Javier Gil-Terrón Rodríguez
MIARFID 2021-2022

Sistemas de sentiment analysis

Aplicaciones de lingüística computacional

En primer lugar, se ha utilizado el parser de minidom para preprocesar los datos para, posteriormente, tokenizar manualmente las urls y menciones de Twitter y luego aplicar el TweetTokenizer y casual_tokenize de nltk. También se ha probado a utilizar 'mi_tokenizador' propuesto en el boletín de la tarea, pero se han obtenido peores resultados en todos los casos, por lo que las tablas que se mostrarán más adelante no lo contemplarán.

En cuanto a la vectorización, se han probado los métodos Tfidf y Count, y al igual que ocurría en el caso anterior, Count ofrecía peores prestaciones en todos los casos, por lo que los esfuerzos se han centrado en maximizar los resultados de Tfidf.

En cuanto a algoritmos de entrenamiento, se han probado distintas parametrizaciones para SVC, LinearSVC, GradientBoostingClassifier, SGDClassifier y KNeighborsClassifier, cuyos resultados se pueden ver reflejados a continuación:

Tabla 1 - Resultados obtenidos con SVC

	P	**R**	**F1**
N	0.58	0.71	0.64
NEU	0.14	0.1	0.12
NONE	0.24	0.16	0.19
P	0.58	0.55	0.57
Accuracy			0.51
Macro AVG	0.39	0.38	0.38

Tabla 2 - Resultados obtenidos con LinearSVC

	P	**R**	**F1**
N	0.58	0.67	0.62
NEU	0.24	0.17	0.20
NONE	0.20	0.13	0.16
P	0.55	0.58	0.56
Accuracy			0.51
Macro AVG	0.39	0.39	0.39

Tabla 3 - Resultados obtenidos con GradientBoostingClassifier

	P	R	F1
N	0.58	0.76	0.66
NEU	0.37	0.16	0.22
NONE	0.29	0.19	0.23
P	0.59	0.54	0.57
Accuracy			0.54
Macro AVG	0.45	0.42	0.42

Tabla 4 - Resultados obtenidos con SGDClassifier

	P	R	F1
N	0.58	0.67	0.64
NEU	0.18	0.12	0.14
NONE	0.22	0.16	0.19
P	0.55	0.58	0.57
Accuracy			0.50
Macro AVG	0.38	0.38	0.38

Tabla 5 - Resultados obtenidos con KNeighborsClassifier

	P	R	F1
N	0.55	0.69	0.61
NEU	0.18	0.12	0.14
NONE	0.25	0.26	0.25
P	0.53	0.42	0.47
Accuracy			0.48
Macro AVG	0.38	0.37	0.37

A la vista de estos resultados, la mejor configuración encontrado en términos de Macro F1 la ha obtenido GradientBoostingClassifier con un 0.42, configuración que ha sido aplicada al conjunto de test y con la que se he generado el archivo .txt adjunto en la tarea.