

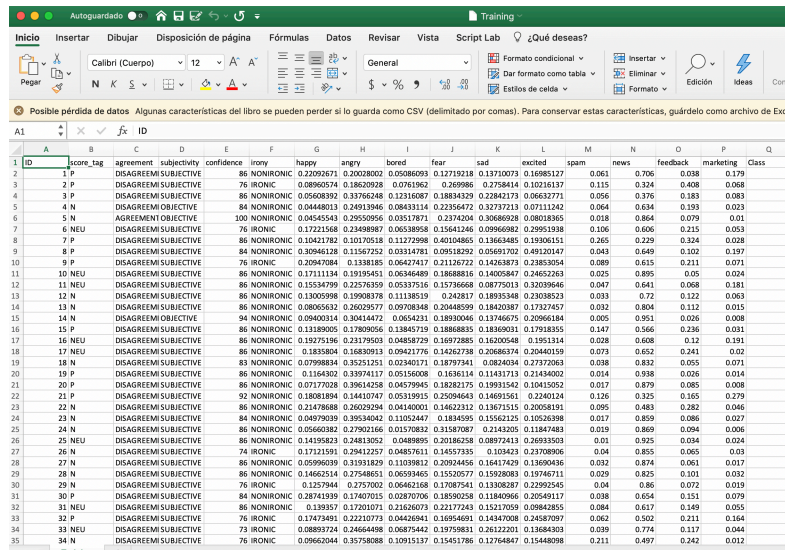
H9 – Testing a selected classifier

Introduction

In this homework, two classifiers of the type Logistic Regression were used: Linear and Multinomial. Both classifiers were used for checking which news are fake and which isn't. With the help of WEKA, this task was easy to realize.

Development & Results

First of all, a new feature was extracted using the intent API of Parallel Dots, this new feature named Marketing was extracted for both the Training and Testing datasets.



ID	score_tag	agreement	subjectivity	confidence	irony	happy	angry	bored	fear	sad	excited	spam	news	feedback	marketing	Class
1	P	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.22092671	0.20028002	0.05086093	0.12719218	0.13710073	0.16985127	0.061	0.706	0.038	0.179	1
2	P	DISAGREEMENT	SUBJECTIVE	76	IRONIC	0.0890574	0.18620928	0.0761862	0.209986	0.2758414	0.10216137	0.115	0.324	0.408	0.008	0
3	P	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.05608392	0.33766248	0.32316087	0.18834329	0.22842173	0.06632771	0.056	0.376	0.183	0.083	1
4	P	DISAGREEMENT	SUBJECTIVE	84	NONIRONIC	0.0448013	0.24913946	0.08433114	0.22356472	0.32737213	0.07111242	0.064	0.634	0.193	0.023	1
5	N	AGREEMENT	OBJECTIVE	100	NONIRONIC	0.04545543	0.29530956	0.03517871	0.2374204	0.30680928	0.08028395	0.018	0.864	0.079	0.031	1
6	NEU	DISAGREEMENT	SUBJECTIVE	76	IRONIC	0.17221568	0.23498987	0.06338058	0.15641246	0.09966987	0.29951938	0.106	0.006	0.215	0.053	0
7	P	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.10421782	0.10170518	0.13272998	0.40104865	0.13663485	0.19306151	0.265	0.229	0.324	0.028	1
8	P	DISAGREEMENT	SUBJECTIVE	84	NONIRONIC	0.30946128	0.11567252	0.03314781	0.09518292	0.0581702	0.49120147	0.043	0.649	0.102	0.197	0
9	P	DISAGREEMENT	SUBJECTIVE	76	IRONIC	0.20947084	0.1338185	0.06427417	0.21126722	0.14263873	0.23853054	0.089	0.615	0.211	0.071	0
10	NEU	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.17111134	0.19195451	0.06346489	0.18688816	0.14005847	0.24652263	0.025	0.895	0.05	0.024	0
11	NEU	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.15534799	0.22576359	0.05337516	0.15736668	0.08775013	0.32039646	0.047	0.641	0.068	0.181	0
12	N	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.33003998	0.19908378	0.11138519	0.242817	0.18955348	0.23028523	0.033	0.72	0.122	0.063	0
13	N	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.08065632	0.26029577	0.09798348	0.20448599	0.18420387	0.17327457	0.032	0.804	0.112	0.015	1
14	N	DISAGREEMENT	SUBJECTIVE	94	NONIRONIC	0.09400314	0.30414472	0.0654231	0.18930046	0.13746675	0.20966184	0.005	0.951	0.026	0.008	1
15	P	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.13380005	0.17809506	0.13845719	0.18888835	0.18360931	0.17918355	0.147	0.566	0.236	0.031	0
16	NEU	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.19275196	0.23179503	0.04858729	0.16972885	0.16200548	0.1951314	0.028	0.608	0.12	0.191	0
17	NEU	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.1835804	0.16830913	0.09421776	0.14262738	0.20686374	0.20440159	0.073	0.652	0.241	0.02	1
18	N	DISAGREEMENT	SUBJECTIVE	83	NONIRONIC	0.07998834	0.35251251	0.02340171	0.18797341	0.0824034	0.27372063	0.038	0.832	0.055	0.071	1
19	P	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.1564302	0.19374117	0.03156008	0.1816114	0.11411713	0.21434002	0.014	0.938	0.006	0.014	0
20	P	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.07177028	0.39614258	0.04579945	0.18282175	0.19931542	0.10415052	0.017	0.879	0.085	0.008	1
21	P	DISAGREEMENT	SUBJECTIVE	92	NONIRONIC	0.18081894	0.14410747	0.03319915	0.25094643	0.14891541	0.2240124	0.126	0.325	0.165	0.279	0
22	N	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.21476608	0.26029504	0.04140001	0.14622312	0.13671515	0.20058191	0.095	0.483	0.282	0.046	0
23	N	DISAGREEMENT	SUBJECTIVE	84	NONIRONIC	0.04979039	0.39534042	0.11052447	0.1834595	0.15562125	0.10526398	0.017	0.859	0.086	0.027	1
24	N	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.05660382	0.27902166	0.01570832	0.31587087	0.2143205	0.11847483	0.019	0.869	0.094	0.006	0
25	NEU	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.14195823	0.24813052	0.0489895	0.20186258	0.08972413	0.26935503	0.01	0.925	0.034	0.024	1
26	N	DISAGREEMENT	SUBJECTIVE	74	IRONIC	0.17121591	0.29412257	0.04857611	0.14557335	0.103423	0.23708906	0.04	0.855	0.065	0.03	0
27	N	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.05996039	0.31931829	0.11039812	0.20924456	0.16417429	0.13690436	0.032	0.874	0.061	0.017	1
28	N	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.14862514	0.27548651	0.06593465	0.15520577	0.13928083	0.19746711	0.029	0.825	0.101	0.032	0
29	N	DISAGREEMENT	SUBJECTIVE	76	IRONIC	0.12572944	0.2757002	0.06462168	0.17987541	0.13936287	0.22952445	0.04	0.86	0.072	0.019	0
30	P	DISAGREEMENT	SUBJECTIVE	84	NONIRONIC	0.28741939	0.17407015	0.02870706	0.18920258	0.11840966	0.20549117	0.038	0.654	0.151	0.079	1
31	NEU	DISAGREEMENT	SUBJECTIVE	86	NONIRONIC	0.139357	0.17201071	0.21626073	0.22172343	0.15217059	0.09842855	0.084	0.617	0.149	0.055	0
32	P	DISAGREEMENT	SUBJECTIVE	76	IRONIC	0.17474501	0.22210773	0.04626041	0.16954691	0.14347008	0.24587097	0.062	0.502	0.211	0.164	0
33	NEU	DISAGREEMENT	SUBJECTIVE	73	IRONIC	0.08893724	0.24664498	0.06875442	0.19759831	0.26122201	0.13684303	0.039	0.774	0.117	0.044	0
34	N	DISAGREEMENT	SUBJECTIVE	76	IRONIC	0.09662044	0.35758088	0.10915137	0.15451786	0.12764827	0.15448098	0.211	0.497	0.242	0.012	0

Figure 1. Training.csv with the new marketing feature.

Linear Logistic Regression

The first classifier used in this homework was a Linear Logistic Regression Classifier. For using this classifier, the option *SimpleLogistic* was selected in WEKA. Then, the following results were obtained.

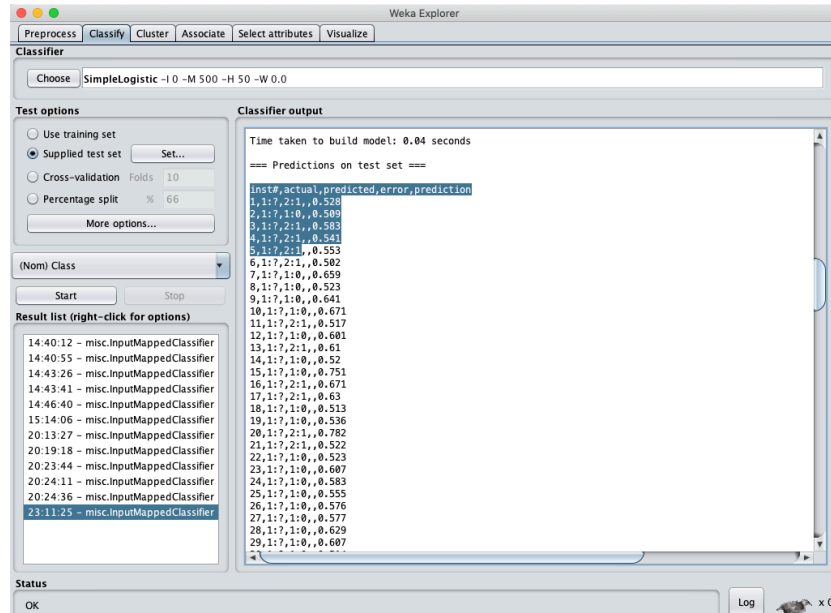


Figure 2. WEKA results obtained from SimpleLogistic classifier.

After that, by using the Bash Script named *Filter.sh*, we cut the unnecessary columns and obtain the format necessary for uploading our results to Kaggle.

Public Leaderboard Private Leaderboard						
This leaderboard is calculated with all of the test data.						
				Raw Data	Refresh	
#	Team Name	Notebook	Team Members	Score	Entries	Last
1	JJoseCortesSarmiento			0.74444	77	4h
2	Oscar Cañongo			0.73333	63	3d
3	Saúl Gheno Hernández			0.72500	106	10h
4	Diego C.			0.72500	44	6d
5	Antonio Sanchez			0.70833	41	11h
6	Begoña Montes Gómez			0.69166	38	7d
7	Daniela Alvarado Pereda			0.68888	130	3h
8	Estefania Pitol			0.68611	37	1d
9	Hector Duran Herrera			0.67777	59	2h
10	Nicolas Albo			0.66666	47	now
Your Best Entry						
Your submission scored 0.66666, which is not an improvement of your best score. Keep trying!						

Figure 3. Score obtained in Kaggle by the SimpleLogistic classifier.

As it can be appreciated, the score obtained was good because it was the same score as the best score, therefore, it can be concluded that the classifier performed well.

Multinomial Logistic Regression

The second classifier used was a Multinomial Logistic Regression Classifier. For using this classifier, the option *Logistic* was selected in WEKA. The following results were obtained.

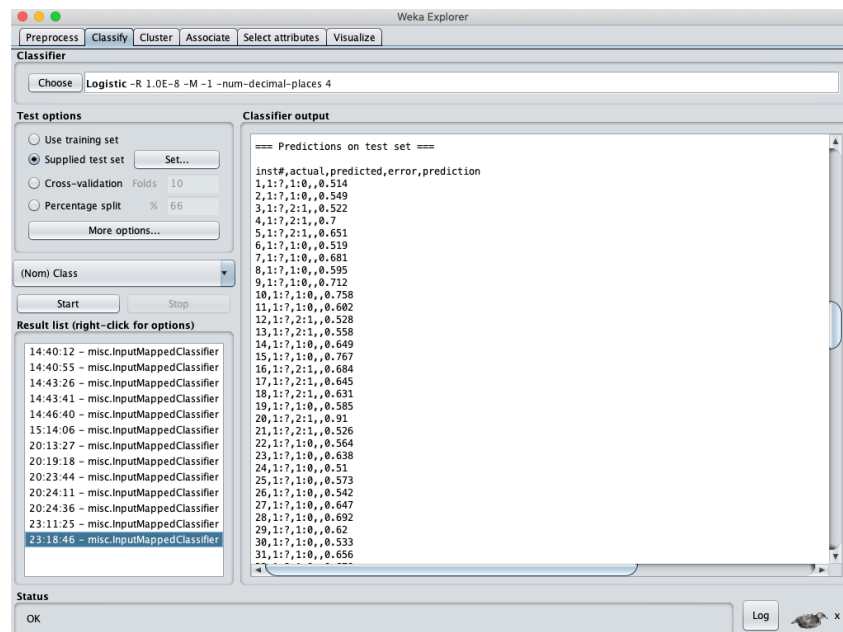


Figure 4. WEKA results obtained from Logistic classifier.

After that, by using the Bash Script named *Filter.sh*, we cut the unnecessary columns and obtain the format necessary for uploading our results to Kaggle.

Public Leaderboard

Private Leaderboard

This leaderboard is calculated with all of the test data.

[Raw Data](#)

[Refresh](#)

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	JJoseCortesSarmiento			0.74444	77	4h
2	Oscar Cañongo			0.73333	63	3d
3	Saúl Ghenno Hernández			0.72500	106	10h
4	Diego C.			0.72500	44	6d
5	Antonio Sanchez			0.70833	41	11h
6	Begoña Montes Gómez			0.69166	38	7d
7	Daniela Alvarado Pereda			0.68888	130	3h
8	Estefania Pitol			0.68611	37	1d
9	Nicolas Albo			0.68055	48	now

Your Best Entry

Your submission scored 0.68055, which is an improvement of your previous score of 0.66666. Great job!

[Tweet this!](#)

Figure 5. Score obtained in Kaggle by the Logistic classifier.

As it can be appreciated, this classifier performed better than the previous classifier used, but not only that, it also performed better than my best score, setting a new best score.

Conclusion

Logistic classifiers may be quite simple if compared with other type of classifiers like Decision Trees or Bayesian Nets, but for this specific problem, they obtained better results than the other type of classifiers mentioned above.