TABLE 12
Classification performance of our proposal (DE) against the rest of the selected methods measured with accuracy (acc), precision (pre), recall (rec), $f1$-score (f1) and area-under-curve ROC (rocauc) when training with 9 features. Results show that DE performs better in 3 out of 4 datasets.

FS method	acc	pre	rec	f1	rocauc
airlines dataset					
CHI2	0.8745	0.6364	0.5251	0.5696	0.7334
DE	0.8664	0.6308	0.4640	0.5295	0.7039
ECE	0.8696	0.6694	0.3795	0.4824	0.6717
f-ANOVA	0.8771	0.6456	0.5406	0.5835	0.7413
GSS	0.8686	0.6450	0.4816	0.5379	0.7123
IG	0.8792	0.6537	0.5383	**0.5869**	0.7415
MI	0.8786	0.6562	0.5257	0.5802	0.7361
OR	0.8679	0.6488	0.4287	0.5113	0.6906
gender dataset					
CHI2	0.5361	0.4954	0.6805	0.5099	0.5417
DE	0.5704	0.5401	0.7084	**0.6100**	0.5759
ECE	0.5430	0.5073	0.7293	0.5706	0.5492
f-ANOVA	0.5478	0.5153	0.7369	0.5685	0.5553
GSS	0.5282	0.5636	0.2052	0.2436	0.5157
IG	0.5372	0.5125	0.6883	0.5232	0.5429
MI	0.5360	0.5022	0.6842	0.5176	0.5416
OR	0.5343	0.5051	0.6631	0.5251	0.5386
imdb dataset					
CHI2	0.5792	0.7027	0.5326	0.4899	0.5829
DE	0.6205	0.6748	0.5447	**0.5790**	0.6216
ECE	0.5505	0.5781	0.4160	0.4635	0.5494
f-ANOVA	0.5825	0.6567	0.6761	0.5598	0.5882
GSS	0.5885	0.6927	0.4466	0.4989	0.5922
IG	0.5852	0.7247	0.5081	0.4933	0.5894
MI	0.5885	0.7319	0.5023	0.4874	0.5919
OR	0.5725	0.6311	0.4234	0.4825	0.5737
tass dataset					
CHI2	0.6780	0.7735	0.3755	0.4729	0.6274
DE	0.6901	0.6779	0.4299	**0.5255**	0.6467
ECE	0.6067	0.5907	0.2251	0.2987	0.5433
f-ANOVA	0.6930	0.7985	0.3438	0.4677	0.6340
GSS	0.6594	0.7222	0.4005	0.4722	0.6153
IG	0.6974	0.8062	0.3217	0.4583	0.6346
MI	0.6945	0.8084	0.3433	0.4685	0.6352
OR	0.6376	0.6158	0.2958	0.3882	0.5804

TABLE 13
Performance dispersion of our proposal (DE) against the rest of the selected methods measured with range, interquartile range (IQR), standard deviation (STD) and coefficient of variation (CV) when training with 9 features. Results show that DE performs better in 2 out of 4 datasets.

FS method	range	IQR	STD	CV (%)
airlines dataset				
CHI2	0.2835	0.0290	0.0671	11.7734
DE	0.1833	0.1018	0.0603	11.3899
ECE	0.2158	0.0185	0.0478	9.9008
f-ANOVA	0.2496	0.0242	0.0511	8.7641
GSS	0.1740	0.0986	0.0570	10.6049
IG	0.2306	0.0364	0.0501	8.5442
MI	0.2273	0.0315	0.0474	**8.1757**
OR	0.2366	0.0639	0.0608	11.8858
gender dataset				
CHI2	0.6253	0.1494	0.2537	49.7470
DE	0.2185	0.0124	0.0471	**7.7200**
ECE	0.5159	0.0188	0.1741	30.5157
f-ANOVA	0.5760	0.0159	0.1939	34.1072
GSS	0.4959	0.0260	0.1373	56.3815
IG	0.5460	0.1397	0.2286	43.6941
MI	0.5921	0.1404	0.2384	46.0551
OR	0.4917	0.1248	0.2024	38.5512
imdb dataset				
CHI2	0.4921	0.3695	0.1849	37.7442
DE	0.2283	0.0722	0.0618	**10.6673**
ECE	0.3600	0.1852	0.1086	23.4350
f-ANOVA	0.4719	0.3198	0.1813	32.3765
GSS	0.2556	0.1483	0.0867	17.3774
IG	0.4180	0.3076	0.1551	31.4443
MI	0.4132	0.3603	0.1691	34.7056
OR	0.3127	0.1125	0.0884	18.3206
tass dataset				
CHI2	0.1926	0.0275	0.0386	8.1571
DE	0.1478	0.0558	0.0360	6.8562
ECE	0.4576	0.0381	0.0880	29.4679
f-ANOVA	0.1692	0.0346	0.0322	6.8889
GSS	0.2002	0.0716	0.0492	10.4178
IG	0.1029	0.0576	0.0298	**6.5040**
MI	0.1725	0.0346	0.0324	6.9091
OR	0.3455	0.1114	0.1015	26.1393

TABLE 14
Classification performance of a deep neural network using ELMo embeddings measured with accuracy (acc), precision (pre), recall (rec), $f1$-score (f1) and area-under-curve ROC (rocauc) without limiting the number of features.

	acc		pre		rec		f1		rocauc	
	mean	std	mean	std	mean	std	mean	std	mean	std
dataset										
airlines	0.8926	0.0064	0.7757	0.0860	0.5037	0.1487	0.5915	0.0888	0.7358	0.0636
gender	0.5893	0.0378	0.6528	0.0845	0.3647	0.1727	0.4376	0.1530	0.5808	0.0387
imdb	0.7020	0.0893	0.6992	0.1518	0.8328	0.1701	0.7355	0.0667	0.7138	0.0799
s140	0.7313	0.0025	0.7249	0.0185	0.7483	0.0449	0.7353	0.0136	0.7312	0.0025
tass	0.5589	0.1094	0.6429	0.2306	0.4950	0.4763	0.3677	0.2659	0.5436	0.0459

TABLE 15
Classification performance against k-nearest neighbours (kNN) complexity for each dataset and feature selection method. The lower the number of neighbours is and the higher the $f1$-score, the better.

	$f1$-score					
no. of neighbours	5	7	9	10	25	$\sum \frac{f1}{k}$
FS method						
airlines dataset						
CHI2	0.2653	0.2763	0.3108	0.2638	0.5918	0.1771
DE	0.5608	0.5604	0.5661	0.2455	0.5688	0.3024
ECE	0.3582	0.3902	0.4426	0.4264	0.5045	0.2394
f-ANOVA	0.3468	0.5773	0.5740	0.5698	0.5970	0.2965
GSS	0.5370	0.5436	0.5359	0.5157	0.5833	0.3195
IG	0.6000	0.6164	0.5931	0.5869	0.5744	**0.3556**
MI	0.6000	0.6164	0.5931	0.5869	0.5744	**0.3556**
OR	0.5335	0.5203	0.4805	0.4581	0.4323	0.2975
gender dataset						
CHI2	0.6425	0.6432	0.6455	0.6447	0.6519	**0.3827**
DE	0.6451	0.6248	0.6452	0.1281	0.6474	0.3287
ECE	0.6157	0.1178	0.1181	0.1029	0.1695	0.1702
f-ANOVA	0.6390	0.6374	0.6324	0.6335	0.6348	0.3779
GSS	0.1579	0.1445	0.1448	0.1258	0.1883	0.0884
IG	0.6375	0.6425	0.6415	0.6403	0.6382	0.3801
MI	0.1532	0.1706	0.1508	0.0940	0.1350	0.0866
OR	0.6452	0.6302	0.6282	0.6280	0.6340	0.3770
imdb dataset						
CHI2	0.6763	0.6763	0.6763	0.4160	0.4160	0.3652
DE	0.6966	0.6966	0.6966	0.6966	0.6966	**0.4138**
ECE	0.6116	0.6160	0.6213	0.3016	0.4354	0.3269
f-ANOVA	0.6788	0.6788	0.6791	0.6791	0.3902	0.3917
GSS	0.6577	0.6577	0.6577	0.5294	0.5294	0.3727
IG	0.6788	0.6840	0.6840	0.4733	0.4733	0.3757
MI	0.6764	0.6764	0.6764	0.3902	0.6764	0.3731
OR	0.6071	0.5442	0.5616	0.4818	0.5478	0.3317
tass dataset						
CHI2	0.4522	0.4522	0.4814	0.4522	0.4814	0.2730
DE	0.3747	0.3747	0.3730	0.3762	0.3762	0.2226
ECE	0.4370	0.3993	0.4007	0.3474	0.3789	0.2389
f-ANOVA	0.4496	0.4062	0.4062	0.4062	0.4010	0.2497
GSS	0.3913	0.3656	0.4666	0.4238	0.4169	0.2414
IG	0.5428	0.5657	0.4554	0.4560	0.4547	**0.3038**
MI	0.4470	0.4489	0.4503	0.4484	0.4475	0.2663
OR	0.4520	0.4225	0.4208	0.3675	0.4236	0.2512

TABLE 16
Performance and complexity measures of DT-based models for each dataset. Number of leaves (rules) and path length (clauses) can be used to compare how comprehensible are the models that result from training with different feature sets. Our proposal is within the best of them and it produce the only feature set that can manage to keep a good classification performance while maintaining a simple model in the gender classification task.

FS method	f1	rules	clauses	complexity	compr. rate
airlines dataset					
CHI2	0.5981	34	7.4706	7.5901	**7.8796**
DE	0.5698	41	7.2683	8.6638	6.5766
ECE	0.5026	69	9.1159	22.9357	2.1912
f-ANOVA	0.6063	44	7.6136	10.2023	5.9430
GSS	0.5821	49	7.4082	10.7567	5.4111
IG	0.6016	47	7.6383	10.9686	5.4845
MI	0.6016	47	7.6383	10.9686	5.4845
OR	0.5515	92	9.0543	30.1691	1.8280
gender dataset					
CHI2	0.6429	309	13.7799	234.6999	0.2739
DE	0.6444	28	6.5000	4.7320	**13.6171**
ECE	0.6328	371	13.7278	279.6620	0.2263
f-ANOVA	0.6381	234	13.0684	159.8524	0.3992
GSS	0.1833	48	8.5417	14.0083	1.3087
IG	0.6423	311	13.3698	222.3661	0.2889
MI	0.6421	302	13.1623	209.2798	0.3068
OR	0.6356	333	13.0541	226.9839	0.2800
imdb dataset					
CHI2	0.6763	15	6.6667	2.6667	25.3597
DE	0.6966	16	6.7500	2.9160	23.8899
ECE	0.4082	44	10.0227	17.6801	2.3086
f-ANOVA	0.6791	14	6.5000	2.3660	28.7026
GSS	0.5976	17	7.1176	3.4449	17.3460
IG	0.6889	16	6.5000	2.7040	25.4767
MI	0.6764	11	5.4545	1.3091	**51.6667**
OR	0.5385	36	8.0000	9.2160	5.8427
tass dataset					
CHI2	0.4807	23	6.8261	4.2868	**11.2146**
DE	0.3735	24	6.7083	4.3202	8.6457
ECE	0.3055	901	14.0954	716.0488	0.0427
f-ANOVA	0.4520	23	7.1739	4.7348	9.5465
GSS	0.4316	36	8.3611	10.0668	4.2870
IG	0.4877	25	7.2800	5.2998	9.2020
MI	0.4509	22	6.9091	4.2007	10.7334
OR	0.3414	834	13.4233	601.0936	0.0568

TABLE 17
Performance and complexity measures of RF-based models for each dataset when training with 5 trees. Number of leaves (rules) and path length (clauses) can be used to compare how comprehensible are the models that result from training with different features selection mechanisms.

FS method	f1	rules	clauses	complexity	compr. rate
airlines dataset					
CHI2	0.6116	58.2000	8.2904	16.0004	**3.8222**
DE	0.5691	65.8000	8.1672	17.5561	3.2417
ECE	0.5057	160.2000	9.8588	62.2831	0.8120
f-ANOVA	0.6057	68.0000	8.5646	19.9519	3.0356
GSS	0.5749	87.4000	8.6319	26.0483	2.2070
IG	0.6194	84.8000	9.2644	29.1130	2.1274
MI	0.6031	84.2000	9.1788	28.3757	2.1255
OR	0.5502	150.2000	9.5490	54.7826	1.0043
gender dataset					
CHI2	0.6417	293.0000	13.8191	223.8153	0.2867
DE	0.6442	41.8000	7.4090	9.1780	**7.0187**
ECE	0.6321	365.6000	13.6071	270.7671	0.2334
f-ANOVA	0.6385	251.2000	12.5725	158.8265	0.4020
GSS	0.1892	53.0000	8.9622	17.0279	1.1109
IG	0.6427	284.2000	12.5812	179.9404	0.3572
MI	0.6411	292.4000	13.6436	217.7181	0.2945
OR	0.6364	337.6000	13.1601	233.8737	0.2721
imdb dataset					
CHI2	0.6810	14.4000	6.3957	2.3561	**28.9034**
DE	0.6966	17.0000	6.6829	3.0370	22.9384
ECE	0.4371	50.2000	8.8899	15.8692	2.7543
f-ANOVA	0.6840	16.4000	6.6628	2.9122	23.4883
GSS	0.5976	21.0000	7.2516	4.4172	13.5280
IG	0.6889	19.4000	7.1179	3.9315	17.5221
MI	0.6764	14.6000	6.4763	2.4494	27.6129
OR	0.5939	43.4000	8.9320	13.8499	4.2884
tass dataset					
CHI2	0.4814	55.6000	8.5205	16.1460	2.9815
DE	0.3730	43.0000	7.7223	10.2572	**3.6366**
ECE	0.2974	899.6000	14.7129	778.9390	0.0382
f-ANOVA	0.4533	57.6000	8.7544	17.6579	2.5673
GSS	0.4324	73.2000	9.0092	23.7654	1.8196
IG	0.4900	56.4000	8.6753	16.9789	2.8861
MI	0.4542	48.2000	8.6188	14.3219	3.1714
OR	0.3432	833.2000	14.4092	691.9755	0.0496