ISSN 1870-4069

Detección del engaño en notas de opinión a través de técnicas tradicionales de clasificación automática de textos

Javier Sánchez-Junquera, Luis Villaseñor-Pineda, Hugo J. Escalante, Manuel Montes-y-Gómez

Instituto Nacional de Astrofísica, Óptica y Electrónica, Laboratorio de Tecnologías del Lenguaje, Puebla, México

jjsjunquera@gmail.com, {villasen,hugojair,mmontesg}@inaoep.mx

Resumen. En este trabajo se realiza un estudio del alcance de técnicas tradicionales usadas en clasificación automática de textos (*v. gr.* bolsa de palabras) para la detección de engaño. Comúnmente las técnicas tradicionales funcionan adecuadamente en clasificación temática. Sin embargo, se desea conocer el rendimiento de dichas técnicas en una tarea intuitivamente no-temática. La colección empleada es un conjunto de notas en inglés de opiniones sobre hoteles, incluyendo notas verdaderas y falsas. Se realizaron experimentos utilizando la representación de bolsa de palabras con esquemas de pesado binario, *tf* y *tf-idf* y entrenando un clasificador probabilista. Los resultados muestran que el engaño puede ser detectado con el enfoque tradicional. Un primer análisis de estos resultados identifica aquellos elementos sobre los que recayó la discriminación.

Palabras clave: Detección de engaño, marcadores sintácticos, clasificación de texto.

Deceptive Detection in Opinion Notes by Traditional Techniques of Automatic Text Classification

Abstract. This work studies the scope of traditional techniques used in automatic text classification (*v. gr.* bag of words) for the deceptive detection. Commonly, traditional techniques work well in thematic classification. However, it is desired to know the performance of these techniques in an intuitively non-thematic task. The collection used is a set of English notes of hotel reviews, including truthful and deceptive notes. Experiments were performed using bag of words with binary weighing schemes, *tf* and *tf-idf* and training a probabilistic classifier. The results show that deception can be detected with the traditional approach. A first analysis of these results identifies those elements on which discrimination fell.

Keywords: Deceptive detection, POS tagging, text classification.

1. Introducción

En la literatura se puede encontrar estudios sobre el engaño como aquella acción en la que una persona intenta convencer de la veracidad de algo que en realidad sabe que es falso. Dichos estudios abarcan desde las respuestas fisiológicas, el lenguaje corporal y el lenguaje natural tanto escrito como hablado. El presente trabajo se centra en la detección de engaño en opiniones que fueron escritas con la intención de resumir o valorar una supuesta experiencia.

El engaño está presente en múltiples actividades en las que el ser humano tiene la posibilidad, necesidad u obligación de expresarse verbalmente. Ejemplo de ello se puede apreciar en la web, donde los usuarios dejan plasmada una reseña, valoración u opinión sobre prácticamente todo. A menudo puede ser útil leer sobre la experiencia de otros para tomar una elección propia, y sabiendo esto a muchos les conviene falsear opiniones en busca de algún beneficio.

Existe gran motivación en detectar automáticamente las opiniones falsas[1], por ejemplo, TripAdvisor[2] posee millones de opiniones de viajeros acerca de alojamientos y tiene particular interés en la detección de opiniones falsas cuyo fin generalmente es aumentar o disminuir la reputación de un establecimiento por parte de propietarios o competidores, respectivamente.

El presente trabajo muestra un análisis de técnicas de procesamiento de lenguaje natural empleando técnicas probabilistas -las cuales han mostrado muy buen desempeño estableciendo un balance entre la calidad de los resultados y la complejidad computacional- en la clasificación de opiniones falsas. En este análisis se utiliza el modelo *Naïve Bayes Multinomial Updatable*, implementado en la plataforma WEKA, el cual se ha usado satisfactoriamente en problemas de clasificación, fundamentalmente en clasificación temática de textos. Con este modelo se analizan algunas representaciones simples de los documentos con el objetivo de estudiar el alcance de técnicas relativamente sencillas para la detección de engaño.

A continuación, en la Sec. 2, se comentará algunos trabajos relacionados a la detección de engaño. En la Sec. 3, se discutirá la metodología llevada a cabo, el corpus con el que se trabajó y la evaluación del sistema. En la Sec. 3.3 se mostrarán los resultados obtenidos y una breve discusión de los mismos. Finalmente se darán las conclusiones y el trabajo futuro en la Sec. 5.

[1] Si bien las opiniones falsas no implican necesariamente la existencia de engaño, o sea, la intención de engañar, en este documento se mencionará "opiniones falsas" como expresión alternativa para referirse a "opiniones engañosas".

[2] TripAdvisor es un sitio web con facilidades a los viajeros. Cuenta con 435 millones de opiniones y comentarios sobre 6,8 millones de alojamientos, restaurantes y atracciones. Para más detalles sobre la moderación de opiniones y detección de fraude en TripAdvisor visitar https://www.tripadvisor.es/vpages/review_mod_fraud_detect.html

2. Trabajos relacionados

Entre los trabajos relacionados más citados se encuentra [7], en el que intentan detectar el engaño desde la psicología y la lingüística computacional. Los autores proponen una colección balanceada de opiniones positivas en inglés, tomada como *gold standard*. Con esta colección entrenan los clasificadores *Naïve Bayes* (NB) y *Support Vector Machine* (SVM), bajo la consideración de trigramas, bigramas y unigramas basados en análisis psicolinguísticos y mediante el recurso externo *Linguistic Inquiry and Word Count* (LIWC). Estos autores concluyen que en las opiniones engañosas predominan las características de la escritura imaginativa y que se carece de información espacial.

Más allá del análisis léxico, en [3] proponen investigar las estructuras sintácticas notando que en las opiniones engañosas las frases verbales, las cláusulas subordinadas y las frases adverbiales con *wh-words* son más frecuentes que en las opiniones verdaderas. En [6] incrementan la colección mencionada con opiniones negativas, igualmente balanceadas entre verdaderas y falsas, también en inglés; detectando que en las opiniones engañosas se usa más la primera persona del singular, así como un lenguaje con tendencia a la exageración.

Mediante aprendizaje semisupervisado, en [4] intentan abordar el problema de la detección a partir de la escasez de ejemplos engañosos. Estos investigadores sugieren la existencia de aspectos comunes en la manera de escribir opiniones positivas y negativas cuando son falsas, ya que sus resultados muestran que un clasificador entrenado con opiniones de ambas polaridades es más efectivo que un clasificador para cada una por separado, confirmando resultados de [6].

Por otra parte, el clasificador *Bayes Multinomial Updatable* ha sido usado en la detección de lenguaje ofensivo [8]; la clasificación de conductas colaborativas en texto [1] y la detección de *spam* a partir del reconocimiento de la personalidad en mensajes cortos [2]. Entre los clasificadores bayesianos disponibles en WEKA 3.8, este alcanzó mejores resultados. Por esta razón, la metodología seguida en este trabajo y los resultados que se reportan son empleando este clasificador.

3. Metodología

3.1. Configuración experimental

Para el entrenamiento y validación del clasificador que se utiliza en este trabajo se emplea el corpus[3] de opiniones sobre 20 hoteles de Chicago construido por [7,6]. Por cada uno de estos hoteles, se cuenta con 20 opiniones negativas falsas, 20 negativas verdaderas, 20 positivas falsas y 20 positivas verdaderas. En total son 1600 opiniones. En este trabajo solamente se tienen en cuenta la información sobre la falsedad o veracidad de las opiniones, proponiéndose como trabajo futuro la inclusión de la polaridad.

El pre-procesamiento de los textos se llevó a cabo con el uso de *NLTK 3.0*. Con los siguientes pasos:

[3] El corpus está disponible gratuitamente en http://myleott.com/op_spam/

- *lowercase*,
- *stemming* (Porter Stemmer),
- eliminación de palabras vacías (según las incluidas en *NLTK* para el inglés),

e ignorándose todos los caracteres que no fueran alfabéticos. Después del preprocesamiento se observaron 6637 *tokens*, encontrándose por documento un promedio de aproximadamente 60 *tokens*, con un número máximo y mínimo de 286 y 12 respectivamente.

3.2. Sobre la selección de términos

Las palabras vacías son aquellas que *carecen* de significado y no aportan sentido distintivo a los textos, o sea, no son útiles para diferenciar unos de otros (por ejemplo: artículos, pronombres y conjunciones). Sin embargo, según [6], en las opiniones positivas, los pronombres en primera persona del singular son más frecuente cuando se trata de opiniones falsas. Por tanto, excluir las palabras vacías del vocabulario de las opiniones podría influir negativamente en la clasificación.

Además de las palabras vacías, existen otras que no son precisamente carentes de significado, pero que por su alta o baja frecuencia tampoco ayudan a separar los textos en clases. Mientras que las palabras más frecuentes tienden a estar indistintamente en todas las clases, las que son poco frecuentes no llegan a ser comunes dentro de una sola clase; estos dos tipos de palabras dificultan la clasificación de una opinión, y son consideradas ruido.

En un intento de observar los términos más representativos para un posterior estudio, se ordenaron las palabras según la diferencia en las que aparecían en ambas clases, así como la ganancia de información (GI) de cada una de ellas. En la Sec 3.3 se discutirá algunas observaciones al respecto. Es importante notar que tanto el ordenamiento como la GI no se emplearon para la selección de términos en el esquema de clasificación propuesto, para evitar un posible sobre-ajuste a la colección empleada.

3.3. Pesado de términos

Tradicionalmente, en la representación de documentos, hay tres tipos de pesado para los términos.

En el pesado binario solo se tiene en cuenta la presencia de los términos en los documentos. De esta forma los pesos son 1 o 0, dependiendo de si el término está en el texto o no, respectivamente.

Por otra parte, el pesado *tf* consiste en la frecuencia del término en el documento. De esta forma, un término que sea muy frecuente tendrá un peso mayor que otro que sea poco frecuente.

El *tf* es que no considera el hecho de que un término sea muy frecuente precisamente en todos los documentos, lo cual indicaría que el término no es útil para representar los documentos. Sin embargo, el *tf-idf* establece un compromiso

entre la frecuencia del término en el documento y la cantidad de documentos que lo contienen. O sea,

$$tf\text{-}idf = tf \times \log \frac{|D|}{|\{d \in D : t \in d\}|},$$

donde $|D|$ es la cantidad total de documentos y $|\{d \in D : t \in d\}|$ es la cantidad de documentos en los que aparece el término t.

4. Resultados

En esta sección se presenta los resultados obtenidos con el clasificador *Bayes Multinomial Updatable* utilizando diferentes esquemas de pesado tradicionales. También se muestran algunas observaciones sobre los términos más relevantes dada la colección de opiniones usada.

El clasificador fue entrenado antes y después del pre-procesamiento indicado en la Sec. 3.1, considerándose unigramas de tokens como rasgos y tres esquemas de pesado (binario, *tf* y *tf-idf*). Los seis modelos obtenidos fueron evaluados mediante una validación cruzada de 10 iteraciones. Para ambos, el entrenamiento y la validación, se usó la implementación de WEKA 3.8.

Los resultados mostrados en la Tabla 1 incluyen precisión, recuerdo y f_1-*measure* por clase, así como la exactitud global y son calculados como sigue:

$$P(c) = \frac{\#\ de\ predicciones\ correctas\ de\ la\ clase\ c}{\#\ de\ \boldsymbol{predicciones}\ para\ la\ clase\ c}, \tag{1}$$

$$R(c) = \frac{\#\ de\ predicciones\ correctas\ de\ la\ clase\ c}{\#\ de\ \boldsymbol{opiniones}\ de\ la\ clase\ c}, \tag{2}$$

$$F_1(c) = \frac{2 \times P(c) \times R(c)}{P(c) + R(c)}, \tag{3}$$

$$E = \frac{\#\ predicciones\ correctas}{\#\ total\ de\ opiniones\ clasificadas}, \tag{4}$$

donde $P(c)$, $R(c)$ y $F_1(c)$ son precisión, recuerdo y f_1-*measure* de la clase c respectivamente, y E es la exactitud global de la clasificación.

Como se puede notar en la Tabla 1, cada esquema muestra una exactitud entre el 70 % y el 88 %, aún cuando no se está usando ningún recurso externo como en los trabajos de [5] y [7], ni desarrollándose un análisis psicolingüístico [7,6] o sintáctico [3]. Sin embargo, los autores de este trabajo consideran más relevante para trabajos futuros tener en cuenta los esquemas que consigan mejor resultado en el recuerdo de las opiniones falsas o en la precisión de las opiniones verdaderas, como es el caso del esquema con pesado *tf* sin pre-procesamiento o con pre-procesamiento, respectivamente.

Tabla 1. Se incluye Precisión, Recuerdo y F_1-*measure* por cada categoría y la Exactitud para cada pesado En cada caso se indica el tipo de pesado y si se hizo pre-procesamiento. Resaltado en negrita los mejores resultados entre los esquema.

			P		R		F_1	
Pesado	**Pre-proc.**	***E***	**V**	**F**	**V**	**F**	**V**	**F**
binario		**87.94 %**	0.914	**0.850**	0.838	0.921	**0.874**	**0.884**
tf		70.38 %	0.931	0.633	0.440	**0.968**	0.598	0.766
tf-idf		82.38 %	0.832	0.816	0.811	0.836	0.822	0.826
binario	√	87.31 %	0.900	**0.850**	**0.840**	0.906	0.869	0.877
tf	√	80.81 %	**0.951**	0.734	0.650	0.966	0.772	0.834
tf-idf	√	82.5 %	0.826	0.824	0.824	0.826	0.825	0.825

4.1. Análisis de los resultados

Palabras relevantes por clase En la colección de opiniones que se procesó en este trabajo, se buscó eliminar primeramente aquellos términos considerados palabras vacías. Sin embargo, un análisis de aquellos términos más útiles para la clasificación confirmaron la conclusión de [7]: la primera persona del singular es más frecuente en las opiniones engañosas que en las verdaderas. Por este motivo se optó por no eliminar las *stopwords* y contemplarlas como posibles términos relevantes para la clasificación.

Para analizar las palabras más relevantes se halló la frecuencia de estas en cada clase. Ordenadas según la diferencia entre estas frecuencias, la Tabla 2 muestra las primeras y algunas de las últimas palabras de esta relación. Como se puede notar, palabras vacías como *i, my, a, to, on, the* y *at* pueden servir para determinar si las opiniones de esta colección son verdaderas o falsas, al igual que palabras con demasiada frecuencia como *chicago* y *hotel*. Por otro lado, hay palabras como *magnificent* y *complimentary* que no son del todo atípicas pero están distribuidas en las clases casi por igual, siendo así irrelevantes para la clasificación.

Sin embargo, del total de 9527 palabras -solamente *lowercase* como pre-procesamiento-, solo 792 tuvieron ganancia de información mayor que cero, algunas coinciden con las primeras listadas en la Tabla 2, y se observan otras como:

- *priceline*: Ocurre 50 veces en las verdaderas y nunca en las falsas.
- *reviews*: En singular y plural, suman por encima de 100 ocurrencias en las verdaderas, mientras que en las falsas solo 24.
- *ave* y *avenue*: Entre ambas ocurren 114 veces en las verdaderas, y solo 23 en las falsas.
- *experience*: Entre plural y singular, ocurre más del doble de veces en las falsas que en las verdaderas.

Tabla 2. Frec. V: Frecuencia en la clase de opiniones verdaderas, **Frec. F**: Frecuencia en la clase de opiniones falsas, **Dif**: Diferencia entre las frecuencias de ambas clases. En la tabla de la izquierda las 20 palabras con mayor diferencia entre las dos clases; en negrita la frecuencia en la clase en la que es mucho más común. En la tabla de la derecha las últimas 20 palabras con menor diferencia de frecuencia entre las clases.

Palabras	Frec. V	Frec. F	Dif
i	2432	**3799**	1367
my	783	**1526**	743
chicago	484	**1020**	536
a	**3463**	3003	460
is	**1257**	875	382
was	2737	**3110**	373
to	3201	**3559**	358
hotel	1485	**1808**	323
we	**1724**	1422	302
on	**951**	686	265
the	**8133**	7882	251
great	**549**	316	233
location	**349**	134	215
for	**1554**	1342	212
very	**752**	582	170
be	436	**600**	164
at	1094	**1257**	163
no	**382**	234	148
floor	**202**	58	144
from	**559**	431	128

Palabras	Frec. V	Frec. F	Dif
magnificent	41	40	1
complimentary	39	38	1
end	38	37	1
fine	33	32	1
arrival	32	31	1
poor	30	31	1
immediately	29	30	1
provided	27	26	1
paying	23	24	1
fresh	23	24	1
seem	22	23	1
loud	22	23	1
standard	20	21	1
system	20	21	1
sink	19	20	1
uncomfortable	20	19	1
makes	19	20	1
surprised	20	19	1
thank	17	18	1
lady	18	17	1

Utilizar estas palabras cuya frecuencia en cada clase, o su ganancia de información, indican que son útiles en la clasificación, podría hacer al modelo sensible de un cambio de colección o dominio. Una cuestión que podría ser menos sensible a este tipo de cambios es el trabajo con patrones o marcadores sintácticos. La siguiente sección tiene algunas observaciones sobre las etiquetas sintácticas de las palabras así como los bigramas de etiquetas.

Marcadores sintácticos por clase El proceso de etiquetar las palabras según la función sintáctica que tienen en la oración recibe el nombre de *POS Tagging*. Un modelo de clasificación que se base en las etiquetas es menos vulnerable a un cambio o incremento de vocabulario respecto a la colección de entrenamiento.

La Tabla 3 muestra la frecuencia de algunas etiquetas dada la clase. Según la colección de opiniones sobre hoteles [7,6] se observa que en la clase de las opiniones verdaderas hay más sustantivos, artículos definidos, adjetivos, verbos en presente y singular, y palabras extranjeras que en la clase de opiniones falsas. Sin embargo, en esta última son más frecuentes los verbos en pasado y los pronombres posesivos.

Tabla 3. Frecuencia de algunas etiquetas según la clase.

POS Tag	**Verdaderas**	**Falsas**
Sustantivo (NN y NNS)	**30,228**	28,802
Artículo Definido (DT)	**14,332**	13,404
Adjetivos (JJ)	**10,843**	9,866
Verbo en presente y singular (VBZ y VBP)	**4,921**	4,041
Palabras Extranjeras (FW)	**28**	5
Verbo en pasado (VBD)	8,263	**9,076**
Pronombre Posesivo (PRP$)	1,911	**2,762**

Al analizar los bigramas de etiquetas (ver Tabla 4 y Tabla 5), o sea, secuencias consecutivas de dos etiquetas, es visible que en algunos casos la diferencia es por encima del doble. A pesar de los errores de cualquier etiquetador automático, se muestra que sí existen —al menos en esta colección— patrones o marcadores sintácticos interesantes como potenciales rasgos para la representación de opiniones.

Tabla 4. Frecuencia de algunos bigramas de etiquetas según la clase.

Bigramas de *Tags*	**Verdaderas**	**Falsas**
NN+FW	**14**	0
EX+VBZ	**74**	34
EX+VBP	**73**	33
WRB+NN	134	**269**
TO+PRP$	106	**214**

Tabla 5. Ejemplos de los bigramas representados en la Tabla 4.

	EX VBP		
NN FW	EX VBZ	WRB NN	TO PRP$
tv etc	there is	why business	to our
policy etc	there are	when dresser	to my
room etc	there isnt	why i	to its
someone knew	there weren't	how rude	to their
las vegas	there wasn't	where they'd	

Las etiquetas *PRP$*, *FW*, *WRB*, *VBZ* y *VB*[4] alcanzan ganancia de información mayor que cero, y si se usan únicamente estas etiquetas como rasgos

[4] Significado de las etiquetas usadas en el documento: artículo definido (DT), *There* (EX), palabra extranjera (FW), adjetivo (JJ), sustantivo en singular o no contable (NN), sustantivo en plural (NNS), pronombre posesivo (PRP$), preposición *to* (TO), verbo en pasado (VBD), verbo en su forma base (VB), verbo

para la clasificación de las opiniones -con pesado *tf* sin normalizar- se logra un 61.31 % de exactitud. Lo cual indica que pueden ser de ayuda para representar las opiniones y distinguir las notas engañosas. Además, usando unigramas y bigramas de etiquetas se obtiene un 66.44 % de exactitud.

5. Conclusiones y trabajos futuros

En este trabajo preliminar se han explorado técnicas simples de representación de las opiniones escritas, sin el empleo de recursos externos, reglas sintácticas ni análisis psicolingüístico, para comprobar su eficacia en la detección del engaño en opiniones. También se observaron algunas características de ambas clases según la colección de opiniones descrita en la Sec. 3.1.

Con los resultados mostrados en la sección anterior podemos afirmar que con técnicas más simples de representación de textos (bolsa de palabras y esquemas de pesados tradicionales) es posible alcanzar un 87.94 % de exactitud; y dependiendo del pesado y el pre-procesamiento, aproximadamente 0.90 en precisión. Seleccionar el pesado y pre-procesamiento adecuado dependerá de cuál es la prioridad: la detección de opiniones engañosas o de opiniones verdaderas.

Además de los pronombres en primera persona del singular [7], se hallaron otras palabras vacías útiles para caracterizar las opiniones falsas, así como otras palabras que por su alta frecuencia suelen ser removidas.

Finalmente se observó que la información sintáctica es importante. Incluso con solo cinco etiquetas sintácticas se puede alcanzar un desempeño arriba del azar.

En nuestro trabajo futuro consideramos pertinente comprobar cada una de estas observaciones sobre otras colecciones de opiniones e incluir la polaridad de las opiniones en el entrenamiento y la clasificación. También se recomienda una fase que consista en clasificar con el pesado *tf* y sin pre-procesamiento (ver Tabla 1), de esta forma se garantiza un buen recuerdo de las opiniones falsas, las cuales pueden ser re-analizadas para detectar las estructuras o características que las hacen ser engañosas.

Agradecimientos. Este trabajo ha sido realizado con el apoyo del Consejo Nacional de Ciencia y Tecnología (CONACyT) a través de la beca No. 613411, y del proyecto CONACYT PDCPN-2014-247870.

Referencias

1. Cincunegui, M., Berdun, F., Armentano, M.G., Amandi, A.: Clasificación de conductas colaborativas a partir de interacciones textuales. In: Argentine Symposium on Artificial Intelligence (ASAI 2015)-JAIIO 44 (Rosario, 2015) (2015)

en presente de la primera o segunda persona del singular (VBP), verbo en presente de la tercera persona del singular (VBZ), adverbio interrogativo (WRB). https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

2. Ezpeleta, E., Zurutuza, U., Hidalgo Gómez, J.M.: Los spammers no piensan: usando reconocimiento de personalidad para el filtrado de spam en mensajes cortos
3. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. pp. 171–175. Association for Computational Linguistics (2012)
4. Fusilier, D.H., Montes-y Gómez, M., Rosso, P., Cabrera, R.G.: Detecting positive and negative deceptive opinions using pu-learning. Information processing & management 51(4), 433–443 (2015)
5. Mihalcea, R., Strapparava, C.: The lie detector: Explorations in the automatic recognition of deceptive language. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. pp. 309–312. Association for Computational Linguistics (2009)
6. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: HLT-NAACL. pp. 497–501 (2013)
7. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 309–319. Association for Computational Linguistics (2011)
8. Razavi, A.H., Inkpen, D., Uritsky, S., Matwin, S.: Offensive language detection using multi-level classification. In: Canadian Conference on Artificial Intelligence. pp. 16–27. Springer (2010)