Comparación de métodos para clasificar comentarios de lugares turísticos por medio de análisis de aentimiento

Luis Guillermo Herrera-Sarmiento

Hoy en día los turistas luego de visitar algún destino, plasman sus experiencias como opiniones en diversas fuentes digitales, siendo información valiosa para empresas turísticas o relacionadas para identificar qué sitios son una oportunidad de mejora para los turistas durante la planificación de sus viajes. En esta investigación se propone la comparación de Support Vector Machine (SVM), Naïve Bayes (NB) y método propuesto basado en SVM y *chi square* como método de selección de características. La técnica híbrida propuesta obtuvo el mejor resultado, seguido de SVM y por último Naïve Bayes, cada una con 76,50 %, 67,53 % y 66,91 % de precisión, respectivamente.

Comparison of Methods for Classifying Comments on Tourist Places by Sentiment Analysis

Nowadays, tourists express their experiences as opinions in various digital sources after visiting a destination, which is considered a valuable information for tourist companies or other related companies to identify which places are an opportunity for improvement, and for tourists when planning their trips. This research proposes the comparison of the support vector machine (SVM), naïve Bayes (NB), and a suggested method based on SVM and chi-square as a feature selection method. The proposed hybrid technique obtained the best result, followed by SVM and finally naïve Bayes, each with 76.50%, 67.53% and 66.91% accuracy, respectively.

COMPARACIÓN DE MÉTODOS PARA CLASIFICAR COMENTARIOS DE LUGARES TURÍSTICOS POR MEDIO DE ANÁLISIS DE SENTIMIENTO

Luis Guillermo Herrera-Sarmiento
20141885@aloe.ulima.edu.pe

Resumen Es común que los turistas, luego de visitar algún destino, plasmen sus experiencias y opiniones en diversas fuentes digitales, convirtiendo esa información en una herramienta valiosa para empresas turísticas o relacionadas porque así podrían identificar qué sitios son una oportunidad de mejora para los turistas durante la planificación de sus viajes. En esta investigación se propone la comparación de Support Vector Machine (SVM), Naïve Bayes (NB) y el método propuesto basado en SVM y *chi square* como método de selección de características. La técnica híbrida propuesta obtuvo el mejor resultado, seguido de SVM y, por último, Naïve Bayes, cada una con 76,50 %, 67,53 % y 66,91 % de precisión, respectivamente.

Introducción

Se ha desarrollado una tendencia que va en aumento sobre el número de viajeros independientes o *backpackers* (Chi, Lo, Chu, y Lin, 2009). El Ministerio de Comercio Exterior y Turismo del Perú (MINCETUR) indica que el crecimiento del turismo en el país en el año 2018 fue de 9,6 % a comparación del año anterior y el Scotiabank proyecta un incremento del 10 % para el año 2019. En el Perú es una de las industrias más grandes, puesto que es una de las principales fuentes de la economía peruana. Por ello se debe tomar en consideración toda información buena o mala acerca de los puntos de interés para poder encontrar puntos de mejora al servicio ofrecido. Lin, Wu, Chen, Ku, y Chen (2014) mencionan que los viajeros que llegan a su destino no encuentran sitios turísticos agradables, lo que conlleva a que estos fomenten una mala imagen del sitio por medios digitales. El propósito de esta investigación es identificar el nivel de agrado de los turistas a partir de sus comentarios en un sitio web turístico y clasificarlo ya sea en negativo o en positivo, puesto que, como menciona Chi *et al.* (2009), un viaje para un turista es una actividad personal que busca objetivos.

Metodología

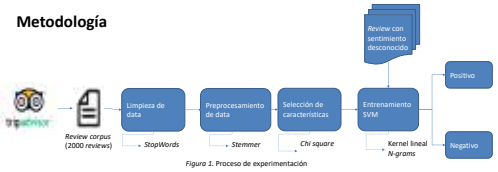

Figura 1. Proceso de experimentación

Recolección de data
Se extrajeron *reviews* de TripAdvisor (Li y Yang, 2017), dado que estos poseen variables como fecha, título, *rating* y ciudad (Parikh, Kestar, Dharia, y Gotmare, 2018) mediante *web scraping* para desarrollar el conjunto de datos a trabajar. Solo se tomaron en consideración los cinco sitios turísticos de la ciudad de Lima, Perú, más comentados y con una antigüedad no mayor a cinco meses (2000 *reviews*).

Limpieza de data
Para la limpieza de la información se reconocieron los elementos que no pudieron ser interpretados dado que en el paso anterior algunos de los elementos extraídos los contenían como se aprecia en la figura 2.

*"Visitamos el lugar en compañía de mi familia en el mes de abril, el complejo arqueológico es hermoso e interesante, muestra la arquitectura preincaica.

 El guía nos comentó acerca de la forma en que se construyó y cómo ha sobrevivido al paso del tiempo

".*

"Visitamos el lugar en compañía de mi familia en el mes de abril, el complejo arqueológico es hermoso e interesante, muestra la arquitectura preincaica. El guía nos comentó acerca de la forma en que se construyó y cómo ha sobrevivido al paso del tiempo".

Figura 2. Review elementos sin regex

Desarrollo del clasificador
Este se basa en el algoritmo propuesto que consiste en dos fases: categorización y clasificación, que a su vez está compuesto de una serie de pasos que usan la medición con TF-IDF (término frecuencia de documento de frecuencia inversa), el modelo de *n-grams* y el estadístico *chi square* como *features selection*:

- Brindar pesos a las palabras mediante TF-IDF.
- Extracción de las características más relevantes con *chi square*.
- Clasificación del sentimiento del *review* con la técnica SVM.
- Comparación de técnicas Naïve Bayes, SVM clásico y SVM propuesto .

Resultados

Los resultados se plantean en dos instancias: 1) en el análisis de la información de TripAdvisor y 2) en el de la experimentación.

Resultados de las visitas
En las siguientes imágenes se observa la proporción de *reviews* por lugar y la de sentimiento por lugar.

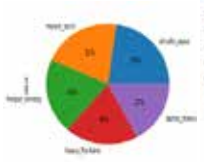

Figura 3. Distribución de reviews por lugar

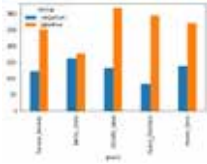

Figura 4. Distribución de sentimiento por lugar

Resultados de la experimentación
En la tabla se muestra la comparación de los niveles de *precision*, *recall*, *f1-score*, *accuracy*.

		precision	*recall*	*f1-score*	*support*
Prop	negative	0.62	0.46	0.53	114
	positive	0.8	0.89	0.84	283
	avg/total	**0.75**	**0.77**	**0.75**	**397**
SVM	negative	0.71	0.05	0.09	128
	positive	0.68	0.99	0.81	266
	avg/total	**0.69**	**0.68**	**0.58**	**394**
NB	negative	0.67	0.03	0.06	128
	positive	0.68	0.99	0.81	266
	avg/total	**0.68**	**0.68**	**0.56**	**394**

Técnica de clasificación de sentimiento	SVM	NB	Método propuesto
Accuracy	**67,53 %**	**66,91 %**	**76,57 %**

Como se observa en las tablas, la técnica híbrida propuesta obtuvo el mejor resultado: 76,50 % de precisión. Además, al momento de la predicción del sentimiento de un comentario fue correcta en un 75 % de las veces, identificó correctamente el 77 % el sentimiento y la exactitud de la clasificación fue del 75 %.

Conclusiones

En esta investigación se aplicaron tres técnicas para la clasificación del texto: SVM propuesto, SVM clásico y Naïve Bayes para la clasificación de comentarios turísticos. Detectando que las técnicas aplicadas pueden detectar y clasificar el sentimiento de los comentarios. En términos de exactitud, el modelo propuesto mostró la mejor *performance* alcanzando el 76,57 % en clasificación seguido del SVM clásico con 69,90 % y, por último, Naïve Bayes con 66,91 % concordando con Thabtah, Eljinini, Zamzeer, y Hadi (2009) que *chi square* como *feature selection* aumenta la exactitud del modelo. Este trabajo de investigación ha demostrado que es factible realizar la clasificación de manera automática e identificar qué lugares poseen el mayor ratio de reseñas.

Referencias

Chi, T. H., Lo, H. H., Chu, Y. H., y Lin, W. C. (2009). A mobile tourism application model based on collective interactive genetic algorithms. *Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology (ICCIT '09). IEEE Computer Society*, 244-249. doi:10.1109/ICCIT.2009.280

Li, J. B., y Yang, L. B. (2017). A Rule-Based Chinese Sentiment Mining System with Self-Expanding Dictionary - Taking TripAdvisor as an Example. *2017 IEEE 14th International Conference on e-Business Engineering (ICEBE)*, Shanghai, 238-242. doi:10.1109/ICEBE.2017.45

Lin, K. C., Wu, S. H., Chen, L. P., Ku, T., y Chen, G. D. (2014). Mining the user clusters on Facebook fan pages based on topic and sentiment analysis. *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, 627-632. doi:10.1109/IRI.2014.7051948

Parikh, V., Kestar, M., Dharia, D., y Gotmare, P. (2018). A tourist place recommendation and recognition system. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 218-222.

Thabtah, F., Eljinini, M. A. H., Zamzeer, M., y Hadi, M. (2009). Naïve Bayesian based on chi square to categorize arabic data. *Communications of the IBIMA*, *10*, 158-163.

Ye, Q., Zhang, Z., y Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*(3), 6527-6535. doi:10.1016/j.eswa.2008.07.035