Congr. Int. en Ing. Electrónica. Mem. ELECTRO, Vol. 39, pp. 288-292, Oct 2017, Chihuahua, Chih. México
http://electro.itchihuahua.edu.mx/memorias_electro/MemoriaElectro2017.zip
ISSN 1405-2172

UNIGRAMAS, BIGRAMAS Y TRIGRAMAS MÁS UTILIZADOS EN EL OPINION SPAM

Donato Hernández[1], Sergio Ledesma[1], Rogelio Castro[1], Araceli Sánchez[2], Fernando Ireta[1] y Heriberto Gutiérrez[1]

[1] Universidad de Guanajuato
División de Ingenierías del Campus Irapuato-Salamanca
Carretera Salamanca-Valle de Santiago km 3.5+1.8

[2]Universidad Tecnológica de León
Electromecánica Industrial
Blvd. Universidad Tecnológica 225 Col. San Carlos

e-mail: donato@ugto.mx, selo@ugto.mx, castro@ugto.mx, salvarez@utleon.edu.mx, fireta@ugto.mx y hgmartin@ugto.mx

RESUMEN.

Hoy en día los consumidores en potencia toman sus decisiones influenciadas, cada vez más por los comentarios de otros usuarios en Internet, ignorando si se trata de un engaño. En esta investigación se obtiene el conjunto de unigramas, bigramas y trigramas empleados en las opiniones falsas, escritas en inglés, vertidas acerca de 20 hoteles situados en el centro de la ciudad de Chicago. Haciendo uso de herramientas de análisis de grandes volúmenes de información, como es la suite de métodos Weka y la metodología para la extracción del vocabulario. Se compararon un total de 800 opiniones verdaderas y 800 falsas, logrando con ello obtener el vocabulario semejante entre ellas y el conjunto de unigramas, bigramas y trigramas más utilizados en el opinion spam.
Palabras clave: n-gramas de palabras, opinión spam.

ABSTRACT.

Today potential consumers make their decisions influenced, more and more by the comments of other users on the Internet, ignoring if it is an opinion spam. In this research we obtain the set of unigramas, bigramas and trigrams used in the opinion spam, about 20 hotels located in downtown Chicago. Making use of tools of analysis of large volumes of information, such as the methods suite Weka and the methodology for the extraction of this vocabulary. A total of 800 true and 800 false opinions were compared, thus obtaining the similar vocabulary between them and the set of unigramas, bigramas and trigrams more used in opinion spam.
Keywords: word n-grams, opinion spam.

1. INTRODUCCIÓN

La cantidad de información en el mundo y en la vida diaria parece no tener fin. Sin embargo las computadoras lo hacen parecer fácil, almacenan nuestras decisiones, nuestros hábitos financieros, entradas y salidas. En otros ámbitos los datos crecen exponencialmente y nos vemos limitados en cuanto a la capacidad de análisis de todo ese volumen de información potencial, del cual no tomamos ventaja.

En el intento de sobresalir entre la competencia, es común observar una lucha desleal entre las empresas que operan bajo el mismo giro comercial. No es de sorprenderse si los consumidores son el objetivo principal. Aprovechando el Internet como un medio de difusión masiva, donde hoy en día la toma de decisiones de compra de los consumidores depende más de los comentarios escritos en el sitio web de la empresa.

Basta con hacer un par de clics o contratar a un grupo de personas para visitar la página de la competencia y comentar de manera negativa los servicios que ahí se ofrecen. Incluso ocurre el caso contrario: auto-criticarse de manera positiva. Hecho que no cae en lo éticamente correcto, pues el consumidor es víctima de engaño comercial y en muchas ocasiones da por sentado como verdadero lo que lee.

Existe una ciencia experimental; La Minería de Datos (Data Mining) que está definida como un proceso que descubre patrones en la información mediante un proceso automático o usualmente semiautomático.

Los patrones descubiertos tienen un significado y conducen a ciertas ventajas, por lo general, económicas. La minería de datos se encarga de resolver problemas analizando grandes volúmenes de información previamente presentes en las bases de datos. Emplea métodos de inteligencia artificial, aprendizaje automático, estadística y transforma la información en una estructura de datos comprensible.

Uno de los métodos más comunes para el análisis de los datos y descubrir los patrones existentes en los diferentes corpus, es el empleo de la suite de métodos denominada Weka [5]. El entorno de Weka es una colección de máquinas de aprendizaje, herramientas de pre-procesamiento y una gran cantidad de algoritmos para el análisis de la información. Este entorno fue diseñado para probar diferentes métodos en conjuntos de datos, de una manera flexible. Weka fue desarrollado en la Universidad de Waikato en Nueva Zelanda y significa: Waikato Environment for Knowledge Analysis, en español Entorno para el análisis del conocimiento de la Universidad de Waikato.

El objetivo de esta investigación es identificar el conjunto de unigramas, bigramas y trigramas más utilizados en el vocabulario del opinion spam, sin hacer distinción entre las opiniones positivas y negativas, mediante el uso de la suite de métodos Weka y otras herramientas desarrolladas para este propósito en particular.

En el análisis de texto es común hacer uso de un procesado llamado n-gramas de palabras, que consta de dividir un texto en sub-secuencias de palabras o letras según sea el estudio, en este caso de palabras. Los Unigramas, bigramas y trigramas son entonces secuencias de una palabra, dos palabras y tres palabras seguidas respectivamente [2].

Por mostrar un ejemplo: para la siguiente opinión "Mi permanencia en el hotel fue excelente". Los unigramas correspondientes serian: "Mi", "permanencia", "en", "el", "hotel", "fue", "excelente"; sus bigramas correspondientes serian: "Mi permanencia", "permanencia en", "en el", "el hotel", "hotel fue", "fue excelente" y su trigramas correspondientes serian: "Mi permanencia en", "permanencia en el", "en el hotel", "el hotel fue", "hotel fue excelente" .

El resto del artículo está organizado de la siguiente manera. La sección 2 presenta algunos trabajos relacionados en el campo del opinión spam. La sección 3 presenta la configuración experimental y el corpus utilizado. En la sección 4 se presentan los resultados experimentales. Finalmente, la sección 5 indica las principales aportaciones y algunas orientaciones de trabajo futuro.

2. TRABAJO RELACIONADO

La detección del opinión spam ha sido tradicionalmente estudiado utilizando varios atributos y métodos de codificación. En el contexto de correo electrónico [1], y páginas Web [2 y 9]. Por ejemplo, [10] propuso la detección de spam mediante la identificación de contenido duplicado. La cantidad de trabajos realizados en los últimos cinco años en la detección del opinión spam ha ido en aumento [3, 4, 6, 7, 8 y 9], todos ellos utilizan los n-gramas de palabras, como los atributos principales. En algunas ocasiones es interesante determinar el sentimiento de la opinión, es decir si es a favor (positiva) o en contra (negativa). Trabajos relacionados en este sentido han sido los publicados por [12, 13, 14 y 15].

Algunos trabajos han agregado atributos que corresponden al estilo de escritura como: LIWC (Lenguage Inquiry Word Count) [15, 16 y 17], series de patrones de tiempo (patterns time series) [11 y 20], lenguaje de modelado estadístico (Statistical models) [17 y 21] y estilometría [3], estos últimos como atributos secundarios para mejorar la eficacia.

Una tarea de la clasificación de texto, donde se han utilizado los n-gramas de palabras, es la atribución de autoría, obras de [18], muestran los resultados del uso de este tipo de atributos en esta tarea.

Este trabajo consiste en la detección de los n-gramas de palabras más utilizados en la redacción de las opiniones falsas de ambas polaridades (positivas y negativas), es decir el conjunto de palabras más comúnmente utilizado para realizar las opiniones que pueden engañar fácilmente a los consumidores.

3. CONFIGURACIÓN EXPERIMENTAL

El procesamiento de las opiniones falsas y verdaderas para la obtención de los n-gramas de palabras se realiza en dos partes: primero formando los n-gramas para las opiniones falsas (incluyendo negativas y positivas). Y segundo su análogo en las opiniones verdaderas. Constando de un total de 6 archivos a analizar en la suite Weka como se muestra en la Tabla 1.

Este procesamiento se realiza en un software ajeno a Weka, diseñado específicamente para este propósito, que consiste en enlistar todas las palabras que aparecen en las opiniones y compararlas contra otro grupo de opiniones.

Tabla 1: Procesamiento de los n-gramas de palabras.

N-gramas	Tipo de opiniones	
	Falsas	Verdaderas
Unigramas	×	×
Bigramas	×	×
Trigramas	×	×

En la Tabla 2 se muestra un ejemplo de cómo el software realiza esta comparación en los unigramas de las opiniones falsas. Se observa un valor de uno (1) en la columna cuando la palabra aparece en aquel tipo de opinión; un valor de cero (0) si no aparece. A este proceso se le conoce como pesado binario. El mismo proceso se realiza también para los bigramas y trigramas.

La suite Weka cuenta con métodos estadísticos, árboles de decisión, filtros para la selección de atributos, etcétera. La selección de atributos se realiza mediante la búsqueda en el espacio de subconjuntos de atributos.

La función InfoGainAttributeEval, realiza esta tarea de selección, consiste en medir cómo cada elemento contribuye en decrementos a la entropía global (análogo a los estudios realizados sobre la entropía en el área de termodinámica). En otras palabras es una medición de la cantidad de información que proporciona una palabra, que nos permita describir una clase en particular.

La entropía mide el *grado de impureza* de cada elemento. Entre más cercano sea a cero, menos impureza hay en el conjunto de datos. Por tanto, *un buen atributo* es aquel que *contiene más información*. En este estudio es aquel que más se usa en las opiniones falsas.

Tabla 2: Pesado binario en las opiniones falsas.

Unigramas (atributos)	Opiniones Falsas (clase)	
	Positiva	Negativa
1. Hotel	1	1
2. Chicago	1	1
3. Reservation	1	1
4. Dirty	0	1
5. Clean	1	0
6. México	0	0

Tabla 3: Comparación entre unigramas falsos y verdaderos.

Número de atributo	Falsas		Verdaderas	
	Entropía	Palabra	Entropía	Palabra
1.	**0.24684**	**not**	**0.11447**	**not**
2.	0.11835	rude	0.08343	told
3.	0.10279	finally	0.07533	great
4.	0.09814	but	0.0652	they
5.	0.08981	comfortable	0.06159	no

Primero se realiza esta selección para los seis casos de los n-gramas y posteriormente se comparan. El corpus con el que cuenta tiene en total 1600 opiniones repartidas en la misma cantidad entre positivas y negativas: 800 falsas y 800 verdaderas. Este corpus fue facilitado por Myle Ott [15] y fue elaborado por trabajadores que se contrataron con el propósito especial de escribir las opiniones con las características correspondientes a cada clase. El apartado denominado *clase* (positiva y negativa), es un distintivo que Weka utiliza para hacer el procesamiento de los *atributos*. Weka, requiere de por lo menos dos clases distintas. Los atributos, son elementos en la primera columna.

Las listas de los n-gramas falsos y positivos para encontrar el vocabulario repetido, como se muestra en la Tabla 3. Se observa que la palabra *not* es utilizada con más frecuencia, nótese cómo *not* tiene un valor más elevado en la columna que corresponde a las opiniones falsas que en la columna de las verdaderas.

4. RESULTADOS

Es posible apreciar en la Tabla 3 la distribución cuantitativa de la entropía, medida mediante el parámetro *informationGain*, de los n-gramas para las opiniones falsas y verdaderas. También es posible apreciar el *grado de impureza* antes descrito de cada atributo o vocablo, en otras palabras la densidad de aparición del mismo.

En la figura 1 se presenta la distribución cuantitativa de cada subconjunto de palabras. En esta figura se puede observar mejor una mayor densidad de vocablos utilizados en las opiniones falsas.

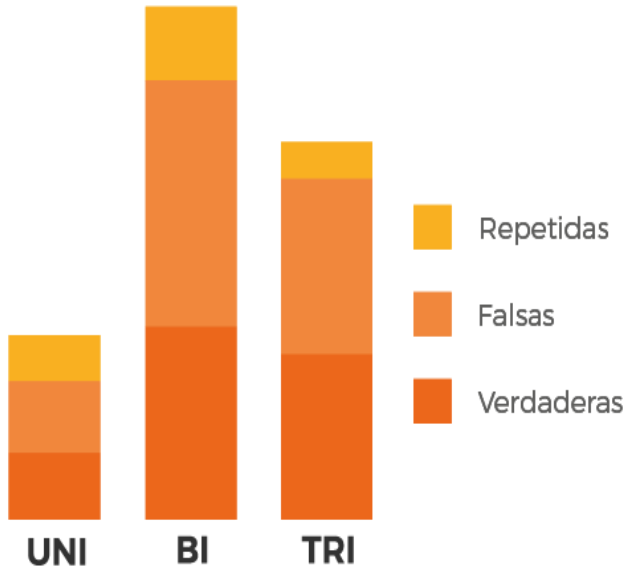

Fig. 1: Distribución de los n-gramas de palabras utilizados en las opiniones falsas.

El listado corto de 20 elementos del vocabulario en este tipo de opiniones se puede ver en la Tabla 4. Donde se puede observar que en las opiniones falsas se hace énfasis en el aspecto de los cuartos o haciendo mención de ellos, también recalca elementos con poca relevancia para el consumidor, como el sitio web del hotel. Además observe que hay alrededor de un 40% más de vocabulario en comparación con una opinión verdadera, (omitiendo el vocabulario repetido).

Tabla 4: Listado corto de vocabulario empleado en opiniones falsas

Vocabulario en opiniones falsas		
Unigramas	Bigramas	Trigramas
room	**to wait**	**had to wait**
arrived	**the room**	to check in
seemed	**the bathroom**	**the rooms were**
website	**the website**	**room was not**
smell	got to	there were no
smelled	**new room**	in the lobby
wait	were no	to the front
about	check in	when we got
bathroom	disappointed with	at the front
noise	close to	when i went
though	a bad	at the desk
for	a different	had not been
got	definitely be	not be staying
modern	there were	won t be
check	be a	would recommend this
found	**room and**	i would have
at	**room but**	i would recommend
elegant	i enjoyed	and there was
people	not very	**the room we**
booked	and there	to be a

Lo que indica que quienes mienten tienden escribir más palabras en una opinión. Para comprobar que el vocabulario encontrado permite predecir si una nueva opinión es falsa, es necesario volver a Weka y hacer uso de los métodos de aprendizaje. Luego obtener un conjunto de opiniones falsas llamada conjunto de prueba y someter la máquina al análisis de dicho conjunto.

Una lista más completa de las palabras utilizadas con mayor frecuencia en las opiniones falsas, puede ser consultada con el autor, ya que en esta publicación sólo es presentada una fracción.

Esta investigación se limitó sólo a encontrar el vocabulario, pasando por alto las reglas gramaticales, estilo de escritura y ambigüedades del idioma inglés. Incluso otras variables como el sexo, la edad y nivel educativo, de quienes realizaron estas opiniones. Estas características generalmente no están disponibles, pero si formaran parte de las opiniones falsas, pueden ser objeto de un estudio a futuro.

5. REFERENCIAS

[1] Drucker, H., Wu, D., & Vapnik, V. N. (2002). Support Vector Machines for Spam Categorization. IEEE Transactions on Neural Networks, August 2002, Volume: 10, Issue: 5, Page(s) 1048-1054.

[2] Feng, S., Banerjee, R. & Choi, Y. (2012). Syntactic Stylometry for Deception Detection. Association for Computational Linguistics, short paper. ACL.

[3] Feng, S., Xing, L., Gogar, A., & Choi, Y. (2012). Distributional Footprintsof Deceptive Product Reviews. In Proceedings of the 2012 International AAAI Conference on WebBlogs and Social Media, June.

[4] Gyongyi, Z., Garcia-Molina, H., & Pedersen, J. (2004). Combating Web Spam with Trust Rank. In Proceedings of the Thirtieth International conference on Very large data bases-Volume 30, 576-587. VLDB Endowment.

[5] Hall,M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA Data Mining Software: an Update. SIGKDD Explor. Newsl., 10-18.

[6] Hernández-Fusilier, D., Guzmán-Cabrera, R., Montes-y-Gómez, M., & Rosso, P. (2013). Using PU-learning to Detect Deceptive Opinion Spam. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, 38-45.

[7] Hernández-Fusilier, D., Montes-y-Gómez, M., Rosso, P., & Guzmán-Cabrera, R. (2014). Detecting Positive and Negative Deceptive Opinions using PU-learning. Information Processing & Management, DOI:10.1016/j.ipm.2014.11.001.

[8] Jindal, N., & Liu, B. (2008). Opinion Spam and Analysis. In Proceedings of the international conference on Web search and Web data mining, 219-230.

[9] Jindal, N., Liu, B., & Lim, E. (2010). Finding Unusual Review Patterns Using Unexpected Rules. CIKM '10, Proceedings of the 19th ACM International conference on Information and knowledge management, October 2010, Page(s) 210-220.

[10] Kanaris, I., Kanaris, K., Houvardas, I. & Stamatatos, E. (2007). Word versus character n-grams for anti-spam filtering. International Journal on Artificial Intelligence Tools, Volume 16 Issue 6, 1047-1067.

[11] Lim, E. P., Nguyen, V. A., Jindal, N., Liu, B., & Lauw, H. W. (2010). Detecting Product Review Spammers Using Rating Behaviours. In CIKM, 939-948.

[12] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lecture on Human Language Technologies. Morgan & Claypool Publishers.

[13] Mukherjee, A., Liu, B., Wang, J., Glance, N. & Jindal, N. (2011). Detecting Group Review Spam. Proceedings of the 20th international conference companion on World wide Web, 93-94.

[14] Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. (2006). Detecting Spam Web Pages through Content Analysis. Transactions on Management Information Systems (TMIS), 83-92.

[15] Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). 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, Portland, Oregon, USA, 309-319.

[16] Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative Deceptive Opinion Spam. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, 309-319.

[17] Raymond, Y. K., Lau, S. Y., Liao, R., Chi-Wai, K., Kaiquan, X., Yunqing, X., & Yuefeng, L. (2011). Text Mining and Probabilistic Modeling for Online Review Spam Detection. ACM Transactions on Management Information Systems, December 2011, Volume: 2, Issue: 4, Article: 25, Page(s) 1-30.

[18] Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. Journal of Law & Policy, Vol. 21 No. 2.

[19] Wu, G., Greene, D. & Cunningham, P. (2010). Merging Multiple Criteria to Identify Suspicious Reviews. RecSys10, 241-244.

[20] Xie, S., Wang, G., Lin, S., & Yu, P. S. (2012). Review Spam Detection via Time Series Pattern Discovery. Proceedings of the 21st international conference companion on World Wide Web, 635-636.

[21] Zhou, L., Sh, Y., & Zhang, D. (2008). A Statistical Language Modeling Approach to Online Deception Detection. IEEE Transactions on Knowledge and Data Engineering, 20(8), 1077-1081.