

Análisis de sentimientos en textos de opinión

Una evaluación práctica

Rutilio Rodolfo López Barbosa



Primera edición: Febrero, 2019.

López Barbosa Rutilio Rodolfo (2019). Análisis de Sentimientos en textos de opinión. Una evaluación práctica. Editorial Plaza y Valdés.

Se permite la copia y distribución por cualquier medio siempre que se mantenga el reconocimiento de sus autores, no se haga uso comercial de las obras y no se realice ninguna modificación de las mismas.

© López Barbosa Rutilio Rodolfo

© Plaza y Valdés S.A. de C.V.
Alfonso Herrera No. 130 Int 11,
Col. San Rafael, Ciudad de México,
C.P. 06470, Delegación Cuauhtémoc.
Teléfono: 50.97.20.70
www.plazayvaldes.com.mx
coediciones@plazayvaldes.com

Plaza y Valdés S.L
Calle Murcia, 2 Colonia de los Ángeles.
Pozuelo de Alarcón 28223, Madrid, España.
Teléfono: 91 812 63 15
madrid@plazayvaldes.com
www.plazayvaldes.es

ISBN: 978-607-8624-26-3

Impreso en México / *Printed in Mexico*

Formato y edición: Jessica Nataly Muñiz Ortiz

Agradecimientos

Agradezco en primer lugar a Dios y dedico el fruto de este esfuerzo a mi familia pero principalmente a todas las mujeres que son y han sido el centro de mi universo desde mi nacimiento y siempre: A Rafaela Barbosa, mi madre, a mi esposa Lore, a mi pequeña Ximenita, a mi hermana Ana y a mi Daniela. Dedico también este trabajo a mi padre, a todos mis hermanos, hermanas y al resto de mi familia.

Agradezco a mis estimados amigos del laboratorio quienes han estado a unos pasos de mí cada vez que los he necesitado, a Eydel por su ímpetu, su amistad, su disponibilidad y su empatía; a Paulo por su hermandad y aprecio, por su sinceridad y por su apoyo, A Enayat por sus consejos, su sabiduría y su amabilidad; A Kary por sus palabras de ánimo, por su amistad y apoyo; A Bernabé por su compañerismo y amistad; A David por su apoyo moral y su empatía; al resto de mis compañeros con los que no tuve tantas oportunidades de convivir pero que demostraron estar dispuestos apoyarme con sus conocimientos y experiencia.

A mis guías Salvador Sánchez Alonso y a Miguel Ángel Sicilia Urbán quienes me condujeron a través de todo el proceso de investigación, redacción y difusión de mis resultados.

Resumen

El presente documento describe el proceso de investigación y desarrollo llevado a cabo en la disciplina del análisis de sentimientos. El objetivo principal de esta investigación fue evaluar la aplicación de las tecnologías del análisis de sentimientos al contenido generado por los usuarios de distintos medios sociales y presentar propuestas de aprovechamiento de los resultados de estas tecnologías a las organizaciones y usuarios. Se estudió el grado de confiabilidad de las herramientas en línea de análisis de sentimientos que trabajan con Twitter como fuente de corpus; se presentó una propuesta heurística que simplifica el análisis de sentimientos de los mensajes de Twitter centrándose en las opiniones directamente relacionadas con los objetos de opinión en lugar de determinar el sentimiento de forma global y que genera información adicional que pudiese resultar útil para el boca a boca electrónico; Finalmente se desarrolló y evaluó una propuesta de predicción de calificaciones cuantitativas de hoteles a partir de las críticas emitidas por los usuarios de sus servicios. Los resultados de esta investigación demuestran que el análisis de sentimientos es una disciplina que en su estado actual puede ser útil para la toma de decisiones para compañías e individuos y que sin embargo es susceptible de ser mejorada para el aprovechamiento de la cantidad masiva de opiniones en texto emitidas por los usuarios de los medios sociales.

Índice General

Resumen	III
Índice General	V
Índice de Figuras	X
Índice de Tablas	XII
Índice de Fórmulas	XIII
Capítulo 1. Análisis de Sentimientos e implicaciones prácticas.	21
1.1. Planteamiento del problema	22
1.1.1. Confiabilidad de las herramientas	22
1.1.2. Combinación simplificadora de enfoques	23
1.1.3. Predicción de calificación de hoteles	24
1.2. Panorama general	26
1.2.1. Aprendizaje automático.	26
1.2.2. Procesamiento de lenguaje natural (PLN)	27
1.2.3. Lingüística computacional.	28
1.2.4. Recuperación de Información.	28
1.2.5. Boca a boca electrónico (electronic word of mouth - eWoM)	28
1.2.6. Medios sociales	29
1.3. Objetivos de Investigación	30
1.3.1. Objetivo general	30
1.3.2. Objetivos específicos	30
1.4. Estructura del documento	31
Capítulo 2. Metodología y Configuración de los Experimentos	33
2.1. Revisión de estado del arte y de la literatura	33
2.2. Identificación de herramientas de análisis de sentimientos	34
2.2.1 Herramientas comerciales	34
2.2.2 Herramientas y recursos de uso libre.	35
2.3. Configuración de los experimentos	35
2.3.1 Confiabilidad de las herramientas de análisis de sentimientos	35

2.3.2	Propuesta de software de análisis de sentimientos específico para Tweets.	37
2.3.3	Predicción de calificaciones de hoteles.	37
	Capítulo 3. Contextualización y Estado del Arte	39
3.1.	Análisis de sentimientos	39
3.1.1.	Definición	39
3.1.2.	Aplicaciones y consideraciones adicionales	43
3.1.3.	Herramientas y recursos	43
3.1.3.1	Comerciales	43
3.1.3.2	De uso libres	44
3.1.4.	Nivel de análisis	46
3.1.5.	Análisis de sentimientos basado en características	47
3.1.6.	Clasificación de métodos	50
3.1.7.	Corpora	51
3.2.	Aprendizaje automático (Machine Learning)	52
3.2.1.	Aprendizaje supervisado	53
3.2.1.1	Árbol de decisiones	54
3.2.1.2	Clasificadores lineales	56
3.2.1.2.1	Máquinas de vectores de soporte (SVM)	58
3.2.1.3	Clasificadores basados en reglas	62
3.2.1.1	Clasificadores probabilísticos	63
3.2.1.1.1	Naive Bayes	63
3.2.1.1.2	Redes bayesianas	65
3.2.1.1.3	Clasificador de entropía máxima	65
3.2.2	Aprendizaje no supervisado	66
3.2.3	Aprendizaje semisupervisado	67
3.3.	Procesamiento de Lenguaje Natural (PLN)	71
3.3.1	Enfoques lingüísticos.	72
3.3.2	Enfoques probabilísticos	74
3.4.	Léxicos de Sentimientos	76
3.4.1	SentiWordNet ^[19]	77

3.4.2	Otros léxicos	79
3.5.	Medios sociales, Web 2.0 y opiniones generadas por los usuarios.	81
3.5.1	Críticas de productos	81
3.5.2	Weblogs y Noticias	82
3.5.3	Twitter	82
3.6.	Boca a boca electrónico (eWoM)	84
3.7.	Resumen	86
	Capítulo 4 Confiabilidad de Análisis de Sentimientos con Twitter	99
4.1.	Formulación del problema	99
4.2.	Metodología	99
4.3.	Configuración de los experimentos	100
4.3.1.	Selección de las herramientas	100
4.3.2.	Selección de entidades para las pruebas	105
4.4.	Recolección de los datos	108
4.4.1.	Descripción del proceso de recolección	108
4.4.2.	Normalización de los datos	108
4.4.3.	Pruebas	108
4.5.	Resultados y Discusión	110
4.5.1.	Primera semana de datos	111
4.5.1.1.	Comparación entre Herramientas	112
4.5.1.2.	Comparación de herramientas con humanos	112
4.5.2.	Segunda semana de datos (Tabla 4.7).	114
4.5.3.	Tercera semana de datos (Tabla 4.8).	116
4.5.4.	Resultados generales	118
4.5.5.	Otros resultados	119
4.6.	Resumen del capítulo	121
	Capítulo 5. Propuesta de un Método Heurístico de Análisis de Sentimientos	123
5.1	Introducción	123
5.2	Planteamiento del problema	124

5.3	Metodología	125
5.4	Descripción de la propuesta	126
5.4.1	Preprocesamiento de tweets	127
5.5.1.1.	Emoticonos	128
5.5.1.2	Diccionarios	128
5.5.1.3	Limpieza y normalización del texto	130
5.4.2	Análisis léxico	131
5.4.3	Léxico de Sentimiento	132
5.4.3.1	Vocabulario subjetivo	133
5.4.3.2	Desambiguación básica con el etiquetador POS	133
5.4.3.3	Coeficiente de neutralidad	134
5.4.3.4	Reentrenamiento del léxico	134
5.4.4	Análisis sintáctico	136
5.4.5	Análisis de frases	137
5.4.5.1	Identificación del objeto de opinión	138
5.4.5.2	Calificación de frases	140
5.4.5.3	Análisis de dependencias entre frases	141
5.4.5.4	Comparaciones y negaciones	142
5.4.5.5	Preguntas	143
5.4.6	Clasificación del sentimiento	143
5.4.6.1	Tweets con oraciones múltiples	143
5.4.6.2	Identificación de términos clave	144
5.5	Configuración de los experimentos	147
5.6	Resultados y Discusión	150
5.6.1	Confiabilidad de clasificación por humanos	150
5.6.2	Medidas de eficiencia	152
5.6.3	Resultados de eficiencia de las herramientas seleccionadas	153
5.6.4	Resultados de eficiencia de la propuesta	155
5.7	Resumen del capítulo	159
Capítulo 6. Predicción de calificación de hoteles mediante el Análisis de Sentimientos		161

6.1.	Introducción	161
6.2.	Descripción detallada del problema	162
6.3.	Metodología	163
6.3.1.	Extracción de datos	164
6.3.1.1	TripAdvisor	164
6.3.1.2	Desarrollo del Web Crawler	165
6.3.1.3	Extracción de datos	169
6.3.2.	Análisis de Sentimientos	170
6.3.2.1	OpinionFinder (OFV2)	171
6.3.2.2	Stanford CoreNLP (RNTN)	175
6.3.2.3	SentUAH	178
6.3.3.	Análisis estadístico	180
6.4.	Resultados y discusiones	181
6.4.1.	Correlación entre las calificaciones globales y los porcentajes	181
6.4.2.	Predicción de calificaciones	185
6.5.	Resumen del capítulo	187
Capítulo 7. Conclusiones y trabajo futuro		189
7.1.	Resumen de conclusiones y resultados	189
7.2.	Conclusiones	190
7.2.1.	Confiabilidad de los resultados del análisis de sentimientos	191
7.2.2.	Simplificación de análisis de sentimientos centrado en el objeto de opinión.	192
7.2.3.	Predicción de calificaciones de hoteles.	193
7.2.4.	Generalización de los resultados.	194
7.3.	Recomendaciones	195
7.3.1.	Orientadas a otros investigadores	195
7.3.1.1.	Relacionadas con los métodos	195
7.3.1.2.	Relacionadas con la evaluación	195
7.3.1.3.	Relacionadas con los medios sociales	196
7.3.2.	Orientadas a los usuarios	197
7.4.	Aportaciones originales	197

7.5. Trabajo futuro	198
7.5.1. Confiabilidad de los resultados del análisis de sentimientos	198
7.5.2. Simplificación del análisis de sentimientos centrado en el objeto de opinión	198
7.5.3. Predicción de calificaciones de hoteles.	199
Referencias	201

Índice de Figuras

Figura 3.1. Clasificación de Métodos de análisis de sentimientos	51
Figura 3.2. Aprendizaje automático Supervisado	55
Figura 3.3. Representación de un conjunto de datos y su árbol de decisión	56
Figura 3.4. Representación del hiperplano en un problema de dos clases	58
Figura 3.5. Niveles de análisis lingüístico y tareas asociadas	72
Figura 4.1. Interfaz de Opinion Crawl	103
Figura 4.2. Interfaz de Sentimentor	103
Figura 4.3. Interfaz de Sentiment140	103
Figura 4.4. Interfaz de StreamCrab	103
Figura 4.5. Interfaz de TweetFeel	104
Figura 4.6. Interfaz de Twitrratr	104
Figura 4.7. Servicios relacionados con tecnologías de información (primera semana)	113
Figura 4.8. Productos no relacionados con tecnologías de información (primera semana)	114
Figura 5.1. Etapas del procesamiento de Sentweet	127
Figura 5.2. Ejemplo de cola de elementos	132
Figura 5.3. Árbol de estructura de las frases de ejemplo	137
Figura 5.4. Grafo para el análisis gramatical	138

Figura 5.5. Nube de palabras para el vocabulario en opiniones positivas	146
Figura 5.6. Nube de palabras para el vocabulario en opiniones negativas	147
Figura 5.7. Resultados de Sentiment140 con tweets de iPad	148
Figura 5.8. Resultados de Twitrratr con tweets de NetFlix	148
Figura 5.9. Resultados de TweetFeel con tweets de Hotmail	149
Figura 5.10. Representación de la eficiencia en la clasificación	152
Figura 6.1. Registro de crítica y calificación para un hotel en TripAdvisor	165
Figura 6.2. Registro de opinión cuantitativa de los servicios del hotel	165
Figura 6.3. Página inicial de TripAdvisor	166
Figura 6.4. Página inicial de TripAdvisor y Firebug	167
Figura 6.5. Ruta de navegación para recolección de datos	167
Figura 6.6. Lista de hoteles de una ciudad	168
Figura 6.7. Lista de comentarios de un hotel	168
Figura 6.8. Datos del usuario	168
Figura 6.9. Base de Datos de crítica de hoteles	169
Figura 6.10. Base de Datos de críticas Analizadas por OpinionFinder	173
Figura 6.11. Ejemplo de clasificación utilizando RNTN	176
Figura 6.12. Base de Datos de crítica Analizadas por RNTN	177
Figura 6.13. Análisis de sentimientos de las críticas usando SentUAH	180
Figura 6.14. Base de datos de las críticas clasificadas por SentUAH	180
Figura 6.15. Correlación entre calificaciones reales y porcentajes positivos (París)	183
Figura 6.16. Correlación entre calificaciones reales y porcentajes positivos (Nueva York)	184
Figura 6.17. Correlación entre calificaciones reales y porcentajes positivos (Londres)	184

Índice de Tablas

Tabla 3.1. Resumen de artículos	88
Tabla 4.1. Información de las herramientas seleccionadas	102
Tabla 4.2. Lista de objetos propuestos para las pruebas	107
Tabla 4.3. Lista final de objetos de prueba	107
Tabla 4.4. Resultados de la primera semana para Gmail	108
Tabla 4.5. Primera Prueba: Datos normalizados positivos de herramientas	109
Tabla 4.6. Primera semana de pruebas de herramientas comparativa con humanos	111
Tabla 4.7. Segunda prueba: Datos normalizados positivos	115
Tabla 4.8. Tercera prueba: Datos normalizados positivos	117
Tabla 4.9. Comportamiento del alfa de Cronbach durante el periodo de pruebas	119
Tabla 5.1. Emoticonos positivos y negativos	128
Tabla 5.2. Ejemplos de errores ortográficos comunes	129
Tabla 5.3. Ejemplos de repetición de letras	129
Tabla 5.4. Ejemplos de Acrónimos, siglas y jerga	130
Tabla 5.5. Ejemplos de tweet clasificados	135
Tabla 5.6. Vocabulario más frecuente en las opiniones positivas hacia Gmail	145
Tabla 5.7. Vocabulario más frecuente en las opiniones negativas hacia Gmail	146
Tabla 5.8. Lista de productos y servicios empleados en los experimentos	149
Tabla 5.9. Clasificación manual de expertos humanos	150
Tabla 5.10. Resultados del análisis de sentimientos por humanos	151
Tabla 5.11. Eficiencia de las herramientas seleccionadas (tweets Positivos)	154
Tabla 5.12. Eficiencia de las herramientas seleccionadas (tweets Negativos)	154

Tabla 5.13. Eficiencia de las herramientas seleccionadas tweets (Neutros/Objetivos)	155
Tabla 5.14. Eficiencia de Sentweet (tweets Positivos)	155
Tabla 5.15. Eficiencia de Sentweet (tweets Negativos)	157
Tabla 5.16. Eficiencia de Sentweet (tweets Neutros)	158
Tabla 5.17. Comparativa de eficiencia en la clasificación de tweets neutros	158
Tabla 5.18. Porcentajes mejorados de eficiencias, clasificación centrada en el objeto	158
Tabla 6.1. Información de las críticas recopiladas	170
Tabla 6.2. Correlaciones entre calificación de hoteles y porcentajes positivos	182
Tabla 6.3. Comparación entre exactitud reportada y correlaciones de esta investigación	185
Tabla 6.4. Alfa de Cronbach, confiabilidad entre calificaciones reales y calculadas	187
Tabla 7.1 Resumen de objetivos y resultados	189

Índice de Fórmulas

Fórmula 3.1. Asignación de la clase para textos con Naive Bayes	64
Fórmula 5.1. Exactitud	153
Fórmula 5.2. Precisión	153
Fórmula 5.3. Recuperación	153
Fórmula 6.1. Conjunto de oraciones	174
Fórmula 6.2. Conjunto de clases	174
Fórmula 6.3. Clasificación de oraciones	174
Fórmula 6.4. Clasificación de críticas	175
Fórmula 6.5. Clasificación de oraciones por SentUAH	179

Análisis de Sentimientos en textos de opinión.

Una evaluación práctica

Se terminó de imprimir en febrero de 2019

Tiraje: 1,000 ejemplares