Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 65-72 recibido 14-01-12 revisado 10-03-12 aceptado 18-03-12

# Detectando la mentira en lenguaje escrito

## *Detecting deception in written language*

**Ángela Almela Sánchez-Lafuente (*), Rafael Valencia-García (‡),**
**Pascual Cantos Gómez (*)**
**Universidad de Murcia**
(*) Departamento de Filología Inglesa
(‡) Departamento de Informática y Sistemas
30071, Murcia (Spain)
{angelalm, valencia, pcantos}@um.es

**Resumen:** La mentira en el lenguaje se ha estudiado desde la perspectiva de varias disciplinas, siendo la más reciente la minería de opiniones. En este contexto, el presente estudio persigue explorar los rasgos sintomáticos de la mentira en lengua escrita en español, lo cual no ha sido aún investigado. Para ello, hemos desarrollado un marco de trabajo basado en un clasificador de máquinas de soporte vectorial (SVM) aplicado a un corpus *ad hoc* de opiniones. Hemos usado las categorías psicolingüísticas definidas en LIWC (Pennebaker, Francis y Booth, 2001) a través de sus cuatro dimensiones fundamentales para entrenar el algoritmo. Los resultados del experimento muestran que es posible separar los textos en lengua española de acuerdo con su condición de verdad, siendo las dos primeras dimensiones, procesos lingüísticos y psicológicos, las más relevantes para la consecución de tal objetivo.
**Palabras clave:** Lenguaje de la mentira, minería de opiniones, extracción de características, máquinas de soporte vectorial, LIWC.

**Abstract:** Deception in language has been studied from the perspective of several disciplines, being the most recent one opinion mining. Within this framework, the present study attempts to explore deception cues in written Spanish, which, to the best of our knowledge, has not been investigated yet. For our purposes, we have developed a framework based on a classifier using a Support Vector Machine (SVM) in order to detect deception in an *ad hoc* opinion corpus. We have used the psycholinguistic categories defined in LIWC (Pennebaker, Francis and Booth, 2001) through its four broad dimensions for the subsequent training of the abovementioned classifier. The findings reveal that truthful and deceptive texts in Spanish are indeed separable, being the two first dimensions, linguistic and psychological processes, the most relevant ones for fulfilling our aim.
**Keywords:** Deception language, opinion mining, feature extraction, support vector machine, LIWC.

## *1 Introducción*

En el contexto de la comunicación humana, la mentira juega un papel activo. A este respecto, DePaulo et al. (1996) afirman que se suele contar de una a dos mentiras al día, ya sea en lenguaje oral o escrito. Por ello, la mentira se ha estudiado desde la perspectiva de varias disciplinas, como la psicología, la lingüística, la psiquiatría, y la filosofía (Granhag y Strömwall, 2004). Más recientemente, la condición de verdad de las opiniones vertidas a través de Internet ha suscitado un interés creciente en el campo de la minería de opiniones (Ott et al., 2011). Esta cuestión es particularmente compleja, ya que el investigador no dispone de más información que el propio lenguaje escrito, cuando la investigación en el área señala que el lenguaje no verbal es el que contiene mayor información sobre la mentira (Vrij, 2010).

Hay características verbales de la mentira que forman parte de herramientas para su detección utilizadas por profesionales e investigadores. Las técnicas lingüísticas

ISSN 1135-5948 

automatizadas se utilizan para examinar los perfiles lingüísticos de la mentira en inglés. Más comúnmente, los investigadores han recurrido a las categorías de palabras definidas en *Linguistic Inquiry and Word Count*, conocido por sus siglas LIWC (Pennebaker, Francis y Booth, 2001); se trata de un programa de análisis de texto que clasifica las palabras en categorías significativas a nivel psicolingüístico. Comprende unas 2.200 palabras y raíces léxicas agrupadas en 75 categorías. Estas se han utilizado para estudiar cuestiones tales como la personalidad humana (Mairesse et al., 2007), los cambios psicológicos (Alpers et al., 2005), los juicios sociales (Leshed et al., 2007), y la salud mental (Rude, Gortner y Pennebaker, 2004). Además de ello, los trabajos que se detallan en el siguiente apartado avalan la utilidad de esta herramienta para la detección y caracterización de la mentira en el lenguaje. La validación del léxico contenido en su diccionario se ha obtenido a través de una comparación de la valoración de gran cantidad de textos escritos por parte de expertos y las puntuaciones obtenidas por medio de su análisis con LIWC.

En este contexto, el presente estudio persigue explorar los rasgos sintomáticos de la mentira en el español escrito, lo cual no se ha investigado aún. El resto del trabajo está organizado del siguiente modo: en la Sección 2 se presenta la investigación previa en el área; en la Sección 3 explicitamos la metodología utilizada para recoger y analizar los datos; en la Sección 4, se presentan y argumentan los resultados; finalmente, en la Sección 5 presentamos las conclusiones y sugerimos algunas directrices para futuras investigaciones en el área.

## *2 Trabajos relacionados*

LIWC fue usado inicialmente por el grupo de Pennebaker para una serie de estudios sobre la detección de la mentira en el lenguaje, publicándose los primeros resultados en (Newman et al., 2003). Para sus propósitos, los autores compilan un corpus con enunciados verdaderos y falsos a través de cinco estudios diferentes. En los tres primeros, los participantes expresan su opinión sobre el aborto. La primera prueba se basa en lenguaje oral, por lo que las opiniones se graban en vídeo, mientras que en la segunda y la tercera los participantes escriben sus opiniones. En el cuarto estudio, los sujetos expresan oralmente sus verdaderos sentimientos hacia su mejor amigo y después la opinión opuesta, mientras que el quinto se basa en un robo simulado en el que los participantes deben negar su supuesta implicación. De las 75 categorías de LIWC se seleccionan 29. Para predecir la mentira, los autores entrenan una validación cruzada de cinco iteraciones, obteniendo una clasificación correcta en un 67% cuando el tema era constante y una tasa de 61% en conjunto. Sin embargo, en dos de los cinco estudios los resultados no superan la pura casualidad. Finalmente, las variables significativas en al menos dos estudios se han usado para evaluar simultáneamente los cinco tests. El bajo rendimiento en algunos de los estudios puede deberse a la mezcla de modos de comunicación, ya que, tal y como afirma Picornell (2011), no todos los indicios verbales del engaño en la comunicación oral se corresponden con la escrita.

A partir de este estudio, LIWC se ha usado en el área forense principalmente para la investigación de la mentira en lenguaje oral. Los primeros estudios en esta línea exploran la utilidad de este programa en comparación con la técnica del *Reality Monitoring* (en adelante, RM). En primer lugar, Bond y Lee (2005) aplican LIWC a muestras aleatorias de un corpus que contiene enunciados orales verdaderos y falsos producidos por 64 reclusos, considerando para el análisis solo las variables seleccionadas por Newman et al. (2003). En conjunto, los resultados muestran que los participantes que mienten puntúan notablemente peor que los que dicen la verdad en lo referido a detalles sensoriales, pero mejor en aspectos espaciales. Este último aspecto se contradice con los resultados previos en la teoría del RM; tal es el caso de (Newman et al., 2003), donde esas categorías no arrojaban resultados significativos. Aparte de esta diferencia, ambos estudios tienen en común el hecho de que, a pesar de considerar la teoría del RM, los autores no realizan codificación manual de sus datos. Por tanto, no ofrecen una comparación directa con la efectividad del análisis automático a través de LIWC.

Este vacío en la investigación ha sido identificado por Vrij et al. (2007). Su hipótesis predecía que la detección de la mentira con LIWC sería menos precisa que mediante una codificación manual con RM. Para comprobarlo, los autores cuentan con un corpus

de entrevistas orales de 120 estudiantes universitarios. A la mitad de los participantes se les asigna la tarea de mentir sobre un episodio escenificado al que acaban de asistir, mientras que el resto deben decir la verdad. El análisis revela que el RM es capaz de distinguir ambos grupos mejor que el Análisis de Contenido Basado en Criterios (CBCA). Además de ello, la codificación manual con RM ofrece más indicios verbales del engaño que su versión automática.

Más recientemente, Fornaciari y Poesio (2011) han realizado un estudio sobre un corpus de transcripciones de testimonios judiciales orales. Este trabajo presenta dos novedades principales: primeramente, el objeto de estudio son muestras de lenguaje producido espontáneamente en vez de enunciados formulados *ad hoc* o en un laboratorio; además de ello, estudia la mentira en una lengua distinta al inglés: el italiano. Los autores retoman la idea de (Newman et al., 2003) del método para clasificar los textos de acuerdo a su cualidad de verdad en vez de estudiar la lengua en términos meramente descriptivos, siendo la unidad de análisis utilizada el enunciado en vez del texto. El fin último es una comparación entre la eficacia discriminatoria de las características léxicas de LIWC y otras variables como son la frecuencia de elementos gramaticales o de *n*-gramas de palabras o de partes de la oración. Para ello, utilizan cinco tipos de vectores, obteniendo como resultado que las características léxicas son ligeramente más eficaces que las gramaticales, aunque no existen diferencias estadísticamente significativas.

LIWC se ha usado también para el estudio de la mentira en lenguaje escrito, en mayor medida en el campo de la lingüística computacional que en el de la ciencia forense. Primeramente, Mihalcea y Strapparava (2009) emplean esta aplicación para un análisis *post hoc*, midiendo algunas de sus categorías en un corpus de 100 opiniones verdaderas y falsas sobre tres temas polémicos; el diseño del cuestionario es, de hecho, muy similar al mencionado anteriormente en (Newman et al., 2003). Como experimento preliminar, utilizan dos clasificadores de aprendizaje automático: un clasificador bayesiano *ingenuo* y máquinas de soporte vectorial (SVMs), tomando las frecuencias de palabras para entrenar ambos algoritmos. Se obtiene un 70% de éxito en la clasificación, lo que es notablemente superior al 50%. A partir de esta información, los autores calculan un índice de preponderancia para cada clase de palabras dentro del conjunto de textos asociados con la mentira. De este modo, consiguen identificar algunas de sus características distintivas, pero en términos meramente descriptivos.

Siguiendo esta línea de investigación, Ott et al. (2011) utilizan los mismos clasificadores de aprendizaje automático. Para su entrenamiento, además de utilizar las categorías de LIWC, evalúan y comparan otros dos enfoques computacionales: la identificación de géneros a través de la distribución de frecuencias de etiquetas gramaticales (*POS-tags*), y la categorización de textos, que les permite modelar tanto el contenido como el contexto con características de *n*-gramas. Su objetivo es el *spam* con opiniones falsas, que es cualitativamente distinto al lenguaje de la mentira en sentido estricto. De acuerdo con los resultados, la categorización del texto basada en *n*-gramas parece la metodología más adecuada, aunque su combinación con las variables de LIWC genera resultados ligeramente mejores.

Estos estudios exploran el lenguaje escrito usado en comunicación asíncrona. Por el contrario, Hancock y su grupo estudian el lenguaje de la mentira en comunicación mediada por ordenador (CMC) síncrona, en la cual todos los participantes se encuentran online de manera simultánea (Bishop, 2009). Concretamente utilizan el chat como medio de comunicación. En su primer estudio con LIWC, Hancock et al. (2004) exploran las diferencias entre el estilo lingüístico del emisor y del receptor dependiendo de si mienten o no. Para el análisis, seleccionan las variables consideradas más relevantes, esto es, frecuencia absoluta de palabra, pronombres, términos relativos a las emociones y a los sentidos, exclusiones, negaciones, y frecuencia de oraciones interrogativas. Los resultados muestran que, en conjunto, cuando los participantes mienten usan más palabras de media, gran cantidad de referencias a los demás, así como más términos sensitivos. En (Hancock et al., 2008) obtienen resultados similares a partir de un experimento comparable. La diferencia principal reside en que los autores incluyen el elemento de la motivación, observando así que los participantes que mienten con cierta motivación tienden a evitar términos causales, mientras que los que carecen de ella aumentan el uso de negaciones.

Todos estos estudios coinciden en la exploración de un corpus a través de un conjunto de variables, pero ninguno de ellos toma la totalidad de categorías de LIWC para la clasificación automática de ambos sublenguajes en el medio escrito. Además, los investigadores suelen tomar el lenguaje de la mentira como un todo, ignorando las características particulares que pueden distinguir a un hablante de otro, asumiendo que todos mienten de manera similar y de forma regular. En vez de comparar cada muestra individual del lenguaje del engaño con su texto de control correspondiente, todo el subconjunto de textos etiquetado como "falso" se compara con el conjunto etiquetado como "verdadero". Por otro lado, la principal desventaja de utilizar un corpus de lenguaje "auténtico" es precisamente la dificultad de obtener una muestra del lenguaje de control en el cual el mismo hablante exprese una idea verdadera con el fin de permitir una comparación idiolectal.

## *3 Metodología*

Para nuestro propósito, hemos desarrollado un marco de trabajo basado en un clasificador de máquinas de soporte vectorial (SVMs); de este modo, pretendemos detectar la mentira en un corpus de opiniones. Estos algoritmos se han aplicado con éxito en diversas tareas de clasificación de texto debido a sus múltiples ventajas: primero, son robustos en grandes espacios dimensionales; segundo, cualquier característica resulta relevante; tercero, son robustos cuando hay un conjunto escaso de muestras; finalmente, la mayoría de los problemas de categorización de texto son separables linealmente (Saleh et al., 2011).

Hemos usado LIWC para obtener los valores de las categorías y posteriormente entrenar el clasificador mencionado. Esta aplicación proporciona un método eficaz para estudiar los componentes emocionales, cognitivos y estructurales contenidos en el lenguaje tomando la palabra como unidad de análisis (Pennebaker, Francis y Booth, 2001). El diccionario interno de LIWC comprende 2.300 palabras y raíces léxicas clasificadas en cuatro grandes dimensiones: procesos lingüísticos estándares, procesos psicológicos, relatividad y asuntos personales. Cada palabra o raíz define una o más de las 75 categorías incluidas por defecto.

Las palabras que constituyen las categorías de LIWC se han seleccionado después de cientos de estudios sobre el comportamiento psicológico (Tausczik y Pennebaker, 2010). En la primera dimensión, procesos lingüísticos, la mayoría de las categorías incluyen palabras gramaticales; por tanto, la selección de elementos es más clara y sencilla, como en el caso de la categoría de los artículos, formada por nueve palabras en español: *el, la, los, las, un, uno, una, unos* y *unas*. Otro ejemplo similar es una categoría incluida en la tercera dimensión: tiempo verbal, que comprende todas las formas verbales en pasado, presente y futuro. Dentro de la misma dimensión se encuentra la categoría de espacio, en la que se incluyen los adverbios y las preposiciones espaciales. Por el contrario, las dos dimensiones restantes contienen categorías más subjetivas, especialmente aquellas referidas a procesos emocionales dentro de la segunda dimensión. De hecho, la selección léxica de estas categorías requirió la colaboración de un grupo de expertos. Más concretamente, se creó una lista inicial de términos a partir de diccionarios y tesauros que fue puntuada por tres jueces independientes. Finalmente, la cuarta dimensión incluye categorías de palabras relacionadas con asuntos personales, intrínsecos a la condición humana. Esta dimensión ha sido excluida en algunos de los estudios mencionados en la sección anterior, alegando que es demasiado dependiente de la temática del texto (Hancock et al., 2004, 2008; Newman et al., 2003). La Tabla 1 ofrece un ejemplo ilustrativo de la lista de categorías del diccionario. El listado completo puede encontrarse en (Pennebaker, Francis y Booth, 2001:17-21), y sus equivalencias en español en (Ramírez-Esparza, Pennebaker y García, 2007:37-39).

Cada opinión se almacenó en un fichero de texto, y los errores ortográficos se corrigieron para no obstaculizar la labor de clasificación de la aplicación. Tras ello, se procedió al análisis de los 800 archivos para crear las muestras para el clasificador. La versión utilizada fue LIWC2001, ya que, a pesar de existir la versión posterior LIWC2007, es la primera la que ha sido totalmente validada para el idioma español en diversos estudios psicolingüísticos (Ramírez-Esparza, Pennebaker y García, 2007).

| I. Dimensión lingüística estándar | II. Procesos psicológ. | III. Relatividad | IV. Asuntos personales |
|---|---|---|---|
| Cuenteo de | Procesos | Espacio | Ocupación |

| | | | |
|---|---|---|---|
| palabras | afectivos | | |
| % de palabras capturadas | Emociones positivas | Inclusiones | Pasatiempo |
| % de palabras > 6 caracteres | Emociones negativas | Exclusiones | Dinero y asuntos financieros |
| Total de pronombres | Procesos cognitivos | Movimiento | Asuntos metafísicos |
| Primera persona singular | Causa y efecto | Tiempo | Estados físicos y funciones |

Tabla 1: Extracto de las variables usadas en LIWC2001

Todos los resultados producidos por el programa se usaron para el experimento, a excepción de una categoría considerada experimental por Pennebaker, Francis y Booth (2001): palabras de relleno (*osearr, esterr, ehh*), ya que es exclusiva del lenguaje oral. La otra categoría experimental, tacos e insultos, ha sido incluida en la primera dimensión por ser ese su sitio en la versión actualizada del programa.

Nuestro experimento ha sido implementado mediante el software para minería de datos de la plataforma Weka (Bouckaert et al., 2010). Más concretamente, aplicamos una máquina de soporte vectorial (SVM) lineal con la configuración predeterminada por la herramienta.

## 3.1 Recogida de datos

El estudio de la distinción entre las opiniones verdaderas y falsas requiere un corpus con una clasificación de los textos según su condición de verdad. Para ello, creamos un cuestionario de diseño similar al de (Mihalcea y Strapparava, 2009). En él participaron 100 sujetos, todos ellos hablantes nativos de la variedad de español peninsular o europeo. Se propusieron cuatro temáticas diferentes: la adopción homosexual, las corridas de toros, el mejor amigo y el mejor profesor. Ya que no se trataba de lengua producida espontáneamente, se consideró necesario minimizar el efecto de la paradoja del observador (Labov, 1972), no revelando el fin último de la investigación a los participantes. Además, se les recalcó la importancia de convencer a su interlocutor de la autenticidad de su argumento con el fin de aumentar su motivación, elemento clave de acuerdo con (Hancock et al., 2008).

Para los dos primeros temas, los participantes recibieron instrucciones de imaginar que estaban participando en un debate y que tenían alrededor de 10 minutos para expresar su opinión. Primero, debían preparar un breve discurso escrito expresando lo que realmente pensaban sobre el tema, y, a continuación, otro discurso defendiendo la postura opuesta. En ambos casos debían escribir un mínimo de 5 oraciones y tantos detalles como fuera posible. Para los otros dos temas, los participantes debían pensar en uno de sus mejores amigos y en uno de sus mejores profesores, respectivamente, e incluir en un relato de extensión similar a los anteriores hechos y anécdotas relevantes en su relación con ambos. Posteriormente, debían pensar en una persona con la que se llevaran mal y describirla como si entre ellos existiese una buena relación. A su vez, debían recordar a un profesor que no tuvieran en buena estima y mentir sobre ello. Conviene señalar que el orden de realización de las tareas era indiferente.

Se recogieron 100 opiniones verdaderas y 100 falsas para cada tema, con una media de 80 palabras por contribución. Se realizó una verificación manual de la calidad de las contribuciones y, a excepción de tres, todas las demás resultaron satisfactorias. Este recurso se encuentra actualmente en proceso de ampliación, y está prevista su publicación en un futuro cercano.

## 3.2 Análisis de datos

Para entrenar el clasificador, el corpus se dividió en opiniones verdaderas y falsas. Para su análisis, consideramos los atributos de cada dimensión de LIWC, descritos en la sección 3, en cada una de las opiniones de manera independiente. Con ello, se obtuvieron diversos clasificadores, cuyos resultados se explican en la siguiente sección. Para cada uno de ellos, se realizó una validación cruzada de 10 iteraciones, teniendo todos los conjuntos una distribución similar de testimonios verdaderos y falsos.

## *4 Resultados y discusión*

Los resultados del experimento se muestran en la Tabla 2. En la primera columna, se indica el número de dimensiones de LIWC utilizadas por cada clasificador. Por ejemplo, 1_2_3_4 indica que se han utilizado todas las dimensiones, mientras que 1_2 indica que se han usado solo las categorías de las dos primeras. Los valores

que aparecen en la tabla corresponden a la medida F1, la media armónica de precisión y exhaustividad.

| | Adopc. homo. | Toros | Mejor amigo | Mejor prof. | Total |
|---|---|---|---|---|---|
| **1** | 0,638 | 0,679 | 0,763 | 0,549 | 0,647 |
| **1_2** | 0,709 | 0,655 | 0,83 | 0,625 | 0,715 |
| **1_2_3** | 0,698 | 0,669 | 0,835 | 0,6 | 0,731 |
| **1_2_3_4** | 0,718 | 0,66 | 0,845 | 0,578 | 0,725 |
| **1_2_4** | 0,728 | 0,63 | 0,83 | 0,61 | 0,702 |
| **1_3** | 0,64 | 0,68 | 0,82 | 0,56 | 0,678 |
| **1_3_4** | 0,657 | 0,643 | 0,815 | 0,564 | 0,667 |
| **1_4** | 0,631 | 0,651 | 0,738 | 0,584 | 0,641 |
| **2** | 0,678 | 0,624 | 0,78 | 0,555 | 0,683 |
| **2_3** | 0,724 | 0,619 | 0,81 | 0,54 | 0,703 |
| **2_3_4** | 0,724 | 0,609 | 0,81 | 0,52 | 0,701 |
| **2_4** | 0,703 | 0,59 | 0,78 | 0,525 | 0,682 |
| **3** | 0,62 | 0,62 | 0,695 | 0,49 | 0,598 |
| **3_4** | 0,611 | 0,595 | 0,684 | 0,487 | 0,621 |
| **4** | 0,506 | 0,525 | 0,639 | 0,513 | 0,579 |

Tabla 2: Resultados del experimento (medida F1)

Los resultados revelan que la dimensión que mejor funciona como clasificador independiente en este caso es la segunda, procesos psicológicos (68,3%). Este dato se corresponde con ciertos resultados de (Newman et al., 2003), donde el vocabulario relacionado con el entendimiento, como *pensar* o *saber*, se encuentra más frecuentemente en opiniones verdaderas, donde los participantes se identifican con su testimonio. En lo que respecta a las palabras dominantes en los testimonios falsos, investigaciones previas señalan las palabras que expresan certeza, tales como *siempre* y *nunca*, probablemente en un intento por parte del hablante de usar palabras contundentes para convencer a su interlocutor y esconder de ese modo la mentira (Bond y Lee, 2005; Mihalcea y Strapparava, 2009). Además, de acuerdo con Burgoon et al. (2003), otra característica asociada con la mentira es la alta frecuencia de vocabulario relativo a emociones negativas. Todas estas categorías se incluyen en la segunda dimensión, confirmándose así su potencial discriminatorio en nuestro experimento de clasificación.

Por su parte, la primera dimensión funciona relativamente bien (64,7%). Parece lógico que así sea, teniendo en cuenta el potencial considerable de las palabras gramaticales, que constituyen buena parte de las dimensiones lingüísticas estándares. La importancia de estos elementos ha suscitado gran interés, no solo en lingüística computacional, sino también en el área de la psicología. Como afirman Chung y Pennebaker (2007:344), las palabras gramaticales "permiten conocer mejor la psique humana". Las variaciones en su uso pueden aportar información valiosa sobre el estado mental de una persona, su edad, su sexo, su estatus social o la condición de verdad de su discurso. En este sentido, llama la atención la elevada frecuencia de la segunda y tercera persona en los textos que contienen engaños, en oposición a la primera persona, frecuente en mayor medida en los testimonios verdaderos. Como apuntan Mihalcea y Strapparava (2009), ello puede deberse al temor del que engaña a identificarse con su testimonio, por lo que suele emplear pronombres y formas verbales que impliquen cierto distanciamiento.

Por el contrario, como cabía esperar a raíz de ciertas investigaciones previas sobre el tema (Fornaciari y Poesio, 2011; Newman et al., 2003), la cuarta dimensión es la que posee por sí misma un menor poder discriminatorio. La razón puede residir en la escasa relación de los temas sugeridos en el cuestionario y el vocabulario contenido en las categorías relativas a asuntos personales. Por último, la tercera dimensión presenta un resultado muy similar de manera independiente: menos de 0,02 puntos de diferencia en la medida del total.

Como muestra la Tabla 2, cuando se entrena el clasificador con ciertas combinaciones de dimensiones, su rendimiento mejora notablemente. Ello se justifica por las siguientes palabras de Vrij (2010:103):

> *No existe un indicio verbal único asociado con la mentira, similar a la nariz de Pinocchio. Sin embargo, algunos indicios verbales pueden considerarse sutiles indicadores a la hora de diagnosticar la mentira.*

De este modo, parece claro que una combinación de características verbales resulta más eficaz que la aplicación de categorías de manera aislada. La agrupación de las tres primeras dimensiones resulta considerablemente exitosa (73,1%), a pesar del bajo rendimiento de la tercera como variable única. Por el contrario, la adición de la cuarta dimensión a esta combinación es

contraproducente, ya que la tasa de éxito empeora en lugar de mejorar.

Otro hecho que llama la atención es que los resultados de la clasificación por medio de estas dimensiones son fuertemente dependientes de la temática de cada subcorpus. Así, por ejemplo, las combinaciones 1_2_4 (72,8%), 2_3 (72,4%) y 2_3_4 (72,4%) son las que mejor discriminan los textos verdaderos de los falsos cuando se aplican al subcorpus sobre la adopción homosexual. Asimismo, las dimensiones 1_2_3 y 1_2_3_4 alcanzan su punto de máxima efectividad cuando se trata de la temática del mejor amigo. Por el contrario, los mejores resultados obtenidos sobre el conjunto de datos de las corridas de toros se consiguen con la combinación 1_3 (68%), mientras que la mejor puntuación en la temática del buen profesor alcanza apenas un 61% con las dimensiones 1_2_4. Por tanto, no cabe duda de que la saturación factorial de las cuatro dimensiones y su correlato con la temática juegan aquí un papel importante, abriéndose una línea de trabajo futuro prometedora.

## *5 Conclusiones e investigación futura*

En el presente trabajo hemos utilizado un clasificador automático para el reconocimiento de la mentira en textos escritos en lengua española, utilizando para su entrenamiento las categorías psicolingüísticas de la aplicación LIWC. A través de un experimento llevado a cabo sobre cuatro conjuntos de datos, hemos probado que es posible separar los textos en dicha lengua de acuerdo con su condición de verdad, siendo las dos primeras dimensiones, procesos lingüísticos y psicológicos, las más relevantes para tal fin.

Nuestras investigaciones futuras en esta línea incluirán una comparación de la metodología empleada en el presente trabajo con un sistema *bag-of-words,* en el cual el texto se represente como una colección desordenada de elementos al margen de factores lingüísticos como la gramática (Lewis, 1998). Además de ello, estamos comparando los resultados de la presente investigación con la aplicación de la misma metodología de análisis a un corpus en inglés, con el fin de elaborar un estudio contrastivo basado en las diferencias y similitudes lingüísticas de ambos idiomas.

## *Agradecimientos*

Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación a través del proyecto SeCloud (TIN2010-18650). Además de ello, Ángela Almela cuenta con la financiación de la Fundación Séneca (12406/FPI/09).

## *Bibliografía*

Alpers, G. W., A. Winzelberg, C. Classen, H. Roberts, P. Dev, C. Koopman, y B. Taylor. 2005. Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior*, 21:361-376.

Bishop, J. 2009. Enhancing the understanding of genres of web-based communities: The role of the ecological cognition framework. *International Journal of Web-Based Communities*, 5(1):4-17.

Bond, G. D. y A. Y. Lee. 2005. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, 19:313-329.

Bouckaert, R. R., E. Frank, M. A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, y I. H. Witten. 2010. WEKA-experiences with a java open-source project. *Journal of Machine Learning Research*, 11:2533-2541.

Burgoon, J. K., J. P. Blair, T. Qin, y J. F. Nunamaker. 2003. Detecting deception through linguistic analysis. *Intelligence and Security Informatics*, 2665:91-101.

Chung, C. y J. W. Pennebaker. 2007. The psychological functions of function words. En K. Fiedler (Ed.), Social Communication, páginas 343-359, *Psychology Press* (New York).

Coulthard. M. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4):431-447.

DePaulo, B. M., D. A. Kashy, S. E. Kirkendol, M. M. Wyer, y J. A. Epstein. 1996. Lying in everyday life. *Journal of Personality and Social Psychology*, 70:979-995.

Fornaciari, T. y M. Poesio. 2011. Lexical vs. Surface Features in Deceptive Language Analysis. En *Proceedings of the ICAIL 2011 Workshop Applying Human Language Technology to the Law*, páginas 2-8, Pittsburgh (Alemania).

Granhag, P. A. y L. A. Strömwall. 2004. *The Detection of Deception in Forensic Contexts*. Cambridge University Press, Cambridge.

Hancock, J. T., L. E. Curry, S. Goorha, y M. T. Woodworth. 2004. Lies in conversation: an examination of deception using automated linguistic analysis. En *Proceedings of the Annual Conference of the Cognitive Science Society*, páginas 1-6, Taylor and Francis Group, Psychology Press, Mahwah (EE.UU.).

Hancock, J. T., L. E. Curry, S. Goorha, y M. T. Woodworth. 2008. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45:1-23.

Labov, W. 1972. *Sociolinguistic Patterns*. Oxford: Blackwell.

Leshed, G., J. T. Hancock, D. Cosley, P. L. McLeod, y G. Gay. 2007. Feedback for guiding reflection on teamwork practices. En *Proceedings of the GROUP'07 Conference on Supporting Group Work*, páginas 217-220, Association for Computing Machinery Press, New York (EE.UU.).

Lewis, D. 1998. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. En *Proceedings of ECML-98, 10th European Conference on Machine Learning*, páginas 4-15, Springer Verlag, Heidelberg (Alemania).

Mairesse, F., M. A. Walker, M. Mehl, y R. K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457-500.

Mihalcea, R. y C. Strapparava. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. En Proceedings of the *Association for Computational Linguistics, ACL-IJCNLP*, páginas 309-312, Singapur (Singapur).

Newman, M. L., J. W. Pennebaker, D. S. Berry, y J. M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29:665-675.

Ott, M., Y. Choi, C. Cardie, y J. T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. En *Proceedings of ACL*, páginas 309-319, Portland (EE.UU.).

Pennebaker, J. W., M. E. Francis, y R. J. Booth. 2001. *Linguistic Inquiry and Word Count*. Erlbaum Publishers, Mahwah (NJ).

Pennebaker, J. W., C. K. Chung, M. Ireland, A. L. Gonzales, y R. J. Booth. 2007. *The Development and Psychometric Properties of LIWC2007*. LIWC.net, Austin (TX).

Picornell, I. 2011. The Rake's Progress: Mapping deception in written witness statements. Comunicación oral presentada en el International Association of Forensic Linguists Tenth Biennial Conference, Birmingham (RU).

Ramírez-Esparza, N., J. W. Pennebaker, y F. A. García. 2007. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología*, 24:85-99.

Rude, S. S., E. M. Gortner, y J. W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18:1121-1133.

Rushdi-Saleh, M., M. T. Martín-Valdivia, A. Montejo-Ráez, y L. A. Ureña-López. 2011. Experiments with SVM to classify opinions in different domains. *Expert System with Applications*, 38(12):14799-14804.

Tausczik, Y. R. y J. W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24-54.

Vrij, A. 2010. *Detecting Lies and Deceit: Pitfalls and Opportunities*. 2nd edition. John Wiley and Sons, Chischester.

Vrij, A., S. Mann, S. Kristen, y R. P. Fisher. 2007. Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior*, 31(5):499-518.