JÓVENES EN LA CIENCIA
REVISTA DE DIVULGACIÓN CIENTÍFICA

VOCABULARIO EMPLEADO EN LAS OPINIONES FALSAS

Gutiérrez Santoyo Fernando1, Hernández Fusilier Donato2

[1] Licenciatura en Comunicaciones y Electrónica, Universidad de Guanajuato | f.gutierrezsantoyo@ugto.mx

[2] Departamento de Ingeniería Electrónica, División de Ingenierías, Campus Irapuato - Salamanca, Universidad de Guanajuato | donato@ugto.mx

RESUMEN

Los consumidores en potencia de hoy en día toman decisiones influenciadas cada vez más por los comentarios de otros usuarios en internet ignorando si se trata de un engaño. La investigación obtiene el vocabulario empleado en las opiniones engañosas de hoteles situados en Chicago. Haciendo uso de herramientas de análisis de grandes volúmenes de información, como Weka, y la metodología para la extracción de dicho vocabulario comparando 800 opiniones verdaderas y 800 falsas en las mismas cantidades; donde se excluye el vocabulario semejante entre ellas.

ABSTRACT

Purchase decision making of nowadays consumers is highly influenced by online reviews that makes people get misleading information. In this study we look for vocabulary used in deceptive opinion on twenty Chicago hotels. We then use Data Mining software based assessment methods to spot deceptive opinions on 800 truthful reviews and 800 deceptive reviews, where we exclude similar vocables.

Palabras Clave DATA MINING; DECEPTIVE; HOTEL; CHICAGO

INTRODUCCIÓN

La cantidad de información en el mundo y en la vida diaria parece no tener fin. Las computadoras lo hacen parecer fácil, almacenan nuestras decisiones, nuestros hábitos financieros, entradas y salidas. Y qué decir de otros ámbitos. Los datos crecen; y nos vemos limitados en cuanto a capacidad de análisis de todo ese volumen de información potencial de la cual no tomamos ventaja. Sin embargo, existe una ciencia experimental que sí lo hace.

Minería de Datos

La minería de datos (Data Mining en inglés) está definida como un proceso que descubre patrones en la información. El proceso debe ser automático o (usualmente) semiautomático.

Los patrones descubiertos tienen un significado y conducen a ciertas ventajas, por lo general, económicas. Por lo cual, la minería de datos se encarga de resolver problemas analizando grandes volúmenes de información previamente presente en bases de datos. Emplea métodos de inteligencia artificial, aprendizaje automático, estadística y transforma la información en una estructura de datos comprensible. [1]

La experiencia muestra que no hay máquinas de aprendizaje universales para todos los problemas del Data Mining y no todos los métodos se acoplan a los diversos tipos de datos.

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. Diseñado para probar diferentes métodos en conjuntos de datos de 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. (Witten, I. H., 2011: 403)

Comportamiento anticompetitivo

En el intento de sobresalir de 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 da por sentado como verdadero lo que lee.

El objetivo de esta investigación identifica el vocabulario empleado en las opiniones falsas (sin hacer distinción entre positivas y negativas), hechas en 20 hoteles en la ciudad de Chicago, EE.UU, mediante el uso de Weka y otras herramientas.

MATERIALES Y MÉTODOS

N-gramas para análisis de texto

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

Por mostrar un ejemplo: “Mi permanencia en el hotel fue excelente” sus bigramas correspondientes son: “Mi permanencia”, “permanencia en”, “en el”, “el hotel”, “hotel fue”, “fue excelente”.

Se poseen en total 1600 opiniones repartidas en la misma cantidad entre positivas y negativas: 800 falsas y 800 verdaderas.

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

Este procesamiento se realiza en un software ajeno a Weka, diseñado por el investigador, que consta en enlistar todas las palabras que aparecen en las opiniones y compararlas contra otro grupo de opiniones.

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

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. (Witten, I.H., 2011: 39-60)

Tabla 2: Pesado binario en las opiniones falsas.

Unigramas (atributos)	Opiniones Falsas (clase)	
	Positiva	Negativa
1. Hotel	1	1
2. Chicago	1	1
3. Reservación	1	1
4. Sucio	0	1
5. Limpio	1	0
6. México	0	0

Filtrado y selección de atributos

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.

Tabla 1: Procesamiento de n-gramas.

N-gramas	Tipo de opiniones	
	Falsas	Verdaderas
1. Unigramas	×	×
2. Bigramas	×	×
3. Trigramas	×	×

La función InfoGainAttributeEval, realiza esta tarea de selección, y consiste en medir cómo cada elemento contribuye en decrementos a la entropía global (análogo a los estudios realizados sobre la entropía termodinámica).

En otras palabras, 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*.[3] Y en este estudio es aquel que más se usa en las opiniones falsas.

Se realiza esta selección para los seis n-gramas. Luego se comparan 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.

El vocabulario repetido es excluido, puesto que se tratan de palabras que inevitablemente serán usadas en ambos tipos de opiniones.

RESULTADOS Y DISCUSIÓN

Se puede apreciar en Tabla 4 la distribución cuantitativa y la suma de la entropía de los n-gramas falsos y los verdaderos. Si contase con más espacio, se podría apreciar el *grado de*

impureza (entropía) antes descrito de cada atributo o vocablo, en otras palabras, la densidad de aparición del mismo.

El Gráfico 1 presenta la distribución cuantitativa de cada subconjunto de palabras. En él se puede observar mejor una mayor densidad de vocablos utilizados en opiniones falsas.

El listado corto de 20 elementos del vocabulario en este tipo de opiniones se puede ver en la Tabla 5. De modo que se nota que las opiniones falsas hacen 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). Lo que indica que quienes mienten tienden escribir más palabras en una opinión.

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

CONCLUSIONES

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 y someter la máquina a prueba.

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, nivel educativo, de quienes realizaron estas opiniones y que pueden ser objeto de estudio a futuro.

El listado completo de palabras puede consultarse con el autor, aquí sólo es presentada una fracción debido al espacio disponible para la publicación.

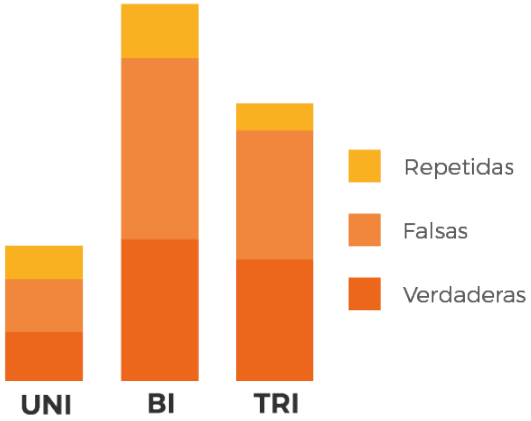

GRÁFICO 1: Distribución cuantitativa de los n-gramas.

AGRADECIMIENTOS

Este trabajo fue posible en gran parte por la Universidad de Guanajuato. Gracias al Dr. Donato Hernández Fusilier por sus consejos y sigerencias. También quiero extender mi agradecimiento a los hoteles que prestaron la información para el análisis.

REFERENCIAS

[1] Witten, I. H., Eibe, Hall (2011) Data mining : practical machine learning tools and techniques. Elseiver, (3rd ed.) (pp. 5)
[2] EcuRed. Conocimiento con todos y para todos. N-gramas. Recuperado de http://www.ecured.cu/N-grama (Julio 2016)
[3] Sección de preguntas en StackOverflow. Recuperado de http://stackoverflow.com/questions/33982943/how-the-selection-happens-in-infogainattributeeval-in-weka-feature-selection (julio 2016)

Tabla 5: 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