ISSN 1870-4069

Un algoritmo para detectar la polaridad de opiniones en los dominios de laptops y restaurantes

Karen L. Vazquez, Mireya Tovar, Darnes Vilariño, Beatriz Beltrán

Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Puebla, México

krnlet@gmail.com, {mtovar, darnes, bbeltran}@cs.buap.mx

Resumen. En este artículo se presenta un estudio para la detección de Polaridad en un conjunto de opiniones de usuarios emitidos hacia Restaurantes en el idioma español e inglés, asi como para opiniones escritas a valoraciones de Laptops en el idioma inglés. Como solución propuesta se emplea Máquina de Soporte Vectorial con validación cruzada. Los mejores resultados se presentaron para el Dominio de Restaurantes en español con un 74% de precisión.

Palabras clave: Minería de opiniones, aprendizaje supervisado, procesamiento de lenguaje natural.

An Algorithm to Detect the Polarity of Opinions in the Domains of Laptops and Restaurants

Abstract. In this article, we present a study for the detection of polarity over a dataset of opinions of users over Restaurants in the Spanish and English language, as well as, opinions about Laptops in the English language. As a proposed solution, we use Vector Support Machine with cross validation. The best results were presented for the Spanish Restaurant Domain with 74% accuracy.

Keywords: Opinion mining, supervised learning, natural language processing.

1. Introducción

El fácil acceso que tenemos a Internet y a su ves, a las grandes cantidades de información que se producen en la Web, la Inteligencia Artificial y más específicamente, el Procesamiento de Lenguaje Natural (PLN) proporcionan mecanismos de extracción de información. La información que se encuentra en Internet se presenta en la mayoría de los casos de manera no estructurada, un ejemplo de ello son las redes sociales, fuente de acceso a opiniones, productos o servicios que la sociedad genera a diario en estos sitios, esta información puede ser una fuente para la aplicación del PLN, que se encarga de la detección automática de los sentimientos

expresados en los textos y su clasificación según la polaridad que tienen, es el área de análisis de sentimientos, también llamada minería de opiniones.

En este documento se muestra la propuesta de solución para el análisis de sentimientos en opiniones expuestas por usuarios a través de las redes sociales, se revisan dos dominios de opiniones, Restaurantes y Laptops, el primero de ellos en el idioma español e inglés y el último únicamente en el idioma inglés. El objetivo es detectar la polaridad de cada opinión, es decir, clasificar la polaridad de la reseña dada por el usuario, siendo esta positiva, negativa, neutra o conflicto. Para el logro de este objetivo, se propone el uso de un clasificador, el seleccionado en esta investigación es Máquina de Soporte Vectorial utilizando validación cruzada para la prueba de datos, los resultados obtenidos muestran un desempeño adecuado con el algoritmo propuesto.

La distribución de este artículo se describe a continuación, en la Sección 2 se presentan los trabajos relacionados con esta investigación, en la Sección 3 se presenta el algoritmo o modelo propuesto, en la Sección 4.1 se presenta la información de los datos utilizados, y los resultados logrados en la Sección 4.2. Finalmente, las conclusiones se presentan en la Sección 5.

2. Trabajo relacionados

El monitoreo de opiniones ha sido estudiado desde hace algunos años a través del Procesamiento del Lenguaje Natural, sin embargo, los estudios e investigaciones realizadas se han logrado en la mayoría de los casos en idiomas diferentes al español, los sistemas para monitoreo de opiniones realizados en esos idiomas y principalmente en el idioma inglés, han tenido resultados favorables. A continuación se describen algunos trabajos relacionados con esta investigación.

En [1] se describe el sistema de minería de opinión llamado “sentiue”, que pretende determinar la polaridad del sentimiento expresado sobre algún aspecto concreto de una entidad. Como sistema de clasificación se utiliza MALLET, y como característica utiliza las palabras textuales y los lemas. Este sistema participó en la tarea 12 de SemEval-2015 obteniendo un 79% de precisión en el dominio de laptops, para la determinación de la polaridad de sentimientos de un texto dado.

En el trabajo de investigación desarrollado por [2] se presenta una contribución a la tarea 5 de SemEval 2016, que se enfocaron en los lenguajes de inglés y francés para el dominio de restaurantes. Su sistema se basa en modelos compuestos, que mezcla rasgos lingüísticos con algoritmos de aprendizaje automático. De acuerdo con los resultados observados, los autores obtuvieron un 88% de precisión para el dominio de restaurantes (idioma inglés), y un 78% de precisión para la determinación de la polaridad en el dominio de restaurantes (lengua francesa).

En [3], los autores describen el sistema que utilizan en la tarea 5 de SemEval 2016. Su sistema se basa en aprendizaje automático supervisado, utilizando un clasificador de máxima entropía, campo aleatorio condicional, y un gran número de características tales como vectores globales, la asignación de Dirichlet latente, bolsa de palabras, iconos gestuales y otros. Se obtuvieron resultados muy competitivos en la tarea 5 de SemEval-2016.

En el trabajo de [4], se propone un esquema de ponderación supervisada en base a dos factores: la importancia de un término en un documento (ITD) y la importancia de

un término para expresar el sentimiento (ITS). Los resultados experimentales muestran que el método produce la mejor precisión en dos de tres conjuntos de datos.

En el estudio [5] se propuso la minería de opiniones basada en Ontologías, enfocado a textos de opiniones provenientes de usuarios con escritura en idioma inglés, proporcionando un nuevo método para el análisis de los sentimientos basados en el análisis vectorial.

Para los estudios realizados en [6], se propuso un método automático para clasificar la polaridad de las opiniones en consumidores de productos de una empresa. El algoritmo se basa en el uso de ontologías para encontrar todas aquellas opiniones realizadas a sus productos incluyendo dentro del análisis de opiniones una fusión de ontologías. Los autores concluyen que conforme aumenta el número de términos de cada ontología, no se nota una mejora en los métodos utilizando ontologías de alto nivel respecto a los métodos directos.

En el trabajo desarrollado en [7] se presentan dos modelos para descubrir la polaridad de mensajes en redes sociales, en particular extraídos del Twitter.

El primer modelo extrae las características léxico-sintácticas de cada tweet. El segundo modelo obtiene las características de cada tweet basándose en la centralidad de grafos.

En el trabajo de [8] clasifican los sentimientos en tweets, de acuerdo a su contenido en positivo, negativo o neutro; donde proponen mejoras a las dependencias objetivo de la clasificación de sentimientos en twitter con la incorporación de características dependientes de objetivos mediante tres pasos, en el primero; se hace una clasificación subjetiva para decidir si hay subjetividad en el tweet o es neutro, en el segundo paso, para los elementos subjetivos se le da su polaridad (positivo o negativo) y finalmente se realiza una optimización basada en grafos usando tweets que están relacionados. Para los dos primeros pasos hacen uso de máquinas de soporte vectorial (SVM-Light). Con las características dependientes de objetivos obtienen una precisión del 85.6%.

Los autores de [9] crean un corpus para los experimentos, añadieron juicios de polaridad contextuales a las anotaciones existentes en el Corpus de Opinión multi-perspectiva Pregunta-Respuesta (MPQA, por sus siglas en inglés), mediante un esquema de anotación. Para los experimentos usaron un lexicón con pistas subjetivas, que fueron agrupadas de acuerdo a su confianza y subjetividad. Posteriormente se expandió el lexicón mediante un diccionario y un tesauro. El algoritmo les permite identificar la polaridad contextual automáticamente para un amplio conjunto de expresiones de sentimientos. El mejor resultado obtenido considera 28 características y logran una precisión de 75.9%.

En Bakliwal et al. [10], construyen un conjunto de datos de tweets políticos. Cada tweet de este conjunto esta anotado como positivo, negativo o neutro. Se incluyeron tweets de sarcasmo. En sus experimentos omitieron el sarcasmo y lograron una precisión de casi el 59% con operaciones simples de búsqueda léxica.

En [11] presentan un enfoque híbrido para determinar el sentimiento de cada tweet. Realizan un pre-procesamiento analizando abreviaturas, lematización, eliminación de palabras cerradas, etc. Probaron seis conjuntos de datos de Twitter, y logran un 83.3% de medida F_1, una precisión de 85.7% y un recall de 82.2%.

En la siguiente sección se describe el algoritmo utilizado para el análisis de sentimientos.

3. Algoritmo propuesto

A continuación se describe el algoritmo propuesto para la determinación de la polaridad de las opiniones proporcionadas en la Tarea 5 de SemEval 2016 [12].

En el algoritmo se inicia con un pre-procesado de datos, en esta fase se realiza una limpieza de la información como es la eliminación de signos de puntuación, eliminación de palabras cerradas, etc. En la segunda fase se realiza la extracción de características utilizando la biblioteca *sci-kit-learn* que es una herramienta de aprendizaje automático en Python [13], [14][1]. En la tercera fase se utiliza un sistema de clasificación utilizando Máquinas de Soporte Vectorial proporcionado por la herramienta de aprendizaje automático sci-kit-learn. En la Fig. 1 se muestra gráficamente, el algoritmo propuesto.

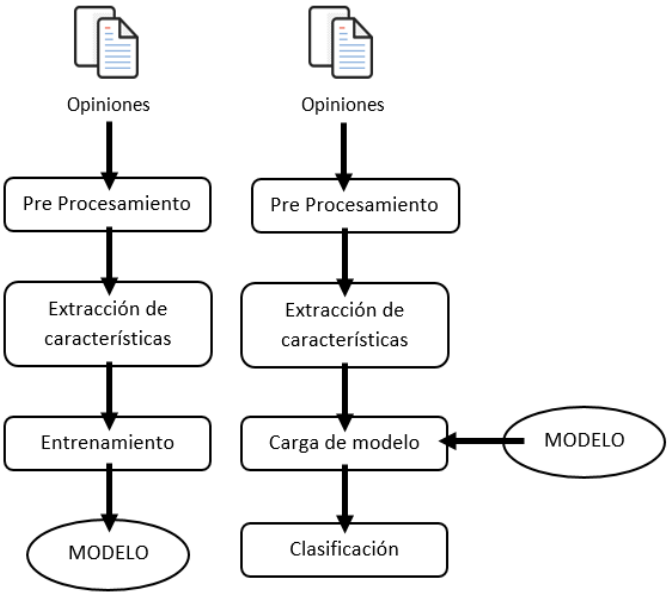

Fig. 1. Algoritmo propuesto.

- Pre procesamiento:
  - Extracción de opiniones del documento en formato XML: Filtrar únicamente opiniones del documento
  - Limpieza de opiniones: Eliminar palabras vacías o cerradas, signos de puntuación, acentos y caracteres aislados.
  - Tokenización: Tokenizar opiniones por palabra.
  - *Stemming*: Proceso heurístico que corta el final de las palabras y con frecuencia incluye la eliminación de los afijos derivativos: coches = coche, fuimos = fui.

[1] http://scikit-learn.org/

- Filtrar Categorías: Filtrar opiniones de los datos de entrenamientos por polaridad como: POSITIVO, NEGATIVO, NEUTRAL y CONFLICTO.

- Extracción de características:
  - Frecuencia y Frecuencia Inversa del Documento para un Término (*TF-IDF*): Este método indica qué tan relevante es la palabra con respecto al documento seleccionado y al corpus en general. Permite también calificar a los documentos del corpus con base en estas palabras clave, es decir, si las palabras tiene mayor peso, entonces el documento está más relacionado con ellas que uno con las mismas palabras pero con menor peso. La herramienta *sci-kit-learn* de Python proporciona varios *vectorizers*[2] de traducción de los documentos de entrada en vectores de características: *TfidfVectorizer*[3] es un vectorizer que usa TF-IDF como esquema de ponderación.
- Sistema de clasificación:
  - En esta se utiliza el método de Máquina de Soporte Vectorial (SVM) que es una técnica de aprendizaje automático. En este caso se utiliza SVC que se basa en libsvm.

4. Resultados

En este apartado se describen los datos utilizados para el análisis de opiniones y los resultados correspondientes a cada dominio, con el algoritmo propuesto.

4.1. Conjunto de datos

Los datos utilizados son los datos de entrenamiento proporcionados por SemEval-2016 para la subtarea 2 de la tarea 5. Los dominios considerados son el de Restaurantes para el idioma español e inglés, y el dominio de Laptops para el idioma inglés. En la Tabla 1 se muestra el total de opiniones por dominio, idioma y polaridad.

Tabla 1. Total de opiniones por dominio, idioma y polaridad.

Dominio	Positivo	Negativo	Neutral	Conflicto
Restaurantes (español)	1,519	443	101	58
Restaurantes (inglés)	1,012	327	55	41
Laptops (inglés)	1,210	707	123	41

4.2. Resultados experimentales

Siguiendo el algoritmo, presentado en la sección 3, se realiza una evaluación de la polaridad detectada con el sistema de clasificación utilizando como característica TF-IDF. En las Tablas 2, 3 y 4 se presentan el total de opiniones evaluadas por Polaridad,

[2] Función utilizada en scikit-learn http://scikit-learn.org/.
[3] Convierte una colección de documentos sin formato a una matriz de características TF-IDF.

el resultado de *Precisión, Recall* y F_1 por dominio, asi como los datos de prueba por polaridad. Cabe mencionar que la realización de la prueba de datos se hizo con validación cruzada[4], en la cual se toma cierto porcentaje para entrenar los datos y el porcentaje restante para la prueba. En este caso se considera el 16% como prueba de datos en todos los dominios y un promedio de 10 repeticiones. Es por eso que en detalle, en la Tabla 2 se muestran los resultados del dominio de Laptops en inglés y en la última columna se desglosa la división de opiniones evaluadas por polaridad, asi mismo se presenta que el total de opiniones evaluadas fueron 304, lo mismo para la Tabla 2 pero en el caso de Restaurantes en idioma inglés, presentando un total de prueba de 222 opiniones evaluadas y por último en la Tabla 3, donde se observa que el total de pruebas para Restaurantes en idioma español fueron 343 opiniones.

Tabla 2. Resultados del algoritmo para el dominio de Laptops del idioma inglés.

Polaridad	Precisión	Recuerdo	F_1	Total de prueba
Conflicto	0.00	0.00	0.00	11
Negativo	0.64	0.63	0.63	62
Neutral	0.00	0.00	0.00	25
Positivo	0.78	0.92	0.84	206
Total	0.66	0.75	0.70	304

Tabla 3. Resultados del dominio de Restaurantes del idioma inglés.

Polaridad	Precisión	Recuerdo	F_1	Total de prueba
Conflicto	0.00	0.00	0.00	5
Negativo	0.67	0.33	0.44	43
Neutral	0.00	0.00	0.00	9
Positivo	0.79	0.96	0.87	165
Total	0.72	0.78	0.73	222

Tabla 4. Resultados del dominio de Restaurantes del idioma en español.

Polaridad	Precisión	Recuerdo	F_1	Total de prueba
Conflicto	0.00	0.00	0.00	14
Negativo	0.67	0.47	0.55	60
Neutral	0.00	0.00	0.00	11
Positivo	0.82	0.96	0.89	258
Total	0.74	0.80	0.76	343

5. Conclusiones

En este artículo se presenta un algoritmo de clasificación automática que permite identificar la polaridad de las opiniones proporcionadas en la tarea 5 de SemEval 2016. Los dominios considerados en las pruebas son de Restaurantes en español e inglés, y el dominio de Laptops para el idioma inglés. En base a los resultados obtenidos se observa que el método de SVM logra obtener resultados satisfactorios al obtener más del 66% de precisión en los tres dominios con la característica de TF-IDF. El algoritmo

[4] Valoración optimista que emplea los propios datos de entrenamiento para evaluar el modelo.

propuesto, no utiliza información adicional para enriquecer las opiniones proporcionadas por Semeval-2016, motivo por el cual el algoritmo no logra clasificar adecuadamente las opiniones de conflicto y neutral. Como trabajo a futuro se tiene considerado el enriquecimiento de los datos de entrenamiento para lograr mejorar la clasificación de las opiniones que tienen poca información. Así como el uso de otros métodos de clasificación supervisada y su prueba con los datos de entrenamiento y pruebas proporcionadas por Semeval-2016.

Agradecimientos. Esta investigación es parcialmente apoyada por el proyecto PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854, por el proyecto ID 00570 VIEP-BUAP. Apoyado por el Fondo Sectorial de Investigación para la Educación, proyecto Conacyt 257357.

Referencias

1. Saias, J.: Sentiue: Target and aspect based sentiment analysis in semeval-2015 task 12. In: Proceedings of the 9th International Workshop on Semantic Evaluation, Denver, Colorado, Association for Computational Linguistics, pp. 767–771 (2015)
2. Brun, C., Perez, J., Roux, C.: Xrce at semeval-2016 task 5: Feedbacked ensemble modeling on syntactico-semantic knowledge for aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation, San Diego, California, Association for Computational Linguistics, pp. 282–286 (2016)
3. Hercig, T., Brychcín, T., Svoboda, L., Konkol, M.: Uwb at semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation, San Diego, California, Association for Computational Linguistics, pp. 354–361 (2016)
4. Deng, Z. H., Luo, K. H., Yu, H. L.: A study of supervised term weighting scheme for sentiment analysis. Expert Systems with Applications, 41, pp. 3506–3513 (2014)
5. Peñalver, I., Garcia, F., Valencia, R., Rodríguez, M. A., Moreno, V., Fraga, A., Sánchez, J. L.: Feature-based opinion mining through ontologies. Expert Systems with Applications, 41, pp. 5995–6008 (2014)
6. Balaguer, E. V., Rosso, P., Locoro, A., Mascardi, V.: Análisis de opiniones con ontologıas. Polibits, 41, pp. 29–36 (2010)
7. Sanzón, Y. M., Vilariño, D., Somodevilla, M. J., Zepeda, C., Tovar, M.: Modelos para detectar la polaridad de los mensajes en redes sociales. Research in Computing Science, 99, pp. 29–42 (2015)
8. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, Portland, Oregon, USA, pp. 151–160 (2011)
9. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase level sentiment analysis. In: HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Vancouver, British Columbia, Canada (2005)
10. Bakliwal, A., Foster, J., van der Puil, J., OBrien, R., Tounsi, L., Hughes, M.: Sentiment analysis of political tweets: Towards an accurate classifier. In: Proceedings of the Workshop on Language in Social Media, Atlanta, Georgia, Association for Computational Linguistics, pp. 49–58 (2013)
11. Khan, F. H., Bashir, S., Qamar, U.: Tom: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems, 57, pp. 245–257 (2014)

12. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez, S. M., Eryigit, G.: Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation, San Diego, California, Association for Computational Linguistics, pp. 19–30 (2016)
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, pp. 2825–2830 (2011)
14. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122 (2013)