Trabajo Final

Extracción, Análisis y Clasificación de una Web con reseñas de películas

Facultad De Ingeniería, Universidad De Cuenca
TEXT MINING
Freddy L. Abad L.
{freddy.abadl}@ucuenca.edu.ec

- **¿Cuál es el tamaño del conjunto de datos luego de eliminar las filas que no contienen información pertinente como por ejemplo valores es castellano u otro información irrelevante.**

Inicialmente se contaba con 1276 registros, sin embargo luego de eliminar los datos no etiquetados se obtuvo 1095 registros, de los cuales 495 registros se utilizan para entrenar el modelo y 495 registros para validar el modelo. En los dos casos se tiene 247 +/- 1 registros positivos o negativos debido a que los datos deben estar equilibrados en su tipo de etiqueta de calificación (Positiva y Negativa).

- **Llenar el cuadro con la precisión del modelo entrenado con cada uno de los algoritmos utilizados:**

En medida del Weight of Information Gain, se analizó varios límites, que acotan el número de características que mejoran/empeoran los resultados del modelo de clasificación.

Algoritmo SVM

Select Weight	Class Recall - True Positive	Class Recall - True Negative	Accuracy	Clasificación Positiva	Clasificación Negativa
>=0.002	70.97	47.97	59.18	176	117
>=0.003	97.18	90.69	93.94	241	224
>=0.004	97.18	92.31	94.74	241	228
>=0.005	95.97	89.88	92.92	238	222
>=0.006	93.95	91.5	92.73	233	226
>=0.007	93.15	85.43	89.30	231	211
>=0.008	82.19	93.15	87.67	231	203
>=0.009	94.35	70.87	82.66	234	175

Algoritmo Naive Bayes (NB)

Select Weight	Class Recall - True	Class Recall - True	Accuracy	Clasificación Positiva	Clasificación Negativa

	Positive	Negative			
>=0.002	58.87	54.25	56.55	146	113
>=0.003	97.58	95.95	96.76	242	237
>=0.004	97.58	95.95	96.76	242	237
>=0.005	95.56	93.93	94.75	237	232
>=0.006	95.56	94.33	94.96	237	233
>=0.007	87.90	91.09	89.50	218	225
>=0.008	88.71	95.14	91.92	220	235
>=0.009	86.69	87.45	87.09	215	216

Conclusión Modelamiento: Del análisis dependiendo del **Weight of Information Gain** obtiene que usando el algoritmo SVM, dados estos datos se obtiene mejor accuracy con valores mayores o igual a 0.004. Y dado el algoritmo de NB se obtienen mejores accuracy con valores mayores o iguales a 0.003 y 0.004.
Al momento de validar los Modelos entrenados dependiendo del Weight of Information Gain, se obtienen los siguientes resultados:

Algoritmo SVM

Select Weight	Class Recall - True Positive	Class Recall - True Negative	Accuracy	Clasificación Positiva	Clasificación Negativa
>=0.002	70.04	51.21	60.61	173	127
>=0.003	70.85	50.00	60.04	175	124
>=0.004	70.85	50.40	60.61	175	125
>=0.005	68.42	54.44	61.41	169	135
>=0.006	65.59	58.47	62.02	162	145
>=0.007	65.99	54.44	60.20	163	135
>=0.008	69.64	46.77	58.18	172	116
>=0.009	74.90	46.77	60.81	185	116

Algoritmo Naive Bayes

Select Weight	Class Recall - True	Class Recall - True	Accuracy	Clasificación Positiva	Clasificación Negativa

	Positive	Negative			
>=0.002	56.68	50.81	53.74	140	126
>=0.003	56.68	51.61	51.14	140	128
>=0.004	56.68	51.61	54.14	140	128
>=0.005	64.78	51.21	57.98	160	127
>=0.006	63.97	50.81	57.37	158	126
>=0.007	58.30	53.63	55.95	144	133
>=0.008	57.09	53.63	55.35	141	133
>=0.009	61.13	56.85	58.99	151	141

Conclusión Validación Modelo: Usando el algoritmo de SVM se obtienen un mejor accuracy con un Weight of Information mayor o igual a 0.006. Y usando el algoritmo de Naive Bayes se obtiene un mejor accuracy con un Weigth equivalente o mayor a 0.009. Dados los resultados, se podría intuir que ante un buen accuracy de un modelo en su entrenamiento, se obtendría mejores accuracy al validar el modelo. Sin embargo en este caso, se obtuvieron mejores resultados en diversos modelos al momento de entrenar y validar.

Cuadro de Resumen - Entrenamiento

Algoritmo	Select Weight	Class Recall - True Positive	Class Recall - True Negative	Accuracy
Naive Bayes	**>=0.004**	97.18	92.31	94.74
SVM	**>=0.004**	97.58	95.95	96.76

Cuadro de Resumen - Validación

Algoritmo	Select Weight	Class Recall - True Positive	Class Recall - True Negative	Accuracy
Naive Bayes	**>=0.006**	65.59	58.47	62.02

SVM	**>=0.009**	61.13	56.85	58.99

- **Identificar cuál es el número de predicciones correctas e incorrectas**

Cuadro de Resumen - Entrenamiento

Algoritmo	Select Weight	Predicciones Correctas	Predicciones Incorrectas
Naive Bayes	**>=0.004**	Predicción Positiva: 242 Predicción Negativa: 237 Total Predicciones: 479	Predicción Positiva: 6 Predicción Negativa: 10 Total Predicciones: 16
SVM	**>=0.004**	Predicción Positiva: 241 Predicción Negativa: 228 Total Predicciones: 469	Predicción Positiva: 7 Predicción Negativa: 19 Total Predicciones: 26

Cuadro de Resumen - Validación

Algoritmo	Select Weight	Predicciones Correctas	Predicciones Incorrectas
Naive Bayes	**>=0.006**	Predicción Positiva: 158 Predicción Negativa: 126 Total Predicciones: 284	Predicción Positiva: 90 Predicción Negativa: 121 Total Predicciones: 211
SVM	**>=0.009**	Predicción Positiva: 185 Predicción Negativa: 116 Total Predicciones: 301	Predicción Positiva: 63 Predicción Negativa: 131 Total Predicciones: 194