



**UNIVERSIDAD
DE GRANADA**

MINERÍA DE MEDIOS SOCIALES
MÁSTER CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

MINERÍA DE TEXTO

ANÁLISIS DE SENTIMIENTOS

Autor

Ignacio Vellido Expósito
ignaciove@correo.ugr.es



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE
TELECOMUNICACIÓN

CURSO 2020-2021

1. Clasificación

1.1. Metodología

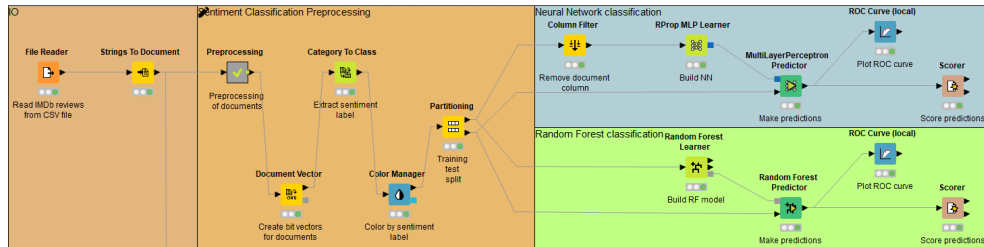


Figura 1: Workflow de la clasificación de sentimientos.

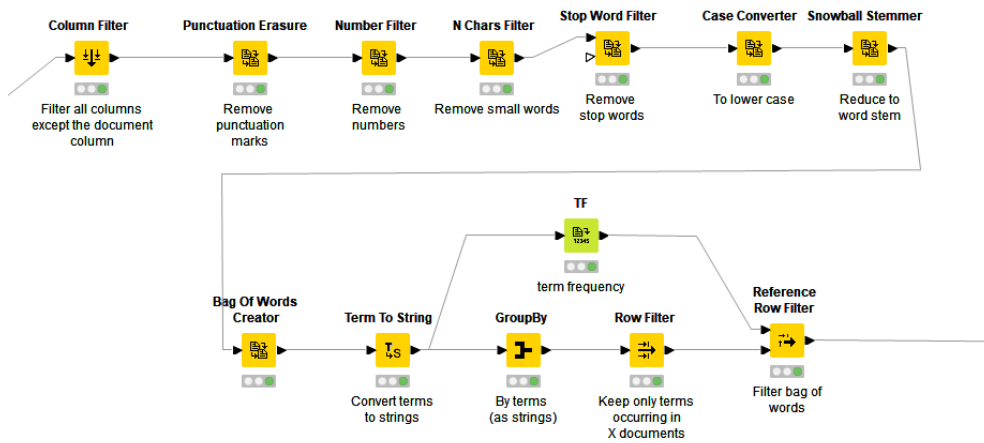


Figura 2: Preprocesamiento de los documentos.

1.2. Resultados

Document ...	NEG	POS
NEG	260	40
POS	65	235

Correct classified: 495

Accuracy: 82.5 %

Cohen's kappa (κ) 0.65

Wrong classified: 105

Error: 17.5 %

Figura 3: Matriz de confusión con Redes Neuronales.

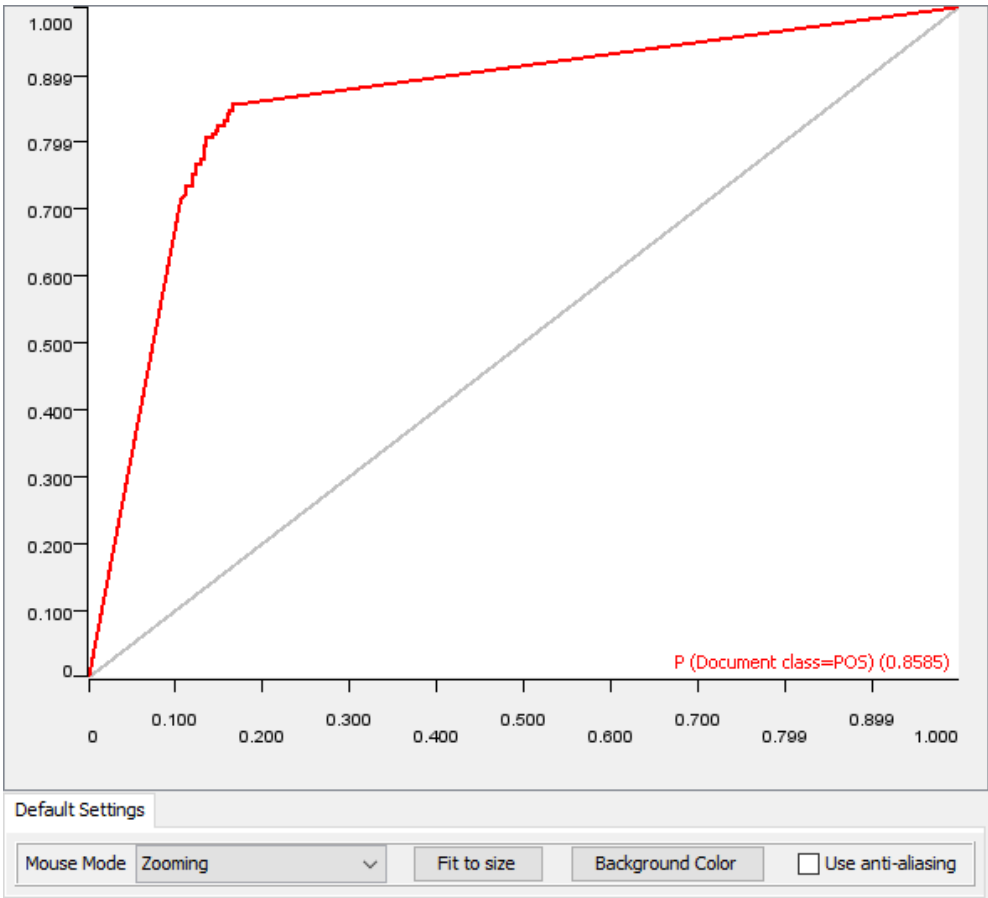


Figura 4: Curva ROC con Redes Neuronales.

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specficity	D F-meas...	D Accuracy	D Cohen'...
NEG	260	65	235	40	0.867	0.8	0.867	0.783	0.832	?	?
POS	235	40	260	65	0.783	0.855	0.783	0.867	0.817	?	?
Overall	?	?	?	?	?	?	?	?	?	0.825	0.65

Figura 5: Medidas estadísticas con Redes Neuronales.

Prediction ...	NEG	POS
NEG	285	39
POS	15	261

Correct classified: 546

Wrong classified: 54

Accuracy: 91 %

Error: 9 %

Cohen's kappa (κ) 0.82

Figura 6: Matriz de confusión con Redes Neuronales.

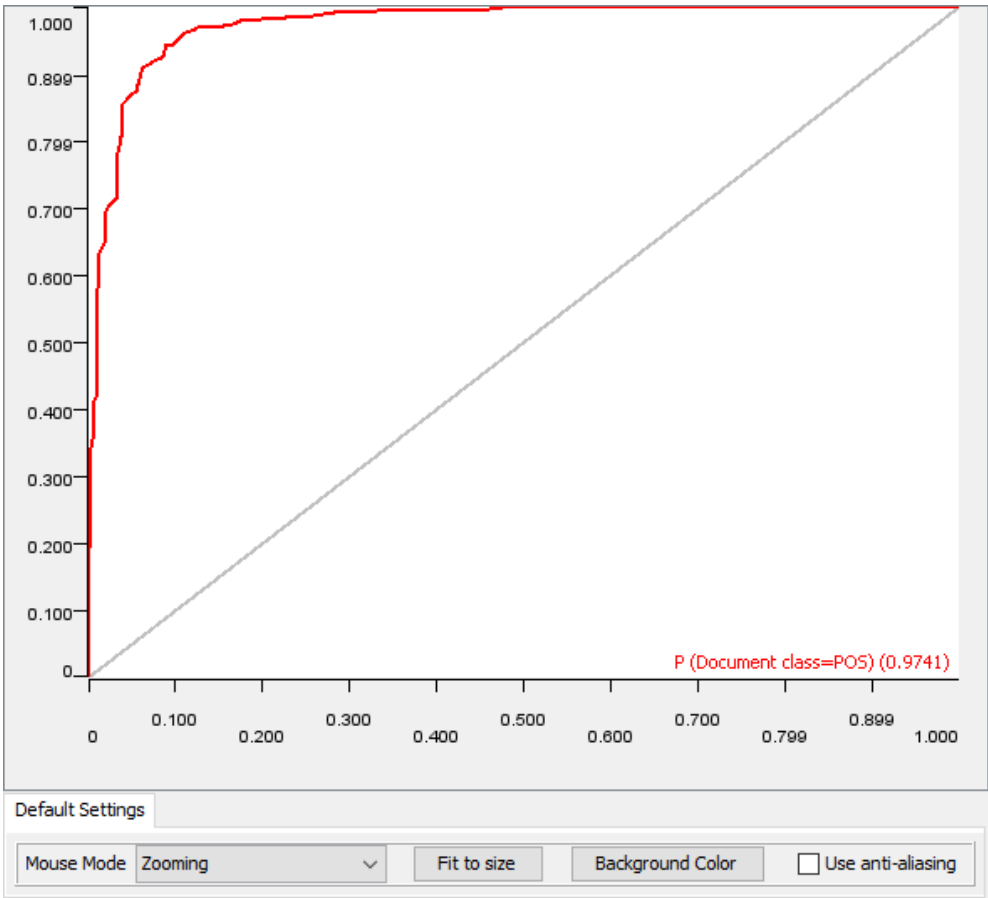


Figura 7: Curva ROC con Random Forest.

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
NEG	285	15	261	39	0.88	0.95	0.88	0.946	0.913	?	?
POS	261	39	285	15	0.946	0.87	0.946	0.88	0.906	?	?
Overall	?	?	?	?	?	?	?	?	?	0.91	0.82

Figura 8: Medidas estadísticas con Random Forest.

2. Análisis de Sentimientos

2.1. Metodología

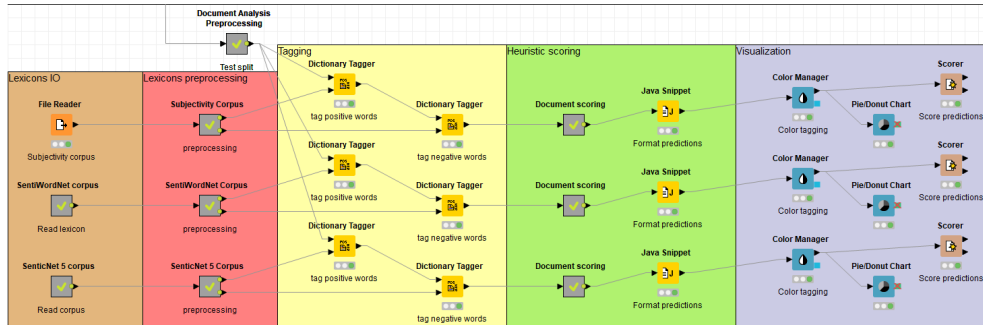


Figura 9: Workflow general del análisis de sentimientos.

Para hacer una comparativa justa de resultados, se aplica el análisis de sentimientos a la misma partición de test con la que se evalúan los métodos de clasificación.

Como preprocesamiento de los documentos aplicamos un Parts Of Speech tagging y Stanford Lemmatizer para quitar las inflexiones de las palabras.

Por otro lado, la lectura de los lexicon varía con cada uno:

- Para el corpus MPQA se usa el mismo procedimiento proporcionado en clase.
- Para SentiWordNet, puesto que cada synterm puede contener más de una palabra, estas se separan en nuevas fila con los mismos valores de sentimiento. Seguidamente

calculamos el valor de objetividad como $1 - (POS + NEG)$ y a cada término le asignamos un valor **neutral** si ambos *PosScore* y *NegScore* son iguales, en otro caso se le asigna la etiqueta con el score más alto. Por último agrupamos las filas dónde

coincidan el synterm, de forma que no se etiqueten algunas palabras con sentimientos diferentes. Para ello hacemos uso del nodo *Group By* y asignamos el sentimiento y el valor de objetividad en base a la moda y a la media respectivamente.

- Para SenticNet modificamos el delimitador de columna por uno nuevo, y le asociamos a cada término el valor de **polarity** indicado.

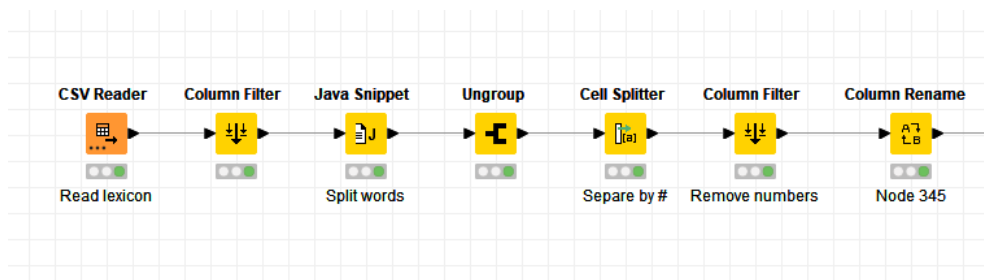


Figura 10: Lectura del lexicon SentiWordNet.

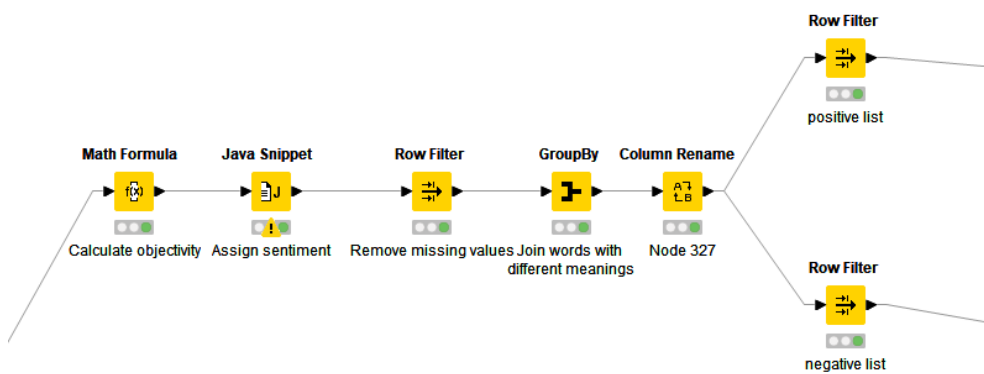


Figura 11: Preprocesamiento del lexicon SentiWordNet.

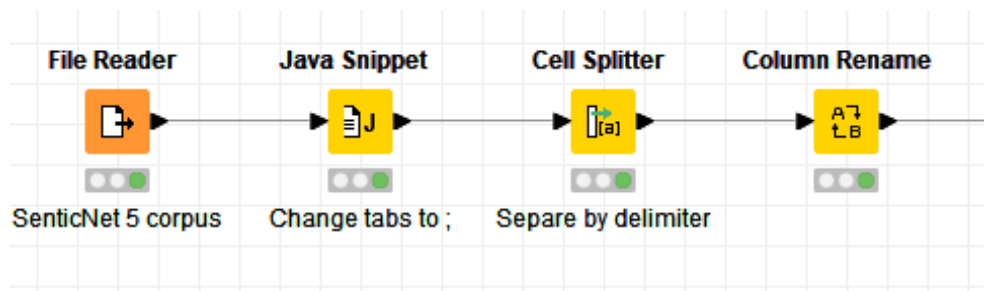


Figura 12: Lectura del lexicon SenticNet.

Finalmente asignamos a cada documento el sentimiento en base al mayor número de palabras que tenga de uno u otro.

2.2. Resultados

Sentiment ...	NEG	POS
NEG	139	171
POS	42	248

Correct classified: 387	Wrong classified: 213
Accuracy: 64.5 %	Error: 35.5 %
Cohen's kappa (κ) 0.299	

Figura 13: Matriz de confusión con MPQA.

Sentiment ...	NEG	POS
NEG	262	38
POS	227	73

Correct classified: 335

Wrong classified: 265

Accuracy: 55.833 %

Error: 44.167 %

Cohen's kappa (κ) 0.117

Figura 14: Matriz de confusión con SentiWordNet.

Sentiment ...	NEG	POS
NEG	116	184
POS	66	234

Correct classified: 350
Accuracy: 58.333 %
Cohen's kappa (κ) 0.167

Wrong classified: 250
Error: 41.667 %

Figura 15: Matriz de confusión con SenticNet.

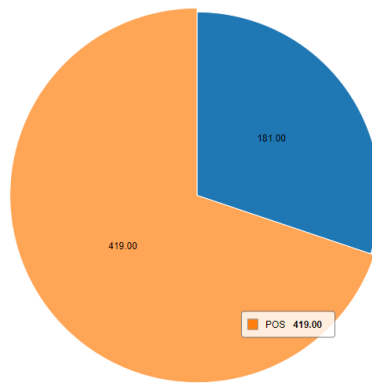


Figura 16: Pie chart con MPQA.

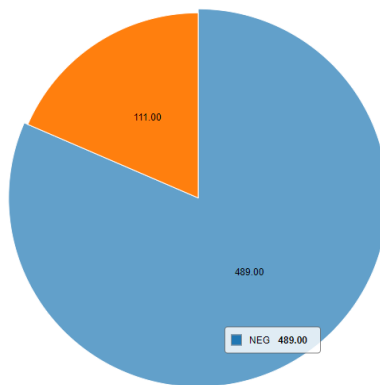


Figura 17: Pie chart con SentiWordNet.

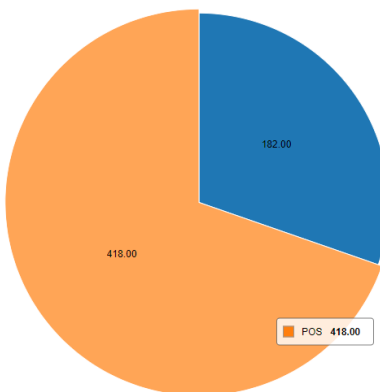


Figura 18: Pie chart con SenticNet.

3. Conclusiones

Método	Precisión	Kappa
MPQA	64.5	0.299
SentiWordNet	55.8	0.117
SenticNet	58.3	0.167
Redes Neuronales	82.5	0.65
Random Forest	91.0	0.82

Cuadro 1: Tabla de resultados generales.