

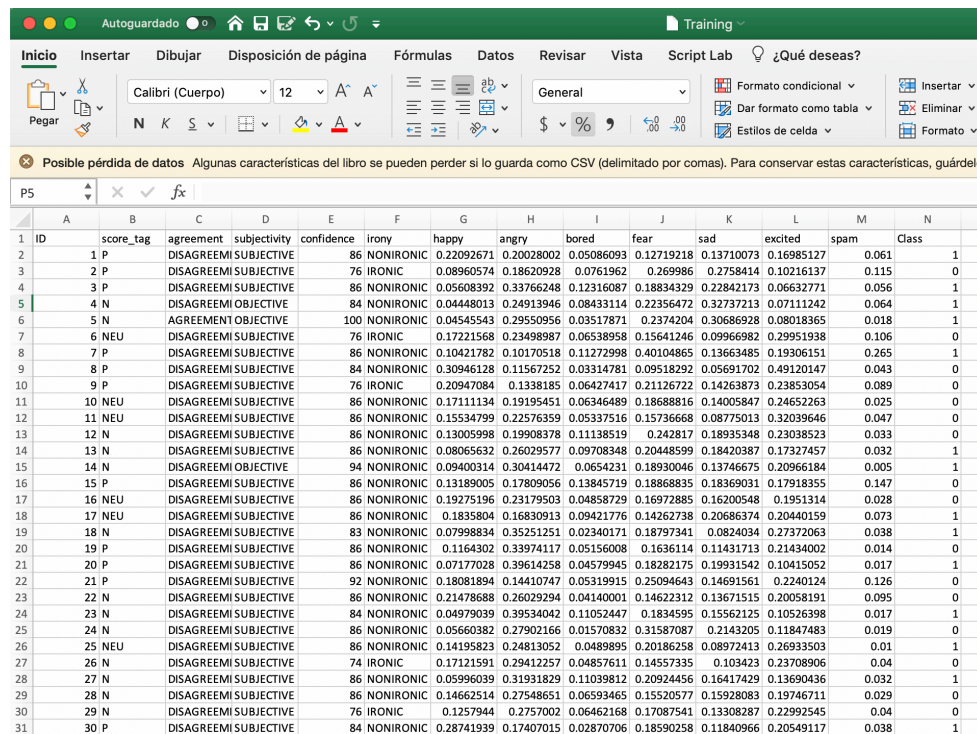
H6 – Testing the bagging and adaboost methods

Introduction

Bagging and boosting methods are useful for dealing with noisy objects in machine learning projects. In WEKA, both methods are available and are going to be used in this homework for solving the Fake News problem.

Development & Results

First of all, a new feature named spam was extracted for the Testing and Training datasets using the Intent Analysis API of Parallel Dots.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	score_tag	agreement	subjectivity	confidence	irony	happy	angry	bored	fear	sad	excited	spam	Class
2	1	P	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.22092671	0.20028002	0.05086093	0.12719218	0.13710073	0.16985127	0.061	1
3	2	P	DISAGREEMI	SUBJECTIVE	76	IRONIC	0.08960574	0.18620928	0.0761962	0.269986	0.2758414	0.10216137	0.115	0
4	3	P	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.05608392	0.33766248	0.12316087	0.18834329	0.22842173	0.06632771	0.056	1
5	4	N	DISAGREEMI	OBJECTIVE	84	NONIRONIC	0.04448013	0.24913946	0.08433114	0.22356472	0.32737213	0.07111242	0.064	1
6	5	N	AGREEMENT	OBJECTIVE	100	NONIRONIC	0.04545543	0.29550956	0.03517871	0.2374204	0.30686928	0.08018365	0.018	1
7	6	NEU	DISAGREEMI	SUBJECTIVE	76	IRONIC	0.17221568	0.23498987	0.06538958	0.15641246	0.09966982	0.29951938	0.106	0
8	7	P	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.10421782	0.10170518	0.11272998	0.40104865	0.13663485	0.19306151	0.265	1
9	8	P	DISAGREEMI	SUBJECTIVE	84	NONIRONIC	0.30946128	0.11567252	0.03314781	0.09518292	0.05691702	0.49120147	0.043	0
10	9	P	DISAGREEMI	SUBJECTIVE	76	IRONIC	0.20947084	0.1338185	0.06427417	0.21126722	0.14263873	0.23853054	0.089	0
11	10	NEU	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.17111134	0.19195451	0.06346489	0.18688816	0.14005847	0.24652263	0.025	0
12	11	NEU	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.15534799	0.22576359	0.05337516	0.15736668	0.08775013	0.32039646	0.047	0
13	12	N	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.13005998	0.19908378	0.11138519	0.242817	0.18935348	0.23038523	0.033	0
14	13	N	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.08065632	0.26029577	0.09708348	0.20448599	0.18420387	0.17327457	0.032	1
15	14	N	DISAGREEMI	OBJECTIVE	94	NONIRONIC	0.09400314	0.30414472	0.0654231	0.18930046	0.13746675	0.20966184	0.005	1
16	15	P	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.13189005	0.17809056	0.13845719	0.18868835	0.18369031	0.17918355	0.147	0
17	16	NEU	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.19275196	0.23179503	0.04858729	0.16972885	0.16200548	0.1951314	0.028	0
18	17	NEU	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.1835804	0.16830913	0.09421776	0.14262738	0.20686374	0.20440159	0.073	1
19	18	N	DISAGREEMI	SUBJECTIVE	83	NONIRONIC	0.07998834	0.35251251	0.02340171	0.18797341	0.0824034	0.27372063	0.038	1
20	19	P	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.1164302	0.33974117	0.05156008	0.1636114	0.11431713	0.21434002	0.014	0
21	20	P	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.07177028	0.39614258	0.04579945	0.18282175	0.19931542	0.10415052	0.017	1
22	21	P	DISAGREEMI	SUBJECTIVE	92	NONIRONIC	0.18081894	0.14410747	0.05319915	0.25094643	0.14691561	0.2240124	0.126	0
23	22	N	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.21478688	0.26029294	0.04140001	0.14622312	0.13671515	0.20058191	0.095	0
24	23	N	DISAGREEMI	SUBJECTIVE	84	NONIRONIC	0.04979039	0.39534042	0.11052447	0.1834595	0.15562125	0.10526398	0.017	1
25	24	N	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.05660382	0.27902166	0.01570832	0.31587087	0.2143205	0.11847483	0.019	0
26	25	NEU	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.14195823	0.24813052	0.0489895	0.20186258	0.08972413	0.26933503	0.01	1
27	26	N	DISAGREEMI	SUBJECTIVE	74	IRONIC	0.17121591	0.29412257	0.04857611	0.14557335	0.103423	0.23708906	0.04	0
28	27	N	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.05996039	0.31931829	0.11039812	0.20924456	0.16417429	0.13690436	0.032	1
29	28	N	DISAGREEMI	SUBJECTIVE	86	NONIRONIC	0.14662514	0.27548651	0.06593465	0.15520577	0.15928083	0.19746711	0.029	0
30	29	N	DISAGREEMI	SUBJECTIVE	76	IRONIC	0.1257944	0.2757002	0.06462168	0.17087541	0.13308287	0.22992545	0.04	0
31	30	P	DISAGREEMI	SUBJECTIVE	84	NONIRONIC	0.28741939	0.17407015	0.02870706	0.18590258	0.11840966	0.20549117	0.038	1

Figure 1. Trainig.csv with the new spam feature.

Then both Training.csv and Testing.csv were transformed into WEKA files.

AdaBoostM1 with DecisionStump

For this method, in the Classify tab of WEKA, AdaBoostM1 is selected and then DecisionStump is picked as the classifier.

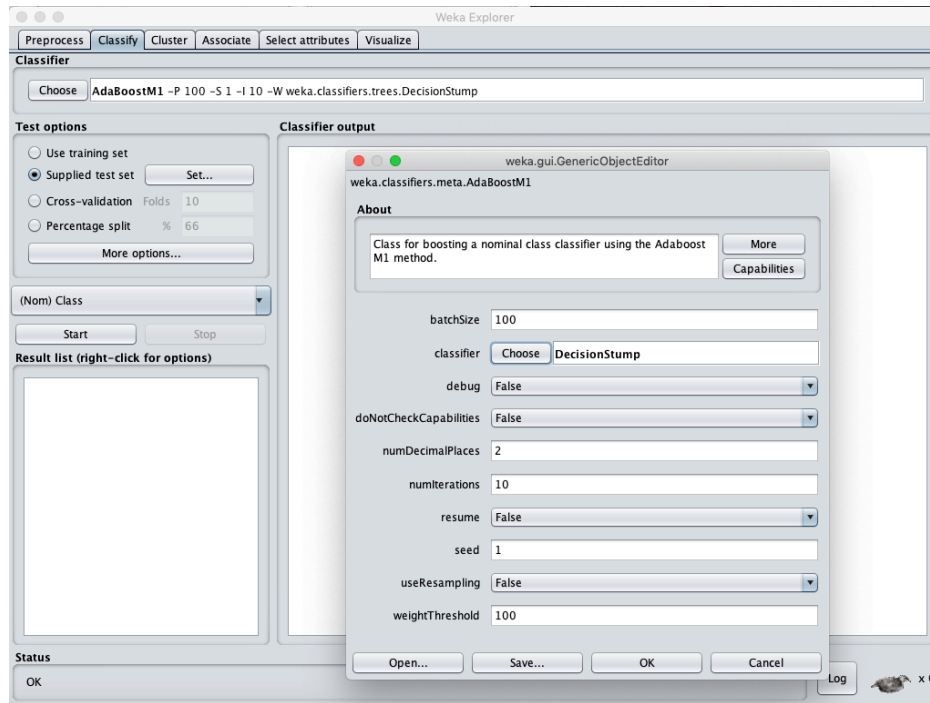


Figure 2. AdaBoostM1 and DecisionStump in WEKA.

The following results are obtained.

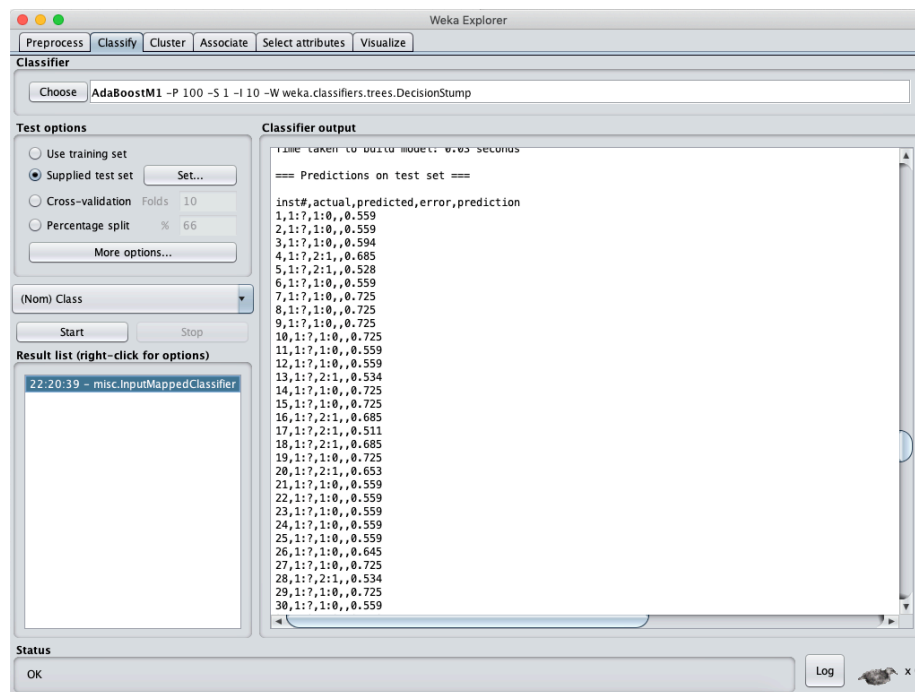


Figure 3. Results of using AdaBoostM1 and DecisionStump in CSV format.

Then, data is filtered and uploaded 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	Saúl Ghenno Hernández			0.72500	90	2d
2	Diego C.			0.71111	28	2d
3	Daniela Alvarado Pereda			0.68611	58	2d
4	Oscar Cañongo			0.68333	45	3d
5	Hector Duran Herrera			0.67777	47	4h
6	Begoña Montes Gómez			0.67500	29	3d
7	JJoseCortesSarmiento			0.67222	63	2d
8	Estefania Pitol			0.67222	31	2d
9	Nicolas Albo			0.66666	38	now
Your Best Entry						
Your submission scored 0.61111, which is not an improvement of your best score. Keep trying!						

Figure 4. Score in Kaggle obtained by the AdaBoostM1 method with DesicionStump classifier.

Although it got a score higher than 0.60000, it couldn't beat my Kaggle's best entry, so maybe another method with another classifier will beat my high score.

Bagging with LMT

For this method, in the Classify tab of WEKA, Bagging is selected and then LMT is picked as the classifier.

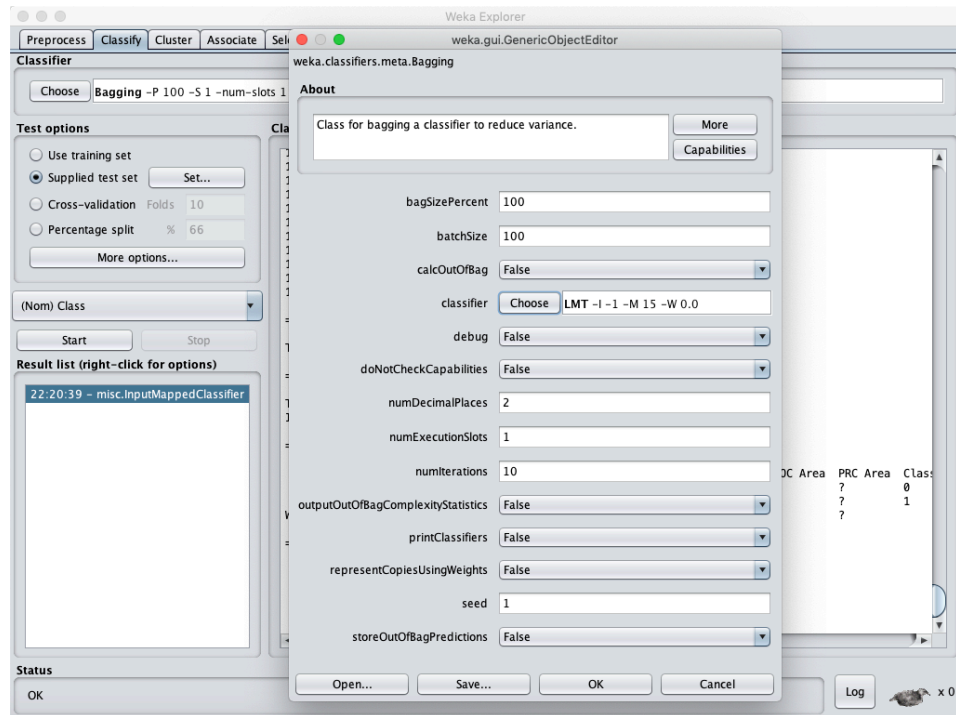


Figure 5. Bagging and LMT in WEKA.

The following results are obtained.

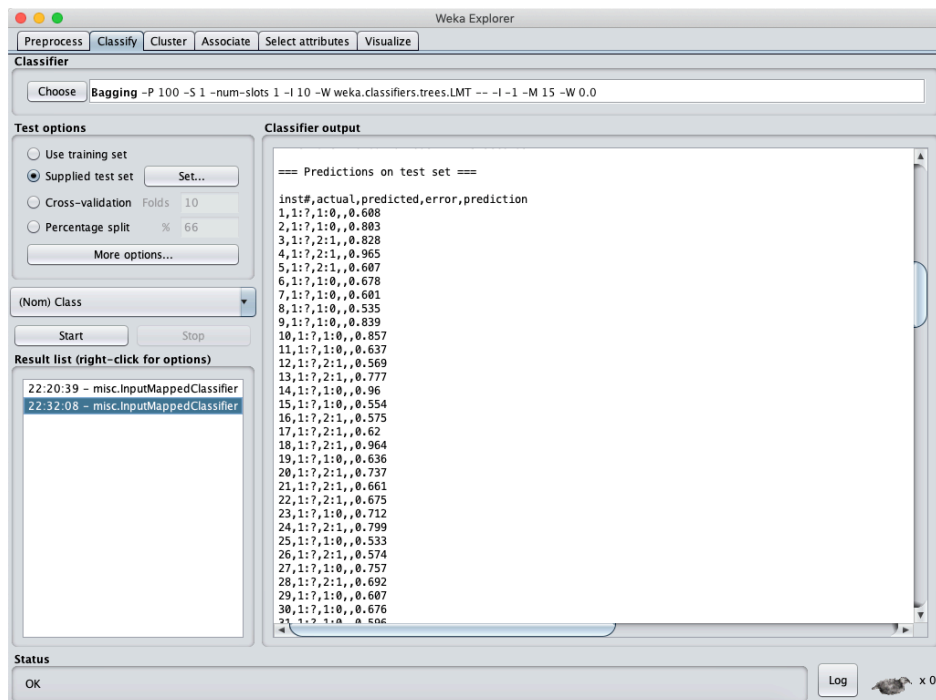


Figure 6. Results of using Bagging and LMT in CSV format.

Then, data is filtered and uploaded to Kaggle.

Overview

Data

Notebooks

Discussion

Leaderboard

Rules

Team

My Submissions

Submit Predictions

This leaderboard is calculated with all of the test data.

Raw Data

Refresh

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Saúl Ghenno Hernández			0.72500	90	2d
2	Diego C.			0.71111	28	2d
3	Daniela Alvarado Pereda			0.68611	58	2d
4	Oscar Cañongo			0.68333	45	4d
5	Hector Duran Herrera			0.67777	47	4h
6	Begoña Montes Gómez			0.67500	29	3d
7	JJoseCortesSarmiento			0.67222	63	2d
8	Estefania Pitol			0.67222	31	2d
9	Nicolas Albo			0.66666	39	now

Your Best Entry

Your submission scored 0.59166, which is not an improvement of your best score. Keep trying!

Figure 7. Score in Kaggle obtained by the Bagging method with LMT classifier.

In this case, the result could not even break the 0.60000 barrier and it was quite far from my best entry.

Conclusion

AdaBoostM1 method with DesicionStump classifier resulted more effective than Bagging method with LMT classifier. Maybe for this specific problem of detecting fake news, AdaBoostM1 is more effective than Bagging, but also the classifiers influence in the results.