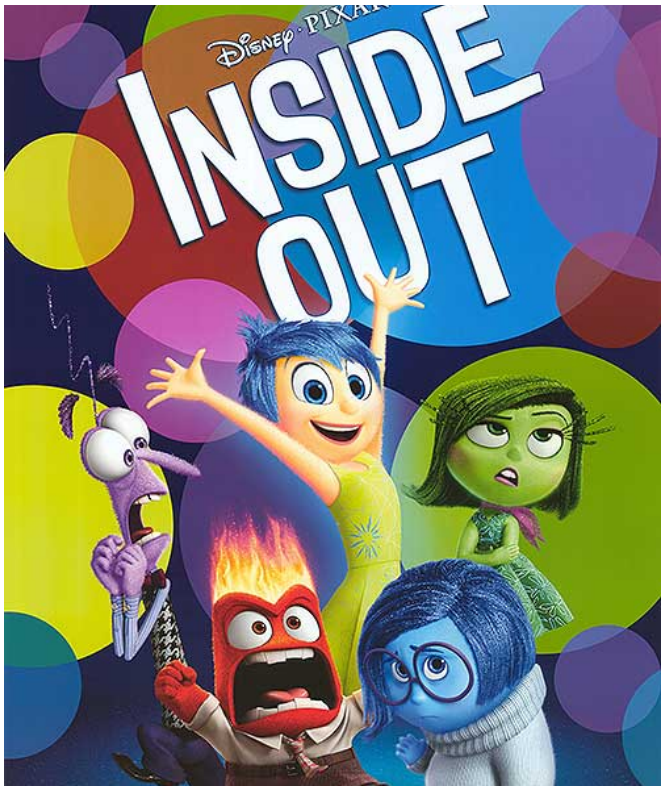
LUND UNIVERSITY

Movie Review Classifications

CONG "OLAF" CHEN

Objective



- Genres according to IMDb: animation, adventure, comedy, drama, family, fantasy. What a mouthful!
- Excerpt from sample critic review: “The level of invention is so high, and the density of detail is so great, that it’s impossible to absorb everything in a single viewing.” –Joe Morgenstern, Wall Street Journal
- Can we get back the movie’s genre(s) if the review snippet is all we have? How about the mystery reviewer’s mood?

LUND UNIVERSITY

Methods used

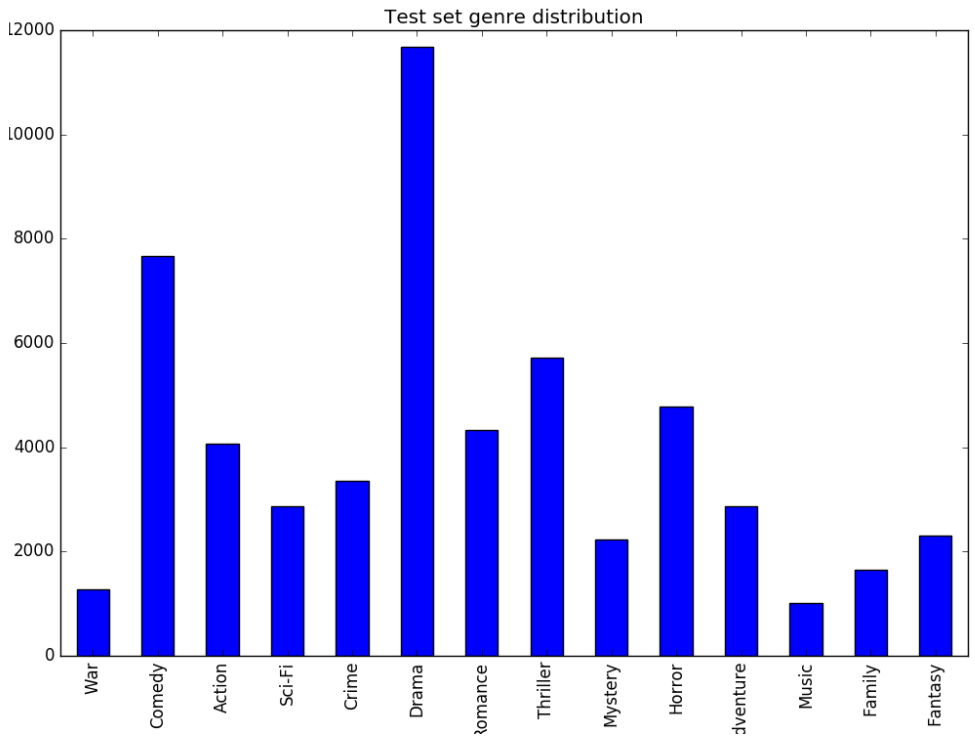
- PRAW (Reddit API)
- IMDbPy
- NLTK
- Scikit Learn

Also considered:

- Gensim's LDA model (dead end)
- Pinterest API

Dataset of genres

- ~75,000 reviews to train with plus ~25,000 to test, each contains IMDb ID of corresponding movie
- Distribution of genres in test set as follows, ignoring all genres present in fewer than 1000 reviews
- On average, a movie corresponds to 7 reviews and a review corresponds to 2.5 genre labels



LUND UNIVERSITY

Libshorttext model

- Popular and trusted package for NLP needs at LTH
- 8 ways to preprocess data: unigram/bigram, with/without stopword removal, with/without Porter stemming
- 4 classification mechanisms: standard/L1/L2 SVC, logistic regression
- 4 ways to weigh features: binary, word count, term frequency, TFIDF
- 6-7 minutes to train a corpus of 190,000 reviews

Classification Accuracy

- Accuracy obtained by comparing lib-shorttext prediction to IMDb listings (in this table, we use bigram features, stemming and stopword filtering)
- Preprocessing and feature selection do not make a major difference by themselves (+/- 1%) but the classification mechanism can, L2SVM and LogReg are better
- However, we have a problem…

	Bin.	Word Count	Term Freq.	TFIDF
SVM	.7164	.7071	.7074	.7190
L1SVM	.7324	.7362	.7362	.7409
L2SVM	.7811	.7791	.7791	.7843
LogReg	.7730	.7761	.7761	.7745

LUND UNIVERSITY

Confusion table snippets

Actual/Predicted	Romance	Horror
Romance	257	70
Horror	5	2444

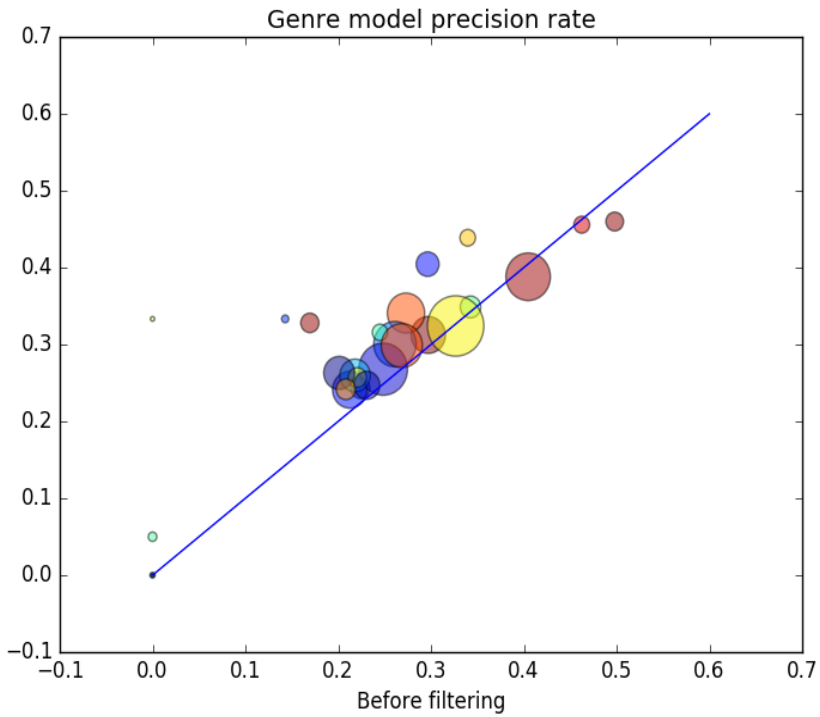
Actual/Predicted	Comedy	Sci-Fi
Comedy	4086	79
Sci-Fi	399	631

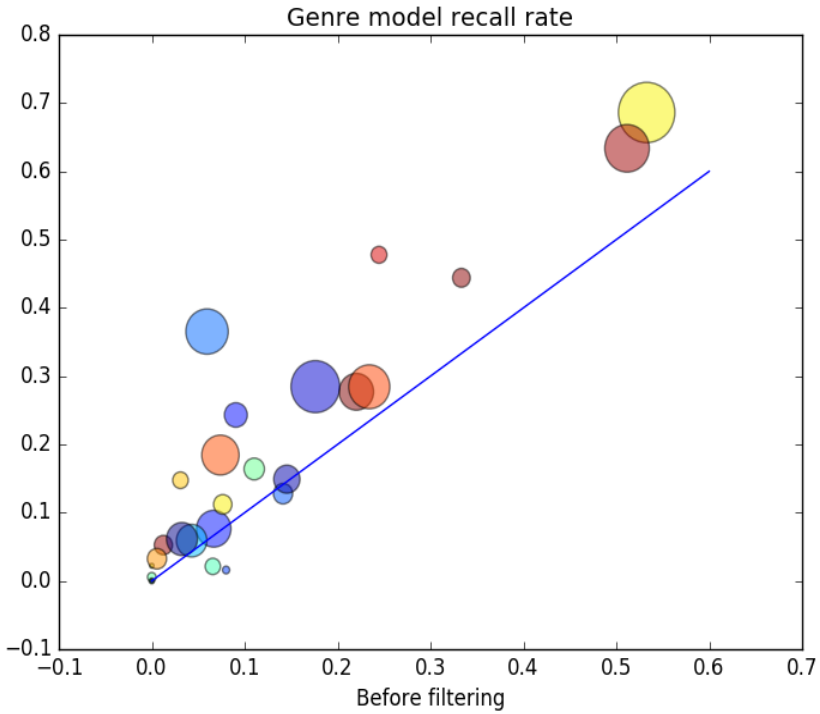
Act./Pred.	Drama	Action	Thriller
Drama	8145	330	574
Action	1180	954	438
Thriller	2192	493	1008

The “Drama” label

- 47% of all our test reviews come from a movie listed by IMDb as a “Drama”. We would expect a similar proportion in our training set. Not surprising given how all movies have to be at least a little dramatic, but this is too high a percentage given we have 27 different genre labels!
- This is not good for our model! We output as many genre labels as IMDb lists for all our reviews, but the same review outputs the same label each time.
- Frequency bias/cold start: As a result of this, 42.75% of all outputted labels are “Drama”—it’s like a security blanket for our model when it sees things it doesn’t recognize. Meanwhile, not a single instance had the predicted label “News”, “Talk Show”, “Game Show”, or “Adult”.
- We must retrain! Remove all instances with the label “Drama”. This alone should not take any individual reviews or movies out of the training or testing sets.

Precision and Recall





LUND UNIVERSITY

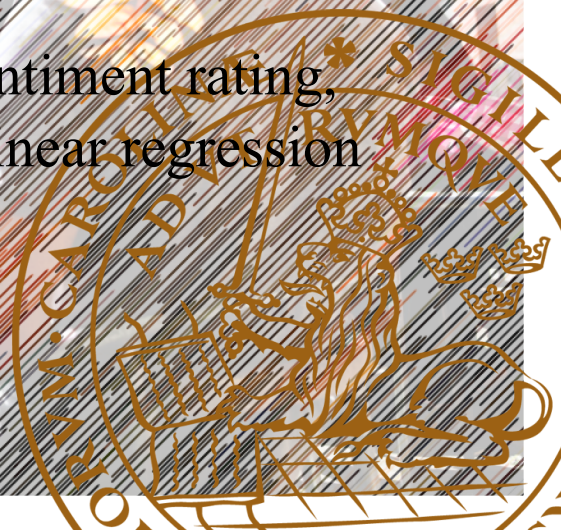
Some observations

- Every tidbit of extraneous info we cut out makes a difference! The vast majority of the points on each scatterplot lies above the blue line in both cases.
- Comedy and horror have excellent recall rates, compared to precision. They are the 2nd and 4th most popular labels in the test set. For romance (5th), this rate somehow went up from 5.95% to 36.53%! Less common labels generally have a higher precision than recall rate, because our model is less likely to guess them—but when that actually happens, it knows.
- The overall accuracy rate actually dropped from 77.59% to 65.47%, once our label could no longer simply output "Drama" knowing it had a decent chance of being correct. Keep in mind that "Drama" is not a very informative label!

LUND UNIVERSITY

Future exploration

- Our model was trained with relatively high-end data. But the same is not necessarily true for data we crawl from clients' social media accounts, such as Reddit. How does our model adapt to different "languages"?
- The next most frequent genres on the "chopping block" are comedy and thriller. Do they provide as much convoluting information about a review as "drama" does?
- If each review was given a multidimensional sentiment rating, each on a scale from 1 to 10, would a SVM or linear regression better classify a client's current mood?



Acknowledgements

- Pierre Nugues
- For critical training data, but more importantly your invaluable real-life application insight:
  - Lars Hård
  - Axel Antonsson
  - Kateryna Wikström
  - Ola Lindberg

  Best of luck in Silicon Valley!
- Andrew Maas, Stanford University (IMDb training data)
- Chih Jen Lin, National Taiwan University (libshorttext)
- Joe Morgenstern, Wall Street Journal (sample review)

LUND UNIVERSITY