

# Índice general

<b>I Introducción</b>	1
1.1 Motivación . . . . .	2
1.2 Descripción del problema. . . . .	4
1.3 Objetivos de la tesis. . . . .	7
1.4 Organización de la tesis. . . . .	9
1.5 Conclusiones del capítulo. . . . .	11
<b>II Categorización de documentos</b>	13
2.1 Aprendizaje automático. . . . .	14
2.2 Aprendizaje semi-supervisado. . . . .	17
2.3 Categorización automática de documentos. . . . .	18
2.3.1 Definición del problema. . . . .	18
2.3.2 Representación de los documentos. . . . .	20
2.3.2.1 Pre-procesamiento. . . . .	20
2.3.2.2 Indexado. . . . .	21
2.3.2.3 Reducción de dimensionalidad. . . . .	23
2.3.2.4 Umbral de frecuencia. . . . .	25
2.3.2.5 Ganancia de información. . . . .	27
2.3.3 Algoritmos de categorización. . . . .	28
2.3.3.1 Naïve Bayes. . . . .	29
2.3.3.2 Máquinas de vectores de soporte. . . . .	31
2.4 Evaluación de un sistema de categorización. . . . .	34
2.4.1 Medidas de evaluación. . . . .	34
2.4.2 Estrategias de evaluación. . . . .	37
2.4.2.1 Validación simple. . . . .	38
2.4.2.2 Validación cruzada. . . . .	39
2.5 Conclusiones del capítulo. . . . .	40

<b>III Trabajo relacionado</b>	41
3.1 Introducción. . . . .	42
3.2 Ensamblajes de clasificadores. . . . .	43
3.2.1 Bagging . . . . .	44
3.2.2 Stacking . . . . .	45
3.2.3 Boosting . . . . .	46
3.3 Categorización semi-supervisada . . . . .	49
3.3.1 Self-training . . . . .	51
3.3.2 Co-training. . . . .	52
3.4 Uso de la Web como corpus. . . . .	53
3.5 Categorización de documentos con clases desbalanceadas. . . . .	57
3.5.1 Método de sub-muestreo. . . . .	58
3.5.2 Método de sobre-muestreo. . . . .	59
3.6 Categorización no temática. . . . .	60
3.6.1 Atribución de autoría. . . . .	61
3.7 Conclusiones del capítulo. . . . .	66
<b>IV Método propuesto</b>	67
4.1 Introducción. . . . .	68
4.2 Arquitectura. . . . .	69
4.2.1 Adquisición de corpus. . . . .	72
4.2.2 Aprendizaje semi-supervisado. . . . .	80
4.3 Aportaciones del método. . . . .	84
4.4 Conclusiones del capítulo. . . . .	85
<b>V Resultados experimentales</b>	87
5.1 Configuración general de los experimentos. . . . .	88
5.1.1 Búsqueda en la Web . . . . .	89
5.1.2 Pre-procesamiento de documentos. . . . .	90
5.1.3 Algoritmos de aprendizaje . . . . .	91
5.1.4 Medidas de evaluación. . . . .	91

5.2 Categorización de noticias sobre desastres naturales. . . . .	94
5.2.1 Objetivo del experimento. . . . .	94
5.2.2 Descripción del corpus. . . . .	95
5.2.3 Resultados. . . . .	96
5.2.3.1 Resultados de referencia. . . . .	96
5.2.3.2 Resultados obtenidos al aplicar el método propuesto. . . . .	98
5.3 Categorización de noticias del corpus de Reuters. . . . .	107
5.3.1 Objetivo del experimento. . . . .	107
5.3.2 Descripción del corpus. . . . .	107
5.3.3 Resultados. . . . .	108
5.3.3.1 Resultados de referencia. . . . .	109
5.3.3.2 Resultados obtenidos al aplicar el método propuesto. . . . .	111
5.4 Atribución de autoría de poemas. . . . .	116
5.4.1 Objetivo del experimento. . . . .	116
5.4.2 Descripción del corpus. . . . .	117
5.4.3 Resultados. . . . .	119
5.4.3.1 Resultados de referencia. . . . .	120
5.4.3.2 Resultados al aplicar el método propuesto. . . . .	121
5.5 Conclusiones del capítulo. . . . .	127
<b>VI Desambiguación del sentido de las palabras</b>	129
6.1 Descripción de la tarea. . . . .	130
6.2 Evaluación y métodos utilizados en WSD. . . . .	134
6.2.1 Evaluación. . . . .	134
6.2.2 Métodos basados en conocimiento . . . . .	137
6.2.3 Métodos basados en corpus: supervisados, no supervisados y semi-supervisados. . . . .	141
6.3 Resultados experimentales. . . . .	144
6.3.1 Objetivo del experimento . . . . .	144
6.3.2 Configuración del experimento. . . . .	145
6.3.3 Descripción del corpus. . . . .	145
6.3.4 Resultados de referencia. . . . .	147
6.3.5 Resultados obtenidos al aplicar el método propuesto en WSD. . . . .	148

6.3.6 Discusión de los resultados. ....	150
6.4 Conclusiones del capítulo. ....	157
<b>VII Conclusiones, aportaciones y trabajo futuro</b>	159
7.1 Conclusiones. ....	160
7.2 Aportaciones . ....	163
7.3 Trabajo futuro. ....	164
7.4 Publicaciones. ....	165
<b>Referencias. ....</b>	167

.

## Índice de figuras

Figura 2.1	Ejemplo de la ley de Zipf . . . . .	26
Figura 2.2	Problema de categorización linealmente separable. . . . .	31
Figura 2.3	Un par de hiperplanos y sus márgenes de riesgo de error. . . . .	32
Figura 3.1	Ejemplos de características estilométricas. . . . .	63
Figura 4.1	Método semi-supervisado de categorización basado en la Web. . . . .	70
Figura 5.1	Gráfica de similitud del conjunto de entrenamiento, colección de desastres. . . . .	103
Figura 5.2	Gráfica de similitud del conjunto de entrenamiento enriquecido, colección de desastres. . . . .	104
Figura 5.3	Gráfica de similitud corpus de entrenamiento, Reuters. . . . .	113
Figura 5.4	Gráfica de similitud corpus de entrenamiento enriquecido, Reuters. . . . .	114
Figura 5.5	Gráfica de similitud usando palabras funcionales, poetas. . . . .	123
Figura 5.6	Gráfica de similitud usando trigramas, poetas. . . . .	123
Figura 5.7	Gráfica de similitud usando bigramas, poetas. . . . .	125
Figura 6.1	Ambigüedad de la palabra <i>age</i> . . . . .	132
Figura 6.2	Gráfica de similitud conjunto de entrenamiento, <i>rate</i> , SemEval. . . . .	152
Figura 6.3	Gráfica de similitud conjunto de entrenamiento enriquecido, <i>rate</i> , SemEval. . . . .	153
Figura 6.4	Gráfica de similitud conjunto de entrenamiento, <i>condition</i> , SemEval. . . . .	153
Figura 6.5	Gráfica de similitud conjunto de entrenamiento enriquecido, <i>condition</i> , SemEval. . . . .	154
Figura 6.6	Gráfica de similitud conjunto de entrenamiento, <i>source</i> , SemEval. . . . .	155
Figura 6.7	Gráfica de similitud conjunto de entrenamiento enriquecido, <i>source</i> , SemEval. . . . .	156

## Índice de Tablas

Tabla 2.1	Ejemplo de la ley de Zipf. . . . .	26
Tabla 4.1	Palabras relevantes para la categoría <i>wheat</i> . . . . .	74
Tabla 4.2	Algunas peticiones para la categoría <i>wheat</i> . . . . .	76
Tabla 4.3	Determinación del peso de una petición. . . . .	77
Tabla 4.4	Número de ejemplos no etiquetados descargados por petición. . . . .	78
Tabla 5.1	Resultado de referencia para la colección de desastres. . . . .	97
Tabla 5.2	Exactitud con NB ( $m=1$ y $m= T $ ), colección desastres. . . . .	98
Tabla 5.3	Exactitud con SVM ( $m=1$ y $m= T $ ), colección desastres. . . . .	99
Tabla 5.4	SLMB y vocabulario de la colección de desastres. . . . .	101
Tabla 5.5	Exactitud sin filtro de selección de snippets usando NB. . . . .	105
Tabla 5.6	Exactitud sin filtro de selección de snippets usando SVM. . . . .	105
Tabla 5.7	Distribución ModApte de Reuters. . . . .	108
Tabla 5.8	Palabras relevantes para la categoría <i>wheat</i> . . . . .	109
Tabla 5.9	Exactitud para diferentes conjuntos de entrenamiento, Reuters. . . . .	111
Tabla 5.10	Exactitud usando 10 y 100 instancias de entrenamiento, Reuters. . . . .	112
Tabla 5.11	SLMB y vocabulario para la colección de Reuters. . . . .	115
Tabla 5.12	Estadísticas del corpus de poetas. . . . .	117
Tabla 5.13	Resultados de referencia atribución de autoría de poemas. . . . .	120
Tabla 5.14	Colección de poetas para la atribución de autoría. . . . .	121
Tabla 5.15	Exactitud aplicando el método propuesto a la atribución de autoría . . .	122
Tabla 5.16	SLMB y vocabulario para el corpus de poetas. . . . .	126
Tabla 6.1	Estadísticas del corpus de entrenamiento, SemEval. . . . .	146
Tabla 6.2	Resultados de referencia usando NB y SVM, SemEval. . . . .	147
Tabla 6.3	Ejemplos de peticiones para <i>drug</i> , SemEval. . . . .	148
Tabla 6.4	Resultados conjunto de entrenamiento enriquecido, SemEval. . . . .	149
Tabla 6.5	Corpus original y enriquecido, medidas SLMB y vocabulario. . . . .	151