ALC – PRÁCTICA 2

SA para el lenguaje español

Cristian Villarroya Sánchez

Contenido

1. Introducción

Para esta práctica se va a realizar un sistema de análisis de sentimientos en español, utilizando los datos proporcionados (T1: TASS2017 (Task 1: Sentiment Analysis at Tweet level) http://www.sepln.org/workshops/tass/2017)

Los datos vienen separados en 3 partes, train, dev y test en formato XML, para su lectura se ha utilizado BeatifulSoup. Después, para el tokenizado de los tweets, se ha utilizado la librería TweetTokenizer de nltk.

2. Preproceso

Antes realizar cualquier experimento, es necesario realizar un preproceso de los datos. En esta ocasión se ha optado por realizar las siguientes acciones:

- Pasar el texto a minúsculas
- Eliminar todos los caracteres no alfanuméricos
- Eliminar saltos de línea
- Sustituir los nombres de usuario (@...) por “usuario”
- Sustituir los hashtags (#...) por “hashtag
- Eliminar las URL
- Sustituir los correos electrónicos por “correo”
- Eliminar las stop-words
- Eliminar los signos de puntuación
- Lematizar los tweets

Se ha probado a eliminar los correos, hashtags y nombres de usuario, en lugar de sustituirlos por las palabras correspondientes, pero el resultado final del sistema empeoraba. Por ejemplo, utilizando un vectorizador de tipo CountVectorizer y un clasificador SVC, la macro F1, sustituyendo los hashtags, correos y nombres de usuario es de 0,41. Mientras que, eliminándolas, es de 0,39.

3. Experimentos

Una vez realizado el preproceso, se ha estudiado la influencia del tipo de vectorizador y del clasificador a utilizar.

Han sido objeto de estudio los tipos de vectorizador CountVectorizer(CV), HashingVectorizer(HV), TFIDFVectorizer(TFIDFV) con n-gramas de tamaño 2,3 y 4 y finalmente TFIDFTransformer(TFIDFT) con CountVectorizer.

En cuanto a los clasificadores, se han estudiado las máquinas de vectores de soporte (SVC), regresión lineal (LR) y random forest (RF).

Para estudiar las diferentes combinaciones, se va a utilizar la métrica Macro F1.

A continuación, se muestra una tabla con los resultados de las diferentes combinaciones:

Los resultados obtenidos con el conjunto de dev han sido los siguientes:

CV + SVC:

	precision	recall	f1-score	support
N	0.60	0.67	0.63	219
NEU	0.19	0.14	0.16	69
NONE	0.35	0.27	0.31	62
P	0.54	0.56	0.55	156
accuracy			0.52	506
macro avg	0.42	0.41	0.41	506
weighted avg	0.50	0.52	0.50	506

HV + SVC:

	precision	recall	f1-score	support
N	0.57	0.73	0.64	219
NEU	0.10	0.03	0.04	69
NONE	0.34	0.16	0.22	62
P	0.54	0.61	0.57	156
accuracy			0.53	506
macro avg	0.39	0.38	0.37	506
weighted avg	0.47	0.53	0.49	506

TFIDFV + SVC

	precision	recall	f1-score	support
N	0.52	0.68	0.59	219
NEU	0.00	0.00	0.00	69
NONE	0.25	0.03	0.06	62
P	0.49	0.66	0.56	156
accuracy			0.50	506
macro avg	0.31	0.34	0.30	506
weighted avg	0.41	0.50	0.43	506

TFIDFT + SVC

	precision	recall	f1-score	support
N	0.59	0.66	0.62	219
NEU	0.11	0.06	0.08	69
NONE	0.29	0.24	0.27	62
P	0.54	0.60	0.57	156
accuracy			0.51	506
macro avg	0.38	0.39	0.38	506
weighted avg	0.47	0.51	0.49	506

CV + LR

	precision	recall	f1-score	support
N	0.56	0.74	0.64	219
NEU	0.11	0.03	0.05	69
NONE	0.34	0.16	0.22	62
P	0.54	0.60	0.57	156
accuracy			0.53	506
macro avg	0.39	0.38	0.37	506
weighted avg	0.47	0.53	0.49	506

HV + LR

	precision	recall	f1-score	support
N	0.54	0.77	0.64	219
NEU	0.33	0.01	0.03	69
NONE	0.36	0.08	0.13	62
P	0.55	0.62	0.58	156
accuracy			0.54	506
macro avg	0.44	0.37	0.34	506
weighted avg	0.49	0.54	0.47	506

TFIDFV + LR

	precision	recall	f1-score	support
N	0.52	0.70	0.59	219
NEU	0.00	0.00	0.00	69
NONE	0.25	0.03	0.06	62
P	0.50	0.64	0.56	156
accuracy			0.50	506
macro avg	0.32	0.34	0.30	506
weighted avg	0.41	0.50	0.44	506

TFIDFT + LR

	precision	recall	f1-score	support
N	0.55	0.73	0.63	219
NEU	0.25	0.01	0.03	69
NONE	0.35	0.18	0.24	62
P	0.54	0.63	0.58	156
accuracy			0.53	506
macro avg	0.42	0.39	0.37	506
weighted avg	0.48	0.53	0.48	506

CV + RF

	precision	recall	f1-score	support
N	0.52	0.74	0.61	219
NEU	0.00	0.00	0.00	69
NONE	0.12	0.02	0.03	62
P	0.52	0.61	0.56	156
accuracy			0.51	506
macro avg	0.29	0.34	0.30	506
weighted avg	0.40	0.51	0.44	506

HV + RF

	precision	recall	f1-score	support
N	0.56	0.81	0.66	219
NEU	0.29	0.03	0.05	69
NONE	0.44	0.06	0.11	62
P	0.55	0.61	0.58	156
accuracy			0.55	506
macro avg	0.46	0.38	0.35	506
weighted avg	0.51	0.55	0.49	506

TFIDFV + RF

	precision	recall	f1-score	support
N	0.52	0.65	0.58	219
NEU	0.33	0.03	0.05	69
NONE	0.15	0.06	0.09	62
P	0.49	0.63	0.55	156
accuracy			0.49	506
macro avg	0.37	0.34	0.32	506
weighted avg	0.44	0.49	0.44	506

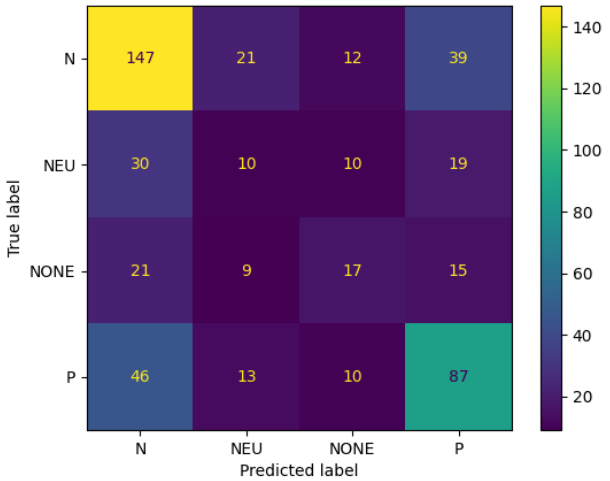
TFIDFT + RF

	precision	recall	f1-score	support
N	0.54	0.74	0.63	219
NEU	0.07	0.01	0.02	69
NONE	0.43	0.05	0.09	62
P	0.51	0.61	0.56	156
accuracy			0.52	506
macro avg	0.39	0.35	0.32	506
weighted avg	0.45	0.52	0.46	506

Como se puede observar, la combinación de CountVectorizer con SVC es el que mejor Macro F1 ha obtenido, qué es la métrica a maximizar y, por tanto, será la combinación final para evaluar el conjunto de test.

Por otro lado, si nos fijásemos en la accuracy, se ve que la mejor combinación es la de HashingVectorizer con RF.

Finalmente, se ha obtenido una matriz de confusión para la combinación ganadora (CV + SVC):



Como se puede ver en la matriz de confusión, el corpus no esta completamente balanceado, pues hay muchas más muestras positivas y negativas que neutras y “None”, lo cual hace que con estas dos últimas, le sea más complicada la clasificación y son las etiquetas con las que peor funciona el sistema.

Como último experimento se ha comprobado la influencia de usar el diccionario de polaridades. Para la combinación de CV + SVC, se han obtenido los siguientes resultados, que como puede verse, son peores que los anteriores usando dicho diccionario:

	precision	recall	f1-score	support
N	0.57	0.67	0.61	219
NEU	0.10	0.06	0.07	69
NONE	0.27	0.21	0.24	62
P	0.54	0.54	0.54	156
accuracy			0.49	506
macro avg	0.37	0.37	0.37	506
weighted avg	0.46	0.49	0.47	506

Para terminar este trabajo, se han predicho las polaridades para el conjunto de test utilizando la combinación CV+SVC, guardando el resultado en el formato especificado en un fichero de nombre CristianVillarroya_SVC.txt.