



UNIVERSIDAD DE MURCIA

FACULTAD DE INFORMÁTICA

Detección de Patrones Psicolingüísticos para el
Análisis de Lenguaje Subjetivo en Español

Dña. María del Pilar Salas Zárate
2017



UNIVERSIDAD DE MURCIA

Facultad de Informática

Tesis Doctoral

**Detección de patrones psicolingüísticos para el análisis de lenguaje
subjetivo en español**

María del Pilar Salas Zárate

Directores:

Dr. Rafael Valencia García

Dr. Miguel Ángel Rodríguez García

2017

Agradecimientos

Le agradezco a Dios por haberme acompañado y guiado a lo largo de este camino, por ser mi fortaleza en los momentos de debilidad y por brindarme una vida llena de aprendizajes, experiencias y sobre todo felicidad.

Quiero agradecer de manera especial al Dr. Rafael Valencia García y al Dr. Miguel Ángel Rodríguez González, directores de esta tesis doctoral, por la gran confianza que han tenido en mí, por su apoyo, sus apreciados y relevantes aportes, sugerencias y consejos durante el desarrollo de esta investigación.

Un sincero agradecimiento al Dr. Ricardo Colomo Palacios de Østfold University College por aceptarme para realizar una estancia de investigación y contribuir al enriquecimiento de este trabajo. Al Dr. Giner Alor Hernández del Instituto tecnológico de Orizaba por su apoyo y consejos desde la maestría y a los profesores miembros del grupo de investigación TECNOMOD de la Universidad de Murcia.

Mi agradecimiento especial a Mario Andrés Paredes Valverde por su comprensión, por inspirarme siempre, por darme empujones constantes, por alentarme cada día para intentar ser una mejor persona, por compartir lo bueno y lo malo y, sobre todo, por creer en mí.

A mi amigo Omar con quien he compartido momentos inolvidables desde hace 10 años. Gracias por su apoyo incondicional y por siempre escucharme.

A mis compañeros del laboratorio, Ginés, Jojo, Manu, Mari Carmen, Frank, Ángel, Philip, Denis, y a todos los que estuvieron en algún momento, Lucia, Astrid, Teddy y muchos más, por todas las horas que hemos pasado juntos, por las charlas y sobre todo por su amistad. Y a todos mis amigos de México que a pesar de la distancia siempre han estado apoyándome.

A Marcelino, Asunción y Paty, por sus consejos, por abrirme las puertas de su hogar y hacerme sentir como si estuviera en México. A la familia de Ginés, Loli y Ángel por su amistad.

Y, por supuesto, el agradecimiento más profundo y sentido va para toda mi familia, quienes siempre han estado al pendiente de mi aun en la lejanía y quienes me han alentado día a día. A mi madre Marina Salas, gracias por tu esfuerzo y por la confianza que depositaste en mí, gracias porque siempre has estado a mi lado y por haberme ayudado a alcanzar todas las metas que me he propuesto. A mi abuelo Delfino Salas y abuela Josefina Zárate(QEPD), por estar siempre en los momentos importantes de mi vida y por todo el apoyo brindado a lo largo de mi vida. A mi tío Rafael Salas, por brindarme su apoyo en todo momento, por ser como un padre para mí y sobre todo por ser el ejemplo para salir adelante. Gracias por los consejos que han sido de gran ayuda para mi vida y crecimiento. A mi tío Isidro Salas, por estar conmigo y apoyarme siempre.

Esta tesis ha sido posible gracias al Consejo Nacional de Ciencia y Tecnología (CONACyT) a través de la beca al extranjero 2013, y a la Secretaría de Educación Pública (SEP) a través de la beca complemento.

ÍNDICE DE CONTENIDO

Capítulo 1. Introducción.....	15
1.1. Organización del documento	16
Capítulo 2. Estado del arte.....	19
2.1. Introducción.....	19
2.2. Lenguaje subjetivo	20
2.3. Procesamiento de Lenguaje natural.....	20
2.3.1. Niveles del procesamiento de lenguaje natural.....	21
2.3.1.1. Nivel fonológico.....	22
2.3.1.2. Nivel morfológico	22
2.3.1.3. Nivel léxico	23
2.3.1.4. Nivel sintáctico	23
2.3.1.5. Nivel semántico	24
2.3.1.6. Nivel de discurso o contextual.....	24
2.3.1.7. Nivel pragmático	25
2.4. Análisis de sentimientos.....	25
2.4.1. Evolución histórica del análisis de sentimientos	26
2.4.2. Niveles de análisis de sentimientos.....	28
2.4.2.1. Análisis de sentimientos a nivel de documento	29
2.4.2.2. Análisis de sentimientos a nivel de sentencia.....	30
2.4.2.3. Análisis de sentimientos a nivel de características	30
2.4.3. Recursos lingüísticos para el análisis de sentimientos.....	31
2.4.3.1. Corpus.....	31
2.4.3.2. Bases de conocimiento	32
2.4.3.2.1. eSOL.....	33
2.4.3.2.2. General Inquierer	33

Índices

2.4.3.2.3.	iSOL.....	34
2.4.3.2.4.	Léxico de opiniones de Bing Liu	34
2.4.3.2.5.	ML-SentiCon	35
2.4.3.2.6.	MPQA (Multi-Perspective Question Answering).....	35
2.4.3.2.7.	SentiWordNet	36
2.4.3.2.8.	Spanish Sentiment Lexicons	36
2.4.4.	Enfoques para el análisis de sentimientos.....	37
2.4.4.1.	Enfoque basado en aprendizaje automático.....	39
2.4.4.1.1.	Extracción de características	40
2.4.4.1.1.1.	Términos y su frecuencia	41
2.4.4.1.1.2.	Características basadas en etiquetas Part of Speech (POS)	42
2.4.4.1.1.3.	Características de dependencia sintáctica.....	43
2.4.4.1.2.	Métodos de selección de características	44
2.4.4.1.3.	Algoritmos de clasificación	46
2.4.4.1.3.1.	Máquina de soporte vectorial.....	46
2.4.4.1.3.2.	Red bayesiana.....	46
2.4.4.1.3.3.	Árboles de decisión.....	47
2.4.4.2.	Enfoque semántico.....	49
2.4.5.	Comparación de métodos para el análisis de sentimientos.....	53
2.4.5.1.	Dominio.....	55
2.4.5.2.	Lenguaje.....	57
2.4.5.3.	Enfoque	58
2.4.5.4.	Características.....	59
2.4.5.5.	Nivel.....	60
2.5.	El lenguaje figurado	60
2.5.1.	Ironía, Sarcasmo, Sátira.....	60
2.5.2.	Comparación de enfoques para la detección de la sátira.....	62

2.5.2.1.	Idioma.....	65
2.5.2.2.	Características.....	65
2.5.2.2.1.	Dominio	66
2.5.2.3.	Enfoque	66
2.6.	Características psicolingüísticas	68
2.6.1.	LIWC en español	68
2.6.1.1.	Categorías de LIWC 2007	69
2.7.	Objetivos de la tesis doctoral.....	74
2.7.1.	Motivación.....	74
2.7.2.	Objetivos	76
2.7.3.	Metodología.....	76
2.8.	Conclusiones	77
Capítulo 3. Método para la clasificación de sentimientos y detección de sátira		
	79	
3.1.	Introducción.....	79
3.2.	Descripción del problema y objetivos específicos	80
3.3.	Método.....	81
3.3.1.	Conjunto de datos	82
3.3.2.	Pre-procesamiento del texto	84
3.3.2.1.	Texto desde sitios web de opiniones	85
3.3.2.2.	Texto de redes sociales	86
3.3.3.	Extracción de características.....	89
3.3.4.	Entrenamiento de algoritmos de aprendizaje automático	101
3.4.	Conclusiones	105
Capítulo 4. Validación y discusión		
	107	
4.1.	Introducción.....	107

Índices

4.2. Evaluación.....	107
4.3. Validación del método de análisis de sentimientos.....	112
4.3.1. Dominio de Películas	112
4.3.2. Dominio de turismo	118
4.3.3. Resultados.....	118
4.3.4. Resultados en el dominio de películas	119
4.3.4.1. Resultados en el dominio de turismo	125
4.3.5. Discusión de los resultados.....	132
4.4. Validación del método de detección de la sátira	136
4.4.1. Corpora.....	137
4.4.2. Resultados.....	140
4.4.3. Discusión de resultados.....	148
4.5. Conclusiones	160
Capítulo 5. Conclusiones y líneas futuras.....	163
5.1. Conclusiones	163
5.2. Contribuciones	163
5.3. Líneas futuras	165
Capítulo 6. Contribuciones científicas.....	167
6.1. Publicaciones JCR.....	167
6.2. Congresos Internacionales	167
6.3. Capítulos en libro	168
Capítulo 7. Resumen extendido en inglés / Extended summary in English....	169
7.1. Introduction	169
7.2. State of the art	170
7.3. Method	171
7.3.1. Pre-processing.....	172

7.3.2.	Feature extraction.....	173
7.3.3.	Training of machine learning algorithms.....	177
7.4.	Evaluation and Results.....	178
7.5.	Discussion	181
7.6.	Conclusions and future work.....	182
Referencias		185

ÍNDICE DE FIGURAS

Figura 2.1. Niveles del procesamiento de lenguaje natural..... 22

Figura 2.2. Análisis léxico. 23

Figura 2.3. Análisis sintáctico..... 24

Figura 2.4. Análisis de sentimientos..... 26

Figura 2.5. Niveles del Análisis de sentimientos..... 29

Figura 2.6. ML-Senticon en español..... 35

Figura 2.7. Extracto de MPQA 35

Figura 2.8. SentiWordNet. 36

Figura 2.9. Extracto de fullStrengthLexicon..... 37

Figura 2.10. Extracto de mediumStrengthLexicon. 37

Figura 2.11. Enfoques del análisis de sentimientos..... 39

Figura 2.12. Proceso de aprendizaje automático. 40

Figura 2.13. Árbol de dependencias..... 44

Figura 2.14. Red bayesiana. 47

Figura 2.15. Tabla de distribución de probabilidad. 47

Figura 2.16. Árbol de decisión. 48

Figura 2.17. Producto con comentarios irónicos..... 62

Figura 3.1. Enfoque propuesto para la clasificación de sentimientos y detección de la
sátira..... 82

Figura 3.2. Opinión muy positiva..... 83

Figura 3.3. Opinión negativa..... 83

Figura 3.4. Tweet no-satírico..... 84

Figura 3.5. Tweet satírico. 84

Figura 3.6. Proceso de tokenización.....	85
Figura 3.7. Ejemplo de opinión con abreviaciones y errores ortográficos.....	86
Figura 3.8. Ejemplo de tweet.....	87
Figura 3.9. Detección de elementos como menciones, hashtag y URL.....	88
Figura 3.10. Texto sin URL, hashtags y menciones.....	89
Figura 3.11. Extracto de categorías de LIWC 2007 en español.....	91
Figura 3.12. Ejemplo de palabras contenidas en el diccionario de LIWC 2007 en español.	92
Figura 3.13. Extracto de una opinión analizada por LIWC.	94
Figura 3.14. Palabras de la opinión detectadas con LIWC.	94
Figura 3.15. Tweet analizado con LIWC.....	95
Figura 3.16. Palabras encontradas con el diccionario LIWC.....	95
Figura 3.17. Palabras categorizadas.....	96
Figura 3.18. Palabras de la opinión categorizadas.....	97
Figura 3.19. Extracto del vector de características resultante del análisis de una opinión.....	98
Figura 3.20. Extracto del vector de características resultante del análisis de un tweet.	98
Figura 3.21. Aprendizaje automático para el análisis de sentimientos.....	103
Figura 3.22. Aprendizaje automático para la detección de tweets satíricos.	104
Figura 4.1. Ejemplo de una opinión muy positiva.....	118
Figura 4.2. Comparación de clases para el dominio de películas.....	133
Figura 4.3. Comparación de clases para el dominio de turismo.....	134
Figura 4.4. Comparación de algoritmos en el dominio de películas.....	135
Figura 4.5. Comparación de algoritmos en el dominio de turismo.	136

Índices

Figura 4.6. Tweet no-satírico	139
Figura 4.7. Tweet satírico.	139
Figura 4.8. Características relevantes para España.....	147
Figura 4.9. Características relevantes para México.	147
Figura 4.10. Porcentaje medio de palabras de tweets satíricos y no satíricos.	148
Figura 4.11. Tweet satírico con emociones positivas.....	149
Figura 4.12 Tweets satíricos con palabras relacionadas a mecanismo cognitivos	150
Figura 4.13. Ejemplos de tweets con uso de hipérbole.....	152
Figura 4.14. Tweets satíricos con adverbios.....	153
Figura 4.15. Tweets satíricos con signos de puntuación.....	154
Figura 4.16. Tweets no satíricos con signos de puntuación.....	155
Figura 4.17. Porcentaje medio de palabras de características de España y México....	158
Figura 4.18. Curva ROC para el conjunto de datos de México.	159
Figura 4.19. Curva ROC para el conjunto de datos de España.....	160
Figure 7.1. Proposed method.....	172
Figure 7.2. Results of the movies domain.....	179
Figure 7.3. Results of the tourism domain.....	180
Figure 7.4. Results for satire detection.....	181

ÍNDICE DE TABLAS

Tabla 2.1. Número de documentos por polaridad del corpus de críticas de cine.....	32
Tabla 2.2. Ejemplo de palabras positivas y negativas de eSOL.....	33
Tabla 2.3. Ejemplo de palabras positivas y negativas de General Inquierer.	33
Tabla 2.4. Palabras positivas y negativas de iSOL.....	34
Tabla 2.5. Palabras positivas y negativas del léxico de opiniones de Bing Liu.	34
Tabla 2.6. Patrones para la extracción de frases.	43
Tabla 2.7. Datos de entrenamiento.	48
Tabla 2.8. Recursos léxicos.....	49
Tabla 2.9. Puntuación de algunos synset en SentiWordNet.....	51
Tabla 2.10. Comparativa de enfoques actuales de análisis de sentimientos.	54
Tabla 2.11. Resumen de los conjuntos de datos.	56
Tabla 2.12. Tabla comparativa de métodos para la detección de la sátira.....	63
Tabla 2.13. Categorías LIWC 2007 de procesos lingüísticos.....	70
Tabla 2.14. Categorías LIWC 2007 de Procesos psicológicos.....	71
Tabla 2.15. Categorías LIWC 2007 de Asuntos personales.	73
Tabla 2.16. Categorías LIWC 2007 de Lenguaje oral.	73
Tabla 2.17. Categorías LIWC de signos de puntuación.....	74
Tabla 3.1. Términos del diccionario.	86
Tabla 3.2. Tokens tweet.....	87
Tabla 3.3. Categorías LIWC.	99
Tabla 3.4. Combinación por pares de categorías LIWC.....	99
Tabla 3.5. Combinación de tres categorías LIWC.	100
Tabla 3.6 Combinación de cuatro categorías de LIWC.....	101

Índices

Tabla 4.1. Matriz de confusión para el análisis de sentimientos.....	110
Tabla 4.2. Matriz de confusión para 5 clases.....	110
Tabla 4.3. Resultados de la clasificación.....	111
Tabla 4.4. Matriz de confusión para la detección de la sátira.....	112
Tabla 4.5. Corpus de películas.....	113
Tabla 4.6. Extracto del corpus de películas.....	114
Tabla 4.7. Experimentos con el algoritmo SMO.	120
Tabla 4.8. Experimentos con el algoritmo BayesNet.	122
Tabla 4.9. Experimentos en el dominio de películas con el algoritmo J48.	124
Tabla 4.10. Experimentos en el dominio turístico con el algoritmo SMO.....	126
Tabla 4.11. Experimentos en el dominio turístico con el algoritmo BayesNet.....	128
Tabla 4.12. Experimentos en el dominio turístico con el algoritmo J48.	130
Tabla 4.13. Cuentas de Twitter.....	137
Tabla 4.14. Experimentos obtenidos con el algoritmo SMO.	141
Tabla 4.15. Experimentos obtenidos con el algoritmo BayesNet.....	143
Tabla 4.16. Experimentos obtenidos con el algoritmo J48.....	144
Tabla 4.17. Ejemplos de tweets satíricos con palabras en presente.....	157
Table 7.1. Examples of LIWC categories.....	174
Table 7.2. All possible combinations of LIWC categories.....	176

LISTA DE ACRÓNIMOS

Término	Significado
ANN	Artificial Neural Networks
IG	Information Gain
LIWC	Linguistic Inquiry and Word Count
MaxEnt	Maximum Entropy
ML	Machine Learning
NLP o PLN	Natural Language Processing
POS-Taggers	Part-Of-Speech Taggers
ROC	Receiver Operating Characteristic
SA	Sentiment Analysis
SMO	Sequential Minimal Optimization
SO	Semantic Orientation
SVM	Support Vector Machine
SWN	SentiWordNet
TF-IDF	Term frequency – Inverse document frequency
URL	Uniform Resource Locator
WEKA	Waikato Environment for Knowledge Analysis
WN	WordNet
XML	eXtensible Markup Language

Capítulo 1. Introducción

Las opiniones son una parte importante en las decisiones del ser humano, cuando una persona desea tomar una decisión se basa en los comentarios de otras personas, por ejemplo, para comprar un producto, seleccionar un destino turístico, incluso para votar por un partido político. Con el surgimiento de la Web 2.0, ya no sólo se dependía de las opiniones de familiares o amigos, sino que se podía acceder a una gran cantidad de información en la Web provista por otros usuarios. Por lo que, actualmente, las personas visitan blogs, foros de discusión o redes sociales con el objetivo de obtener las experiencias de otros usuarios antes de tomar una decisión.

La lingüística es una de las áreas que se ha enfocado en el estudio de la opinión, o mejor dicho del lenguaje subjetivo. Este tipo de lenguaje se emplea para expresar estados personales en el contexto de una conversación o un texto (J. Wiebe, Wilson, Bruce, Bell, & Martin, 2004) (Martínez Cámara, 2016). Por otro lado, el análisis de sentimientos, también conocido como minería de opiniones, se ha convertido en un tema muy popular que tiene como objetivo el procesar opiniones públicas disponibles en la Web a través de técnicas de procesamiento de lenguaje natural. En este contexto, diferentes propuestas basadas en aprendizaje automático y orientación semántica han surgido en los últimos años. Estos trabajos abordan problemas tales como análisis y construcción de lexicones de sentimientos, evaluación y clasificación de mensajes de Twitter, negación, por mencionar algunos. Otros trabajos se centran en analizar las opiniones en diferentes niveles, a saber, documento, sentencia y aspectos.

A pesar de los esfuerzos llevados a cabo en el análisis de sentimientos existen diversas características psicológicas y lingüísticas que no han sido abordadas. Por lo tanto, la clasificación automática de opiniones requiere un esfuerzo multidisciplinario, donde la lingüística y el procesamiento del lenguaje natural juegan un rol importante. Gracias a estas disciplinas es posible entender el lenguaje humano, clasificar las opiniones y resumir los sentimientos expresados acerca de un producto, servicio o cualquier otro aspecto.

Capítulo 1. Introducción

El lenguaje figurado tal como la ironía, el sarcasmo y la sátira juega un papel muy importante en los sistemas de análisis de sentimientos. El doble sentido expresado en una opinión o comentario a través de este lenguaje puede invertir la polaridad de la opinión. Aunque, el lenguaje figurado ha sido ampliamente estudiado por diversas áreas como la lingüística, solo pocos estudios se han enfocado en la detección automática.

Por otro lado, es importante mencionar que pocos trabajos para el análisis de sentimientos, y lenguaje figurado, se han enfocado en el idioma español, quizá debido a la carencia de recursos lingüísticos en ese idioma. Sin embargo, el estudio del español es de lo más importante ya que es uno de los idiomas más utilizados en internet.

Las razones expuestas en los párrafos anteriores han sido la principal motivación para la realización de la investigación que se describe en esta tesis doctoral. Es por ello que se propone un método para la detección de patrones psicolingüísticos para el análisis de sentimientos y la detección de la sátira en español. Este método permitirá, a través de un enfoque automático supervisado, clasificar textos como positivo, negativo, neutro, muy positivo o muy negativo y como satíricos y no satíricos.

1.1. Organización del documento

La presente tesis doctoral describe la investigación que se ha llevado a cabo en el campo del análisis de sentimientos y procesamiento del lenguaje natural. El documento se divide en siete capítulos que exponen el estudio que se realizó.

El capítulo 1, en el que nos encontramos, provee una breve introducción a la tesis y su organización.

El capítulo 2 consiste en un detallado estudio del arte de las tecnologías base para el desarrollo del método propuesto. El estudio inicia con una introducción al lenguaje subjetivo. Posteriormente, se presenta el campo del procesamiento del lenguaje natural, así como los diferentes niveles de procesamiento. Después, se describe el campo del análisis de sentimientos y se proporcionan las definiciones más utilizadas por la comunidad investigadora. Además, se presenta su evolución histórica desde sus inicios en el siglo XX hasta la fecha. Asimismo, se presentan los

diferentes niveles de análisis de opiniones, así como los dos principales enfoques en los cuales se basan la mayoría de estudios, a saber, orientación semántica y aprendizaje automático. En el penúltimo apartado, se provee una introducción al lenguaje figurado, específicamente la ironía, el sarcasmo y la sátira. Finalmente, se presenta la importancia de las características psicolingüísticas en el lenguaje humano, y se introduce la herramienta LIWC, la cual permite obtener variables psicolingüísticas desde un texto escrito.

El capítulo 3 describe el método para el análisis de sentimientos y la detección de la sátira propuesto en este trabajo de investigación.

El capítulo 4 se centra en la validación del método propuesto para el análisis de sentimientos y detección de la sátira. Este capítulo se divide en dos apartados. En el primer apartado se presentan y discuten los resultados del análisis de sentimientos en dos dominios, a saber, películas y turismo. En el segundo apartado se presentan los resultados y discusión para la detección de la sátira, el cual fue validado en el dominio de noticias.

En el capítulo 5 se presentan las conclusiones obtenidas del trabajo de investigación y las posibles vías futuras.

El capítulo 6 presenta un listado de las publicaciones científicas obtenidas a partir de este trabajo.

Finalmente, el capítulo 7 incluye un resumen breve de la tesis en inglés.

Capítulo 2. Estado del arte

2.1. Introducción

En el presente capítulo se introducen las diferentes tecnologías que son parte fundamental para el desarrollo de esta tesis doctoral, entre las que destacan, análisis de sentimientos, procesamiento de lenguaje natural (PLN), y la psicolingüística.

En primer lugar, se da una introducción al lenguaje subjetivo, el cual es el tema central del presente estudio. En segundo lugar, se presenta la disciplina del procesamiento de lenguaje natural, una de las principales tecnologías de la presente tesis doctoral. Posteriormente, se analizan los diferentes niveles de análisis del procesamiento de lenguaje natural.

En tercer lugar, se muestra un estudio de las principales técnicas y enfoques del análisis de sentimientos. Este estudio se inicia con la definición del término “análisis de sentimientos”, este término ha sido definido por varios investigadores, sin embargo, se presentan las más utilizadas en la literatura. Posteriormente, se presenta una reseña historia desde sus inicios hasta la actualidad. Además, se introducen los diferentes niveles y los enfoques principales para abordar el análisis de opiniones, así como los recursos lingüísticos más importantes disponibles para la comunidad investigadora. Finalmente, se lleva a cabo un análisis de los trabajos más relacionados y se comparan con el método propuesto en este trabajo de tesis doctoral.

En cuarto lugar, se aborda el tema del lenguaje figurado. Concretamente, se da un estudio sobre la ironía, el sarcasmo y la sátira. Debido a que este estudio está principalmente centrado en la sátira, se realiza un análisis de los principales métodos propuestos para su detección y se lleva a cabo una comparación con el método propuesto.

En quinto lugar, se presenta un análisis relacionado con la psicolingüística, y se introduce la herramienta LIWC, la cual ha demostrado ser eficiente en estudios relacionados con el lenguaje humano.

Capítulo 2. Estado del arte

Finalmente, se presenta las principales motivaciones para el desarrollo del presente estudio, así como los objetivos planteados, y la metodología a seguir para el correcto desarrollo del problema planteado.

2.2. Lenguaje subjetivo

La subjetividad en el lenguaje natural se refiere a aspectos del lenguaje utilizados para expresar opiniones, evaluaciones y especulaciones (J. M. Wiebe, 1994). El significado subjetivo y las dimensiones interpersonales, sociales y afectivas del lenguaje, es un tema importante en la lingüística computacional que ha sido discutido por varios autores desde hace varios años (J. Wiebe et al., 2004) .

Las emociones han sido estudiadas en muchos campos, por ejemplo, la psicología, filosofía, sociología, biología, etc. Sin embargo, no hay un conjunto de emociones básicas en las que coincidan los investigadores. De acuerdo a (Parrott, 2001), las personas tienen seis tipos de emociones primarias, es decir, el amor, alegría, sorpresa, ira, tristeza y miedo, que pueden subdividirse en emociones secundarias o terciarias. Además, cada emoción puede tener diferentes intensidades, los puntos fuertes de una opinión se relacionan con la intensidad de las emociones. Los conceptos de emociones y opiniones no son equivalentes, aunque tienen una gran relación. Cuando se habla de sentimientos subjetivos de emociones u opiniones, es necesario distinguir dos aspectos diferentes, los estados mentales o sentimientos de las personas y las expresiones del lenguaje utilizadas para describir los estados mentales. Aunque solo hay seis tipos de emociones, hay un gran número de expresiones del lenguaje que se pueden usar para expresarlos. Del mismo modo, hay un gran número de expresiones de opinión que describen sentimientos positivos y negativos. El análisis de sentimientos trata de inferir los sentimientos de las personas basándose en sus expresiones idiomáticas. Sin embargo este campo no solo pretende inferir opiniones positivas o negativas, sino también descubrir otras piezas de información asociadas que son importantes para las aplicaciones de análisis de opiniones (Indurkha & Damerau, 2010).

2.3. Procesamiento de Lenguaje natural

El procesamiento de lenguaje natural (PLN) es el campo que se basa en técnicas computacionales para analizar y representar textos que ocurren naturalmente en

uno o más niveles del análisis lingüístico con el propósito de obtener el procesamiento de lenguaje similar al humano para una serie de tareas o aplicaciones.

Varias definiciones han sido propuestas por diferentes autores para el término PLN, las más conocidas son las de (Liddy, 2001) y (Chowdhury, 2003).

(Liddy, 2001) propone la siguiente definición: *“El procesamiento de lenguaje natural es un conjunto de técnicas computacionales teóricamente motivadas para analizar y representar naturalmente textos de origen natural en uno o más niveles de análisis lingüísticos para lograr el propósito de procesar el lenguaje humano por una serie de tareas o aplicaciones”*.

(Chowdhury, 2003) define PLN como: *“un área de investigación que explora cómo las computadoras pueden utilizarse para entender y manipular texto escrito en lenguaje natural o del habla para hacer operaciones útiles”*.

En base a las definiciones anteriores se puede determinar que el objetivo del PLN es el desarrollo de métodos, recursos, y herramientas que permitan analizar y entender el lenguaje natural de tal forma que permita la comunicación entre el humano y máquinas.

2.3.1. Niveles del procesamiento de lenguaje natural

Según (Feldman, 1999), el estudio del lenguaje se estructura normalmente en 7 niveles de análisis: 1) Nivel fonológico, 2) Nivel morfológico, 3) Nivel léxico, 4) Nivel sintáctico, 5) Análisis semántico, 6) Nivel del discurso y 7) Nivel pragmático. La Figura 2.1 muestra un resumen de los diferentes niveles de lenguaje natural.

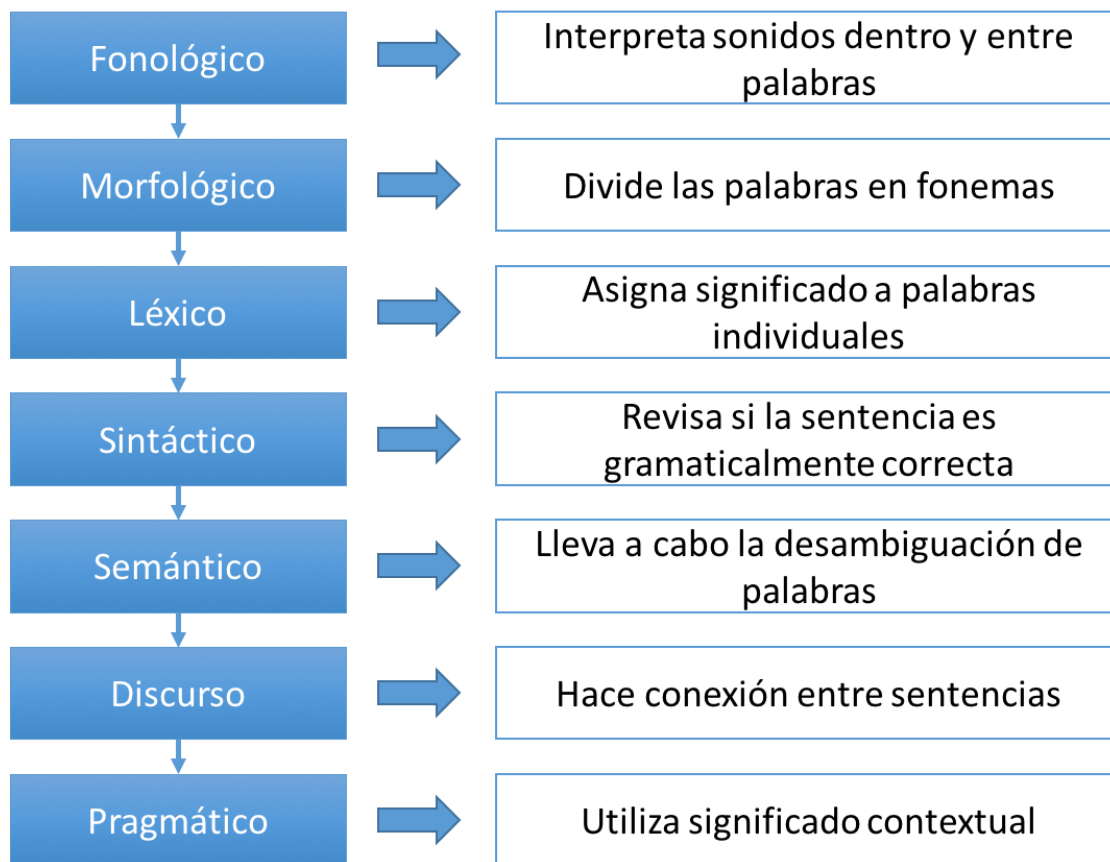


Figura 2.1. Niveles del procesamiento de lenguaje natural.

2.3.1.1. Nivel fonológico

Este nivel se ocupa de la interpretación de los sonidos del habla. Existen tres tipos de reglas usadas en el análisis fonológico: 1) Reglas fonéticas, utilizadas para los sonidos dentro de las palabras; 2) reglas de fonema, para las variaciones de pronunciación cuando las palabras se hablan juntas; y 3) reglas prosódicas, las cuales son utilizadas para la fluctuación del acento y entonación cuando se enuncia una sentencia. En los sistemas PLN que aceptan como entrada la voz humana, las onda sonoras se analizan y codifican en una señal digitalizada para su interpretación a través de diversas reglas o para comparar con otro modelo de lenguaje (Liddy, 2001).

2.3.1.2. Nivel morfológico

El nivel morfológico consiste en detectar la relación que se establece entre las unidades mínimas que forman una palabra, por ejemplo, sufijos o prefijos. Para

ello, los sistemas de procesamiento morfológico transforman cada secuencia de caracteres en una secuencia de morfema, a través de técnicas como *stemmers*, lematizadores, o *POS-taggers*. Los seres humanos pueden descomponer una palabra en morfemas para comprender su significado. De la misma manera, un sistema PLN puede reconocer el significado de cada morfema para obtener el significado total de la palabra (Rodríguez, 2014).

2.3.1.3. Nivel léxico

El nivel léxico interpreta el significado de las palabras individuales. Existen varios tipos de procesamiento que permiten el análisis léxico de la palabra. Una de las técnicas principales se basa en la asignación de etiquetas individuales a cada palabra del texto. Estas etiquetas determinan la categoría léxica de la palabra y en el caso de que una palabra pueda tener diferentes funciones dentro de la sentencia, será etiquetado con la categoría léxica más probable en función del contexto.

Por ejemplo, si se quiere analizar la oración “María compró unos zapatos” (ver Figura 2.2), el análisis léxico nos estaría diciendo que “compró” es un verbo en pasado, y que “María” es un nombre propio.

María	compró	unos	zapatos
maría	comprar	uno	zapato
NP00000	VMIS3S0	DI0MP0	NCMP000

Figura 2.2. Análisis léxico.

2.3.1.4. Nivel sintáctico

El análisis sintáctico se centra en analizar las palabras de una oración con el objetivo de determinar la estructura gramatical de una sentencia. Los resultados de este nivel es una representación de la estructura de la sentencia donde se muestran las relaciones de dependencia entre las palabras. Siguiendo con el ejemplo mostrado en la sección anterior, si se utiliza un análisis de dependencias se obtendría una estructura gramatical como se presenta en la Figura 2.3. En este análisis se están relacionando las palabras entre sí a través de flechas. Donde se puede obtener que la palabra “María” es el sujeto de “compró”, y que el objeto de

“compró” es “unos zapatos”. Por tanto, este análisis proporciona información necesaria para extraer semántica de las sentencias.

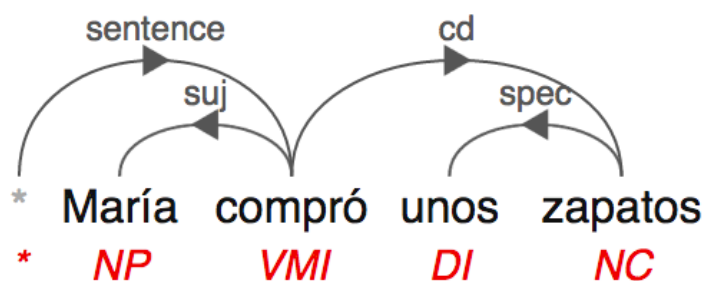


Figura 2.3. Análisis sintáctico.

2.3.1.5. Nivel semántico

El análisis semántico determina los posibles significados de una sentencia. El principal objetivo del procesamiento semántico es determinar los posibles significados de una oración centrándose en las interacciones entre los significados de nivel de palabra en la sentencia. Este nivel puede incluir la desambiguación semántica de palabras con múltiples sentidos. La desambiguación semántica permite que solo un sentido de palabras polisémicas sea seleccionado. Por ejemplo, la palabra “bolsa” tiene dos significados, artículo con asas para guardar cosas y una institución financiera. Actualmente, existen dos métodos para llevar a cabo la desambiguación: basado en corpus anotados y basados en fuentes de conocimiento. En el primer método cada palabra tiene su significado asociado basado en una fuente léxica por ejemplo, WordNet (Miller, 1995). El segundo método consiste en asignar sentidos a las palabras en base a reglas heurísticas que asocian a la palabra que se quiere desambiguar con sus descripciones (Martí, 2015).

2.3.1.6. Nivel de discurso o contextual

A diferencia del análisis sintáctico y semántico que trabajan con unidades de longitud de sentencia, el análisis de discurso trabaja con unidades de texto más largas que una sentencia, es decir, no interpreta los textos como conjuntos independientes de sentencias concatenadas, sino que analiza las propiedades del texto como un todo que transmite el significado haciendo conexiones entre las

sentencias que lo componen. Existen dos tipos de procesamiento del discurso que se pueden dar en este nivel, la resolución anafórica y el reconocimiento de la estructura del discurso. El primero consiste en reemplazar palabras vacías semánticamente como pronombres, por la entidad apropiada a la que se hace referencia. El segundo, determina las funciones de las sentencias en el texto y añade representación significativa al texto (Liddy, 2001).

2.3.1.7. Nivel pragmático

El análisis pragmático es uno de los niveles más complejos donde se analiza como las sentencias son usadas en distintas situaciones y de cómo el uso afecta al significado de las sentencias. El análisis que se realiza en este nivel se encuentra relacionado con los factores extralingüísticos que condicionan el uso del lenguaje en situaciones concretas. Algunas aplicaciones PLN utilizan bases de conocimiento y módulos de inferencia que permiten interpretar las intenciones, los planes y los objetivos de un texto (Rodríguez, 2014).

2.4. Análisis de sentimientos

El término “análisis de sentimientos” ha sido definido por diversos autores, sin embargo en la comunidad investigadora, los más utilizados son los propuestos por (Liu, 2012) y (Cambria & Hussain, 2012).

(Liu, 2012) define el análisis de sentimientos como: *“El campo de estudio que analiza las opiniones, sentimientos, evaluaciones, valoraciones, actitudes y emociones de las personas hacia entidades tales como productos, servicios, organizaciones, individuos, cuestiones, eventos, temas y sus atributos”*.

(Cambria & Hussain, 2012) lo define como *“Conjunto de técnicas computacionales para la extracción, clasificación, comprensión y evaluación de opiniones expresadas en fuentes publicadas en Internet, comentarios en portales web y en otros contenidos generados por usuarios”*.

El análisis de sentimientos, también conocido como minería de opiniones, es el área que a través de técnicas de PLN y análisis de texto identifica y extrae opiniones y sentimientos desde diversas fuentes de información. Para lograr dicho objetivo, trabajos en esta área se han basado en dos principales enfoques y tres

Capítulo 2. Estado del arte

principales niveles de análisis de la opinión (ver Figura 2.4). En las siguientes secciones se da una explicación detallada de los enfoques y niveles de análisis de sentimientos.

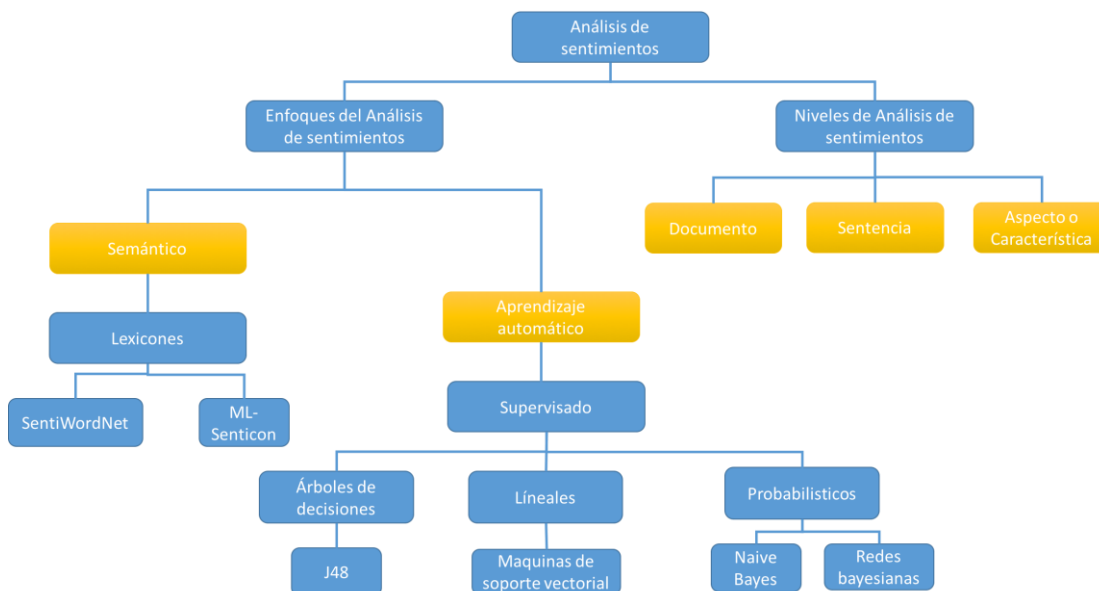


Figura 2.4. Análisis de sentimientos.

2.4.1. Evolución histórica del análisis de sentimientos

El análisis de opiniones o minería de opiniones ha sido ampliamente estudiado en los últimos años y se ha convertido en un tema que ha atraído la atención de la comunidad investigadora del procesamiento de lenguaje natural. Sin embargo, sus inicios surgen en el siglo XX, donde se publicaron algunos trabajos que se consideran como los precursores del área. (Carbonell, 1979) propone un modelo computacional para representar el pensamiento subjetivo de las personas, el cual clasificó en dos tipos de opinión: liberal o conservador. Este trabajo está basado en el dominio político, el cual fue elegido debido a que se presenta la subjetividad de las personas. (Wilks & Bien, 1983) presentaron un modelo para interpretar el lenguaje natural y tratar la subjetividad de las personas, los autores llevan a cabo un análisis para entender el concepto metafórico y las relaciones entre sentencias. Fue de los primeros trabajos en donde se daba a conocer el concepto de “sentidos de las palabras (en inglés, *word senses*)”, donde una palabra puede tener diferentes significados dependiendo del dominio o contexto. En (Wiebe & Rapaport, 1988) introducen un proceso que reconoce los pensamientos y las percepciones de los personajes en las narrativas en tercera persona. Su enfoque es capaz de identificar

las sentencias subjetivas y a los personajes a los que se atribuye. En (Wiebe, 1990) el algoritmo fue mejorado para extraer palabras subjetivas e identificar sus personajes en el texto narrativo de ficción. Los autores construyen un conjunto de reglas basados en el análisis de pasajes de más de cuarenta novelas y cuentos. Trabajos posteriores se centraron principalmente en la interpretación de la metáfora, narrativa, puntos de vista, afectos, evidencia en el texto y áreas relacionadas.

A principios del siglo XXI inicia una amplia investigación relacionada con el análisis de sentimientos, donde diferentes técnicas son propuestas con el objetivo de procesar automáticamente la opinión de las personas. La primera vez que fue utilizado el término *Opinion Mining* fue en (Dave, Lawrence, & Pennock, 2003). De acuerdo a los autores, una herramienta de minería de opinión debería procesar un conjunto de resultados de búsqueda para un elemento determinado, generando una lista de atributos del producto y agregando opinión sobre cada uno de ellos. Mientras tanto, *Sentiment Analysis* fue utilizado por primera vez en (Nasukawa & Yi, 2003). Los autores presentaron un enfoque para extraer sentimientos asociados con polaridades positivas o negativas para específicas entidades.

Posteriormente desde el año 2004 al año 2010, los investigadores se centraron en diferentes aspectos del análisis de sentimientos, tales como técnicas de clasificación y extracción, aplicación a diferentes dominios, construcción de recursos léxicos, entre otros. Por ejemplo, en (Mullen & Collier, 2004) introducen un enfoque para clasificar los textos como positivos o negativos utilizando SVM (*Support Vector Machine*, máquinas de soporte vectorial). En (T. Wilson, Wiebe, & Hoffmann, 2005) presentan un nuevo enfoque para el análisis del sentimiento a nivel de frase que determina primero si una expresión es neutra o polar y luego desambigua la polaridad de las expresiones polares. Con este enfoque, el sistema es capaz de identificar automáticamente la polaridad contextual para un gran subconjunto de expresiones de sentimiento. Mientras tanto, en (Melville, Gryc, & Lawrence, 2009) desarrollan un marco de trabajo para incorporar el conocimiento léxico en el aprendizaje supervisado para la categorización del texto. Además, los autores aplican con éxito el enfoque desarrollado a la tarea de la clasificación de sentimientos, ampliando el estado del arte en el campo, que se ha centrado

Capítulo 2. Estado del arte

principalmente en el aprendizaje supervisado de forma aislada. (Baccianella, Esuli, & Sebastiani, 2010) presentan SentiWordNet 3.0, un recurso léxico mejorado explícitamente diseñado para apoyar la clasificación de sentimientos y aplicaciones de minería de opinión.

En años más recientes, una gran cantidad de trabajos de análisis de sentimientos han surgido aplicando técnicas sofisticadas con el objetivo de mejorar la precisión de la clasificación a diferentes niveles de granularidad. Estos enfoques se han centrado en diferentes dominios. Sin embargo, una especial atención ha surgido por las redes sociales donde las personas suelen escribir sobre sus experiencias y opiniones acerca de cualquier tema, entre los enfoques que destacan se encuentran los presentados en (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011), (Kouloumpis, Wilson, & Moore, 2011), (Rosenthal, Ritter, Nakov, & Stoyanov, 2014) y (Nakov, Rosenthal, Sebastiani, & Stoyanov, 2016).

Como conclusión, el año 2001 marcó el comienzo de problemas de investigación relacionados al análisis de sentimientos y posteriormente se han publicado cientos de artículos sobre el tema. Según lo publicado en (Pang & Lee, 2008), existen tres factores determinantes que originaron el surgimiento de nuevos enfoques en esta área.

- El surgimiento de métodos de aprendizaje automático en el PLN y la recuperación de la información.
- La disponibilidad de conjuntos de datos para entrenar algoritmos de aprendizaje automático. El surgimiento de plataformas donde los usuarios podían compartir sus opiniones contribuyó a la obtención de esos conjuntos de datos.
- La realización de desafíos intelectuales y aplicaciones comerciales e inteligentes que el área ofrece.

2.4.2. Niveles de análisis de sentimientos

El análisis de sentimientos se puede realizar a diferentes niveles de granularidad, es decir, a nivel de documento, a nivel de frase o sentencia y a nivel de aspecto o

característica (Ver Figura 2.5). A continuación, cada uno de dichos niveles se explica brevemente.

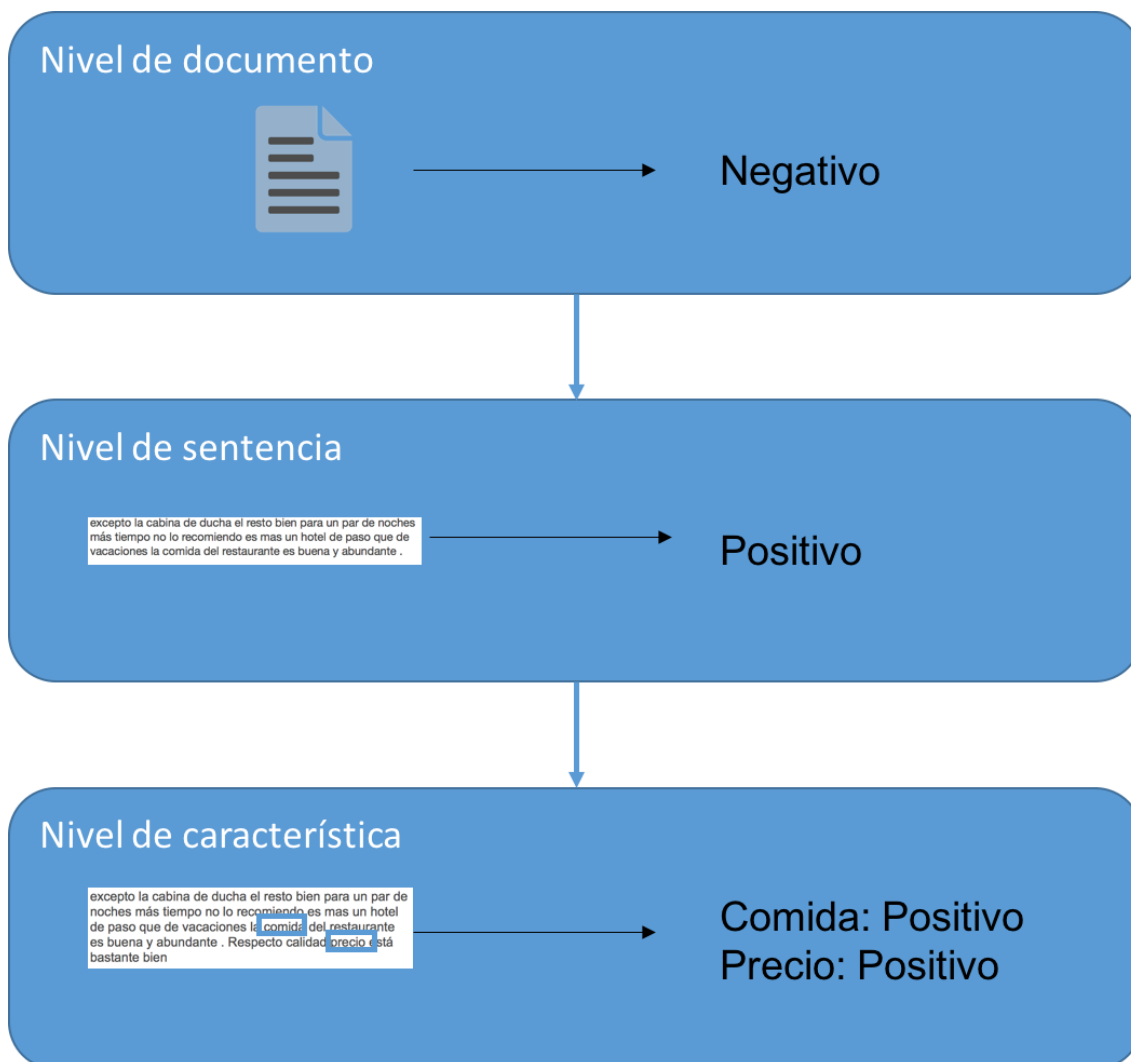


Figura 2.5. Niveles del Análisis de sentimientos.

2.4.2.1. Análisis de sentimientos a nivel de documento

El análisis de sentimiento a este nivel considera todo el documento para llevar a cabo la detección de la polaridad. Este nivel es comúnmente conocido como “clasificación de sentimientos a nivel de documento”. El objetivo de este análisis es determinar si un documento de opinión completo expresa un sentimiento positivo o negativo (Pang, Lee, & Vaithyanathan, 2002) (Turney, 2002). Por ejemplo, si un usuario provee la opinión de un producto, el sistema de análisis de sentimiento determina si la opinión general del usuario expresa una opinión positiva o negativa. Este nivel de análisis asume que cada documento contiene opiniones sobre una sola entidad. Por ejemplo, un producto. Por lo que no es aplicable a

documentos donde se evalúa o comparan varias entidades. Algunos ejemplos de enfoques que llevan a cabo el análisis a nivel de documento son los presentados en (Pang et al., 2002), (Turney, 2002) y (Moraes, Valiati, & Gavião Neto, 2013)

2.4.2.2. Análisis de sentimientos a nivel de sentencia

El análisis de sentimientos a este nivel determina si cada sentencia expresa una opinión positiva, negativa, o neutral (Comúnmente significa que no se ha dado ninguna opinión). Este nivel de análisis se relaciona con la clasificación de la subjetividad (Wiebe, Bruce, & O'Hara, 1999), el cual consiste en distinguir sentencias objetivas que proporcionan información fáctica de sentencias subjetivas que expresan puntos de vista y opiniones. Sin embargo, es importante aclarar que la subjetividad no es equivalente a sentimientos, ya que muchas sentencias objetivas pueden implicar opiniones. Algunos trabajos como los de (T. Wilson et al., 2005), (Meena & Prabhakar, 2007), (Fu & Wang, 2010) y (Zhang & He, 2015) tratan de extraer la polaridad de cada oración. Además, otros enfoques tales como el de (T. Wilson, Wiebe, & Hwa, 2004), han profundizado a nivel de sintagmas y obtienen la polaridad de una sentencia como la combinación de las polaridades de los sintagmas que la componen (Liu, 2012).

2.4.2.3. Análisis de sentimientos a nivel de características

El análisis a nivel de documento también llamado análisis a nivel de característica lleva a cabo un análisis de grano fino (Hu & Liu, 2004). La diferencia entre este nivel y los anteriores es que con este nivel se puede determinar qué es exactamente lo que a las personas les gusta o no les gusta. Por lo que el objetivo de este nivel de análisis es descubrir sentimientos sobre entidades y/o sus aspectos. Por ejemplo, en la frase “El restaurante es caro, pero el menú es muy bueno”, se evalúan dos aspectos del restaurante, el precio y la comida. En cuanto al primer aspecto tiene un sentimiento negativo, y el segundo aspecto un sentimiento positivo (Liu, 2012). Un desafío en este análisis, es con respecto a los tipos de opiniones (opiniones regulares y opiniones comparativas) (Jindal & Liu, 2006). Una opinión regular expresa una opinión sobre una entidad en particular. Por ejemplo, “la cámara del iPhone 7 es buena”, donde se expresa un sentimiento positivo sobre el aspecto cámara del iPhone 7. Por otro lado, una opinión

comparativa se basa en comparar múltiples entidades con respecto a algunos de los aspectos que comparten, por ejemplo, “la cámara del iPhone 7 es mejor que la del Galaxy S7”, en el cual se compara el iPhone 7 y el Galaxy S7.

2.4.3. Recursos lingüísticos para el análisis de sentimientos

2.4.3.1. Corpus

Hoy en día existen diversos corpus disponibles para la comunidad investigadora, con el objetivo de llevar a cabo experimentos y evaluar los enfoques de análisis de sentimientos. No obstante, varios trabajos en la literatura llevan a cabo la colección de documentos de opinión, los cuales casi siempre son etiquetados manualmente por expertos. Sin embargo, esta tarea requiere mucho esfuerzo y tiempo. Por esta razón, es de suma importancia para la comunidad investigadora, tener conocimiento de los corpora que están actualmente disponibles, ya que esto además permite llevar a cabo una correcta comparación y medir el rendimiento de un nuevo enfoque de análisis de sentimientos. A continuación, se presentan algunos corpora de opiniones disponibles.

- **Corpus MPQA:** Este corpus contiene artículos de noticias desde una gran variedad de fuentes desde junio del 2001 hasta mayo del 2002, cada artículo fue manualmente etiquetado, asociando etiquetas de opinión, y otros estados personales como creencias, emociones, sentimientos, especulaciones, entre otros. La primera versión del corpus fue presentado en (J. Wiebe, Wilson, & Cardie, 2005), el cual contiene 535 documentos, un total de 11114 sentencias. La segunda versión contiene 157 documentos adicionales y dos nuevos tipos de anotaciones (actitudes y objetivos) (T. Wilson, 2008). Finalmente, la tercera versión del corpus fue extendida incluyendo un nuevo tipo de anotación a entidades específicas y eventos (L. Deng & Wiebe, 2015).
- **Corpus MCE:** Es la versión en inglés traducida del corpus MuchoCine. El corpus contiene 3878 opiniones, las cuales han sido clasificadas en una escala desde 1 (muy negativo) hasta 5 (muy positivo) (Martín-Valdivia, Martínez-Cámara, Perea-Ortega, & Ureña-López, 2013).

- **Corpus OCA:** Corpus en árabe sobre comentarios de películas. El corpus ha sido generado a partir de diferentes páginas Web de películas tales como Cinema, Al Rasid, Film Reader, Elcinema, Grind House, Mzyondubai, Aflamee, Grind Film, Cinema Gate, Emad Ozery Blog, entre otras (Rushdi Saleh, Martín-Valdivia, Montejo-Ráez, & Ureña-López, 2011). El corpus contiene 500 opiniones, de las cuales 250 son consideradas positivas y 250 negativas.
- **Corpus Criticas Cine:** Corpus en español sobre críticas de cine, el cual ha sido extraído desde la Web de muchocine¹. El corpus contiene un total de 3878 críticas y aproximadamente 2 millones de palabras, con una media de 546 palabras por crítica. Cada crítica está clasificada con una puntuación de 1 a 5 donde 1 es muy negativo y 5 muy positivo (Cruz Mata, Troyano Jiménez, Enríquez de Salamanca Ros, & Ortega Rodríguez, 2008). A continuación, en la Tabla 2.1. se muestra el número de documentos por cada polaridad establecida.

Tabla 2.1. Número de documentos por polaridad del corpus de críticas de cine.

Polaridad	Número de documentos
1	351
2	923
3	1253
4	890
5	461

2.4.3.2. Bases de conocimiento

Las bases de conocimiento al igual que los corpora son un recurso importante en el análisis de sentimientos. Estas consisten desde una simple lista de palabras de opinión a bases de conocimiento léxicas. A continuación, se describe algunas bases de conocimiento, las cuales son comúnmente aplicadas en la literatura.

¹ muchocine.net

2.4.3.2.1. eSOL

Es una lista de palabras indicadoras de opinión en español dependientes del dominio de críticas de cine. Su elaboración siguió un enfoque basado en corpus. El corpus utilizado para este objetivo fue el *Spanish Movie Reviews*. La lista está formada por 2535 palabras positivas y 5639 palabras negativas (Molina-González, Martínez-Cámara, Martín-Valdivia, & Perea-Ortega, 2013). A continuación, la Tabla 2.2. muestra un ejemplo de palabras positivas y palabras negativas de eSOL.

Tabla 2.2. Ejemplo de palabras positivas y negativas de eSOL.

Positivas	Negativas
gran	menos
buen	fallida
imprescindible	previsible
original	falta

2.4.3.2.2. General Inquierer

Es un léxico que proporciona información sintáctica, semántica y pragmática a cada palabra etiquetada morfológicamente. Fue desarrollado por el departamento de ciencias sociales en Harvard. El diccionario contiene 1915 palabras positivas y 2291 palabras negativas (Philip J., Dexter C., & Marshall S., 1966). A continuación, en la Tabla 2.3. se presentan algunos ejemplos de palabras positivas y negativas de General Inquierer.

Tabla 2.3. Ejemplo de palabras positivas y negativas de General Inquierer.

Positivas	Negativas
Abundant	Abnormal
Accept	Addict
Advantage	Aggressor
Affirmative	Angry

2.4.3.2.3. iSOL

Al igual que eSOL, iSOL es una lista de palabras indicadoras de opinión en español, sin embargo, esta es independiente del dominio. La elaboración del recurso partió de la lista de palabras propuestas por el profesor Bing Liu (Liu, Hu, & Cheng, 2005). La lista de palabras fue traducida automáticamente usando el traductor Reverso y posteriormente corregido manualmente. La lista consta de un total de 2509 palabras positivas y 5626 negativas. La Tabla 2.4. muestra algunos ejemplos de palabras contenidas en iSOL.

Tabla 2.4. Palabras positivas y negativas de iSOL.

Positivas	Negativas
abiertamente	añicos
abrazo	abandonada
abrazos	abandonadas
absorbente	abandonado

2.4.3.2.4. Léxico de opiniones de Bing Liu

Es un léxico de opiniones en inglés, el cual consiste de 2006 palabras positivas, y 4783 palabras negativas. Este recurso incluye errores de ortografía, variaciones morfológicas, jerga, y contenido frecuente de redes sociales (Hu & Liu, 2004)(Liu et al., 2005). La Tabla 2.5. muestra algunos ejemplos de palabras contenidas en el léxico.

Tabla 2.5. Palabras positivas y negativas del léxico de opiniones de Bing Liu.

Positivas	Negativas
a+	2-faced
abound	2-faces
abounds	abnormal
abundance	abolish

2.4.3.2.5. ML-SentiCon

Este recurso contiene lexicones de sentimientos a nivel de lema para el inglés, español, catalán, vasco y gallego. Para cada lema se provee una estimación de polaridad de muy negativo (-1.0) hasta muy positivo (+1.0) y una desviación estándar relacionada con la ambigüedad de la estimación de la polaridad (Cruz, Troyano, Pontes, & Ortega, 2014). La Figura 2.6 muestra un extracto del lexicón ML-Senticon en español.

```
<?xml version="1.0" encoding="UTF-8" ?>
<senticon lang="es">
  <layer level="1">
    <positive>
      <lemma pos="a" pol="0.708" std="0.149"> acertado </lemma>
      <lemma pos="a" pol="0.906" std="0.125"> admirable </lemma>
      <lemma pos="n" pol="0.45" std="0.331"> admiración </lemma>
      <lemma pos="v" pol="0.75" std="0.177"> admirar </lemma>
      <lemma pos="a" pol="0.375" std="0.0"> afectivo </lemma>
      <lemma pos="n" pol="0.321" std="0.112"> afecto </lemma>
      <lemma pos="a" pol="0.563" std="0.131"> afectuoso </lemma>
      <lemma pos="n" pol="0.5" std="0.425"> afición </lemma>
      <lemma pos="a" pol="0.813" std="0.063"> afortunado </lemma>
      <lemma pos="a" pol="0.75" std="0.181"> agradable </lemma>
      <lemma pos="a" pol="0.5" std="0.125"> agradecido </lemma>
      <lemma pos="v" pol="0.458" std="0.315"> alegrar </lemma>
```

Figura 2.6. ML-Senticon en español.

2.4.3.2.6. MPQA (Multi-Perspective Question Answering)

Este léxico es parte de OpinionFinder, el cual contiene una lista de palabras y expresiones indicadoras de subjetividad. Las expresiones fueron coleccionadas desde un conjunto de fuentes. La mayoría de expresiones se obtuvieron desde el trabajo reportado en (Riloff & Wiebe, 2003). En la Figura 2.7 se muestra un extracto del léxico. Cada una de las expresiones está acompañada por su nivel de subjetividad (fuerte o débil), longitud en palabras, categoría morfológica, y una polaridad, positiva, negativa o neutra (T. Wilson et al., 2005).

```
type=weaksubj len=1 word1=abandoned pos1=adj stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=abandonment pos1=noun stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=abandon pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abase pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abasement pos1=anypos stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abash pos1=verb stemmed1=y priorpolarity=negative
type=weaksubj len=1 word1=abate pos1=verb stemmed1=y priorpolarity=negative
type=weaksubj len=1 word1=abdicate pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=aberration pos1=adj stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=abhor pos1=anypos stemmed1=y priorpolarity=negative
```

Figura 2.7. Extracto de MPQA

2.4.3.2.7. SentiWordNet

Es un recurso léxico construido a partir de WordNet, el cual asocia tres valores numéricos a cada *synset* de WordNet: positivo, negativo y objetivo. Los tres puntajes van desde 0 hasta 1 y la suma de los tres es igual a 1. Los diferentes sentidos de una palabra pueden tener diferentes propiedades relacionadas con la opinión. La estructura del diccionario es la siguiente: POS (*part of speech*), ID (id del *synset* correspondiente a WordNet), PosScore (puntuación positiva), NegScore (puntuación negativa), SynsetTerms (representa el *synset* y el número del sentido de la palabra que pertenece a el *synset*), Gloss (una definición sobre el término, y una frase de ejemplo usando el término) (Esuli & Sebastiani, 2006). Aunque no se muestra la puntuación neutral de un *synset*, se puede deducir mediante la ecuación (2.1)

$$NeuScore = 1 - (PosScore + NegScore) \quad (2.1)$$

Donde *PosScore* es la puntuación positiva, y *NegScore* es la puntuación negativa contenidas en el lexicón.

La Figura 2.8 muestra un extracto del léxico SentiWordNet.

#	POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a		00001740	0.125	0	able#1	(usually followed by `to') having the necessary means or skill or know-how or authority to do something; "able to swim"; "she was able to program her computer"; "we were at last able to buy a car"; "able to get a grant for the project"
a		00002098	0	0.75	unable#1	(usually followed by `to') not having the necessary means or skill or know-how; "unable to get to town without a car"; "unable to obtain funds"
a		00002312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ or organism; "the abaxial surface of a leaf is the underside or side facing away from the stem"
a		00002527	0	0	ventral#2 adaxial#1	nearest to or facing toward the axis of an organ or organism; "the upper side of a leaf is known as the adaxial surface"
a		00002730	0	0	acrosopic#1	facing or on the side toward the apex
a		00002843	0	0	basisopic#1	facing or on the side toward the base
a		00002956	0	0	abducting#1 abducent#1	especially of muscles; drawing away from the midline of the body or from an adjacent part
a		00003131	0	0	adductive#1 adducting#1 adducent#1	especially of muscles; bringing together or drawing toward the midline of the body or toward an adjacent part
a		00003356	0	0	nascent#1	being born or beginning; "the nascent chicks"; "a nascent insurgency"

Figura 2.8. SentiWordNet.

2.4.3.2.8. Spanish Sentiment Lexicons

Este recurso provee dos lexicones para el español. El formato de ambos lexicones tiene la siguiente estructura: palabra en español, posición del *synset* en WordNet 1.6, anotación en español y anotación en inglés. La anotación en inglés se corresponde a la anotación generada automáticamente por los lexicones OpinionFider y SentiWordNet. La anotación en español, es la anotación manual de sentimientos llevada a cabo por dos hablantes nativos de español (esta anotación

es disponible solo para las primeras 100 palabras en cada lexicón) (Perez-Rosas, Banea, & Rada, 2012). El primer lexicón es nombrado fullStrengthLexicon (ver Figura 2.9), el cual es más robusto y aprovecha las anotaciones manuales de sentimientos hechas para el lexicón OpinionFinder (J. Wiebe et al., 2005).

abundancia	5119096	neg	pos	
abundante	13887	pos	pos	
adecuado	51373	pos	pos	
admirable	1807605	pos	pos	
adorar	1777817	pos	pos	
agitacion	7513508	neg	neg	
agotamiento	14016361		neg	neg
agradable	89051	pos	pos	
agudo	1214430	neg	pos	
alerta	4664778	pos	neg	
amenaza	7226015	neg	neg	
apreciado	1462625	pos	pos	
apropiado	51373	pos	pos	
aviso	7226015	neg	pos	
belleza	10613996	pos	pos	

Figura 2.9. Extracto de fullStrengthLexicon.

El segundo lexicón mediumStrengthLexicon (ver Figura 2.10), aprovecha las anotaciones de sentimiento realizadas para SentiWordNet (Esuli & Sebastiani, 2006).

accesible	498892	pos	pos	
acumulacion	9625489	neg	neg	
adaptar	1843927	pos	pos	
aliviar	1204852	pos	pos	
aproximado	869222	neg	neg	
autoritario	678397	neg	neg	
basico	856572	pos	neg	
bondad	3800378	pos	pos	
calibre	3717355	pos	neg	
caligrafia	4826894	pos	pos	
cambio	3733220	pos	pos	
centenario	1585840	pos	pos	
cerrado	282089	neg	neg	
complejo	2093020	neg	neg	
convencional	1630148	pos	pos	

Figura 2.10. Extracto de mediumStrengthLexicon.

2.4.4. Enfoques para el análisis de sentimientos

Con el objetivo de crear sistemas automáticos que lleven a cabo un análisis de sentimientos efectivo, varios autores han basado sus trabajos en dos principales enfoques: enfoque basado en aprendizaje automático y enfoque semántico (Peñalver-Martinez et al., 2014). Los métodos basados en aprendizaje automático supervisados requieren de un conjunto de características para posteriormente entrenar un algoritmo de clasificación. El método n-gramas es frecuentemente

Capítulo 2. Estado del arte

utilizado para la extracción de características (Arafat, Elawady, Barakat, & Elrashidy, 2014) (Moraes et al., 2013), aunque existen algunos otros métodos tales como frecuencia de términos, características basadas en su etiqueta POS, entre otros. Sin embargo, la mayoría de métodos proveen un vector de características de alta dimensión, lo cual reduce la efectividad de este enfoque. Para abordar este problema es importante utilizar un método de selección de características que permita eliminar el ruido y características irrelevantes, lo que mejora la precisión de la clasificación (Gelbukh, 2013). Entre los métodos de selección de características comúnmente utilizados en un enfoque supervisado son IG (*Information Gain*, ganancia de información) y Mmrm (*minimun redundancy maximun relevance*, mínima redundancia máxima relevancia) (Arafat et al., 2014) (Moraes et al., 2013) (Habernal, Ptáček, & Steinberger, 2014). Por otro lado, los métodos basados en un enfoque semántico hacen uso de diccionarios de palabras conocidos como lexicones, donde las palabras que denotan sentimiento en una opinión son comparadas con las contenidas en el diccionario con el objetivo de obtener su valor de polaridad. Entre los lexicones más utilizados en la literatura se encuentran SentiWordNet (Baccianella et al., 2010) y WordNet-affect (Valitutti, 2004). En los siguientes apartados se realizará una descripción detallada de cada uno de los enfoques para el análisis de sentimientos. A continuación, la Figura 2.11 presenta una descripción grafica de los principales enfoques del análisis de sentimientos.

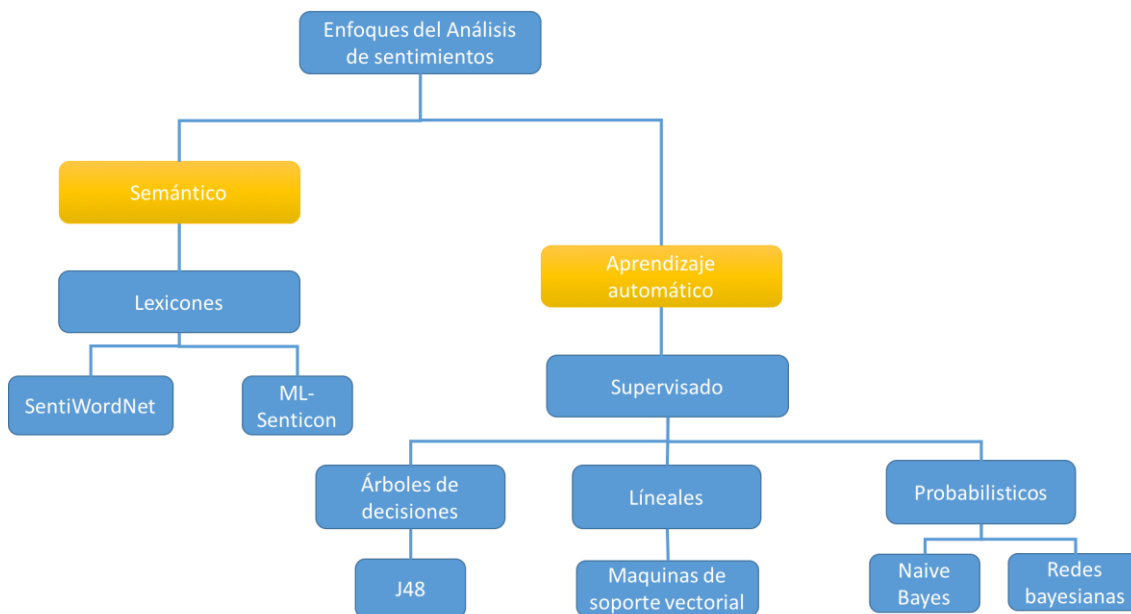


Figura 2.11. Enfoques del análisis de sentimientos.

2.4.4.1. Enfoque basado en aprendizaje automático

El enfoque basado en aprendizaje automático consiste en entrenar un algoritmo de clasificación supervisado a partir de una colección de datos etiquetados. La Figura 2.12 muestra el proceso del aprendizaje automático. Este enfoque consiste de dos principales pasos. El primer paso es la extracción de características desde un conjunto de datos de entrenamiento, las cuales se representa como un vector de características. Las características son extraídas a través de diversos métodos como bolsa de palabras, n-gramas, TF-IDF (*Term frequency - Inverse document frequency*, frecuencia de término - frecuencia inversa de documento) (Salton & Buckley, 1988) , entre otros. Estos métodos son explicados en detalle en la sección 2.4.4.2.2. El segundo paso es el entrenamiento del algoritmo de clasificación a través de los vectores de características. Comúnmente se requiere de dos conjuntos de datos uno de entrenamiento y otro para pruebas. El primer conjunto es utilizado para que el algoritmo “aprenda” de las diversas características de los documentos y el segundo conjunto sirve para evaluar el rendimiento del modelo obtenido. Dicho modelo, permite clasificar nuevos documentos, en este caso como positivos o negativos. Entre los algoritmos de aprendizaje más utilizados para la clasificación de opiniones se encuentran la máquina de soporte vectorial (SVM),

redes bayesianas (Naive Bayes, BayesNet), árboles de decisión (J48) y K-vecinos más cercanos (KNN).

En este enfoque el método para la extracción de características y el algoritmo de clasificación seleccionados juegan un rol importante en la precisión de la clasificación de la opinión. Es decir, la elección correcta del algoritmo de aprendizaje y la correcta determinación de las características que representan los ejemplos es más que fundamental para obtener un buen resultado en la clasificación (Martínez Cámara, 2016).

Uno de los primeros trabajos de análisis de sentimientos en usar el enfoque de aprendizaje automático es el de (Pang et al., 2002), ellos llevan a cabo experimentos aplicando varios algoritmos de clasificación supervisados tales como el clasificador Bayesiano (*Naive Bayes*), máxima entropía, y máquinas de soporte vectorial en el dominio de opiniones de películas.

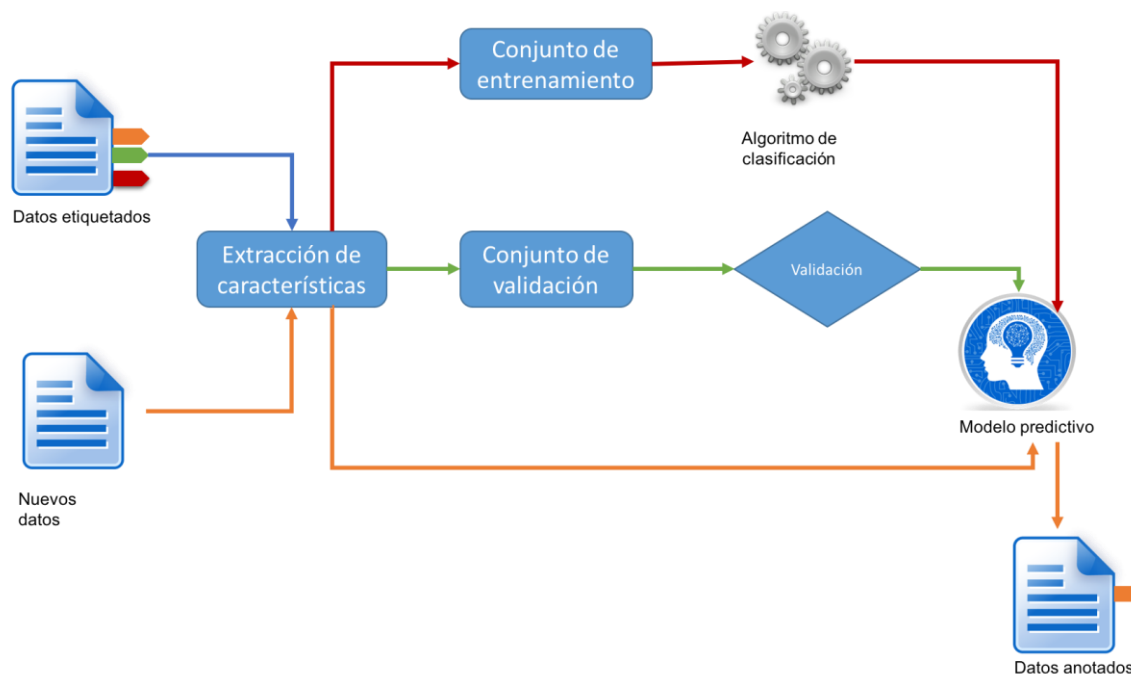


Figura 2.12. Proceso de aprendizaje automático.

2.4.4.1.1. Extracción de características

Los enfoques basados en máquinas de aprendizaje han sido ampliamente utilizados en el análisis de sentimientos. Para mejorar el rendimiento de la clasificación, es muy importante construir un vector que contenga importantes características para el análisis de sentimientos. Entre los métodos comúnmente

utilizados en la literatura se encuentran N-gramas, TF-IDF, características basadas en dependencias sintácticas y características basadas en etiquetas POS. A continuación, se describe cada uno de estos métodos de extracción de características.

2.4.4.1.1.1. Términos y su frecuencia

Este tipo de características son palabras individuales o palabras n-gramas (bigramas y trigramas) y su frecuencia. El método unigramas obtiene las características eliminando espacios adicionales y caracteres ruidosos entre dos palabras. Por ejemplo, en la frase “El móvil tiene excelentes características”, se obtendrían las características “El”, “móvil”, “tiene”, “excelentes”, “características”. El método bigramas obtiene las características a través de la extracción de dos palabras consecutivas en el texto. Por ejemplo, en la frase “Los actores son muy malos”. Las características bigramas obtenidas son: “Los_actores”, “actores_son”, “son_muy”, y “muy_malos”.

El esquema de ponderación TF-IDF también es ampliamente utilizado. Esta es una medida numérica que representa que tan relevante es una palabra para un documento en una colección. El valor TF-IDF aumenta proporcionalmente al número de veces que aparece una palabra en un documento, pero es compensada por la frecuencia de la palabra en el conjunto de documentos, lo que permite determinar las palabras que son más comunes que otras. Por lo que, TF representa la importancia de un término en un documento, e IDF representa la importancia del término en la colección de documentos (Kaleel & Abhari, 2015). Dicho esto, la frecuencia del termino TF se puede definir como:

$$tf_{ij} = \frac{n_{ij}}{N} \quad (2.2)$$

Donde n_{ij} es el número de veces que la palabra i aparece en el documento j y N es el numero total de palabras en el documento j .

$$N = \sum_k n_{kj} \quad (2.3)$$

Capítulo 2. Estado del arte

La frecuencia inversa del documento *IDF* se calcula con la siguiente ecuación. Donde d_i es el número de documentos que contiene la palabra i y D es el número total de documentos.

$$idf_i = \log\left(\frac{D}{d_i}\right) \quad (2.4)$$

Por tanto, el valor TF-IDF para una palabra w en un documento d es calculado por la siguiente ecuación.

$$tf - idf = tf_{ij} * idf_i \quad (2.5)$$

2.4.4.1.1.2. Características basadas en etiquetas Part of Speech (POS)

La información POS es comúnmente utilizado en el análisis de sentimientos. Los adjetivos han sido ampliamente usados con características por un gran número de investigadores (Mullen & Collier, 2004) (Whitelaw, Garg, & Argamon, 2005). Trabajos basados en la detección de la subjetividad demuestran una alta relación entre la presencia de adjetivos y sentencias subjetivas (Hatzivassiloglou & Wiebe, 2000). Este hallazgo ha sido tomado como evidencia de que algunos adjetivos son buenos indicadores de opinión y han sido usados para la selección de características para la clasificación de sentimientos (Pang & Lee, 2008). Algunos trabajos tales como el de (Turney, 2002), en lugar de centrarse únicamente en adjetivos propone seleccionar frases en base a una serie de patrones, la mayoría incluyendo un adjetivo o un adverbio. Tabla 2.6. muestra algunos ejemplos de patrones para la extracción de frases, donde la etiqueta JJ indica adjetivos, NN sustantivos, RB adverbios y VB verbos. El segundo patrón, por ejemplo, indica que dos palabras consecutivas son extraídas si la primera palabra es un adverbio, la segunda palabra es un adjetivo, pero la tercera palabra la cual no es extraída, no puede ser un sustantivo.

Tabla 2.6. Patrones para la extracción de frases.

	Primera palabra	Segunda palabra	Tercera palabra (no extraída)
1	JJ	NN or NNS	Ninguno
2	RB, RBR, o RBS	JJ	No NN ni NNS
3	JJ	JJ	No NN ni NNS
4	NN o NNS	JJ	No NN ni NNS
5	RB, RBR, o RBS	VB, VBD, VBN, o VBG	Ninguno

2.4.4.1.1.3. Características de dependencia sintáctica

Las características basadas en dependencia de palabras obtenidas a partir de árboles de análisis de dependencia, se han utilizado en el análisis de sentimientos por varios investigadores. Los estudios han demostrado que las características basadas en árboles de dependencia superan a la bolsa de palabras (en inglés, *bag-of-words*). Stanford NLP (Manning et al., 2014) y FreeLing (Padró, 2012) son algunas de las herramientas más utilizadas para la obtención de los árboles de dependencia. La Figura 2.13 muestra un ejemplo del árbol de dependencia obtenido con FreeLing.

Algunos enfoques establecen reglas para la selección de características de dependencia enriquecidas con sentimiento. Es decir, se lleva a cabo una selección de características sintácticas a partir de un conjunto de reglas, lo cual permite obtener únicamente las relaciones que pueden contribuir a la detección del sentimiento.

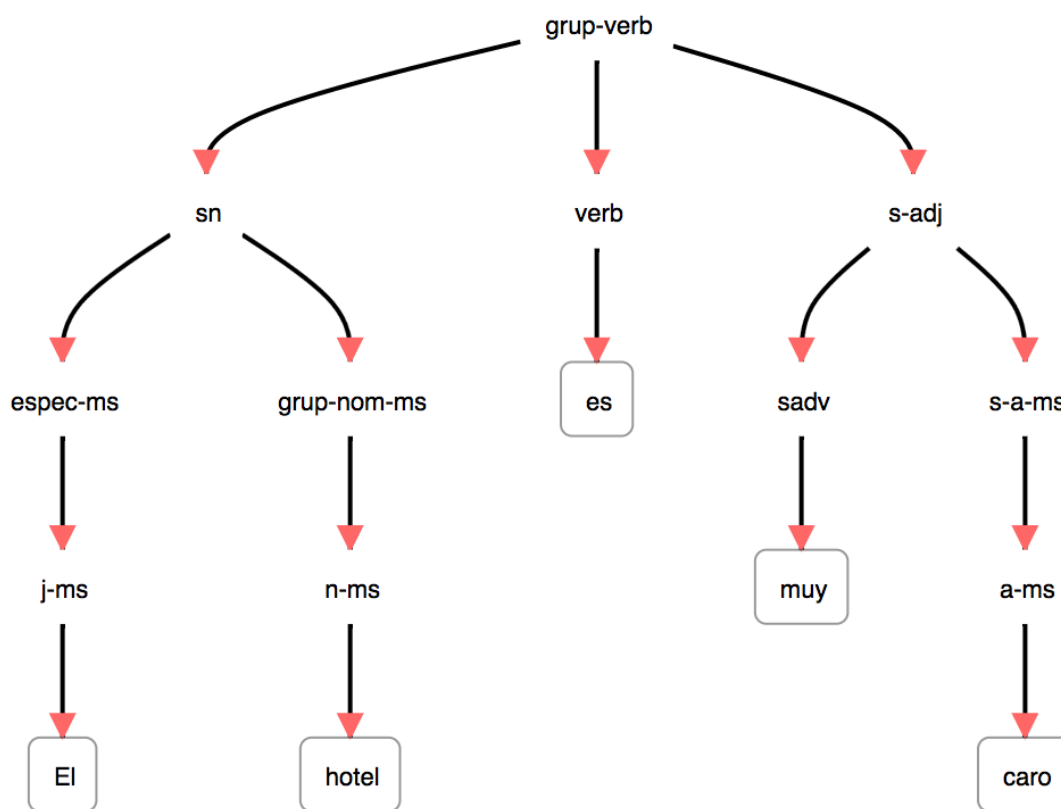


Figura 2.13. Árbol de dependencias.

2.4.4.1.2. Métodos de selección de características

Los métodos de selección de características tienen como objetivo crear un modelo predictivo más preciso a través de la selección de características más relevantes y discriminantes para la clasificación. En la clasificación de sentimientos basada en máquinas de aprendizaje, un problema es la alta dimensión de los vectores de características. En este sentido un método de selección de características es necesario para eliminar características irrelevantes y redundantes del vector. Las características irrelevantes son aquellas que no contribuyen en la clasificación y las características redundantes son esas que están altamente correlacionadas. En un gran conjunto de características, es común que muchas características no sean informativas debido a la irrelevancia o redundancia con respecto a la clase. Por lo que remover este tipo de características permite mejorar los resultados del rendimiento de los algoritmos de clasificación.

La dimensión del vector de características puede ser reducido usando métodos tales como IG (*Information Gain*, Ganancia de información), MI (*Mutual*

Information, información mutua), CHI (*Chi-square*, chi-cuadrado). A continuación, se detallan cada uno de estos métodos.

- **Ganancia de Información:** Es la medida de selección de características ampliamente usado en el análisis de sentimientos. Determina la relevancia de una característica dada, estudiando la presencia o ausencia de la característica en el documento. Normalmente, las características con los valores más altos son seleccionadas para reducir el tamaño del vector (Arafat et al., 2014). IG, es definida por la ecuación (2.6)

$$IG(t) = \sum_{K=1}^C P(C_k) \log \frac{1}{P(C_k)} - \sum_{t \in \{t_p, t_n\}} P(t) \sum_{K=1}^C P(t|C_k) \log \frac{1}{P(t|C_k)} \quad (2.6)$$

Donde $P(C_k)$ es la probabilidad a priori de un documento que ocurre en la clase C_k , $P(t)$ es la probabilidad de que el termino t ocurra o no en un documento. $P(t|C_k)$ es la probabilidad condicional del termino t que ocurre o no en un documento de la clase C_k , y C es el número de clases.

- **Información mutua:** Es un criterio ampliamente utilizado en el modelado estadístico de las asociaciones de palabras. Dado el término t_i y un conjunto de documentos C^k , la información mutua entre ellos es definida como:

$$MI(t_i, C^k) = \log \frac{P(t_i, C^k)}{P(t_i) \times P(C^k)} \quad (2.7)$$

Donde $P(t_i, C^k)$, es la probabilidad conjunta de que un documento contenga el termino t_i y pertenezca al conjunto de documentos C^k , $P(t_i)$ es la probabilidad de que un término t_i ocurra en un documento. $P(C^k)$ es la probabilidad de que un documento pertenezca a un conjunto de documentos C^k

- **Chi-cuadrado:** Mide la divergencia de los datos observados de los valores que se esperan bajo la hipótesis nula de ausencia de asociación. Esto requiere el cálculo de los valores esperados basados en los datos.

$$X^2 = \sum \frac{(O - E)^2}{E} \quad (2.8)$$

Donde O define la puntuación observada, y E especifica la puntuación esperada.

2.4.4.1.3. Algoritmos de clasificación

En esta sección se lleva a cabo una revisión de algoritmos de aprendizaje supervisado, también conocidos como clasificadores: Máquina de soporte vectorial (SVM), Redes bayesianas (BN) y árboles de decisión.

2.4.4.1.3.1. Máquina de soporte vectorial

La máquina de soporte vectorial es una técnica de aprendizaje automático, el cual tiene una sólida base teórica y realiza una clasificación más precisa que la mayoría de otros algoritmos. Muchos investigadores han informado que SVM es el método más preciso para la clasificación de texto y es comúnmente utilizado en la clasificación de sentimientos (Pang et al., 2002) (Shboul, Al-Ayyoub, & Jararweh, 2015) (Yijing, Haixiang, Xiao, Yanan, & Jinling, 2016). SVM es un método de aprendizaje lineal que encuentra un hiperplano óptimo para separar dos clases. SVM busca maximizar la distancia al punto de entrenamiento más cercano de cada clase para lograr un mejor rendimiento de clasificación (Moraes et al., 2013).

2.4.4.1.3.2. Red bayesiana

La red bayesiana es estructurada como una combinación de un grafo acíclico dirigido de nodos y enlaces, y un conjunto de tablas de probabilidad condicional. Los nodos representan características o clases, mientras que los enlaces entre los nodos representan la relación entre ellos. Las tablas de probabilidad condicional determinan la fuerza de los enlaces. Hay una tabla de probabilidad para cada nodo (característica) que define la distribución de probabilidad para el nodo, dados sus nodos principales (ver Figura 2.15). Si un nodo no tiene padres, la distribución de probabilidad es incondicional. Si un nodo tiene uno o más padres, la distribución de probabilidad es una distribución condicional, donde la probabilidad de cada valor de característica depende de los valores de los padres (Williams, Zander, & Armitage, 2006). La Figura 2.14 muestra un ejemplo de una red bayesiana, además como se mencionó anteriormente existe una tabla de probabilidad para cada nodo,

en este ejemplo se muestra únicamente la tabla para el nodo “pronóstico” (ver Figura 2.15).

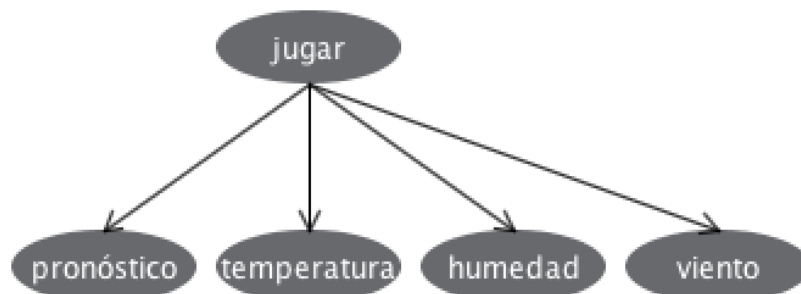


Figura 2.14. Red bayesiana.

jugar soleado	nublado	lluvioso	
sí	0.238	0.429	0.333
no	0.538	0.077	0.385

Figura 2.15. Tabla de distribución de probabilidad.

2.4.4.1.3.3. Árboles de decisión

Para problemas de clasificación donde se utilizan conjuntos grandes de datos y la información contenida es compleja, los árboles de decisión proporcionan una solución útil (Aitkenhead, 2008). En un árbol de decisiones cada nodo del árbol representa una característica en una instancia para ser clasificada y cada rama representa un valor que el nodo puede asumir. Las instancias se clasifican a partir del nodo raíz y se ordenan en función de sus valores de la característica. El algoritmo más conocido en la literatura para construir árboles de decisión es el C4.5 (Quinlan, 2014), que es una extensión del algoritmo ID3 (Quinlan, 1979).

Una de las principales ventajas de los árboles de decisión es que son fáciles de entender e interpretar, es decir, las personas son capaces de comprender los modelos de árboles de decisión después de una breve explicación. Además, los árboles de decisión tienden a funcionar mejor cuando se trata de características discretas. La Figura 2.16 muestra un ejemplo de un árbol de decisión para el

Capítulo 2. Estado del arte

conjunto de entrenamiento de la Tabla 2.7.. A partir de este árbol se puede determinar, por ejemplo, que si el pronóstico es soleado y la humedad es menor o igual que 75, entonces sí se puede jugar, o que si el pronóstico es lluvioso y hay viento se determina que no se puede jugar.

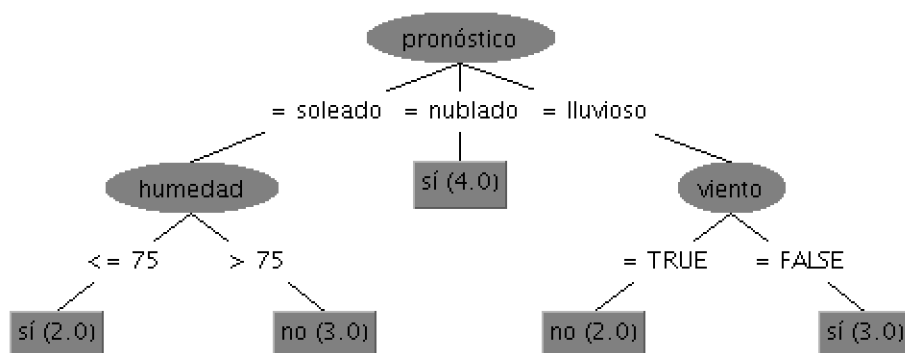


Figura 2.16. Árbol de decisión.

Tabla 2.7. Datos de entrenamiento.

pronóstico	temperatura	humedad	viento	clase
soleado	85	85	FALSE	no
soleado	80	90	TRUE	no
nublado	83	86	FALSE	sí
lluvioso	70	96	FALSE	sí
lluvioso	68	80	FALSE	sí
lluvioso	65	70	TRUE	no
nublado	64	65	TRUE	sí
soleado	72	95	FALSE	no
soleado	69	70	FALSE	sí
lluvioso	75	80	FALSE	sí
soleado	75	70	TRUE	sí
nublado	72	90	TRUE	sí
nublado	81	75	FALSE	sí
lluvioso	71	91	TRUE	no

2.4.4.2. Enfoque semántico

El enfoque de orientación semántica es dividido en dos principales categorías: basado en diccionario y basado en corpus. En el enfoque basado en diccionario, el sentimiento es identificado usando sinónimos y antónimos desde un diccionario léxico como WordNet. En el enfoque basado en corpus, palabras que denotan opinión son identificadas en base a una lista de palabras.

La mayoría de estudios están orientados principalmente en un enfoque basado en lexicones. Los lexicones son diccionarios de palabras que contienen una orientación semántica o polaridad, es decir un valor positivo o negativo. Por tanto, los recursos léxicos son una parte importante del enfoque semántico. Actualmente existen diversos léxicos disponibles (ver Tabla 2.8.). En la sección 2.4.3.2 se describe cada uno a detalle. SentiWordNet es el lexicón que ha sido más ampliamente utilizado. Este es un recurso que deriva de WordNet, y que provee tres valores numéricos (positivos, negativos y objetivos) para cada *synset* de WordNet.

Tabla 2.8. Recursos léxicos.

Lexicones
ML-SentiCon (Cruz et al., 2014)
SentiWordNet (Esuli & Sebastiani, 2006)
SpanishSentimentLexicons (J. Wiebe et al., 2005)
General Inquirer (Philip J. et al., 1966)
Léxico de opiniones de Bing Liu (Hu & Liu, 2004)(Liu et al., 2005)

El proceso de detección de la polaridad consiste en procesar el texto y detectar palabras que denoten opinión, por lo que las palabras que no proporcionan ninguna información tales como palabras vacías (en inglés, *stop words*) son eliminadas. Una vez que se han obtenido las palabras de opinión se comprueba su aparición en el lexicón para asignar una polaridad. Posteriormente se lleva a cabo la suma de los valores de polaridad de los términos y se obtiene una polaridad global de la opinión.

Capítulo 2. Estado del arte

Comúnmente en los lexicones, una palabra puede tener diferentes significados, y por tanto diferentes valores de polaridad. Por ejemplo, en la Tabla 2.9. se muestra un extracto de los resultados obtenidos para la palabra “mejor” (*better*, en inglés) desde *SentiWordNet*. Esta tabla muestra 13 significados, 4 que pertenecen a la categoría “adjetivo”, 4 que pertenecen a la categoría “sustantivo”, 2 que pertenecen a la categoría “adverbio”, y 3 que pertenecen a la categoría “verbo”. Para resolver este problema y mejorar la precisión de la clasificación, varios enfoques determinan la categoría gramatical de las palabras y en base a la categoría que se le fue asignada realizan la suma de todos sus sentidos (Peñalver-Martinez et al., 2014). Algunos otros enfoques llevan a cabo WSD (*Word Sense Desambiguation*, desambiguación de los significados de las palabras) obteniendo un único valor de polaridad para cada palabra.

Tabla 2.9. Puntuación de algunos synset en SentiWordNet.

Etiqueta	Identificador	Positividad	Negatividad	Synset	Descripción	Ejemplos
a	00067379	0.75	0	better#3 best#2	comparative and superlative of `well') wiser or more advantageous and hence advisable	"it would be better to speak to him"; "the White House thought it best not to respond"
a	00230335	0.875	0	better#1	(comparative of `good') superior to another (of the same class or set or kind) in excellence or quality or desirability or suitability; more highly skilled than another	"You're a better man than I am, Gunga Din"; "a better coat"; "a better type of car"; "a suit with a better fit"; "a better chance of success"; "produced a better mousetrap"; "she's better in math than in history"
a	00231761	0.875	0	better#2	(comparative of `good') changed for the better in health or fitness	"her health is better now"; "I feel better"
a	01471260	0	0	better#4	more than half	"argued for the better part of an hour"
n	05143558	0.5	0	better#1	something superior in quality or condition or effect	"a change for the better"
n	05143690	0.5	0	better#4	the superior one of two alternatives	"chose the better of the two"
n	09851719	0.5	0	better#3	a superior person having	"the common man has been kept

Etiqueta	Identificador	Positividad	Negatividad	Synset	Descripción	Ejemplos
					claim to precedence	in his place by his betters"
n	09851876	0	0	wagerer#1 punter#3 bettor#1 better#2	someone who bets	
r	00509846	0	0	better#2 best#3	from a position of superiority or authority	"father knows best"; "I know better."
r	00059607	0.875	0	better#1	comparative of `well'; in a better or more excellent manner or more advantageously or attractively or to a greater degree etc.;	"She had never sung better"; "a deed better left undone"; "better suited to the job"
v	00205046	0.375	0	meliorate#2 improve#2 better#3 ameliorate#2	get better	"The weather improved toward evening"
v	00205885	0.375	0	meliorate#1 improve#1 better#2 amend#2 ameliorate#1	to make better	"The editor improved the manuscript with his changes"
v	01106864	0.75	0	break#14 better#1	surpass in excellence	She bettered her own record"; "break a record"

En algunos casos este enfoque no ha resultado ser tan óptimo debido a que una palabra puede tener diferentes sentidos (positivo o negativo) dependiendo del contexto donde son utilizadas. Para hacer frente a este hecho, lexicones dependientes del contexto han sido propuestos por diferentes trabajos (Molina-Gonzalez, 2016) (Park, Lee, & Moon, 2015) (Deng, Sinha, & Zhao, 2016).

2.4.5. Comparación de métodos para el análisis de sentimientos

En los últimos años varios enfoques han sido propuestos para el análisis de opiniones desde varias fuentes de opinión (blogs, foros, sitios Web comerciales). Aunque recientemente un interés especial ha surgido sobre las redes sociales, donde la gente comparte sus opiniones acerca de diversos temas. A continuación, en la Tabla 2.10. se presentan los trabajos más actuales para el análisis de sentimientos, los cuales están basados en dos principales enfoques: orientación semántica y aprendizaje automático supervisado. Llevar una comparación de los diferentes enfoques es una tarea difícil debido a que están orientados a objetivos específicos, a diferentes idiomas y a diferentes dominios. Por lo tanto, el corpus utilizado para evaluar la mayoría de sistemas difiere en tópicos, tamaño e idioma. Sin embargo, se han determinado una serie de características a través de las cuales podemos diferenciar sus principales aportaciones. Las características que se han determinado son las siguientes:

- **Dominio:** Se refiere al contexto en el que ha sido principalmente enfocado el sistema.
- **Lenguaje:** Se refiere al idioma o idiomas hacia el cual fue enfocado el método.
- **Enfoque:** Es la técnica utilizada para llevar a cabo el análisis de sentimientos, aprendizaje automático o de orientación semántica.
- **Características:** Determina que características han sido obtenidas para posteriormente ser utilizadas para la clasificación de sentimientos.
- **Nivel:** Esta característica determina el nivel de clasificación de los sentimientos, es decir, a nivel de documento, sentencia o aspecto.

Tabla 2.10. Comparativa de enfoques actuales de análisis de sentimientos.

	Dominio	Lenguaje	Enfoque	Características	Nivel
(Deng, Luo, & Yu, 2014)	Películas Productos	Inglés	Aprendizaje automático supervisado (SVM)	Frecuencia de términos	Documento
(Ghiassi, Skinner, & Zimbra, 2013)	Twitter	Inglés	Aprendizaje automático supervisado (DAN2 y SVM)	Frecuencia de términos	
(Habernal et al., 2014)	Facebook Películas Productos	Checo	Aprendizaje automático supervisado (MaxEnt y SVM)	n-gramas, características basadas en etiquetas POS, emoticones y TF-IDF	
(Moraes et al., 2013)	Películas Productos	Inglés	Aprendizaje automático supervisado (Naive Bayes, SVM, ANN)	TF-IDF	Documento
(Singh, Piryani, Uddin, & Waila, 2013)	Películas	Inglés	Semántico (SentiWordNet)	Adverbios, adjetivos, verbos	Documento Aspecto
(Al-Ayyoub, Essa, & Alsmadi, 2015)	Twitter	Árabe	Semántico (Lexicón árabe)	Palabras que denotan opinión	
(Peñalver-Martinez et al., 2014)	Películas	Inglés	Semántico (SentiWordNet)	n-gramas	Sentencia Aspecto
(Montejo-Ráez, Martínez-Cámara, Martín-Valdivia, & Ureña-López, 2014)	Twitter	Inglés	Semántico (SentiWordNet)	Adverbios, adjetivos, verbos	
(Molina-González et al., 2013)	Películas	Español	Semántico (iSOL, eSOL)	Palabras que denotan opinión	Documento

2.4.5.1. Dominio

Los trabajos anteriormente presentados se han enfocado principalmente en tres contextos: películas, productos, y conjuntos de datos obtenidos desde las redes sociales. Por ejemplo, en (Z.-H. Deng et al., 2014), (Moraes et al., 2013) y (Peñalver-Martinez et al., 2014), utilizan el corpus de películas presentado en (Pang et al., 2002). El corpus contiene 1000 opiniones positivas y 1000 opiniones negativas de películas extraídas de una base de datos de películas de internet (IMDb). Además, (Z.-H. Deng et al., 2014) utiliza el corpus proporcionado por (Maas et al., 2011). Este conjunto de datos contiene 50000 opiniones de películas. Por otro lado, en (Habernal et al., 2014) obtuvieron opiniones de la base de datos de películas checa CSFD². El conjunto de datos contiene un total de 30897 opiniones positivas, 30768 neutrales y 29716 negativas. En (Singh et al., 2013) coleccionan opiniones de 100 películas hindú desde la base de datos de opiniones de películas IMDb obteniendo un total de 760 positivas y 240 negativas. Finalmente, (Molina-González et al., 2013) utilizaron el corpus MuchoCine presentado en (Cruz Mata et al., 2008). El corpus consiste de 3878 opiniones

Con respecto al dominio de productos, en (Z.-H. Deng et al., 2014) obtienen un conjunto de datos de opiniones de productos desde Amazon, el cual contiene un total de 16000 opiniones. Por su parte, en (Habernal et al., 2014) obtienen opiniones de una tienda de compras online checa³. El corpus final consiste de 102977 post positivos, 31943 post neutrales, y 10387 post negativos. Mientras tanto, en (Moraes et al., 2013) coleccionan desde Amazon opiniones de distintos productos como GPS, libros, y cámaras. El conjunto de datos contiene 2000 opiniones, 1000 positivas y 1000 negativas.

Por otro lado, con respecto a los corpora de medios sociales, en (Ghiassi et al., 2013) coleccionan un conjunto de tweets. El sujeto seleccionado en su investigación fue “Justin Bieber”. Mientras tanto en, (Habernal et al., 2014)

² <http://www.csfd.cz/>

³ <http://www.mall.cz>

Capítulo 2. Estado del arte

coleccionan post desde nueve diferentes cuentas de Facebook⁴ de marcas, servicios o productos. El corpus contiene 2587 post positivos, 5174 post neutros, y 1991 post negativos. En (Al-Ayyoub et al., 2015), obtienen un corpus de tweets en árabe. Finalmente, (Montejo-Ráez et al., 2014) coleccionaron un corpus de 181492 tweets positivos y 194804 tweets negativos. A continuación, se presenta en la Tabla 2.11 un resumen de los conjuntos de datos utilizados en cada trabajo presentado anteriormente.

Tabla 2.11. Resumen de los conjuntos de datos.

	Dominio	Lugar de procedencia	Tamaño del corpus
(Z.-H. Deng et al., 2014)	-Películas -Productos	-Base de datos de películas en Internet (IMDb). -Amazon	-2000 opiniones (1000 positivas, 1000 negativas) -16000 opiniones
(Ghiassi et al., 2013)	-Justin Bieber.	-Twitter	-10345184 tweets
(Habernal et al., 2014)	-Czench posts -Películas -Productos	-Facebook -CSFD -e-shop Mall	-10000 posts -91381 opiniones (30897 positivas, 30768 neutras, 29716 negativas) -145307 posts (102977 positivas, 31943 neutras y 10387 negativas)
(Moraes et al., 2013)	-Películas -Productos	-Base de datos de películas en Internet (IMDb) -Amazon	-2000 opiniones (1000 positivas, 1000 negativas)
(Singh et al., 2013)	-Películas	-Base de datos de películas en Internet (IMDb)	-1000 opiniones (760 positivas, 240 negativas)
(Al-Ayyoub et al., 2015)	-Twitter	-Twitter	-900 tweets (300 positivos, 300 negativos y 300 neutros)
(Peñalver-Martinez et al.,	-Películas	-Base de datos de películas en Internet	-2000 opiniones (1000 positivas, 1000 negativas)

⁴ <https://www.facebook.com/>

	Dominio	Lugar de procedencia	Tamaño del corpus
2014)		(IMDb)	
(Montejo-Ráez et al., 2014)	-Twitter	Twitter	-376296 tweets (181492 positivos, 194804 negativos)
(Molina-González et al., 2013)	-Películas	-MuchoCine	-3878 opiniones (1274 positivas, 2625 negativas)

La propuesta presentada en este trabajo de tesis se puede aplicar a cualquier dominio, sin embargo, se requiere de un corpus etiquetado. El método presentado en este trabajo ha sido validado en el dominio de películas y en el dominio turístico. Ambos dominios son de gran importancia debido a que con frecuencia las personas buscan comentarios de películas para tomar la decisión sobre ver una película y respecto a viajes, para elegir su próximo destino de viaje. El primero se ha utilizado como un corpus de referencia para la evaluación del propuesto método, el cual ha sido ampliamente utilizado en la comunidad investigadora. El segundo fue construido desde la Web TripAdvisor⁵.

2.4.5.2. Lenguaje

La mayoría de métodos propuestos para el análisis de sentimientos están enfocados principalmente en el idioma inglés (Z.-H. Deng et al., 2014) (Ghiassi et al., 2013) (Moraes et al., 2013) (Singh et al., 2013) (Peñalver-Martinez et al., 2014) (Montejo-Ráez et al., 2014). Algunas otras propuestas están enfocadas en idiomas tales como el árabe (Al-Ayyoub et al., 2015), checo (Habernal et al., 2014) y español (Molina-González et al., 2013).

El método propuesto en este trabajo de tesis está enfocado en el idioma español. El español es el tercer idioma más utilizado en Internet por lo que el análisis de opiniones en este idioma es de lo más importante. Además, el método propuesto puede ser fácilmente adaptado a cualquier idioma.

⁵ <https://www.tripadvisor.com>

2.4.5.3. Enfoque

Como se mencionó en la sección 2.4.4 existen dos enfoques principales en los que se basan la mayoría de trabajos de análisis de sentimientos, el aprendizaje automático y el de orientación semántica. Por un lado, los trabajos enfocados en aprendizaje automático requieren de un corpus de opiniones etiquetado para entrenar un algoritmo de clasificación. Por ejemplo, en (Z.-H. Deng et al., 2014) utilizan el algoritmo de clasificación SVM, y los parámetros por defecto con un kernel lineal. (Ghiassi et al., 2013) particionan el conjunto de datos en conjunto de entrenamiento el cual consta del 51% de los datos y en conjunto de prueba con el 49% de datos. Los algoritmos de clasificación utilizados son DAN2, (*A Dynamic Architecture for Artificial Neural Networks*) y SVM. En (Habernal et al., 2014) dos algoritmos de clasificación, MaxEnt, y SVM son entrenados utilizando los valores por defecto para cada algoritmo y un kernel lineal para SVM. Finalmente, en (Moraes et al., 2013) entrenan tres algoritmos de clasificación (SVM, ANN (*Artificial neural network*, red neuronal artificial) y BN) con dos conjuntos de datos uno balanceado y uno no balanceado. Los resultados son obtenidos en términos de una validación cruzada de 10.

Por otro lado, en el enfoque semántico se requiere comúnmente de lexicones de palabras. Por ejemplo, en (Singh et al., 2013) utilizan el lexicón SentiWordNet para asignar valores a las palabras basadas en su etiqueta POS. Por otro lado, en (Al-Ayyoub et al., 2015) obtienen la polaridad de las palabras basados en un lexicón en árabe propuesto en el mismo trabajo, el cual contiene alrededor de 120000 distintos términos en árabe. En (Peñalver-Martinez et al., 2014) la polaridad de un aspecto es obtenida a través de la suma de los valores de polaridad de las palabras relacionadas a dicho aspecto. El valor de cada palabra es obtenido a través de SentiWordNet y en base a la categoría gramatical de la palabra. Además, (Montejo-Ráez et al., 2014) resuelven el problema de clasificación de polaridad combinando los valores de SentiWordNet con un análisis *random walk* (Toutanova, Manning, & Ng, 2004) de los conceptos encontrados en el texto sobre el grafo de WordNet. Finalmente, en (Molina-González et al., 2013) la polaridad de la opinión es obtenida en base a dos lexicones, uno de propósito general (iSOL) y uno dependiente del dominio (eSOL).

Cada enfoque en la clasificación de sentimientos tiene sus ventajas y desventajas. Por un lado, el enfoque de orientación semántica es no supervisado e independiente del dominio. Comúnmente un lexicón es construido para todos los dominios, aunque actualmente han surgido varias propuestas dependientes del dominio, donde se ha demostrado una mejor precisión. Por otro lado, el enfoque de aprendizaje automático supervisado requiere de un corpus específico del dominio etiquetado para el entrenamiento del clasificador. Un factor clave para determinar el éxito de un enfoque automático depende de la calidad y cantidad de los datos de entrenamiento. Existen varios corpus disponibles que han demostrado ser efectivos. Sin embargo, hay carencia de conjuntos de datos para ciertos dominios y la obtención de un corpus es tediosa. No obstante, si se obtiene un conjunto de datos de calidad que representen el dominio, el aprendizaje automático suele ser más preciso que el enfoque de orientación semántica. Por las razones antes mencionadas, este trabajo de tesis doctoral se base en un enfoque de aprendizaje automático supervisado. Tres algoritmos de clasificación han sido utilizados SVM, BN y J48 en los experimentos.

2.4.5.4. Características

La extracción de características es un proceso de suma importancia para la correcta clasificación de las opiniones. Por un lado, los enfoques basados en aprendizaje automático supervisado, requieren de un vector de características, las cuales son utilizadas para entrenar el clasificador y para que este aprenda de las características. Por tanto, la precisión de la clasificación depende de la correcta clasificación del vector de características. Los principales métodos utilizado son n-gramas (Habernal et al., 2014), TF-IDF (Habernal et al., 2014) (Moraes et al., 2013). Y más recientemente algunos otros enfoques obtienen características más sofisticadas como basadas en etiquetas POS (Habernal et al., 2014). Por otro lado, los enfoques basados en orientación semántica requieren de la extracción de palabras que denoten opinión, con el objetivo de asignarles un valor obtenido de un lexicón. Comúnmente estos enfoques descartan palabras vacías y consideran adjetivos, adverbios, y verbos (Singh et al., 2013) (Montejo-Ráez et al., 2014). Algunos otros autores (Peñalver-Martinez et al., 2014) utilizan n-gramas tales como *n-gram before*, *n-gram after* y *n-gram around*.

Capítulo 2. Estado del arte

El método propuesto en esta tesis doctoral, a diferencia de los trabajos anteriormente presentados se basa en la extracción de características psicolingüísticas. Una de las cuestiones clave en la psicolingüística es el reflejo de cuestiones emocionales y cognitivas de los seres humanos en el lenguaje oral y escrito. Además, una investigación en este campo demuestra que hay una relación entre el lenguaje y el estado del ánimo de las personas (Stiles, 1992).

2.4.5.5. Nivel

Como se mencionó en la sección 2.4.2, existen tres niveles de clasificación de los sentimientos, nivel de documento, nivel de sentencia y nivel de aspecto. Como una primera aproximación este trabajo de tesis doctoral está enfocado a nivel de documento. No obstante, a futuro se planea llevar una clasificación a nivel de sentencia y aspectos a partir de patrones psicolingüísticos.

2.5. El lenguaje figurado

El lenguaje figurado es uno de los temas más difíciles del PLN. A diferencia del lenguaje literal, el lenguaje figurado utiliza recursos lingüísticos tales como la metáfora, la analogía, la ambigüedad, entre otros, para proyectar significados más complejos, que comúnmente es difícil de entender no sólo para las computadoras, sino también para el ser humano (Reyes, Rosso, & Buscaldi, 2012).

El uso del lenguaje figurado como la ironía, el sarcasmo y la sátira han sido considerados como un fenómeno importante en las plataformas sociales como Twitter. Por lo que, la detección del lenguaje figurado ha sido fundamental para enfoques en campos como el análisis de sentimientos, donde la polaridad afectiva del significado literal puede contrastar notablemente con el afecto creado por el significado figurado.

2.5.1. Ironía, Sarcasmo, Sátira

(Wilson & Sperber, 2007) definen la ironía como un acto comunicativo que expresa lo contrario a lo que se dice literalmente. Varios tipos de ironía han sido distinguidos tales como la ironía dramática, la ironía trágica entre otros. Sin embargo, la mayoría de expertos consideran dos tipos de ironía: ironía verbal e ironía situacional. La ironía verbal comunica un sentido opuesto, es decir, una

persona dice lo contrario a lo que se quiere referir. La ironía situacional, es una propiedad inesperada o incongruente en una situación o evento, es decir, situaciones que no sucederán.

La mayoría de los estudios sobre la ironía verbal y situacional se basan en un enfoque lingüístico. Con respecto a la ironía verbal, está ampliamente relacionada con el significado del sarcasmo y la sátira. Por ejemplo, (Colston, 2000) considera el sarcasmo como un término que se utiliza comúnmente para describir una expresión irónica verbal. (Gibbs, 2000) menciona que el sarcasmo en combinación con humor, hipérbole, preguntas retóricas y subestimaciones son sólo tipos de ironía. Por lo que la ironía es considerada como una macro categoría, donde el sarcasmo y la sátira son más específicas. Sin embargo, existen diferencias significativas que distinguen a cada una. Por ejemplo, el sarcasmo tiene como objetivo atacar, es más ofensivo y está muy asociado con estados afectivos negativos. Por lo que un comentario sarcástico se distingue por contener palabras agresivas. Por lo contrario, la ironía se considera no ofensiva. Finalmente, la sátira es usada para criticar o ridiculizar algo o alguien, por lo que es más utilizada para expresar situaciones de indignación utilizando parodia, exageración, comparación, y doble sentido.

El uso del lenguaje figurado en algunos casos ha resultado ser muy interesante, por ejemplo, como disparador de ventas. Uno de los casos más interesantes es el de la camiseta de la luna de tres lobos (Ver Figura 2.17). Este artículo se convirtió en uno de los productos más populares de Amazon debido al comentario irónico que hizo una persona. A partir de ahí se convirtió viral y esto provocó el aumento de la venta de dicho producto.

Capítulo 2. Estado del arte

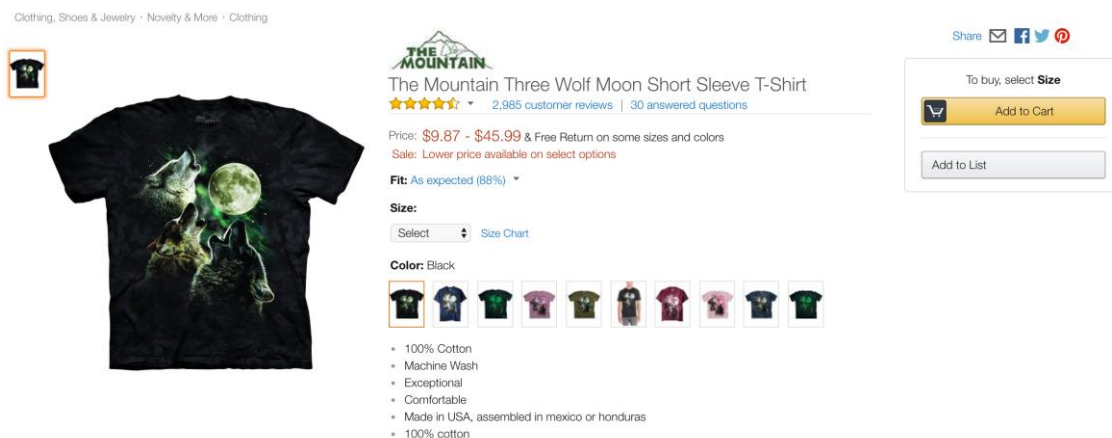


Figura 2.17. Producto con comentarios irónicos.

2.5.2. Comparación de enfoques para la detección de la sátira

En los últimos años diversos enfoques han surgido con el objetivo de llevar a cabo la detección automática de la ironía y el sarcasmo. Sin embargo, pocos estudios han abordado la detección de la sátira, es por ello que este trabajo de tesis doctoral se enfoca en la detección de la sátira.

Con el objetivo de realizar una comparativa entre métodos para la detección de la sátira más importantes en la actualidad, se han definido los siguientes parámetros:

Idioma: Es el idioma o los idiomas en lo que han sido aplicados los métodos propuestos en cada uno de los enfoques presentados.

Características: Son el conjunto de características extraídas para su posterior uso en la detección de la sátira.

Dominio: Es el dominio o dominios en los que han sido validados los métodos propuesto en cada uno de los enfoques presentados.

Enfoque: Son las técnicas utilizadas para la detección automática de la sátira.

En la Tabla 2.12. se muestra un resumen de la comparación realizada a partir de las características comentadas anteriormente. Posteriormente se analizarán por separado cada una de las características para establecer una comparativa entre los diferentes métodos estudiados.

Tabla 2.12. Tabla comparativa de métodos para la detección de la sátira.

Propuesta	Idioma	Características	Dominio	Enfoque
(Rubin, Conroy, Chen, & Cornwell, 2016)	Inglés	-Afectos negativos -Pronombres -Adjetivos -Adverbios	Noticias	Aprendizaje automático supervisado (SVM)
(Ahmad, Akhtar, Chopra, & Akhtar, 2014)	Inglés	-F-IDF	Noticias	Aprendizaje automático supervisado (SVM)
(Burfoot & Baldwin, 2009)	Inglés	-Pesos de características binarias -Características BNS	Noticias	Aprendizaje automático supervisado (SVM)
(Barbieri, Ronzano, & Saggion, 2015b)	Español	-Frecuencia -Ambigüedad -Verbos -Adjetivos -Adverbios -Interjecciones -Determinadores -Pronombres -Sinónimos	Noticias	Aprendizaje automático supervisado (SVM)

Propuesta	Idioma	Características	Dominio	Enfoque
(Owais, Nafis, & Khanna, 2015)	Inglés	-Palabras positivas y negativas -Malas palabras -Pesos de características binarias -Características BNS -Jerga -Malas palabras	Noticias	Aprendizaje automático supervisado (SVM)
(Barbieri, Ronzano, & Saggion, 2015a)	Inglés Español Italiano	-Frecuencia -Ambigüedad -Verbos -Adjetivos -Adverbios -Interjecciones -Determinadores -Pronombres -Sinónimos -Palabras positivas y negativas -Malas palabras	Noticias	Aprendizaje automático supervisado (SVM)

2.5.2.1. Idioma

La mayoría de enfoques para la detección de la sátira han sido diseñadas para funcionar con el idioma inglés (Rubin et al., 2016) (Ahmad et al., 2014) (Burfoot & Baldwin, 2009) (Owais et al., 2015) (Barbieri et al., 2015a), quizá debido a la carencia de recursos en otros idiomas. Sin embargo, hay algunas otras investigaciones que se centran en el idioma italiano y español (Barbieri et al., 2015a) y (Barbieri et al., 2015b).

El método propuesto en esta tesis doctoral es enfocado al español, el cual es uno de los principales idiomas utilizados en internet. Por lo tanto, la información de los dominios de internet en este idioma es de suma importancia. Aunque trabajos tales como los de (Barbieri et al., 2015a) y (Barbieri et al., 2015b) están enfocados al español de España. Este trabajo de tesis doctoral está enfocado al español de España y México. Y además se lleva a cabo un análisis para determinar si hay una diferencia significativa entre los patrones de características usadas para detectar la sátira entre estas dos culturas.

2.5.2.2. Características

Las propuestas anteriormente presentadas se basan en diferentes enfoques de extracción de características tales como TF-IDF (Ahmad et al., 2014), pesos de características binarias y características BNS (Burfoot & Baldwin, 2009) (Owais et al., 2015), mientras que algunos otros estudios exploran características más sofisticadas tales como lingüísticas y afectivas (Rubin et al., 2016) (Barbieri et al., 2015b) (Barbieri et al., 2015a). Sin embargo, respecto a estas últimas características han sido seleccionadas sobre la base de otros estudios suponiendo que serán relevantes en la detección de la sátira. Desde esta perspectiva, en este trabajo de tesis doctoral se analiza un conjunto más amplio de características relacionadas con aspectos psicolingüísticos, signos de puntuación, asuntos personales y categorías del lenguaje oral con el fin de encontrar características importantes que puedan contribuir a determinar si un texto es satírico o no.

2.5.2.2.1. Dominio

Las aportaciones en la detección de la sátira se han enfocado principalmente en las noticias satíricas. Por ejemplo, (Rubin et al., 2016) recopilan y analizan un corpus de 360 artículos de noticias. El conjunto de datos se recopiló de dos sitios satíricos y dos no satíricos de noticias estadounidenses y canadienses, *The Onion* y *The Beaverton*, y *The Toronto Star* y *The New York Times*. Por su parte, (Ahmad et al., 2014), (Burfoot & Baldwin, 2009) y (Owais et al., 2015) obtienen noticias satíricas y reales desde fuentes como *The Onion* y del Corpus *Giwawor*. En (Ahmad et al., 2014) utilizan un corpus de un total de 2624 noticias verdaderas y 171 artículos de noticias satíricas. Mientras que el corpus presentado en (Burfoot & Baldwin, 2009) consta de un total de 4000 noticias reales y 233 noticias satíricas. En (Owais et al., 2015), el corpus obtenido contiene un total de 2500 noticias reales y 110 noticias satíricas.

Por otro lado, en (Barbieri et al., 2015b) y (Barbieri et al., 2015a) obtienen un corpus desde fuentes satíricas españolas tales como *El mundo Today* y *El jueves*. Además en (Barbieri et al., 2015a) también se obtienen desde fuentes italianas como *Spinoza* y *Lercio*, y fuentes en inglés como *NewsBiscuit* y *The Daily Mash*.

Este trabajo al igual que los trabajos anteriormente presentados se centra en noticias satíricas, ya que es de suma importancia debido a que son particularmente populares en la Web y específicamente en las redes sociales, donde es relativamente fácil imitar una fuente de noticias real. Estas noticias pueden ser ampliamente distribuidas desde casi cualquier sitio en internet. Comúnmente las noticias satíricas se confunden con legítimas, específicamente cuando se disocia desde su fuente original, como ocurre cuando se recomienda una noticia satírica a través de la red social *Facebook*⁶. En este trabajo se obtuvieron dos conjuntos de datos uno desde fuentes españolas y otro desde fuentes mexicanas.

2.5.2.3. Enfoque

Todos los enfoques presentados para la detección automática de la sátira están basados en un enfoque de aprendizaje automático supervisado. En (Rubin et al.,

⁶ <https://www.facebook.com/>

2016) de su conjunto de datos que consta de 360 artículos de noticias reservan el 25% de datos para llevar a cabo sus pruebas y realizan una validación cruzada de 10 para el conjunto de entrenamiento. Por lo que el modelo fue entrenado con 270 artículos y probado con un conjunto de 90 artículos con proporciones iguales de noticias satíricas y legítimas. Ellos asignan dos clases, satíricas (1) y reales (0), por lo que llevan a cabo un entrenamiento supervisado utilizando el algoritmo SVM y un Kernel lineal. En (Ahmad et al., 2014) entrenan el algoritmo de clasificación SVM con un conjunto de 1495 artículos de noticias verdaderas y 100 noticias satíricas. Además, ellos llevan a cabo la evaluación del modelo con un conjunto de 1129 artículos de noticias verdaderas, y 71 de noticias satíricas. En (Burfoot & Baldwin, 2009) también entrenan el algoritmo SVM utilizando un Kernel lineal y con los valores por defecto de los parámetros. El conjunto utilizado para el entrenamiento consta de 2505 artículos de noticias verdaderas y 133 satíricas y el conjunto para pruebas consta de 1495 noticias verdaderas y 100 satíricas. Por otro lado, (Barbieri et al., 2015b) y (Barbieri et al., 2015a), llevan el entrenamiento automático supervisado utilizando el algoritmo SVM. Estos autores llevan a cabo una clasificación binaria balanceada, es decir un conjunto de datos con la misma cantidad de datos satíricos y no satíricos es utilizado tanto para entrenar como para validar. Por lo que un conjunto compuesto del 80% de noticias satíricas y otro conjunto del 80% de noticias no satíricas es usado para el entrenamiento. Entonces la validación fue llevada a cabo sobre un conjunto que incluye el 20% de las noticias satíricas y 20% de las noticias no satíricas. Finalmente, en (Owais et al., 2015) utilizan un conjunto de datos de 1800 noticias verdaderas y 70 noticias satíricas para entrenar el algoritmo SVM, y un conjunto de 700 noticias verdaderas y 40 satíricas para la validación.

El método presentado en este trabajo también está basado en un enfoque de aprendizaje automático supervisado. Sin embargo, la diferencia radica en que en este trabajo se lleva el entrenamiento de tres algoritmos de clasificación: J48, BayesNet y SVM con el objetivo de evaluar su rendimiento para este problema específico de clasificación. La elección de estos algoritmos se basó en que han sido usados exitosamente en tareas de clasificación. Además, se llevó una validación cruzada de 10, la cual es una de las mejores técnicas utilizadas para la evaluación de los resultados obtenidos por el clasificador.

2.6. Características psicolingüísticas

Cuando se habla de sentimientos subjetivos de emociones u opiniones, es necesario distinguir dos aspectos diferentes, los estados mentales o sentimientos de las personas y las expresiones del lenguaje utilizadas para describir los estados mentales. Aunque solo hay seis tipos de emociones (amor, alegría, sorpresa, ira, tristeza y miedo). Hay un gran número de expresiones del lenguaje que se pueden usar para expresarlos. Del mismo modo, hay un gran número de expresiones de opinión que describen sentimientos positivos y negativos. Las emociones han sido estudiadas en varios campos como psicología y sociología. Además, investigaciones en el campo de la psicolingüística provee evidencias empíricas de la relación entre el lenguaje y el estado de la mente de las personas o incluso su salud mental. La herramienta LIWC fue desarrollada precisamente para proveer un método eficiente para estudiar asuntos psicolingüísticos a través del análisis de un corpus.

2.6.1. LIWC en español

La fiabilidad de LIWC (Francis & Pennebaker, 1993) ha sido demostrada ampliamente en el ámbito del idioma inglés, desafortunadamente la mayoría de investigaciones que están relacionadas al uso de palabras y su impacto psicológico esta principalmente limitado a la lengua inglesa. Con el objetivo de abordar dicho problema, (Ramírez-Esparza, Pennebaker, García, & Suriá Martínez, 2007) desarrollaron la versión en español del diccionario. El proceso de traducción del diccionario de inglés a español involucro los siguientes pasos:

Las palabras fueron traducidas del inglés al español. Por ejemplo, el pronombre "I" en ingles se tradujo en español como "YO" y se mantuvo en sus categorías originales (pronombre, primera persona del singular y total primera persona). Esta fase se llevó a cabo con la ayuda de diversos diccionarios. Sin embargo, algunas palabras en español no se dejaron en las mismas categorías debido a que transmiten ideas que son congruentes en un idioma, pero no en otro. Por ejemplo, la palabra "blue" significa azul y estar triste en inglés. Sin embargo, en español solo significa azul.

Además, algunas otras palabras se agregaron en el diccionario en español que no existían en el diccionario en inglés, esto con el objetivo de mejorar el

diccionario en español. Por ejemplo, la palabra “rimbombante” no existe en el diccionario en inglés, pero si en la versión en español. Los verbos fueron conjugados a los tiempos presente, pasado y futuro, y tanto como en la forma singular como en el plural.

El tercer paso consistió en confirmar que las categorías asignadas en inglés tuvieran sentido en español incluyendo las modificaciones pertinentes. Por ejemplo, el verbo *take* en inglés, se puede traducir al español como tomar, agarrar, coger. Finalmente, una revisión por tres jueces fue llevada a cabo.

Con el objetivo de llevar a cabo la equivalencia entre ambos diccionarios, analizó textos en inglés y su traducción en español (Ramírez-Esparza et al., 2007). Esto les permitió analizar textos que tuvieran la misma información en ambos idiomas. Un total de 83 textos se obtuvieron de Internet, entre los temas que destacan se pueden mencionar: política, noticias, salud, bibliografías, propaganda, anuncios, canciones, reflexiones, recetas, poemas, por mencionar algunos. En su estudio los autores encontraron que las versiones correlacionaron en la mayoría de las categorías. Sin embargo, al analizar la diferencia de medias encontraron que no todas las categorías son comparables entre las versiones principalmente debido a problemas de gramática. Una de las categorías que difirió significativamente fue yo (o primera persona del singular). Esto debido a que, en inglés, al referirse a la primera persona en singular, siempre se incluye el pronombre yo. Sin embargo, en español se omite. Otro grupo de categorías que defirieron fueron los tiempos verbales, así como también categorías que incluyen personas. Por ejemplo, amigos, familia, hermanos. Los autores atribuyen esto a que en español existen más conjugaciones de verbos y a que una gran cantidad de palabras son usadas para indicar el género (por ejemplo., primo-prima, amigo-amiga, adulto-adulta).

2.6.1.1. Categorías de LIWC 2007

La versión final del diccionario de LIWC en español cuenta con un total de 7515 palabras y raíces de palabras, y con el mismo número de categorías y subcategorías que el LIWC en inglés, es decir un total de 72 categorías. Además, las categorías están clasificadas en cuatro grupos principales: 1) dimensiones lingüísticas, 2) procesos psicológicos, 3) asuntos personales y 4) categorías del

Capítulo 2. Estado del arte

lenguaje oral. Además, LIWC contiene un grupo que incluye 12 categorías de puntuación tales como punto, coma, exclamación, entre otros.

El primer grupo “dimensiones lingüísticas” involucra palabras funcionales e información gramatical. Por ejemplo, artículos, negación, verbos auxiliares. A continuación Tabla 2.13. muestra una lista completa de las categorías lingüísticas, así como algunos ejemplos de palabras que pertenecen a cada categoría.

Tabla 2.13. Categorías LIWC 2007 de procesos lingüísticos.

Categorías	Abreviación	Ejemplos
Conteo de palabras	wc	--
Palabras por oración	wps	--
% de palabras capturadas	dic	--
% de palabras > seis letras	sixltr	--
Total de palabras funcionales	Funcnt	--
Total de pronombres	Totpron	Yo, tu, mi
Pronombres personales	Pronper	Se, tú, yo
Primera persona del singular	Yo	Yo, mi, mis
Primera persona del plural	Nosotros	Nosotros, nos, nuestro
Segunda persona	Tuutd	Tu, usted
Tercera persona del singular	Ellella	Ella, él, su
Tercera persona del plural	Ellos	Ellos, su, sus
Pronombres impersonales	Pronimp	Todos, varios, aquellos
Artículos	Articulo	El, la, los, las
Verbos	Verbos	Irá, caminaré, verás
Verbos auxiliares	VerbAux	Pudiera, pudiésemos, pudimos
Tiempo pasado	Pasado	Procesé, quitó, recordé
Tiempo presente	Present	Abrimos, aceptas, preparas
Tiempo futuro	Futuro	Produciré, reunirá, tendrás
Adverbios	Adverb	Pronto, recientemente, siempre
Preposiciones	Prepos	Según, tras, por
Conjunciones	Conjunc	Tanto, y, excepto

Categorías	Abreviación	Ejemplos
Negaciones	Negacio	Jamás, no, nunca
Cuantificadores	Cuatif	Sección, suficiente, varias
Números	Numeros	Seis, once, mil
Palabras ofensivas	Maldec	Subnormal, tonta, pendejo
Verbos en primera persona del singular	VerbYo	Tuve, veo, contaba
Verbos en segunda persona del singular	VerbTu	Contabas, dices, estarás
Verbos en primera persona del plural	VerbNos	Contaremos, decimos, fuimos
Verbos tercera persona del singular	VerbosEl	Contó, era, fue
Verbos tercera persona del plural	VerbELLOS	Dirían, creen, saben
Subjuntivo	Subjuntiv	Creyéramos, descubráis, estemos
segunda persona del plural	VosUtds	Les, vosotros, ustedes
Formal	Formal	Usted, ustedes
Informal	Informal	Tu, tú, tus, tuya
Verbos Segunda persona del plural (España)	VerbVos	Sabéis, vengáis, creéis

El segundo grupo “procesos psicológicos” es capaz de estimar emociones positivas, emociones negativas, procesos sociales, procesos cognitivos, entre otros. Dentro de esta dimensión, por ejemplo, la emoción o cuestiones afectivas se basaron en palabras de varias fuentes tales como PANAS (Watson, Clark, & Tellegen, 1988) y el tesoro de Roget, siendo posteriormente evaluado por grupos de tres jueces que trabajaron de forma independiente. Las categorías que componen la segunda dimensión, procesos psicológicos, se muestra en Tabla 2.14..

Tabla 2.14. Categorías LIWC 2007 de Procesos psicológicos.

Categorías	Abreviación	Ejemplos
Procesos sociales	Social	Hablar, asiste, amigable
Familia	Familia	Mamá, papá, prima
Amigos	Amigos	Amiga, camarada, carnal
Humanos	Humanos	Ciudadano, dama, gente

Capítulo 2. Estado del arte

Categorías	Abreviación	Ejemplos
Procesos afectivos	Afect	Gilipollas, herir, llorar
Emociones positivas	EmoPos	Interesante, listo, sereno
Emociones negativas	EmoNeg	Agobia, burla, destruye
Ansiedad	Ansiedad	Nervioso, miedo, tenso
Enojo	Enfado	Odiar, matar, cabrón
Tristeza	Triste	Luto, llorar, tristeza
Procesos cognitivos	MecCog	Análisis, saber, tentar
Entendimiento	Insight	Sentimiento, opina, percata
Causa y efecto	Causa	Solución, declaró, forzar
Discrepancia	Discrep	Supondrá, debería, podría
Tentativos	Tentat	Supone, quizá, chance
Certeza	Certeza	Todo, verdad, total
Inhibiciones	Inhib	Abstención, dejar, impedir
Inclusive	Incl	Adentro, con, incluyendo
Exclusive	Excl	Obstante, pero, sin
Procesos sensoriales y perceptuales	Percept	Sonido, tener, ver
Ver	Ver	Testimonio, ver, vigilo
Escuchar	Oir	Voces, sonido, ruido
Sentir	Sentir	Agarra, besar, caricia
Procesos biológicos	Biolog	Agita, bulimia, gripa
Cuerpo	Cuerpo	Antebrazo, apéndice, esqueleto
Salud	Salud	Ceguera, doctor, inflamación
Sexual	Sexual	Condón, enrollaron, polvo
Ingerir	Ingerir	Consumes, desayunar, golosina
Relatividad	Relativ	Continúa, descenso, meses
Movimiento	Movim	Aprisa, bailar, caminar
Espacio	Espacio	Debajo, encerrar, interior
Tiempo	Tiempo	Día, empieza, hasta

El tercer grupo “asuntos personales” implica categorías de palabras relacionadas con asuntos personales intrínsecas a la condición humana. Por ejemplo, trabajo, pasatiempos, religión, por mencionar unos pocos. A continuación, en Tabla 2.15. se presenta la lista completa de categorías relacionadas a este grupo, así como algunos ejemplos.

Tabla 2.15. Categorías LIWC 2007 de Asuntos personales.

Categorías	Abreviación	Ejemplos
Trabajo	Trabajo	Eficaz, fuerza, obra
Logro	Logro	Elevar, ganar, logro
Pasatiempos	Placer	Entrena, fútbol, jazz
Hogar	Hogar	Estufa, hipoteca, tele
Dinero y asuntos financieros	Dinero	Euro, franquicia, pagar
Religión	Relig	Evangelio, iglesia, mandamiento
Muerte	Muerte	Fallecimiento, guerra, morir

El cuarto grupo “categorías del lenguaje oral”, se incluyó en LIWC con el objetivo de tener en cuenta categorías relacionadas al lenguaje hablado. Aunque LIWC no ha sido diseñado para el lenguaje oral, (Francis & Pennebaker, 1993) han encontrado que LIWC es útil en análisis de conversaciones y entrevistas. Esta dimensión contiene tres categorías paralingüísticas: asentir, no-fluidez, y relleno. Tabla 2.16. muestra estas tres categorías junto con algunos ejemplos.

Tabla 2.16. Categorías LIWC 2007 de Lenguaje oral.

Categorías	Abreviación	Ejemplos
Asentir	Asentir	Acepta, efectivamente, genial
no fluidez	Nofluen	Ah, ajá, eh, em
Relleno	Relleno	Osea, uf

Finalmente, LIWC incluye 12 categorías de puntuación tales como punto, coma, punto y coma, etc. A continuación, en la Tabla 2.17. las categorías que componen este grupo son presentadas. Para paréntesis, LIWC cuenta pares de

Capítulo 2. Estado del arte

paréntesis (es decir, el paréntesis que abre (“y el paréntesis que cierra “)” se cuenta como uno). Además, la categoría “otra puntuación” incluye todas las marcas de puntuación no contabilizados en las otras categorías de puntuación, así como caracteres ASCII desde 33-47, 58-64, 91-96 y 123-126.

Tabla 2.17. Categorías LIWC de signos de puntuación.

Categorías	Abreviación	Ejemplos
Total puntuación	Allpunc	--
Punto	Period	.
Comas	Comma	,
Dos puntos	Colon	:
Punto y coma	Semic	;
Signos de interrogación	Qmark	¿?
Signos de exclamación	Exclam	¡!
Guiones	Dash	- - — —
Comillas	Quote	“”
Apóstrofes	Apostro	' '
Paréntesis	Parenth	()
Otras puntuaciones	Otherp	+ / < @

2.7. Objetivos de la tesis doctoral

2.7.1. Motivación

Las opiniones siempre han sido parte importante para el humano en la toma de decisiones, desde comprar un producto, votar por un partido político, mejorar los productos o servicios, etc. Antes de la Web 2.0, cuando un usuario necesitaba tomar una decisión para adquirir un producto o servicio normalmente se basaba en las opiniones de sus amigos o familiares; cuando una organización necesitaba saber las opiniones del público en general acerca de sus productos y servicios, se llevaba a cabo encuestas y sondeos de opinión. Sin embargo, con la Web 2.0, especialmente con el crecimiento drástico del contenido generado por el usuario en los últimos años, el mundo se ha transformado. El drástico crecimiento de

Internet en la sociedad ha cambiado sustancialmente la forma en que la gente expresa sus opiniones. Ellos pueden ahora postear opiniones de productos en sitios comerciales y expresar sus puntos de vista sobre casi cualquier cosa en los foros de Internet y blogs, lo cual es conocido como “contenido generado por el usuario”.

Las redes sociales tales como Twitter y Facebook son otro medio importante donde actualmente las personas suelen expresar una gran cantidad de contenido subjetivo en tiempo real. Los usuarios expresan todo tipo de cuestiones desde experiencias con productos que adquieren, opiniones sobre política, deportes, etc.

El boca a boca online representa nuevas fuentes de información medibles. Hoy en día si una persona quiere comprar un producto, no necesita limitarse a las opiniones de familiares y amigos debido a que hay muchas opiniones de productos en la Web. Por otro lado, para las empresas ya no es necesario llevar a cabo encuestas, organizar grupos de enfoque con el fin de conocer las opiniones de los consumidores, ya que el contenido generado por los usuarios en la Web ya les da esta información.

Sin embargo, la búsqueda de opiniones y el seguimiento en la Web es una tarea formidable, debido a que hay una gran cantidad de fuentes y cada fuente contiene un gran volumen de texto dogmático. Por lo que para un humano es difícil encontrar las fuentes pertinentes, leerlas, resumirlas y organizarlas de forma utilizables. Por lo tanto, esta es la primera justificación para llevar a cabo el análisis de las opiniones. Las personas requieren de sistemas automáticos que les permitan acceder a información útil para tomar una decisión.

Por otro lado, uno de los principales problemas en el análisis de las opiniones es el lenguaje figurado, dado que este puede crear un significado totalmente diferente de la opinión. Es decir, si una persona provee un comentario positivo acerca de un producto, pero está siendo irónico, es probable, que en realidad este dando un comentario negativo acerca de ese producto. Por lo que la detección automática del lenguaje figurado se ha convertido en un reto en los últimos años. Esta es otra de las motivaciones planteadas en el presente estudio.

Actualmente existe diversos esfuerzos para el análisis de sentimientos y el lenguaje figurado. Sin embargo, la mayoría han sido enfocados al idioma inglés y

Capítulo 2. Estado del arte

solo algunos pocos se han enfocado en otros idiomas como el español, el cual es el tercer idioma más importante de internet. Este es la tercera motivación de la presente tesis doctoral, la cual esta principalmente enfocada en el idioma español.

2.7.2. Objetivos

El objetivo principal de esta tesis es la detección de patrones psicolingüísticos para el análisis de lenguaje subjetivo en español. Para conseguir este objetivo se definen las siguientes tareas:

- Diseño de un método para la detección de patrones psicolingüísticos para el análisis de sentimientos.
- Diseño de un método para la detección de patrones psicolingüísticos para la detección de la sátira.
- Validación del método para el análisis de sentimientos en diversos dominios como el turístico y películas.
- Validación del método para la detección automática de la sátira en el dominio de noticias.

2.7.3. Metodología

La metodología a seguir se basa en el estudio del estado del arte, diseño e implementación de los métodos para el análisis del lenguaje subjetivo y su validación en un dominio de aplicación.

- Estudio del estado del arte: Estudio de las tecnologías del PLN, análisis de sentimientos y lenguaje subjetivo. Específicamente, los diferentes niveles de procesamiento, principales enfoques del análisis de sentimientos, niveles de procesamiento de la opinión, bases de conocimiento, recursos lingüísticos disponibles y principales técnicas para la detección del lenguaje figurado.
- Diseño e implementación de métodos: Diseño e implementación de un método para el análisis de sentimientos y detección de la sátira basados en características psicolingüísticas.
- Validación de la propuesta a través de la definición de escenarios de validación en diversos dominios. Concretamente, el método de análisis de

sentimientos se aplicará al dominio turístico y de películas; y el método de detección de la sátira se aplicará en el dominio de noticias en redes sociales.

2.8. Conclusiones

En la primera sección se dio una introducción al lenguaje subjetivo. Posteriormente, en la segunda sección se incluyó un estudio acerca del procesamiento de lenguaje. Se proporciona un análisis de las diferentes definiciones propuestas por diferentes autores. Finalmente, en esta sección se presentaron los diferentes niveles lingüísticos que cubre el PLN.

En la tercera sección, se analizaron los principales enfoques y niveles de procesamiento del análisis de sentimientos. Como primer punto, se llevó a cabo un estudio de las diferentes definiciones del término “análisis de sentimientos”. Segundo, se describe la evolución histórica de esta área de investigación desde sus orígenes en 1979 hasta los últimos años. En tercer lugar, se presentó un análisis de los diferentes niveles de análisis de la opinión. En el cuarto punto, se presentaron los recursos lingüísticos existentes. Posteriormente, se presentaron los principales enfoques sobre los cuales se basan la mayoría de estudios del análisis de sentimientos y las principales técnicas para abordarlos. Finalmente, se llevó a cabo un estudio de los métodos actuales para el análisis de sentimientos y se compararon con el método propuesto en esta tesis doctoral.

En la cuarta sección, se dio una introducción al lenguaje subjetivo, concretamente sobre la ironía el sarcasmo y la sátira. Entonces, dado que este estudio está principalmente centrado en la sátira, se llevó a cabo un análisis de las propuestas actuales y se hizo una comparación con el método presentado en esta tesis. En la quinta sección, se analizó la importancia de la psicolingüística en el lenguaje subjetivo. Además, se presentó la herramienta LIWC la cual es una base importante para la realización del presente estudio.

Finalmente, la última sección hizo hincapié en el problema que se resuelve en esta tesis doctoral, presentando la motivación, los objetivos establecidos y la metodología a seguir para lograr la correcta ejecución del estudio propuesto.

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

3.1. Introducción

El análisis de sentimientos se ha convertido en un tema muy popular que permite entender la opinión pública, a partir del análisis de comentarios y opiniones expresados en Internet. Sin embargo, uno de los principales problemas en la clasificación de sentimientos es la existencia de muchas reglas conceptuales que rigen la expresión lingüística de los sentimientos. La psicología humana, que se relaciona con los aspectos sociales, culturales, personales y biológicos, pueden ser características importantes a considerar en el análisis de sentimientos. Para ello, LIWC es una herramienta útil, ya que permite la extracción de características psicológicas y lingüísticas a partir de texto en lenguaje natural.

Por otro lado, el lenguaje figurado tal como la ironía, sarcasmo, y sátira ha sido considerado un importante tema en las redes sociales como Twitter⁷. Por tanto, el lenguaje figurado ha sido fundamental para varios campos de investigación tales como el análisis de sentimientos, desde que este lenguaje puede invertir la polaridad de un texto. Diversos enfoques han sido propuestos en los últimos años para la detección del sarcasmo e ironía. Sin embargo, hay una carencia de enfoques para la detección de la sátira. La sátira consiste en el uso del humor y la ironía para criticar y ridiculizar a algo o a alguien. Aunque algunas características lingüísticas y emocionales han sido utilizadas en la detección del lenguaje figurado, no se ha llevado a cabo una exploración a fondo de estas características. Por ello, se seleccionó la herramienta LIWC, la cual ha sido utilizada con éxito para explorar el estilo lingüístico de los contenidos en medios sociales.

⁷ <https://twitter.com>

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

En este trabajo de tesis doctoral, se propone un método para el análisis de sentimientos y para la detección de la sátira. Este método clasifica una opinión como positiva, negativa, neutra, muy positiva o muy negativa. Además, permite determinar qué características son más discriminantes en cada uno de los problemas de clasificación, es decir, para la clasificación de sentimientos y la clasificación de la sátira. El capítulo explica en profundidad el método propuesto, que combina técnicas de aprendizaje automático y PLN.

El sistema recibe como entrada un conjunto de datos en lenguaje natural etiquetados. Posteriormente, los datos son analizados y pre-procesados con el objetivo de normalizar los datos. Para entonces, llevar a cabo la extracción de características psicolingüísticas del texto. Finalmente, las características previamente obtenidas son utilizadas para entrenar un algoritmo de clasificación. Como resultado se obtiene un modelo que permite clasificar datos nuevos, en este caso opiniones o tweets.

3.2. Descripción del problema y objetivos específicos

Como se ha mencionado en la sección anterior, el objetivo principal de esta tesis doctoral es la detección de patrones psicolingüísticos para el análisis de lenguaje subjetivo en español. Específicamente, se propone el desarrollo de un método para el análisis de sentimientos y la detección de textos satíricos y no satíricos. Por lo tanto, los siguientes puntos serán abordados en este trabajo.

- Determinar qué tan relevantes son las características psicolingüísticas en la clasificación de sentimientos.
- Determinar qué tan relevantes son las características psicolingüísticas en la clasificación de textos satíricos.
- Identificar cuáles son las características más relevantes para el análisis de sentimientos.
- Identificar cuáles son las características más relevantes para la detección de la sátira.

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

- Identificar si existen características psicolingüísticas específicas que permitan diferenciar los textos satíricos de diferentes culturas.

3.3. Método

El método propuesto se basa en un enfoque de máquina de aprendizaje, el cual recibe como entrada un corpus de texto en lenguaje natural que puede estar compuesto de textos extraídos desde blogs o redes sociales como Twitter. El método se divide en 3 etapas principales: 1) pre-procesamiento del texto a través de técnicas PLN; 2) extracción de las características psicolingüísticas a través de LIWC; y 3) entrenamiento de algoritmos de aprendizaje.

El método propuesto es adaptable a diversos problemas de clasificación e idiomas. En esta tesis doctoral ha sido utilizado para la clasificación de opiniones de diversos blogs y para la detección de la sátira en Twitter. Ambos, enfocados en el idioma español.

A continuación, la Figura 3.1 muestra los módulos del enfoque propuesto. Como se puede observar este consta de tres módulos principales: 1) pre-procesamiento del corpus recibido, el cual consiste en eliminar información innecesaria y en corregir el texto para su posterior procesamiento; 2) extracción de características psicolingüísticas a través de LIWC; y 3) entrenamiento de algoritmos de clasificación con la finalidad de obtener un modelo predictivo que permita clasificar nueva información.

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

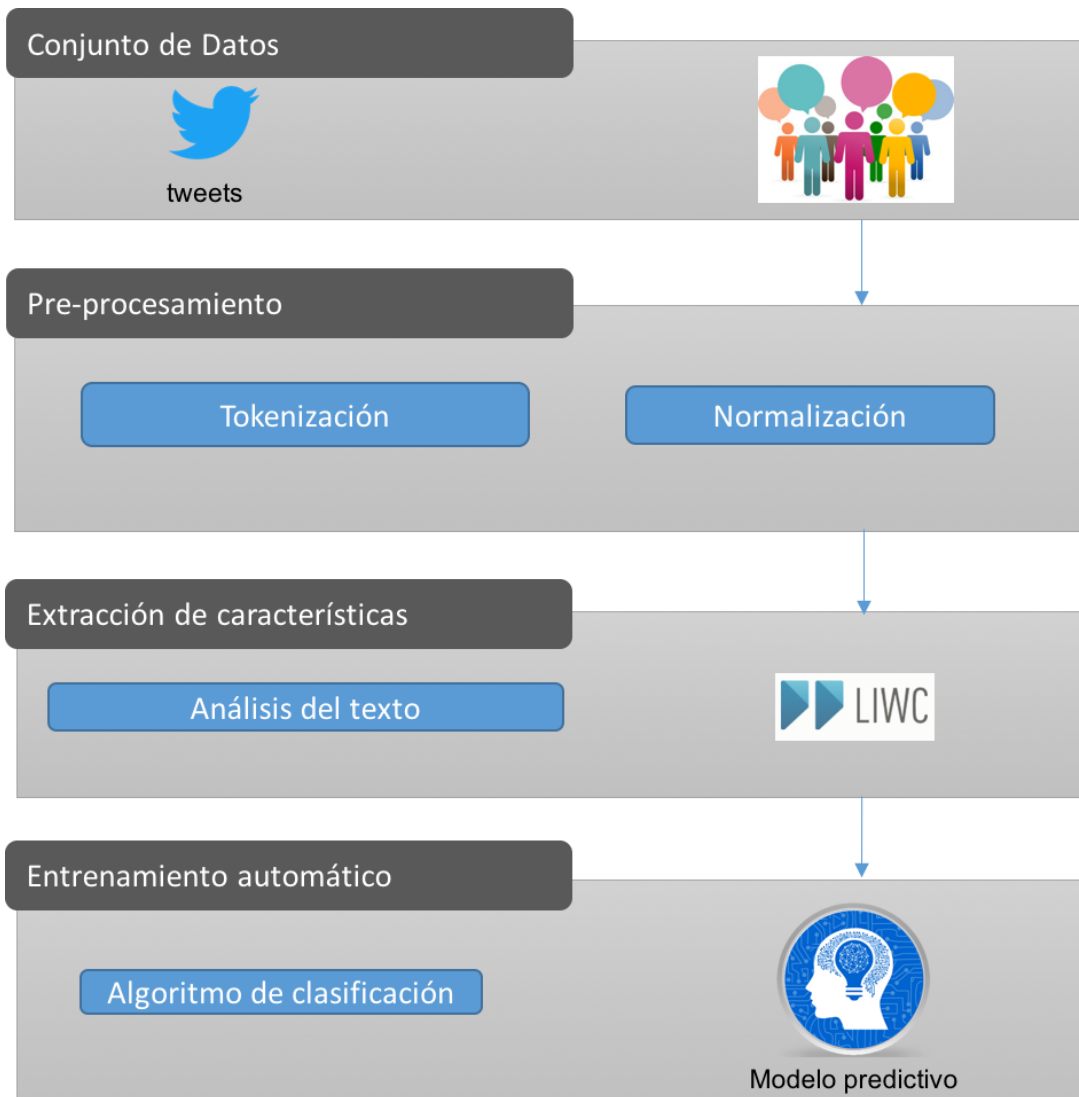


Figura 3.1. Enfoque propuesto para la clasificación de sentimientos y detección de la sátira.

3.3.1. Conjunto de datos

En el caso del análisis de sentimientos, el corpus consiste de un conjunto de documentos en este caso opiniones, las cuales pueden estar etiquetados como: positivo, negativo, neutro, muy positivo, muy negativo. A continuación, se presentan dos ejemplos de opiniones extraídas de diferentes corpus de opiniones. Como se puede observar, la Figura 3.2 presenta una opinión de películas extraída de la página Muchocine, clasificada con un valor de 5, es decir, como “muy positiva”. Por otro lado,

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

la Figura 3.3, muestra una opinión de TripAdvisor, específicamente de alojamiento con un valor de 2, es decir, negativa.

```
<review author="Caro" title="Transamerica" rank="5" maxRank="5" source="muchocine">
  <summary>Gran actuación de Felicity Huffman</summary>
  <body>No tengo palabras para describir esta película, simplemente es GENIAL. Felicity hace un gran trabajo y demuestra lo gran actriz que es y por qué se ha llevado todos esos premios (el Oscar lo tiene que ganar), yo sólo la había visto en Mujeres Desesperadas y el cambio de registro me ha encantado y si ves la película en VO se aprecia mucho más. Punto a favor: la misma que la dobla en la serie se ha encargado de ponerle la voz en la película, pero no la pone lo suficientemente grave, al principio un poquito, pero luego se le olvida. La primera media hora resulta un poquito lenta, pero una vez se van desarrollando las historias de los personajes te quedas enganchada en ese mundo. Decir también que me gusta más el tagline americano que el que han puesto en España; Life is more than the sum of its parts. Bree, antes Stan, está a punto de conseguir la autorización para someterse a su operación de cambio de sexo, pero recibe la llamada de un chico que dice ser su hijo. Su psiquiatra le dice que debe conocer al chico antes de firmar la autorización. Bree a regañadientes acepta ir a Nueva York a sacar al chico de la cárcel, aunque la fianza es de 1$, Toby no es ningún angelito, con 17 años trapichea con drogas y ha sido detenido por prostitución. El chaval le convence para que le lleve a Los Ángeles, donde se hará actor porno, y compran un coche. El viaje está lleno de incidentes y mentiras. Bree no le ha dicho que es su verdadero padre y menos que en realidad es un hombre (le ha dicho que es de la Iglesia del Padre Potencial) y Toby acaba descubriendo su secreto. Aunque al principio reacciona mal, termina aceptando a Bree. Un hippie les roba el coche y los deja tirados en la carretera sin más ropa que la que llevaban puesta. En un bar, Bree liga con un hombre que los deja en Phoenix, en casa de los padres de Bree. La madre reacciona fatal al ver a su hija, pero el resto de la familia no parece tener ningún problema (sobre todo la hermana). Cuando les confiesa que Toby es su hijo, la madre empieza a tratarlo estupendamente, e incluso quiere que se quede a vivir con ellos como condición a pagarle el billete de avión a Bree para que llegue a tiempo a su operación. Bree, por fin le cuenta a Toby quien es en realidad y reacciona fatal, llegando a darle un puñetazo en un ojo. Ya en Los Ángeles, Bree se opera, pero ese día no es el más feliz de su vida, echa de menos a Toby. Lo mejor sin duda es la gran actuación de Felicity Huffman. Muchas de las frases "mi cuerpo puede estar aún en proceso, pero no hay nada malo con mi alma". La transexual rubia y sus amigos. La madre de Bree</body>
</review>
```

Figura 3.2. Opinión muy positiva.

```
<review id=62 rank="2" topic="Alojamiento">
  <content>
    Estuvimos alojados en este hotel dos días en el mes de septiembre ,era un destino de paso y el precio interesante para un hotel de 5 estrellas ,es un Hotel ,bastante impersonal , nada acogedor , la sala de desayunos en el ultimo piso no tiene ningun atractivo y el restaurante que está al lado tampoco seduce a pesar de las vistas. Las habitaciones y sobre todo los cuartos de baño son muy oscuros ,el calor de los focos alojenos en el baño llegan a agobiar. Pero lo peor es que no hay ninguna seguridad de la ropa que tengas en tu habitación ,hay una pequeña caja fuerte para las joyas , pero ,la persona que limpia la habitación puede robarte descaradamente la ropa con total impunidad , como ha sido en nuestro caso. El hotel no se hace responsable de nada porque no conoce quienes son los empleados de la limpieza , no hay gobernantas ni responsables , ni nada de nada . Si te alojas allí mejor ir sin nada.
  </content>
</review>
```

Figura 3.3. Opinión negativa.

Por otro lado, con respecto a la sátira, el corpus requerido debe contener un conjunto de tweets etiquetados como satíricos o no-satíricos. A continuación, en la Figura 3.4 se presenta un ejemplo de tweet clasificado como no-satírico, y en la Figura 3.5 se muestra un ejemplo de un tweet clasificado como satírico.

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

```
<tweet id="782512990012702720" fuente="@el_país" clasificación="no-satírica">
  <content>
    Roger Waters, exlíder de Pink Floyd lanza un duro mensaje a Peña Nieto en México:
    "Renuncia ya"
  </content>
</tweet>
```

Figura 3.4. Tweet no-satírico.

```
<tweet id="782618385712570368" fuente="@eldeforma" clasificación="satírica">
  <content>
    Confirman que mensaje de Roger Waters llegó al corazón de Peña. México ahora es
    un mejor país. Se acabó la violencia y hay opulencia
  </content>
</tweet>
```

Figura 3.5. Tweet satírico.

3.3.2. Pre-procesamiento del texto

La primera etapa del método de análisis de sentimientos consiste en el pre-procesamiento de cada una de las opiniones. El objetivo es preparar los textos eliminando aquellos elementos que se consideran superfluos. Las dos fases básicas que se realizan son: 1) Tokenización y 2) Normalización del texto. El pre-procesamiento del texto dependerá en gran medida de la fuente de donde sea obtenido. Por ejemplo, para los textos obtenidos de Twitter, es necesario llevar a cabo un pre-procesamiento más detallado, debido a que el lenguaje utilizado en esta red social se caracteriza principalmente por un estilo de escritura informal y libre en el que las abreviaciones, repetición de caracteres y el uso indebido de letras mayúsculas se utilizan comúnmente para ahorrar caracteres o hacer énfasis en los tweets. Los usuarios de twitter suelen cometer errores de ortografía y gramática, así como crear textos cortos que son difíciles de analizar (Cotelo, Cruz, Troyano, & Ortega, 2015). Estos problemas de variación léxica se abordaron en base al trabajo presentado en (Vilares, Alonso, & Vilares, 2013). En las siguientes secciones se explica de manera detallada, cual es el procedimiento para el procesamiento de texto obtenido desde sitios web de opiniones y desde redes sociales como twitter.

3.3.2.1. Texto desde sitios web de opiniones

Para la fase de tokenización se utilizó el tokenizer de Stanford NLP (Manning et al., 2014), el cual divide el texto en una secuencia de *tokens*, los cuales corresponden a “palabras”. La Figura 3.6 muestra un extracto de una opinión, con los resultados obtenidos al llevar a cabo este proceso.

No	tengo	palabras	para	describir	esta	película	,	simplemente	es	genial	.
----	-------	----------	------	-----------	------	----------	---	-------------	----	--------	---

Figura 3.6. Proceso de tokenización.

La normalización de textos consiste en homogenizar todas las opiniones. Para ello, se considera la extensión de abreviaturas y acrónimos, eliminación de palabras vacías, corrección de errores ortográficos, por mencionar algunos. En este caso, solo se consideraron tareas como extensión de abreviaturas y corrección de errores ortográficos. Debido a que la jerga online es universal y comúnmente utilizada en redes sociales, blogs y post online, etc., se obtuvo una lista con las principales abreviaciones y acrónimos, la cual fue construida a partir de fuentes tales como netLingo⁸. La Tabla 3.1. presenta algunos términos contenidos en el diccionario. Finalmente, el texto se analiza utilizando el diccionario Hunspell para corregir los errores de ortografía.

⁸ <http://www.netlingo.com/>

Tabla 3.1. Términos del diccionario.

Abreviación	Termino correcto
pq	Por qué/Porque
spro	Espero
tb	También
xq	Porque/Por qué
Ad+	Además
iwal	igual
aki	Aquí

Por ejemplo, considerando el ejemplo de la Figura 3.7, las palabras “pq” sería corregida y reemplazada por “porque” y la palabra “dolo” se reemplaza por “solo”.

Decidimos venir a esta cala un día de mucho viento y fue un acierto, está más resguardada que el Playazo (donde estábamos y nos fuimos por como pegaba). El mar estaba algo revuelto pero aún así se puede hacer snorkel y disfrutar. El camino se hace largo pq es de piedras pero se puede ir en un coche normal, dolo hace falta paciencia y ganas de disfrutar de las vacaciones.

Figura 3.7. Ejemplo de opinión con abreviaciones y errores ortográficos.

3.3.2.2. Texto de redes sociales

El primer paso para el pre-procesamiento de textos de redes sociales consiste en un proceso de tokenización de todos los tweets mediante el uso de la herramienta Twokenize (Owoputi et al., 2013), un tokenizador específico de twitter que no solo divide el texto en *tokens*, sino que también permite la detección de entidades de interés tales como menciones, etiquetas (*hashtags*) y URLs. Por ejemplo, tras aplicar la técnica de tokenización al tweet presentado en la Figura 3.8, “#Entérate Cesan a maestro de secundaria en #Puebla que usaba sellos con frases despectivas <https://t.co/xO4030n1LZ> <https://t.co/liABVQWhmp>”, los resultados obtenidos serían los presentados en la Tabla 3.2. y Figura 3.9.

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

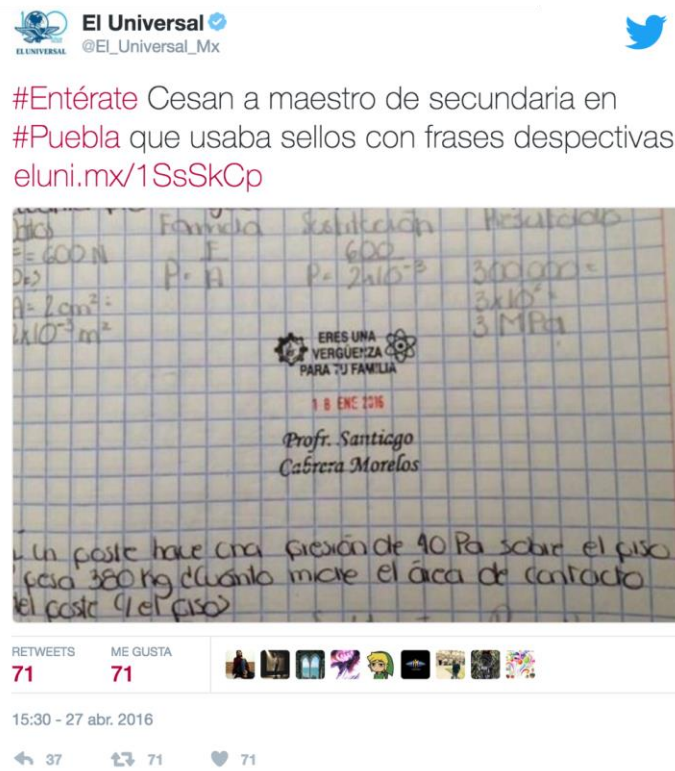


Figura 3.8. Ejemplo de tweet.

La Tabla 3.2. muestra los resultados del proceso de tokenización para un tweet, utilizando la herramienta Twokenize. Como se puede observar, esta sentencia es dividida en 16 *tokens*.

Tabla 3.2. Tokens tweet.

Id	palabra
1	#Entérate
2	Cesan
3	a
4	maestro
5	de
6	secundaria
7	en

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

Id	palabra
8	#Puebla
9	que
10	usaba
11	sellos
12	con
13	frases
14	despectivas
15	https://t.co/x04030n1LZ
16	https://t.co/liABVQWhmp

La Figura 3.9 presenta la detección de elementos tales como menciones, hashtags y URLs. Para el ejemplo presentado, Twokenize detecta dos hashtags: “#Entérate” y “#Puebla”; y dos URLs: “https://t.co/x04030n1LZ” y “https://t.co/liABVQWhmp”.

#Entérate Cesan a maestro de secundaria en #Puebla que usaba sellos con frases despectivas https://t.co/x04030n1LZ https://t.co/liABVQWhmp	Mención Hashtag URL
--	---------------------------

Figura 3.9. Detección de elementos como menciones, hashtag y URL.

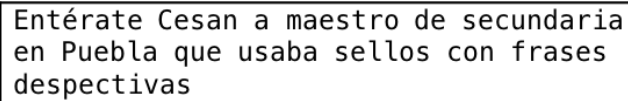
Con el objetivo de reducir el ruido del conjunto de datos se eliminan los siguientes elementos que previamente fueron detectados por Twokenize:

1. Eliminar menciones y respuestas a tweets de usuarios que se representan mediante cadenas que empiezan por @.
2. Eliminar URL, es decir, cadenas que empiezan con http://
3. Eliminar el carácter “#” de las etiquetas, ya que frecuentemente sólo el resto de la cadena forma una palabra legible que contribuye a una mejor comprensión.

Siguiendo con el ejemplo presentado anteriormente, Twokenize detecta 2 hashtags (#Entérate y #Puebla) y 2 URL (<https://t.co/x04030n1LZ> y

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

<https://t.co/liABVQWhmp>), por lo que las URLs son eliminadas y únicamente el “#” es eliminado. El resultado obtenido se presenta en la Figura 3.10.



Entérate Cesan a maestro de secundaria
en Puebla que usaba sellos con frases
despectivas

Figura 3.10. Texto sin URL, hashtags y menciones.

El uso de abreviaturas es común en los tweets dada la limitación de 140 caracteres. En este sentido, la segunda etapa consiste en sustituir elementos por sus expansiones. Por ejemplo, “también” en lugar de “tb”, “por favor” en lugar de “xfa”. Esta tarea se llevó a cabo utilizando el diccionario mencionado en la sección 3.3.2.1, el cual se basa en fuentes tales como NetLingo (Jansen, 2014).

Finalmente, los errores de ortografía son corregidos utilizando Hunspell (Németh, 2005), un corrector ortográfico de código abierto y analizador morfológico basado en MySpell, diseñado para los idiomas con una morfología rica (Németh et al., 2004). Por ejemplo, "palabar", "daibetes", y "compelto" son corregidos a "palabra", "diabetes", y "completo", respectivamente.

Como se mencionó anteriormente, los recursos base utilizados para esta tarea son Hunspell, Twokenize y NetLingo. Estos recursos fueron seleccionados debido a que han sido ampliamente utilizados para pre-procesar tweets en otros trabajos de investigación tales como: (Cotelo et al., 2015) (Vilares et al., 2013) (Anta, Chiroque, Morere, & Santos, 2013) (Farhan Hassan Khan, Bashir, & Qamar, 2014) (Cordobés et al., 2014) (Wiley, Jin, Hristidis, & Esterling, 2014) (F. H. Khan, Qamar, & Javed, 2014).

3.3.3. Extracción de características

La segunda fase consiste en la extracción de características. Esta fase es muy importante ya que permite obtener vectores de características que posteriormente serán utilizados para entrenar algoritmos de aprendizaje supervisado. La exactitud de los modelos generados en el aprendizaje automático dependerá de la construcción del vector que contenga importantes características para cada uno de los problemas de

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

clasificación de texto planteados (clasificación de sentimientos y clasificación de textos satíricos).

LIWC se utilizó con el objetivo de obtener las características psicolingüísticas del texto y explorar que categorías LIWC pueden proporcionar patrones para la detección de la polaridad y la detección de la sátira. Como se mencionó en la sección 2.6, LIWC contiene un total de 72 categorías las cuales están clasificadas en 5 grupos principales: 1) procesos lingüísticos, 2) procesos psicológicos, 3) asuntos personales, 4) categorías del lenguaje oral y 5) signos de puntuación. En este trabajo de tesis doctoral se utilizó el diccionario LIWC 2007 en español. A continuación, la Figura 3.11 y Figura 3.12 muestran un extracto de este diccionario. La Figura 3.11 presenta las categorías junto con un valor que identifica a esa categoría. Mientras que la Figura 3.12 muestra algunas palabras del diccionario que pueden ser palabras completas o palabras raíces. Estas últimas se identifican con un *. Por ejemplo, la palabra “desdichad*” indica que se tomaran en cuenta palabras como “desdichado”, “desdichada”, “desdichados”, “desdichadas”.

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

1	Funct
2	TotPron
3	PronPer
4	Yo
5	Nosotros
6	TuUtd
7	ELlElla
8	Ellos
9	PronImp
10	Articulo
11	Verbos
12	VerbAux
13	Pasado
14	Present
15	Futuro
16	Adverb
17	Prepos
18	Conjunc
19	Negacio
20	Cuantif
21	Numeros
22	Maldec
23	verbYO
24	verbTU
25	verbNOS
26	verbosEL
27	verbELLOS
28	Subjuntiv
29	VosUtds
30	formal
31	informal
32	verbVos
121	Social
122	Familia
123	Amigos
124	Humanos
125	Afect
126	EmoPos
127	EmoNeg
128	Ansiedad
129	Enfado
130	Triste
131	MecCog

Figura 3.11. Extracto de categorías de LIWC 2007 en español.

Cada una de las palabras tienen especificado de uno a más números, lo cual significa que pueden pertenecer a más de una categoría. Por ejemplo, la palabra “desea” tiene relacionados 4 números (11, 14, 125 y 126), lo cual indica que esta palabra pertenece a 4 grupos de categorías de LIWC: verbos “verbos”, presente “present”, afectivo “afect” y emociones positivas “emopos”.

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

describís	11	14	121	131	
describiste*	11	13	121	131	
describo	11	14	121	131	
descubra*	11	14	28	131	132
descubráis	11	14	28	131	132
descubre*	11	14	131	132	
descubrí	11	13	131	132	133
descubría*	11	13	131	132	
descubriera*	11	13	28	131	132
descubriéramos	11	13	28	131	132
descubrieron	11	13	131	132	133
descubriese*	11	13	28	131	132
descubriésemos	11	13	28	131	132
descubrimos	11	13	14	131	132
descubrió	11	13	131	132	133
descubrir	11	131	132		
descubrirá*	11	15	131	132	
descubrire*	11	15	131	132	
descubriré*	11	15	131	132	
descubris	11	14	131	132	
descubriste*	11	13	131	132	
descubro	11	14	131	132	
descuento	258				
descuidad*	125				
descuidar	125	131	137		
desde	1	17			
desdichad*	125	127	128	130	
desea	11	14	125	126	
desea*	131	134			
deseá*	131	134			
deseaba*	11	13			
deseáis	11	131			
deseamos	11	13	14	125	126
desean	11	14	131		
deseando	131				
desear*	131				
deseara*	11	13	28	131	
deseará*	11	15	131		
deseáramos	11	13	28	131	
deseare*	11	15	131		
desearé*	11	15	131		
desearon	11	13	131		
deseas	11	14	131		
desease*	11	13	28	131	
deseásemos	11	13	28	131	
deseaste	131	134			
deseaste*	11	13	131		
desech*	125	127			
deseé	11	13	131	134	
desee*	11	14	28	131	
desempeña	11	14	254		
desempeñáis	11	14	254		
desempeñamos	11	13	14	254	
desempeñan	11	14	254		

Figura 3.12. Ejemplo de palabras contenidas en el diccionario de LIWC 2007 en español.

Con respecto al lenguaje figurado, las categorías LIWC se han utilizado en la clasificación del sarcasmo (Justo, Corcoran, Lukin, Walker, & Torres, 2014), la clasificación de la sátira (Rubin et al., 2016) (Skalicky & Crossley, 2015) y tareas similares, como la identificación de textos humorísticos (Mihalcea & Pulman, 2009). Sin embargo, en el caso de la detección de la sátira únicamente se han investigado características tales como emociones negativas, negaciones, cuantificadores, certeza,

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

tentativa, exclusión, inclusión, discrepancia, causalidad, pasado, presente y signos de puntuación. Los autores de los trabajos antes mencionados seleccionaron el conjunto de características basado en estudios previos en la literatura asumiendo que las características serían discriminantes en cuanto a la detección de la sátira. Por esta razón, en este trabajo se ha examinado la lista completa de las 72 categorías disponibles en LIWC con el fin de detectar cuáles de las características del lenguaje figurado mencionadas en la literatura y qué otras características pueden contribuir para la detección de la sátira.

En cuanto al análisis de sentimientos, LIWC ha sido utilizado únicamente como un lexicón basado en orientación semántica, el cual es construido a partir de 2 categorías (emociones positivas y emociones negativas) de las 72 categorías de LIWC (Hsu, See, & Wu, 2010) (Filho, Pardo, & Alusio, 2013) (Gonçalves, Araújo, Benevenuto, & Cha, 2013) (Hutto & Gilbert, 2014). Muy pocos trabajos han estudiado la importancia de este tipo de características en el análisis de sentimientos. Aunque desde los inicios de esta disciplina, los autores claman que existe una amplia relación entre la opinión expresada y el estado mental de las personas. Por lo tanto, la psicología humana, que se relaciona con aspectos sociales, culturales, personales, y biológicos pueden ser características importantes a considerar en el análisis de sentimientos, es por ello que en este trabajo se estudian las 72 categorías de LIWC, con el objetivo de determinar qué características son más discriminantes en la clasificación de sentimientos.

Con el objetivo de llevar a cabo el procesamiento del texto con LIWC, primero se convirtió cada corpus en archivos de texto plano, posteriormente cada opinión y tweet fueron analizados individualmente. LIWC procesa el texto en tres etapas: 1) detección de palabras contenidas en el diccionario; 2) asociación de las palabras del texto con las categorías LIWC; y 3) cálculo del porcentaje de palabras contenidas en cada categoría.

En la primera etapa, el módulo de análisis de LIWC compara cada una de las palabras contenidas en la opinión o el tweet con el diccionario definido por el usuario, en este caso se seleccionó el diccionario de LIWC 2007 para el español. A

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

continuación, se presentan dos ejemplos uno de una opinión de un sitio Web y un tweet, los cuales son procesados con LIWC. Considerando el extracto de la opinión presentada en la Figura 3.13, LIWC encuentra 57 palabras de las 63 palabras contenidas en la opinión (ver Figura 3.14).

La cinta no derrocha acción como se espera y tiene una profundidad de guion que no me pareció necesaria. Por cierto, la escena de la tortura es un poco patética. Resumiendo, película de poca acción sobre un agente secreto que empieza en esto de las licencias para matar. Casi mejor nos esperamos a su versión en DVD y la disfrutamos un día lluvioso.

Figura 3.13. Extracto de una opinión analizada por LIWC.

La cinta no derrocha acción como se espera y tiene una profundidad de guion que no me pareció necesaria. Por cierto, la escena de la tortura es un poco patética. Resumiendo, película de poca acción sobre un agente secreto que empieza en esto de las licencias para matar. Casi mejor nos esperamos a su versión en DVD y la disfrutamos un día lluvioso.

Figura 3.14. Palabras de la opinión detectadas con LIWC.

Con respecto al ejemplo del tweet presentado en la Figura 3.15 “TV Azteca lanza “Plim”, la plataforma digital donde podrás volver a ver tus programas favoritos de la televisora”. LIWC encuentra 15 de las 18 palabras del texto (ver Figura 3.16).

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira



Figura 3.15. Tweet analizado con LIWC.

TV Azteca **lanza** "Plim", **la plataforma** digital **donde podrás volver a ver tus programas favoritos de la televisora**

Figura 3.16. Palabras encontradas con el diccionario LIWC.

En la segunda etapa, el diccionario identifica los términos que se encuentran asociadas con categorías de LIWC. Cabe mencionar que el diccionario de LIWC contiene palabras y *word stems*. Estos últimos finalizan con un (*) en el diccionario, lo cual indica que LIWC ignora todas las letras posteriores. Por ejemplo, la palabra "favorit*" incluirá favorito", "favorita", "favoritos", "favoritas". La Figura 3.17 presenta un extracto de los resultados de la opinión presentada en la Figura 3.13. Como se puede observar, LIWC detecta 4 artículos "Articulo" (la, una, un y las), 3 adverbios "Adverb" (no, como y casi), 1 negación "Negacio" (no), 6 palabras afectivas "Afect" (cierto, tortura, patética, matar, mejor y disfrutamos), 3 emociones positivas "EmoPos" (cierto, mejor, disfrutamos), 3 emociones negativas "EmoNeg" (tortura,

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

patética, y matar), 2 palabras relacionadas con enfado “Enfado” (tortura y matar), 1 palabra relacionada a tristeza “Triste” (patética) y 12 palabras relacionadas con procesos cognitivos “MecCog” (como, espera, y, tiene, que, pareció, cierto, secreto, esto, casi y esperamos).

	TotPron	Articulo	EmoPos	MecCog
TV				
Azteca				
lanza				x
plim				
la	x	x		
plataforma				
digital				
donde				
podrás				
volver				
a				
ver				x
tus	x			
programas				
favoritos			x	
de				
la	x	x		
televisora				

Figura 3.17. Palabras categorizadas.

La Figura 3.18 presenta un extracto de los resultados del ejemplo del tweet presentado en la Figura 3.15. Como se puede observar, las palabras se clasifican de la siguiente manera: “lanza” como un proceso cognitivo (“MecCog”); “la” como un pronombre (“TotPron”) y un artículo (“Artículo”); “ver” como un proceso cognitivo (“MecCog”); “tus” como un pronombre (“TotPron”) y “favoritos” como una emoción positiva (“EmoPos”). Por tanto, LIWC encuentra que son usados tres pronombres (“la”, “tus”, y “la”); dos artículos (“la”, “la”); una emoción positiva (“favoritos”), y dos mecanismos cognitivos (“lanza” y “ver”).

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

Word	Articulo	Adverb	Negacio	Afect	EmoPos	EmoNeg	Enfado	Triste	MecCog
la	X								
cinta									
no		X	X						
derrocha									
acción									
como		X							X
se									
espera									X
y									X
tiene									X
una	X								
profundidad									
de									
guion									
que									X
me									
pareció									X
necesaria									X
por									
cierto				X	X				X
escena									
tortura				X		X	X		
es									
un	X								
poco									
patética				X		X		X	
resumiendo									
película									
poca									
sobre									
agente									
secreto									X
empieza									
en									
esto									X
las	X								
licencias									
para									
matar				X		X	X		
casi		X							X
mejor				X	X				
nos									
esperamos									X
a									
su									
versión									
dvd									
disfrutamos				X	X				
día									
lluvioso									

Figura 3.18. Palabras de la opinión categorizadas.

Después de procesar las palabras, el módulo calcula el porcentaje de palabras totales que coinciden con cada categoría del diccionario. Por ejemplo, para la opinión

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

presentada anteriormente, LIWC detectó que la opinión contiene 57 palabras del diccionario: 4 artículos, 3 adverbios, 1 negación, 6 palabras afectivas, 3 emociones positivas, 3 emociones negativas, 2 palabras relacionadas a enfado, 1 palabra relacionada a tristeza y 12 palabras relacionadas a procesos cognitivos. Con respecto al tweet, LIWC detectó 18 palabras del diccionario: 3 pronombres, 2 artículos, 1 emoción positiva y 2 palabras relacionadas a mecanismos cognitivos. Finalmente, estos números son convertidos a porcentajes. La Figura 3.19 muestra el resultado para el extracto de la opinión y la Figura 3.20 presenta los resultados para el tweet, los cuales se han utilizado anteriormente como ejemplos. El proceso mencionado anteriormente se realizó para todas las opiniones.

Finalmente, se construyeron vectores de características para su posterior uso en el aprendizaje automático, los cuales consisten de las características obtenidas del análisis de opiniones con LIWC y sus respectivos porcentajes, con la excepción de variables de resumen: palabras totales “WC”, palabras por oración “WPS”, palabras de más de 6 letras “Sixtlr” y conteo de palabras del diccionario “Dic”.

Filename	Segment	Articulo	Adverb	Negacio	Afect	EmoPos	EmoNeg	Enfado	Triste	MecCog
Opinion.txt	1	14.29	6.35	3.17	9.52	4.76	4.76	3.17	1.59	22.22

Figura 3.19. Extracto del vector de características resultante del análisis de una opinión.

Filename	TotPron	Articulo	EmoPos	MecCog
706547424177889280.txt	16.67	11.11	5.56	11.11

Figura 3.20. Extracto del vector de características resultante del análisis de un tweet.

Además, se llevó a cabo el análisis de textos para identificar que categorías, en conjunto e individualmente contribuyen mejor a la detección de la sátira y al análisis de sentimientos, Primero se realizó el análisis del texto por cada conjunto de categorías tal como muestra la Tabla 3.3.. Obteniendo un total de cinco vectores de características.

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

Tabla 3.3. Categorías LIWC.

Conjunto de categorías	Descripción
LP	Procesos lingüísticos
PP	Procesos psicológicos
PC	Asuntos personales
SC	Categorías del lenguaje oral
PM	Signos de puntuación

En según lugar, se estudiaron las características a través de todas las posibles combinaciones de pares obtenidas de los cinco conjuntos de categorías LIWC (ver Tabla 3.4.). Por lo que se obtuvieron un total de 10 vectores de características, cada uno con las características correspondientes a la combinación de las categorías.

Tabla 3.4. Combinación por pares de categorías LIWC.

Conjuntos de categorías	Descripción
LP-PP	Combinación de procesos lingüísticos y procesos psicológicos
LP-PC	Combinación de procesos lingüísticos y asuntos personales
LP-SC	Combinación de procesos lingüísticos y categorías del lenguaje oral
LP-PM	Combinación de procesos lingüísticos y signos de puntuación
PP-PC	Combinación de procesos psicológicos y asuntos personales
PP-SC	Combinación de procesos psicológicos y categorías del lenguaje oral
PP-PM	Combinación de procesos psicológicos y signos de puntuación
PC-SC	Combinación de asuntos personales y categorías del lenguaje oral
PC-PM	Combinación de asuntos personales y signos de puntuación
SC-PM	Combinación de categorías del lenguaje oral y signos de puntuación

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

En tercer lugar, se analizan las características a partir de la combinación de tres elementos. La Tabla 3.5. muestra las combinaciones obtenidas. Por lo que un total de 10 vectores son obtenidos.

Tabla 3.5. Combinación de tres categorías LIWC.

Conjuntos de categorías	Descripción
LP-PP-PC	Combinación de procesos lingüísticos, procesos psicológicos y asuntos personales
LP-PP-SC	Combinación de procesos lingüísticos, procesos psicológicos y categorías del lenguaje oral
LP-PP-PM	Combinación de procesos lingüísticos, procesos psicológicos y marcas de puntuación
LP-PC-SC	Combinación de procesos lingüísticos, asuntos personales y categorías de lenguaje oral
LP-PC-PM	Combinación de procesos lingüísticos, asuntos personales y signos de puntuación
LP-SC-PM	Combinación de procesos lingüísticos, categorías de lenguaje oral y signos de puntuación
PP-PC-SC	Combinación de procesos psicológicos, asuntos personales y categorías de lenguaje oral
PP-PC-PM	Combinación de procesos psicológicos, asuntos personales y signos de puntuación
PP-SC-PM	Combinación de procesos psicológicos, categorías de lenguaje oral y signos de puntuación
PC-SC-PM	Combinación de asuntos personales, categorías de lenguaje oral y signos de puntuación

El cuarto análisis consiste en el estudio de las características combinando cuatro elementos tales como se muestra en la Tabla 3.6. Un total de 5 vectores de características son obtenidos en este análisis.

Tabla 3.6 Combinación de cuatro categorías de LIWC

Conjuntos de categorías	Descripción
LP-PP-PC-SC	Combinación de procesos lingüísticos, procesos psicológicos, asuntos personales y categorías del lenguaje oral
LP-PP-PC-PM	Combinación de procesos lingüísticos, procesos psicológicos, asuntos personales y signos de puntuación
LP-PP-SC-PM	Combinación de procesos lingüísticos, procesos psicológicos, categorías del lenguaje oral y signos de puntuación
LP-PC-SC-PM	Combinación de procesos lingüísticos, asuntos personales, categorías del lenguaje oral y signos de puntuación
PP-PC-SC-PM	Combinación de procesos psicológicos, asuntos personales, categorías del lenguaje oral y signos de puntuación

Finalmente, se estudian todas las características en conjunto. Por lo que se obtiene un único vector de características con una dimensión de 72.

Los vectores de características obtenidos con cada una de las combinaciones se utilizaron para el entrenamiento de tres algoritmos de clasificación. En la siguiente sección, se explica este proceso en detalle.

3.3.4. Entrenamiento de algoritmos de aprendizaje automático

En el presente trabajo de tesis se utilizó la herramienta WEKA (Bouckaert et al., 2010) para el aprendizaje automático supervisado. WEKA provee varios algoritmos de clasificación que permiten la creación de modelos de acuerdo a los datos y al propósito de análisis. Los clasificadores se categorizan en 5 grupos: 1) Bayesiano (Naive Bayes, BayesNet, etc.); 2) funciones (regresión lineal, SMO, logistic, etc.); perezosos (IBk, LWL, etc.); 3) meta clasificadores (Bagging, Vote, etc.); 4) varios (SerializedClassifier and InputMappedClassifier); 5) reglas (DecisionTable, OneR, etc.); y 6) árboles (J48, RandomTree, etc.)

Se seleccionaron tres algoritmos para el presente estudio, el árbol de decisiones C4.5 (J48), el algoritmo red de Bayes (BayesNet) y el algoritmo de máquina de soporte

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

vectorial (SMO). Estos algoritmos fueron seleccionados debido a que han sido usados en experimentos de diversos trabajos obteniendo buenos resultados en la clasificación de datos. Los valores por defecto proporcionados por WEKA se consideran para cada algoritmo de aprendizaje.

La clasificación basada en aprendizaje automático supervisado requiere de dos conjuntos de datos: conjunto de entrenamiento y conjunto de validación. De manera general, los datos de entrenamiento consisten en un conjunto de información etiquetada, normalmente vectores de características. El algoritmo analiza estos datos y crea un modelo a partir de esta información, el cual se representa a través de reglas de clasificación, árboles de decisión o fórmulas matemáticas dependiendo del algoritmo empleado. Entonces, el modelo obtenido se valida con el conjunto de datos de prueba. Para este fin, se realizó una validación cruzada de 10 iteraciones, la cual consistió en dividir en 10 subconjuntos los datos de muestra. Uno de los subconjuntos se utiliza para las pruebas y el resto para el entrenamiento. El proceso se repite durante 10 iteraciones. Finalmente, se realiza una media aritmética de los resultados con el fin de obtener un único resultado (Martínez Cámara, 2016). El método de validación cruzada se seleccionó ya que se utiliza comúnmente por muchos estudios para el proceso de validación. Además, este método es muy preciso puesto que se evalúan a partir de K combinaciones de datos de entrenamiento y de prueba. Es decir, en cada iteración se utiliza un conjunto de datos de validación independiente para evaluar el desempeño de un modelo construido usando un conjunto de datos de entrenamiento (Su et al., 2015).

A continuación, la Figura 3.21 muestra el flujo de trabajo del aprendizaje automático supervisado para la clasificación de sentimientos. El conjunto de datos consiste en los vectores de características obtenidas en la sección 3.3.3. Estos vectores son utilizados para entrenar un clasificador y validar el modelo obtenido. Finalmente, el modelo generado es usado para clasificar datos desconocidos. En este caso el sistema es capaz de distinguir entre opiniones positivas, negativas, neutras, muy positivas y muy negativas.

Capítulo 3. Método para la clasificación de sentimientos y detección de sátira

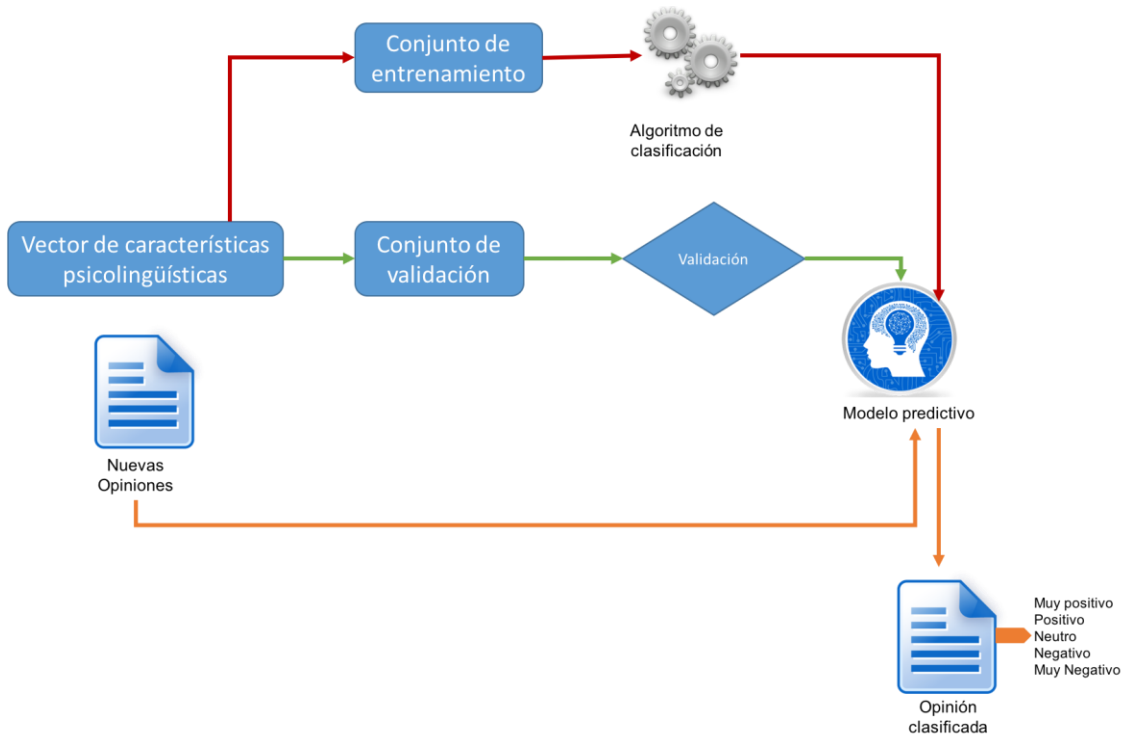


Figura 3.21. Aprendizaje automático para el análisis de sentimientos.

Por otro lado, la Figura 3.22 muestra el proceso de aprendizaje automático para la detección de la sátira. Como se especifica en la sección 3.3.3 se requiere de un vector de características, los cuales posteriormente se utilizan para el entrenamiento del clasificador y la evaluación del modelo generado. En este caso, el modelo es capaz de distinguir tweets satíricos y no satíricos.

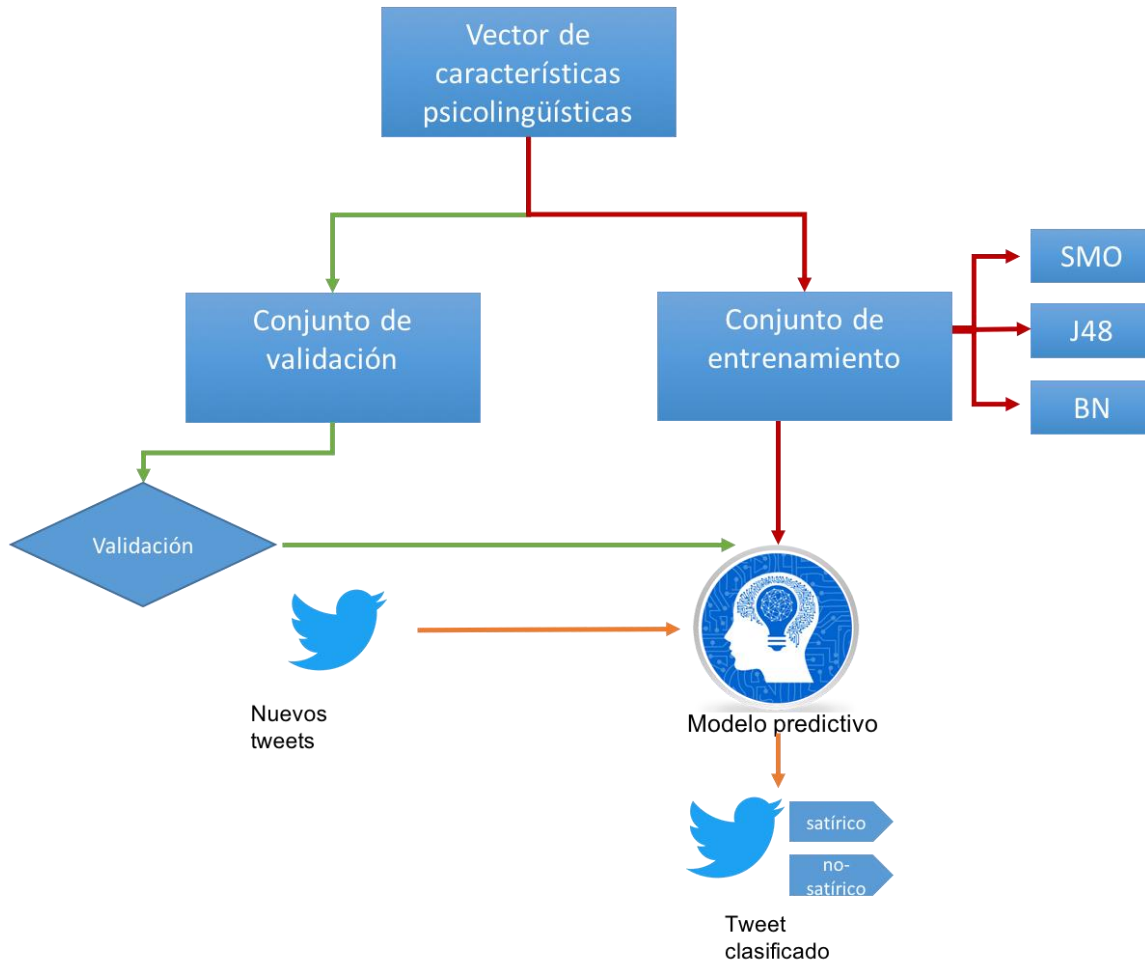


Figura 3.22. Aprendizaje automático para la detección de tweets satíricos.

Finalmente se utilizó un método de selección de características con el objetivo de identificar qué características son más relevantes para la clasificación. Este proceso se llevó a cabo usando el criterio de ganancia de información (en inglés, *Information Gain (IG)*). Los valores obtenidos pueden variar desde 0 a 1, donde las características que son más relevantes para la detección de la sátira o clasificación de la polaridad tienen un alto valor de ganancia de información. El objetivo es seleccionar las características que son más discriminantes con respecto al problema de clasificación. Por lo tanto, se seleccionaron las características que obtuvieron la ganancia de información más elevada.

3.4. Conclusiones

En este capítulo se presentó un método para el análisis de sentimientos y detección automática de la sátira. El capítulo inicia con una introducción donde se provee una descripción de las principales tecnologías que han sido utilizadas para el desarrollo del sistema. Posteriormente, en la sección 3.2 se plantearon los objetivos específicos de esta investigación.

Finalmente, en la sección 3.3, se presentó la parte central de esta tesis doctoral donde se da una descripción detallada de cada uno de los pasos del método propuesto. Esta sección comenzó con una descripción del método en general. Posteriormente, se presentaron las tres principales etapas del método. En la sección 3.3.1 se describió el conjunto de datos requerido como entrada al sistema, el cual debe estar compuesto de opiniones o tweets clasificados con una polaridad “positivo”, “negativo”, “neutro”, “muy positivo” y “muy negativo” o como “satírico” y “no satírico”. En la sección 3.3.3 se presentó la fase del pre-procesamiento del texto, el cual será tratado dependiendo de la fuente donde haya sido extraído. Esta sección consiste en dos apartados: el pre-procesamiento del texto desde sitios web de opiniones y pre-procesamiento de texto de redes sociales. Ambos consisten en normalizar el texto, sin embargo, la diferencia radica en que un texto proveniente desde redes sociales como twitter, contienen más errores de ortografía y es más difícil de analizar debido a la jerga utilizada, además de contener caracteres especiales que, en este caso, no son útiles para las etapas posteriores de procesamiento. La sección 3.3.3 presentó el procedimiento para llevar a cabo la extracción de características psicolingüísticas con LIWC. La extracción de características se llevó a cabo de manera individual y en combinación obteniendo un conjunto de vectores con las características y sus porcentajes correspondientes tras el análisis. Finalmente, en la sección 3.3.4 se describió la fase del aprendizaje automático. En este módulo recibe como entrada los vectores obtenidos en la fase anterior, para posteriormente entrenar diferentes clasificadores. El modelo generado permite clasificar nuevos datos, es decir, clasificar opiniones como positivas, negativas, muy positivas, muy negativas, neutras, o tweets como satíricos o no satíricos.

Capítulo 4. Validación y discusión

4.1. Introducción

El presente capítulo trata del análisis de los resultados obtenidos en la evaluación del método para el análisis de sentimientos y detección de la sátira. Estos resultados se presentan, analizan y evalúan con el objetivo de determinar la importancia de las características psicolingüísticas.

Para la evaluación se utilizan las medidas de precisión, exhaustividad y medida-F (Salton & McGill, 1983). Estas medidas se utilizan ampliamente para la validación de sistemas de análisis de sentimientos y en general para cualquier sistema de clasificación de texto.

Respecto a el análisis de sentimientos, el primer experimento se realiza en el dominio de películas, un corpus de referencia ampliamente utilizado para la evaluación de sistemas de análisis de sentimientos. El segundo experimento se realiza en el dominio de turismo. En este experimento, se requiere de un corpus etiquetado, el cual fue obtenido en este trabajo de tesis para dicho fin.

Finalmente, se presentan los experimentos para la evaluación del método de la sátira. Para este fin, se obtuvo un corpus de noticias etiquetado manualmente como satírico y no satírico.

4.2. Evaluación

Los sistemas de clasificación presentados en esta tesis están basados en enfoque de aprendizaje supervisado, los cuales se han evaluado a través de una validación cruzada. La validación cruzada es un método de evaluación que consiste en reducir la relación de dependencia entre los datos para entrenar el algoritmo y los datos de prueba. Este enfoque divide el conjunto de datos en k subconjuntos, utilizando $k-1$ particiones para construir el modelo y una sola para evaluar el modelo obtenido. El proceso se repite k veces, y en cada iteración el subconjunto de evaluación es

Capítulo 4. Validación y discusión

sustituido por alguna de las otras fracciones del corpus. En este trabajo, se eligió una validación cruzada de 10 (Martínez Cámara, 2016).

Con el objetivo de evaluar el rendimiento de los métodos propuestos en esta tesis doctoral, se utilizaron las métricas precisión, exhaustividad, y medida-F. Estas métricas fueron inicialmente propuesta por (Salton & McGill, 1983). La precisión representa la proporción de casos positivos que son realmente positivos. La precisión se puede calcular a partir de la ecuación (4.1).

$$Precisión = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ positivos} \quad (4.1)$$

La exhaustividad es la proporción de casos actuales positivos que se predijeron correctamente como tales. La exhaustividad se puede calcular a partir de la ecuación (4.2).

$$Exhaustividad = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos} \quad (4.2)$$

La medida-F es la medida armónica de precisión y exhaustividad. La ecuación de la medida-F es:

$$medida - F = 2 * \frac{Precision * Exhaustividad}{Precision + Exhaustividad} \quad (4.3)$$

En una clasificación multiclase, como es la clasificación de la polaridad y la detección de la sátira, se calcula la precisión, exhaustividad y la medida-F para cada clase. Es decir, se obtiene una precisión para la clase positiva, negativa, neutra, muy positiva y muy negativa, de igual manera con la exhaustividad y la medida-F. Por lo tanto, para obtener una evaluación completa del sistema se deben combinar los resultados de la evaluación de cada clase. Para este fin se aplican las medida de macroaveraging (Lewis, 1992). Se puede decir que esta métrica es la media aritmética de la precisión, exhaustividad y medida-F, siendo el cociente el número de clases con

las que cuenta el problema. Las ecuaciones de Macro-Precisión y Macro-Exhaustividad son presentadas a continuación:

$$Macro - Precisión = \frac{\sum_{i=1}^{|C|} Precision}{|C|} \quad (4.4)$$

$$Macro - Exhaustividad = \frac{\sum_{i=1}^{|C|} Exhaustividad}{|C|} \quad (4.5)$$

El Macro de la medida-F será simplemente la media armónica de macro-precisión y macro-exhaustividad.

Para el caso de ambos métodos la métricas anteriormente presentadas fueron obtenidas a partir de una matriz de confusión (ver Tabla 4.1. y Tabla 4.4), donde cada columna de la matriz representa el número de predicciones de cada clase (muy positivo, muy negativo, positivo, neutro, negativo) (satírico, no-satírico), mientras que cada fila representa a las instancias en la clase real.

A continuación, en la Tabla 4.1. se muestra la matriz de confusión para el caso de la clasificación de sentimientos. Como se puede observar el texto remarcado “verdaderos positivos” indica que los textos fueron clasificados correctamente como positivos, negativos, neutros, muy positivos, muy negativos. Mientras que los datos de forma horizontal sin remarcar “falsos negativos” indican que las opiniones fueron clasificadas como otra clase, pero en realidad pertenecen a la clase que se está evaluando. Por ejemplo, considerando la clase “positivo”, horizontalmente los datos sin remarcar indican que las opiniones fueron clasificadas como (negativa, neutra, muy positiva, muy negativa), pero en realidad pertenecían a la clase “positiva”. Finalmente, los datos en vertical sin remarcar “falsos positivos” indican las opiniones que pertenecen a otra clase, pero fueron clasificadas como la clase positiva.

Capítulo 4. Validación y discusión

Tabla 4.1. Matriz de confusión para el análisis de sentimientos.

	Positivo	Negativo	Neutro	Muy positivo	Muy negativo
Positivo	Verdaderos positivos	Falsos negativos	Falsos neutros	Falsos muy positivos	Falsos muy negativos
Negativo	Falsos positivos	Verdaderos negativos	Falsos neutros	Falsos muy positivos	Falsos muy negativos
Neutro	Falsos positivos	Falsos negativos	Verdaderos neutros	Falsos muy positivos	Falsos muy negativos
Muy positivo	Falsos positivos	Falsos negativos	Falsos neutros	Verdaderos muy positivos	Falsos muy negativos
Muy negativo	Falsos positivos	Falsos negativos	Falsos neutros	Falsos muy positivos	Verdaderos muy negativos

Con el objetivo de clarificar lo antes mencionado, consideramos la matriz de confusión presentada en la Tabla 4.2. Como se puede observar el número de opiniones analizadas son 500 (100 positivas, 100 negativas, 100 neutras, 100 muy positivas, y 100 muy negativas). Los números remarcados indican aquellos casos que fueron correctamente clasificados de acuerdo a cada clase.

Tabla 4.2. Matriz de confusión para 5 clases.

	Positivo	Negativo	Neutro	Muy positivo	Muy negativo	Total, de opiniones
Positivo	60	10	8	15	7	100
Negativo	7	75	5	4	9	100
Neutro	12	16	40	19	13	100
Muy Positivo	15	8	5	63	9	100
Muy Negativo	3	17	5	6	69	100

Para obtener las medidas de precisión (P), exhaustividad (R), y medida-F (F1) a partir de esta matriz, se deben obtener los casos verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN). La Tabla 4.3 muestra los resultados obtenidos para la matriz de confusión presentada en la Tabla 4.2. Por ejemplo, para el caso de la clase “Positivo” los verdaderos positivos (TP) son 60, es decir, 60 opiniones fueron correctamente clasificadas como positivas; los falsos negativos (FN) son 40, es decir, aquellos casos positivos que fueron clasificados como otra clase (10 negativos, 8 neutros, 15 muy positivos, 7 muy negativos); los falsos positivos (FP) son 37, es decir, aquellos casos que no eran positivos, pero fueron clasificados como tales (7 negativos, 12 neutros, 15 muy positivos, y 3 muy negativos).

Tabla 4.3. Resultados de la clasificación.

Clase	TP	FN	FP	P	R	F1
Positivo	60	40	37	0.619	0.600	0.609
Negativo	75	25	51	0.595	0.750	0.664
Neutro	40	60	23	0.635	0.400	0.491
Muy positivo	63	37	44	0.589	0.630	0.609
Muy negativo	69	31	38	0.645	0.690	0.667
<i>macroaving</i>				0.616	0.614	0.608

La Tabla 4.4 presenta la matriz de confusión para la clasificación de textos satíricos. Como se puede observar el texto remarcado “verdaderos positivos” indica que los textos fueron clasificados correctamente como satíricos o como no satíricos (Verdaderas satíricas y Verdaderas no-satíricas). Mientras que los datos de forma horizontal sin remarcar “falsos negativos” indican que los textos fueron clasificados como no satíricas, pero en realidad pertenecen a la clase satírica; o que son no satíricos, pero fueron clasificados como satíricos. Finalmente, los datos en vertical sin remarcar “falsos positivos” indica los textos que pertenecen a otra clase, pero fueron clasificadas como la clase satírica o por lo contrario como no satírica.

Tabla 4.4. Matriz de confusión para la detección de la sátira.

	satírica	no-satírica
satírica	Verdaderas satíricas	Falsas no-satíricas
no-satírica	Falsas satíricas	Verdaderas no-satíricas

4.3. Validación del método de análisis de sentimientos

El método de análisis de sentimientos se evaluó en dos dominios, películas y turismo en el idioma español. El corpus de películas ha sido utilizado como un corpus de referencia debido a que ha sido utilizado por varios enfoques para evaluar sus sistemas. El segundo corpus es del dominio turístico y fue obtenido en este trabajo de tesis.

4.3.1. Dominio de Películas

Como se mencionó anteriormente, el corpus *Spanish Movie Review*, ha sido utilizado como un corpus de referencia para la evaluación del enfoque de análisis de sentimientos. Este corpus fue presentado por primera vez en (Cruz Mata et al., 2008) y está disponible para su uso⁹. Este conjunto de datos consiste de 3878 opiniones y aproximadamente 2 millones de palabras, con una media de 546 palabras por crítica.

Tal como muestra la Tabla 4.5. , el corpus esta categorizado en 5 conjuntos con una puntuación de 1 a 5, con un total de 351 opiniones muy negativas, 923 opiniones negativas, 1253 opiniones neutrales, 890 opiniones positivas y 461 opiniones muy positivas.

⁹ <http://www.lsi.us.es/~fermin/corpusCine.zip>

Tabla 4.5. Corpus de películas.

Puntuación	Número de opiniones
1	351
2	923
3	1253
4	890
5	461
Total	3878

Para este experimento, se seleccionaron 200 opiniones aleatorias para cada categoría con el objetivo de obtener un corpus balanceado. A continuación, la Tabla 4.6. muestra algunos ejemplos de opiniones del corpus y la polaridad que tienen asignada.

Tabla 4.6. Extracto del corpus de películas.

Opinión	Polaridad
<p><i>“Noche de esas sin ganas de nada, algo de malhumor y el nunca bien ponderado "odio a todo ser humano que habita este planeta". ¿Qué mejor que ver una comedia para sacar el nubarrón de la cabeza y despejarse un rato?. Buena propuesta, lástima que me lo sigo preguntando. Tengo tendencia a ver malas películas a sabiendas que lo son. Por lo general lo encuentro gracioso, pero siempre hay una excepción a la regla. Para corroborar la magnitud de mi limitación (nunca puedo dejar una película por la mitad) vi completas las dos horas y diez minutos de esta absolutamente dolorosa "comedia". Dejo constancia que llegué hasta la parodia de Hair y "Aquarius". Quiero mis 130 minutos de vuelta. Virgen a los 40 años es la peor película que vi en mucho tiempo (mucho).P.D.: Catherine Keener te sigo queriendo.”</i></p>	1
<p><i>“Rowan Atkinson protagoniza esta comedia al más puro estilo británico en la que toma el papel de un pastor protestante, párroco de un pueblo perdido en el paisaje inglés. En este lugar idílico convive en su comunidad como el párroco perfecto. Sin embargo, su abstracción del mundo y su obsesión por escribir un discurso para una importante convención le impide ver la aventura de su mujer con el profesor de golf, la facilidad con la que la hija cambia de novio o que su hijo no se relaciona con otros chicos más que para recibir palizas. Es entonces cuando llega la nueva ama de llaves (Maggie Smith), una <>dulce y tradicional señora con un oscuro pasado que no duda en arreglar de arriba a abajo la vida de esta familia. El único inconveniente serán su "peculiares" métodos de poner fin a los problemas. Por fin vemos a Rowan haciendo un papel que no nos recuerde a Mr. Bean. Le ha costado, pero por fin lo ha conseguido. De hecho, este cambio no será por mucho tiempo, ya que el ingeniero inglés más divertido</i></p>	2

Opinión	Polaridad
<p><i>está preparando Mr. Bean 2. Secretos de Familia es una comedia inglesa, es decir, de humor inglés, con todo lo que esto implica, para no dejar lugar a dudas. El personaje que encarna la elogiada por su interpretación, Maggie Smith, introduce el calificativo de negro a la comedia, incluso macabra en determinadas ocasiones. El resultado es una película entretenida, sin grandes pretensiones más que pasar un rato agradable, idóneo para que sea en familia, o para los que lo prefieran, ver en dvd en una tarde lluviosa de domingo.”</i></p>	
<p><i>“Yo creía que iba a ver algo nuevo de Tim Burton, pero me he encontrado con una película que es entretenida ,pero que se parece bastante a su predecesora Pesadilla antes de Navidad. De todas maneras la película como indicaba anteriormente esta bien, tiene ese ambiente gótico tan característico de Burton, cabe destacar el juego de luces y algunos encuadres que la hacen parecer una película de actores de carne y hueso, la historia es original y tiene bastantes toques de humor, pero lo que no me gusta y me pasa también con las películas de Disney es que canten, por este tipo de formula y la caracterización de los personajes sigo afirmando y concluyo aquí, que La Novia Cadáver parece la segunda parte de pesadilla antes de navidad.”</i></p>	3
<p><i>“Hay películas muy buenas, buenas, malas, muy malas, Señales, etc. Hay películas que gustan a un grupo de personas con ciertos criterios, esas mismas a otro grupo les hace vomitar. Pero hay películas que, independientemente de si te gustan o no, tienes que decir "qué pelicolón". Esta es una de ellas. La película trata sobre los diferentes grupúsculos raciales que hay en la ciudad de Los Ángeles. Una colisión de coche inicial, nos lleva a una trama increíble sobre como interactúan las personas y como cada uno de nuestros actos termina por desencadenar unos sucesos u otros. Un policía xenófobo, otro pardillo, un par de negros ladrones, un par de blancos "respetables", un inspector con ideales, un hispano bueno, un persa (iraní) un poco esquizofrénico, etc. Todos y</i></p>	4

Opinión

Polaridad

cada uno de ellos tiene relación con un suceso concreto de la película. Y de esto trata. Explicar esta película es imposible. Esto hay que verlo. Da muy mal rollo, durante casi toda ella, hasta el punto de querer dejar de verla. Pero tengo que quitarme el sombrero y decir, menudo peliculón. La trama se entrelaza de un modo increíble, sin hacerse pesada pero sí angustiosa. Los actores están sublimes, en papeles de personas horribles que han sido absorbidos por una ciudad injusta y cruel. Y el guión es buenísimo, ¿que más puedo decir? Resumiendo, una película muy dura, con un guión magistral y un desarrollo increíble. Sólo tiene una escena bonita, en la que lloré como una magdalena. Que maravilla.”

“No tengo palabras para describir esta película, simplemente es GENIAL. Felicity hace un gran trabajo y demuestra 5 lo gran actriz que es y por qué se ha llevado todos esos premios (el Oscar lo tiene que ganar), yo sólo la había visto en Mujeres Desesperadas y el cambio de registro me ha encantado y si ves la película en VO se aprecia mucho más. Punto a favor: la misma que la dobla en la serie se ha encargado de ponerle la voz en la película, pero no la pone lo suficientemente grave, al principio un poquito, pero luego se le olvida. La primera media hora resulta un poquito lenta, pero una vez se van desarrollando las historias de los personajes te quedas enganchada en ese mundo. Decir también que me gusta más el tagline americano que el que han puesto en España; Life is more than the sum of its parts. Bree, antes Stan, está a punto de conseguir la autorización para someterse a su operación de cambio de sexo, pero recibe la llamada de un chico que dice ser su hijo. Su psiquiatra le dice que debe conocer al chico antes de firmarle la autorización. Bree a regañadientes acepta ir a Nueva York a sacar al chico de la cárcel, aunque la fianza es de 1\$, Toby no es ningún angelito, con 17 años trapichea con drogas y ha sido detenido por prostitución. El chaval le convence para que le lleve a Los Ángeles, donde se hará actor porno, y compran un coche. El viaje está

Opinión**Polaridad**

lleno de incidentes y mentiras. Bree no le ha dicho que es su verdadero padre y menos que en realidad es un hombre (le ha dicho que es de la Iglesia del Padre Potencial) y Toby acaba descubriendo su secreto. Aunque al principio reacciona mal, termina aceptando a Bree. Un hippie les roba el coche y los deja tirados en la carretera sin más ropa que la que llevaban puesta. En un bar, Bree liga con un hombre que los deja en Phoenix, en casa de los padres de Bree. La madre reacciona fatal al ver a su hija, pero el resto de la familia no parece tener ningún problema (sobre todo la hermana). Cuando les confiesa que Toby es su hijo, la madre empieza a tratarlo estupendamente, e incluso quiere que se quede a vivir con ellos como condición a pagarle el billete de avión a Bree para que llegue a tiempo a su operación. Bree, por fin le cuenta a Toby quien es en realidad y reacciona fatal, llegando a darle un puñetazo en un ojo. Ya en Los Ángeles, Bree se opera, pero ese día no es el más feliz de su vida, echa de menos a Toby. Lo mejor sin duda es la gran actuación de Felicity Huffman. Muchas de las frases "mi cuerpo puede estar aún en proceso, pero no hay nada malo con mi alma". La transexual rubia y sus amigos. La madre de Bree"

Capítulo 4. Validación y discusión

4.3.2. Dominio de turismo

El segundo corpus fue obtenido desde el sitio Web TripAdvisor, el sitio de viajes más grande del mundo con 315 millones de usuarios únicos por mes, aproximadamente 200 millones de opiniones de más de 4.4 millones de alojamientos, restaurantes y atracciones.

El corpus contiene 1600 opiniones sobre hoteles, restaurantes, museos, entre otros temas. Cada opinión fue manualmente examinada y clasificada por un grupo de expertos en este dominio, con el objetivo de asegurar la calidad del corpus. Cada opinión se almacenó en un formato XML (*eXtensible Markup Language* - Lenguaje de Marcado Extensible). Además, se asignó un valor de 1 a 5 a cada opinión de acuerdo con los siguientes criterios:

- 5, cuando la revisión tiene un sentimiento muy positivo.
- 4, cuando la revisión tiene un sentimiento positivo.
- 3, cuando la revisión no tiene sentimiento.
- 2, cuando la revisión tiene un sentimiento negativo.
- 1, cuando la revisión tiene un sentimiento muy negativo.

A continuación, la Figura 4.1 presenta un ejemplo de una opinión almacenada en formato XML y con su respectivo valor de polaridad.

```
<review id="3" rank="5" topic="Alojamiento">
  <abstract>
    Gran hallazgo
  </abstract>
  <content>
    Fantástico hotel, personal muy amable, habitaciones espaciosas, desayuno excelente
    y muy silencioso para descansar. El vecindario es uno de los mejores de Barna, el
    jardín en la zona posterior es una delicia. Muy pero muy recomendable.
  </content>
</review>
```

Figura 4.1. Ejemplo de una opinión muy positiva.

4.3.3. Resultados

Esta sección presenta los resultados obtenidos por el método propuesto para el análisis de sentimientos en los dos dominios anteriormente mencionados, películas y

turismo. Los resultados son representados de la siguiente manera en cada tabla: la primera columna representa el conjunto de categorías utilizadas para el entrenamiento, como se mencionó en la sección 3.3.3 se utiliza cada conjunto de características (LP “Procesos lingüísticos”, PP “Procesos psicológicos”, PC “Asuntos personales”, SC “Categorías del lenguaje oral”, PM “Signos de puntuación”) individualmente, así como todas sus posibles combinaciones. Posteriormente en las siguientes columnas se presentan los resultados de precisión, exhaustividad y medida-F cuando son utilizadas dos clases (positiva, negativa), tres clases (positiva, negativa, neutra) y cinco clases (positiva, negativa, neutra, muy positiva, muy negativa).

4.3.4. Resultados en el dominio de películas

A continuación, en esta sección se presentan los resultados obtenidos para el dominio de películas. Para este experimento se utilizó el corpus de referencia *Spanish Movie Review* (para más información ver la sección 4.3.1). La Tabla 4.7., Tabla 4.8. y Tabla 4.9. presentan los resultados de evaluación, como se mencionó anteriormente usando, dos, tres y cinco clases.

La Tabla 4.7. muestra los resultados obtenidos por el algoritmo SMO, como se puede observar, los experimentos con dos clases obtienen mejores resultados que con tres y cinco clases. Cuando los experimentos involucran una categoría, el mejor resultado se obtiene para la categoría PP “procesos psicológicos” con una medida-F de 82%, 78% y 58.6% para dos, tres y cinco clases, respectivamente. Mientras que los peores resultados son obtenidos por la categoría SC “categorías del lenguaje oral” con una medida-F de 62%, 62.4% y 45.6%. Con respecto a la combinación de categorías los mejores resultados son obtenidos cuando se involucran las categorías individuales con los resultados más altos. Por ejemplo, en la combinación por pares, los mejores resultados son obtenidos cuando se combinan las categorías LP “procesos lingüísticos” y PP “procesos psicológicos” obteniendo un valor considerablemente mejor con una medida-F de 84.1% y 78.9%, para dos y tres clases. Sin embargo, para cinco clases los mejores resultados son obtenidos combinando la categoría PP

Capítulo 4. Validación y discusión

“Procesos psicológicos” y PM “signos de puntuación” con una medida-F de 61.6%. Finalmente, los resultados para dos y tres clases muestran que la combinación de todas las categorías provee mejores resultados, lo cual indica que a pesar de ser mejores predictores las categorías LP “Procesos lingüísticos” y PP “Procesos psicológicos” existen características de PC “asuntos personales”, SC “categorías del lenguaje oral” y PM “signos de puntuación” que son también importantes para la clasificación de sentimientos.

Tabla 4.7. Experimentos con el algoritmo SMO.

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP	0.806	0.804	0.803	0.719	0.818	0.760	0.572	0.576	0.568
PP	0.821	0.820	0.820	0.735	0.840	0.780	0.585	0.59	0.586
PC	0.703	0.703	0.703	0.661	0.751	0.701	0.513	0.532	0.511
SC	0.705	0.680	0.620	0.646	0.721	0.624	0.574	0.498	0.456
PM	0.731	0.730	0.729	0.664	0.755	0.703	0.538	0.519	0.507
LP-PP	0.842	0.841	0.841	0.798	0.846	0.789	0.591	0.598	0.593
LP-PC	0.801	0.800	0.799	0.723	0.825	0.766	0.585	0.596	0.587
LP-SC	0.817	0.814	0.812	0.720	0.820	0.761	0.574	0.577	0.57
LP-PM	0.826	0.824	0.823	0.736	0.842	0.781	0.586	0.586	0.581
PP-PC	0.820	0.819	0.818	0.739	0.844	0.784	0.588	0.599	0.591
PP-SC	0.816	0.815	0.814	0.728	0.833	0.774	0.592	0.597	0.593
PP-PM	0.832	0.831	0.831	0.736	0.841	0.781	0.615	0.618	0.616
PC-SC	0.718	0.718	0.717	0.664	0.755	0.704	0.519	0.531	0.52
PC-PM	0.746	0.746	0.746	0.677	0.771	0.719	0.537	0.546	0.539
SC-PM	0.745	0.741	0.738	0.665	0.756	0.702	0.543	0.536	0.52
LP-PP-PC	0.843	0.843	0.842	0.746	0.846	0.789	0.596	0.606	0.599
LP-PP-SC	0.846	0.845	0.845	0.771	0.845	0.790	0.595	0.601	0.596
LP-PP-PM	0.846	0.845	0.845	0.784	0.852	0.795	0.611	0.614	0.611

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP-PC-SC	0.806	0.805	0.804	0.724	0.827	0.768	0.588	0.6	0.59
LP-PC-PM	0.824	0.823	0.822	0.738	0.846	0.785	0.593	0.6	0.595
LP-SC-PM	0.829	0.828	0.827	0.733	0.838	0.778	0.584	0.586	0.581
PP-PC-SC	0.820	0.819	0.818	0.734	0.838	0.778	0.591	0.6	0.593
PP-PC-PM	0.820	0.819	0.818	0.729	0.832	0.773	0.613	0.62	0.615
PP-SC-PM	0.841	0.840	0.840	0.733	0.839	0.779	0.612	0.616	0.613
PC-SC-PM	0.761	0.761	0.761	0.685	0.781	0.728	0.553	0.562	0.554
LP-PP-PC-SC	0.843	0.843	0.842	0.761	0.851	0.796	0.595	0.604	0.598
LP-PP-PC-PM	0.856	0.856	0.856	0.787	0.853	0.799	0.608	0.613	0.609
LP-PP-SC-PM	0.848	0.848	0.847	0.779	0.849	0.795	0.613	0.616	0.613
LP-PC-SC-PM	0.833	0.831	0.831	0.740	0.848	0.787	0.582	0.591	0.584
PP-PC-SC-PM	0.839	0.838	0.837	0.737	0.843	0.783	0.598	0.604	0.6
LP-PP-PC-SC-PM	0.864	0.864	0.864	0.781	0.856	0.803	0.6	0.606	0.602

La Tabla 4.8. presenta los resultados obtenidos por el algoritmo BayesNet para el dominio de películas. Como se puede observar, los resultados obtenidos por este algoritmo son menores que los obtenidos por el algoritmo SMO.

En general cuando los experimentos involucran una categoría, el mejor resultado es obtenido por la categoría LP “procesos lingüísticos” con una medida-F de 77.3% para dos clases, 75% para tres clases y 53.5% para cinco clases. Por lo contrario, los peores resultados son obtenidos por la categoría PC “asuntos personales” para dos clases con una medida-F de 54.6%, mientras tanto para tres y cinco clases los peores resultados son obtenidos por dos categorías, PC “asuntos personales” y SC “categorías del lenguaje oral” con una medida-F de 52.9% y 36.7%. En cuanto a la combinación de categorías los mejores resultados para dos y tres clases son los obtenidos por la combinación de LP, PP y PM, así como la combinación de todas las categorías, con una medida-F de 79.6% y 75.5%. Para cinco clases, diversas

Capítulo 4. Validación y discusión

combinaciones (LP-PP, LP-PP-PC, LP-PP-SC y LP-PP-PC-SC) obtienen los mejores resultados con una medida-F de 60.5%. Además, los resultados muestran que al igual que con SMO los resultados obtenidos para dos clases son mejores que para tres y cinco clases.

Tabla 4.8. Experimentos con el algoritmo BayesNet.

	2 clases			3clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP	0.775	0.774	0.773	0.705	0.806	0.75	0.562	0.565	0.535
PP	0.698	0.698	0.697	0.692	0.723	0.674	0.46	0.567	0.465
PC	0.596	0.624	0.546	0.46	0.7	0.529	0.34	0.5	0.367
SC	0.65	0.65	0.65	0.46	0.7	0.529	0.34	0.5	0.367
PM	0.729	0.701	0.658	0.647	0.716	0.613	0.402	0.499	0.392
LP-PP	0.778	0.778	0.777	0.736	0.796	0.751	0.616	0.612	0.605
LP-PC	0.761	0.76	0.759	0.705	0.806	0.75	0.562	0.565	0.535
LP-SC	0.775	0.774	0.773	0.705	0.806	0.75	0.562	0.565	0.535
LP-PM	0.787	0.786	0.786	0.698	0.797	0.742	0.56	0.563	0.533
PP-PC	0.698	0.698	0.698	0.692	0.723	0.674	0.46	0.567	0.465
PP-SC	0.698	0.698	0.697	0.692	0.723	0.674	0.46	0.567	0.465
PP-PM	0.721	0.721	0.721	0.685	0.727	0.671	0.459	0.566	0.465
PC-SC	0.596	0.624	0.546	0.46	0.7	0.529	0.34	0.5	0.367
PC-PM	0.678	0.674	0.657	0.647	0.716	0.613	0.402	0.499	0.392
SC-PM	0.727	0.7	0.657	0.647	0.716	0.613	0.402	0.499	0.392
LP-PP-PC	0.784	0.784	0.783	0.736	0.796	0.751	0.616	0.612	0.605
LP-PP-SC	0.778	0.778	0.778	0.736	0.796	0.751	0.616	0.612	0.605
LP-PP-PM	0.796	0.796	0.796	0.738	0.8	0.755	0.609	0.608	0.604
LP-PC-SC	0.761	0.76	0.759	0.705	0.806	0.75	0.562	0.565	0.535
LP-PC-PM	0.78	0.78	0.78	0.698	0.797	0.742	0.56	0.563	0.533
LP-SC-PM	0.787	0.786	0.786	0.698	0.797	0.742	0.56	0.563	0.533

	2 clases			3clases			5 clases		
	P	R	F	P	R	F	P	R	F
PP-PC-SC	0.696	0.696	0.696	0.692	0.723	0.674	0.46	0.567	0.465
PP-PC-PM	0.72	0.72	0.72	0.685	0.727	0.671	0.459	0.566	0.465
PP-SC-PM	0.722	0.721	0.721	0.685	0.727	0.671	0.459	0.566	0.465
PC-SC-PM	0.678	0.674	0.657	0.647	0.716	0.613	0.402	0.499	0.392
LP-PP-PC-SC	0.784	0.784	0.783	0.736	0.796	0.751	0.616	0.612	0.605
LP-PP-PC-PM	0.795	0.795	0.795	0.738	0.8	0.755	0.609	0.608	0.604
LP-PP-SC-PM	0.795	0.795	0.795	0.738	0.8	0.755	0.609	0.608	0.604
LP-PC-SC-PM	0.779	0.779	0.779	0.698	0.797	0.742	0.56	0.563	0.533
PP-PC-SC-PM	0.719	0.719	0.719	0.685	0.727	0.671	0.459	0.566	0.465
LP-PP-PC-SC-PM	0.796	0.796	0.796	0.738	0.8	0.755	0.609	0.608	0.604

Los resultados obtenidos para el dominio de películas con el algoritmo J48 son presentados en Tabla 4.9.. Estos resultados son menores que los obtenidos por el algoritmo SMO y BayesNet. Cuando los experimentos involucran una categoría, el mejor resultado para dos clases es obtenido por la categoría LP “procesos lingüísticos” con una medida-F de 72.2% y el peor es obtenido por la categoría SC “categorías del lenguaje oral” con una medida-F de 48.3%. Para tres clases, la mejor categoría es PC “asuntos personales” con una medida-F de 70.6%, y la peor categoría es SC “categorías del lenguaje oral” con una medida-F de 56.2%. Finalmente, para cinco clases, la categoría más discriminante es PP “procesos psicológicos” con una medida-F de 55%, mientras que SC “categorías del lenguaje oral” es la peor con una medida-F de 42.4%.

Con respecto a la combinación de categorías, cuando se hace uso de dos clases el mejor resultado es obtenido con la combinación de las categorías LP “procesos lingüísticos”, PC “asuntos personales” y PM “signos de puntuación” con una medida-F de 75%, y el peor resultado es obtenido con la combinación de categorías PP “procesos psicológicos”, PC “asuntos personales” y SC “categorías del lenguaje oral” con una medida-F de 69.5%. Por otro lado, cuando los experimentos se realizan con

Capítulo 4. Validación y discusión

tres clases, los mejores resultados son obtenidos con la combinación LP “procesos lingüísticos”, PP “procesos psicológicos” y PC “asuntos personales”, con una medida-F de 74.9%, mientras el peor resultado es obtenido por la combinación de categorías SC “categorías del lenguaje oral” y PM “marcas de puntuación” con una medida-F de 66.8%. Finalmente, para cinco clases la combinación de categorías LP “procesos lingüísticos”, PP “procesos psicológicos” y SC “categorías del lenguaje oral” obtiene los mejores resultados con una medida-F de 55.6%, y el peor resultado es obtenido por la combinación de LP “procesos lingüísticos”, PC “asuntos personales” y PM “signos de puntuación” con una medida-F de 49%.

Tabla 4.9. Experimentos en el dominio de películas con el algoritmo J48.

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP	0.723	0.723	0.722	0.69	0.689	0.689	0.531	0.531	0.531
PP	0.705	0.705	0.705	0.65	0.671	0.67	0.551	0.55	0.55
PC	0.704	0.701	0.695	0.696	0.734	0.706	0.532	0.53	0.53
SC	0.4	0.65	0.483	0.684	0.707	0.562	0.678	0.499	0.422
PM	0.694	0.685	0.66	0.669	0.704	0.662	0.508	0.509	0.508
LP-PP	0.725	0.725	0.725	0.73	0.728	0.729	0.532	0.531	0.531
LP-PC	0.723	0.723	0.722	0.739	0.746	0.742	0.519	0.519	0.519
LP-SC	0.719	0.719	0.719	0.696	0.697	0.696	0.523	0.523	0.523
LP-PM	0.744	0.744	0.743	0.717	0.714	0.715	0.515	0.515	0.515
PP-PC	0.693	0.693	0.692	0.675	0.675	0.675	0.519	0.519	0.519
PP-SC	0.703	0.703	0.702	0.7	0.701	0.701	0.548	0.545	0.546
PP-PM	0.714	0.714	0.713	0.703	0.704	0.703	0.542	0.54	0.54
PC-SC	0.706	0.703	0.696	0.708	0.732	0.713	0.533	0.533	0.533
PC-PM	0.712	0.711	0.71	0.699	0.702	0.699	0.527	0.525	0.526
SC-PM	0.699	0.689	0.663	0.678	0.691	0.668	0.534	0.534	0.534
LP-PP-PC	0.724	0.724	0.723	0.749	0.748	0.749	0.536	0.536	0.536

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP-PP-SC	0.72	0.72	0.72	0.731	0.728	0.729	0.556	0.556	0.556
LP-PP-PM	0.749	0.749	0.749	0.727	0.725	0.726	0.506	0.507	0.506
LP-PC-SC	0.733	0.733	0.732	0.732	0.736	0.734	0.522	0.522	0.521
LP-PC-PM	0.75	0.75	0.75	0.712	0.713	0.712	0.491	0.489	0.49
LP-SC-PM	0.744	0.744	0.743	0.719	0.713	0.715	0.523	0.521	0.521
PP-PC-SC	0.695	0.695	0.695	0.699	0.695	0.697	0.523	0.524	0.524
PP-PC-PM	0.723	0.722	0.722	0.698	0.695	0.696	0.529	0.528	0.528
PP-SC-PM	0.708	0.708	0.707	0.711	0.713	0.712	0.529	0.527	0.526
PC-SC-PM	0.713	0.713	0.711	0.689	0.694	0.691	0.512	0.514	0.513
LP-PP-PC-SC	0.725	0.725	0.725	0.731	0.731	0.731	0.541	0.542	0.541
LP-PP-PC-PM	0.748	0.748	0.747	0.737	0.735	0.736	0.527	0.528	0.527
LP-PP-SC-PM	0.746	0.746	0.746	0.731	0.73	0.73	0.516	0.516	0.516
LP-PC-SC-PM	0.747	0.746	0.746	0.705	0.708	0.706	0.542	0.542	0.542
PP-PC-SC-PM	0.721	0.721	0.721	0.7	0.7	0.7	0.519	0.518	0.518
LP-PP-PC-SC-PM	0.749	0.749	0.749	0.721	0.721	0.721	0.502	0.501	0.501

4.3.4.1. Resultados en el dominio de turismo

En esta sección se presentan los resultados obtenidos de la evaluación del método de análisis de sentimientos en el dominio turístico. Para llevar a cabo los experimentos, se utilizó el corpus presentado en la sección 4.3.2 .

A continuación, la Tabla 4.10., Tabla 4.11. y Tabla 4.12. muestran los resultados obtenidos por el algoritmo SMO, BayesNet, y J48. La primera columna indica la categoría o combinación de categorías utilizadas para entrenar el clasificador y las columnas posteriores muestran las medidas precisión, exhaustividad, y medida-F obtenidas cuando son utilizadas 2 clases (positiva, negativa), 3 clases (positiva, neutras, negativa) y 5 clases (positiva, neutra, negativa, muy positiva, muy negativa).

Capítulo 4. Validación y discusión

Los resultados obtenidos por el algoritmo SMO se presentan en la Tabla 4.10.. Estos resultados muestran que cuando las categorías se utilizan individualmente, la categoría PP “procesos lingüísticos” es más discriminante para dos, tres y cinco clases con una medida-F de 84.6%, 78.2% y 58%, respectivamente. El peor resultado para dos y tres clases se obtiene con la categoría PM “signos de puntuación” con una medida-F de 57.7% y 59.6%. Para cinco clases, el peor resultado es obtenido por SC “categorías del lenguaje oral” con una medida-F de 38.3%.

Con respecto a la combinación de categorías para dos y tres clases, el mejor resultado es obtenido por la combinación de categorías LP “procesos lingüísticos”, PP “procesos psicológicos”, PC “asuntos personales” y PM “signos de puntuación” con una medida-F de 90.9% y 84.6%. Por otro lado, para cinco clases, el mejor resultado se obtiene con la combinación de todas las categorías, es decir, LP “procesos lingüísticos”, PP “procesos psicológicos”, PC “asuntos personales”, SC “categorías del lenguaje oral” y PM “signos de puntuación”, con una medida-F de 62.8. Finalmente, la combinación de categorías menos discriminante para dos, tres y cinco clases es SC “categorías del lenguaje oral” y PM “signos de puntuación” con una medida-F de 64.1% para dos clases, 63.9% para tres clases, y 45% para cinco clases.

Finalmente, los resultados muestran de manera general que se obtienen mejores resultados para dos clases que para tres y cinco clases.

Tabla 4.10. Experimentos en el dominio turístico con el algoritmo SMO.

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP	0.823	0.822	0.822	0.92	0.837	0.772	0.582	0.586	0.579
PP	0.846	0.846	0.846	0.729	0.848	0.782	0.578	0.59	0.58
PC	0.691	0.676	0.66	0.636	0.719	0.655	0.491	0.496	0.467
SC	0.638	0.623	0.582	0.604	0.682	0.606	0.445	0.464	0.383
PM	0.673	0.635	0.577	0.626	0.684	0.596	0.457	0.47	0.431
LP-PP	0.902	0.902	0.902	0.84	0.895	0.84	0.606	0.613	0.607
LP-PC	0.847	0.845	0.845	0.797	0.849	0.784	0.578	0.583	0.577

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP-SC	0.825	0.824	0.824	0.802	0.839	0.776	0.582	0.586	0.579
LP-PM	0.82	0.819	0.819	0.769	0.836	0.772	0.573	0.579	0.573
PP-PC	0.857	0.857	0.857	0.744	0.866	0.798	0.579	0.59	0.581
PP-SC	0.842	0.842	0.842	0.765	0.843	0.778	0.581	0.593	0.583
PP-PM	0.855	0.855	0.855	0.737	0.856	0.79	0.57	0.583	0.573
PC-SC	0.697	0.69	0.683	0.637	0.733	0.673	0.484	0.498	0.475
PC-PM	0.698	0.695	0.693	0.639	0.735	0.68	0.495	0.494	0.475
SC-PM	0.657	0.651	0.641	0.611	0.698	0.639	0.481	0.479	0.45
LP-PP-PC	0.906	0.905	0.905	0.832	0.893	0.843	0.615	0.621	0.616
LP-PP-SC	0.899	0.899	0.899	0.845	0.895	0.844	0.61	0.616	0.611
LP-PP-PM	0.901	0.9	0.9	0.839	0.891	0.839	0.612	0.618	0.612
LP-PC-SC	0.844	0.843	0.843	0.808	0.853	0.789	0.576	0.581	0.575
LP-PC-PM	0.834	0.833	0.833	0.764	0.848	0.785	0.561	0.568	0.562
LP-SC-PM	0.828	0.827	0.826	0.825	0.843	0.781	0.573	0.578	0.573
PP-PC-SC	0.859	0.859	0.859	0.785	0.861	0.796	0.582	0.591	0.584
PP-PC-PM	0.851	0.851	0.851	0.766	0.864	0.798	0.574	0.584	0.576
PP-SC-PM	0.851	0.851	0.851	0.756	0.854	0.789	0.576	0.588	0.578
PC-SC-PM	0.716	0.713	0.71	0.653	0.752	0.694	0.498	0.502	0.489
LP-PP-PC-SC	0.903	0.903	0.902	0.838	0.895	0.845	0.619	0.624	0.619
LP-PP-PC-PM	0.91	0.91	0.909	0.836	0.893	0.846	0.625	0.631	0.626
LP-PP-SC-PM	0.902	0.901	0.901	0.843	0.891	0.841	0.615	0.622	0.616
LP-PC-SC-PM	0.837	0.837	0.837	0.814	0.859	0.798	0.573	0.58	0.574
PP-PC-SC-PM	0.869	0.869	0.869	0.754	0.858	0.794	0.577	0.586	0.579
LP-PP-PC-SC-PM	0.904	0.903	0.903	0.834	0.89	0.845	0.627	0.633	0.628

La Tabla 4.11. muestra que BayesNet obtiene menores resultados que SMO, cuando la evaluación involucra una sola categoría. Los mejores resultados son obtenidos con PP

Capítulo 4. Validación y discusión

“procesos psicológicos” para dos, tres y cinco clases con una medida-F de 78.8%, 76.6% y 55.2%, respectivamente. Por lo contrario, los peores resultados son obtenidos por la categoría PM “signos de puntuación” con una medida-F de 61.9%, 57.3% y 34.4%.

Por otro lado, la mejor combinación para dos clases es la obtenida por LP “procesos lingüísticos” y PP “procesos psicológicos” con una medida-F de 81.5%, mientras que la peor es la obtenida por SC “categorías del lenguaje oral” y PM “signos de puntuación” con una medida-F de 66.2%. Con respecto a tres clases, la mejor combinación es la obtenida por LP “procesos lingüísticos”, PP “procesos psicológicos” y PC “asuntos personales” con una medida-F de 79.3% y la peor es la obtenida por la combinación de SC “categorías del lenguaje oral” y PM “signos de puntuación” con una medida-F de 63.4%. Finalmente, para tres clases, la mejor combinación de categorías es LP “procesos lingüísticos”, PP “procesos psicológicos”, y PM “signos de puntuación” con una medida-F de 57.4% y la peor combinación es PC y PM con una medida-F de 37.1%.

Tabla 4.11. Experimentos en el dominio turístico con el algoritmo BayesNet.

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP	0.778	0.778	0.778	0.733	0.795	0.751	0.476	0.531	0.514
PP	0.788	0.788	0.788	0.755	0.818	0.766	0.558	0.578	0.552
PC	0.673	0.671	0.668	0.632	0.724	0.669	0.365	0.457	0.371
SC	0.633	0.633	0.633	0.6	0.688	0.639	0.332	0.453	0.352
PM	0.668	0.649	0.619	0.582	0.658	0.573	0.351	0.448	0.344
LP-PP	0.816	0.815	0.815	0.776	0.818	0.789	0.576	0.589	0.572
LP-PC	0.775	0.775	0.775	0.751	0.772	0.761	0.522	0.532	0.518
LP-SC	0.785	0.785	0.784	0.741	0.794	0.753	0.515	0.529	0.513
LP-PM	0.777	0.777	0.777	0.727	0.786	0.743	0.516	0.529	0.514
PP-PC	0.797	0.797	0.797	0.768	0.828	0.776	0.567	0.584	0.559

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
PP-SC	0.787	0.786	0.786	0.767	0.82	0.771	0.56	0.579	0.554
PP-PM	0.796	0.796	0.796	0.768	0.822	0.772	0.556	0.576	0.55
PC-SC	0.695	0.694	0.693	0.638	0.733	0.678	0.36	0.457	0.374
PC-PM	0.684	0.682	0.68	0.632	0.723	0.668	0.365	0.457	0.371
SC-PM	0.668	0.666	0.662	0.601	0.686	0.634	0.371	0.451	0.376
LP-PP-PC	0.813	0.813	0.812	0.78	0.821	0.793	0.566	0.583	0.564
LP-PP-SC	0.81	0.809	0.808	0.773	0.815	0.786	0.575	0.589	0.571
LP-PP-PM	0.813	0.812	0.812	0.775	0.815	0.788	0.578	0.591	0.574
LP-PC-SC	0.776	0.775	0.775	0.756	0.803	0.761	0.52	0.531	0.517
LP-PC-PM	0.776	0.775	0.775	0.749	0.797	0.757	0.52	0.529	0.517
LP-SC-PM	0.776	0.775	0.775	0.735	0.789	0.747	0.514	0.528	0.512
PP-PC-SC	0.796	0.795	0.795	0.774	0.823	0.778	0.57	0.587	0.562
PP-PC-PM	0.797	0.796	0.796	0.764	0.822	0.775	0.565	0.581	0.557
PP-SC-PM	0.797	0.796	0.796	0.75	0.817	0.766	0.559	0.578	0.552
PC-SC-PM	0.703	0.702	0.701	0.641	0.734	0.677	0.36	0.457	0.374
LP-PP-PC-SC	0.809	0.808	0.808	0.777	0.82	0.791	0.565	0.582	0.563
LP-PP-PC-PM	0.812	0.811	0.811	0.774	0.818	0.788	0.567	0.583	0.565
LP-PP-SC-PM	0.811	0.81	0.81	0.773	0.815	0.786	0.576	0.59	0.573
LP-PC-SC-PM	0.777	0.776	0.776	0.749	0.798	0.757	0.519	0.529	0.516
PP-PC-SC-PM	0.799	0.799	0.798	0.778	0.824	0.779	0.568	0.584	0.56
LP-PP-PC-SC-PM	0.806	0.805	0.805	0.777	0.818	0.79	0.566	0.583	0.564

La Tabla 4.12. muestra los resultados obtenidos con el algoritmo J48, como se puede observar los resultados son inferiores que los obtenidos con el algoritmo SMO y BayesNet. Individualmente, la categoría PP “procesos psicológicos” es la más discriminante para dos, tres y cinco clases con una medida-F de 77%, 75.4% y 54.6%,

Capítulo 4. Validación y discusión

respectivamente. Por lo contrario, la peor categoría es SC “categorías del lenguaje oral” con una medida-F de 62% para dos clases, 63.7% para tres clases, y 41.6% para cinco clases. Por otro lado, la combinación de todas las categorías provee mejores resultados para dos clases con una medida-F de 81.9%, mientras que la combinación de SC “categorías del lenguaje oral” y PM “signos de puntuación” cosecha los peores resultados con una medida-F de 65.6%. Con respecto a tres clases los mejores resultados son los obtenidos por la combinación de LP “procesos lingüísticos”, PP “procesos psicológicos”, PC “asuntos personales” y SC “categorías del lenguaje oral” con una medida-F de 77.4% y los peores son los obtenidos por la combinación de PC “asuntos personales” Y PM “marcas de puntuación” con una medida-F de 64.6%. Finalmente, para cinco clases, los mejores resultados son los obtenidos por la combinación de PP “procesos psicológicos” y PM “signos de puntuación” con una medida-F de 55.8%, mientras que los peores son obtenidos por la combinación de SC “categorías del lenguaje oral” y PM “signos de puntuación” con una medida-F de 46%.

Finalmente, estos experimentos también demuestran que utilizando dos clases (positiva, negativa) se obtienen mejores resultados que utilizando tres clases (positiva, neutra, negativa) y cinco clases (positiva, neutra, negativa, muy positiva, muy negativa).

Tabla 4.12. Experimentos en el dominio turístico con el algoritmo J48.

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP	0.753	0.753	0.753	0.729	0.732	0.73	0.513	0.514	0.513
PP	0.77	0.77	0.77	0.751	0.758	0.754	0.546	0.547	0.546
PC	0.666	0.666	0.666	0.68	0.71	0.677	0.478	0.478	0.477
SC	0.646	0.62	0.62	0.606	0.693	0.637	0.461	0.472	0.416
PM	0.649	0.649	0.649	0.643	0.702	0.656	0.474	0.473	0.471
LP-PP	0.767	0.767	0.766	0.77	0.77	0.77	0.555	0.557	0.556
LP-PC	0.767	0.767	0.766	0.721	0.723	0.722	0.499	0.499	0.499
LP-SC	0.746	0.746	0.746	0.737	0.739	0.738	0.523	0.524	0.522

	2 clases			3 clases			5 clases		
	P	R	F	P	R	F	P	R	F
LP-PM	0.751	0.751	0.751	0.727	0.729	0.728	0.531	0.53	0.53
PP-PC	0.745	0.745	0.745	0.746	0.751	0.748	0.545	0.545	0.544
PP-SC	0.77	0.77	0.77	0.769	0.774	0.772	0.546	0.546	0.546
PP-PM	0.763	0.763	0.763	0.755	0.76	0.757	0.559	0.558	0.558
PC-SC	0.688	0.688	0.688	0.683	0.729	0.691	0.482	0.484	0.482
PC-PM	0.671	0.67	0.67	0.638	0.666	0.646	0.498	0.498	0.498
SC-PM	0.656	0.656	0.656	0.654	0.693	0.649	0.461	0.46	0.46
LP-PP-PC	0.811	0.811	0.811	0.763	0.761	0.762	0.533	0.536	0.534
LP-PP-SC	0.815	0.815	0.815	0.765	0.764	0.764	0.55	0.551	0.55
LP-PP-PM	0.813	0.813	0.813	0.751	0.75	0.751	0.517	0.517	0.516
LP-PC-SC	0.765	0.765	0.765	0.74	0.742	0.741	0.493	0.493	0.493
LP-PC-PM	0.764	0.763	0.763	0.741	0.746	0.743	0.503	0.503	0.503
LP-SC-PM	0.768	0.768	0.768	0.734	0.739	0.736	0.543	0.542	0.542
PP-PC-SC	0.763	0.763	0.763	0.76	0.764	0.762	0.549	0.55	0.549
PP-PC-PM	0.779	0.779	0.779	0.75	0.752	0.751	0.539	0.539	0.539
PP-SC-PM	0.779	0.779	0.779	0.772	0.776	0.774	0.553	0.553	0.553
PC-SC-PM	0.68	0.68	0.679	0.667	0.688	0.673	0.486	0.486	0.486
LP-PP-PC-SC	0.81	0.81	0.81	0.776	0.773	0.774	0.53	0.533	0.531
LP-PP-PC-PM	0.818	0.818	0.818	0.761	0.759	0.76	0.525	0.524	0.525
LP-PP-SC-PM	0.816	0.816	0.816	0.763	0.762	0.763	0.529	0.529	0.528
LP-PC-SC-PM	0.767	0.767	0.767	0.731	0.736	0.733	0.499	0.498	0.498
PP-PC-SC-PM	0.765	0.765	0.765	0.749	0.751	0.75	0.536	0.536	0.536
LP-PP-PC-SC-PM	0.819	0.819	0.819	0.772	0.769	0.771	0.536	0.534	0.535

4.3.5. Discusión de los resultados

Los resultados de clasificación para el dominio de películas (ver Figura 4.2) y de turismo (Figura 4.3) muestran que, reduciendo el número de clases, la precisión de los clasificadores incrementa, es decir, la clasificación con dos clases (positivo, negativo) provee mejores resultados que la clasificación con tres clases (positivo, neutro, negativo) y cinco clases (positivo, neutro, negativo, muy positivo, muy negativo). Por lo tanto, con menos clases el algoritmo de clasificación realiza una mejor clasificación. Esto significa que se requieren criterios y características adicionales para obtener una clasificación de grano fino, por ejemplo, en 5 categorías

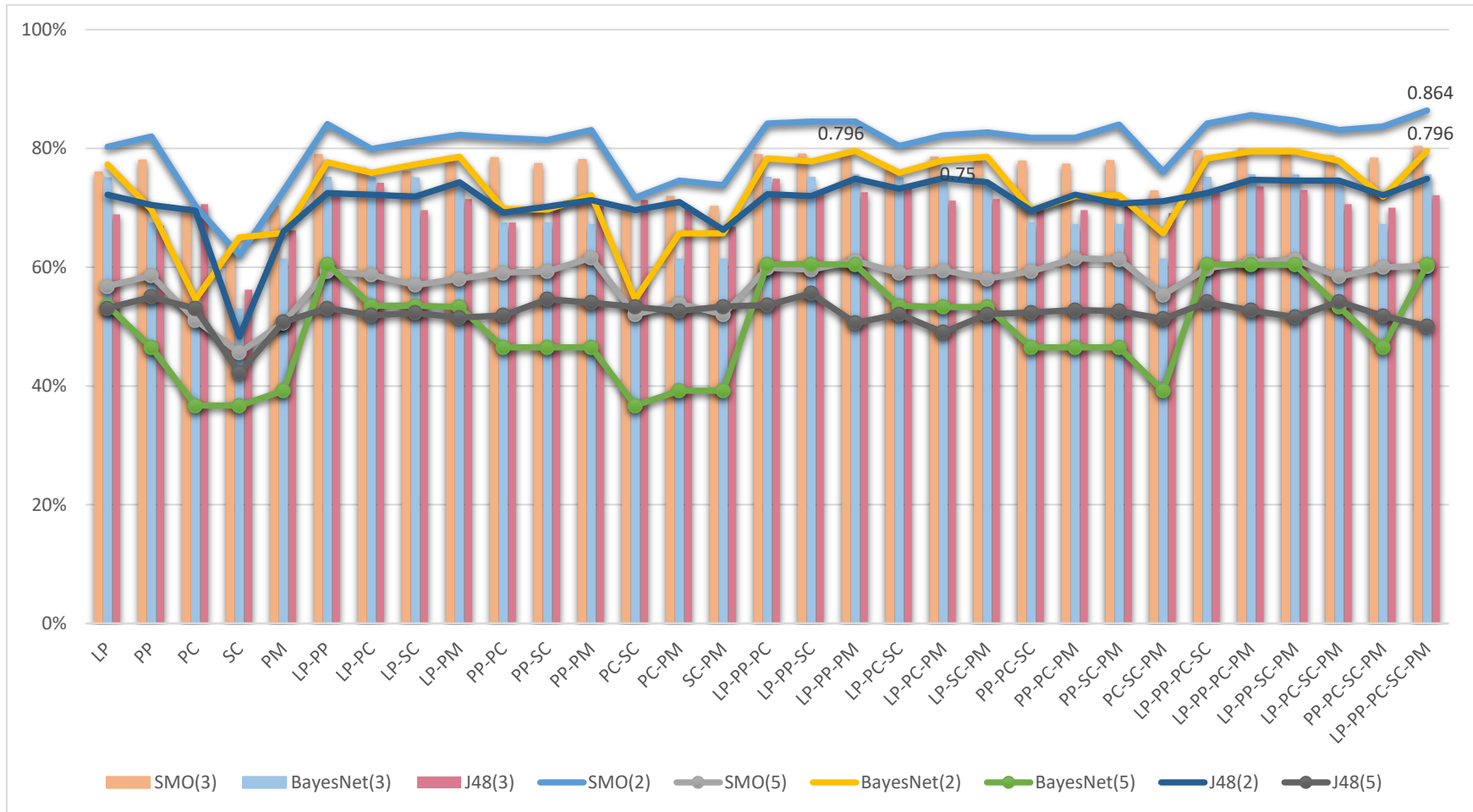


Figura 4.2. Comparación de clases para el dominio de películas.

Capítulo 4. Validación y discusión

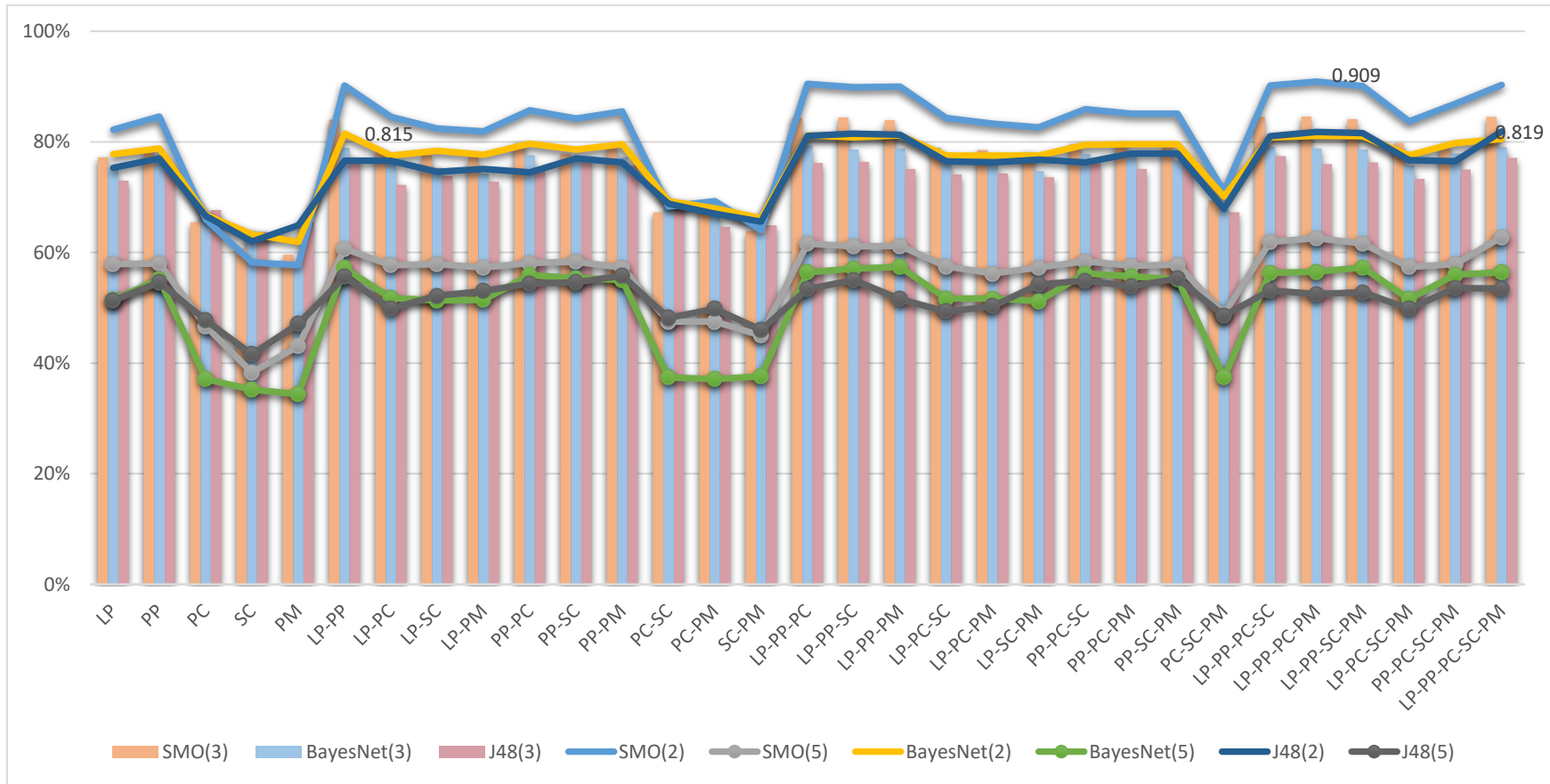


Figura 4.3. Comparación de clases para el dominio de turismo.

La Figura 4.4 presenta los resultados obtenidos por cada algoritmo en el dominio de películas. Como se puede observar, el algoritmo SMO tiene un mejor rendimiento que el algoritmo BayesNet y J48.

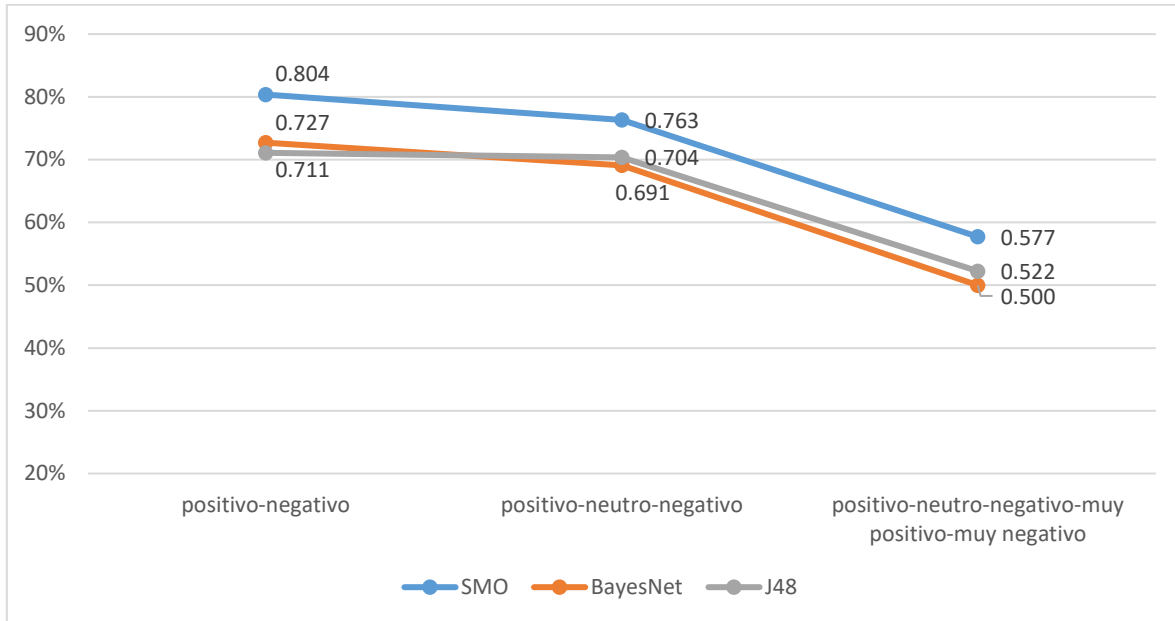


Figura 4.4. Comparación de algoritmos en el dominio de películas.

La Figura 4.5 muestra que SMO obtiene mejores resultados de clasificación que los algoritmos BayesNet y J48 para el dominio de turismo.

En resumen, La Figura 4.4 y Figura 4.5 muestran que para ambos dominios: películas y turismo, SMO obtiene mejores resultados que los algoritmos de clasificación BayesNet y J48. Estos resultados pueden ser justificados por el análisis presentado en (Bhavsar & Ganatra, 2012), donde claramente se muestra que los modelos SVM son más precisos en comparación de otros algoritmos tales como árboles de decisión, redes bayesianas, redes neuronales, por mencionar algunos.

Capítulo 4. Validación y discusión

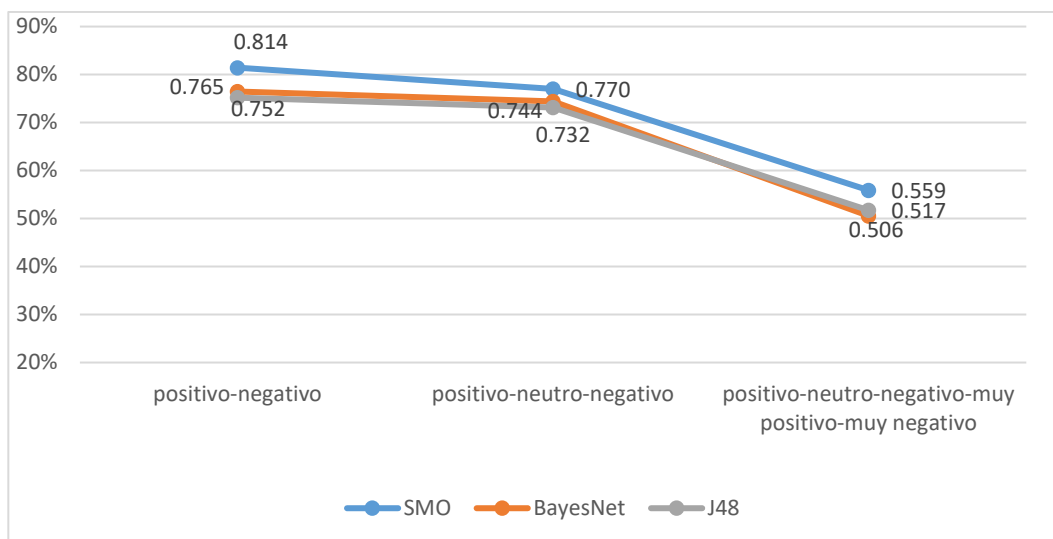


Figura 4.5. Comparación de algoritmos en el dominio de turismo.

Por otro lado, los resultados en ambos dominios muestran que individualmente, las categorías LP “procesos lingüísticos” y PP “procesos psicológicos” obtienen los mejores resultados. La primera categoría, LP “procesos lingüísticos” contiene palabras gramaticales y la segunda categoría palabras relacionadas a los procesos psicológicos tales como: sentimientos positivos, optimismo, ansiedad, tristeza, depresión, entre otras. Las opiniones escritas con frecuencia contienen una gran cantidad de palabras gramaticales y palabras relacionadas con el estado emocional del autor, lo que confirma el potencial discriminatorio de estas dos categorías en los experimentos de clasificación.

4.4. Validación del método de detección de la sátira

En esta sección se presentan los resultados de validación del método para la detección de la sátira. Este método fue evaluado en el dominio de noticias. Las noticias satíricas, también conocidas como noticias falsas en inglés *fake news*, es un tipo de parodia presentada en formato típico del periodismo. Las noticias satíricas son particularmente populares en la Web y específicamente en redes sociales, en las que es fácil imitar una fuente de noticias creíble, por lo que las historias pueden crear una amplia distribución desde cualquier sitio. Las noticias satíricas se confunden a menudo con las noticias legítimas, específicamente cuando se disocia de su fuente

original, como ocurre cuando se recomienda un tweet desde una cuenta de Twitter a través de Facebook.

4.4.1. Corpora

Como se mencionó anteriormente, este método está basado en un enfoque de aprendizaje automático, por lo que se requiere de un corpus etiquetado con el objetivo de distinguir entre una noticia satírica y no satírica. Debido a que la detección de la sátira ha sido poco estudiada, solo algunos corpus están disponibles. Por ejemplo, en (Barbieri et al., 2015b) presentan un corpus en español. Sin embargo, este sólo contiene información en lenguaje español de España y actualmente ya no se encuentra disponible. Además, dado que nuestro objetivo es también llevar a cabo un análisis de la sátira en lenguaje mexicano, se coleccionaron los dos conjuntos de datos.

La colección de tweets desde la API de Twitter, se llevó a cabo usando Twitter4J, una Librería basada en Java que facilita el uso de la API de Twitter. Los tweets fueron obtenidos desde 8 cuentas de Twitter, 4 satíricas (2 de España y 2 de México), y 4 no satíricas (2 de España, y 2 de México) (Ver Tabla 4.13.). Se obtuvieron un total de 10000 tweets satíricos (5000 de España y 5000 de México), y 10000 tweets no satíricos (5000 de España y 5000 de México).

Tabla 4.13. Cuentas de Twitter.

	País	Cuenta de Twitter
Satírico	México	El deforma (@eldeforma)
		El dizque (@eldizque)
	España	El jueves (@eljueves)
		El mundo today (@elmundotoday)
No-satírico	México	El universal (@El_Universal_Mx)
		Excelsior (@Excelsior)
	España	El país (@el_pais)
		El mundo (@elmundoes)

Capítulo 4. Validación y discusión

Una vez que los tweets se obtuvieron, se aplicó un filtrado para eliminar *retweets*, duplicados, tweets escritos en otros idiomas y tweets con solo URLs. Finalmente, se llevó a cabo una revisión manual de los tweets con el objetivo de asegurar que los obtenidos fueran relevantes para el estudio, lo cual resultó en 6829 tweets satíricos (3107 de México y 3722 de España) y 7716 tweets no satíricos (4561 de México y 3155 de España). El corpus final contiene un total de 5000 tweets satíricos y 5000 tweets no satíricos, ya que únicamente se seleccionaron 2500 tweets satíricos de México, 2500 tweets satíricos de España, 2500 tweets no satíricos de México y 2500 tweets no satíricos de España con el objetivo de obtener un corpus balanceado.

A continuación, la Figura 4.6 muestra un ejemplo de un tweet no-satírico, mientras la Figura 4.7 muestra un ejemplo de tweet satírico desde el corpus. El corpus es compartido como una lista de IDs de tweets¹⁰ debido a que por políticas de privacidad de Twitter no es posible compartir el contenido.

¹⁰ <http://dis.um.es/~valencia/SatiricalTwitterDataSet.zip>



Figura 4.6. Tweet no-satírico¹¹



Figura 4.7. Tweet satírico¹².

¹¹ https://twitter.com/el_pais/status/782512990012702720

¹² <https://twitter.com/eldeforma/status/782618385712570368>

4.4.2. Resultados

Como se mencionó anteriormente LIWC contiene un total de 72 características, las cuales son clasificadas en 5 principales conjuntos: procesos lingüísticos (LP), procesos psicológicos (PP), asuntos personales (PC), categorías del lenguaje oral (SC) y signos de puntuación (PM). Los experimentos involucraron el uso de cada conjunto individualmente y todas sus posibles combinaciones.

Este experimento indica qué grupo de características es más interesante y cuál clasificador es el más adecuado para este problema específico. Como se puede observar en la Tabla 4.14., los resultados generales obtenidos por el algoritmo SMO muestran que los conjuntos más relevantes tanto para España como para México son los signos de puntuación (PM), procesos lingüísticos (LP) y procesos psicológicos (PP), y los peores son asuntos personales (PC) y categorías del lenguaje oral (SC). Específicamente, para la clasificación de la sátira mexicana con el algoritmo SMO se obtienen los mejores resultados con LP, PP, y PM con una medida-F de 80.2%, 75.1% y 79.6%, respectivamente. Por lo contrario, los peores resultados son obtenidos por asuntos personales (PC) y categorías del lenguaje oral (SC) con una medida-F de 65.8% y 55.6%. Por otro lado, para la sátira española, se obtienen de igual manera los mejores resultados con procesos lingüísticos (LP), procesos psicológicos (PP) y signos de puntuación (PM), con una medida-F de 77%, 69.1% y 84.9% y los peores resultados son obtenidos con una medida-F de 57.5% para asuntos personales (PC) y de 48.1 para categorías del lenguaje oral (SC).

Los resultados confirman claramente la utilidad de la adopción de categorías lingüísticas, psicológicas y signos de puntuación en la distinción de textos satíricos y no satíricos. Con respecto a las combinaciones de los cinco conjuntos, los resultados muestran que son mejores que los obtenidos por las categorías individuales. Se puede observar que no es suficiente usar solo una categoría de características. Los resultados incrementan añadiendo las características de las categorías lingüísticas y emocionales. Por lo contrario, las categorías del lenguaje oral y los asuntos personales no son tan relevantes.

Tabla 4.14. Experimentos obtenidos con el algoritmo SMO.

	México			España		
	P	R	F	P	R	F
LP	0.802	0.802	0.802	0.771	0.771	0.77
PP	0.751	0.751	0.751	0.664	0.693	0.691
PC	0.679	0.679	0.658	0.633	0.623	0.575
SC	0.606	0.606	0.556	0.69	0.617	0.481
PM	0.818	0.802	0.796	0.877	0.854	0.849
LP-PP	0.825	0.825	0.825	0.778	0.776	0.776
LP-PC	0.813	0.813	0.813	0.773	0.772	0.772
LP-SC	0.8	0.8	0.8	0.772	0.771	0.771
LP-PM	0.865	0.864	0.864	0.885	0.872	0.869
PP-PC	0.764	0.764	0.764	0.702	0.7	0.698
PP-SC	0.751	0.751	0.751	0.695	0.694	0.692
PP-PM	0.839	0.835	0.833	0.878	0.855	0.85
PC-SC	0.683	0.667	0.642	0.651	0.642	0.619
PC-PM	0.822	0.807	0.802	0.878	0.855	0.849
SC-PM	0.818	0.801	0.794	0.878	0.855	0.85
LP-PP-PC	0.828	0.828	0.828	0.78	0.779	0.779
LP-PP-SC	0.825	0.825	0.825	0.777	0.776	0.775
LP-PP-PM	0.879	0.879	0.879	0.891	0.881	0.879
LP-PC-SC	0.812	0.812	0.812	0.773	0.772	0.772
LP-PC-PM	0.868	0.867	0.867	0.885	0.873	0.87
LP-SC-PM	0.865	0.864	0.864	0.886	0.873	0.871
PP-PC-SC	0.765	0.756	0.765	0.701	0.701	0.697
PP-PC-PM	0.841	0.838	0.837	0.878	0.855	0.85
PP-SC-PM	0.84	0.835	0.834	0.878	0.856	0.85
PC-SC-PM	0.823	0.808	0.803	0.878	0.855	0.85

	México			España		
	P	R	F	P	R	F
LP-PP-PC-SC	0.828	0.828	0.828	0.781	0.78	0.779
LP-PP-PC-PM	0.886	0.885	0.885	0.893	0.883	0.881
LP-PP-SC-PM	0.879	0.879	0.879	0.893	0.884	0.882
LP-PC-SC-PM	0.87	0.869	0.869	0.886	0.875	0.872
PP-PC-SC-PM	0.842	0.839	0.839	0.878	0.856	0.851
LP-PP-PC-SC-PM	0.885	0.885	0.885	0.894	0.884	0.883

La Tabla 4.15. presenta los resultados obtenidos por el algoritmo BayesNet. Como se puede observar estos resultados son menores que los obtenidos por el algoritmo SMO. Con el algoritmo SMO los mejores resultados obtenidos son de 88.5% y 88.3% para México y España, en comparación con los obtenidos por el algoritmo BayesNet de 78.7% para México y 76.5 para España. Al igual que en el anterior experimento se muestra que los conjuntos más relevantes tanto para España como para México son los signos de puntuación (PM) con una medida-F de 76.5% y 75.1%; procesos lingüísticos (LP) con una medida-F de 70.3% y 68.8%; y procesos psicológicos (PP) con una medida-F de 67.1% y 65%. Por lo contrario, los peores resultados son asuntos personales (PC) con una medida-F de 59.2% y 55.5%; y categorías del lenguaje oral (SC) con una medida-F de 38.3% y 42.1%. Además, mejores resultados son obtenidos si se lleva a cabo una combinación de las categorías individuales más discriminantes. Estos resultados también confirman claramente la utilidad de la adopción de categorías lingüísticas, psicológicas y signos de puntuación en la distinción de textos satíricos y no satíricos.

Tabla 4.15. Experimentos obtenidos con el algoritmo BayesNet.

	México			España		
	P	R	F	P	R	F
LP	0.703	0.703	0.703	0.688	0.688	0.688
PP	0.671	0.671	0.671	0.651	0.651	0.65
PC	0.617	0.608	0.592	0.669	0.613	0.555
SC	0.3	0.55	0.383	0.671	0.562	0.421
PM	0.78	0.769	0.765	0.769	0.755	0.751
LP-PP	0.722	0.722	0.722	0.691	0.691	0.691
LP-PC	0.709	0.709	0.709	0.69	0.69	0.69
LP-SC	0.703	0.703	0.703	0.689	0.689	0.689
LP-PM	0.778	0.778	0.778	0.728	0.726	0.726
PP-PC	0.686	0.685	0.685	0.661	0.66	0.66
PP-SC	0.671	0.671	0.671	0.653	0.653	0.653
PP-PM	0.773	0.773	0.773	0.735	0.733	0.732
PC-SC	0.617	0.608	0.592	0.645	0.61	0.564
PC-PM	0.766	0.759	0.756	0.763	0.748	0.744
SC-PM	0.78	0.769	0.765	0.77	0.758	0.755
LP-PP-PC	0.731	0.731	0.731	0.695	0.694	0.694
LP-PP-SC	0.722	0.722	0.722	0.693	0.693	0.693
LP-PP-PM	0.783	0.783	0.783	0.763	0.762	0.762
LP-PC-SC	0.709	0.709	0.709	0.692	0.692	0.692
LP-PC-PM	0.781	0.781	0.781	0.723	0.722	0.722
LP-SC-PM	0.778	0.778	0.778	0.73	0.728	0.728
PP-PC-SC	0.686	0.685	0.685	0.661	0.661	0.66
PP-PC-PM	0.774	0.773	0.773	0.729	0.728	0.728
PP-SC-PM	0.773	0.773	0.773	0.738	0.736	0.735
PC-SC-PM	0.766	0.759	0.756	0.764	0.75	0.747

	México			España		
	P	R	F	P	R	F
LP-PP-PC-SC	0.731	0.731	0.731	0.695	0.695	0.695
LP-PP-PC-PM	0.788	0.787	0.787	0.763	0.763	0.763
LP-PP-SC-PM	0.783	0.783	0.783	0.765	0.765	0.765
LP-PC-SC-PM	0.781	0.781	0.781	0.725	0.724	0.724
PP-PC-SC-PM	0.774	0.773	0.773	0.732	0.731	0.731
LP-PP-PC-SC-PM	0.788	0.787	0.787	0.765	0.765	0.765

La Tabla 4.16. muestra los resultados obtenidos por el algoritmo J48. Estos resultados son en general mejores que los obtenidos por BayesNet, pero muy bajos comparados con los obtenidos por el algoritmo SMO. El mejor resultado obtenido por J48 es de una medida-F de 79.4% y 80.5% relativamente mejores que los obtenidos por BayesNet (78.7% para México y 76.5 para España). Sin embargo, los resultados obtenidos por J48 son menores que los obtenidos por SMO (88.5% para México y 88.3% para España). Al igual que en los demás experimentos se muestra el poder discriminatorio de las características lingüísticas (70.7% para México y 68.4% para España) y signos de puntuación (77% para México y 75.9% para España) en la clasificación de la sátira.

Tabla 4.16. Experimentos obtenidos con el algoritmo J48

	México			España		
	P	R	F	P	R	F
LP	0.707	0.707	0.707	0.684	0.684	0.684
PP	0.662	0.661	0.66	0.65	0.65	0.649
PC	0.626	0.618	0.605	0.632	0.624	0.615
SC	0.598	0.554	0.408	0.671	0.562	0.421
PM	0.791	0.775	0.77	0.773	0.762	0.759
LP-PP	0.719	0.719	0.719	0.699	0.699	0.699

	México			España		
	P	R	F	P	R	F
LP-PC	0.698	0.698	0.698	0.688	0.688	0.688
LP-SC	0.704	0.704	0.704	0.691	0.691	0.691
LP-PM	0.793	0.793	0.793	0.764	0.763	0.763
PP-PC	0.691	0.69	0.69	0.654	0.653	0.652
PP-SC	0.672	0.671	0.669	0.654	0.652	0.651
PP-PM	0.771	0.771	0.77	0.749	0.748	0.748
PC-SC	0.634	0.625	0.612	0.639	0.633	0.626
PC-PM	0.782	0.775	0.772	0.761	0.754	0.752
SC-PM	0.791	0.775	0.77	0.78	0.767	0.764
LP-PP-PC	0.716	0.716	0.716	0.691	0.691	0.691
LP-PP-SC	0.72	0.72	0.72	0.698	0.698	0.698
LP-PP-PM	0.776	0.776	0.776	0.797	0.797	0.797
LP-PC-SC	0.698	0.698	0.698	0.695	0.695	0.695
LP-PC-PM	0.794	0.794	0.794	0.75	0.75	0.75
LP-SC-PM	0.794	0.794	0.794	0.765	0.765	0.765
PP-PC-SC	0.689	0.688	0.688	0.654	0.653	0.653
PP-PC-PM	0.77	0.77	0.77	0.755	0.755	0.755
PP-SC-PM	0.772	0.771	0.771	0.746	0.746	0.745
PC-SC-PM	0.781	0.774	0.771	0.773	0.762	0.76
LP-PP-PC-SC	0.716	0.716	0.716	0.695	0.695	0.695
LP-PP-PC-PM	0.784	0.784	0.784	0.803	0.803	0.803
LP-PP-SC-PM	0.777	0.777	0.777	0.799	0.799	0.799
LP-PC-SC-PM	0.793	0.793	0.793	0.765	0.765	0.765
PP-PC-SC-PM	0.768	0.768	0.768	0.749	0.749	0.749
LP-PP-PC-SC-PM	0.783	0.783	0.783	0.805	0.805	0.805

Capítulo 4. Validación y discusión

Como conclusión, los resultados obtenidos muestran que las categorías más relevantes son las de signos de puntuación (PM) con una medida-F de 79.6% (SMO), 76.5% (BayesNet), 77% (J48) para México y 84.9%(SMO), 75.1% (BayesNet), 75.9% (J48) para España y procesos lingüísticos (LP) con una medida-F de 80.2% (SMO), 70.3% (BayesNet), 70.7% (J48) para México y 77% (SMO), 68.8%(BayesNet) 68.4% (J48) para España. Los resultados también muestran que la combinación de características de las categorías relevantes antes mencionadas, proporcionan mejores resultados. Por ejemplo, si solo se usan las características de signos de puntuación (PM) y el clasificador SMO, se obtiene una medida-F de 79.60% y 84.90 para España y México, respectivamente. Sin embargo, si procesos lingüísticos (LP) y signos de puntuación (PM) se combinan, entonces la medida-F incrementa a 86.4% y 86.9% para México y España. Por último, los experimentos muestran que con todas las categorías se obtienen los mejores resultados con una precisión de 88.5%, exhaustividad de 88.5 %, una medida-F de 88.3% para México y una precisión de 89.4%, exhaustividad de 88.4 %, una medida-F de 88.3% para España, lo que indica que algunas características contenidas en otras categorías son también relevantes, como las contenidas en el proceso psicológico (PP).

Finalmente, como se menciona en la sección 3.3.4 se lleva a cabo una etapa de análisis de características para comprender la importancia de cada característica individualmente en cada modelo obtenido y para obtener las características más discriminantes para la detección de la sátira. Este proceso consistió en seleccionar las características con los valores más altos en base a los resultados obtenidos por el método ganancia de información. La Figura 4.8 y Figura 4.9 muestra las 13 características más relevantes para la detección de la sátira para España y México, a saber: "Adverb", "Afect", "Apostro", "Certeza", "Cuantif", "Colon", "EmoPos", "Exclam", "Informal", "MecCog", "Present", "Quote" y "Social". Las características con mejor valor asignado están relacionadas con el proceso lingüístico, el proceso psicológico y los signos de puntuación. Esto confirma la importancia de este tipo de características para la detección de la sátira.

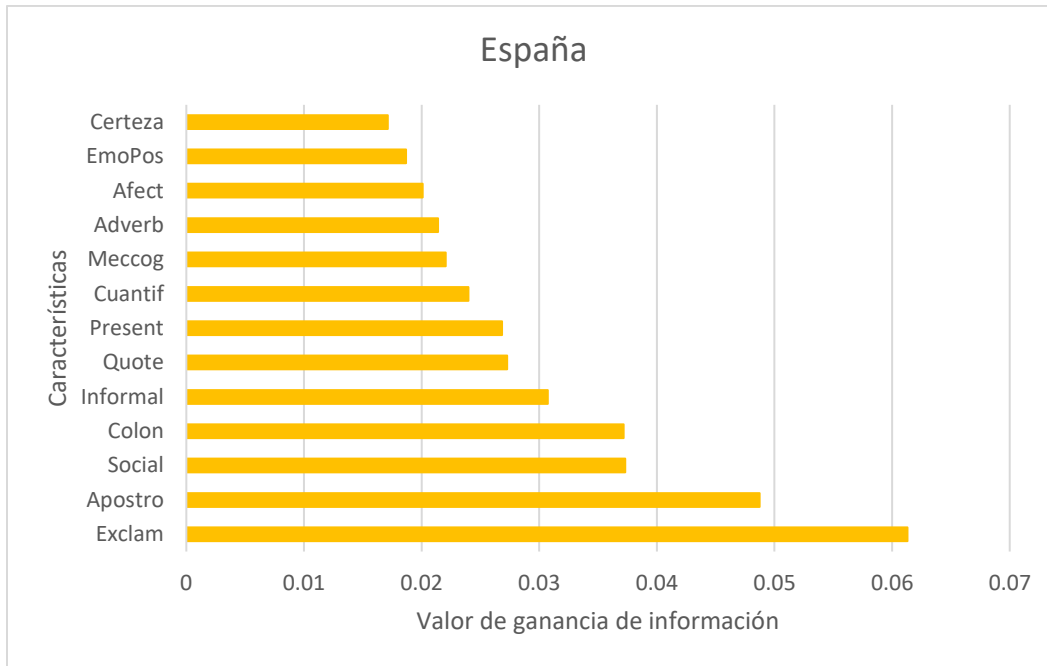


Figura 4.8. Características relevantes para España.

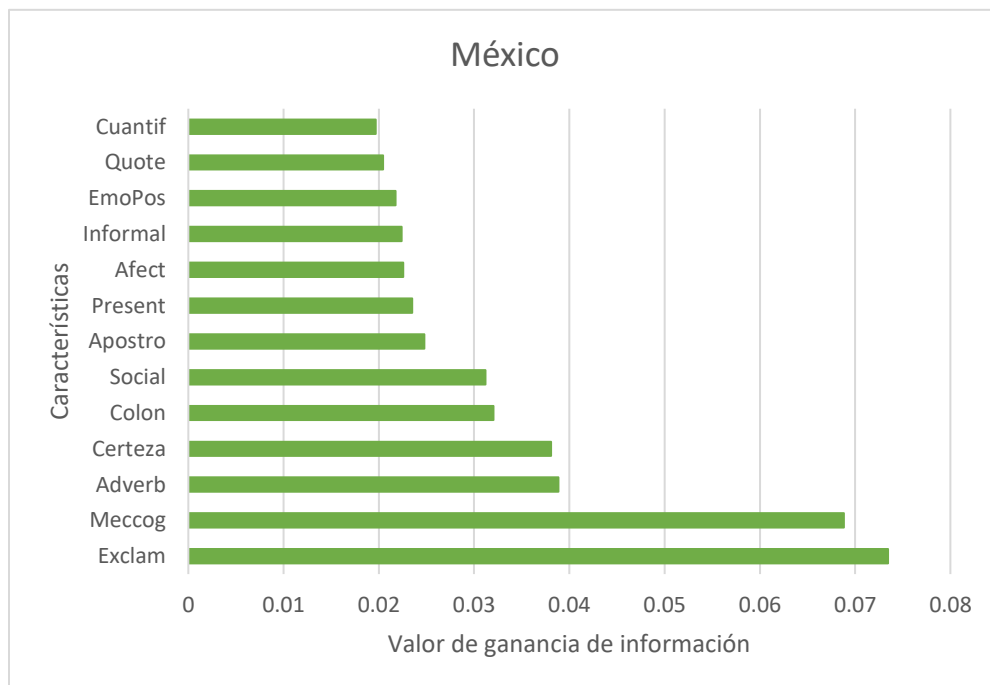


Figura 4.9. Características relevantes para México.

4.4.3. Discusión de resultados

Los resultados muestran que el método propuesto obtiene resultados prometedores para la detección de tweets satíricos y no satíricos. Además, confirman la importancia de características lingüísticas, emocionales y signos de puntuación para la detección de la sátira.

Como se puede observar en la Figura 4.10, los tweets satíricos contienen más adverbios, cuantificadores y palabras en tiempo presente que los tweets no satíricos. En el caso de procesos psicológicos, cinco características fueron predictores significativos: procesos sociales, procesos afectivos, emociones positivas, procesos cognitivos y certeza. Finalmente, con respecto a los signos de puntuación, dos puntos, exclamación, apóstrofes y las comillas fueron predictores significativos. Las conclusiones que se pueden extraer de estos resultados son:

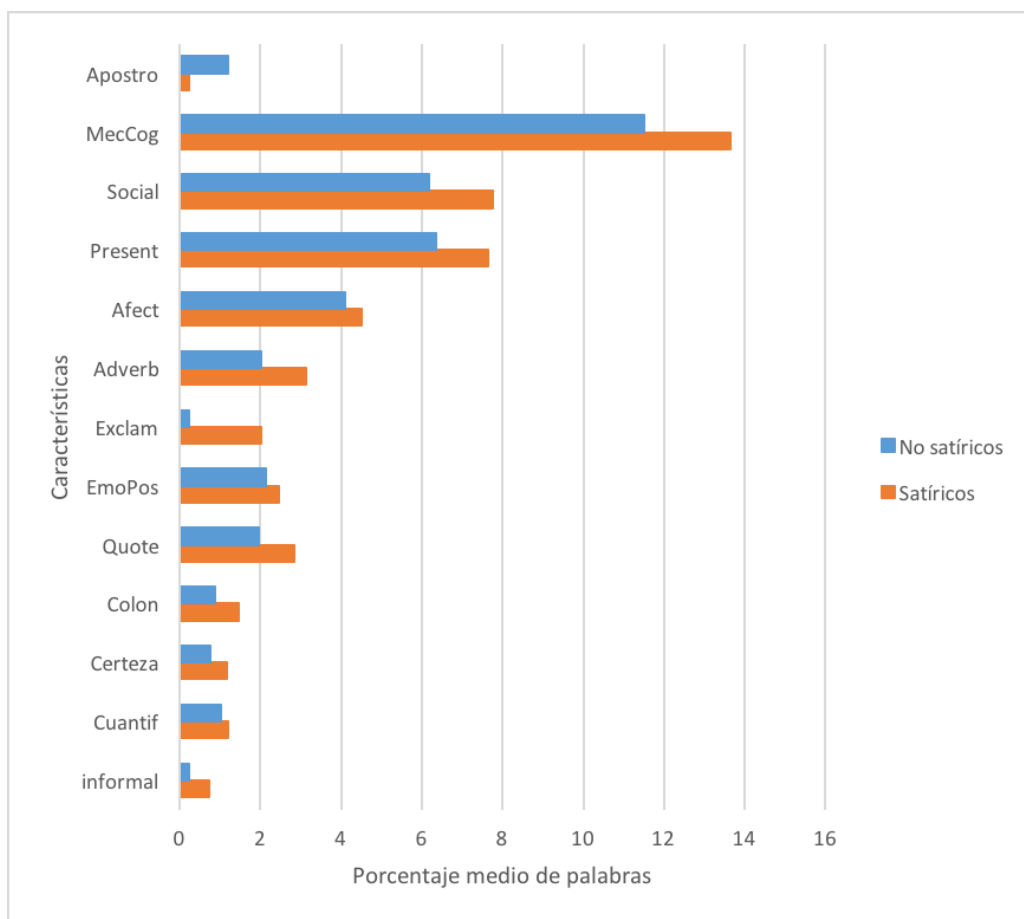


Figura 4.10. Porcentaje medio de palabras de tweets satíricos y no satíricos.

Los tweets satíricos contienen más palabras positivas (“EmoPos”) con el objetivo de dar un significado distinto a una cuestión negativa, es decir, un problema social, abuso o cualquier situación indignante que el usuario desee divulgar. Por ejemplo, en el tweet presentado en la Figura 4.11 “Confirman que mensaje de Roger Waters llegó al corazón de Peña. México ahora es un mejor país. Se acabó la violencia y hay opulencia”. El tweet está compuesto de palabras positivas. Sin embargo, este tweet está en realidad intentando mostrar indignación ante la situación que el país enfrenta debido al mal gobierno.



Figura 4.11. Tweet satírico con emociones positivas.

Los mecanismos cognitivos (“MecCog”). Por ejemplo: causa, saber, deber son indicativos de un lenguaje más complejo (Tausczik & Pennebaker, 2010). Por lo tanto, la alta tasa de procesos cognitivos se atribuye al hecho de que la sátira es más compleja y difícil de identificar que el lenguaje literal. Por ejemplo, los tweets presentados en la Figura 4.12 contienen varias palabras como “sabe”, “que”, “sin”, “pero”, “prefiere”, “preguntar”, “igual”, “porque”, “que”, “siempre” y “pensáis”, los cuales pertenecen a los mecanismos cognitivos.

Capítulo 4. Validación y discusión



Figura 4.12 Tweets satíricos con palabras relacionadas a mecanismo cognitivos

Las categorías certeza (“Certeza”), informal (“Informal”), Afectiva (“Afect”) y cuantificadores (“Cuantif”) están fuertemente relacionadas con la hipérbole, que es un fenómeno recurrente en las noticias satíricas en las que una situación real es exagerada hasta convertirse en absurda. Se atribuye esto a las siguientes razones: 1)

Las palabras de certeza tal como “siempre” y “nunca” son algunos ejemplos que generalmente exageran una situación; 2) El lenguaje informal es narrativo, emocional e interactivo. En la literatura estas características se ha asociado con la exageración (Lindey, 2016) (McCarthy & Carter, 2004); 3) Los hallazgos en el trabajo de (McCarthy & Carter, 2004) revelaron que la exageración está asociada con un conjunto de características afectivas; y 4) los cuantificadores de masas tales como algo, todo y bastante son características que normalmente aparecen en textos hiperbólicos. Los tweets presentados en Figura 4.13 son ejemplos de hipérbole en los cuales una situación es exagerada. En el primer ejemplo, lenguaje informal y palabras de certeza tales como “siempre” fueron usadas para este objetivo. En el segundo tweet, palabras afectivas tales como “teme” y cuantificadores tales como “todo” fueron usadas para aumentar la veracidad.

Capítulo 4. Validación y discusión



Figura 4.13. Ejemplos de tweets con uso de hipérbole.

Los adverbios son también buenos indicadores de tweets satíricos, los cuales son usados para exagerar o minimizar una situación (Kovaz, Kreuz, & Riordan, 2013). En el caso de los tweets presentados en la Figura 4.14, los adverbios “tan” y “muy” se utilizan para intensificar el significado.



Figura 4.14. Tweets satíricos con adverbios.

Los signos de puntuación tales como: signos de exclamación (“Exclam”), dos puntos (“Colon”) y comillas se utilizan con frecuencia para denotar sátira. En base a los hallazgos de (Kreuz & Caucci, 2007) y (Sulis, Hernández Farías, Rosso, Patti, & Ruffo, 2016), se concluye que la sátira comparte estas características con ironía y

Capítulo 4. Validación y discusión

sarcasmo. Los tweets presentados en la Figura 4.15 son claros ejemplos de tweets satíricos que contienen signos de puntuación como exclamación (!), dos puntos (:), y comillas (“”, «»).



Figura 4.15. Tweets satíricos con signos de puntuación.

Además, los resultados muestran que el apóstrofe (“Apostro”) es comúnmente más usado en tweets no satíricos. Por ejemplo, la Figura 4.16 muestra dos ejemplos de tweets no satíricos en los cuales apóstrofes son utilizados.



Figura 4.16. Tweets no satíricos con signos de puntuación.

La categoría social es también relevante en estos resultados. Sin embargo, esta es más dependiente del contenido. Es decir, la sátira de noticias suele representar

Capítulo 4. Validación y discusión

problemas sociales actuales y por lo tanto se considera que realmente no es discriminante para su detección. Por ejemplo, los tweets presentados anteriormente (ver Figura 4.12 Figura 4.13 y Figura 4.15) se relacionan con aspectos sociales como la política.

La situación de la categoría “Presente” es similar a la categoría “Social” porque es dependiente del contexto. Por ejemplo, los autores en (Skalicky & Crossley, 2015) encontraron que el tiempo pasado es más discriminante en las opiniones de Amazon¹³ debido a que los textos satíricos presentan una historia o narrativa relacionada con el producto del cual se estaba opinando, y como tal se encontraron más marcas verbales en tiempo pasado. Sin embargo, nuestros resultados muestran que las noticias satíricas tienen más marcas verbales en tiempo presente. Lo cual se atribuye al hecho de que las noticias frecuentemente expresan situaciones actuales. Por ejemplo, considerando algunos de los tweets presentados anteriormente, se puede observar que el escritor satiriza noticias actuales (en el momento de la publicación). Como se puede ver en la Tabla 4.17., se identificaron varios verbos en el tiempo presente, lo que demuestra que los tweets de noticias satíricas se escriben comúnmente en ese tiempo.

¹³ <https://www.amazon.com>

Tabla 4.17. Ejemplos de tweets satíricos con palabras en presente.

Tweet	Categoría Presente
"Rajoy no sabe por qué lleva un año sin trabajar pero prefiere no preguntar nada para no romper la racha"	Sabe, lleva, prefiere
"Igual el gobierno está vaciando la hucha de las pensiones porque debajo del dinero hay un Pokemon. Es que, joder, siempre pensáis lo peor."	Está, hay, es, pensáis
"Sony lanza un teléfono muy resistente porque cree que eres imbécil y se te va a caer"	Lanza, cree, eres, va
"Los anuncios de la Lotería son tan falsos que ya ni siquiera ponen personas de verdad en ellos"	Son, ponen
"Si escuchas el discurso de Pedro Sánchez al revés se oye el mensaje: "¡Ayuda, Albert Rivera es un reptiliano!""	Oye, es

Como se puede observar en la Figura 4.17, no hay diferencias significativas entre la sátira mexicana y la sátira española. Aunque las mismas características discriminantes fueron obtenidas para ambos conjuntos de datos, existen diferencias con respecto al porcentaje de su uso en las noticias satíricas. Por ejemplo, hay una gran diferencia entre las noticias satíricas españolas y mexicanas con respecto al uso de signos de exclamación y mecanismos cognitivos, ya que son más utilizados en la sátira mexicana. Es decir, los escritores mexicanos usan con más frecuencia signos de exclamación para modificar el tono y denotar sátira en el texto. Además, el uso de palabras relacionadas con mecanismos cognitivos ("MecCog") como "admitir", "afecta", "sabe", "pero", "algún", "causa" son frecuentemente más usados en la sátira mexicana que en la española.

Con respecto a otras características como social, presente ("Present"); emociones positivas ("EmoPos"); signos de puntuación como "Colon" y "Quote"; informal, cuantificadores ("Cuanti"), certeza, y adverbios ("Adverb") no hay una diferencia significativa en su uso. Por tanto, estas características son usadas con la misma frecuencia en noticias satíricas españolas y mexicanas. Además, los escritores satirizan problemas sociales actuales tales como políticos, por lo que características

Capítulo 4. Validación y discusión

como presente y social desempeñan un papel importante. Por último, los adverbios y las palabras de certeza (“Certeza”) como “siempre”, “jamás”, “muy”, “nunca” son utilizados por los escritores españoles para exagerar o minimizar una situación.

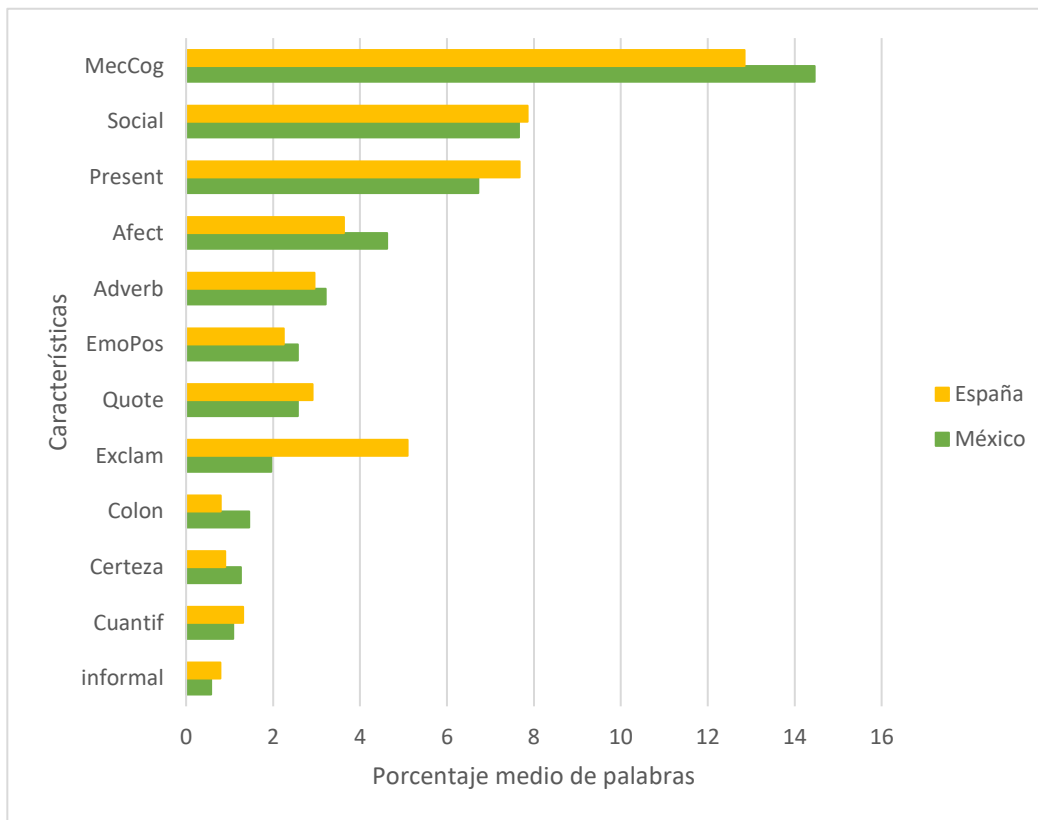


Figura 4.17. Porcentaje medio de palabras de características de España y México.

La métrica del área ROC se utiliza en este trabajo para evaluar el desempeño y diseño de cada clasificador. Esta métrica se representa de forma equivalente trazando la fracción de positivos verdaderos (*TPR-true positive rate*), también conocida como sensibilidad con respecto a la fracción de falsos positivos (*FPR- false positive rate*), que se calcula como (1- especificidad). La representación de (1-especificidad, sensibilidad) se denomina curva ROC. Una curva que está más cerca del punto superior izquierdo significa que el clasificador es más preciso, mientras que una curva que está más cerca de la diagonal significa que el clasificador es menos preciso (Gao, Lee, & Lim, 2006) (Gurcan & Birturk, 2016). En este trabajo de tesis doctoral se utiliza la interfaz de flujo de conocimiento de WEKA con el objetivo de obtener la curva ROC para cada clasificador y realizar una comparación entre el rendimiento de los clasificadores.

La Figura 4.18 y Figura 4.19 presentan los resultados de evaluación obtenidos por los tres algoritmos a través de la curva ROC. Estas figuras muestran tres curvas, correspondiente a cada algoritmo. La curva más delgada representa el algoritmo SMO, la curva más gruesa representa al clasificador BayesNet y la otra representa el algoritmo J48. Como se puede observar, la curva SMO está más cerca del punto superior izquierdo en ambos casos, es decir, en el caso de conjunto de datos de España y México. Esto indica que SMO es más preciso que los otros clasificadores. Las curvas J48 y BayesNet están por su parte más cerca de la diagonal lo que indica que estos clasificadores son menos precisos. Los resultados coinciden con los reportados por otros trabajos en la literatura donde se concluye que SVM es uno de los mejores algoritmos de clasificación y supera algoritmos como J48, BayesNet, entre otros (Shboul et al., 2015) (Kaya, Fidan, & Toroslu, 2013) (Yijing et al., 2016) (Moradi & Rostami, 2015). Además, estos resultados de la evaluación se justifican con el análisis de varios clasificadores reportado en (Bhavsar & Ganatra, 2012), en el que se ha demostrado claramente que los modelos SVM son más robustos y precisos que otros algoritmos, incluidos los comparados en este trabajo. Otros hechos que justifican estos resultados corresponden a la robustez de SMO en espacios de alta dimensión y en conjuntos de muestras escasos donde la mayoría de los problemas de categorización de texto son linealmente separables (Rushdi Saleh et al., 2011).

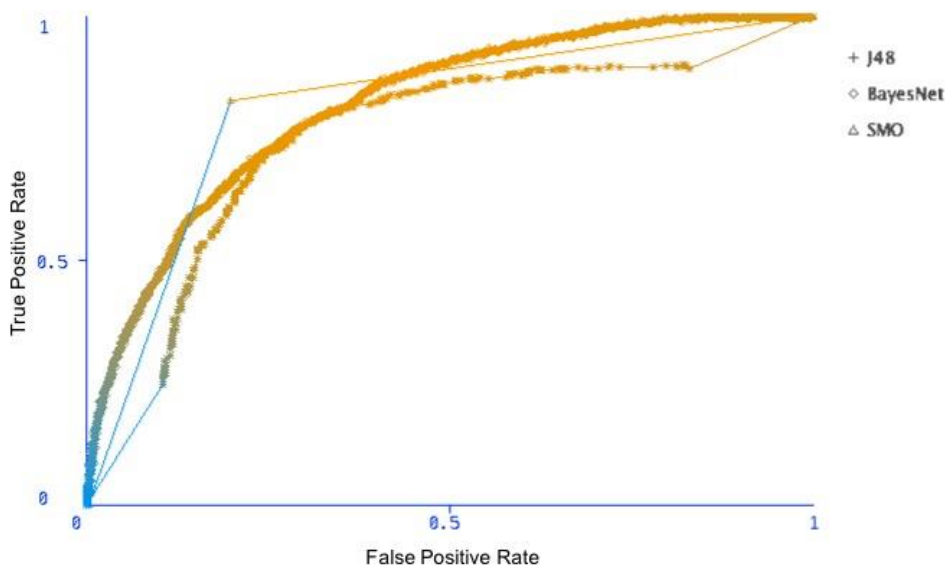


Figura 4.18. Curva ROC para el conjunto de datos de México.

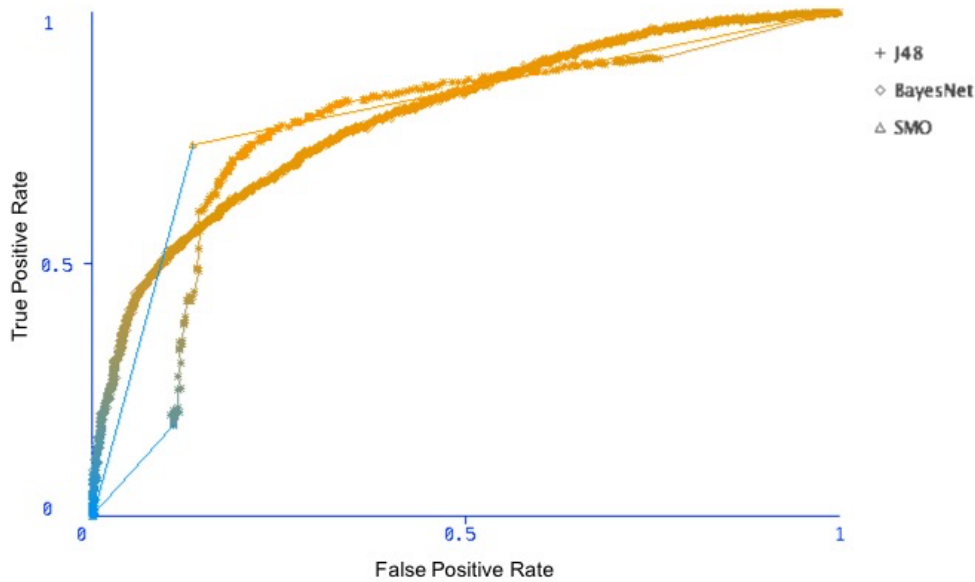


Figura 4.19. Curva ROC para el conjunto de datos de España.

4.5. Conclusiones

En este capítulo se presentó la validación del método propuesto para el análisis de sentimientos y la detección de la sátira. Por un lado, la evaluación de la clasificación de sentimientos se llevó a cabo en dos dominios diferentes: películas y turismo. El corpus de películas es un corpus de referencia que ha sido utilizado por investigadores con la finalidad de evaluar sus métodos. Esto nos permite demostrar la efectividad del método propuesto. Por otro lado, la evaluación del método de la sátira se llevó a cabo en el dominio de las noticias.

En la sección 4.2 se presentan las diferentes medidas de evaluación que fueron utilizadas durante el proceso de evaluación del método propuesto, a saber, precisión, exhaustividad y medida-F.

En la sección 4.3 se evaluó el método del análisis de sentimientos. En primer lugar, en la sección 4.3.4 se describe el corpus utilizado en el dominio de películas y a continuación se presentan los resultados obtenidos. Estos resultados demostraron que el algoritmo que provee mejores resultados es SMO respecto a J48 y BayesNet. Mientras que el algoritmo que obtuvo peores resultados fue J48. Por otro lado, los resultados también demostraron la importancia de las categorías lingüísticas y

psicológicas para el análisis de sentimientos. Además, se pudo observar que la combinación de estas características provee mejores resultados en combinación que individualmente.

Posteriormente, en la sección 4.3.4.1 se evalúa el método en el dominio turístico. Primero, se describió el corpus utilizado, el cual fue construido con opiniones del sitio Web TripAdvisor, luego se presentaron los resultados y una discusión de las categorías más discriminantes. Los resultados obtenidos para este dominio son prometedores y demuestran al igual que en el dominio de películas la importancia de las características lingüísticas y psicológicas para el análisis de sentimientos. Además, estos resultados confirman que SMO es el algoritmo más adecuado para este problema de clasificación.

En la sección 4.4 se evaluó el método para la detección de la sátira. Primero, se describió el corpus, el cual fue obtenido desde diversas cuentas Twitter de noticias satíricas y verídicas. Posteriormente se presentaron los resultados, donde claramente se confirma la relevancia de los signos de puntuación y características lingüísticas y emocionales para la detección de tweets satíricos y no satíricos. Además, los resultados también demostraron que SMO es el algoritmo más indicado para este problema de clasificación, superando los resultados obtenidos por los algoritmos BayesNet y J48.

Capítulo 5. Conclusiones y líneas futuras

5.1. Conclusiones

El análisis de sentimientos es un problema que requiere de una amplia labor de investigación. Existen problemas que aun necesitan ser resueltos. Por ejemplo, el lenguaje figurado es un importante aspecto a ser considerado en los sistemas de análisis de sentimientos. Sin embargo, este tipo de lenguaje es muy difícil de detectar, y aún más las variaciones que existen como la ironía, el sarcasmo, la sátira, entre otros.

El trabajo en esta tesis describió un método para la detección de patrones psicolingüísticos para el análisis de sentimientos y la detección automática de la sátira. Las características psicolingüísticas resultaron ser efectivas junto con técnicas de procesamiento de lenguaje natural y minería de datos para la detección de sentimientos y de la sátira. Además, la validación de los métodos en diversos dominios ha demostrado su efectividad para la clasificación de opiniones y tweets.

5.2. Contribuciones

Es esta tesis se presenta soluciones para el análisis de sentimientos y detección de la sátira a través de la extracción de características psicolingüísticas. Las soluciones presentadas se han aplicado con éxito en varios dominios. Las principales aportaciones de esta tesis son las siguientes:

- **Desarrollo de un método para la clasificación de sentimientos y detección de la sátira.** Este método permite clasificar opiniones como positivas, negativas, neutras, muy positivas y muy negativas y tweets como satíricos y no satíricos. El método puede ser adaptado a diversos problemas de clasificación de textos e idiomas. Sin embargo, este requiere un corpus etiquetado como entrada.
- **Proceso para el pre-procesamiento de tweets en español:** La normalización de textos extraídos de redes sociales tal como Twitter suele ser

más difícil debido a que los usuarios suelen abreviar palabras y usar jerga debido a la limitación de 140 caracteres que tienen los tweets. Actualmente, existen pocas herramientas del procesamiento del lenguaje natural que permiten normalizar estos textos en español. Para ello, nosotros hemos definido un proceso que permite normalizar los tweets para procesarlos posteriormente como un texto normal. El proceso consiste de tres principales pasos: 1) tokenización del texto y detección de entidades tales como URLs, menciones y etiquetas; 2) eliminar los elementos detectados en el paso 1 con excepción de etiquetas donde sólo es eliminado el “#”; y 3) extensión de abreviaturas y corrección de ortografía. Este proceso es de suma importancia, ya que actualmente Twitter está siendo un foco de investigación debido a la gran cantidad de información subjetiva contenida en estas redes sociales, la cual está constituida principalmente por opiniones.

- **Desarrollo de un corpus en el dominio del turismo.** Los corpora son un recurso importante en el análisis de sentimientos. Por un lado, los métodos basados en un enfoque de aprendizaje automático requieren de corpus etiquetados con el objetivo de entrenar algoritmos de clasificación. Por otro lado, estos corpus sirven como base para la evaluación de sistemas de análisis de sentimientos. El desarrollo de un corpus requiere esfuerzo y tiempo debido a que el etiquetado se realiza manualmente con el objetivo de obtener un corpus de calidad. Sin embargo, hoy en día existen pocos corpus disponibles en español en la comunidad investigadora, es por ello, que el corpus obtenido en este trabajo de tesis supone una gran aportación.
- **Desarrollo de un corpus de tweets satíricos.** Este corpus consiste en un conjunto de tweets etiquetados como satíricos y no satíricos extraídos desde diversas cuentas de Twitter. Actualmente existen algunos corpus del lenguaje figurado como ironía y sarcasmo. Sin embargo, hay una carencia de corpus con información satírica, sobre todo en español. Este corpus además de ser en este idioma, está dividido en sátira mexicana y satírica española.
- **Detección de características psicolingüísticas para el análisis de sentimientos.** Otra aportación relevante de esta tesis se centra en la

identificación y extracción de características psicolingüísticas que son más discriminantes para el análisis de sentimientos y detección de la sátira.

5.3. Líneas futuras

Con respecto a investigación futura, se proveen varios aspectos que no han sido considerados como parte de esta tesis. Sin embargo, son considerados como líneas de investigación futuras a explorar. A continuación, se detalla cada uno de estos aspectos.

- **Integrar técnicas que permitan proveer un mejor soporte del proceso de normalización ante casos como tweets.**

La normalización de textos como tweets es una tarea muy difícil debido a que normalmente son textos con palabras abreviadas y con faltas de ortografía. En este trabajo de tesis, se propone un proceso para su normalización. Sin embargo, tiene una limitación en cuanto al procesamiento de etiquetas en inglés *hashtags*. Una etiqueta puede contener múltiples palabras juntas. Por lo que considerar técnicas tales como la presentada en (Bejcek, Stranák, & Pecina, 2013), permitirá abordar este problema.

- **Aplicación del método a diversos dominios.**

El método de análisis de sentimiento desarrollado en esta tesis ha sido favorablemente aplicado en los dominios turístico y de películas. Por lo que la aplicación a otros dominios sería otra de las líneas de investigación a explorar como posible línea futura a tener en cuenta. Sin embargo, como se mencionó anteriormente, se requiere de un corpus del dominio etiquetado. Es por ello, que se propone el desarrollo de nuevos corpus en diversos dominios. Un área de especial interés es el dominio médico, el cual ha sido poco explorado. Sin embargo, las opiniones pueden ser de gran interés entre pacientes sobre todo cuando padecen de enfermedades que requieren de autogestión como la diabetes, asma, cáncer, hipertensión, etc. Por otro lado, en cuanto al lenguaje figurado, este trabajo está enfocado en la detección de la sátira, por lo que la creación de nuevos corpus para la ironía, sarcasmo y sátira en diversos dominios, permitiría el desarrollo y evaluación de nuevos sistemas de análisis de sentimiento.

- **Aplicación del método en otros idiomas.**

La aplicación del método se ha enfocado en el idioma español, por lo que como trabajo a futuro se aplicará este método a otros idiomas tales como inglés, francés, árabe y a diversas variedades de español, como Argentina, Uruguay, Venezuela, etc. Esto permitirá determinar si los patrones psicolingüísticos detectados en esta tesis pueden contribuir también a la detección del análisis de sentimientos y sátira en diferentes idiomas y culturas.

- **Detección de patrones psicolingüísticos para el sarcasmo y la ironía.**

El procedimiento para detectar patrones psicolingüísticos únicamente ha sido diseñado para la sátira. Por lo que sería muy interesante también detectar patrones psicolingüísticos para el sarcasmo e ironía. Además, esto permitiría determinar el nivel de similitud entre estos tipos de lenguaje figurado, es decir, determinar qué categorías psicolingüísticas comparten.

- **Integración del sistema de detección del lenguaje figurado en el análisis de sentimientos.**

Los sistemas presentados en este trabajo de tesis doctoral son independientes. Por lo que la incorporación de un módulo que permita detectar la ironía, el sarcasmo y la sátira, así como otros tipos de lenguaje figurado como el humor en el sistema de análisis de sentimientos, permitirá no sólo detectar la polaridad de la opinión, sino también detectar si el texto es literal, irónico, sarcástico o satírico.

- **Contribuir al enriquecimiento de LIWC en español.**

La extracción de características depende en gran medida del diccionario de LIWC. Sin embargo, este diccionario carece de algunas palabras del español como verbos y de una gran variedad de palabras utilizadas en diversos países como Venezuela, Colombia, Ecuador, etc. El diccionario de LIWC puede ser enriquecido con otras palabras, lo cual se traduciría en una mejor extracción de características y, por tanto, una mejor precisión del sistema.

Capítulo 6. Contribuciones científicas

6.1. Publicaciones JCR

1. Salas-Zárate María del Pilar, Paredes-Valverde Mario Andrés, Rodríguez-García Miguel Ángel, Valencia-García Rafael, Alor-Hernández Giner. Automatic Detection of Satire in Twitter: A psycholinguistic-based approach. Knowledge-Based Systems. **Impact Factor: 3.325. (Por aparecer)**
2. Salas-Zárate María del Pilar, Valencia-García Rafael, Ruiz-Martínez Antonio, Colomo-Palacios Ricardo. Feature-based opinion mining in financial news: An ontology-driven approach. Journal of Information Science, 2016. DOI: 10.1177/0165551516645528. **Impact Factor: 1.158.**
3. Salas-Zarate María del Pilar, Paredes-Valverde Mario Andrés, Limon-Romero Jorge, Tlapa Diego, Baez-Lopez Yolanda. Sentiment Classification of Spanish Reviews: An Approach based on Feature Selection and Machine Learning Methods. Journal of Universal Computer Science, 2016. 22(5), 691-708. **Impact Factor: 0.466.**
4. Salas-Zárate María del Pilar, López-López Estanistalo, Valencia-García Rafael, Aussenac-Gilles Nathalie, Almela Ángela, Alor-Hernández Giner. A study on LIWC categories for opinion mining in Spanish reviews. Journal of Information Science, 2015. 40(6), 749-760. ISSN: 0165-5515. **Impact Factor: 1.158.**

6.2. Congresos Internacionales

1. Salas-Zárate María del Pilar, Medina-Moreira José, Álvarez-Sagubay Paul Javier, Lagos-Ortiz Katty, Paredes-Valverde Mario Andrés, Valencia-García Rafael. Sentiment Analysis and Trend Detection in Twitter. Technologies and Innovation, vol. 658, pp. 63-76. Springer International Publishing. 2016.
2. López-López Estanistalo, Salas-Zárate María del Pilar, Almela Ángela, Rodríguez-García Miguel Ángel, Valencia-García Rafael, Alor-Hernández Giner. LIWC-Based Sentiment Analysis in Spanish Product Reviews. In Distributed Computing and Artificial Intelligence, 11th International Conference, vol. 290,

Capítulo 6. Contribuciones científicas

pp. 379-386. Springer International Publishing. ISSN: 2194-5357. 2014.
Indexado en Scopus.

3. Salas-Zárate María del Pilar, Almela Ángela, Rodríguez-García Miguel Ángel, Valencia-García Rafael. Estudio de las categorías LIWC para el análisis de sentimientos en español. TIMM 2014, pp. 33-36.

6.3. Capítulos en libro

1. Salas-Zárate María del Pilar, Paredes-Valverde Mario Andrés, Rodríguez-García Miguel Ángel, Valencia-García Rafael, Alor-Hernández Giner. Sentiment Analysis based on Psychological and Linguistic Features for Spanish language. Current Trends on Knowledge-Based Systems, vol. 120. ISBN: 978-3-319-51904-3

Capítulo 7. Resumen extendido en inglés / Extended summary in English

7.1. Introduction

The dramatic spread of the Internet in society has substantially changed the forms of communication, entertainment, knowledge acquisition and consumption. There is a constant increase in the number of people who consider the Internet as a medium for answering their queries (García-Crespo, Colomo-Palacios, Gómez-Berbís, & Ruiz-Mezcua, 2010) in addition to using it as a powerful means of communication. Indeed, on the one hand, the reviews expressed in forums, blogs and social networks are having greater importance to make a decision to buy a product, hire a service, and vote for a political party, among others. On the other hand, for providers, this information is also important to get some feedback about their clients' expectations and needs, clients' feelings about their products or services and then to improve them.

Linguistics is one of the areas that has focused on the study of opinion or rather subjective language. Subjective language is defined as the language used to express personal states in the context of a text or conversation (J. Wiebe et al., 2004)(Martínez Cámara, 2016). On the other hand, the sentiment analysis or opinion mining is a type of subjectivity analysis, which aims at identifying opinions, emotions and evaluations expressed in natural language. The main goal is to predict the sentiment orientation (i.e. positive, negative or neutral) of a review by analysing sentiment or opinion words and expressions in sentences and documents. Therefore, the automatic classification of opinions requires the collaboration of two branches such as linguistics and the natural language processing, which allows to understand human language, classify opinions and summarize the sentiment expressed about a product, service or any other aspect.

Figurative language such as irony, sarcasm, and satire play a very important role in sentiment analysis systems. The double meaning expressed in an opinion can

reverse the polarity of this. Although figurative language has been extensively studied by several areas such as linguistics, only few studies have focused on automatic detection.

On the other hand, it is important to mention that few works for sentiment analysis, and figurative language have been focused on the Spanish language, perhaps due to the lack of linguistic resources in this language. However, the study of Spanish is of the most important since it is one of the most used languages on the internet.

The stated reasons in the above paragraphs have been the main motivation to carry out the research described in this thesis. This work proposes a method for the detection of psycholinguistic patterns for sentiment analysis and satire in Spanish. This proposal is based on a supervised machine learning approach that classifies texts as positive, negative, neutral, high positive or high negative and as satirical, non-satirical.

7.2. State of the art

In recent years, several proposals have presented studies for sentiment analysis based on two main approaches: Semantic Orientation and Machine Learning. The Semantic Orientation (SO) approach uses lexical resources such as lexicons, which have been automatically or semi-automatically generated. These lexicons have usually extended the WordNet lexical database. Some works based on this approach used lexicons such as SentiWordNet (Montejo-Ráez et al., 2014) (Peñalver-Martinez et al., 2014) (Singh et al., 2013), iSOL, eSOL (Molina-González et al., 2013), lexicon in Arabic (Al-Ayyoub et al., 2015), for mention a few.

With regard to sentiment analysis methods based on Machine Learning, commonly they rely on supervised classification. These approaches use a collection of data to train the classification algorithms. Among the machine learning techniques commonly used in the sentiment polarity classification are Support Vector Machine (SVM) (Deng, Luo, & Yu, 2014) (Ghiassi et al., 2013), Naive Bayes (NB) (Moraes et al., 2013), Maximum Entropy (MaxEnt) (Habernal et al., 2014), among others. On the other hand, machine learning methods require good, representative features for

delivering good performance. Therefore, their success relies on the effectiveness of the feature extraction process. Most of the existing research focuses on features extracted through methods such as term frequency (Deng, Luo, & Yu, 2014) (Ghiassi et al., 2013) (Moraes et al., 2013), character n-grams, and word n-grams (like bigrams and trigrams) (Habernal et al., 2014) (Peñalver-Martinez et al., 2014), and features based on POS tagging (Habernal et al., 2014).

All these proposals have tried to introduce sentiment classification problem in different levels i.e. document-level (Deng, Luo, & Yu, 2014) (Moraes et al., 2013) (Molina-González et al., 2013), sentence-level (Peñalver-Martinez et al., 2014), and aspect-level (Singh et al., 2013) (Peñalver-Martinez et al., 2014).

In literature, the automatic detection of figurative language is considered a difficult problem and has been addressed in only a few studies. Approaches with which to identify sarcasm and irony have appeared in recent years (Poria, Cambria, Hazarika, & Vij, 2016) (Davidov, Tsur, & Rappoport, 2010) (Sulis et al., 2016). However, satire has rarely been studied in literature. The proposals are based mainly in a supervised machine learning algorithm, where the classification algorithm more used is SVM (Rubin et al., 2016) (Ahmad et al., 2014) (Barbieri et al., 2015b) (Owais et al., 2015) (Barbieri et al., 2015a). The aforementioned proposals are based on different feature extraction approaches. For example, some works are based on TF-IDF (Ahmad et al., 2014), Binary feature weights (Burfoot & Baldwin, 2009) (Owais et al., 2015), while others explore more sophisticated features such as linguistic, affective and punctuation marks (Barbieri et al., 2015a) (Barbieri et al., 2015b) (Rubin et al., 2016). However, these features have been selected on the basis of other studies, and on the assumption that the features have a great influence on sarcasm or satire detection.

7.3. Method

The proposed method is based on a machine learning approach, which requires as input a labeled corpus. The method is divided in three main steps: 1) text pre-processing; 2) psycholinguistic feature extraction; and 3) machine learning algorithm.

Capítulo 7. Extend summary in English

This method can be adapted to several classification problems and languages. In this doctoral thesis, has been used for sentiment classification and satire detection. Next, the Figure 7.1 presents the steps of proposed method and subsequently each step is described.

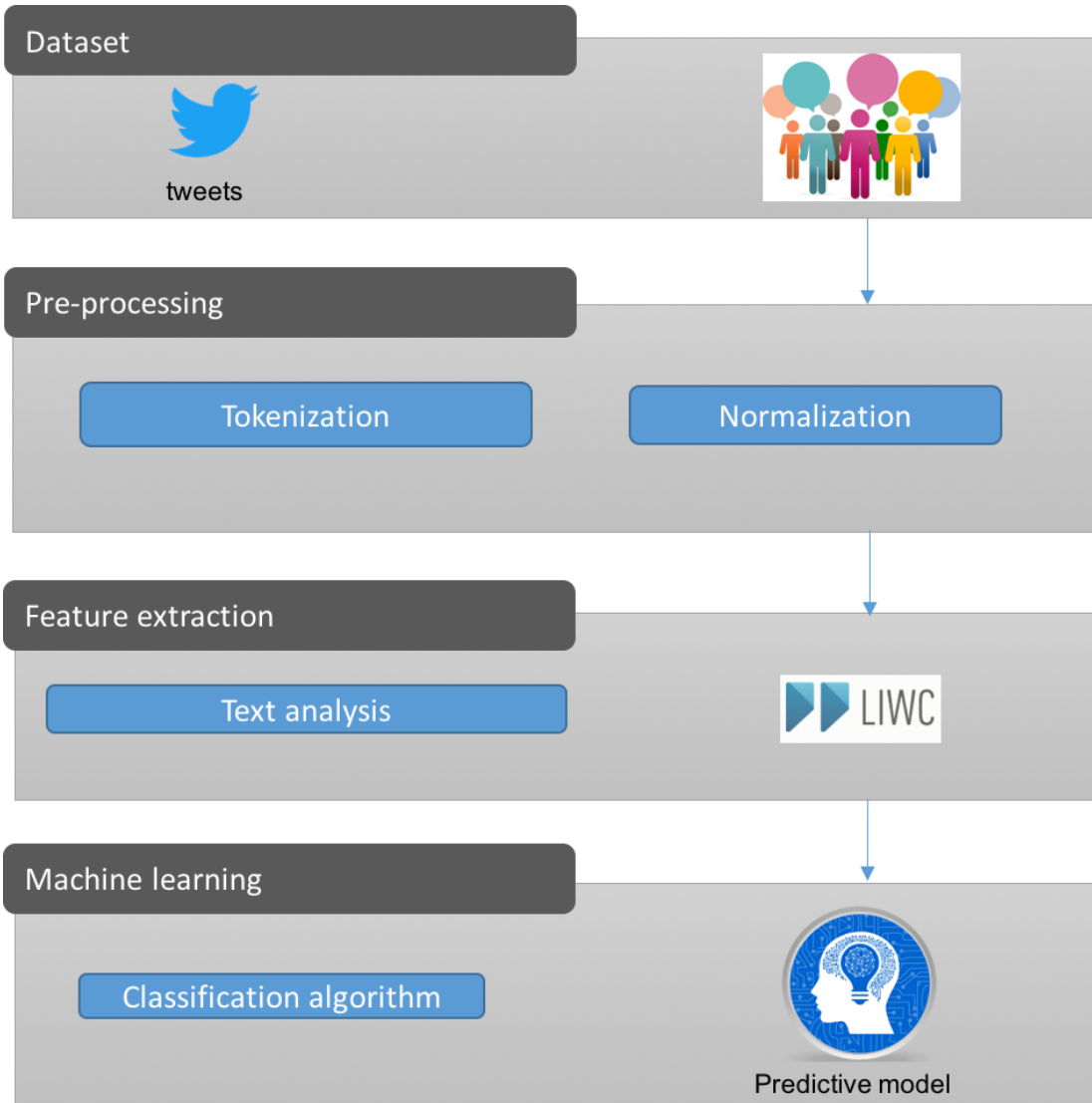


Figure 7.1. Proposed method.

7.3.1. Pre-processing

The first step of the method consists in the pre-processing of each text. The main goal of this step is the normalization the text by removing those elements that are considered superfluous. The two main phases are: 1) Tokenization and 2) Normalization of the text. The techniques used for the pre-processing of each text

depend on the information sources from they were obtained. For example, the text obtained from blogs or forums are considered normal text, meanwhile, tweets are considered special text because the language used in twitter is mainly characterized by an informal and free writing style.

With regards to normal text two steps are carried out, 1) Tokenization of the text with Stanford NLP, 2) Normalization of the text by replacing the abbreviations and correcting the spelling errors with Hunspell dictionary. About special text, three steps were performed, 1) Tokenization process of all tweets and detection of entities such as mentions and hashtags by using the Twokenize tool, 2) Remove mentions, URLs, and the character “#” of the hashtags, 3) Replace elements with their expansions and correct spelling errors with Hunspell dictionary.

7.3.2. Feature extraction

The second step refers to the psycholinguistic feature extraction. LIWC was used aiming to obtain the psycholinguistic features of the texts and explore which LIWC categories can provide patterns for sentiment analysis and satire detection.

LIWC contains dictionaries in several languages such as English, Portuguese, Arabic, Spanish, to mention but a few. These dictionaries have been translated from the English version. In this work, the 2007 Spanish version (Ramírez-Esparza et al., 2007) was used, which was overseen by a native speaker of Mexican Spanish who was closely helped by a Colombian Spanish speaker. The final version involved the collaboration of a native Spanish speaker from Spain. This dictionary is composed of 7515 words and word stems. Each word can be classified into one or more of the 72 categories included by default in LIWC. These categories are additionally classified into the five main sets: 1) linguistic processes; 2) psychological processes; 3) personal concerns; 4) spoken categories; and 5) Punctuation. Table 7.1 present some examples of the LIWC categories.

Capítulo 7. Extend summary in English

Table 7.1. Examples of LIWC categories.

Set	Categories	Examples
Linguistic process (LP)	totpron/ total of pronouns	I, we, you
	artículo/Articles	The
	Negacio/negations	Never
Psychological process (PP)	Emopos	Happy, pretty, good
	Emoneg	Hate, enemy, ugly
	Meccog/cognitive process	know, should
Personal concerns (PC)	Trabajo/work	Employee, boss, career
	Dinero/money	Money, profit
	Relig/Religion	God, heaven, church
Spoken categories (SC)	Asentir/assent	Agree, cool
	Nofluen/nonfluencies	Uh, um
	Relleno/filler	blah
Punctuation marks (PM)	Periods, commas, colons, semicolons, question marks, exclamation marks, dashes, quotation marks, apostrophes, parentheses, other punctuation.	

The linguistic process categories involve grammatical information such as total pronouns, articles, negations, word count, auxiliary verbs, among others. The second set contains the psychological process, which can estimate positive emotions, negative emotions, social processes, cognitive processes, among others. The third set consists of word categories related to personal concerns intrinsic to the human condition. The fourth set contains “spoken categories”, and has been included to accommodate certain dimensions of spoken language. Although LIWC has not been designed for the spoken language, the authors have found it to be useful when analyzing conversations and interviews. Finally, the fifth set consists of twelve punctuation categories (periods, commas, colons, etc.). Regarding parentheses, LIWC counts pairs of parentheses. Moreover, the “Other Punctuation” category contains all those

punctuation marks not included in the other punctuation categories, along with ASCII characters from 33-47, 58-64, 91-96 and 123-126.

Aiming to perform the processing with LIWC, the corpus was converted into plain text files, after which each text was analyzed individually. The processing is carried out in three steps:

1. LIWC compares each word in the text with the dictionary.
2. The dictionary then identifies which words are associated with the LIWC categories.
3. The percentage of total words that match each of the dictionary categories is subsequently calculated.

Therefore, the resulting feature vector consists of the features with their respective percentages, except for the summary variables WC (Total words count), WPS (Words per sentence), Sixltr (words longer than 6 letters), and DIC (Dictionary words count).

The feature extraction was obtained from the individual categories and their combinations. Table 7.2 shows all possible combinations obtained with the LIWC categories.

Capítulo 7. Extend summary in English

Table 7.2 All possible combinations of LIWC categories

Combination	Description
LP	Linguistic process
PP	Psychological process
PC	Personal concerns
SC	Spoken categories
PM	Punctuation marks
LP-PP	Combination of Linguistic process and Psychological process
LP-PC	Combination of Linguistic process and Personal concerns
LP-SC	Combination of Linguistic process and Spoken categories
LP-PM	Combination of Linguistic process and Punctuation marks
PP-PC	Combination of Psychological process and Personal concerns
PP-SC	Combination of Psychological process and Spoken categories
PP-PM	Combination of Psychological process and Punctuation marks
PC-SC	Combination of Personal concerns and Spoken categories
PC-PM	Combination of Personal concerns and Punctuation marks
SC-PM	Combination of Spoken categories and Punctuation marks
LP-PP-PC	Combination of Linguistic process, Psychological process and Personal concerns
LP-PP-SC	Combination of Linguistic process, Psychological process and Spoken categories
LP-PP-PM	Combination of Linguistic process, Psychological process and Punctuation marks
LP-PC-SC	Combination of Linguistic process, Personal concerns and Spoken categories
LP-PC-PM	Combination of Linguistic process, Personal concerns and Punctuation marks
LP-SC-PM	Combination of Linguistic process, Spoken categories, and Punctuation marks
PP-PC-SC	Combination of Psychological process, Personal concerns and Spoken categories

Combination	Description
PP-PC-PM	Combination of Psychological process, Personal concerns and Punctuation marks
PP-SC-PM	Combination of Psychological process, Spoken categories and Punctuation marks
PC-SC-PM	Combination of Personal concerns, Spoken categories and Punctuation marks
LP-PP-PC-SC	Combination of Linguistic process, Psychological process, Personal concerns and Spoken categories
LP-PP-PC-PM	Combination of Linguistic process, Psychological process, Personal concerns and Punctuation marks
LP-PP-SC-PM	Combination of Linguistic process, Psychological process, Spoken categories and Punctuation marks
LP-PC-SC-PM	Combination of Linguistic process, Personal concerns, Spoken categories and Punctuation marks
PP-PC-SC-PM	Combination of Psychological process, Personal concerns, Spoken categories and Punctuation marks
LP-PP-PC-SC-PM	Combination of Linguistic process, Psychological process, Personal concerns, Spoken categories and Punctuation marks

7.3.3. Training of machine learning algorithms

The final step of the proposed method consists of the training of the machine learning algorithms. WEKA (Bouckaert et al., 2010), a collection of machine learning algorithms was used in this work. Three algorithms were selected for the present study, namely, SMO, BayesNet and J48. In the supervised machine learning two datasets are required: training data and testing data.

The training phase consist in the built of a model based on a feature vector, in this case obtained from the previous section. The model is represented by means of classification rules, decision trees or mathematical formulae. The obtained model is evaluated with the testing data set, and subsequently used to classify unknown data. Therefore, the model can classify text as a positive, negative, neutral, high positive and high negative; and satirical, no satirical.

7.4. Evaluation and Results

To measure of the performance of the method for sentiment analysis and satire detection, the precision, recall and F-measure metrics were used. With regards to sentiment analysis, the validation was addressed in two domains: movies and tourism. For the first domain, a benchmark corpus was used, meanwhile, in the second domain a corpus was collected from TripAdvisor website. About satire detection, the validation was carried out in the news domain. Therefore, a corpus was generated with news satirical from Spain and Mexican.

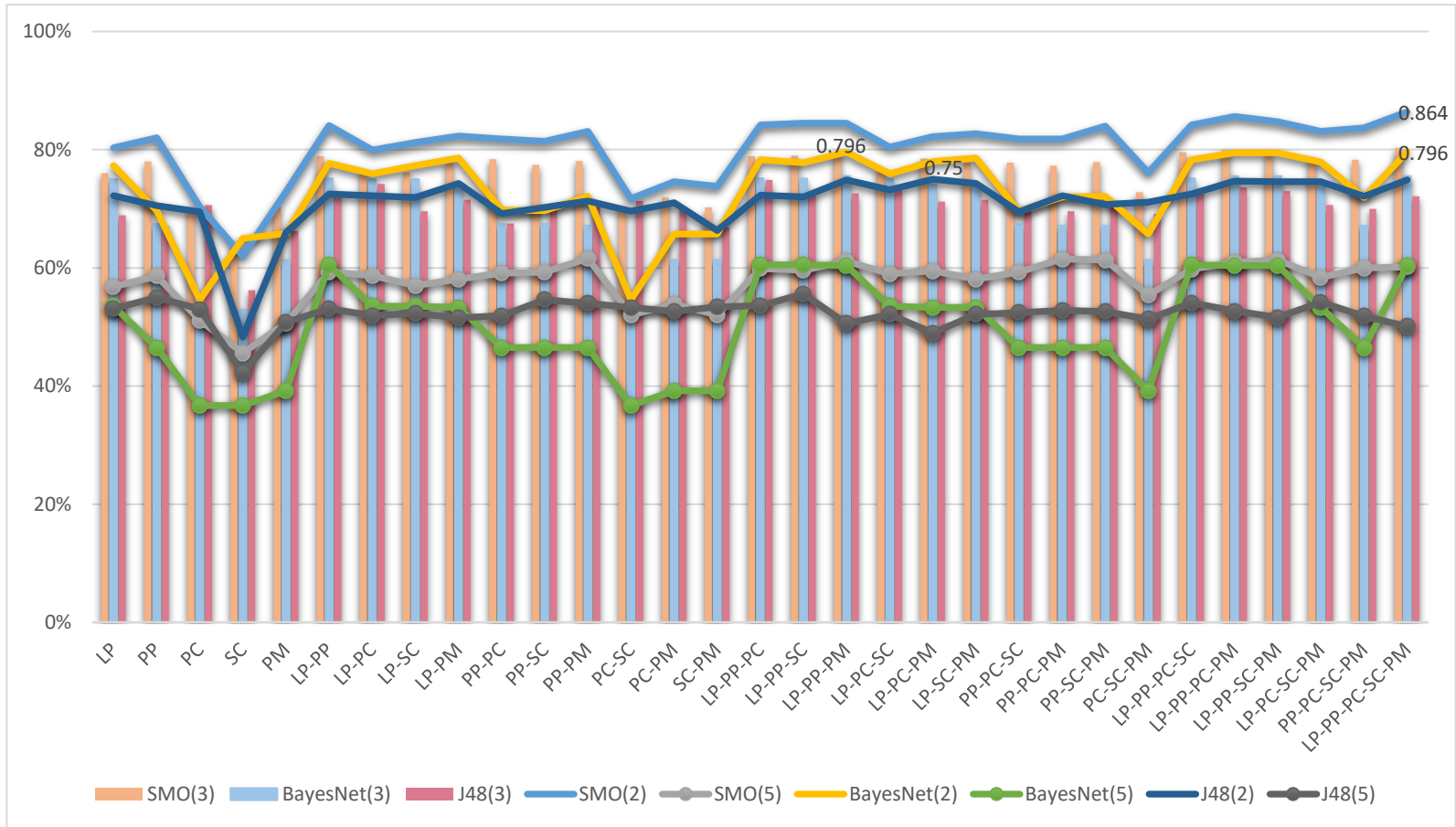
The experiments for the sentiment analysis involve the use the two classes (positive, negative), three classes (positive, negative, and neutral) and five classes (positive, negative, neutral, high positive, and high negative).

The results obtained for the classification of movies reviews are presented in Figure 7.2. The results show that the SMO algorithm provides better results than J48 and BayesNet algorithms. When the evaluation involved only one category, the best results are obtained with the PP “Psychological process” and LP “Linguistic process”. Conversely, the SC “Spoken categories”, provides the worst results. Furthermore, the combination of LIWC dimensions provides better results than individually.

On the other hand, the Figure 7.3 presents the obtained results for the classification of reviews about tourism. As it can be seen, the best results are also obtained for the SMO algorithm, just like in the movies domain. Individually, the PP “Physiological process” category obtained the best results, meanwhile, the SC “Spoken categories” and PM “Punctuation marks” obtained the worst results. Furthermore, the combination of categories provides better results that individually.

For both domains, the best classification results are obtained considering two classes (positive and negative). On the contrary, the results obtained with three classes (positive, negative, and neutral) and five classes (positive, negative, neutral, high positive, and high negative) are poor.

Figure 7.2. Results of the experiments



Capítulo 7. Extend summary in English

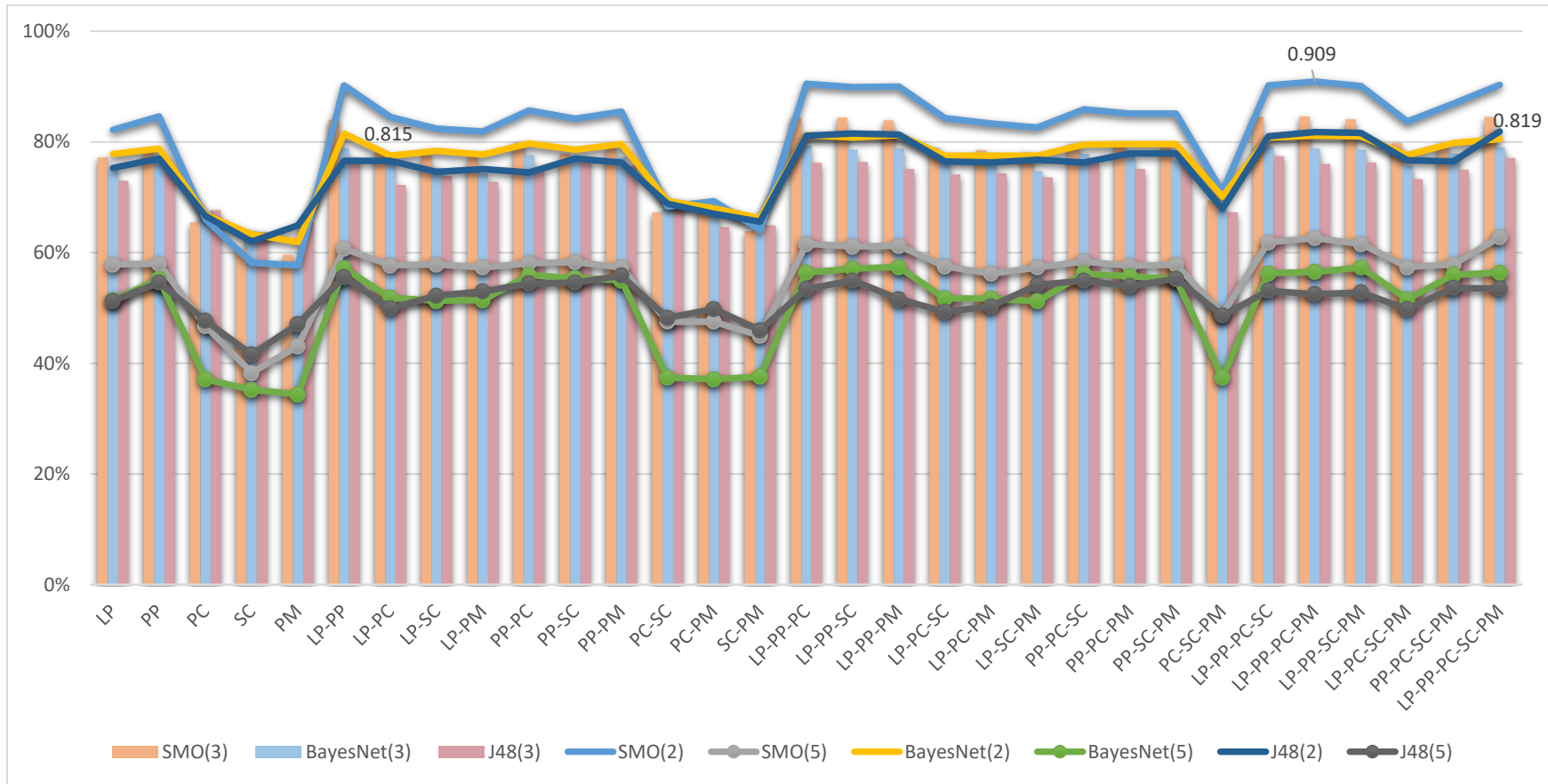


Figure 7.3. Results of the tourism domain.

The experimental results obtained for satire detection allowed us to state which group of features is most interesting and which classifier performs best for this specific problem. As it can be seen in Figure 7.4, the general results show that the most relevant sets for both Spain and Mexico are punctuation marks (PM), linguistic process (LP), and psychological process (PP), while the worst are personal concerns (PC), and spoken categories (SC). About classifiers, SMO obtained better results than J48 and BayesNet. These results clearly confirm the usefulness of adopting the linguistic, psychological and punctuation mark categories in the distinction of satirical and non-satirical texts. Also, the results show that it is not sufficient to use only one category of features. As can be observed, the results are increased by simply adding the features from the linguistic and emotional categories.

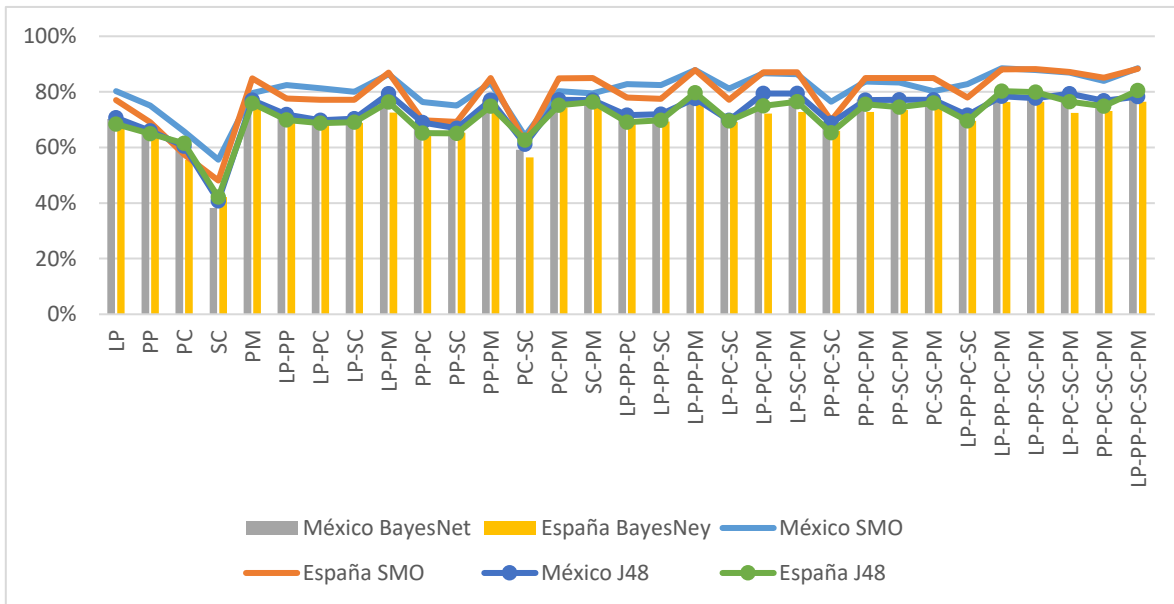


Figure 7.4. Results for satire detection.

7.5. Discussion

The results obtained for sentiment classification show that reducing the number of classes the algorithm precision increases i.e. the classification with two categories (positive-negative) provided better results than the classification with three (positive-neutral-negative) and five (highly positive-positive-neutral-negative-highly negative).

Capítulo 7. Extend summary in English

Regarding to satire, our approach obtained encouraging results for the detection of satirical and non-satirical tweets. They also confirm the importance of linguistic features, emotional features and punctuation marks for satire detection. Regarding the linguistic features, the results demonstrate that satirical tweets contain more adverbs, quantifiers and present tenses than non-satirical tweets. In the case of the physiological process, five features were significant predictors: social process, affective process, positive emotions, cognitive process and certainty. Finally, about punctuation marks, colons, exclamation marks, apostrophes and quotation marks were significant predictors.

With regard to the classification algorithms, for both sentiment analysis and satire detection, SMO Algorithm for SVM classifier obtained the best results. The results can be justified by the analysis presented in (Bhavsar & Ganatra, 2012), where it is clearly shown how SVM models are more accurate in comparison to other classification algorithms such as: decision trees, neural network, Bayesian network, nearest neighbor.

7.6. Conclusions and future work

Sentiment analysis is a problem that requires extensive research. There are problems that still need to be solved, for example figurative language is an important aspect to be considered in sentiment analysis systems. However, this type of language is very difficult to detect, and even more the variations that exist as irony, sarcasm, satire, among others.

The work in this thesis described a method for the detection of psycholinguistic patterns for sentiment analysis and the automatic detection of satire. The psycholinguistic features demonstrated to be effective along with natural language processing techniques and data mining for the detection of sentiments and satire. In addition, the validation of the method in different domains verified its effectiveness for the classification of opinions and tweets.

The main contributions of this thesis are the following: 1) Development of a method for the classification of sentiments and detection of satire, 2) A process for the

pre-processing of tweets in Spanish, 3) Development of a corpus in the tourism domain, 4) Development of a corpus of satirical tweets, and 5) Detection of psycholinguistic features for the sentiment analysis and satire.

As regards to feature work, several aspects that have not been considered as part of this thesis are provided.

- Integrate techniques NLP that allow improve the support for the normalization process in cases of informal language such, Twitter posts.
- Apply the developed method in different domains. This requires the develop of new corpora in several domains. The medical domain is an area of special interest for sentiment analysis which has been little explored. However, opinions can be of great interest to patients especially when they suffer from diseases that require self-management such as diabetes.
- Apply the develop method in other languages, as well as in the varieties of the same language.
- Detect psycholinguistic patterns for sarcasm and irony aiming to determine the similarities and differences among them.
- Integrate the figurative language detection system in the sentiment analysis.
- Contribute to the enrichment of the LIWC dictionary in Spanish with new vocabulary variations such as Venezuela, Ecuador, etc.

Referencias

- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30–38). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=2021109.2021114>
- Ahmad, T., Akhtar, H., Chopra, A., & Akhtar, M. W. (2014). Satire Detection from Web Documents Using Machine Learning Methods. In *Proceedings of the 2014 International Conference on Soft Computing and Machine Intelligence* (pp. 102–105). Washington, DC, USA: IEEE Computer Society. <https://doi.org/10.1109/ISCMI.2014.34>
- Aitkenhead, M. J. (2008). A co-evolving decision tree classification method. *Expert Systems with Applications*, 34(1), 18–25. <https://doi.org/10.1016/j.eswa.2006.08.008>
- Al-Ayyoub, M., Essa, S. B., & Alsmadi, I. (2015). Lexicon-based sentiment analysis of Arabic tweets. *International Journal of Social Network Mining*, 2(2), 101–114. <https://doi.org/10.1504/IJSNM.2015.072280>
- Anta, A. F., Chiroque, L. N., Morere, P., & Santos, A. (2013). Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of of NLP Techniques. *Procesamiento del Lenguaje Natural*, 50(0), 45–52.
- Arafat, H., Elawady, R. M., Barakat, S., & Elrashidy, N. M. (2014). Different feature selection for sentiment classification, 3(1), 137–150.
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Presented at the Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf
- Barbieri, F., Ronzano, F., & Saggion, H. (2015a). Do We Criticise (and Laugh) in the

Referencias

- Same Way? Automatic Detection of Multi-lingual Satirical News in Twitter. In *Proceedings of the 24th International Conference on Artificial Intelligence* (pp. 1215–1221). Buenos Aires, Argentina: AAAI Press. Retrieved from <http://dl.acm.org/citation.cfm?id=2832415.2832418>
- Barbieri, F., Ronzano, F., & Saggion, H. (2015b). Is this Tweet Satirical? A Computational Approach for Satire Detection in Spanish. *Procesamiento del Lenguaje Natural*, 55(0), 135–142.
- Bejcek, E., Stranák, P., & Pecina, P. (2013). Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proc. of the 9th Workshop on Multiword Expressions* (pp. 106–115).
- Bhavsar, H., & Ganatra, A. (2012). A Comparative Study of Training Algorithms for Supervised Machine Learning, 2(4), 74–81.
- Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2010). WEKA—Experiences with a Java Open-Source Project. *Journal of Machine Learning Research*, 11, 2533–2541.
- Burfoot, C., & Baldwin, T. (2009). Automatic Satire Detection: Are You Having a Laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 161–164). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=1667583.1667633>
- Cambria, E., & Hussain, A. (2012). *Sentic Computing: Techniques, Tools, and Applications*. Springer Science & Business Media.
- Carbonell, J. G. (1979). *Subjective Understanding: Computer Models of Belief Systems*.
- Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), 51–89. <https://doi.org/10.1002/aris.1440370103>
- Colston, H. L. (2000). On necessary conditions for verbal irony comprehension. *Pragmatics & Cognition*, 8(2), 277–324. <https://doi.org/10.1075/pc.8.2.02col>
- Cordobés, H., Anta, A. F., Chiroque, L. F., García, F. P., Redondo, T., & Santos, A. (2014). Graph-based Techniques for Topic Classification of Tweets in Spanish. *IJIMAI*, 2(5), 32–38.

- Cotelo, J. M., Cruz, F. L., Troyano, J. A., & Ortega, F. J. (2015). A modular approach for lexical normalization applied to Spanish tweets. *Expert Systems with Applications*, 42(10), 4743–4754. <https://doi.org/10.1016/j.eswa.2015.02.003>
- Cruz, F. L., Troyano, J. A., Pontes, B., & Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13), 5984–5994. <https://doi.org/10.1016/j.eswa.2014.04.005>
- Cruz Mata, F., Troyano Jiménez, J. A., Enríquez de Salamanca Ros, F., & Ortega Rodríguez, F. J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. Retrieved from <http://rua.ua.es/dspace/handle/10045/8067>
- Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of the 12th International Conference on World Wide Web* (pp. 519–528). New York, NY, USA: ACM. <https://doi.org/10.1145/775152.775226>
- Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning* (pp. 107–116). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=1870568.1870582>
- Deng, L., & Wiebe, J. (2015). Mpqa 3.0: An entity/event-level sentiment corpus. In *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1323–1328).
- Deng, S., Sinha, A. P., & Zhao, H. (2016). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*. <https://doi.org/10.1016/j.dss.2016.11.001>
- Deng, Z.-H., Luo, K.-H., & Yu, H.-L. (2014). A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*, 41(7), 3506–3513. <https://doi.org/10.1016/j.eswa.2013.10.056>
- Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th Conference on*

Referencias

- Language Resources and Evaluation (LREC'06* (pp. 417–422).
- Feldman, S. (1999). NLP meets the jabberwocky: Natural language processing in information retrieval : Search Engine Section. *Online*, 23(3), 62–72.
- Filho, P. P. B., Pardo, T. A., & Alusio, S. M. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. Presented at the In 9th Brazilian Symposium in Information and Human Language Technology, Fortaleza, Ceara.
- Francis, M. E., & Pennebaker, J. W. (1993). LIWC: Linguistic Inquiry and Word Count. Dallas, TX: Southern Methodist University.
- Fu, G., & Wang, X. (2010). Chinese Sentence-level Sentiment Classification Based on Fuzzy Sets. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 312–319). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=1944566.1944602>
- Gao, S., Lee, C.-H., & Lim, J. H. (2006). An ensemble classifier learning approach to ROC optimization. In *18th International Conference on Pattern Recognition (ICPR'06)* (Vol. 2, pp. 679–682). <https://doi.org/10.1109/ICPR.2006.246>
- García-Crespo, Á., Colomo-Palacios, R., Gómez-Berbís, J. M., & Ruiz-Mezcua, B. (2010). SEMO: a framework for customer social networks analysis based on semantics. *Journal of Information Technology*, 25(2), 178–188. <https://doi.org/10.1057/jit.2010.1>
- Gelbukh, A. (2013). *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Karlovasi, Samos, Greece, March 24-30, 2013, Proceedings*. Springer.
- Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16), 6266–6282. <https://doi.org/10.1016/j.eswa.2013.05.057>
- Gibbs, R. W. (2000). Irony in Talk Among Friends. *Metaphor and Symbol*, 15(1–2), 5–27. <https://doi.org/10.1080/10926488.2000.9678862>
- Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and Combining

- Sentiment Analysis Methods. In *Proceedings of the First ACM Conference on Online Social Networks* (pp. 27–38). New York, NY, USA: ACM. <https://doi.org/10.1145/2512938.2512951>
- Gurcan, F., & Birturk, A. A. (2016). A Hybrid Movie Recommender Using Dynamic Fuzzy Clustering. In O. H. Abdelrahman, E. Gelenbe, G. Gorbil, & R. Lent (Eds.), *Information Sciences and Systems 2015* (pp. 159–169). Springer International Publishing. https://doi.org/10.1007/978-3-319-22635-4_14
- Habernal, I., Ptáček, T., & Steinberger, J. (2014). Supervised sentiment analysis in Czech social media. *Information Processing & Management*, 50(5), 693–707. <https://doi.org/10.1016/j.ipm.2014.05.001>
- Hatzivassiloglou, V., & Wiebe, J. M. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1* (pp. 299–305). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/990820.990864>
- Hsu, R., See, B., & Wu, A. (2010). Machine Learning for Sentiment Analysis on the Experience Project.
- Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177). New York, NY, USA: ACM. <https://doi.org/10.1145/1014052.1014073>
- Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Presented at the Eighth International AAAI Conference on Weblogs and Social Media.
- Indurkha, N., & Damerau, F. J. (2010). *Handbook of Natural Language Processing* (2nd ed.). Chapman & Hall/CRC.
- Jansen, E. (2014). *NetLingo: The Largest List of Chat Acronyms and Text Shorthand*. NetLingo Inc.
- Jindal, N., & Liu, B. (2006). Mining Comparative Sentences and Relations. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2* (pp. 1331–1336). Boston, Massachusetts: AAAI Press. Retrieved from

Referencias

- <http://dl.acm.org/citation.cfm?id=1597348.1597400>
- Justo, R., Corcoran, T., Lukin, S. M., Walker, M., & Torres, M. I. (2014). Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69, 124–133. <https://doi.org/10.1016/j.knosys.2014.05.021>
- Kaleel, S. B., & Abhari, A. (2015). Cluster-discovery of Twitter messages for event detection and trending. *Journal of Computational Science*, 6, 47–57. <https://doi.org/10.1016/j.jocs.2014.11.004>
- Kaya, M., Fidan, G., & Toroslu, I. H. (2013). Transfer Learning Using Twitter Data for Improving Sentiment Classification of Turkish Political News. In E. Gelenbe & R. Lent (Eds.), *Information Sciences and Systems 2013* (pp. 139–148). Springer International Publishing. https://doi.org/10.1007/978-3-319-01604-7_14
- Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57, 245–257. <https://doi.org/10.1016/j.dss.2013.09.004>
- Khan, F. H., Qamar, U., & Javed, M. Y. (2014). SentiView: A visual sentiment analysis framework. In *International Conference on Information Society (i-Society 2014)* (pp. 291–296). <https://doi.org/10.1109/i-Society.2014.7009062>
- Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Fifth International AAAI Conference on Weblogs and Social Media*. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2857>
- Kovaz, D., Kreuz, R. J., & Riordan, M. A. (2013). Distinguishing Sarcasm From Literal Language: Evidence From Books and Blogging. *Discourse Processes*, 50(8), 598–615. <https://doi.org/10.1080/0163853X.2013.849525>
- Kreuz, R. J., & Caucci, G. M. (2007). Lexical Influences on the Perception of Sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language* (pp. 1–4). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=1611528.1611529>
- Lewis, D. D. (1992). *Representation and Learning in Information Retrieval*. University

- of Massachusetts, Amherst, MA, USA.
- Liddy, E. (2001). Natural Language Processing. *The School of Information Studies: Faculty Scholarship*. Retrieved from <http://surface.syr.edu/istpub/63>
- Lindey, J. (2016). Literal versus exaggerated always and never, *21*(2), 219–249.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. San Rafael, Calif.: Morgan & Claypool Publishers.
- Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th International Conference on World Wide Web* (pp. 342–351). New York, NY, USA: ACM. <https://doi.org/10.1145/1060745.1060797>
- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (pp. 142–150). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=2002472.2002491>
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit (pp. 55–60). Presented at the Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.
- Martí, M. A. (2015). *Semántica y pragmática*. Universitat de Barcelona.
- Martínez Cámara, E. (2016). Análisis de Opiniones en Español. Retrieved from <http://rua.ua.es/dspace/handle/10045/53569>
- Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., & Ureña-López, L. A. (2013). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, *40*(10), 3934–3942. <https://doi.org/10.1016/j.eswa.2012.12.084>
- McCarthy, M., & Carter, R. (2004). ‘There’s millions of them’: hyperbole in everyday conversation. *Journal of Pragmatics*, *36*(2), 149–184. [https://doi.org/10.1016/S0378-2166\(03\)00116-4](https://doi.org/10.1016/S0378-2166(03)00116-4)
- Meena, A., & Prabhakar, T. V. (2007). Sentence Level Sentiment Analysis in the

Referencias

- Presence of Conjuncts Using Linguistic Analysis. In G. Amati, C. Carpineto, & G. Romano (Eds.), *Advances in Information Retrieval* (pp. 573–580). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-71496-5_53
- Melville, P., Gryc, W., & Lawrence, R. D. (2009). Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1275–1284). New York, NY, USA: ACM. <https://doi.org/10.1145/1557019.1557156>
- Mihalcea, R., & Pulman, S. (2009). Linguistic Ethnography: Identifying Dominant Word Classes in Text. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 594–602). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00382-0_48
- Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11), 39–41. <https://doi.org/10.1145/219717.219748>
- Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M.-T., & Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, 40(18), 7250–7257. <https://doi.org/10.1016/j.eswa.2013.06.076>
- Molina-Gonzalez, M.-D. (2016). *Generación de recursos para Análisis de Opiniones en español*. Jaén: Universidad de Jaén. Retrieved from <http://ruja.ujaen.es/handle/10953/699>
- Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L. A. (2014). Ranked WordNet graph for Sentiment Polarity Classification in Twitter. *Computer Speech & Language*, 28(1), 93–107. <https://doi.org/10.1016/j.csl.2013.04.001>
- Moradi, P., & Rostami, M. (2015). Integration of graph clustering with ant colony optimization for feature selection. *Knowledge-Based Systems*, 84, 144–161. <https://doi.org/10.1016/j.knosys.2015.04.007>
- Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems*

- with Applications*, 40(2), 621–633.
<https://doi.org/10.1016/j.eswa.2012.07.059>
- Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. *Microsoft Research*. Retrieved from <https://www.microsoft.com/en-us/research/publication/sentiment-analysis-using-support-vector-machines-with-diverse-information-sources/>
- Nakov, P., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of SemEval (2016)* (pp. 1–18).
- Nasukawa, T., & Yi, J. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In *Proceedings of the 2Nd International Conference on Knowledge Capture* (pp. 70–77). New York, NY, USA: ACM.
<https://doi.org/10.1145/945645.945658>
- Németh, L. (2005). Hunspell. Retrieved 15 September 2016, from <https://hunspell.github.io/>
- Németh, L., Trón, V., Halácsy, P., Kornai, A., Rung, A., & Szakadát, I. (2004). Leveraging the open source ispell codebase for minority language analysis. In *First Steps in Language Documentation for Minority Languages* (pp. 56–59).
- Owais, S. T., Nafis, T., & Khanna, S. (2015). An Improved Method for Detection of Satire from User-Generated Content, 6(3), 2084–2088.
- Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 380–391.
- Padró, L. (2012). Analizadores Multilingües en FreeLing. *Linguamática*, 3(2), 13–20.
- Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1–135.
<https://doi.org/10.1561/1500000011>
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10* (pp. 79–86).

Referencias

- Stroudsburg, PA, USA: Association for Computational Linguistics.
<https://doi.org/10.3115/1118693.1118704>
- Park, S., Lee, W., & Moon, I.-C. (2015). Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recognition Letters*, 56, 38–44.
<https://doi.org/10.1016/j.patrec.2015.01.004>
- Parrott, W. G. (2001). *Emotions in Social Psychology: Essential Readings*. Psychology Press.
- Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M. Á., Moreno, V., Fraga, A., & Sánchez-Cervantes, J. L. (2014). Feature-based opinion mining through ontologies. *Expert Systems with Applications*, 41(13), 5995–6008. <https://doi.org/10.1016/j.eswa.2014.03.022>
- Perez-Rosas, V., Banea, C., & Rada, M. (2012). Learning Sentiment Lexicons in Spanish.
- Philip J., S., Dexter C., D., & Marshall S., S. (1966). *The general inquirer: A computer approach to content analysis*. Oxford, England: M.I.T. Press.
- Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In *Proceedings of COLING 2016* (pp. 1601–1612). Retrieved from <http://arxiv.org/abs/1610.08815>
- Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples, 168–201.
- Quinlan, J. R. (2014). *C4.5: Programs for Machine Learning*. Elsevier.
- Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá Martínez, R. (2007). La psicología del uso de las palabras: un programa de computadora que analiza textos en español. Retrieved from <http://rua.ua.es/dspace/handle/10045/25918>
- Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74, 1–12. <https://doi.org/10.1016/j.datak.2012.02.005>
- Riloff, E., & Wiebe, J. (2003). Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 105–112). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/1119355.1119369>

- Rodriguez, M. Á. G. (2014). Extracción semántica de información basada en evolución de ontologías. *Proyecto de investigación*: Retrieved from <https://digitum.um.es/xmlui/handle/10201/41246>
- Rosenthal, S., Ritter, A., Nakov, P., & Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 73–80).
- Rubin, V. L., Conroy, N. J., Chen, Y., & Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. (pp. 7–17). Presented at the Proceedings of the Workshop on Computational Approaches to Deception Detection at the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-CADD2016), San Diego, California.
- Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, 38(12), 14799–14804. <https://doi.org/10.1016/j.eswa.2011.05.070>
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.
- Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. Retrieved 17 December 2016, from <http://cds.cern.ch/record/100508>
- Shboul, B. A., Al-Ayyoub, M., & Jararweh, Y. (2015). Multi-way sentiment classification of Arabic reviews. In *2015 6th International Conference on Information and Communication Systems (ICICS)* (pp. 206–211). <https://doi.org/10.1109/IACS.2015.7103228>
- Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)* (pp. 712–717). <https://doi.org/10.1109/iMac4s.2013.6526500>
- Skalicky, S., & Crossley, S. (2015). A statistical analysis of satirical Amazon.com product reviews. *The European Journal of Humour Research*, 2(3), 66–85.

Referencias

- <https://doi.org/10.7592/EJHR2014.2.3.skalicki>
- Stiles, W. B. (1992). *Describing Talk: A Taxonomy of Verbal Response Modes* (Har/Dskt edition). Newbury Park: SAGE Publications, Inc.
- Su, J. G., Hopke, P. K., Tian, Y., Baldwin, N., Thurston, S. W., Evans, K., & Rich, D. Q. (2015). Modeling particulate matter concentrations measured through mobile monitoring in a deletion/substitution/addition approach. *Atmospheric Environment*, *122*, 477–483. <https://doi.org/10.1016/j.atmosenv.2015.10.002>
- Sulis, E., Hernández Farías, D. I., Rosso, P., Patti, V., & Ruffo, G. (2016). Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, *108*, 132–143. <https://doi.org/10.1016/j.knosys.2016.05.035>
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. <https://doi.org/10.1177/0261927X09351676>
- Toutanova, K., Manning, C. D., & Ng, A. Y. (2004). Learning random walk models for inducing word dependency distributions. In *Proceedings of the twenty-first international conference on Machine learning* (p. 103). ACM.
- Turney, P. D. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 417–424). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/1073083.1073153>
- Valitutti, R. (2004). WordNet-Affect: an Affective Extension of WordNet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 1083–1086).
- Vilares, J., Alonso, M. A., & Vilares, D. (2013). Prototipado Rápido de un Sistema de Normalización de Tuits: Una Aproximación Léxica (pp. 39–43). Presented at the Tweet-Norm@ SEPLN 2013.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

- Whitelaw, C., Garg, N., & Argamon, S. (2005). Using Appraisal Groups for Sentiment Analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (pp. 625–631). New York, NY, USA: ACM. <https://doi.org/10.1145/1099554.1099714>
- Wiebe, J. M. (1990). Identifying Subjective Characters in Narrative. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2* (pp. 401–406). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/997939.998008>
- Wiebe, J. M. (1994). Tracking Point of View in Narrative. *Comput. Linguist.*, *20*(2), 233–287.
- Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and Use of a Gold-standard Data Set for Subjectivity Classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 246–253). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/1034678.1034721>
- Wiebe, J. M., & Rapaport, W. J. (1988). A Computational Theory of Perspective and Reference in Narrative. In *Proceedings of the 26th Annual Meeting on Association for Computational Linguistics* (pp. 131–138). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/982023.982039>
- Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning Subjective Language. *Comput. Linguist.*, *30*(3), 277–308. <https://doi.org/10.1162/0891201041850885>
- Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, *39*(2–3), 165–210. <https://doi.org/10.1007/s10579-005-7880-9>
- Wiley, M. T., Jin, C., Hristidis, V., & Esterling, K. M. (2014). Pharmaceutical drugs chatter on Online Social Networks. *Journal of Biomedical Informatics*, *49*, 245–254. <https://doi.org/10.1016/j.jbi.2014.03.006>
- Wilks, Y., & Bien, J. (1983). Beliefs, points of view, and multiple environments. *Cognitive Science*, *7*(2), 95–119. <https://doi.org/10.1016/S0364->

Referencias

0213(83)80007-X

- Williams, N., Zander, S., & Armitage, G. (2006). A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. *SIGCOMM Comput. Commun. Rev.*, 36(5), 5–16. <https://doi.org/10.1145/1163593.1163596>
- Wilson, D., & Sperber, D. (2007). *On verbal irony*. Psychology Press.
- Wilson, T. (2008). *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. ProQuest.
- Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347–354). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/1220575.1220619>
- Wilson, T., Wiebe, J., & Hwa, R. (2004). Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In *Proceedings of the 19th National Conference on Artificial Intelligence* (pp. 761–767). San Jose, California: AAAI Press. Retrieved from <http://dl.acm.org/citation.cfm?id=1597148.1597270>
- Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, 94, 88–104. <https://doi.org/10.1016/j.knosys.2015.11.013>
- Zhang, P., & He, Z. (2015). Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification. *Journal of Information Science*, 41(4), 531–549. <https://doi.org/10.1177/0165551515585264>