ISSN 2007-9737

Clasificación automática de opiniones en dominios cruzados

Rafael Guzmán Cabrera

Universidad de Guanajuato Departamento de Ingeniería Eléctrica,
División de Ingenierías campus Irapuato Salamanca,
México

guzmanc@ugto.mx

Resumen. Cada vez es más común que los usuarios de internet tengan acceso a blog´s y redes sociales. En estos sitios es común emitir opiniones. Las opiniones permiten medir la percepción de las personas respecto a un tópico o producto determinado. Cuando el número de opiniones es muy grande su análisis se hace más complicado y generalmente se busca recurrir a herramientas que permitan realizar esta tarea de manera automática. En el presente trabajo se lleva a cabo la categorización automática de opiniones de texto. Estas opiniones corresponden a cuatro productos: libros, dvds, cocinas y electrónicos. Se tienen tanto opiniones positivas como negativas. Se presentan resultados de categorización usando dominios cruzados como entrenamiento y prueba, utilizando diferentes métodos de aprendizaje y se complementan con graficas de similitud, las cuales nos permiten tener una referencia visual de la proximidad del lenguaje entre los distintos dominios bajo estudio. Los resultados obtenidos permiten ver la viabilidad de la metodología propuesta.

Palabras clave. Clasificación de opiniones, aprendizaje automático, subjetivo.

Automatic Classification of Cross-Domain Opinions

Abstract. It is increasingly common for internet users to have access to blogs and social networks. On these sites, it is common to issue opinions. Opinions allow people to measure the perception of a particular topic or product. When the number of opinions is very large, its analysis becomes more complicated and it is generally sought to resort to tools that allow this task to be carried out automatically. In the present work, the automatic categorization of text opinions is carried out. These opinions correspond to four products: books, DVDs, kitchens and electronics. You have both positive and negative opinions. Categorization results are presented using cross domains as training and testing, using different learning methods and are complemented with similarity graphs, which allow us to have a visual reference of the proximity of the language between the different domains under study. The results obtained allow us to see the feasibility of the proposed methodology.

Keyword. Opinion classification, machine learning, subjective.

1. Introducción

El internet tiene un impacto profundo en el mundo laboral, en el ocio y en la búsqueda de conocimiento a nivel mundial. Uno de los servicios que más éxito ha tenido en internet ha sido la World Wide Web (mejor conocida como www ó la web). Gracias a la web, millones de personas tienen acceso fácil e inmediato a una cantidad extensa y diversa de información en línea.

Comparado con enciclopedias y bibliotecas tradicionales, la web ha permitido una descentralización repentina y extrema de la información y los datos. Algunas compañías e individuos han adoptado el uso de los blogs disponibles en línea con la finalidad de conocer y compartir las opiniones sobre sus empleos, jefes, compañeros, familia, algún tópico, productos determinados, etc., creando una gran cantidad de información y de datos [1]. Estas cantidades masivas de opiniones emitidas a través de la web, permiten medir la percepción de los usuarios por lo que su análisis se vuelve complicado.

Por esta razón, se busca recurrir a herramientas que permitan realizar esta tarea de manera automática [2].

En el contexto de aprendizaje automático, entendemos por categorización uno de los dos escenarios siguientes:

- A partir de una serie de observaciones, categorizar consiste en establecer la existencia de clases ó grupos de datos (aprendizaje no supervisado).
- Sabiendo la existencia de ciertas clases, categorizar consiste en establecer una regla para ubicar nuevas observaciones en alguna de las clases existentes (aprendizaje supervisado).

La categorización de documentos puede ser vista como la tarea de asignar un valor de 0 o 1 en cada elemento de una matriz de decisión. Donde los documentos a categorizar están representados por el conjunto $D = \{d_1, .., d_m\}$, mientras que el conjunto de posibles categorías a asignar al conjunto de documentos está representada por $C = \{c_1, \dots, c_m\}$.

De esta manera, un valor $a_{ij} = 1$ sería interpretado como que el elemento d_j pertenece al elemento c_i. En la tabla 1 se muestra el esquema utilizado para la categorización de documentos. La columna de la izquierda contiene las categorías, previamente definidas, y la fila superior los documentos a categorizar [3].

El llevar a cabo el procesamiento automático de opiniones es algo que potencialmente puede ser muy útil para sustentar la toma de decisiones en una empresa. Al tener una retroalimentación directa por parte de los usuarios del producto o servicio que proporciona dicha empresa. También pueden ser utilizadas para ver la percepción mediática de un determinado personaje de la vida pública, por ejemplo, la percepción que los ciudadanos tienen de los gobernantes o lo que opinan de un determinado actor. El enfoque dominante en la comunidad investigadora para abordar el problema de la clasificación automática de documentos se basa en la aplicación de técnicas de aprendizaje máquina [4-7].

El presente trabajo se centra en el análisis de opiniones sobre cuatro productos:

Libros, películas en general (DVDs), electrodomésticos y aplicaciones de cocinas. Todas estas opiniones fueron extraídas de un blog. Se tienen tanto opiniones positivas como negativas. La finalidad es construir un sistema de aprendizaje automático que permita clasificar una nueva opinión (no vista antes por el conjunto de entrenamiento) como positiva o negativa de una manera rápida y eficiente. Pero, además, se realizaron experimentos multidominio, esto es, se entrena en un dominio y se prueba en otro distinto, la idea intuitiva que se pretende probar en el presente trabajo es que comúnmente utilizamos palabras o frases similares para emitir opiniones, tanto positivas como negativas, independientemente del dominio del que se trate.

2. Corpus

El corpus utilizado en el presente trabajo fue creado dentro del departamento de computación y ciencias informáticas de la universidad de Pennsylvania[1]. Contiene 8000 documentos en formato electrónico correspondientes a opiniones subjetivas de diferentes productos del sitio web Amazon. En la tabla 2 se muestra el número de documentos por categorías y los tipos de dominios. Cada uno de estos productos de Amazon es tratado como un dominio distinto. Este corpus se encuentra disponible[2].

3. Algoritmos de clasificación utilizados

Para el desarrollo del presente trabajo se utilizaron tres métodos de clasificación, los cuales se describen a continuación.

3.1. Naive Bayes

El clasificador bayesiano se considera como parte de los clasificadores probabilísticos, los cuales se basan en la suposición de 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.

En tareas como la clasificación de textos este algoritmo se encuentra entre los más utilizados. El algoritmo de Naive Bayes usa el conjunto de

[1] http://www.upenn.edu/

[2] http://www.cs.jhu.edu/~mdredze/datasets/ sentiment/

Tabla 1. Categorización de documentos

		Documentos a categorizar						
		d_1	⋯	d_j	⋯	⋯	⋯	d_n
Categorías predefinidas	c_1	a_{11}	⋯	⋯	⋯	⋯	⋯	a_{1n}
	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	c_i	a_{i1}	⋯	a_{ij}	⋯	⋯	⋯	a_{in}
	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	c_n	a_{m1}	⋯	a_{mj}	⋯	⋯	⋯	a_{mn}

Tabla 2. Organización del corpus utilizado

Dominios	**Nº de documentos por Categorías**	
	Opiniones Negativas	**Opiniones Positivas**
Books	1000	1000
DVDs	1000	1000
Electronics	1000	1000
Kitchen	1000	1000
Total Documentos	4000	4000
	8000	

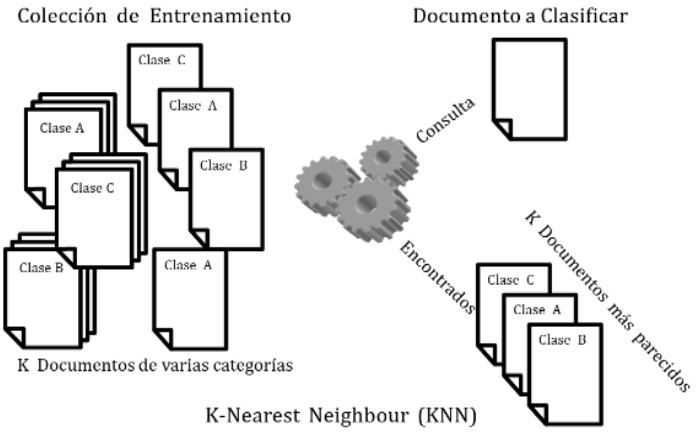

Fig.1 Algoritmo del vecino más cercano

entrenamiento para estimar los parámetros de una distribución de probabilidad que describa el conjunto de entrenamiento. Al documento con la probabilidad más alta le es asignada la categoría.

En este esquema el clasificador es construido estimando la probabilidad de cada clase, la cual es representada por T_r.

Entonces, cuando una nueva instancia i_j es presentada, el clasificador le asigna la categoría $c \in C$ más probable, después de aplicar la regla $c = argmax_{c_i \in C} P(c_i|i_j)$, y utilizando el teorema de Bayes para estimar la probabilidad tenemos :

$$c = argmax_{c_i \in C} \frac{P(i_j|c_i)P(c_i)}{P(i_j)}. \quad (1)$$

Considerando que el denominador de esta ecuación no cambia entre categorías, tenemos:

Tabla 3. Resultados de clasificación en dominios cruzados

M	eTr	C1				Fc	C2				Fc	C3				Fc	TC
		B	D	E	K		B	D	E	K		B	D	E	K		
1	B	72	69.5	65.75	70.5	1779	72.5	69.5	68	72	720	73.5	70	67.75	72.5	603	N-B
		72.75	71	68.5	75.25	1779	73.25	73.75	69.5	76	720	72.75	71.25	67	76	603	SVM
		64.75	66.5	65.75	65	1779	67	65.5	64.5	62.25	720	66.25	66.75	66.75	71	603	KNN
	D	72	79.5	74.5	75.5	1641	72.25	78.25	72.75	76	651	69.25	78.25	71.5	76.75	421	N-B
		70.25	75.25	69.25	72	1641	69.5	77.25	70.5	69.5	651	69	75	69	71.25	421	SVM
		61.5	69.75	62.25	65.75	1641	64.25	70	65.25	70.5	651	63.25	68.75	61.5	67.5	421	KNN
	E	62.5	62.75	78.25	75.5	1142	65.5	63.5	77.75	78.25	496	64.75	64.75	77	78	421	N-B
		59.25	65	82.5	79.5	1142	60.75	67.75	81.75	79.75	496	65	66.25	80.75	80.25	421	SVM
		57.25	63.75	72.75	75.5	1142	60	65.25	73	75.75	496	59.5	63.5	77.5	77.5	421	KNN
	K	66.25	69.75	78.25	80.75	950	68	70.25	78.25	82	411	67	71.25	78.25	81	336	N-B
		65	69	77	84.25	950	67.5	65.75	76.75	82.5	411	65	67.5	76.75	81	336	SVM
		65	66	76	76	950	64	66	77.5	77.75	411	61.25	67.5	74	75.25	336	KNN

$$c = argmax_{c_i \in C} P(i_j|c_i) P(c_i). \quad (2)$$

Tomando en cuenta que el esquema es llamado “naive” debido al supuesto de independencia entre atributos; esto es, se asume que las características son condicionalmente independientes dadas las clases:

$$c = argmax_{c_i \in C} P(c_i) \prod_{k=1}^{n} P(a_{kj}|c_i), \quad (3)$$

donde $P(c_i)$ es la fracción de ejemplos en T_r que pertenecen a la clase c_i, y $P(a_{kj}|c_i)$ se calcula usando el teorema de Bayes [8].

3.2. Vecinos más cercanos (IBK)

El algoritmo se basa en localizar el documento más parecido al que se desea clasificar.

Para esto, se debe utilizar ese documento como si fuera una consulta sobre la colección de entrenamiento. Una vez localizado el documento de entrenamiento más similar, se asigna la categoría.

Una de las variantes más conocidas de este algoritmo es la del *k-nearest neighbour* (KNN) que consiste en tomar los k documentos más parecidos, en lugar de sólo el primero.

Como en esos k documentos los habrá, presumiblemente, de varias categorías, se suman los coeficientes de los de cada una de ellas asignando la categoría a la que tenga mayor puntaje, ver figura 1.

KNN es espacialmente eficaz cuando el número de categorías posibles es alto, y cuando los documentos son heterogéneos y difusos [8].

3.3. Máquinas de vectores de soporte (SVM)

SVM ofrece la ventaja de ser utilizadas para resolver tanto problemas lineales como no lineales, buscando entre todos los híper planos separadores, aquel que maximice la distancia de separación entre dos o más conjuntos [9]. La manera más simple de realizar la separación es mediante una línea recta, un plano recto o un hiperplano N-dimensional, pero los universos a clasificar no se suelen presentar en el ideal de las dos dimensiones como ocurre en el ejemplo gráfico, sino que un algoritmo SVM debe tratar con más de dos variables predictoras, curvas no lineales de separación, casos donde los conjuntos de datos no pueden ser completamente separados, este problema se incrementa cuando el número de categorías que se pretende clasificar crece.

4. Resultados

En la Tabla 3 se muestran los resultados obtenidos en el desarrollo del presente trabajo. Se implementaron tres clasificadores, distintos con las siguientes características: El clasificador 1 (C1) representa la línea base.

En este caso los documentos son pre-procesados únicamente eliminando las etiquetas HTML y signos de puntuación. Para los clasificadores 2 y 3 (C2 y C3), además, se eliminan las palabras auxiliares (*stop words*) utilizando la lista corta y lista larga respectivamente, con la finalidad de medir el impacto de las palabras de paro en la clasificación de opiniones.

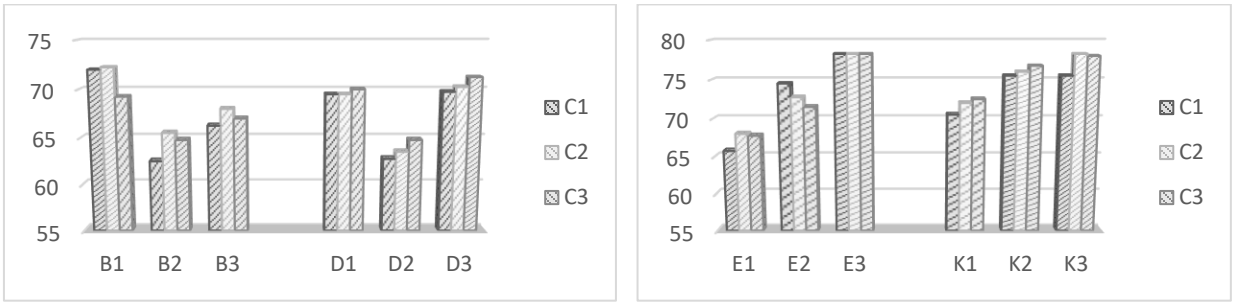

Fig. 2. Resultados usando Naive Bayes

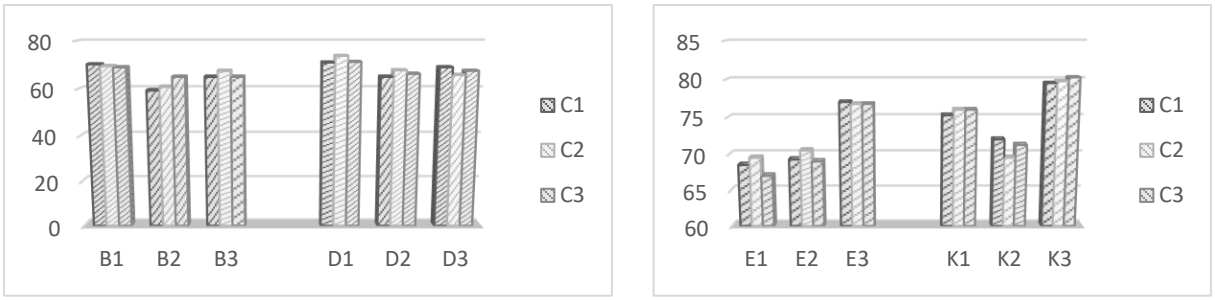

Fig. 3. Resultados usando SVM

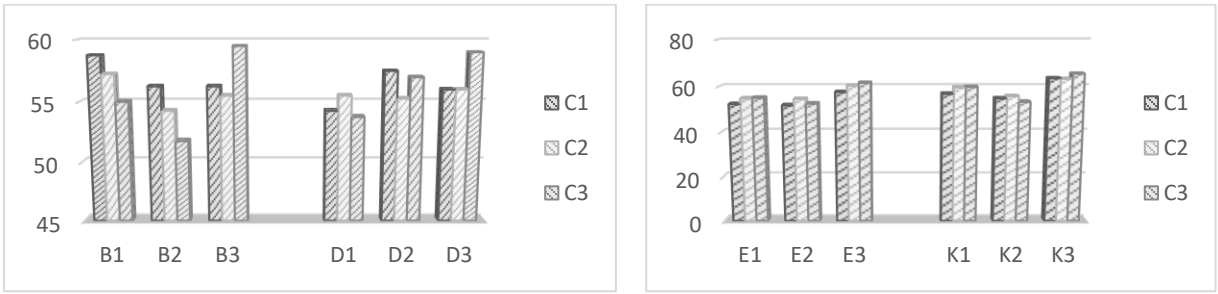

Fig. 4. Resultados usando SVM

En cada caso se utilizaron tres métodos de clasificación descritos en párrafos anteriores: Naive Bayes (N-B), Vecinos más cercanos (KNN) y SVM.

Los símbolos B (Libros), DVDs (D), electrodomésticos (E) y cocinas (K) que están al inicio de las filas (izquierda) representan los dominios de entrenamiento (Tr) y los mismos símbolos que están en la parte superior de las columnas representan los dominios de prueba (Te). Así podemos apreciar en la tabla los resultados tanto de entrenar y probar en el mismo domino, así como entrenar con un dominio y probar en otro.

Para cada clasificador, en la tabla 3, se muestra la frecuencia de corte utilizada en cada dominio en específico. Con la finalidad de medir el desempeño de cada método de clasificación en los diferentes escenarios de prueba.

Se puede apreciar que es posible realizar la clasificación en dominios cruzados (entrenar con un dominio y probar en otro) por ejemplo, cuando se utiliza el conjunto de electrónicos como entrenamiento y el conjunto de cocinas como

Tabla 4. Medidas de evaluación

Grupo	Subgrupo	TC	Matriz de confusión	Medidas de Evaluación			Categoría
				Precisión	*Recall*	*Fallout*	
B	B1	N-B	a b $<--$ $classified$ 146 54 | $a = Negativas$ 57 143 | $b = Positivas$	0.719	0.73	0.285	Negativa
				0.726	0.715	0.27	Positiva
E	E3		a b $<--$ $classified$ 183 17 | $a = Negativas$ 70 130 | $b = Positivas$	0.723	0.915	0.35	Negativa
				0.884	0.65	0.085	Positiva
B	B1	SVM	a b $<--$ $classified$ 149 51 | $a = Negativas$ 71 129 | $b = Positivas$	0.677	0.745	0.355	Negativa
				0.717	0.645	0.255	Positiva
K	K3		a b $<--$ $classified$ 163 37 | $a = Negativas$ 42 158 | $b = Positivas$	0.795	0.815	0.21	Negativa
				0.81	0.79	0.185	Positiva
D	D1	J48	a b $<--$ $classified$ 141 59 | $a = Negativas$ 74 126 | $b = Positivas$	0.656	0.705	0.37	Negativa
				0.681	0.63	0.295	Positiva
K	K3		a b $<--$ $classified$ 148 52 | $a = Negativas$ 38 162 | $b = Positivas$	0.796	0.74	0.19	Negativa
				0.757	0.81	0.26	Positiva

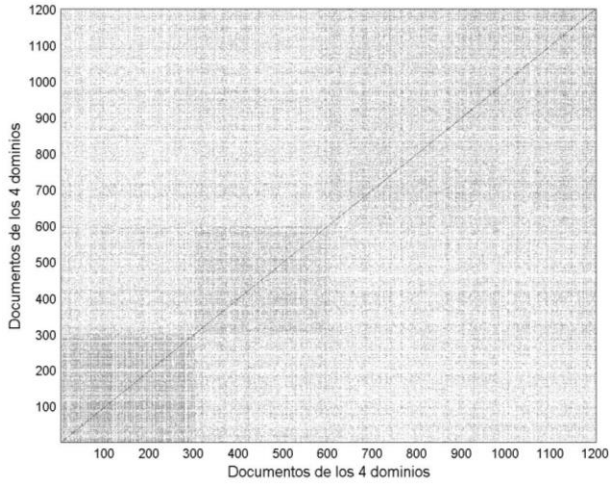

Fig. 5. Gráfica de similitud

prueba se logra tener una mejora en la línea base utilizando los tres clasificadores.

A continuación, se muestran algunas graficas que ilustran tanto los conjuntos de entrenamiento/prueba, así como método de clasificación utilizado y el escenario de clasificación, identificado por C1, C2 Y C3.

La Figura 2 muestra los resultados obtenidos de la categorización de todos los pares de dominios usando Naive Bayes.

Como se puede apreciar el mejor desempeño se tiene en los conjuntos de electrónicos y cocinas utilizando C2. También se presenta una pequeña mejora en dvds, utilizando C3 del orden de 1.25%.

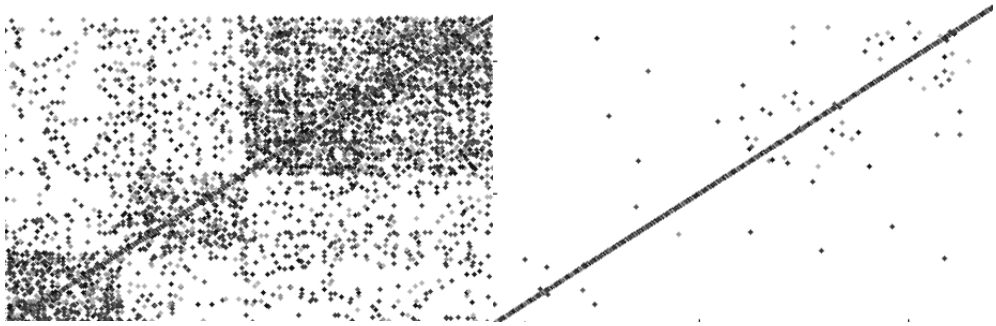

Fig. 6. Gráfica de similitud con umbral de 0.305556 (Superior) y 0.5056 (Inferior)

La figura 3 muestra los resultados obtenidos de la categorización de todos los pares de dominios usando SVM. En este caso el mejor desempeño lo tiene electrónicos y cocinas.

En la figura 4 se muestra los resultados obtenidos de la categorización de todos los pares de dominios utilizando vecinos más cercanos. En este acaso se aprecia una pequeña mejora en libros, utilizando C3, mientras que para DVD la mejora es del orden del 4% en la exactitud utilizando C3.

El que tiene un mejor desempeño con una mejora del 6% es electrónicos con los clasificadores C2 y C3. En el caso de cocinas la mejora es del 5% utilizando C3.

En la tabla 4 se muestran las medidas de evaluación, así como la matriz de confusión para los mejores resultados obtenidos en cada técnica de categorización. Excepto para cocinas se clasifican mejor las opiniones positivas que las negativas. La exactitud más alta se obtiene para el dominio de electrónicos que coincide con el *Fallout* más bajo utilizando Naive Bayes.

El *Recall* más alto se obtiene para electrónicos (opiniones negativas) mientras que el más bajo se presenta para DVDs (positivas) usando KNN. Las gráficas de similitud son una representación gráfica que permite observar el traslape del vocabulario de los documentos pertenecientes a un corpus. En la gráfica de similitud mostrada en la figura 5 se compara un dominio con los demás (incluido el mismo) para cada dominio se utilizaron 300 archivos (total 1,200) la diagonal representa la comparación de un archivo con el mismo.

En la figura 6 se muestra la misma gráfica, pero definiendo un umbral, esta acción permite identificar al vocabulario “propio” del dominio.

Los puntos conglomerados en la esquina superior derecha describen corresponden a los conjuntos de electrónicos y cocinas fueron los que dominaron en la categorización, y en algunos casos también el dominio de Libros, como se aprecia en la esquina inferior izquierda.

Por otra parte, con un umbral de ‘0.5056’ se limpia, lo cual nos indica que son pocas las palabras distintas utilizadas entre los diferentes dominios.

5. Conclusiones

Hay muchos problemas en los que no se cuenta con suficientes instancias de entrenamiento.

La información presentada en este trabajo puede ser útil en este escenario. Cuando se requiera entrenar en un dominio y probar en otro. En el caso de opiniones se cuenta con información de tipo subjetiva. Los resultados obtenidos permiten apreciar la similitud de lenguaje entre dominios tanto para opiniones positivas como negativas.

De los resultados obtenidos se desprende que las palabras de paro no ayudan a clasificar información subjetiva en dominios cruzados. Al realizar la eliminación de las palabras de paro se observa una mejora significativa principalmente utilizando un preprocesamiento en el cual se eliminaron los signos de puntuación, etiquetas HTML y se realiza la eliminación de las palabras de paro utilizando la lista larga en idioma inglés.

El clasificador que muestra un mejor desempeño para la clasificación de opiniones es Naive Bayes. También se observa de los

resultados obtenidos que es posible realizar la clasificación de opiniones en dominios cruzados.

En este caso se obtienen mejores resultados cuando existe una cercanía en el lenguaje entre los dominios, por ejemplo, para el caso de electrónicos y cocinas.

Referencias

1. **Ayasamy, R.K. (2013).** Organizing Information in the Blogosphere: The Use of Unsupervised Approach. *International Journal of Soft Computing and Engineering (IJSCE),* Vol. 3, No. 5, pp. 194**–**198.
2. **Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015).** Antisocial Behavior in Online Discussion Communities. *AAAI International Conference on Weblogs and Social Media*. pp. 61–70.
3. **Aphinyanaphongs, Y., Fu, L.D., Li, Z., Peskin, E. R., Efstathiadis, E., Aliferis, C.F., & Statnikov, A. (2014).** A Comprehensive Empirical Comparison of Modern Supervised Classification and Feature Selection Methods for Text Categorization. *Journal of the Association for Information Science and Technology,* Vol. 65, No. 10, pp. 1964–1987. DOI: 10.1002/asi.23110.
4. **Behera, R.N., Manan, R., & Dash, S. (2016).** Ensemble based hybrid machine learning approach for sentiment classification – a review. *International Journal of Computer Applications,* Vol. 146, No. 6. pp. 31–36.
5. **Carbonell, J.G., Michalski, R.S., & Mitchell, T.M. (1983).** An overview of machine learning. *Machine learning,* Springer Berlin Heidelberg, Vol. 1, pp. 3–23. DOI: 10.1016/B978-0-08-051054-5.50005-4.
6. **Sebastiani, F. (2002).** Machine learning in automated text categorization. *ACM Computing Surveys (CSUR),* Vol. 34, No. 1, pp. 1–47. DOI: 10.1145/505282.505283.
7. **Guzmán-Cabrera, R., Rosso, P., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2012).** *Clasificación automática de textos*.
8. **Hall, A.M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009).** The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter,* Vol. 11, No. 1, pp. 10–18. DOI: 10.1145/1656274.1656278.
9. **Vapnik, V.N. (1998).** *Statistical Learning Theory*. John Wiley & Sons, Inc.
10. **Kompan, M. & Bielikova, M. (2011).** News Article Classification Based on a Vector Representation Including Words' Collocations. *Advances in Intelligent and Soft Computing.* Springer, Vol. 101, pp. 1–8. DOI: 10.1007/978-3-642-23163-6_1.

Article received on 07/09/2019; accepted on 14/11/2019.
Corresponding author is Rafael Guzman-Cabrera.