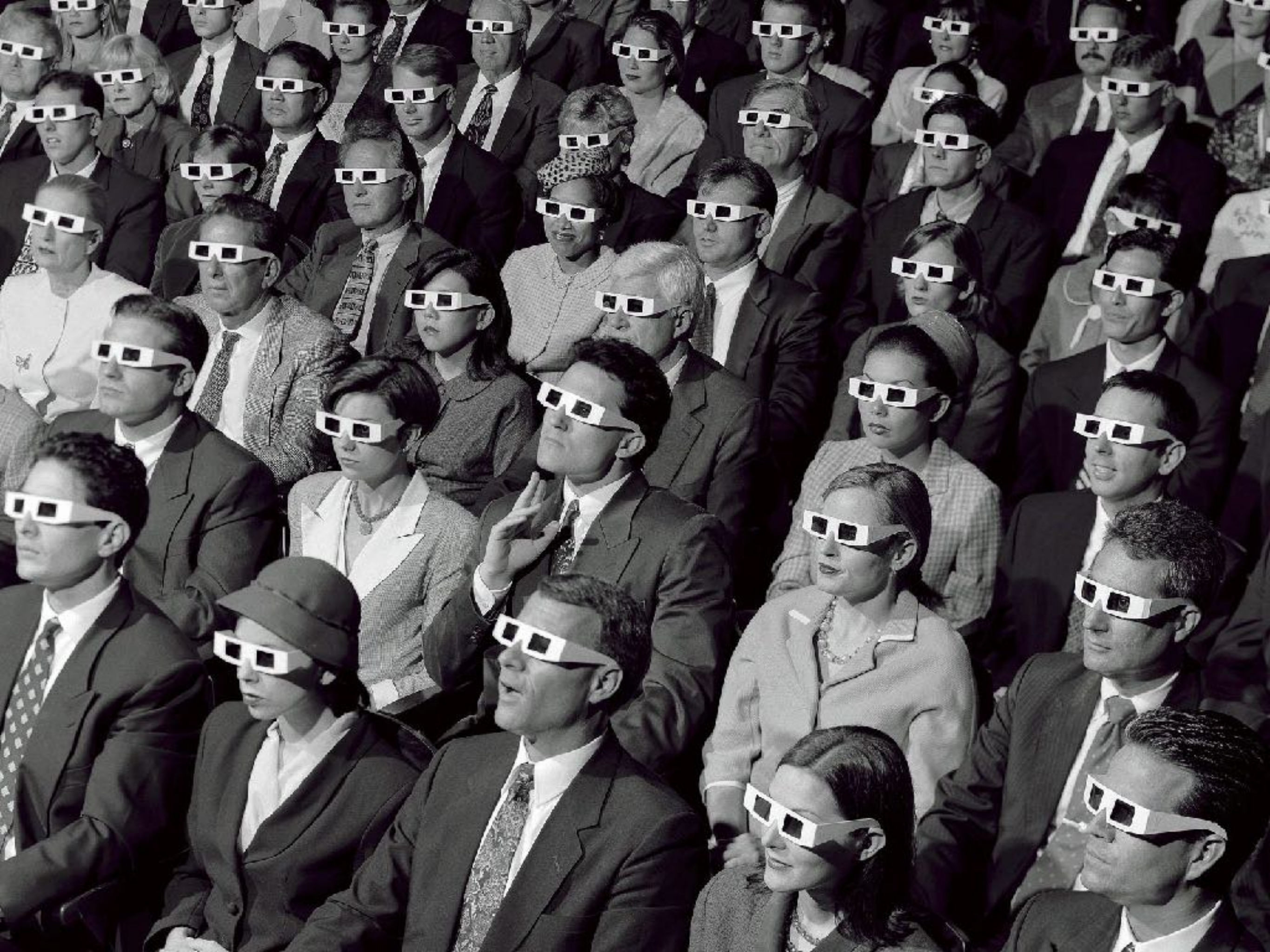


PYDATA KAUNAS #1

MOVIE AUDIENCE SCORE CALCULATION

MOTIVATION







MOVIE CRITICS



CRITICS SCORE

 32%

AUDIENCE SCORE

 79%

CRITICS V FANS

DAWN OF A FRANCHISE

CRITICS

**"ONE OF YOU
IS LYING TO ME..."**

FANS

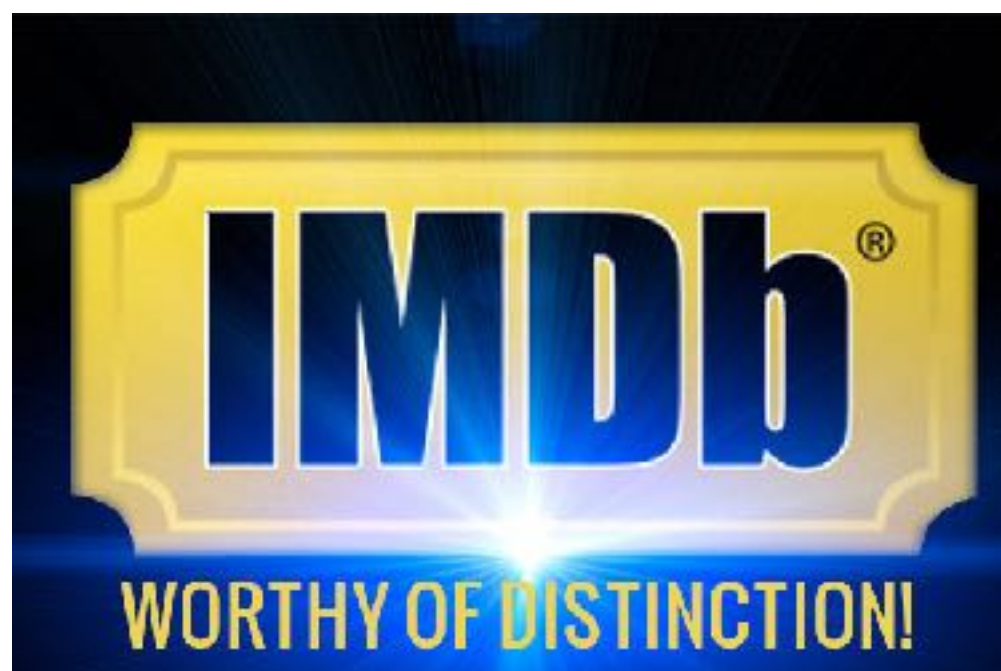
ME

VIA 9GAG.COM

ŠARŪNAS NAVICKAS

- ▶ Software Engineer @ Adform - mostly Scala and Python
- ▶ Interested in Recommendation systems and Movies - hobby project <https://mintly.eu/>
- ▶ (Newbie) Brazilian Jiu Jitsu practitioner
- ▶ Doing some Long-Distance running (if we consider 21km long)

CURRENT SOLUTIONS



**Rotten
Tomatoes™**

FOUNDED IN 1979 BY ED MITZ

CINEMAScore

No	Yes On DVD	Yes On VHS	No	I'll Wait For It To Appear On Free TV	Yes On DVD	Yes On VHS	Actor In Lead Role	Actress In Lead Role	Type Of Movie (Comedy, Horror)	Subject Matter Characters Or Plot	Director
1	2	3	1	2	3	4	5	6	7	8	9
Would you buy this movie on DVD or VHS?			Would you rent this movie on DVD or VHS? (choose 1 answer)				Reason(s) for attending this movie.				
CINEMAScore [®] TM											
AUDIENCE REACTION SURVEY											
PLEASE FOLD BACK THOSE TABS THAT APPLY, AND RETURN THIS BALLOT TO THE CINEMAScore [®] POLLSTER LOCATED OUTSIDE THIS THEATRE.											
1	2	3	4	5	1	2	1	2	3	4	5
GRADES					GENDER		YOUR AGE				
A	B	C	D	F	MALE	FEMALE	Under 18	18-24	25-34	35-49	50 & Over
OR BETTER											
OR WORSE											

PROBLEM

AUDIENCE IS NOT HOMOGENEOUS

(But these sites treat as such)



SOLUTION

SENTIMENTAL ANALYSIS ON LITHUANIAN COMMENTS

Movie	obuolys.lt	filmai.in	linkomanija	torrent.ai
A Clockwork Orange	-	22	74	50
Requiem for a Dream	-	-	339	160
The Human Centipede	-	73	141	44
The Passion of the Christ	66	66	75	30
South Park: Bigger Longer and Uncut	0	26	92	-
Zeitgeist the Movie	-	28	56	685
The Silence of the Lambs	28	93	117	3
Borat	250	57	144	14
SAW	102	81	12	15

THE PLAN

- ▶ Get initial data
- ▶ Annotate
- ▶ Transform data
- ▶ Train
- ▶ Use specific movie data on trained model

VOWPAL WABBIT

- ▶ Artificial Neural Networks (ANN)
- ▶ Naïve Bayes (Bayes)
- ▶ Support Vector Machines (SVM)

(NOT SO) HONORABLE MENTIONS

- ▶ Weka
- ▶ Apache Mahout
- ▶ Apache Spark



**MACHINE LEARNING
SPEAKS NUMBERS**

BAG OF WORDS (BOW)

	filmas	žiauriai	geras
žiauriai geras filmas	1	1	1
ziauriai geras	0	0	1
belenkoks FILMAS!	0	0	0

NORMALIZE

- ▶ `normalize(žiauriai geras filmas) = ziauriai geras filmas`
- ▶ `normalize(ziauriai geras) = ziauriai geras`
- ▶ `normalize(belenkoks FILMAS!) = belenkoks filmas`

LEVENSHTEIN DISTANCE

- ▶ `levenshtein(ziauriai, zeurei) = 2`
- ▶ `levenshtein(sveikas, sveiks) = 1`
- ▶ `levenshtein(ačiū, ačių) = 1`
- ▶ ...
- ▶

```
def equals(w1, w2):  
    return levenshtein(w1, w2) <= 2
```

STEMMING

- ▶ “**stemming** is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form”
- ▶ stem(geri vyrai gera girą gėrė) = ger vyr ger gir gėr

LEMMA

- ▶ “In morphology and lexicography, a **lemma** (plural *lemmas* or *lemmata*) is the **canonical form, dictionary form, or citation form** of a set of words”
- ▶ lemma(geri vyrai gerą girą gėrė) = geras vyras geras gira gerti
- ▶ http://donelaitis.vdu.lt/main.php?id=4&nr=7_2

NGRAM

- ▶ $N=1$: word1, word2, word3, word4...wordN
- ▶ $N=2$: (word1, word2), (word2, word3), (wordN-1, wordN)
- ▶ $N=3$: (word1, word2, word3), (word2, word3, word4), ...

**HOW TO JOIN
EVERYTHING?**

{w1, stem(w1),
lemma(w1),
synonyms(w1)}

{w2, stem(w2),
lemma(w2),
synonyms(w2)}

A brown fox
jumps over lazy
dog

1

0



Good

Bad

HOW TO MEASURE

**WHAT IS GOOD, AND WHAT
IS BAD?**

CONFUSION MATRIX

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	TP True Positive	FP False Positive
	negatives	FN False Negative	TN True Negative

COOL STUFF YOU CAN CALCULATE OUT OF IT

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\text{TPR} = \frac{\text{TP}}{P} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

specificity or true negative rate (TNR)

$$\text{TNR} = \frac{\text{TN}}{N} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

precision or positive predictive value (PPV)

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

negative predictive value (NPV)

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

miss rate or false negative rate (FNR)

$$\text{FNR} = \frac{\text{FN}}{P} = \frac{\text{FN}}{\text{FN} + \text{TP}} = 1 - \text{TPR}$$

fall-out or false positive rate (FPR)

$$\text{FPR} = \frac{\text{FP}}{N} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{TNR}$$

false discovery rate (FDR)

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}} = 1 - \text{PPV}$$

false omission rate (FOR)

$$\text{FOR} = \frac{\text{FN}}{\text{FN} + \text{TN}} = 1 - \text{NPV}$$

accuracy (ACC)

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{P + N} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Matthews correlation coefficient (MCC)

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

Informedness or Bookmaker Informedness (BM)

$$\text{BM} = \text{TPR} + \text{TNR} - 1$$

Markedness (MK)

$$\text{MK} = \text{PPV} + \text{NPV} - 1$$

Sources: Fawcett (2006), Powers (2011), and Ting (2011) [\[1\]](#) [\[2\]](#) [\[3\]](#)

THE

ARCHITECTURE

BOW

WORD_GETTER

WORD_COUNTER

NGRAM

ITER

PAR

FILTER

VOWPAL WABBIT

WORDNET

CRAWLER

STEMMER

LEVENSHTEIN

NORMALIZE

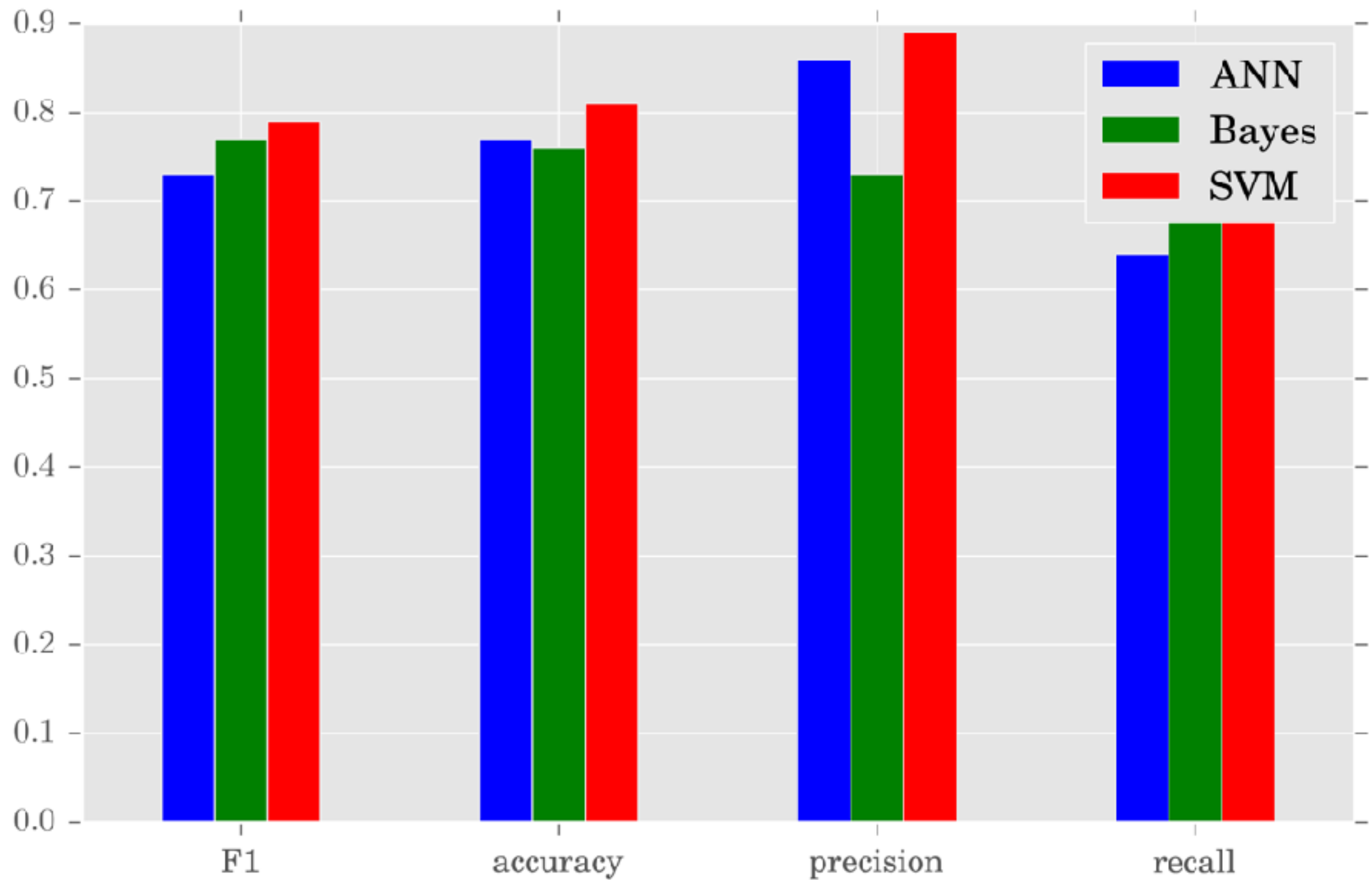
LEMMA

SCORE

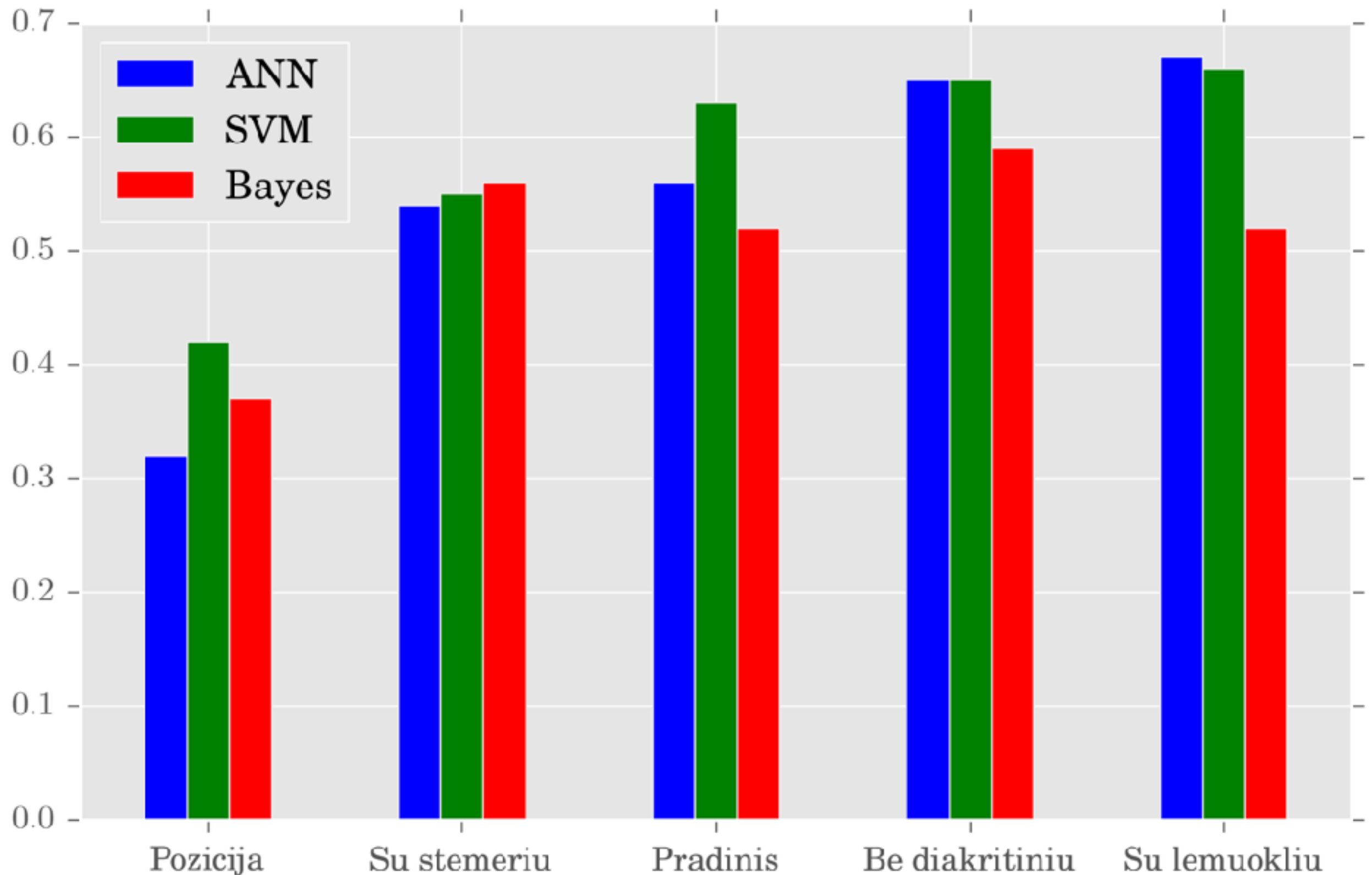
JUPYTER
NOTEBOOK

RESULTS

NO TRANSFORMATIONS



MCC SCORE COMPARING TRANSFORMATIONS





TOMATOMETER



85%

Average Rating: 7.7/10
Reviews Counted: 257
Fresh: 226
Rotten: 41

All Critics | Top Critics



Critics Consensus: Brooding and dark, but also exciting and smart, Batman Begins is a film that understands the essence of one of the definitive superheroes.

AUDIENCE SCORE



94%
liked it

Average Rating: 3.9/5
User Ratings: 1,109,496

ADD YOUR RATING



+ WANT TO SEE

NOT INTERESTED



Add a Review (Optional)

Share on Facebook

Batman Begins 2005

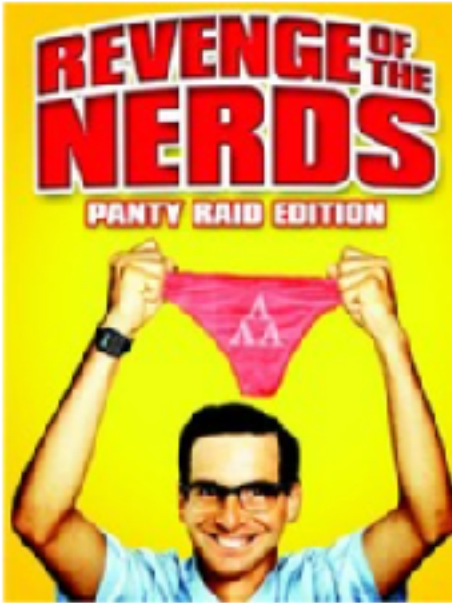


Score: 7.21


Comment Count: 137

Top Comments:

- tikrai geras filmas, tikrai netoks, kaip ankstesnėse dalyse sis yra rimtas ir su gera vaidyba ir siužetu ir t.t. ir jo tesinys taip pat yra puikus kas nemate, linkiu malonaus žiurejimo
- Tikrai geras filmas,manau bus klasika ir dar ilgai jo nepamirs :D
- filmas tikrai geras, atsimenu siunciausi ji galvojau ai ailine betmeno versija... bet kai paziurejau, buvo vov.... geras, pritraukiantis, tikrai neprailgsta laikas galeciau prilyginti ji the dark knight. super 10-10!!
- labai geras filmas ir aktorius geras
- tikrai pats geriausias matytas batmanas, apskritai man sitas filmas yra vienas is geriausiu



TOMATOMETER ?


 **69%**

Average Rating: 5.9/10
Reviews Counted: 42
Fresh: 29
Rotten: 13

All Critics | **Top Critics**


Critics Consensus: Undeniably lowbrow but surprisingly sly, Revenge of the Nerds has enough big laughs to qualify as a minor classic in the slob-vs.-snobs subgenre.

AUDIENCE SCORE ?

 **73%**
liked it

Average Rating: 3.3/5
User Ratings: 55,924

ADD YOUR RATING



[+ WANT TO SEE](#)

[NOT INTERESTED](#)

☆ ☆ ☆ ☆ ☆

Add a Review (Optional)

[Share on Facebook](#)

Revenge of the Nerds 1984



Score: 9.00

Comment Count: 26

Top Comments:

- SUPER SITAS FILMAS rekomenduoju labai geras
- Kaip jūs drįstat sakyt,kad filmas geras.Kaip jis toks nera,jeigu toki filmai patinka,tai jūs ne prie kultūros,prie tos dumos kvailių masės kur tik papai ir alus patinka.Netgi IMDB reitingas mazas.Ten tokia politika 1-5 balas (Nieko vertas filmas), 6-6,5 (popkornas),7-10 (geras filmas)
- geras filmas man taip atrodo
- senas geras filmas
- man asmeniskai tikrai jis patiko bet zinoma butu labai gerai jei jus imestumet visas likuses dalis Aciu



TOMATOMETER

61%

Average Rating: 5.8/10
Reviews Counted: 123
Fresh: 75
Rotten: 48

All Critics | Top Critics

Critics Consensus: So embarrassing it's believable, American Pie succeeds in bringing back the teen movie genre.

AUDIENCE SCORE

61%
liked it

Average Rating: 3.4/5
User Ratings: 33,780,579

ADD YOUR RATING



WANT TO SEE

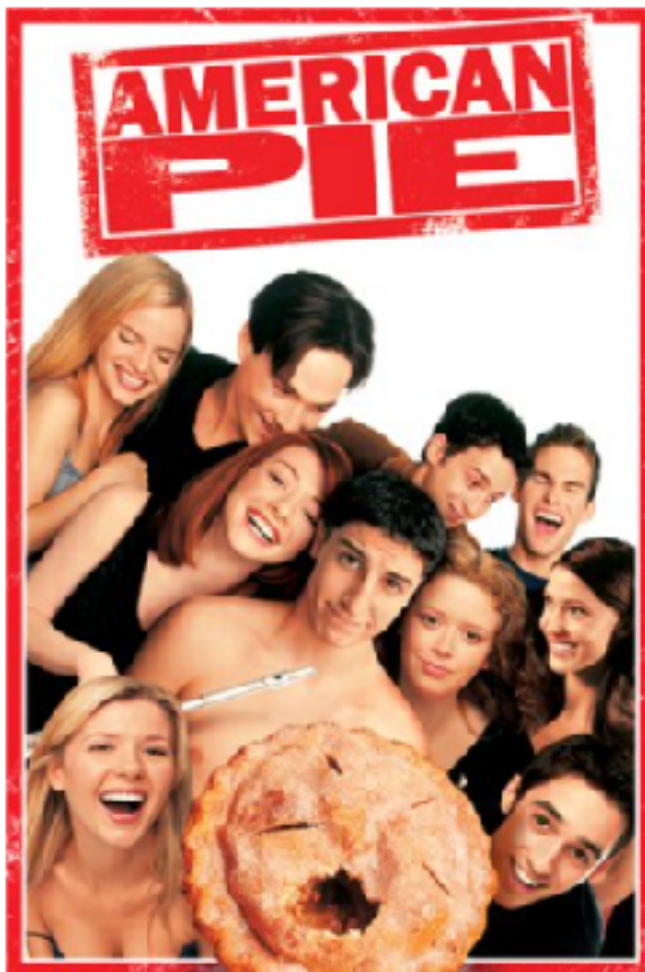
NOT INTERESTED



Add a Review (Optional)

Share on Facebook

American Pie 1999



Score: 6.17

Comment Count: 56

Top Comments:

- Labai geras filmas. labai patiko tikrai verta paziureti ir smagiai pasijuokti!
- Visu laiku Komedija ziaurei geras filmas
- truksta 7 serijos bet pack tikraj geras dekul...
- Su drauge ziurejom tai zvengem visa nakty ziurejom visas. Labai geras filmas ♥
- Tikrai linkamas ir geras filmas. Kai nori geros nuotaikos ir linksmi praleisto laiko, verta atsisiusti ir pasižiūrėti.

KLAUSIMAI?