Sentiment Analysis of Movie Reviews Using Multiple Machine Learning Methods

Jack St. Clair & Zachary Toner

★ 1/10

The Worst Film Ever Made
MattClayton92   27 April 2016

When someone calls a film "terrible", they're usually either exaggerating or have no concept of what a terrible film truly is. The Room embodies that concept. It's not just poorly written, poorly acted, poorly directed - poorly everything - it honestly looks like

The Dataset

- Large Movie Review Dataset: http://ai.stanford.edu/~amaas/data/sentiment/
- 50,000 movie reviews: 25,000 training, 25,000 testing
- "Highly polar": Only contains scores of 1-4 and 7-10
- Bag of words format
- Goals

Bag of words

1	the	2616	jobs	
2	and	2617	sir	
3	a	2618	stayed	
4	of	2619	stays	
5	to	2620	code	
6	is	2621	ned	
7	it	2622	significant	
8	in	2623	wear	
9	i	2624	mature	
10	this	2625	rights	
11	that	2626	lifetime	
12	was	2627	examples	
13	as	2628	paying	
14	for	2629	studios	
15	with	2630	jungle	
16	movie	2631	ghosts	
17	but			
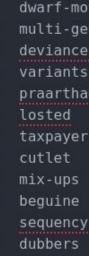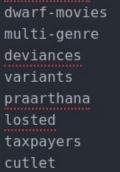
21737	upstage	89457	googling
21738	compensates	89458	skg
21739	conceiving	89459	hit-persons
21740	dashiell	89460	timecrimes
21741	crystals	89461	dwarf-movies
21742	locataire	89462	multi-genre
21743	exacting	89463	deviances
21744	italian-americans	89464	variants
21745	twenty-somethings	89465	praarthana
21746	extracting	89466	losted
21747	assisting	89467	taxpayers
21748	consumes	89468	cutlet
21749	anxieties	89469	mix-ups
21750	sinker	89470	beguine
21751	thornfield	89471	sequency
21752	flourished	89472	dubbers
21753	stettner	89473	zillionaire
21754	droids	89474	alt-rock
21755	boomer	89475	i-don't-need-revenge-nor-my-husband-
		89476	catepillar

Bag of words 2

8   10 0:8 1:4 2:3 3:5 4:7 5:1 6:1 8:5 9:6 10:2 11:3 13:2 14:1 17:4 19:1 20:1 22:1 25:1 26:1 27:2 33:1 34:1 35:2 36:1 37:2 39:1 42:1 47:1 50:1 53:1 56:3 58:1 62:
    2718:1 3397:1 4499:2 4642:1 7140:1 8084:1 9748:1 22982:1 27765:1
9   9 0:3 1:3 2:1 3:2 5:2 6:1 7:1 8:1 9:1 15:2 26:1 35:1 37:1 40:1 44:1 46:1 64:1 78:1 84:1 98:1 116:1 153:1 174:1 185:1 297:2 298:1 323:1 389:1 540:2 568:1 1104
10 4 0:12 1:4 2:6 3:4 4:5 5:4 6:2 7:2 8:3 9:2 10:3 11:4 12:1 13:3 14:2 16:3 17:2 18:1 22:1 23:1 24:1 25:1 29:1 30:1 31:1 32:1 33:1 38:2 41:1 43:2 45:1 50:1 53:1
    1505:1 1522:1 1590:1 1735:1 1793:1 1920:1 2081:1 2106:1 2268:2 2631:1 3307:1 3308:1 3316:1 3663:1 4143:1 4172:1 4643:1 5458:1 5703:2 6089:1 6729:1 7544:1 8249
11 10 1:1 2:2 3:2 4:1 5:2 6:1 7:2 9:1 10:1 12:1 13:1 14:1 17:1 22:1 23:1 114:1 115:1 132:1 148:2 281:1 290:1 306:1 327:1 543:1 565:1 594:1 667:1 677:1 681:1 687
12 8 0:16 1:11 2:12 3:14 4:11 5:6 6:4 7:9 8:14 9:2 10:1 11:2 13:4 14:2 16:4 17:13 18:4 19:2 20:3 21:5 22:3 23:3 25:2 26:1 27:11 28:4 32:3 34:3 35:1 36:5 37:2 39
    313:1 341:1 345:1 357:1 363:1 369:1 384:1 393:1 427:2 435:2 439:1 461:1 479:1 483:1 489:2 511:1 548:1 570:2 579:1 732:1 796:1 838:1 845:1 868:1 942:1 1031:2 1
    6774:2 7908:1 8095:1 8105:1 8938:1 10013:1 10067:1 10238:1 11469:1 11544:2 13079:1 15181:1 16140:1 84632:1

Pickle

- Allows python data structures to be stored in a file

K-Nearest Neighbors

Binary	Predicted Bad Reviews	Predicted Good Reviews
True Bad Reviews	550	471
True Good Reviews	312	667

Accuracy: 0.6085

Hyperparameters
Weights: Uniform
Nearest neighbors: 5

K-Nearest Neighbors

Accuracy: 0.2465

Predicted Labels

Multiclass	1	2	3	4	7	8	9	10
1	211.	25.	7.	20.	6.	39.	12.	99.
2	90.	8.	2.	9.	6.	25.	9.	47.
3	63.	14.	6.	20.	7.	26.	5.	62.
4	60.	4.	9.	22.	9.	36.	10.	53.
7	51.	8.	4.	13.	12.	37.	4.	39.
8	50.	5.	8.	24.	5.	54.	12.	58.
9	30.	8.	7.	15.	12.	40.	10.	69.
10	109.	10.	7.	29.	5.	56.	18.	170.

True Labels

Random Forests

Binary	Predicted Bad Reviews	Predicted Good Reviews
True Bad Reviews	847	174
True Good Reviews	179	800

Accuracy: 0.8235

Hyperparameters
Max_Features: sqrt

Random Forests

Accuracy: 0.363

Predicted Labels

Multiclass	1	2	3	4	7	8	9	10
1	329.	3.	3.	5.	2.	6.	0.	71.
2	133.	2.	6.	3.	3.	4.	0.	45.
3	124.	0.	3.	6.	7.	8.	0.	55.
4	79.	3.	5.	15.	9.	14.	0.	78.
7	40.	1.	2.	9.	6.	33.	0.	77.
8	32.	1.	1.	3.	10.	42.	2.	125.
9	26.	0.	1.	0.	2.	21.	1.	140
10	42.	0.	1.	4.	4.	25.	0.	328.

True Labels

Support Vector Machines

Binary	Predicted Bad Reviews	Predicted Good Reviews
True Bad Reviews	811	210
True Good Reviews	180	799

Accuracy: 0.805

Hyperparameters
C: 1.0
Gamma: 1/n Features

Support Vector Machines

Accuracy: 0.3485

Predicted Labels

Multiclass	1	2	3	4	7	8	9	10
1	290.	23.	20.	30.	14.	5.	7.	30.
2	92.	20.	24.	27.	7.	5.	5.	16.
3	76.	21.	22.	37.	13.	12.	4.	18.
4	45.	13.	26.	52.	10.	26.	10.	21.
7	20.	6.	16.	24.	26.	30.	17.	29.
8	18.	2.	11.	29.	27.	33.	36.	60
9	9.	1.	12.	21.	17.	29.	22.	80.
10	49.	9.	9.	16.	12.	35.	42.	232.

True Labels

Best performing method:

Random Forests

How can these results be improved? What's next?