

Análisis de polaridad en Twitter

Gilberto Leon Martagón, Esaú Villatoro-Tello,
Héctor Jiménez-Salazar y Christian Sánchez-Sánchez

Departamento de Tecnologías de la Información,
División de Ciencias de la Comunicación y Diseño,
Universidad Autónoma Metropolitana Unidad Cuajimalpa, México DF
{2123805605, evillatoro, hjimenez, csanchez}@correo.cua.uam.mx

Resumen En este artículo se propone una técnica supervisada para determinar la polaridad/sentimiento de mensajes publicados en Twitter. Nuestro método propone una forma de representación vectorial, donde a partir de una selección de términos basada en la Ley de Zipf se muestra que es posible obtener resultados favorables al momento de realizar la clasificación de mensajes de acuerdo a su polaridad. La representación propuesta obedece a las características léxicas propias que surgen del uso de esta plataforma de microblogueo, así como las limitantes impuestas por la misma, tal y como el número limitado de caracteres permitidos, las etiquetas, las menciones y los enlaces Web. Los experimentos que se realizaron, sobre una colección de tuits de temáticas variadas, muestran que los atributos empleados para representar tuits permiten obtener un desempeño favorable al momento de identificar el sentimiento de los usuarios en sus publicaciones.

Palabras clave: análisis de polaridad, clasificación de textos, procesamiento de lenguaje natural, aprendizaje automático.

1. Introducción

Twitter, lanzado en 2006, sin lugar a dudas es una de las plataformas de microblogueo más importantes hoy en día. Las últimas cifras sobre su uso revelan que hay más de 200 millones de usuarios que escriben más de 400 millones de tuits cada día [1], sobre una gran variedad de temas. Debido a la aceptación que esta plataforma ha conseguido, diversas entidades, entre las que podemos contar a empresas comerciales, celebridades y políticos, se encuentran inmersas en una constante búsqueda de posicionamiento para poder influenciar de forma eficiente en la opinión de algún grupo demográfico de usuarios suscritos a Twitter.

Por lo tanto, uno de los esfuerzos involucrados en este posicionamiento consiste en obtener una idea de la opinión¹ (*i.e.*, positiva, negativa y/o neutral) sobre un tema, producto o servicio. Sin embargo, hacerlo a mano se convierte en un proceso imposible debido a la gran cantidad de información que es producida

¹ El análisis de opinión es también conocido como análisis de polaridad y/o de sentimientos.

en este tipo de redes sociales. De esto surge la necesidad de contar con herramientas automáticas que permitan hacer un análisis de polaridad de manera confiable y eficiente.

Actualmente, existen diversas técnicas para hacer análisis de polaridad en textos largos en diversos idiomas y dominios. Sin embargo, las características impuestas a través de la plataforma de Twitter, como lo es la longitud máxima de los mensajes, motivan a los usuarios a innovar en su forma de expresión, provocando con esto el surgimiento de un lenguaje propio de la plataforma, el cual varía dependiendo del idioma y del contexto (social y/o cultural) en el que estas formas de expresión son publicadas. Es importante mencionar que todo esto tienen un impacto negativo en el desempeño de técnicas tradicionales de clasificación de polaridad en textos largos, por lo tanto, cualquier intento de análisis de tuits debe tener en cuenta las características propias de Twitter.

En este trabajo proponemos una representación de las publicaciones en Twitter que toma en cuenta las características del lenguaje empleado en la plataforma. Así por ejemplo, nuestra propuesta toma en cuenta el uso de etiquetas (*hashtags*), las cuales pueden formar parte sintáctica y/o semántica de un tuit; agregado a esto también consideramos las menciones a otros usuarios, con las que ocurre lo mismo que con las etiquetas; finalmente, los enlaces a sitios Web son también considerados por nuestra forma de representación propuesta.

Adicionalmente, se propone como forma de selección de las palabras más representativas de un tuit, el uso del punto de transición, *i.e.*, se toma en cuenta la frontera entre las palabras de baja y alta frecuencia, derivado de la Ley de Zipf. Los resultados experimentales muestran que las consideraciones hechas sobre el lenguaje propio de twitter así como la forma de selección de atributos permiten obtener un desempeño favorable al momento de hacer la clasificación de polaridad.

El resto de este trabajo está organizado de la siguiente manera: la sección 2 describe brevemente el trabajo relacionado, la sección 3 describe el método propuesto mientras que la sección 4 describe los datos empleados así como los experimentos realizados. Finalmente, la sección 5 expresa las conclusiones obtenidas durante la realización de este trabajo.

2. Trabajo relacionado

En los trabajos propuestos previamente, existen algunos que toman en cuenta algunas o varias de las características propias de Twitter, como lo es el uso de emoticonos. En el trabajo propuesto en [2] de un conjunto inicial de 60 emoticonos identificados automáticamente dentro de mensajes de twitter, se decidió conservar únicamente 15 de uso común, los cuales fueron asociando manualmente con polaridades (*i.e.*, sentimientos) específicas. Una de las características principales de este subconjunto de emoticonos es que son representados a partir de pocos caracteres ASCII y cuentan con un significado claro y preciso, *e.g.*, los mostrados en las primeras dos columnas de la Tabla 1.

El problema de considerar un conjunto de emoticonos con estas características es que no refleja la variedad de emoticonos que se usan actualmente en la plataforma de twitter. Es normal encontrar emoticonos de más de 5 caracteres que representan una idea más compleja y menos específica que aquellas que reflejan los emoticonos simples. De esta forma, cualquier intento de considerar emoticonos para hacer análisis de polaridad debe contar con un diccionario extenso de estos nuevos emoticonos, los cuales a su vez pueden combinarse para generar nuevas expresiones y/o emociones (Tabla 1), lo cual no es el propósito del presente trabajo.

Tabla 1. Comparación de emoticonos simples VS emoticonos compuestos.

Emoticonos Simples	Significado	Emoticonos Compuestos	Significado
:)	Alegría		Persona enojada lanzado un escritorio
:(Tristeza		Persona cantando
;))	Guiño		Persona cantando
:D	Risa		Amigos chocando manos
XD	Burla		Oso inexpressivo

Además de la limitante anterior, hay muchos tuits que utilizan uno o más emoticonos de manera irónica, resultando en una publicación que contiene una idea contraria a la polaridad definida por el emoticono. Dado esto, sistemas como el propuesto en [2] requieren considerar información del contexto para poder hacer una adecuada clasificación.

Algunos otros trabajos consideran las palabras contenidas en los tuits para realizar la clasificación de polaridad [3,4,5]. En éstos se concluye que la forma más eficiente de representación de los tuits es por medio del uso de unigramas de palabras en vez de bigramas. Sin embargo, una debilidad de este tipo de trabajos es que realizan experimentos sobre dominios y/o temáticas bien definidas, y además dependen de algún recurso léxico.

La propuesta presentada en este artículo, a diferencia de trabajos previos, no considera el uso de emoticonos, ni ningún otro tipo de recurso léxico para la clasificación de polaridad en tuits. Nuestra propuesta considera que el uso de las etiquetas (*#hashtags*), las menciones a otros usuarios (*@Usuario*) y los enlaces a páginas Web que aparecen dentro de los tuits pueden ser útiles para determinar la polaridad de los mismos. Agregado a esto, mostramos que considerando las palabras cercanas al punto de transición, es posible lograr un desempeño adecuado en la clasificación de la polaridad de tuits de temáticas diversas.

3. Método propuesto

3.1. Preprocesamiento de los tuits

Previo al proceso de entrenamiento y clasificación se realizó un preprocesamiento a los tuits el cual consistió en los siguientes pasos:

- Los tuits se convierten a minúsculas con la finalidad de normalizar el vocabulario.
- Cualquier secuencia de espacios en blanco se convierte en un solo espacio.
- Cualquier mención a usuarios encontrada se convierte a la cadena “PARA-USUARIO”, de modo que se conserve la intención de las menciones, pero como un único elemento del vocabulario.
- Cualquier enlace encontrado se convierte a la cadena “ENLACE-SALIENTE”, de modo que se conserve la intención de los enlaces, pero como un mismo elemento del vocabulario. Es importante mencionar que no se hace uso de la información contenida en cada sitio, sin embargo, se cree que el uso de esta información puede enriquecer el modelo propuesto.
- Se eliminan los signos de puntuación. Esto también elimina cualquier emoticono que pudiera aparecer, puesto que no fueron tomados en cuenta para el funcionamiento de este modelo.
- Cada palabra de un tuit es llevada a su raíz léxica. Este proceso se lleva a cabo mediante el lematizador Porter [6].
- Se eliminan las palabras vacías y/o funcionales.

3.2. Forma de representación de los tuits

En este trabajo se aborda la problemática del análisis de polaridad en tuits desde el paradigma de clasificación de textos (CT)². Bajo este paradigma un primer paso necesario es el *indexado* de los documentos de entrenamiento (Tr), actividad que denota hacer el mapeo de un documento d_j en una forma compacta de su contenido. La representación más comúnmente utilizada para representar cada documento es un vector con términos ponderados como entradas, concepto tomado del modelo de espacio vectorial usado en recuperación de información [8]. Es decir, un texto d_j es representado como el vector $\vec{d}_j = \langle w_{kj}, \dots, w_{|\tau|j} \rangle$, donde τ es el *diccionario*, *i.e.*, el conjunto de términos que ocurren al menos una vez en algún documento de Tr , mientras que w_{kj} representa la importancia del término t_k dentro del contenido del documento d_j . En ocasiones τ es el resultado de filtrar las palabras del vocabulario, *i.e.*, resultado de un preprocesamiento (Sección 3.1). Una vez que hemos hecho los filtrados necesarios, el diccionario τ puede definirse de acuerdo a diferentes criterios, sin embargo el que se empleó en esta propuesta corresponde a la Bolsa de Palabras.

² La Clasificación de Textos es la tarea de asociar automáticamente categorías predefinidas con documentos a partir del análisis de su contenido [7].

La Bolsa de Palabras (BOW)³ es la forma tradicionalmente utilizada para representar los documentos [7]. Este método de representación utiliza a las palabras simples como los elementos del vector de términos.

Con respecto al peso w_{kj} , se tienen diferentes formas de calcularlo, entre las más usadas en la comunidad científica se tienen el ponderado booleano y ponderado por frecuencia relativa de términos. Una breve descripción es dada a continuación:

- *Ponderado Booleano*: Consiste en asignar el peso de 1 si la palabra ocurre en el documento y 0 en otro caso.

$$w_{kj} = \begin{cases} 1, & \text{si } t_k \in d_j \\ 0, & \text{en otro caso} \end{cases} \quad (1)$$

- *Ponderado por frecuencia relativa (TF-IDF)*: Este tipo de ponderado es una variación del tipo anterior y se calcula de la siguiente forma:

$$w_{kj} = TF(t_k) \times IDF(t_k) \quad (2)$$

donde $TF(t_k)$ es la frecuencia del termino t_k en el documento d_j . IDF es conocido como la “frecuencia inversa” del termino t_k dentro del documento d_j . El valor de IDF es una manera de medir la “rareza” del termino t_k . Para calcular el valor de IDF se utiliza la siguiente ecuación:

$$IDF(t_k) = \log \frac{|D|}{|\{d_j \in D : t_k \in d_j\}|} \quad (3)$$

donde D representa la colección de documentos que está siendo indexada.

Este tipo de técnicas trae un costo agregado, que es el producir un espacio de términos (atributos) τ de alta dimensionalidad (i.e., $|\tau| \rightarrow \infty$). Este problema puede ocasionar problemas de *sobre-ajuste* en el proceso de aprendizaje, i.e., ocurre el fenómeno por medio del cual un clasificador se adapta a las características contingentes de Tr , en lugar de únicamente a las características constitutivas de las categorías, provocando problemas de efectividad debido a que el clasificador tiende a comportarse mejor sobre los datos con los que ha sido entrenado y sin conservar la tendencia en aquellos no vistos.

Uno de los métodos que ha mostrado ser efectivo como técnica de reducción de dimensionalidad es por medio de conservar los términos que se encuentran alrededor del punto de transición (pt_T) [9]. El tp_T es un valor de frecuencia que divide a los términos del vocabulario τ en dos conjuntos de términos, los de baja y alta frecuencia [10,11]. En los estudios realizados en [9,10,11] se demuestra que los términos de frecuencia media están fuertemente relacionados con el contenido de los documentos, lo cual permite resolver efectivamente tareas como la clasificación temática de textos. Nuestra hipótesis es que por medio de utilizar

³ En inglés se conoce como Bag Of Words.

un subconjunto de términos de frecuencias medias a altas es posible clasificar la polaridad en los tuits. La forma tradicional de calcular el punto de transición es:

$$tp_T = \frac{\sqrt{8 * I_1 + 1} - 1}{2} \quad (4)$$

donde I_1 representa el número de palabras con frecuencia 1 en el vocabulario τ .

3.3. Métodos de clasificación

El clasificador Naïve Bayes (*NB*) se considera como parte de los clasificadores probabilísticos, los cuales se basan en la suposición que las cantidades de interés se rigen por distribuciones de probabilidad, y que la decisión óptima puede tomarse por medio de razonar acerca de esas probabilidades junto con los datos observados [12]. Para nuestro trabajo empleamos el Naïve Bayes tradicional, el cual se describe a continuación.

En este esquema el clasificador es construido usando Tr para estimar la probabilidad de cada clase. Entonces, cuando una nueva instancia (documento) d_j es presentada, el clasificador le asigna la categoría $c \in C$ más probable por aplicar la regla:

$$c = \arg \max_{c_i \in C} P(c_i | d_j) \quad (5)$$

utilizando el teorema de Bayes para estimar la probabilidad tenemos:

$$c = \arg \max_{c_i \in C} \frac{P(d_j | c_i) P(c_i)}{P(d_j)} \quad (6)$$

dado que el denominador en la ecuación anterior no difiere entre categorías puede omitirse quedando de la siguiente forma:

$$c = \arg \max_{c_i \in C} P(d_j | c_i) P(c_i) \quad (7)$$

tomando en cuenta que el esquema es “naïve” (*i.e.*, se presupone la independencia entre atributos), se asume que las características son condicionalmente independientes dadas las clases. Esto simplifica los cálculos produciendo:

$$c = \arg \max_{c_i \in C} P(c_i) \prod_{k=1}^n P(d_{kj} | c_i) \quad (8)$$

donde $P(c_i)$ es la fracción de documentos en Tr que pertenecen a la clase c_i , es decir:

$$P(c_i) = \frac{|Tr_{c_i}|}{|Tr|} \quad (9)$$

y $P(d_{kj} | c_i)$ se calcula de acuerdo a:

$$P(d_{kj} | c_i) = \frac{1 + |Tr_{ki}|}{|A| + \sum_{l=1}^{|A|} |Tr_{li}|} \quad (10)$$

donde $|Tr_{ki}|$ es el número de documentos dentro del conjunto de entrenamiento que poseen el atributo k y además pertenecen a la clase c_i (i.e., d_{ki}), $|A|$ es el número total de atributos, de esta forma $|Tr_{li}|$ es el número de ejemplos de entrenamiento que poseen el atributo l y pertenecen a la clase c_i .

En resumen, la tarea de aprendizaje en el clasificador Naïve Bayes consiste en construir una hipótesis por medio de estimar las diferentes probabilidades $P(c_i)$ y $P(d_{kj}|c_i)$ en términos de sus frecuencias sobre Tr .

3.4. Evaluación

Para evaluar un sistema de clasificación de texto se utilizan las medidas de *Precisión* y *Recuerdo*, que son medidas comunes en el área de recuperación de información. La precisión (P) es la proporción de documentos clasificados correctamente en una clase c_i con respecto a la cantidad de documentos clasificados en esa misma clase. El recuerdo (R), la proporción de documentos clasificados correctamente en una clase c_i con respecto a la cantidad de documentos que realmente pertenecen a esa clase. Así, la precisión se puede ver como una medida de la corrección del sistema, mientras que el recuerdo da una medida de cobertura o completitud.

Adicionalmente, es común emplear la medida- F para describir el comportamiento de la clasificación, la cual se define como:

$$medida - F = \frac{(1 + \beta^2)Precision * Recuerdo}{\beta^2 Precision + Recuerdo} \quad (11)$$

donde con $\beta = 1$ representa la media armónica entre la precisión y el recuerdo. La función de β es la de controlar la importancia relativa entre las medidas de precisión y recuerdo. Es común asignar un valor de 1 indicando igual importancia a ambas medidas.

4. Pruebas realizadas

En esta sección se describen el conjunto de datos empleados para la realización de nuestros experimentos así como la configuración empleada en cada caso propuesto. Es importante mencionar que para todos los experimentos se empleó como método de clasificación el algoritmo de Naïve Bayes.

4.1. Conjunto de datos

Para la realización de nuestros experimentos trabajamos con los Tuits proporcionados por el SemEval 2013⁴. Los tuits proporcionados por el foro de evaluación representan mensajes de una gran variedad de temas entre los cuales se encuentra información de personajes públicos, productos y eventos.

⁴ <http://www.cs.york.ac.uk/semeval-2013/task2/>

Los tuits proporcionados para entrenamiento están etiquetados en tres clases, *positivo*, *negativo* y *neutral*. A diferencia de trabajos previos, el trabajar con esta colección representa un reto mayor pues el problema de clasificación no es binario. La distribución de los datos de entrenamiento fue de 950 tuits por clase. Estos datos fueron utilizados para entrenar y generar nuestros modelos de clasificación.

Agregado a los datos de entrenamiento, el SemEval 2013 proporcionó un conjunto de datos de prueba (2,845 en total) distribuidos de la siguiente manera: 740 negativos, 975 positivos, y 1,135 neutros. Estos datos fueron utilizados para evaluar nuestra propuesta.

4.2. Método base

Como método base se empleó una forma de representación la típica BOW empleando un esquema de pesado *booleano* (Sección 3). El objetivo de este experimento fue determinar la complejidad del problema de clasificación de polaridad por medio de técnicas tradicionales de CT.

4.3. Experimentos

Para los siguientes experimentos se utilizó como forma de representación una BOW empleando un esquema de pesado TF-IDF. Al mismo tiempo se utilizó como estrategia de reducción de dimensionalidad la selección de los términos alrededor del tp_T . En la tabla 2 aparecen como renglones: el porcentaje de frecuencia tomado alrededor del tp_T , la frecuencia más baja considerada, la frecuencia más alta considerada, y en el último renglón el tipo de frecuencias usadas por cada experimento.

Tabla 2. Resultados obtenidos bajo las diferentes configuraciones propuestas.

Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
23 %	23 %	79 %	89 %	94 %
69	69	20	10	5
117	759	759	759	759
medias	media-alta	media-baja	baja	muy-bajas

Es importante mencionar que sólo para el experimento 1 se utilizó el rango alrededor del tp_T para definir los límites inferiores y superiores de la frecuencia de los términos, mientras que para los experimentos 2 a 5 el límite superior se dejó abierto a la frecuencia más alta, *i.e.*, 759.

4.4. Resultados

La tabla 3 muestra los resultados obtenidos de los experimentos planteados. La primer columna hace referencia al nombre del experimento, mientras que la

Tabla 3. Resultados obtenidos bajo las diferentes configuraciones propuestas.

Nombre experimento	Num. atributos	Tipo pesado	Precisión (P)	Recuerdo (R)	Medida (F_β)
Base	7,620	Booleano	0.35	0.35	0.35
Exp 1	22	TF-IDF	0.36	0.35	0.35
Exp 2	41	TF-IDF	0.44	0.44	0.44
Exp 3	238	TF-IDF	0.54	0.54	0.54
Exp 4	539	TF-IDF	0.59	0.60	0.60
Exp 5	1,399	TF-IDF	0.64	0.65	0.65

segunda y tercer columna indican el esquema de pesado empleado en la representación y el número de atributos (términos) considerados en la representación.

Como es posible observar en la Tabla 3 las técnicas tradicionales de CT (*i.e.*, Experimento 1) no resultan ser apropiadas para el problema de clasificación de polaridad. Este hecho es mostrado una vez más con los resultados obtenidos en el Experimento 2, donde al usar el punto de transición en su forma tradicional, se obtiene un desempeño similar al Experimento 1, mostrando con esto que para identificar la polaridad de los tuits es necesario incluir términos que van más allá del contenido de los mensajes.

Los resultados del Experimento 2 muestran que al ampliar el rango de los términos alrededor del punto de transición es posible mejorar los resultados de clasificación. Estos resultados nos motivaron a hacer experimentos ampliando el rango de frecuencias bajas. Como es posible ver en los resultados del Experimento 5, por medio de utilizar tan sólo un vector de atributos de 1399 elementos (18.35 % del tamaño original del vector de términos) es posible alcanzar un valor de *medida* – $F = 0,65$, lo que significó una mejora de más de un 82 % respecto de la configuración base.

5. Conclusiones

La clasificación de polaridad de tuits comparte algunos elementos y características que tiene la clasificación de polaridad de textos más largos. Sin embargo, hay suficientes diferencias como para que las técnicas útiles en el segundo caso necesiten ser ajustadas para probar ser eficientes en el contexto de Twitter.

Como puede verse en los experimentos realizados, la representación adecuada de cada elemento a clasificar es crucial para esta labor, y esto involucra tanto el tamaño de la representación como la calidad de la información que contenga.

El método base, que utiliza una BOW *binaria* resulta en un modelo con mucha información poco relevante, que tarda mucho en entrenarse y que produce un modelo de clasificación altamente sesgado. Por el contrario, el método propuesto aprovecha la cantidad de información que cada palabra puede aportar en el caso de cada tuit en relación al conjunto de tuits en general, por lo que se muestra como una mejor solución para el problema de la clasificación de polaridad automática.

Nuestros experimentos mostraron que es necesario definir formas automáticas para definir el número de atributos que se utilizarán para la representación de

los elementos a clasificar, ya que, por un lado, un número pequeño genera un conjunto de atributos poco informativos para el modelo de clasificación. Por otro lado, si el número de atributos es demasiado grande, estos atributos introducen ruido que resulta en un sesgo significativo del clasificador, resultando en un desempeño deficiente.

Los resultados obtenidos nos motivan a seguir trabajando sobre esta línea. Como trabajo futuro se planea experimentar con atributos más complejos, por ejemplo bi-gramas de palabras, con los cuales se cree que la representación de los tuits puede ser enriquecida. Agregado a esto, se considera experimentar con el uso de los emoticonos, los cuales han mostrado aportar información en trabajos previos.

Agradecimientos. Agradecemos a la Universidad Autónoma Metropolitana Unidad Cuajimalpa y al proyecto SEP-PROMEP número 48510294 (UAM-CA-31) por el apoyo para la asistencia a este evento.

Referencias

1. Wickre, K. (2013). Celebrating #Twitter7. En <http://blog.twitter.com/2013/03/celebrating-twitter7.html> (Última visita en Marzo 21 de 2013.)
2. Davidiv, D., Tsur, O. y Rappoport, A. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. En *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (Coling 2010)*. pp. 241-249
3. Pang, B., Lee, L., y Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. En *Proceedings of the Empirical methods in natural language processing EMNLP'02*. pp. 79-86.
4. Pang, B. y Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. En *Proceedings de la Association for Computational Linguistics*. pp. 271-278.
5. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., y Gordon, J. (2012) Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. LNAI Vol. 7629, pp. 1-14.
6. Porter, M. F. (1997) An algorithm for suffix stripping. Morgan Kaufmann Publishers Inc. pp. 313-316.
7. Sebastiani F. (2002) Machine Learning in Automated Text Categorization. En *ACM Computing Surveys*, Vol. 34, No. 1, March 2002, pp. 1-47.
8. Baeza-Yates, R., y Ribeiro-Neto, B. (1999) Modern Information Retrieval, Addison Wesley.
9. Reyes-Aguirre, B., Moyotl-Hernández, E., y Jiménez-Salazar, H. (2003) Reducción de términos índice usando el punto de transición. En *Avances en Ciencias de la Computación*. pp. 127-130.
10. Zipf, G. K. (1949) Human Behavior and the Principle of Least-Effort, Addison-Wesley, Cambridge MA.
11. Booth, A. D. (1967) A law of occurrences for words of low frequency. En *Information and Control*. Vol. 10(4). pp. 386-393.
12. Mitchell, T. (1997) Machine Learning. McGraw-Hill.