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

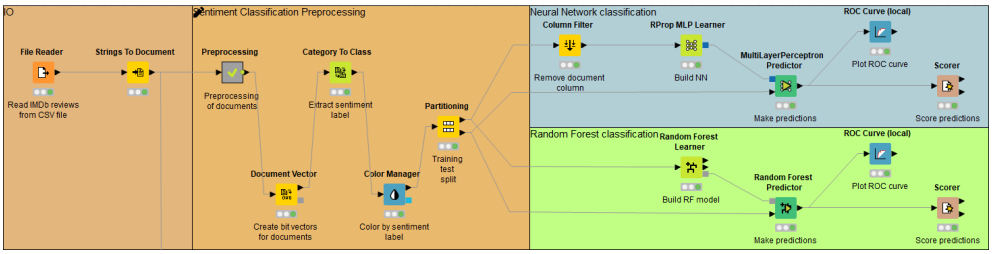

Figura 1: Workflow de la clasificación de sentimientos.

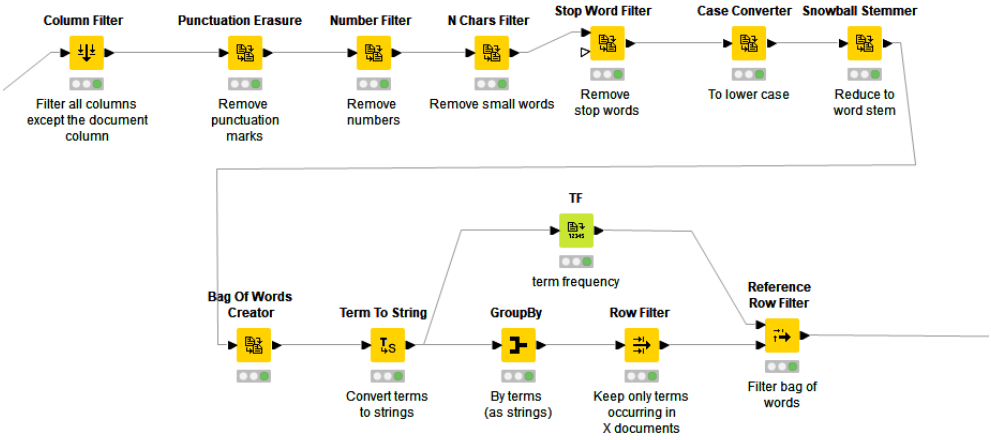

Figura 2: Preprocesamiento de los documentos.

1.2. Resultados

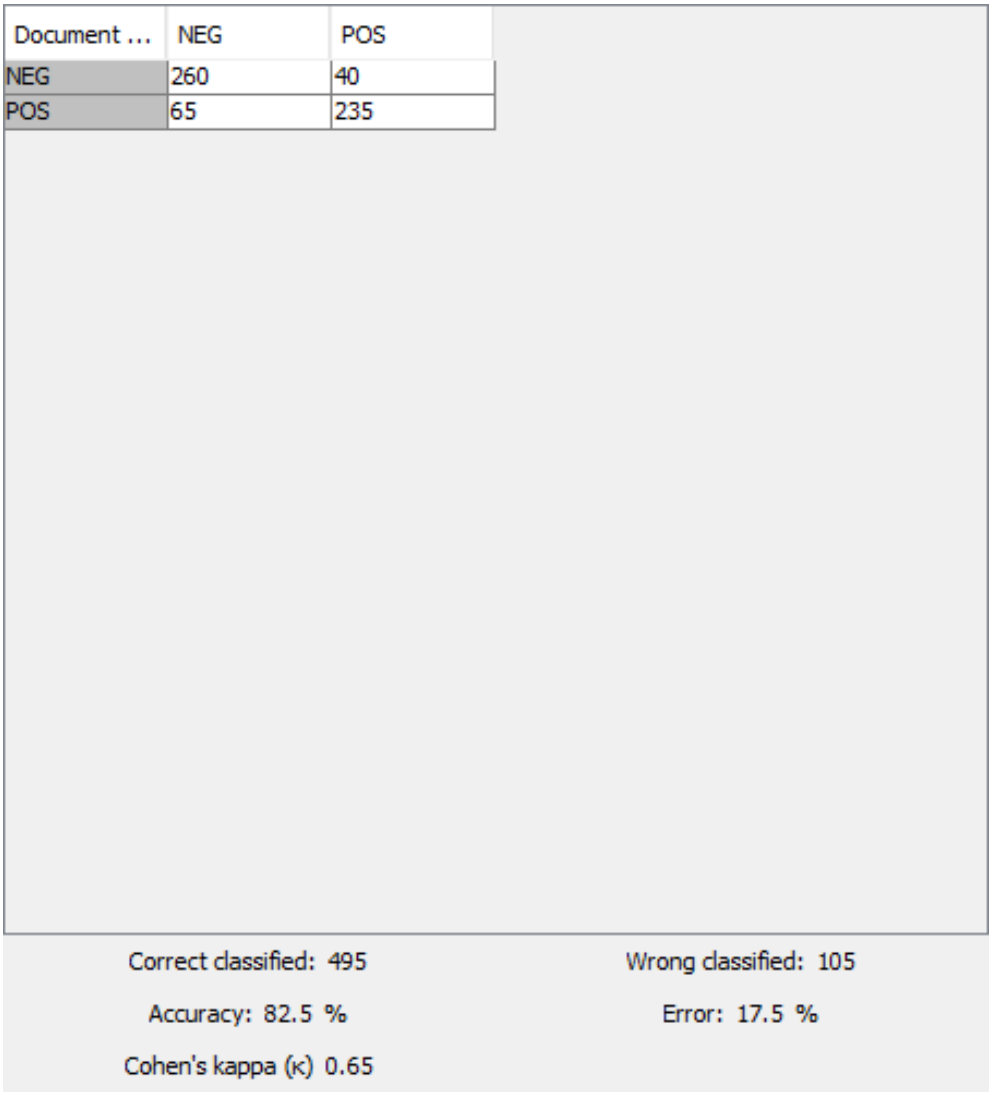

Figura 3: Matriz de confusión con Redes Neuronales.

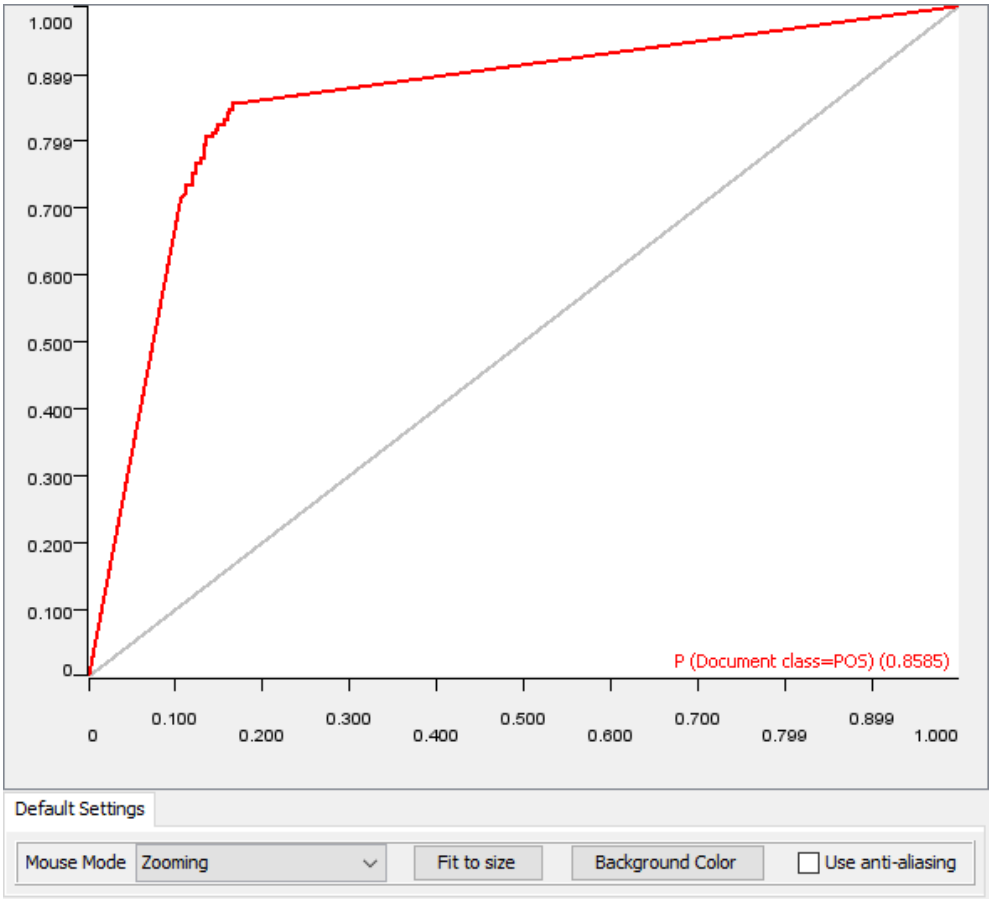

Figura 4: Curva ROC con Redes Neuronales.

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	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.

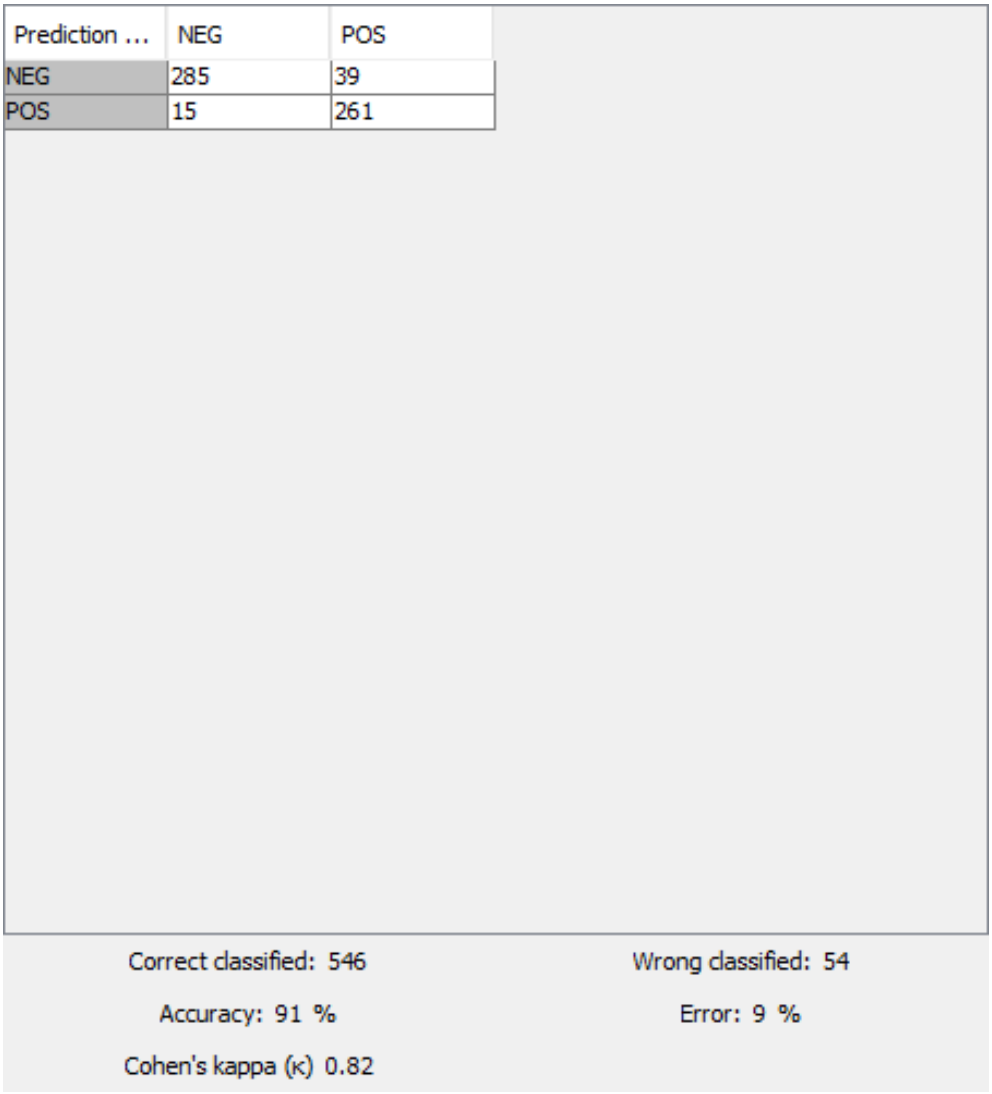

Figura 6: Matriz de confusión con Random Forest.

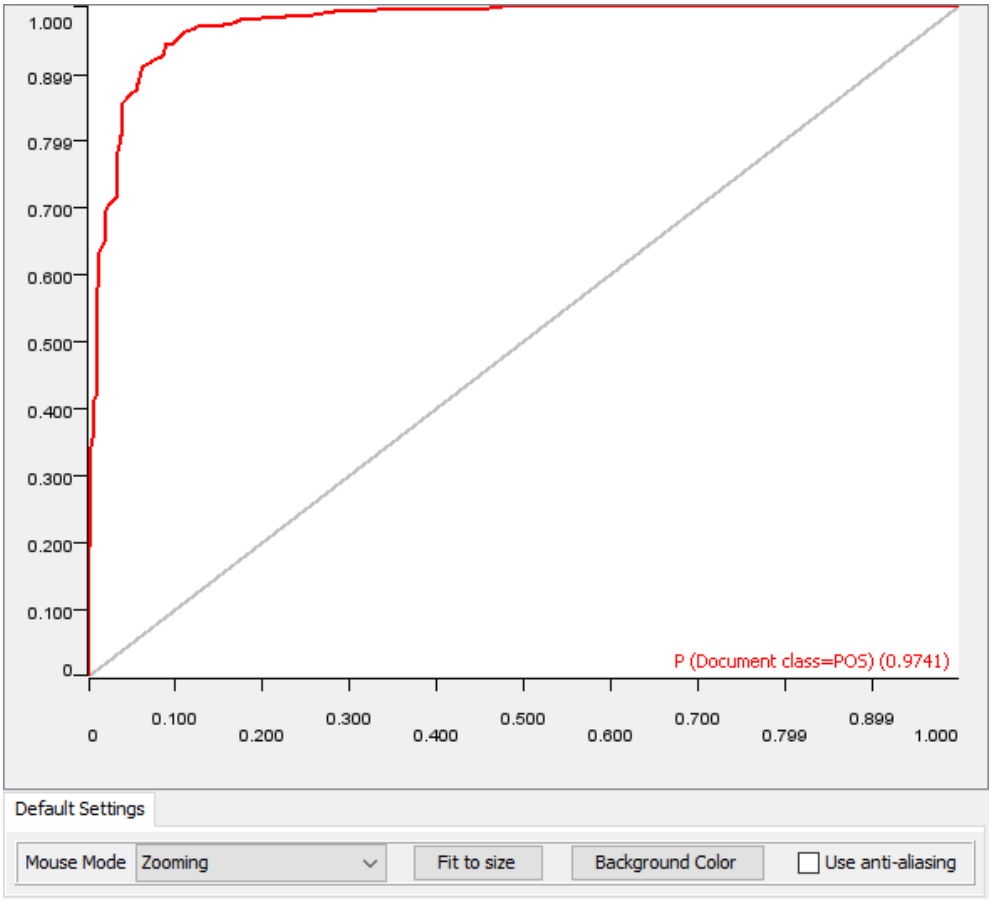

Figura 7: Curva ROC con Random Forest.

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	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

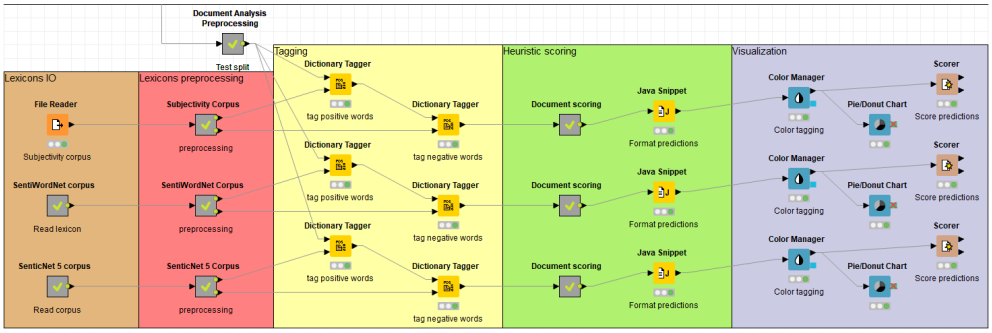

Figura 9: Workflow general del análisis de sentimientos.

Para hacer una comparativa justa de resultados, se realiza el análisis de sentimientos sobre 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 synsetTerm 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 synsetTerm, 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** proporcionado.

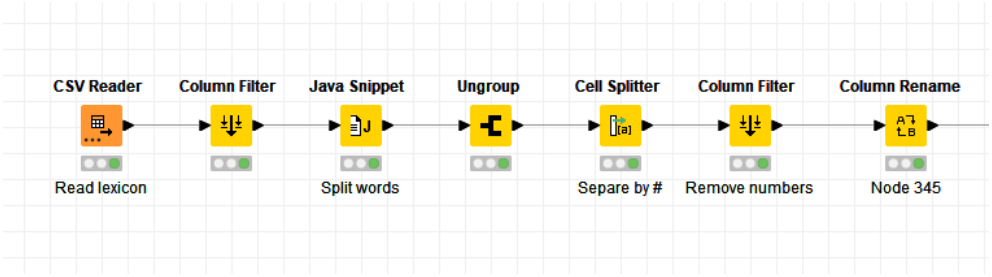

Figura 10: Lectura del lexicon SentiWordNet.

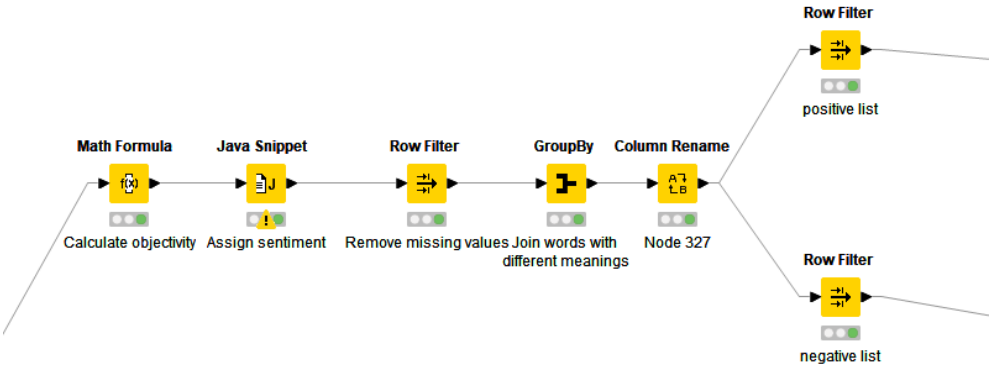

Figura 11: Preprocesamiento del lexicon SentiWordNet.

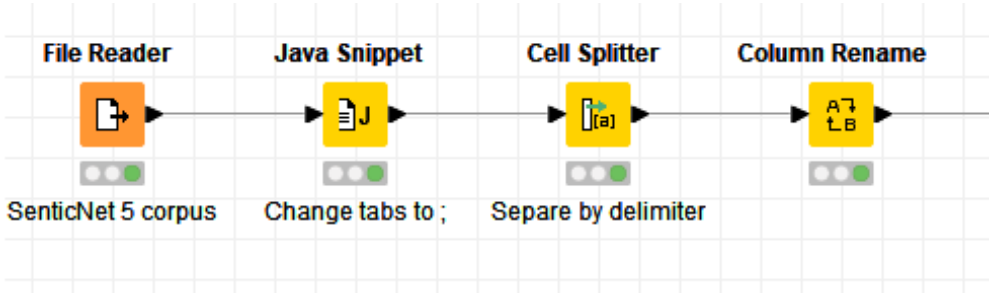

Figura 12: Lectura del lexicon SenticNet.

Finalmente para la asignación de sentimiento a cada documento se han usado dos heurísticas diferentes:

1. Seleccionado en base al mayor número de palabras que tenga de una etiqueta u otra.
2. Ponderización según la función del término en la frase (POS), de la siguiente manera: nombres y adjetivos (tanto en singular como en plural) junto a adverbios reciben el doble de peso; conjunciones comparativas (como la cláusula *but*) reciben el triple.

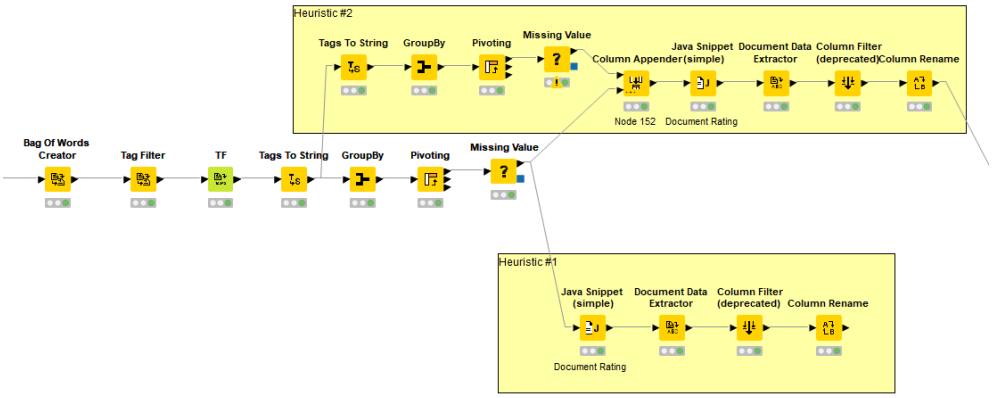

Figura 13: Cálculo de las diferentes heurísticas. Para la segunda, agrupamos no solo con la etiqueta de sentimiento sino además con la de POS.

2.2. Resultados

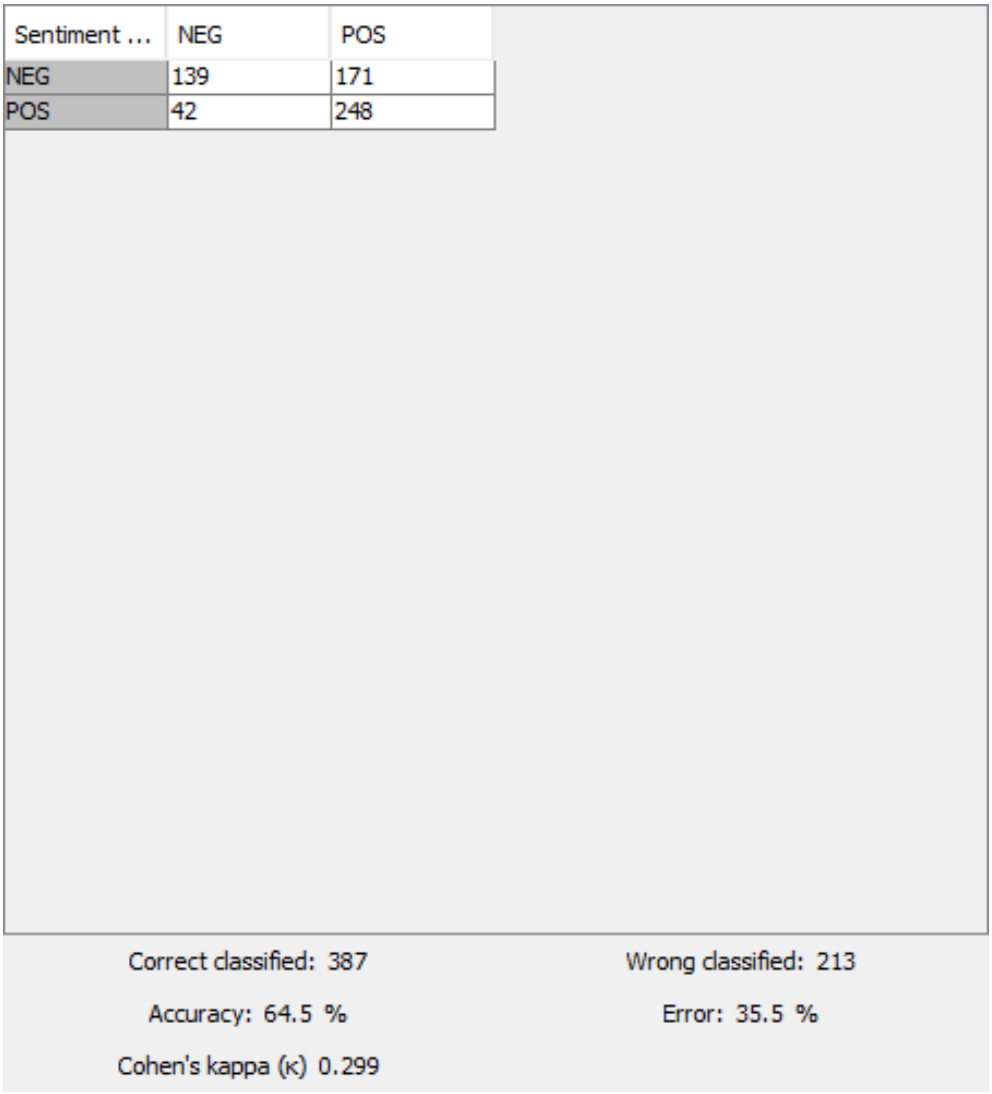

Figura 14: Matriz de confusión con MPQA y primera heurística.

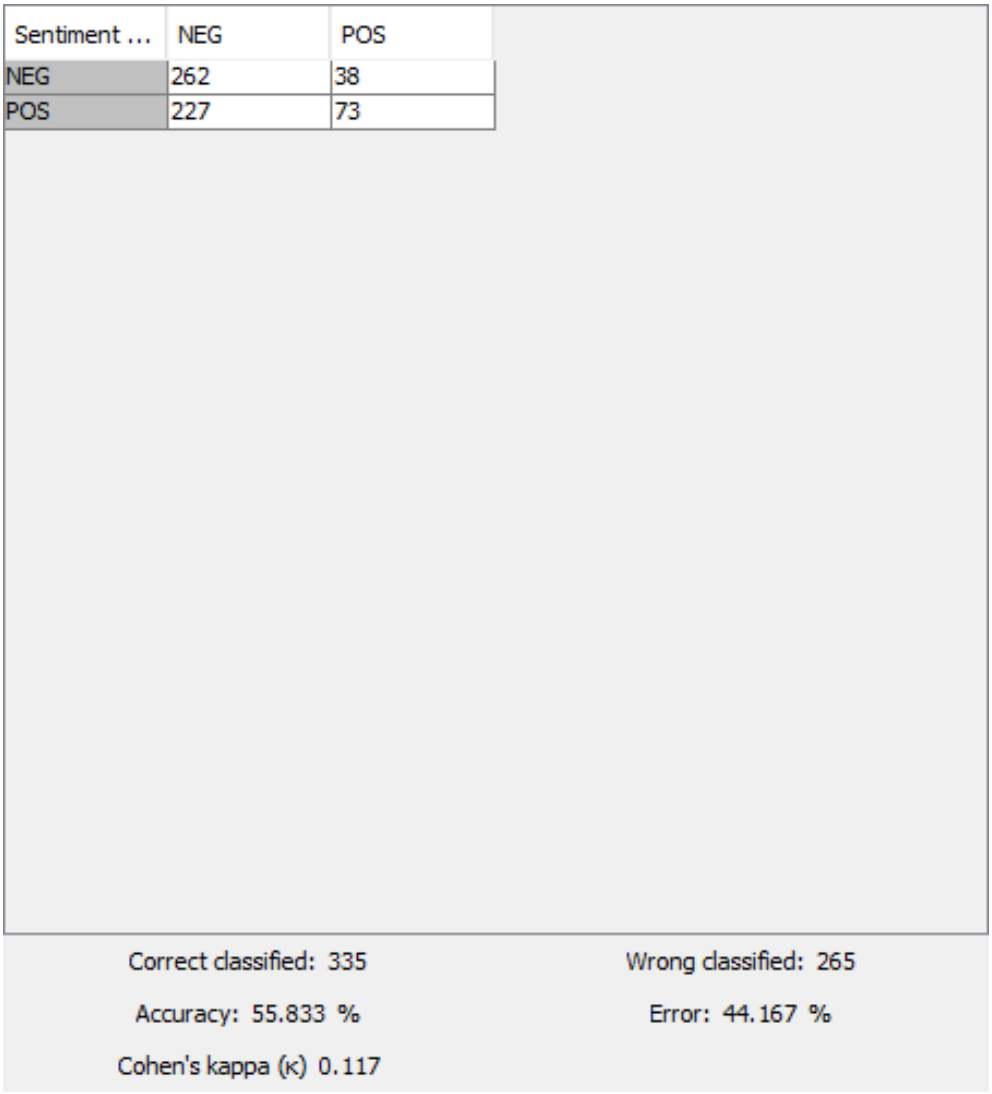

Figura 15: Matriz de confusión con SentiWordNet y primera heurística.

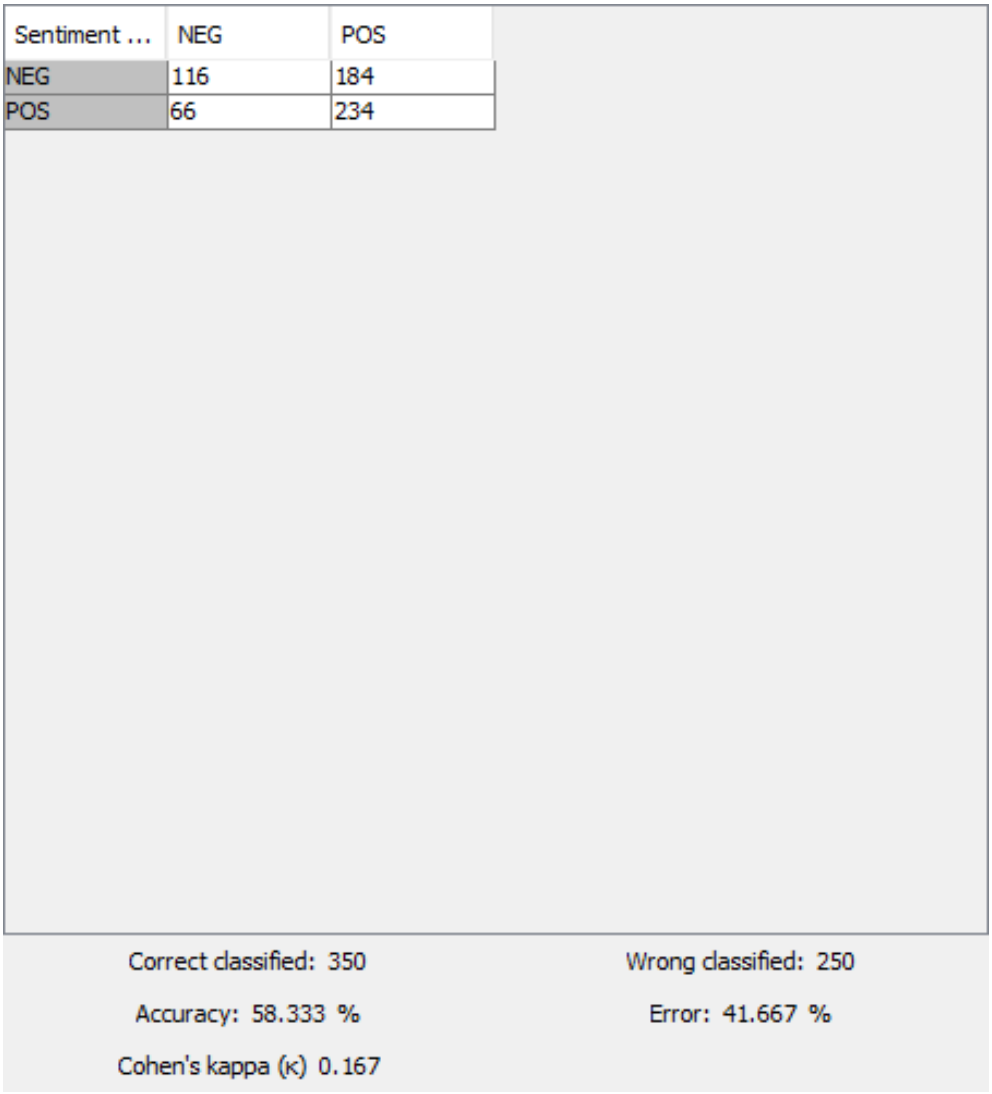

Figura 16: Matriz de confusión con SenticNet y primera heurística.

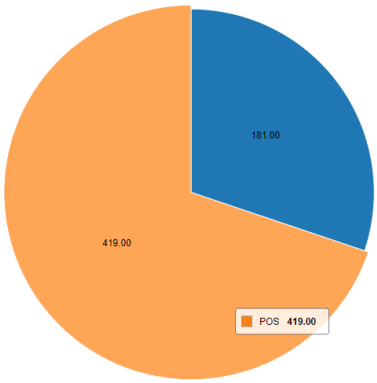

Figura 17: Pie chart con MPQA y primera heurística.

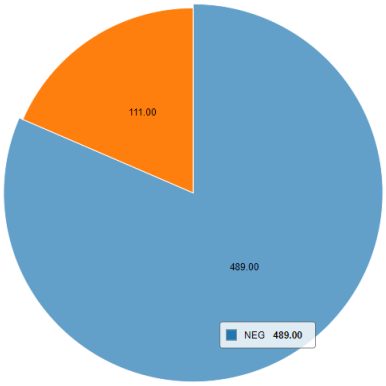

Figura 18: Pie chart con SentiWordNet y primera heurística.

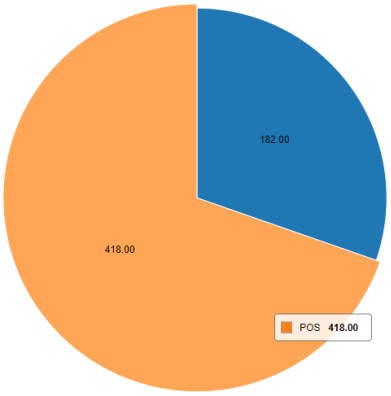

Figura 19: Pie chart con SenticNet y primera heurística.

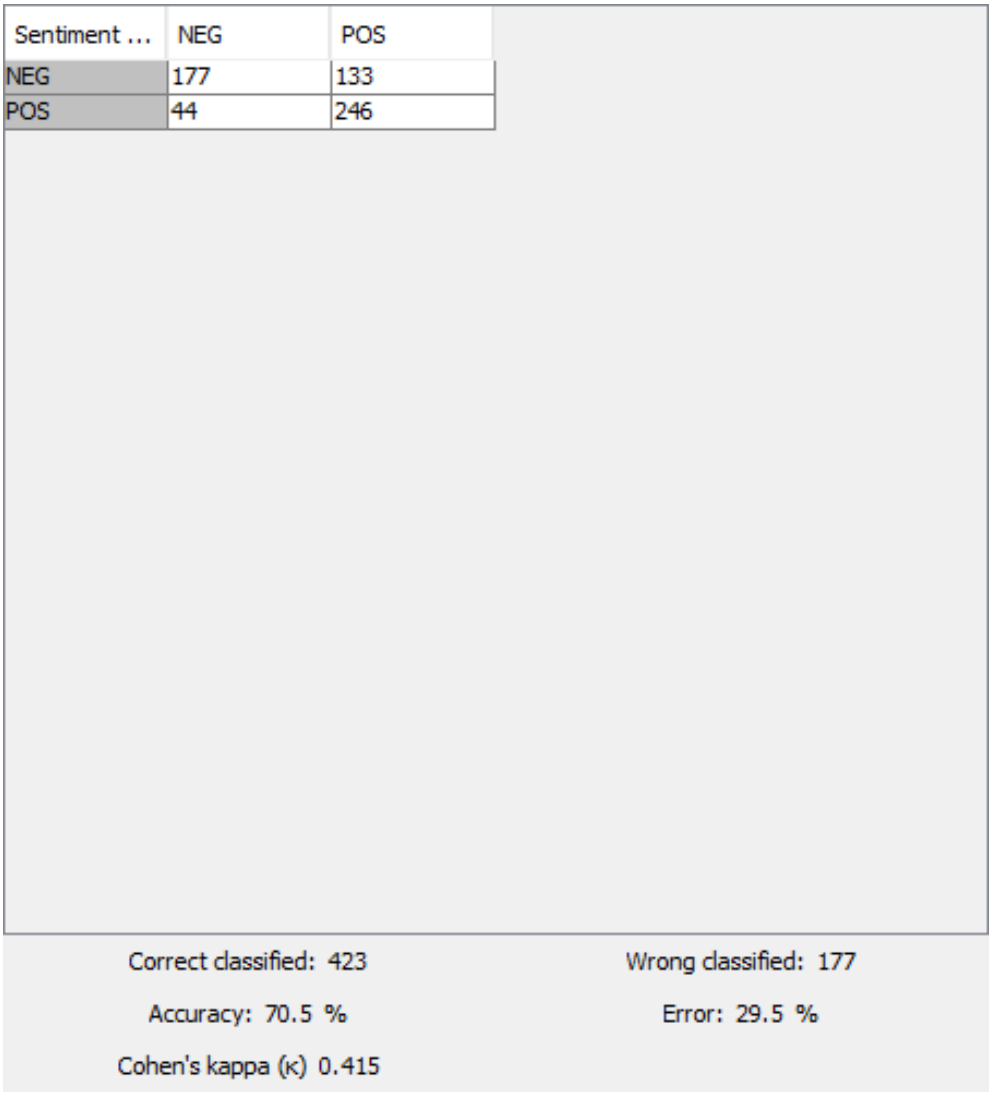

Figura 20: Matriz de confusión con MPQA y segunda heurística.

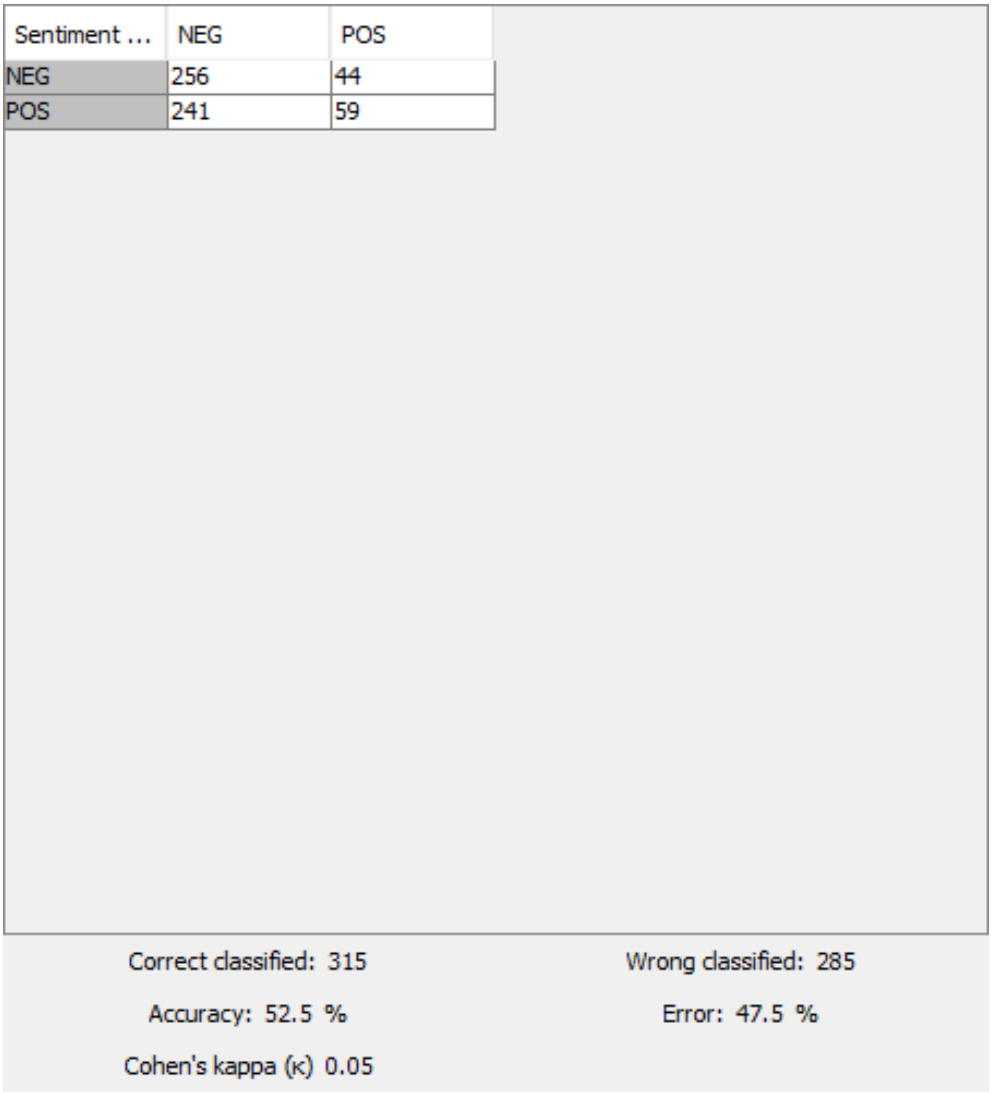

Figura 21: Matriz de confusión con SentiWordNet y segunda heurística.

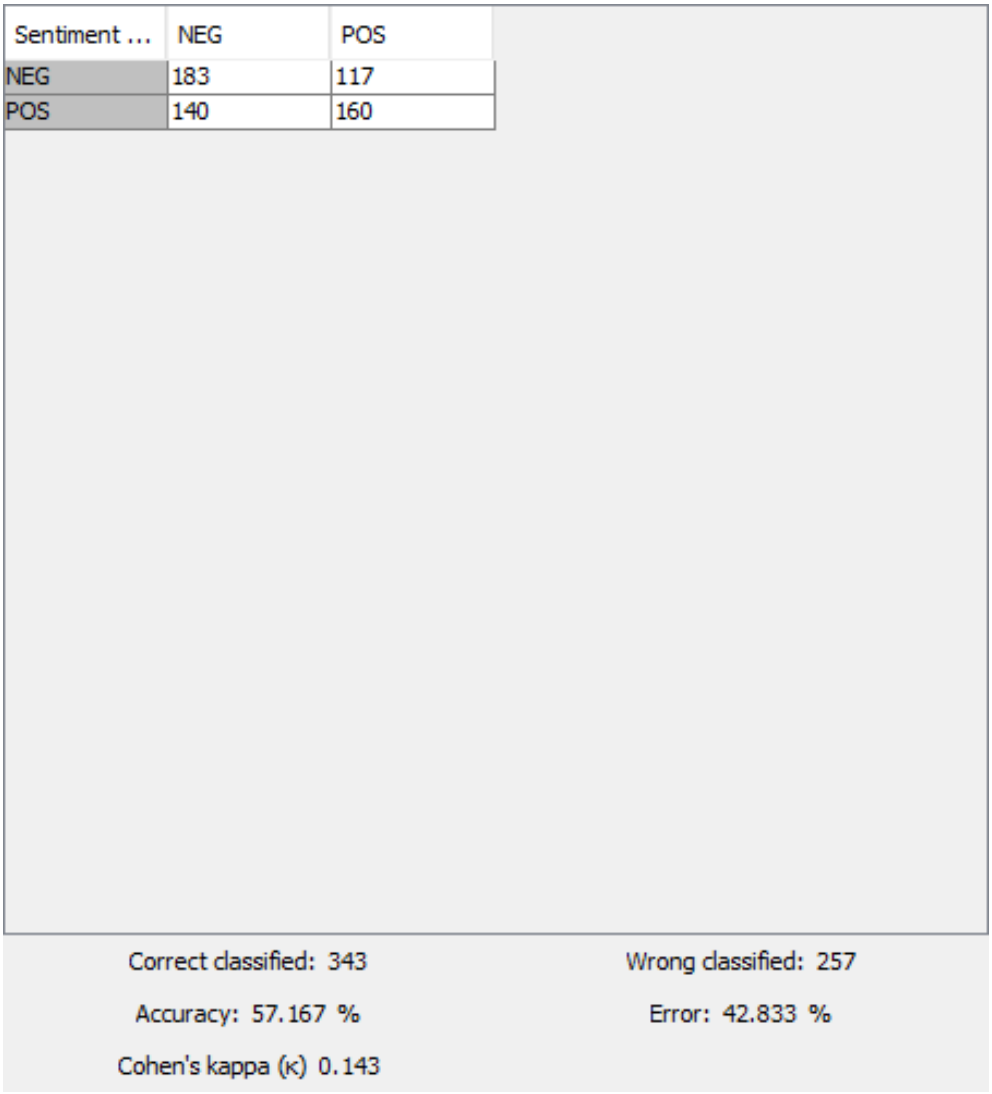

Figura 22: Matriz de confusión con SenticNet y segunda heurística.

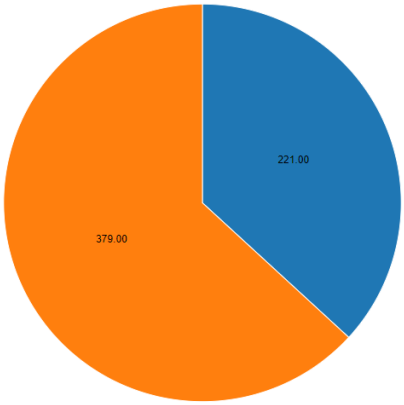

Figura 23: Pie chart con MPQA y segunda heurística.

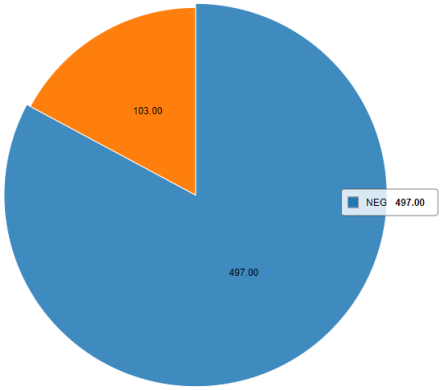

Figura 24: Pie chart con SentiWordNet y segunda heurística.

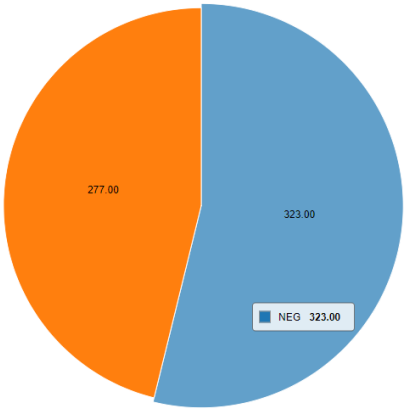

Figura 25: Pie chart con SenticNet y segunda heurística.

3. Conclusiones

Método	**Accuracy**	**Kappa**	**Specificity**	**Sensitivity**	**F-measure**
MPQA	64.5	0.299	0.591	0.768	0.566
SentiWordNet	55.8	0.117	0.657	0.525	0.664
SenticNet	58.3	0.167	0.559	0.637	0.481
MPQA	70.5	0.415	0.649	0.800	0.666
SentiWordNet	52.5	0.05	0.572	0.515	0.642
SenticNet	57.1	0.143	0.577	0.566	0.587
Redes Neuronales	82.5	0.65	0.854	0.800	0.832
Random Forest	91.0	0.82	0.870	0.950	0.913

Cuadro 1: Tabla de resultados generales.

Los resultados nos muestran una clara victoria de las técnicas de clasificación frente a los diccionarios.

Ambas heurísticas utilizadas en los diccionarios muestran efectos pobres, donde se ve una predicción excesiva de un solo sentimiento (dependiente según el lexicon). Con la segunda vemos que la predicción está más balanceada, mayormente para SenticNet, pero por desgracia esto no hace que mejore su accuracy, aunque sí el F-score. A pesar de esto la mejoría si es notoria en el corpus MPQA.

Los valores de *specificity* y *sensitivity* nos muestran las tasas de verdaderos negativos/positivos respectivamente. A partir de ellas vemos una mayor tendencia general de los diferentes métodos a valorar positivamente cada una de las reviews.