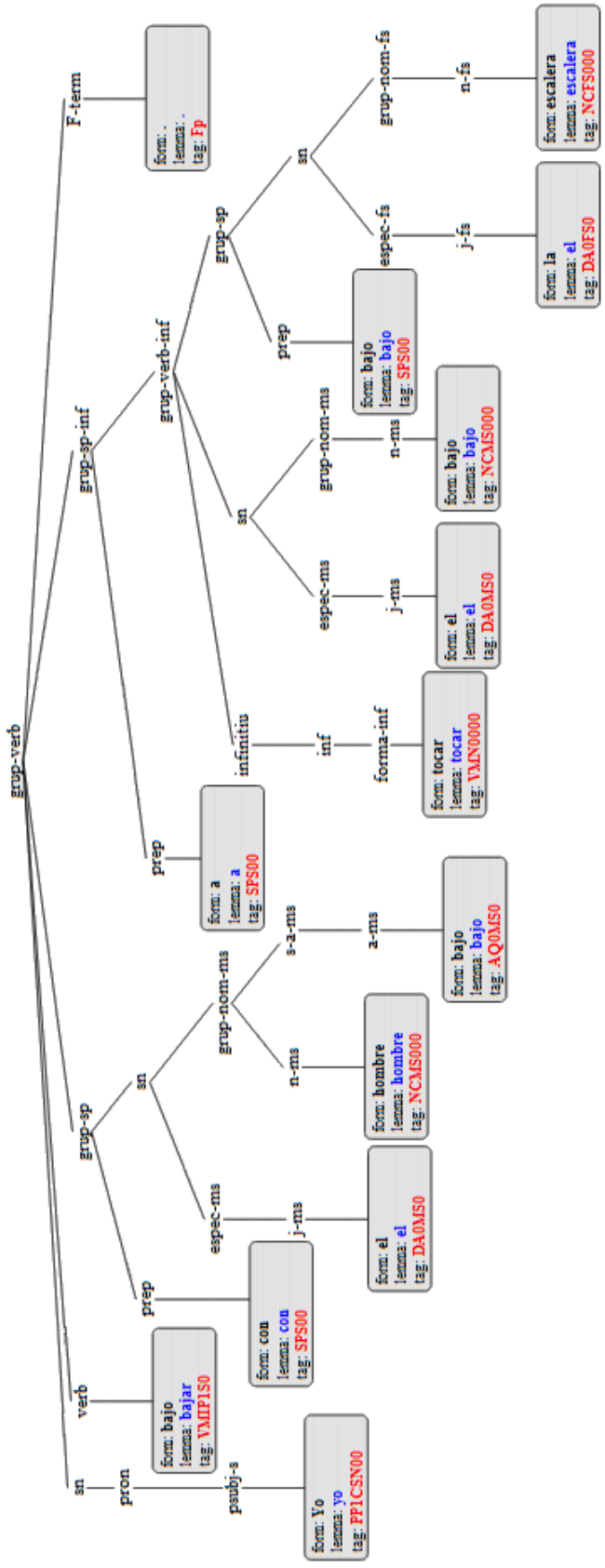

Fig. 2.2: *Resultado de procesar la frase: Yo bajo con el hombre bajo a tocar el bajo bajo la escalera con full parsing.*

Además, se hubiera considerado unívocamente qué acepción de cada palabra se está usando. Lamentablemente no es computacionalmente realizable la tarea de *full parsing* sobre textos tan extensos. En el caso de *Wikipedia*, para un párrafo seleccionado al azar la herramienta demoró en promedio 3 segundos. Dado que cuenta con 9.276.844 párrafos, procesar todo el cuerpo de datos hubiera llevado 27.830.532 segundos, o más de 322 días. El tiempo necesario para hacer *full parsing* sobre el cuerpo de datos hizo que se descarte la idea.

En la imagen 2.2 de la página 10, vemos como sería el resultado del *full parsing* para la frase *Yo bajo con el hombre bajo a tocar el bajo bajo la escalera*.

2.2.3. Creación de la lista de palabras

Una vez procesados los cuerpos de datos con el analizador morfológico, el siguiente paso fue generar la lista de palabras. Recordemos el formato de salida obtenido con la opción *morfo*:

palabra (lema tag probabilidad)+

Utilizando la etiqueta *tag* se filtró, por ejemplo, las preposiciones omitiendo aquellas palabras cuyo *tag* comenzaban con la letra *S*[5]. De esta manera se especificó los distintos tipos de palabras que se eliminaron.

La etapa de análisis morfológico mostró algunas particularidades con respecto a la salida generada. El lema, en algunos casos, estaba compuesto por 2 partes separadas por el símbolo +. Por un lado, estaba el lema propiamente dicho y por otro lado, se agregaba un sufijo. En la mayoría de los casos el sufijo correspondió a un pronombre. Un ejemplo es la palabra *considerarse*, donde se obtiene *considerar+se*. Aquí se separa por un lado el verbo, en este caso *considerar* y por otro lado el pronombre, *se*. En esas ocasiones se tomó sólo la parte izquierda hasta el símbolo +, dado que es la parte correspondiente a la raíz de la palabra.

[5] Para más información sobre que características se pueden obtener a través del *tag* ver http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

La lista de palabras fue hecha en orden descendente por la cantidad de apariciones de cada palabra. Esas palabras, sacadas de *Wikipedia* y *Los Cuentos*, aparecían una vez por cada categoría gramatical. Las categorías consideradas en esta tesis fueron: adjetivo, sustantivo, verbo y adverbio.

Cuando se realizó el análisis morfológico las palabras no tenían un contexto. Es decir, no estaban en una oración, sino que eran palabras sueltas. Por lo tanto, se decidió utilizar el dato *probabilidad* para estimar con mejor precisión la cantidad de apariciones de los pares *(palabra, categoría gramatical)*. Por ejemplo, para analizar la palabra *bajo* sin contexto, el resultado fue el de la tabla 2.1. La interpretación de esta

lema	**tag**	**categoría gramatical**	**probabilidad**
bajo	SPS00	Preposición	0.879562
bajo	AQ0MS0	Adjetivo	0.0766423
bajo	NCMS000	Sustantivo	0.040146
bajar	VMIP1S0	Verbo	0.00364964

Tab. 2.1: *Análisis morfológico de la palabra bajo.*

salida corresponde, en primer lugar a la palabra *bajo* como preposición (por ejemplo: *Jose estaba* ***bajo*** *la mesa*). En este caso, la salida muestra que la probabilidad de que la palabra corresponda a una preposición es de 0.879562. La segunda posibilidad es que corresponda a un adjetivo (ejemplo: *Federico es el más* ***bajo*** *del equipo*) con una probabilidad de 0.0766423. La tercera alternativa es para el caso en que *bajo* resulte ser un sustantivo (ejemplo: *El* ***bajo*** *suena mejor que la guitarra en algunas bandas de música*) con probabilidad de 0.040146. Finalmente la probabilidad de que sea verbo es 0.00364964 (ejemplo: *Yo* ***bajo*** *por las escaleras solo en caso de emergencia.*)

2.2.4. Recálculo de la cantidad de apariciones

Luego de correr el analizador morfológico, se decidió recalcular la cantidad de apariciones teniendo en cuenta las distintas categorías de palabras. El recálculo utilizó el dato *probabilidad* y la cantidad de apariciones calculadas previamente al ejecutar el analizador del lenguaje. En el **Algoritmo 1** se muestra con pseudo-código cómo se calculó la cantidad de apariciones para una palabra.

Algoritmo 1 Calcula la cantidad de apariciones que tienen los lemas de un texto.

para cada palabra P **hacer**
  para cada lema L de P **hacer**
    $C \leftarrow$ categoria gramatical de L
    $cantAp(L, C) = cantAp(L, C) + cantAp(P) * prob(P, C)$
  fin para
fin para

donde:

- *cantAp(L,C)* es la cantidad de apariciones del lema L con la categoría gramatical C. Inicialmente se encuentra inicializado con 0 para todo (L, C).
- *cantAp(P)* es la cantidad de apariciones de la palabra P previo a ejecutar el analizador morfológico.
- *prob(P,C)* es la probabilidad obtenida del analizador morfológico para la palabra P que tiene la categoría C .

Por ejemplo, suponiendo que la palabra *bajo* aparece 1000 veces, la cantidad de apariciones en sus distintas formas serían:

(bajo , preposicion)	= (1000*0.880)	= 880
(bajo , adjetivo)	= (1000*0.077)	= 77
(bajo , sustantivo)	= (1000*0.040)	= 40
(bajar , verbo)	= (1000*0.004)	= 4

Siguiendo con el ejemplo, supongamos que la palabra *bajar* aparece 300 veces, según el algoritmo 1 el cálculo sería el siguiente:

(bajar , verbo)	= 4 + (300*1)	= 304

Donde el valor 4 utilizado corresponde al cálculo hecho para (bajar , verbo) de la palabra *bajo*.

2.2.5. Combinación de palabras de Wikipedia y Los Cuentos

A continuación procedimos a integrar los dos cuerpos de datos (*Wikipedia* y *Los Cuentos*). En primer lugar se normalizó la cantidad de apariciones de palabras que tenía cada uno. El proceso de normalización consistió en reasignar el valor de

cantidad de apariciones para cada palabra mediante la siguiente formula:

$$\frac{C_p}{\sum\limits_{p \in palabras} C_p} \tag{2.1}$$

donde C_p es la cantidad de apariciones de la palabra p y *palabras* corresponde al conjunto de todas las palabras de ese cuerpo de datos. Es decir, por cada cuerpo de datos se tenía un listado de palabras con sus respectivas cantidades normalizadas (entre 0 y 1). Para combinarlos se hizo la unión de ambos listados y se los ordenó de manera descendente en función del valor obtenido. En los casos que algunas de las palabras pertenecían a ambos cuerpos, se aplicó una interpolación entre ambos valores, quedando la palabra con el nuevo valor calculado. Como resultado se obtuvo un cuerpo de datos combinado con palabras de *Wikipedia* y *Los Cuentos*.

2.3. Estadísticas de las palabras

A medida que se fueron procesando los distintos filtros sobre los cuerpos de datos la cantidad de palabras fue variando. En un principio la cantidad total de vocablos disminuyó pero luego al utilizar el analizador morfológico se vio un incremento. En algunos casos, palabras que no aparecieron en el cuerpo sí fueron consideradas en la etapa del analizador del lenguaje agregando un nuevo vocablo. Por ejemplo, supongamos que el cuerpo de datos no tiene la palabra *bajar* pero sí el vocablo *bajo*. En este caso, al hacer el análisis morfológico de *bajo* (ver tabla 2.1 de la página 12) se agrega la palabra *bajar* al listado de palabras, incrementando la cantidad.

Luego de aplicar los filtros a cada cuerpo de datos se realizó la unión de las palabras dejando un total de **175.413** palabras. Una vez obtenido el listado final de palabras se decidió visualizar a través de un histograma la distribución que existía con respecto a la cantidad de apariciones de una palabra. Estos histogramas nos sirven para ver cómo están distribuidas las palabras en cada cuerpo de datos. Las figuras 2.3 y 2.4 muestran los histogramas relativos a los datos usados para *Wikipedia* y *Los Cuentos* respectivamente.

De las imágenes se puede ver que, a grandes rasgos ambos gráficos tienen el mismo patrón. Se encuentra un aumento de la frecuencia a medida que se va aumentando la cantidad de apariciones hasta llegar a un punto máximo y luego comienza a bajar. Analizando los histogramas vemos que, los valores más grandes del eje X son los que tienen mayor cantidad de apariciones. Aquí se puede ver que a medida que son mayores los valores, menor cantidad de palabras con esa cantidad de apariciones hubo. En otras palabras a mayor cantidad de apariciones menor frecuencia. Por lo tanto, nosotros hemos dado prioridad al conjunto de palabras con más apariciones.

2.3.1. Cubrimiento

El cubrimiento se define como el porcentaje de elementos de un conjunto que aparecen en otro conjunto. En nuestro caso, vamos a considerar el porcentaje de palabras de un texto que aparece en el listado de palabras a puntuar. Con la intención de mostrar qué cantidad de palabras del cuerpo de datos son necesarias para lograr un buen cubrimiento, se decidió procesar textos extraídos del sitio *Wikinews*[6]. Estos textos son artículos periodísticos, un dominio diferente al de *Wikipedia* y *Los Cuentos*

Se fueron generando cuerpos de datos tomando los primeros k elementos de cada cuerpo (*Wikipedia* y *Los Cuentos*) y aumentado gradualmente el valor de k. Se comenzó con $k = 100$ palabras y se terminó con la cantidad de palabras que había en el cuerpo más grande. El valor de k se fue incrementando de a 100 palabras para apreciar la evolución del cubrimiento según la cantidad de palabras que tenía el cuerpo de datos.

El algoritmo para determinar el cubrimiento consiste en recorrer cada archivo con el detalle del *full parsing* generado por *Freeling*. Para ello, se recorren todos los lemas y se ve si pertenecen o no al cuerpo de palabras combinado. El **Algoritmo 2** (página 17) muestra el pseudo-código utilizado.

[6] http://www.wikinews.org/

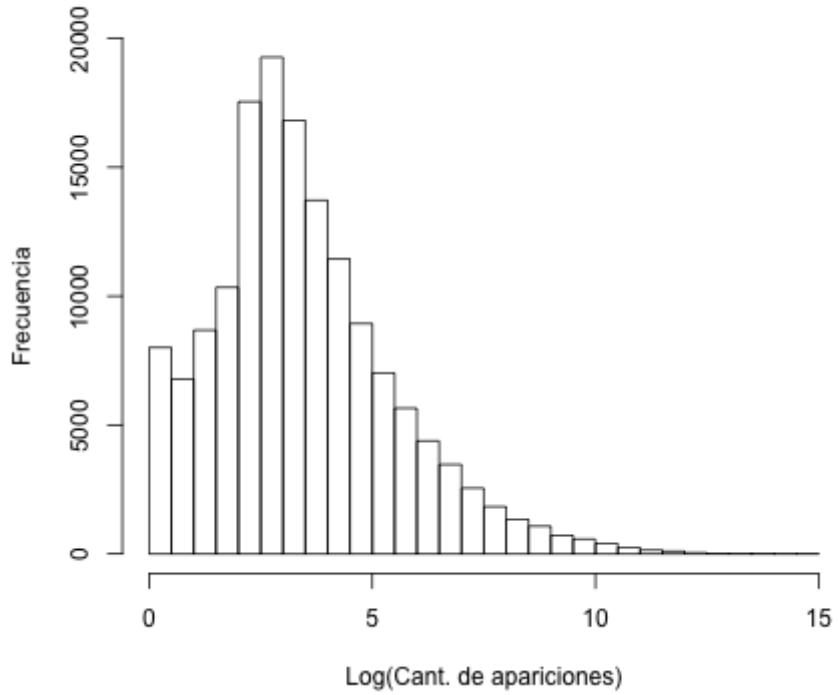

Fig. 2.3: *Histograma de Wikipedia. Cantidad de palabras que aparecen e^x veces en el cuerpo de datos.*

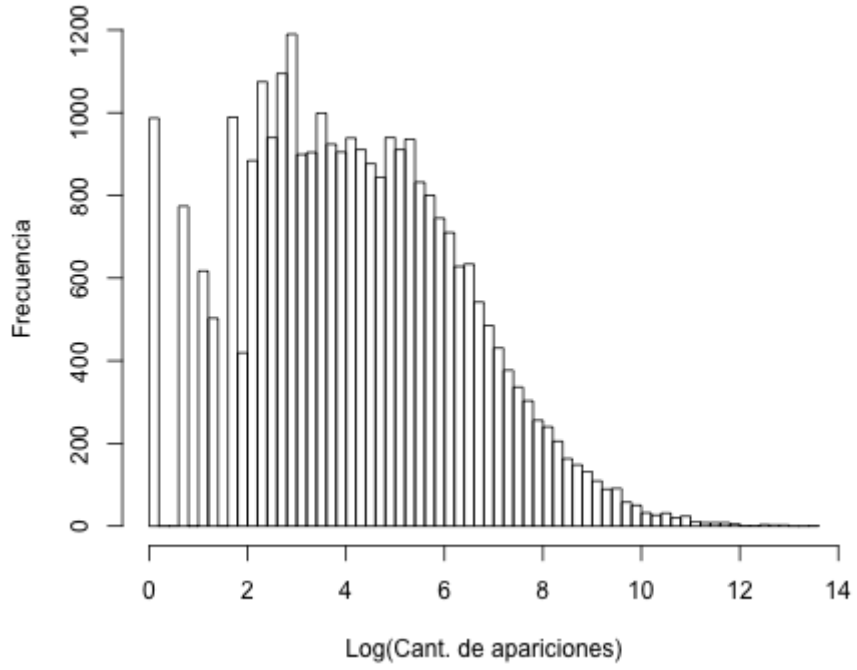

Fig. 2.4: *Histograma de Los Cuentos. Cantidad de palabras que aparecen e^x veces en el cuerpo de datos.*

Algoritmo 2 Calcula el porcentaje de cubrimiento de un cuerpo de datos sobre otro cuerpo procesado con *full parsing*.

```
palabrasEnElCuerpo ← 0
palabrasTotales ← 0
crear diccionario vacio
para cada lema en el archivo generado por Freeling hacer
  agregar lema en diccionario
fin para
para cada palabra p en el diccionario hacer
  si p pertenece al cuerpo de datos entonces
    palabrasEnElCuerpo ← palabrasEnElCuerpo + 1
  fin si
  palabrasTotales ← palabrasTotales + 1
fin para
devolver palabrasEnElCuerpo/palabrasTotales
```

En la figura 2.5 se puede ver el gráfico de cubrimiento sobre *Wikinews*. En el eje X aparece la cantidad de palabras que fueron consideras en el cuerpo de datos, mientras que en el eje Y aparece el porcentaje que el cuerpo cubrió sobre *Wikinews*.

De este gráfico se puede ver cómo aumenta el porcentaje de cubrimiento a medida que se va considerando mayor cantidad de palabras en el cuerpo de datos. Otro aspecto a destacar es cómo se va formando una meseta luego de cierta cantidad de palabras incluidas en el cuerpo. Así, pensando en una etapa posterior de puntuación de palabras, se podría pensar que para analizar *Wikinews* no es necesario ponderar las 175.413 palabras. Con lograr, digamos, 20 mil palabras con mayor cantidad de apariciones, se obtendría un cubrimiento casi óptimo.

Por otro lado es necesario mencionar que incluso considerar todas las palabras de los cuerpos de datos de *Wikipedia* y *Los Cuentos* no alcanza para lograr un 100 % del cubrimiento de *Wikinews*. Esto puede deberse a que *Wikinews*, al tener un formato de artículo periodístico, contiene gran cantidad de fechas, nombres y lugares, entre otras palabras que no se consideran para puntuar en el cuerpo de datos original.

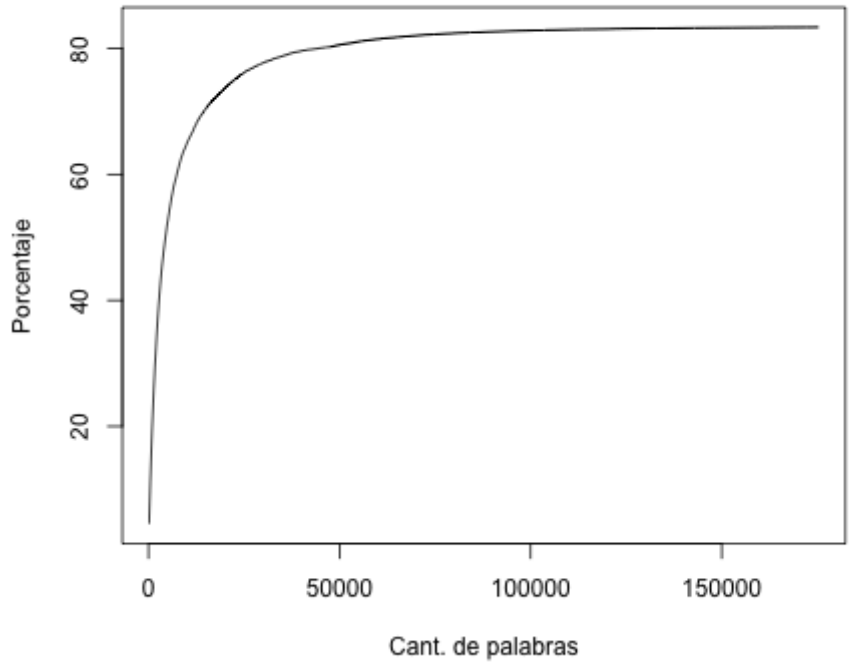

Fig. 2.5: *Cubrimiento sobre Wikinews. El eje X determina la cantidad de palabras que fueron consideradas en el cuerpo de datos, mientras que en el eje Y aparece el porcentaje que se cubrió sobre Wikinews. El porcentaje debe interpretarse en centecimos.*

3. PUNTUACIÓN DE PALABRAS

Una vez procesados los cuerpos de datos (*Wikipedia* y *Los Cuentos*) se llegó al listado de palabras sobre el cual se va a trabajar. Para poder estimar el afecto de un texto fue necesario ponderar las palabras obtenidas, para ello se contó con la colaboración de voluntarios. Para la puntuación se contruyó un sitio web[1] donde pudieron calificar las palabras. A continuación, se detalla cómo fue la construcción del sitio para puntuar un vocablo en las 3 dimensiones: *agrado, activación e imaginabilidad.*

Para esta etapa es necesario mencionar algunas diferencias con el trabajo realizado por Whissell. Estas diferencias se deben en gran medida al hecho que en esa época no había un uso masivo de Internet. El uso de las redes sociales que permiten actualmente una mayor difusión en cantidad y diversidad no estaban en auge como en estos días.

3.1. Valores de puntuación

Haciendo uso de las 3 dimensiones del trabajo de Whissell, decidimos utilizar una escala similar. Para las 3 dimensiones a ponderar se usó una escala de 3 valores. En la tabla 3.1 se pueden ver los valores asignados a cada opción que se votó.

Agrado	**Activación**	**Imaginabilidad**
(1) Desagradable	(1) Pasivo	(1) Difícil de imaginar
(2) Ni agradable ni desagradable	(2) Ni activo ni pasivo	(2) Ni difícil ni fácil de imaginar
(3) Agradable	(3) Activo	(3) Fácil de imaginar

Tab. 3.1: *Valores de puntuación para cada una de las distintas dimensiones.*

3.2. Construcción del sitio de puntuación

Para agilizar la puntuación decidimos pedir al participante poca información, pero relevante. Entre los datos que se requirieron figuraba: la fecha de nacimiento

[1] http://www.ocovinu.com.ar

(en formato mes y año), el máximo nivel completo de educación y si el participante tenía el idioma español como idioma nativo. En el caso que el voluntario marcara que no tiene como idioma nativo el español, no se le permitió completar la encuesta. Además, se le pidió a los participantes completar un *recaptcha*[11], el cual aportó seguridad al sitio, evitando la generación de datos erróneos, por parte de *web crawlers* o programas maliciosos que podrían haber hecho peligrar los resultados.

En la figura 3.1 se puede ver la página de inicio del sitio construido con la información necesaria.

Fig. 3.1: *Página de inicio del Sistema de Puntuación*

Una vez que el voluntario ingresó los datos requeridos y el *recaptcha* se validó correctamente se presenta una página de introducción. En ésta página se explicó de manera simple y con poco detalle (para no sesgar al voluntario) lo que debía hacer para puntuar. En esta página se le pidió al voluntario la puntuación de 20 palabras. En la figura 3.2 se puede ver la imagen de la página en cuestión.

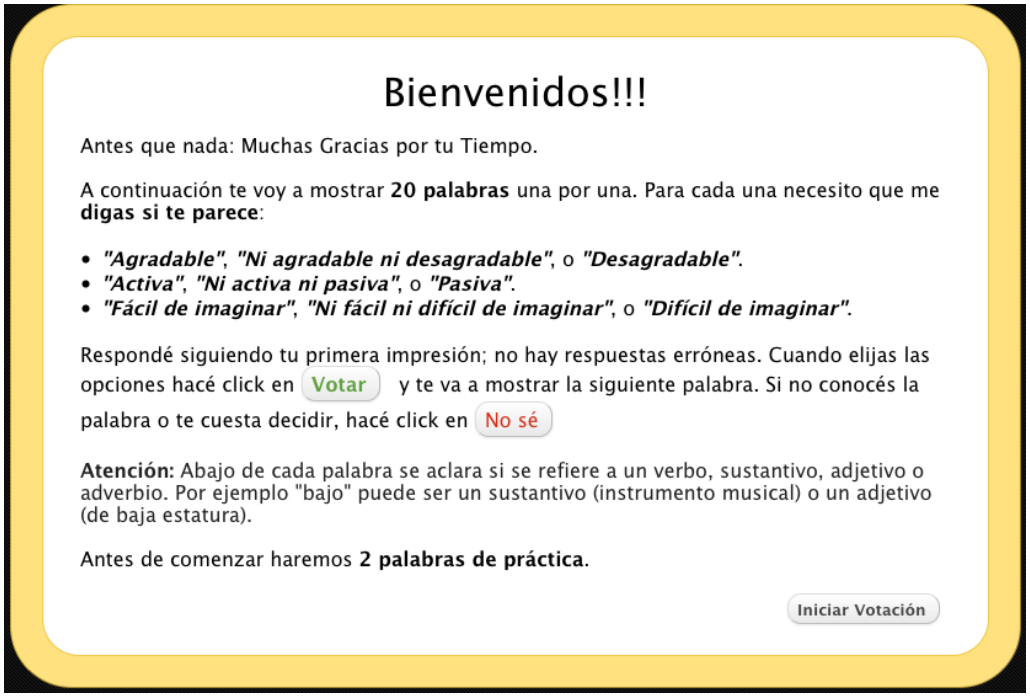

Fig. 3.2: *Página de introducción del Sistema de Puntuación*

En la página de introducción el voluntario fue notificado, que al hacer click en *Iniciar Votación* le aparecerían 2 palabras de entrenamiento. De esa manera, se logró un primer contacto del voluntario con el sistema de puntuación. Las 2 palabras de entrenamiento fueron *inmundicia* y *fastidiar*, las cuales fueron elegidas al azar. Terminado el entrenamiento de las 2 palabras, se mostró un mensaje diciendo que se iba a iniciar la votación requerida para el experimento. Creemos necesario aclarar que estas 2 palabras de entrenamiento no fueron tenidas en cuenta para las etapas posteriores.

Mientras los voluntarios realizaban la puntuación se les mostró, debajo de cada palabra, la categoría gramatical correspondiente a la misma. Además, debajo de la categoría se informó el número de palabra que estaba completando de las 20 necesarias. En la figura 3.3 se puede ver como se mostró una palabra para puntuar del experimento.

Al finalizar las 20 palabras, al voluntario se le dio la opción de seguir colaborando

Fig. 3.3: *Página de puntuación para la palabra Navegar con su forma de uso verbal en el Sistema de Puntuación*

con otras 20 palabras o finalizar la puntuación. En caso de haber optado por continuar colaborando era redirigido a la página de puntuación con un nuevo conjunto de palabras.

3.3. Asignación de palabras para cada voluntario

Las **175.413** palabras que conformaron nuestro cuerpo de datos fueron ingresadas a una base de datos. Estas palabras fueron ordenadas en forma descendente de acuerdo a la cantidad de apariciones normalizada entre *Wikipedia* y *Los Cuentos*.

Para cada voluntario, se eligieron las 20 palabras de la siguiente manera:

1. Se descartaron las palabras con 5 o más votos.
2. Se descartaron las palabras votadas por aquellos voluntarios que tenían la misma fecha de nacimiento.
3. Finalmente, se reordenaron de manera descendente las palabras según la cantidad de votaciones.

El método de asignación nos permitió mantener acotada la cantidad de palabras puntuadas con menos de 5 votaciones. Buscamos de esa manera tener la mayor can-

tidad de palabras completas, es decir, con 5 votaciones.

Por último, nótese que no hizo falta identificar con datos personales a los voluntarios para registrar unívocamente el voto y así evitar que un mismo voluntario puntuara dos veces la misma palabra. Estimamos que la cantidad de voluntarios hubiera sido mucho menor en el caso que se hubiera pedido por ejemplo el email para identificarlos.

3.4. Voluntarios

A través del sitio de puntuación ingresaron un total de **662** personas. Del nivel de educación de los voluntarios podemos decir que el 0.91 % de los mismos hizo sólo la primaria, el 22.71 % tenía completo el secundario, el 7.92 % tenía hecho un terciario, y el 68.45 % tenía una carrera universitaria completa.

En cuanto a la cantidad de palabras puntuadas al día de hoy asciende a **2566**. Cada una de esas palabras tuvo al menos 5 puntuaciones hechas por 5 voluntarios distintos. Vale repetir que, por la forma en que se hizo la asignación de palabras, no pudo haber participantes que hayan votado 2 veces la misma palabra.

Como resultado de la puntuación de los voluntarios comentaremos algunos datos obtenidos. Dada la variedad de personas de la muestra decidimos buscar las primeras 5 palabras que lograron un acuerdo perfecto. Es decir, aquellas palabras que fueron ponderadas 5 veces (cada vez por una persona distinta) en cada dimensión y todas las votaciones coincidieron. Estos vocablos se pueden ver en la tabla 3.2. En la tabla 3.3 se pueden ver las palabras que tuvieron menor y mayor valor en la media de agrado por sobre el resto. En otras palabras, los vocablos más desagradables y más agradables, respectivamente.

Palabra	Clase	Media agrado	Media activación	Media imaginabilidad
cabeza	sustantivo	2	2	3
canción	sustantivo	3	2	3
jugar	verbo	3	3	3
época	sustantivo	2	2	2
situar	verbo	2	2	2

Tab. 3.2: *Palabras extraídas de la base de conocimiento con acuerdo perfecto en todas sus dimensiones.*

Palabra	Clase
jugar	verbo
beso	sustantivo
sonrisa	sustantivo
compañía	sustantivo
reír	sustantivo
celebrar	verbo

Palabra	Clase
asesinato	sustantivo
caro	adjetivo
ahogar	verbo
herida	sustantivo
cigarro	sustantivo
lastimar	verbo

Tab. 3.3: *Palabras con* ***mayor*** *(izquierda) y* ***menor*** *(derecha) valor en la media de agrado.*

3.5. Estadística descriptiva de la puntuación

Presentamos aquí algunas estadísticas sobre los resultados obtenidos en las puntuaciones. En primer lugar estudiamos la correlación entre las distintas dimensiones, para lo cual armamos la matriz de correlación.

	Agrado	Activación	Imaginabilidad
Agrado	1.0	0.14	0.10
Activación	0.14	1.0	0.11
Imaginabilidad	0.10	0.11	1.0

Tab. 3.4: *Correlación entre las distintas dimensiones*

La tabla 3.4 muestra cómo la correlación entre las distintas dimensiones es muy débil. Esto confirma que las 3 dimensiones del estudio (*agrado, activación e imaginabilidad*) elegidas por Whissell tienen alto grado de independencia entre sí. Es decir, la percepción del agrado de las palabras tiende a ser independiente (u ortogonal) a la percepción de activación e imaginabilidad.

Para medir el grado de acuerdo de los voluntarios con respecto a las puntuaciones realizadas utilizamos la métrica Kappa. Esta métrica estima cuánto coinciden las

opiniones de diferentes personas sobre escalas nominales. La diferencia entre Kappa y el porcentaje simple de acuerdo o coincidencias es que Kappa excluye aquellos acuerdos que fueron por azar. En la formula de la figura 3.4 se puede ver como se excluyen aquellos casos que se dan por azar.

$$k = \frac{P_o - P_e}{1 - P_e}$$

Fig. 3.4: *Formula para unweighted Kappa*

Donde:

- P_o: Proporción de unidades en las cuales hubo acuerdo entre los voluntarios que votaron.
- P_e: Proporción de unidades para la cual el acuerdo se espera que sea por azar.

Llamaremos al coeficiente k *unweighted* Kappa[5]. La utilización de *unweighted* Kappa debe hacerse bajo las siguiente hipótesis con respecto a los coeficientes de acuerdo (en este caso los valores de las dimensiones):

- Deben ser independientes.
- Las opciones de la escala nominal deben ser independientes, mutuamente exclusivas y exhaustivas.
- Los voluntarios que hacen las puntuaciones deben operar de manera independiente.

Todas estas condiciones son cubiertas en nuestro proceso de puntuación. Sin embargo, resulta más conveniente por el tipo de escala que se utilizó el uso de *weighted* Kappa[4]. En la formula de la figura 3.5 se puede ver como se hace el cálculo del coeficiente k para *weighted* Kappa.

$$k = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$$

Fig. 3.5: *Formula para weighted Kappa*

Donde:

- Se usa cuando la escala es ordinal.
- La idea es tener en cuenta las diferencias de selección.
- k: Cantidad de valores en la escala ordinal.
- w: Matriz con los pesos de cada valor (en este caso se usó la escala de la tabla 3.1 de la página 19).
- x: Matriz con los valores observados (votaciones).
- m: Matriz con los valores esperados según el azar.

Este nuevo coeficiente hace una mejora sobre el anterior, teniendo en cuenta las escalas de valores para pesar el acuerdo o desacuerdo que hubo. Por ejemplo, supongamos que hay dos voluntarios que ponderan la misma palabra en cuanto a agradabilidad. El voluntario A decide que es *Agradable*(3) y el voluntario B decide que es *Ni agradable ni desagradable*(2). En este caso, *unweighted* Kappa arrojaría un valor de desacuerdo que sería el mismo que si el voluntario B hubiera seleccionado *Desagradable*(1). Sin embargo, con *weighted* Kappa se distingue que *Agradable* y *Desagradable* dista más que *Agradable* de *Ni agradable ni desagradable*.

En la tabla 3.5 se muestra la interpretación de los valores que pueden arrojar las dos variantes de Kappa, y en la tabla 3.6 se muestran los resultados obtenidos de la base de conocimiento.

Valor	Significado
< 0	Acuerdo pobre
0.01 - 0.2	Acuerdo leve
0.21 - 0.4	Acuerdo justo
0.41 - 0.6	Acuerdo moderado
0.61 - 0.8	Acuerdo sustancial
0.81 - 1.00	Acuerdo perfecto

Tab. 3.5: *Interpretación de los posibles resultados arrojados por la métrica Kappa*

	unweighted Kappa	*weighted* Kappa
Agrado	0.30	0.42
Activación	0.23	0.30
Imaginabilidad	0.21	0.14

Tab. 3.6: *Kappa weighted y unweighted aplicado a cada una de las dimensiones*

De la tabla 3.6 se puede ver que el mayor grado de acuerdo logrado fue en la dimensión agrado, tanto para *unweighted* como para *weighted* Kappa. Interpretando este valor con la escala descripta en la tabla 3.5 se llega a la conclusión que el coeficiente mostró un acuerdo moderado entre los voluntarios, lo cual resulta significativo teniendo en cuenta que la puntuación en las 3 dimensiones es una tarea extremadamente subjetiva.

4. SISTEMA DE ESTIMACIÓN DEL AFECTO

Llamamos *Sistema de Estimación del Afecto* al programa que calcula los valores de *agrado, activación e imaginabilidad* de un texto en función de la base de conocimiento adquirida a través de voluntarios. A continuación se detallará qué componentes fueron necesarios para la construcción y puesta en funcionamiento del sistema.

Es importante aclarar que el Sistema de Estimación del Afecto es una prueba de concepto. La motivación principal fue ver qué resultados se obtenían de la base de conocimiento conseguida, usando un sistema muy simple que pueda servir como baseline contra el cual comprar otros sistemas. Además, si los resultados son satisfactorios con un sistema tan simple, eso daría pie a experimentar con sistemas más complejos que usen la base de conocimiento.

4.1. Construcción del sistema

Para la construcción del sistema se utilizó la base de conocimiento. Se tomaron las palabras ponderadas a través del sitio de puntuación y fueron volcadas en una base de datos. Cada palabra quedó asociada con 6 valores, 3 correspondientes a los promedios de cada una de las dimensiones y los otros 3 a la desviación estándar, también de cada dimensión. En la tabla 4.1 presentamos un ejemplo, con la palabra *dormir*, de cómo se realizó el cálculo del promedio y la desviación estándar. Los valores puntuados en cada dimensión fueron los que se pueden ver en la tabla 3.1 (página 19). Los cálculos tanto del promedio como de la desviación estándar se realizaron con la información de cada fila.

Previo a la ejecución del sistema, se utilizó el analizador morfológico para hacer un *full parsing* del texto a estimar. Ese procesamiento ayudó a estimar con bajo error la acepción que se estaba usando en cada palabra. Recordamos que en la figura 2.2 (página 10) hay un ejemplo de la salida generada al hacer *full parsing* sobre un

	vot 1	vot 2	vot 3	vot 4	vot 5	promedio	desv. est.
Agrado	3	3	3	3	3	3	0
Activación	1	3	1	1	1	1.4	0.8
Imaginabilidad	3	3	3	3	3	3	0

Tab. 4.1: *Valores puntuados para el verbo dormir junto con el promedio y la desviación estándar de cada dimensión.*

texto. Una vez obtenida la nueva información (promedio y desviación estándar) de las palabras de la base de conocimiento, y corrido el analizador morfológico sobre el texto a estimar, se calcularon los valores promedios de las dimensiones para todas las palabras del texto.

A continuación mostramos un ejemplo de cómo estimaría el sistema el texto: *Mi amigo espera poder terminar la prueba a tiempo.* Para este caso supondremos que nuestra base de conocimiento tiene la información de la tabla 4.2. Recordemos que cada palabra tiene asociado los valores de las medias y las desviaciónes estándar para cada dimensión, en este ejemplo por simplicidad dejaremos afuera las desviaciones.

Palabra	Clase	Media agrado	Media activación	Media imaginabilidad
amigo	sustantivo	3.0	2.4	3
esperar	verbo	1.2	1	2.8
poder	verbo	2.8	2.8	2.2
terminar	verbo	2.2	3	2.8
prueba	sustantivo	1.8	2.4	2.2
tiempo	sustantivo	2	2	2.2

Tab. 4.2: *Base de conocimiento para analizar el texto:* ***Mi amigo espera poder teminar la prueba a tiempo****.*

Primero se obtienen los lemas de cada palabra y luego se busca si ese lema pertenece a la base de conocimiento para hacer el cálculo de la estimación. En la tabla 4.3 se pueden ver los lemas de cada palabra.

	Mi	amigo	espera	poder	terminar	la	prueba	a	tiempo
lema	mi	amigo	esperar	poder	terminar	el	prueba	a	tiempo

Tab. 4.3: *Lemas que se desprenden del texto de ejemplo.*

Obtenidos los lemas se buscan los mismos en la base de conocimiento y se calculan las medias de cada dimensión. En la tabla 4.4 se ven los resultados que arrojaría el sistema para la estimación de la frase.

Media agrado	Media activación	Media imaginabilidad
2.16	2.26	2.53

Tab. 4.4: *Valores estimados por el sistema para el texto de ejemplo.*

4.2. Evaluación del Sistema

La evaluación del sistema fue realizada sobre dos cuerpos de datos para ver en cada uno distintas características. El primer cuerpo de datos consistió en un conjunto de textos, y sirvió para contrastar la opinión de un grupo de personas con las predicciones del sistema. En segundo lugar se utilizaron opiniones de usuarios con respecto a productos para ver cuál era la efectividad del sistema al momento de predecir si las opiniones eran positivas o negativas.

4.2.1. Construcción de un Gold Standard

La primera evaluación contrastó las predicciones del sistema con las puntuaciones de un grupo de personas (a las que llamaremos Gold Standard). El experimento consistió en seleccionar un grupo de textos para evaluar en *agrado, activación e imaginabilidad*. Este grupo de textos estaba formado por oraciones y párrafos obtenidos de los cuerpos de datos de *Wikipedia* y *Los Cuentos*.

Se buscaron treinta oraciones que tuvieran más de 10 palabras. El total de las 30 oraciones estaba conformado por: 15 oraciones sacadas de *Wikipedia* y las quince restantes obtenidas de *Los Cuentos*. La condición de tener más de 10 palabras por oración, fue para asegurar un cubrimiento mayor al 0 % en cada oración.

Por otro lado, se tomaron diez párrafos de los cuales una mitad fue obtenido de *Wikipedia* y la otra se tomó de *Los Cuentos*, para ellos se buscó primero algunos

al azar y luego sobre esos se descartaron aquellos que no llegaban a desarrollar una idea. Este filtrado fue realizado debido a que en algunos textos un párrafo no tenía sentido por sí mismo, y era necesario leer también el anterior o el posterior. Los textos seleccionados se pueden ver en el apéndice B.

Los textos de test fueron entregados a cinco personas, de los cuales dos eran hombres y tres mujeres. Estas personas puntuaron cada oración y cada párrafo en *agrado, activación e imaginabilidad* utilizando la misma escala de valores de la tabla 3.1 (página 19). La principal diferencia con la tarea de puntuación era que para este caso la puntuación tenía otra granularidad, debido a que fue hecha sobre la oración o el párrafo, según el caso, en lugar de hacerlo sobre palabras. Luego, procedimos a comparar las puntuaciones de las personas con respecto a los resultados que arrojó el sistema.

4.2.2. Resultados

Para medir el nivel de acuerdo que hubo entre las votaciones del Gold Standard se calcularon los índices Kappa. En la tabla 4.5 se muestran los valores obtenidos para cada dimensión. El índice *unweighted* Kappa tiene un menor grado de acuerdo,

	Agrado	Activación	Imaginabilidad
unweighted Kappa	0.06	0.06	0.16
weighted Kappa	0.17	0.17	0.22

Tab. 4.5: *Índice Kappa para las distintas dimensiones.*

en todas las dimensiones, que el weighted Kappa. De todas maneras, si recordamos la escala de valores que tiene Kappa (tabla 3.5 de la página 27) se puede llegar a la conclusión que hubo un bajo grado de acuerdo entre los voluntarios. Esto muestra la extrema subjetividad de puntuar en las 3 dimensiones del estudio.

Para estimar el grado de relación entre los valores calculados por el sistema y las medias del Gold Standard, se corrieron tests de correlación de Pearson. En el apéndice C.1 se pueden ver los coeficientes ρ obtenidos para cada dimensión.

Sistema\Gold Standard	Agrado	Activación	Imaginabilidad
Agrado	0.59	0.15	-0.18
Activación	0.13	0.40	0.14
Imaginabilidad	0.16	0.19	0.07

Tab. 4.6: *Correlación entre las distintas dimensiones.*

De la tabla 4.6 los valores que no son parte de la diagonal, al ser menores que 0.2 muestran una débil relación, lo cual muestra la independencia entre las distintas dimensiones. El valor ρ de la dimensión agrado se ve el valor más alto, lo cual nos da indicios de que el sistema tiene una buena estimación del agrado si lo comparamos con el Gold Standard. Por otro lado se vio, en menor medida que el agrado, que la dimensión activación mostró valores significativos de correlación. La dimensión imaginabilidad mostró el valor más bajo de todos los ρ, lo cual resulta llamativo. Aquí se puede ver que no hubo casi relación alguna entre los valores de la estimación y el Gold Standard para la imaginabilidad. Las figuras 4.1, 4.2 y 4.3 muestran los gráficos contrastando los resultados del sistema con el Gold Standard para cada una de las dimensiones. Para facilitar la lectura de la tendencia en cada gráfico se agregó un modelo lineal para cada curva.

En el gráfico de la figura 4.1, se puede ver que hay un cierto grado de concordancia entre los resultados del sistema y el Gold Standard. Con concordancia nos referimos a que si bien varían los valores, las pendientes de ambas curvas son marcadamente descendentes. El motivo de la similitud está respaldado con los valores obtenidos en la correlación para la dimensión *agrado*.

Una explicación posible para la buena correlación es el hecho que entre los voluntarios se vio un acuerdo moderado (a través del índice Kappa). Es decir, si entre los voluntarios del Gold Standard hubo acuerdo significativo para el agrado de los textos, probablemente esos textos tenían palabras que en la base de conocimiento también tuvieron un buen acuerdo. Por lo tanto, al ponderar con el Sistema de Puntuación (usando otros voluntarios) esas palabras, hizo que mejorara la relación entre el Sistema y el Gold Standard.

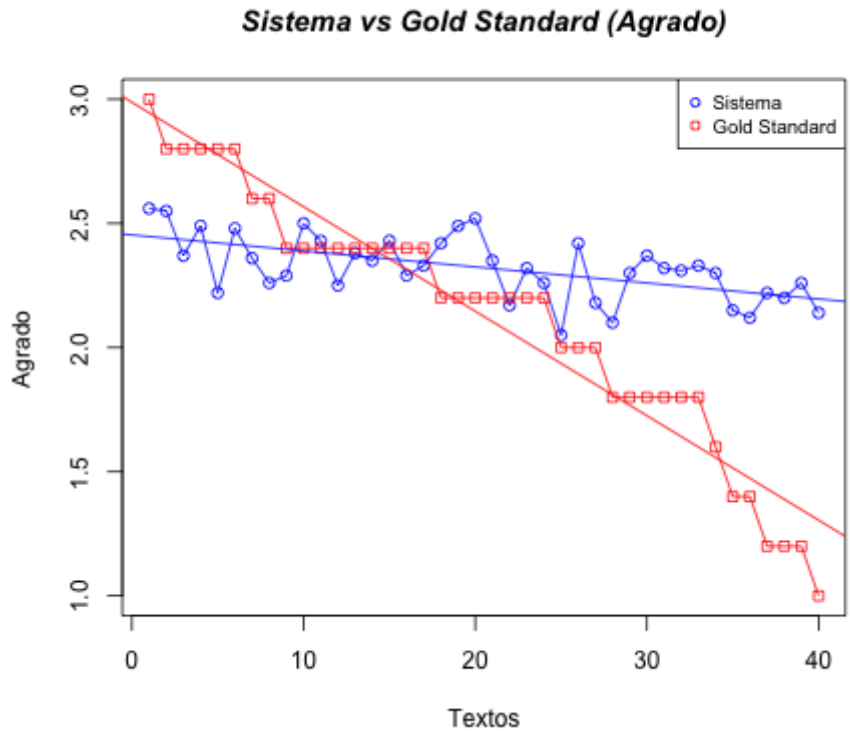

Fig. 4.1: *Comparación entre el Sistema y el Gold Standard para la variable Agrado.*

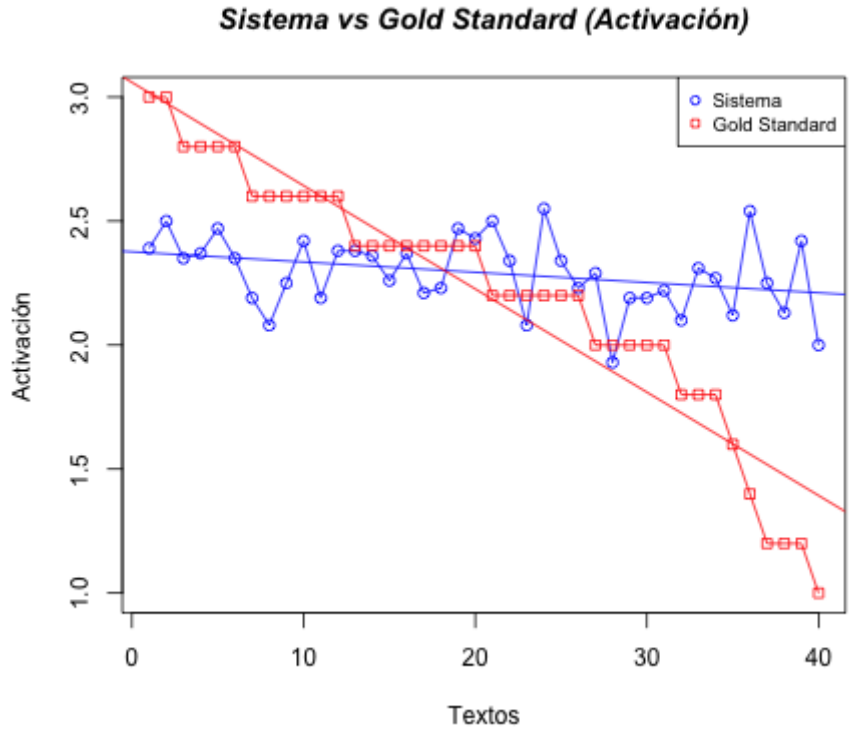

Fig. 4.2: *Comparación entre el Sistema y el Gold Standard para la variable Activación.*

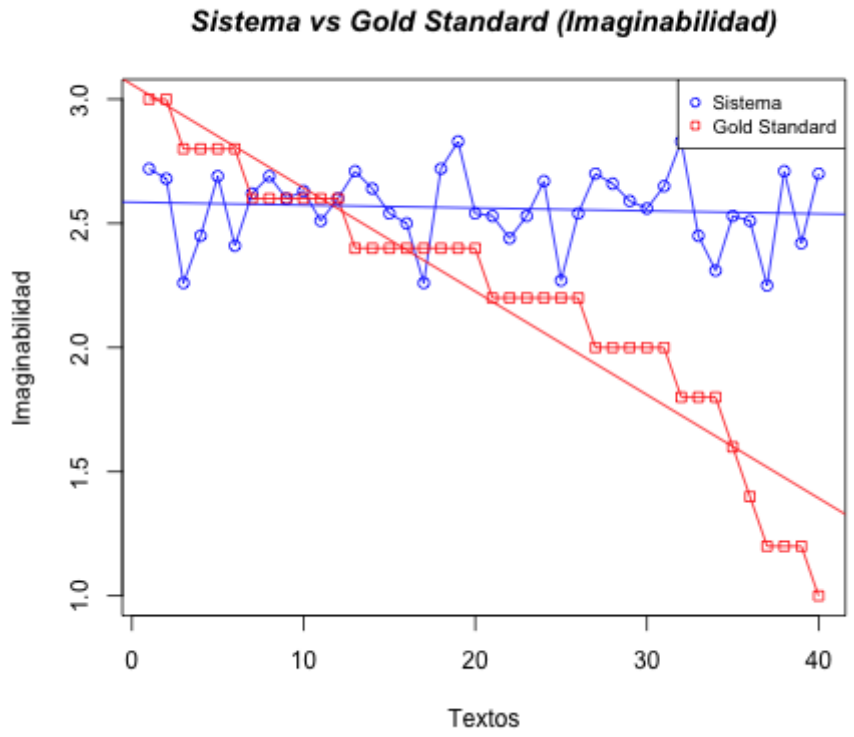

Fig. 4.3: *Comparación entre el Sistema y el Gold Standard para la variable Imaginabilidad.*

En las figuras 4.2 y 4.3 (página 33) el grado de concordancia es menor que el de la figura 4.1. Aquí se puede observar una tendencia suave en la dimensión *activación* del sistema, acompañando la pendiente del Gold Standard. Para el caso de *imaginabilidad* es más difícil de ver esa tendencia a través del modelo lineal.

De todas maneras, consideramos que estos resultados son muy buenos, dada la simpleza de la idea detrás del cálculo. La base de conocimiento es útil, en el sentido que captura información relevante de las tres dimensiones de estudio. El método de estimación sirve para el idioma español de la misma manera que el trabajo de Whissell para el inglés.

4.2.3. Performance vs. Cubrimiento

Por último, estudiamos la evolución del desempeño del sistema en función de la cantidad de palabras tenidas en cuenta de la base de conocimiento. En el gráfico de la figura 4.4 se puede ver la evolución que tuvo el sistema con respecto al cubrimiento. A medida que se fueron agregando 100 palabras puntuadas en el sistema había mayor cantidad de información que se podía ponderar sobre el Gold Standard. Inicialmente, con 250 palabras, se logró cubrir un poco más del 18 % de los textos y se finalizó por encima del 44 % cuando había 2500 palabras.

En cuanto a la correlación también se ve una progresión a medida que el sistema iba incorporando conocimiento a través de las palabras puntuadas. El comportamiento varió según la dimensión: la pendiente positiva de las curvas de agrado y activación es más marcada que la de imaginabilidad.[1] A cada curva de correlación se le ajustó un modelo logarítmico. Cuanto más cerca de 1 está R^2, el valor más se parece a la función que se esta ajustando, en este caso logarítmica.

Además, el gráfico de la figura 4.4 nos muestra una tendencia positiva, en cuanto a que, a mayor conocimiento en la base, mejor es la correlación. Esto último determina un acercamiento logarítmico entre la forma en que califica un ser humano un texto y el funcionamiento del sistema. La importancia de este acercamiento logarítmico viene a consecuencia, que para hacer una mejora sustancial al sistema no va a alcanzar con puntuar más palabras, dado que la curva muestra que a mayor cantidad de palabras más lenta es la mejora del sistema. Por lo tanto, será necesario abordar el problema desde otro ángulo utilizando la información de la base de conocimiento de otra manera.

[1] En el apéndice C se pueden ver los valores obtenidos de la correlación del Sistema de Estimación vs. el Gold Standard.

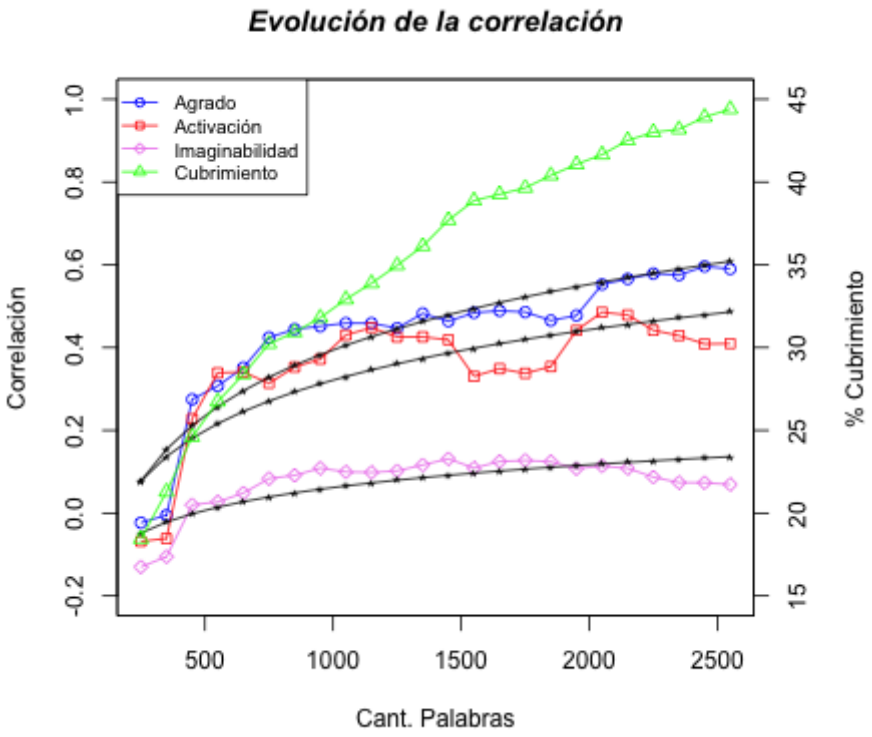

Fig. 4.4: *Evolución de la correlación entre el Sistema de Estimación y el Gold Standard. El eje Y de la izquierda muestra la escala de valores de correlación. El eje Y de la derecha es la escala de valores para el porcentaje de cubrimiento logrado por el sistema. La línea color verde marca el porcentaje de cubrimiento; las otras tres curvas corresponden a los valores de correlación entre el sistema y el Gold Standard en cada dimensión. Luego, la curva que se aproxima a la correlación de cada dimensión representa la función de ajuste logarítmico. Los R^2 de agrado, activación e imaginabilidad fueron: 0.900, 0.717 y 0.681 respectivamente.*

4.3. Ciao.es

Para la segunda evaluación que se corrió sobre el sistema, se tomó como cuerpo de datos opiniones sobre productos. Estas opiniones fueron tomadas del sitio *Ciao*[2] y corresponden al mismo cuerpo utilizado en el trabajo de Brooke[3], donde se usaron los recursos disponibles de análisis de sentimiento (*sentiment analysis*) en inglés para puntuar textos en español por medio de una traducción automática. Los productos de los cuales se obtuvo opiniones fueron: coches, hoteles, lavadoras, libros, teléfonos celulares, música, ordenadores y películas.

[2] http://ciao.es

Originalmente, el sitio *Ciao* permite al usuario adosar a la opinión una calificación, la cual tiene una escala de 1 a 5 estrellas. Aquí supusimos que hay una fuerte relación entre la calificación que aportó el usuario a su opinión y la cuantificación del agrado. Creemos que a mayor cantidad de estrellas, más agradable le resultó el producto y a menor, más disconforme estuvo, por lo tanto menos agradable le parece. Dada esa suposición, decidimos que aquellas opiniones que tenían 1 ó 2 estrellas fueran catalogadas como "desagradables" (o negativas) y aquellas que tenían 4 ó 5 fueran "agradables" (o positivas). Para el experimento se omitieron aquellas opiniones que estaban calificadas con 3 estrellas, dado que se las consideró neutras en cuanto a su nivel de agrado.

En esta ocasión se buscó evaluar el sistema en cuanto a los aciertos de si una opinión es agradable o no. Dicho de otra manera, queremos ver qué porcentajes de aciertos tiene el sistema considerando sólo la dimensión agrado. Para ello se implementó un 5-Fold Cross Validation tomando siempre el 80 % de las opiniones como datos de entrenamiento al sistema y el 20 % restante como datos de test. En la figura 4.5 se muestra como se utilizaban los datos de entrenamiento para hacer la predicción sobre el 20 % restante.

Usamos nuestro Sistema de Estimación para predecir si una opinión es positiva o negativa de la siguiente manera. Del 80 % de los datos de entrenamiento se calculó la media M de agrado para todas las oraciones y párrafos. Luego, si la media de agrado de una opinión del 20 % de test era mayor a M, la predicción era positiva, o bien negativa en caso contrario. En el gráfico de la figura 4.6 se muestran los porcentajes de aciertos de las predicciones y el porcentaje de cubrimiento de los textos. Para las predicciones que se realizaron figura el *porcentaje total*, que muestra del total de opiniones a predecir cuántas se predijeron correctamente. De igual manera que se hizo con los resultados del Gold Standard, se ajustaron las curvas con una función logarítmica para marcar la tendencia y ver con R^2 que tan bien se ajustan los datos obtenidos.

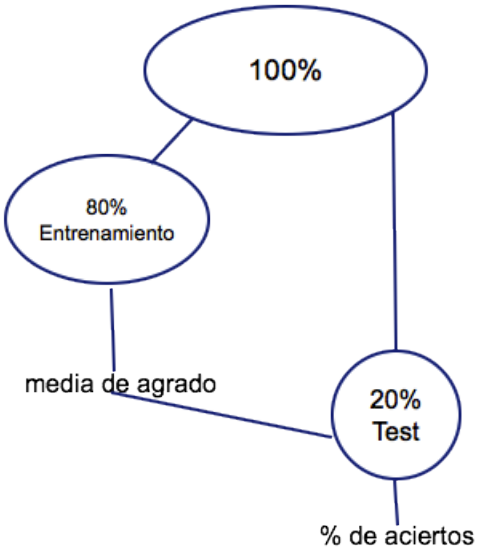

Fig. 4.5: *5-Fold Cross Validation. Con el 80 % de entrenamiento se calculaba la media de agrado y con el 20 % restante se verificaba si cada opinión estaba por encima o debajo de esa media. En los casos que la media de la opinión estaba por debajo de la obtenida en el entrenamiento y la opinión tenia 1 ó 2 estrellas la opinión era considerada un acierto para el caso negativo. En cambio, si la media de la opinión estaba por encima de la media obtenida con el 80 % de entrenamiento y tenia 4 ó 5 estrellas era un acierto para el caso positivo. Esto se realizaba variando el 80 % y el 20 % de las muestras 5 veces. Los porcentajes finales de aciertos eran el promedio de esas 5 ejecuciones.*

Se nota un incremento en el porcentaje de aciertos a medida que se incrementan las palabras. Aquí se debe interpretar como valor inicial del eje Y de la izquierda el 50 %, dado que en el peor de los casos el sistema funcionaría de la misma manera que arrojar una moneda y acertar qué cara va a salir. Dicho de otra manera, al ser una decisión binaria el porcentaje de aciertos debe ser mejor al 50 % para ser considerado mejor que el azar.

Del gráfico anterior destacamos como, con un bajo cubrimiento de poco más del 44 % se puede ver picos de aciertos de hasta 66 %, lo cual representa una mejora de 16 puntos sobre la base de 50 %. Este gráfico nos da indicios de que al incrementar la cantidad de palabras sobre el cuerpo de datos se obtendría una mejora en el cubrimiento y en el porcentaje de aciertos.

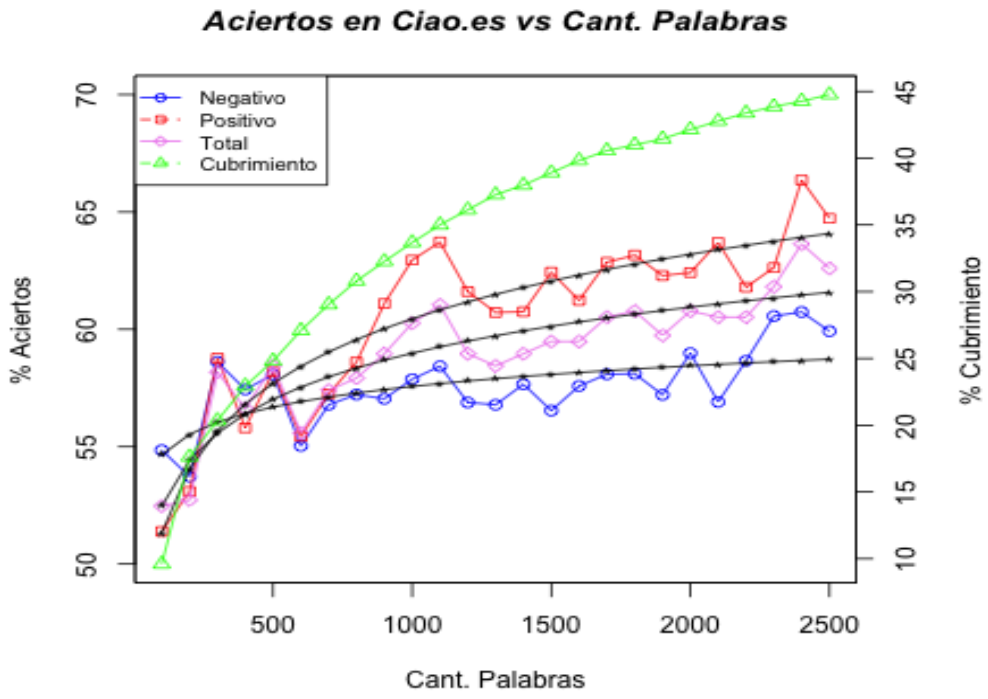

Fig. 4.6: *En el eje Y de la izquierda se puede ver la escala de porcentajes de aciertos positivos, negativos y totales. En el eje Y de la derecha se puede ver la escala de valores relativa al porcentaje de cubrimiento sobre los textos. La linea verde de cubrimiento debe interpretase con el eje Y de la derecha, mientras que las las otras curvas corresponden con el eje Y de la izquierda. Luego, la curva que se ajusta a las curvas positiva, negativa y totales, representa la función de ajuste logarítmico. Los R^2 para positivo, negativo y total fueron: 0.824, 0.427 y 0.802 respectivamente*

5. CONCLUSIONES

En esta tesis se presentó un sistema mediante el cual es posible estimar valores de *agrado, activación e imaginabilidad* sobre un texto. Esta herramienta resulta una base sobre la cual se pueden implementar varias mejoras, e incluso puede ser utilizada como sistema *baseline* para comparar el desempeño de otros sistemas. Si bien la idea detrás del sistema resulta simple, creemos que los resultados han sido satisfactorios.

En cuanto a la evolución del sistema los resultados han sido concluyentes: a mayor cantidad de palabras puntuadas el resultado más se aproxima a la opinión de las personas. Sin embargo, la relación antes mencionada mostró un crecimiento de tendencia logarítmica, con lo cual sería necesario utilizar otro método de estimación para hacer una mejora sustancial.

Durante las distintas evaluaciones notamos que los resultados del sistema para las tres dimensiones fueron, en general, con resultados mayores a 2. Si recordamos la tabla de valores con los que se puntuaron las palabras (tabla 3.1 de la página 19), podemos decir que tuvieron una connotación entre positiva (3) y media (2) para cada dimensión. Este resultado nos lleva a pensar que para el sistema el ser humano tiende a expresarse de manera positiva. En otras palabras, el sistema mostró un balance de tendencia positiva sobre los textos cuando se consideraron votaciones por palabras. En consecuencia, en el caso de la dimensión agrado, si un texto resultaba calificado por el sistema con un valor menor que dos mostraba que había un desagrado bastante marcado.

Teniendo en cuenta los buenos resultados del sistema, también es necesario remarcar que el lenguaje natural tiene ciertas características que son difíciles de detectar. La comunicación de mensajes con ironías, sarcasmos, doble sentido, entre otros, engañan al sistema, dado que éste no detecta las características en cuestión y pierde

la verdadera intención del mensaje.

Como trabajo de mejora creemos que se puede agregar bastante a este sistema *baseline*. La utilización de palabras relativas al contexto pueden significar una mejora en los valores de las dimensiones. Por ejemplo, palabras como *muy*, *re*, *extremadamente*, etc. podrían aportar un coeficiente por el cual se multiplica la palabra que le sigue.

Otra mejora al sistema podría ser multiplicar cada palabra por un coeficiente en función del grado de acuerdo que hubo para esa palabra. Es decir, en aquellos casos en que todos los voluntarios hayan estado de acuerdo en una dimensión, tomar como coeficiente el valor 1 para multiplicar la palabra. En el caso que hubo desacuerdo, multiplicar la palabra por el índice Kappa, o por el porcentaje de acuerdo, por ejemplo. De este modo, se le daría mayor importancia a palabras con un afecto más claro, por sobre aquellas más ambiguas.

Los cuerpos de datos, tanto *Wikipedia* como *Los Cuentos*, fueron los que determinaron el listado de palabras que fueron puntuadas. La selección de las palabras no fue pensada para un dominio en particular. Creemos que esto pudo hacer el listado más extenso de lo necesario. Por lo tanto, para aquellos que deseen hacer el sistema aplicado a un dominio específico, sugerimos la búsqueda de cuerpos de datos acordes a los dominios sobre los cuales se va a aplicar el sistema. Una estrategia que creemos que puede ser abordada es, primero buscar un porcentaje de cubrimiento que satisfaga las expectativas y luego puntuar sobre ese listado de palabras. Con esto sería posible encontrar un listado de palabras que sea más chico que el utilizado e incluso que mejore el porcentaje de cubrimiento.

Por último, cabe aclarar que los porcentajes de cubrimiento deben ser interpretados de manera relativa sobre los textos, debido a que siempre va a haber un conjunto de palabras que no van a aportar valor en las dimensiones. La búsqueda del cubrimiento del 100 % no sería necesaria dado que siempre va a haber ruido que

no es necesario puntuar, tal es el caso de las fechas, nombres, apellidos, etc. Para llegar a una conclusión definitiva en cuanto al porcentaje óptimo de cubrimiento, creemos que cada dominio debe ser analizado con baterías de test hasta que se logre el cubrimiento deseado para el dominio.

Apéndice

A. PALABRAS Y SÍMBOLOS EXCLUIDOS

A.1. Símbolos excluidos

A continuación se muestran los signos eliminados durante el proceso de selección de palabras, que se encuentra detallado en la sección 2.2 del capítulo 2. Los signos que se eliminaron fueron:

!	“	#	$	%	&	’
(	)	*	+	”	—	–
,	-	.	/	:	;	¡
¡	¿	=	?	@	[	]
<	>	\	|	ˆ	_	‘
{	}	˜	”	»	«	_

Dígitos y carácteres de espacios en blanco como tabs, enters, etc. también fueron eliminados

A.2. Palabras eliminadas manualmente

En el proceso manual de filtrado, explicado en la sección 2.2.1 del capítulo 2, se comenta que ciertas palabras fueron eliminadas. A continuación se puede ver las palabras eliminadas ordenadas alfabéticamente:

abu achával action adobe africana africanas africano africanos afroamericana afroamericano afroamericanos again age ages aguilera airlines airways alejandro alemana alemanas alemanes alemán alexandre all allan allöder almagro alp alsacia alvear america american americas amerindia amerindias amerindio amerindios américas and andrade ann antártica antárticas antártico antárticos apple apurímac ara ardenas argentinas argentino argentinos asiática asiáticas asiático asiáticos association athletic atp august australiana australianas australiano australianos auvernia award baby back bad balcanes band basilea batista battle baviera beat belga belgas ben bermúdez berna best big billboard birmania bob bobs bogotá bolonia bolívar bolívares book boom borbón bosco box boxes boy boys brasileña brasileñas brasileño brasileños bridge bridges british británica británicas británico británicos bros bruselas bucarest burdeos business calderón camacho canadiense canadienses cantábrica cantábrico capitel capiteles car carlista carlistas carrasco castellón castro catalana catalanas catalanes catalán celta celtas celtics centroamérica cerdeña championship channel chart checa checo checos chilena chilenas chileno chilenos chinas chinita chinos chris church ciento cientos cities city codice codigo col cole college colocolo colombiana colombianas colombiano colombianos communauté company comédie corporation corpus countries country county cover covers cpu cpus creta cretas croata croatas cubana cubanas cubano cubanos cup cádiz cárdenas céline córdoba dark days dead deep dehesa dehesas dehesilla dei della der die digimon doc dos drive dublín duero dylan east edge eeuu emiliano encyclopédie end enríquez escocia español española españolas españoles estadounidense estadounidenses estatúder estrada eta eugenio eurocopa europe europea europeas europeo europeos euskera fbi ferro filipo filme filmes fire first folk for foral forales fort foundation four francesa francesas franceses francisca francés free freud friedrich from fráncfort gabriela galia game garcía gastón gerona gijón ginebra giorgio girl girls godoy google goëland gral grammy grand great grey greys griega griegas griego griegos group guadalquivir guipúzcoa génova hab habkm hard hardcore hawái head heart heavy heinrich henar henares here hip hispana hispanas hispano hispanoamérica hispanos hispánica hispánico hispánicos hit hits hms hop hot ian ibérica ibéricas ibérico ibéricos iii inc indie ine inglesa inglesas ingleses inglés intel inter interamericana interamericano international irlandesa irlandesas irlandeses irlandés isidro islam islámica islámicas islámico islámicos israelí israelíes italiana italianas italianita italiano italianos its itunes james jan janes japonesa japonesas japoneses japonés jason jaén jedi jerusalén johan journal judas judía judías judío judíos jules juniors junín just kings knéset kong krai kurt ladies lady land lands last latinoamericana latinoamericanas latinoamericano latinoamericanos let libra libras librilla life line little live long lope lord lost luca luigi lérida lópez mac machine madame major marcel marge maria marines mario marín match matches maya mayas mejía merindad merindades mexicana mexicanas mexicano mexicanos michoacán mil milanito millar millares milán misuri montero monteros moscú msnm muhammad málaga múnich nasa nathan network never new nick nicole nilo norteamericana norteamericanas norteamericano norteamericanos nápoles obregón ocampo office one open opus orchestra oregón orinoco osorio otan otero out over pachacámac panamericana panamericano panamericanos papúa paraguaya paraguayas paraguayo paraguayos party pat pekín people per pere persa persas peruana peruanas peruano peruanos peseta pesetas peter piamonte pib pietro pinar pirineo pirineos play playoff playoffs plus poe pompeyo port portuguesas portugueses portugués potosí precolombina precolombinas precolombino precolombinos prehispánica prehispánicas prehispánico prehispánicos press pri pro psoe pta ptas pts pía pías pío píos quilmes quintana quintanas quintanilla quintanillas quiroga race racing raión records reich remake research road rob romana romanas romano romanos rumana rumanas rumania rumano rumanos run rusa rusas rusos rímac saavedra saboya sam san sarmiento sarmientos school science senior serbias serbio serbios serranía serranías set sets sexta sexto side single singles sir sláine småland society solís songs sonic sony sound soviética soviéticas soviético soviéticos space special spiderman sport sporting sports sri stan stars state stones studios sudamericana sudamericanas sudamericano sudamericanos sueca suecas sueco suecos suite suites summer system systems sánchez taifa taifas tarragona team tebas tecámac ted that the three tolosa tom torreones torreón township tragédie transilvania tucumán uci uefa und under unesco uruguaya uruguayas uruguayo uruguayos uss val valparaíso venezolana venezolanas venezolano venezolanos vergara vii viii villareal villegas villena vol von vélez vólost way what when william wrestling xbox xii xiii xix xvi xvii xviii xxi you your youtube zelanda zúrich ábside ábsides álava árabe árabes ártica árticas ártico árticos åland æsir école évangéline óblast öland

A.3. Adverbios eliminados

El proceso de eliminación de palabras, sección 2.2.1 del capítulo 2, se comenta que entre las palabras que se eliminaron figuran algunos adverbios. A continuación se detallan los adverbios eliminados.

Lugar	Tiempo	Modo	Cantidad	Afirmación	Negación	Duda
Aquí	Ahora	Bien	Más	Sí	No	Quizás
Ahí	Luego	Mal	Menos	También	Tampoco	Tal vez
Allí	Después	Así	Poco	Cierto	Nunca	Acaso
Cerca	Ayer	Aprisa	Mucho			
Lejos	Hoy	Deprisa	Bastante			
Arriba	Mañana	Despacio	Muy			
Abajo	Entonces		Casi			
Alrededor	Pronto					
Dentro	Tarde					
Fuera	Siempre					

A.4. Conectores eliminados

En el capítulo 2 se mencionan aquellos tipos de palabras que no aportan valor al estudio en la variables *agrado, activación e imaginabilidad*. De esas palabras en la sección 2.2.1 se habla que algunos conectores fueron eliminados. A continuación se muestran los conectores eliminados ordenados alfabéticamente.

a causa de, a la vez, a menos que, a menudo a no ser que, actualmente, además, ahí, al comienzo, al lado de, al mismo tiempo, al principio, allí, anteriormente, antes bien, antes de que, antes, aparte de, aquí, asimismo, así pues, así que, aun cuando, aunque, aunque, bien, con el paso del tiempo, con que, con referencia a, conque, contrariamente, cuando, dado que, de ahí que, de la misma manera, de la misma manera, de manera que, de manera que, de modo que, de todas maneras, debido a que, del mismo modo, delante de, dicho de otra manera, dicho de otro modo, en aquel, en cambio, en caso de, en conclusión, en consecuencia, en cuanto a, en definitiva, en ese, en este, en lo que concierne a, en medio de, en otras palabras, en otro orden de cosas, en pocas palabras, en primer lugar, en resumen, en segundo lugar, en síntesis, encima de, entre otras, entre otras, es decir, excepto, finalmente, finalmente, fue entonces cuando, había una vez, hace tiempo, igualmente, incluso, inicialmente, inversamente, mas aun, mas, mejor dicho, menos, mientras tanto, mientras, más precisamente, más tarde, no obstante, o sea, otro caso más, para concluir, paralelamente, pero, pese a, por arriba de, por consiguiente, por debajo de, por el contrario, por ello, por ende, por eso, por esta razón, por este motivo, por esto, por lo cual, por lo dicho, por lo que sigue, por lo que, por lo tanto, por más que, por otra parte, por otro lado, por una parte, por último, porque, posteriormente, previamente, pues bien, pues, que, respecto de, resumiendo, salvo, si bien, si no que, siempre que, simultáneamente, sin embargo, sino, sintetizando, sumado a, también, tiempo atrás, ya que, ya sea.

B. TEXTOS PUNTUADOS POR VOLUNTARIOS

Como se menciona en el capítulo 4 y sección 4.2.1, a un grupo de 5 voluntarios se le entregaron 40 textos para que los midan en las dimensiones del estudio (ver tabla 3.1 en la página 19). Esos resultados formaron el Gold Standard. A continuación las 30 oraciones y luego los 10 párrafos que fueron entregados para puntuar.

1. En su largo descenso desde la meseta de Anatolia hasta Tarsos, corría a través del desfiladero entre las paredes de roca llamadas las puertas Cilicias.
2. El riesgo soberano, que se asimila al riesgo país, es el término usual para referirse a la calificación de riesgo dada a un Estado Soberano.
3. Por la noche, en las instalaciones de la Sociedad Comercio e Industria, se llevó a cabo un gran banquete en honor de las autoridades que visitaron la ciudad.
4. Más tarde encontraría su fórmula ideal de la novela, al insertar aquel costumbrismo en una visión enamorada del paisaje y de las gentes de la montaña, con sus pasiones y su lenguaje característico.
5. Los ríos que nacen en estas sierras se encajan formando algunas gargantas excavadas en granito, desprovistas de acumulaciones en su fondo, en las que se pueden observar algunos saltos de agua.
6. La Terapia Familiar es una disciplina terapéutica que aborda la intervención y el tratamiento de la familia en su conjunto, y no de un miembro individual.
7. Con los viajes de Marco Polo y el descubrimiento del continente americano se diversificó la música de culto religioso.
8. Los autores presumen que la inhibición latente baja puede ser positiva cuando está combinada con alta inteligencia y buena memoria de trabajo - la capacidad de pensar en muchas cosas simultáneamente - pero negativa en individuos con escasa capacidad intelectual.
9. El sueño de la organización era crear un periódico que fuera una voz para las organizaciones sin fines de lucro (término al que también se le refiere abreviadamente como “OSFL”).
10. El protagonista es un chico que acaba de cumplir 14 años, y por ello le es regalada la legendaria espada que su padre usase en vida.
11. Un reemplazo sintético del corazón es una de las aspiraciones más anheladas de la medicina moderna.
12. Para compensar la fuerza de gravedad y la fuerza magnética contrapuesta posee anillos a modo de contrapesos que deben colocarse pacientemente hasta lograr un equilibrio determinado.
13. El fin justifica los medios es una frase atribuida a Maquiavelo y significa que gobernantes y otros poderes han de estar por encima de la ética y la moral dominante para conseguir sus objetivos o llevar a cabo sus planes.
14. Era el nieto preferido de Fernando el Católico, y fue educado a la española por su abuelo; en un principio fue investido como regente, en un testamento dado en 1512, hasta la llegada de Carlos a España, pero el anciano rey lo revocó antes de morir favoreciendo a su hermano Carlos, educado en Borgoña.
15. En el jardín se ubica un Instituto de investigación y diversos departamentos de seguimiento de los cultivos de plantas y su desarrollo, de Botánica, Floricultura así como una Biblioteca.
16. En homenaje a tantos compañeros que entregaron parte de su vida, y algunos la vida misma, por el retorno a la democracia en mi país.

17. En el televisor se podía ver a un perro parlante quejándose de ser muy pequeño y no poder alcanzar los estantes más altos de una cocina; Mario reía cada vez que el perro intentaba un inverosímil plan para subir.
18. De su apariencia física no recuerdo mucho, ella era mas bien normal, tez clara ojos pardos, mas bien café, labios pequeños y muy rojos.
19. Cada mañana se sentaba sobre ese sillón tan bajo, pegadito a la pared y al lado de su trinche colmado de recuerdos de cristal.
20. Yo no soy el sueño que sueña; yo no soy el reflejo de una imagen en un cristal; a mí no me aniquila la cerrazón de una conciencia o de toda conciencia posible.
21. El demonio de negro vestía, con un escudo calcinado que en eras anteriores emitía dulce brillo, ahora opaco, la espada aún en la vaina roída por las sombras y una lanza que amenazaba con su aguda punta.
22. El observó de manera atenta y nostálgica la figura de su sombra dibujada en el muro de la vieja casa de adobe, las rosas secas plasmaban su silueta en la pared que era pintada por los colores sepia del sol agonizante.
23. No había dormido bien esta noche porque la tarde anterior luchando con otro pedigüeño, me había quedado con una porción de cartón demasiado corta para la medida de mis piernas, así que había estado metiendo mis pies, sin calcetines, entre las bolsas de basura que habían allí tiradas.
24. Es un absceso con pus hedionda, ósea, muchachos: una herida fétida que cualquier persona con nariz puede oler, ver, descubrir, incluso recetar.
25. Es una lástima, porque soy joven, libre, puedo ir y volver, besarme con alguien hasta que me duelan los labios, y tantas cosas más ...
26. Solo y Sentir, son dos palabras, una tiene algo de soledad, pero la otra acompaña con lo del corazón, y muy importante el sentir.
27. Es indiscutible que jamás podré olvidar, si tan solo me hubiese advertido de lo fatídico que resultaría todo para mi, no estaría ahora deshojando una margarita, cual adolescente ridículo y encaprichado por una noviecita.
28. Alguna vez me gustaría que existiera la conciencia colectiva como una gran masa que operase a voluntad.
29. Era un viernes, y se acercaba la hora de cerrar todo, Roberto ansiaba volver a casa, a darle cariños a su perro, que era su única compañía.
30. Todo empezó un frío día de otoño cuando yo miraba fijamente a través de la ventana del salón.

1. El salchichón de chocolate tiene la misma forma cilíndrica como un salami, pero no es un producto de carne. Se sirve en rodajas en secciones transversales, el marrón oscuro del chocolate sustituye la carne roja, y las partes pequeñas rotas de galletas sustituyen la grasa del salami. Algunas variedades también contienen nueces picadas, como almendras o hazelnuts. También pueden hacerse en bolas pequeñas.
2. Primero fueron excavadas dos tumbas que estaban cubiertas por el gran montículo artificial. La primera tenía una pequeña estructura rectangular; tres de sus paredes estaban pintadas con la representación del rapto de Perséfone y Plutón conduciendo una cuadriga de caballos blancos, al mismo tiempo que sujeta a Perséfone. También en la escena está representada Deméter y tres Parcas. Las figuras son de gran tamaño, sobre todo la de Plutón. Son unas pinturas de gran fuerza y calidad y los eruditos creen que se deben a un gran maestro que conocía la perspectiva y que tenía mucha facilidad para el dibujo y para el color. Esta tumba había sido saqueada en la antigüedad. La segunda era una gran tumba con bóveda de cañón, que fue identificada como la tumba de Filipo.
3. A temperatura ambiente los lípidos pueden ser sólidos, a lo que se le denomina "grasas concretas"(o simplemente grasas), si a temperatura ambiente los lípidos se presentan como líquidos se denominan .ªceites". El aceite de oliva es líquido a temperatura ambiente. El aceite de oliva posee algunas propiedades características de todos los aceites vegetales, así como otras particulares de la aceituna. Una de las principales propiedades se deriva de su alto contenido de ácido oleico (llegando de media a un 75 %) Las propiedades dependerán en gran medida de la variedad de aceituna empleada, de la forma en la que se procesó el aceite y de los procedimientos de almacenado.

4. El diseño universal es parte esencial de la estrategia para conseguir una sociedad en la que todas las personas pueden participar. Un modelo de sociedad que se está redefiniendo tomando como base la inclusión de todos y que deriva, en gran medida, de la reflexión acerca del modo que la sociedad quiere acoger a la persona en toda su diversidad. Un ingrediente de esta diversidad es la discapacidad. En este modelo social, se priman los valores de la igualdad de oportunidades y el respeto de los derechos de todos.

5. Organizaciones como el Fondo Mundial para la Naturaleza, fueron creadas con el objeto de preservar a las especies de la extinción. Algunos países también han intentado evitar la destrucción de hábitats, la sobreexplotación de los suelos, y la polución, mediante la promulgación de leyes y decretos. Aunque muchas extinciones provocadas por los seres humanos podrían ser calificadas como accidentales, también hay otras que se han concretado de manera deliberada, como la destrucción de algunos virus peligrosos, o la extirpación de especies problemáticas.

6. A los pocos días, el perrito recorría la modesta covacha, husmeando por todos los rincones. Su estirpe era incierta porque tenía las características de un fox terrier, pero, visto de otro ángulo, parecía más bien un perro de aguas y hasta podría confundirse con un labrador. En buenas cuentas, el can no pertenecía ni a una ni a otra raza, sino que era sólo un quiltrito de poca monta, un perrito en el cual se vislumbraba, sí, una inteligencia manifiesta que se reflejaba en sus enormes ojos pardos.

7. Camino por el cielo, y observo como los pájaros revolotean sobre mis manos. Las sacudo y migas de pan se elevan hacia las aves, miro un rayo de sol que se acerca piadoso, entibia mi rostro saludándolo como una niña saluda al mar. Mis pies se mueven sin buscar sustento, y siento que asiendo a cada paso, El universo me congratula expectante y doy la bienvenida a este día como ningún otro. Entro a una nube por su puerta, se abre de par en par, La lluvia me da la bienvenida a su hogar entregándome su fresca sonrisa. Veo las estrellas que se esconden en el techo del mundo, una nueva vida sin vida, ayer muero para encontrar lo que siempre busqué.

8. Miguel saca una botella de vino tinto, sirve en algunos vasos, menos a los niños y agradece a Alamiro por aceptar la invitación. Alamiro, a su vez, le da las gracias por recibirlo. Cuando acaba el almuerzo, los hijos salen disparados con distintos rumbos, cada uno con una manzana. José y Alamiro quedan solos, es como si hubiese habido un acuerdo. Alamiro se sienta más cerca de Miguel.

9. Las personas que van dentro del ómnibus son apenas distinguibles unas de otras, no hay iluminación interior y en el cielo hace años que se ve el sol. Se puede ver el polvo que se acumula en las ventanas y en los asientos, y también en las personas, que aunque menos visible en su cuerpo, es un polvo más visible en sus rostros. Por la ventana derecha del ómnibus se puede ver como los de negro golpean a un hombre tirado en el piso. Se ve que se resistió. Pobre desgraciado, por suerte no soy yo piensa cada uno de los que va dentro del ómnibus, antes o después, sintiéndose más o menos culpables al hacerlo.

10. Cuando ella tomaba vino en copas transparentes y fumaba, él se ponía incómodo y resoplando afirmaba que eso era signo de libertinaje, todos perdían la voz cuando yo preguntaba qué quería decir esa palabra. Creo que el primo Alberto la envidiaba sólo porque ella había conocido otros lugares, vivía sola y sabía tocar el piano.

C. VALORES CORRESPONDIENTES A LA CORRELACIÓN

En la sección 4.2.3 del capítulo 4, se muestra el comportamiento de la correlación que hay entre el Sistema y el Gold Standard. A continuación se muestran los valores obtenidos de la correlación en las 3 dimensiones del estudio. Con estos valores también se formó el gráfico 4.4.

Cant. Palabras	Agrado	Activación	Imaginabilidad
250	-0.023	-0.068	-0.130
350	-0.004	-0.06	-0.105
450	0.274	0.227	0.019
550	0.307	0.338	0.027
650	0.351	0.340	0.048
750	0.424	0.313	0.083
850	0.444	0.352	0.091
950	0.451	0.371	0.108
1050	0.458	0.430	0.099
1150	0.458	0.448	0.098
1250	0.446	0.425	0.102
1350	0.482	0.425	0.116
1450	0.463	0.418	0.131
1550	0.484	0.330	0.108
1650	0.488	0.349	0.124
1750	0.485	0.337	0.127
1850	0.465	0.354	0.124
1950	0.477	0.441	0.107
2050	0.552	0.485	0.113
2150	0.565	0.478	0.108
2250	0.578	0.441	0.087
2350	0.575	0.428	0.073
2450	0.595	0.408	0.073
2550	0.589	0.409	0.069

Tab. C.1: *Correlación entre las distintas dimensiones*

Bibliografía

[1] A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12, 2008.

[2] J. Altarriba. Cognitive approaches to the study of emotion-laden and emotion words in monolingual and bilingual memory. *Bilingual Education and Bilingualism*, 56:232, 2006.

[3] J. Brooke, M. Tofiloski, and M. Taboada. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria*, pages 50–54, 2009.

[4] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[5] J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[6] W.N. Francis and H. Kucera. Brown corpus. *Department of Linguistics, Brown University, Providence, Rhode Island*, 1, 1964.

[7] J.R. Gray, T.S. Braver, and M.E. Raichle. Integration of emotion and cognition in the lateral prefrontal cortex. *Proceedings of the National Academy of Sciences*, 99(6):4115, 2002.

[8] Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. Freeling 2.1: Five years of open-source language processing tools. In *LREC*, 2010.

[9] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*, volume 2010, 2010.

[10] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[11] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[12] C. Whissell. The dictionary of affect in language. *Emotion: Theory, research, and experience*, 4:113–131, 1989.

[13] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics, 2005.

[14] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427–434. IEEE, 2003.