

Estudio de las categorías LIWC para el análisis de sentimientos en español

María del Pilar Salas-Zárate, Miguel Ángel Rodríguez-García, Rafael Valencia-García

Departamento de Informática y Sistemas.

Campus de Espinardo s/n 30100 Murcia. España

mariapilar.salas@um.es, miguelangel.rodriguez@um.es, valencia@um.es

Ángela Almela

Departamento de Idiomas.

Universidad Católica San Antonio de Murcia, España

aalmela@ucam.edu

Abstract

Las opiniones expresadas en redes sociales o blogs son actualmente un medio en el cual se basan los usuarios para la toma de decisiones en situaciones tales como la compra de un producto o en la contratación de un servicio; recientemente, el número de opiniones expresadas ha aumentado exponencialmente en la Web. La minería de opiniones tiene como objetivo la extracción de información subjetiva a partir de contenido generado por usuarios, es decir, permite extraer un valor directo, tal como positivo o negativo, a partir de un comentario textual. Este trabajo presenta un estudio sobre la eficacia de la clasificación de opiniones en español en cinco categorías utilizando la combinación de características lingüísticas y psicológicas de LIWC. Además se presenta una evaluación comparativa de los resultados de las técnicas de clasificación J48, SMO y BayesNet utilizando la medida-F.

1 Introducción

Las opiniones expresadas en foros, blogs y redes sociales están teniendo un gran impacto en la toma de decisiones para comprar un producto, contratar un servicio, votar por un partido político, entre otras. Además, para las empresas también es importante la información expresada en estos medios para mejorar un producto o servicio. Sin embargo, el número de opiniones ha incrementado exponencialmente en la Web, por lo que leer todas las opiniones resulta imposible para los usuarios. Por estos motivos, han surgido diferentes tecnologías tal como la minería de opiniones, con la finalidad de procesar automáticamente las opiniones y saber si se está hablando de forma positiva, negativa o neutra sobre un producto o servicio y medir la intensidad de dicha opinión. En este trabajo se realiza un estudio de las distintas dimensiones lingüístico-psicológicas obtenidas por el programa LIWC (por sus siglas en inglés *Linguistic Inquiry and Word Count*, Buscador Lingüístico y Contador de Palabras) para clasificar opiniones en español en cinco categorías: positiva, negativa, neutra, muy positiva y muy negativa. Para

Copyright © by the paper's authors. Copying permitted only for private and academic purposes.

In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en <http://ceur-ws.org>

este propósito un corpus de productos tecnológicos fue compilado. Este corpus contiene opiniones obtenidas de sitios de comercio electrónico, posteriormente el corpus se procesó en LIWC para extraer las características lingüísticas. Finalmente, para evaluar los resultados de clasificación se utilizaron los clasificadores J48, SMO y BayesNet de WEKA [Bou10].

Este trabajo está estructurado de la siguiente manera: la sección 2 describe y ofrece una discusión del análisis de textos con dimensiones LIWC, la sección 3 presenta la evaluación realizada con WEKA con un experimento. Finalmente, la sección 4 describe las conclusiones.

2 LIWC

LIWC es un software que ofrece una herramienta eficaz para estudiar componentes emocionales, cognitivos y estructurales contenidos en un texto [Bou10]. Este software contiene un diccionario en español compuesto por 7.515 palabras. Cada palabra se puede clasificar en una o más de las 72 categorías incluidas por omisión en LIWC. Además, las categorías se clasifican en cuatro dimensiones: 1) procesos lingüísticos estándar, 2) procesos psicológicos, 3) relatividad, y 4) asuntos personales.

Para el presente estudio se obtuvo un corpus de 600 opiniones, específicamente 100 muy negativas, 150 negativas, 100 neutras, 150 positivas y 100 muy positivas de productos tecnológicos tales como dispositivos móviles; con el propósito de analizar los textos a través de todas las posibles combinaciones de las dimensiones de LIWC y la clasificación de opiniones: 1) positiva y negativa, 2) positiva, neutra y negativa, y 3) muy positiva, positiva, neutra, muy negativa y negativa. Una vez realizado el análisis, todos los resultados obtenidos por el programa LIWC se usaron para entrenar el clasificador de aprendizaje automático.

3 Evaluación y resultados

WEKA [Bou10] ofrece diferentes clasificadores, los cuales permiten la creación de modelos de acuerdo con los datos y el propósito del análisis. Los clasificadores se dividen en siete grupos: redes bayesianas (Nave Bayes, etc.), funciones (regresión lineal, SMO, regresión logística, etc.), aprendizaje vago (IBk, LWL, etc.), meta-clasificadores (Bagging, Vote, etc.), reglas (DecisionTable, OneR, etc.), árboles de decisión (J48, RandomTree, etc.) y otros (SerializedClassifier e InputMappedClassifier).

En este trabajo, WEKA se utilizó para evaluar los resultados de clasificación de las opiniones basándose en las categorías de LIWC. El experimento se llevó a cabo utilizando tres algoritmos: el árbol de decisión J48, la red bayesiana (BayesNet) y el algoritmo SMO para clasificadores SVM [Kee01]. Estos algoritmos fueron seleccionados debido a que han sido utilizados en otros experimentos [Nah12] [Che12] obteniendo buenos resultados en la clasificación de los datos. Para cada clasificador se realizó una validación cruzada de 10 iteraciones. Dicha validación se aplicó con el objetivo de estimar la precisión de los modelos predictivos.

Los resultados del experimento se muestran en la Table 1. En la primera columna se indica qué dimensiones de LIWC se utilizan. Por ejemplo, 1_2_3_4 indica que se han utilizado todas las dimensiones, mientras que 1_2 indica que solo se utilizaron las categorías de las dos primeras dimensiones. Posteriormente se muestran los resultados para cada clasificador J48, BayesNet y SMO con la clasificación de opiniones 2 (positiva y negativa), 3 (positiva, neutra y negativa) y 5 (muy positiva, positiva, neutra, muy negativa y negativa). Los valores que se presentan corresponden a la medida-F (F1), la media armónica de precisión y exhaustividad.

Table 1: Sample Table

	J48			BayerNet			SMO		
	2	3	5	2	3	5	2	3	5
1	0.74	0.682	0.41	0.797	0.692	0.447	0.843	0.744	0.489
2	0.799	0.670	0.462	0.833	0.706	0.49	0.822	0.722	0.469
3	0.73	0.619	0.395	0.781	0.620	0.376	0.79	0.628	0.409
4	0.741	0.618	0.377	0.761	0.636	0.397	0.755	0.602	0.461
1_2	0.803	0.704	0.496	0.882	0.761	0.521	0.886	0.777	0.539
1_3	0.751	0.741	0.418	0.819	0.776	0.457	0.832	0.710	0.493
1_4	0.771	0.676	0.424	0.812	0.713	0.466	0.832	0.722	0.496
2_3	0.819	0.699	0.498	0.878	0.747	0.523	0.862	0.741	0.495
2_4	0.809	0.671	0.478	0.853	0.740	0.515	0.844	0.737	0.49
3_4	0.737	0.655	0.422	0.811	0.678	0.416	0.817	0.714	0.48
1_2_3	0.816	0.677	0.466	0.885	0.755	0.519	0.881	0.780	0.536
1_2_4	0.82	0.701	0.498	0.866	0.766	0.523	0.879	0.774	0.53
1_3_4	0.802	0.668	0.423	0.828	0.723	0.463	0.837	0.743	0.505
2_3_4	0.804	0.690	0.452	0.875	0.759	0.528	0.867	0.762	0.502
1_2_3_4	0.83	0.682	0.513	0.875	0.759	0.532	0.904	0.780	0.571

Los resultados demuestran que los diferentes algoritmos de clasificación resultaron similares, aunque los mejores resultados se obtuvieron por los SVM. Los modelos SVM se han aplicado con éxito en muchas tareas de clasificación de texto [Rus11], debido a sus ventajas principales tales como 1) su robustez en espacios dimensionales elevados, 2) la relevancia de cualquier característica, y 3) su robustez en conjuntos escasos de muestras. Además, basados en las categorías de clasificación los mejores resultados se obtuvieron con dos categorías (positiva y negativa), es decir, con la combinación de un menor número de categorías el algoritmo realiza una mejor clasificación, debido a que al existir menos categorías el algoritmo asigna los casos con mayor exactitud a una clase u otra. Por otra parte, la combinación de todas las dimensiones de LIWC aporta el mejor resultado de clasificación con una medida-F de 90,4%. De forma individual la primera y la segunda dimensión obtienen los mejores resultados debido a la gran cantidad de palabras gramaticales que son parte de la dimensión lingüística, y al hecho de que las opiniones frecuentemente contienen palabras relacionadas con el estado emocional del autor. Finalmente, la cuarta dimensión es la que arroja los peores resultados, debido a que el tema elegido para este estudio tiene poca relación con el vocabulario correspondiente con asuntos personales.

3.1 Conclusiones

En el presente trabajo se llevó a cabo un experimento basado en la clasificación de sentimientos con el objetivo de evaluar el potencial de la clasificación de las dimensiones LIWC. Con el propósito de realizar un estudio exhaustivo, consideramos dos categorías (positiva, negativa), tres categorías (positiva, negativa y neutra) y cinco categorías (muy positiva, muy negativa, positiva, negativa y neutra) para la clasificación de opiniones en español. Por otro lado, para evaluar la eficacia de las características de LIWC se utilizó la plataforma WEKA, concretamente los clasificadores J48, BayesNet y SMO. Los resultados muestran que la clasificación de opiniones con dos categorías (positiva, negativa) obtiene mejores resultados, siendo el clasificador SMO el que tiene un mejor comportamiento.

3.1.1 Agradecimientos

Este trabajo ha sido financiado por el Ministerio español de Economía y Competitividad y la Comisión Europea (FEDER) a través del proyecto SeCloud (TIN2010- 18650)

References

- [Bou10] R. R. Bouckaert, E. Frank, M. A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten
WEKA experiences with a java open source project *Journal of Machine Learning Research*, 11:2533–

2541, 2010.

- [Pen01] J. W. Pennebaker, M. E. Francis, R. J. Booth. *Linguistic Inquiry and Word Count*. Mahwah NJ: Erlbaum Publishers, 2001.
- [Kee01] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy. Improvements to Platt's SMO Algorithm for SVM Classifier Design *Neural Computation*, 13(3):637–649, 2001.
- [Nah12] J. Nahar, K. Tickle, S. Ali, P. Chen. Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer *Expert Systems with Applications*, 39:12371–12377, June 2012.
- [Che12] L. Chen, L. Qi, F. Wang. Comparison of feature-level learning methods for mining online consumer reviews *Expert Systems with Applications*, 9588–9601, 2012.
- [Rus11] M. Rushdi Saleh, M. T. Martn Valdivia, A. Montejo, L. A. Urea. Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, 38(12):14799–14804, 2011.