The Linked Data Mining Challenge 2015

Petar Ristoski, Heiko Paulheim, Vojtěch Svátek, Václav Zeman
Challenge Overview

• Task: Predict the review class of movies
  – Binary classification problem: “good” and “bad”
Challenge Overview

• Task: Predict the review class of movies
  – Binary classification problem: “good” and “bad”

• Dataset source: metacritic.com
  – Average rating of all time reviews for a list of movies
Challenge Overview

• Task: Predict the review class of movies
  – Binary classification problem: “good” and “bad”

• Dataset source: metacritic.com
  – Average rating of all time reviews for a list of movies

• Objective: Use LOD and publicly available datasets to build predictive models for the given task
## Source Dataset

metacritic.com

<table>
<thead>
<tr>
<th>#</th>
<th>Store</th>
<th>Title</th>
<th>User</th>
<th>Watch Now</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>100</td>
<td>Best Kept Secret</td>
<td>6.1</td>
<td>Watch Now</td>
<td>Sep 6, 2013</td>
</tr>
<tr>
<td>2.</td>
<td>100</td>
<td>Boyhood</td>
<td>7.7</td>
<td>Watch Now</td>
<td>Jul 11, 2014</td>
</tr>
<tr>
<td>3.</td>
<td>98</td>
<td>Pan’s Labyrinth</td>
<td>8.5</td>
<td>Watch Now</td>
<td>Dec 29, 2006</td>
</tr>
<tr>
<td>4.</td>
<td>98</td>
<td>Carol</td>
<td>tbd</td>
<td>Watch Now</td>
<td>Dec 18, 2015</td>
</tr>
<tr>
<td>8201.</td>
<td>41</td>
<td>Drillbit Taylor</td>
<td>6.3</td>
<td>Watch Now</td>
<td>Mar 21, 2008</td>
</tr>
<tr>
<td>8202.</td>
<td>41</td>
<td>Dumb and Dumber</td>
<td>7.4</td>
<td>Watch Now</td>
<td>Dec 16, 1994</td>
</tr>
<tr>
<td>8203.</td>
<td>41</td>
<td>Stoned</td>
<td>4.5</td>
<td>Watch Now</td>
<td>Mar 24, 2006</td>
</tr>
<tr>
<td>8204.</td>
<td>41</td>
<td>Lucky</td>
<td>tbd</td>
<td>Watch Now</td>
<td>Jul 15, 2011</td>
</tr>
<tr>
<td>10201.</td>
<td>11</td>
<td>Date Movie</td>
<td>2.9</td>
<td>Watch Now</td>
<td>Feb 17, 2006</td>
</tr>
<tr>
<td>10202.</td>
<td>11</td>
<td>Pinocchio</td>
<td>2.9</td>
<td>Watch Now</td>
<td>Dec 25, 2002</td>
</tr>
<tr>
<td>10203.</td>
<td>11</td>
<td>Last Ounce of Courage</td>
<td>5.6</td>
<td>Watch Now</td>
<td>Sep 14, 2012</td>
</tr>
<tr>
<td>10204.</td>
<td>11</td>
<td>Scary Movie 5</td>
<td>2.5</td>
<td>Watch Now</td>
<td>Apr 12, 2013</td>
</tr>
</tbody>
</table>
Source Dataset

metacritic.com

Best Kept Secret
Argot Pictures | Release Date: Sep 6, 2013

Summary
Critic Reviews | User Reviews | Details & Credits | Trailers & Videos

1. 100 Best Kept Secret
2. 100 Boyhood
3. 98 Pan's Labyrinth
4. 98 Carol

8201. 41 Drillbit Taylor
8202. 41 Dumb and Dumber
8203. 41 Stoned
8204. 41 Lucky

10201. 11 Date Movie
10202. 11 Pinocchio
10203. 11 Last Ounce of Courage
10204. 11 Scary Movie 5

User Score
Generally favorable reviews based on 29 Ratings

Metascore
Universal acclaim based on 4 Critics

User: 2.9 Watch Now Feb 17, 2006
User: 2.9 Watch Now Dec 25, 2002
User: 5.6 Watch Now Sep 14, 2012
User: 2.5 Watch Now Apr 12, 2013

Summary: JFK High School, located in the midst of a rundown area in Newark, New Jersey, is a public school for all types of students with special education needs, ranging from those on the autism spectrum to those with multiple disabilities. Janet Mino has taught her class of young men with autism... Expand ▼

Director: Samantha Buck
Genre(s): Drama, Documentary
Rating: Not Rated
Runtime: 86 min
More Details and Credits »
Source Dataset

- Link each movie to DBpedia
  - Movie’s title and movie’s director
• Link each movie to DBpedia
  – Movie’s title and movie’s director

• Movies are divided into three classes
  – Good: rating > 60
  – Neutral: 60 > rating > 40
  – Bad: rating < 40

Source Dataset

6/18/2015
Source Dataset

- Link each movie to DBpedia
  - Movie’s title and movie’s director
- Movies are divided into three classes
  - Good: rating > 60
  - Bad: rating < 40

4525 good
1865 bad
Final Dataset

• Select top 1,000 and bottom 1,000 movies

• Stratified 80/20 split
  – Training dataset: 1,600 movies
    • id; Movie; Release date; DBpedia_URI; Label
  – Test dataset: 400 movies
    • id; Movie; Release date; DBpedia_URI; Label
1. Performance
   
   - $accuracy = \frac{tp+tn}{tp+fp+fn+tn}$
Evaluation

1. Performance
   - \( accuracy = \frac{tp+tn}{tp+fp+fn+tn} \)

2. Originality of the approach and techniques
   - Data preprocessing
   - Feature engineering
   - Machine learning approaches
Evaluation

1. Performance
   - $accuracy = \frac{tp+tn}{tp+fp+fn+tn}$

2. Originality of the approach and techniques
   - Data preprocessing
   - Feature engineering
   - Machine learning approaches

3. Validity of the hypotheses and findings
Allowed Dataset

• Any dataset that follows the LOD principles
  – DBpedia, LinkedMDB, DBropes, etc.
Allowed Dataset

• Any dataset that follows the LOD principles
  – DBpedia, LinkedMDB, DBropes, etc.

• Any non-LOD dataset **IF**
  – Made it accessible via standard SW technologies
  – Add corresponding LOD links
Allowed Dataset

• Any dataset that follows the LOD principles
  – DBpedia, LinkedMDB, DBropes, etc.

• Any non-LOD dataset **IF**
  – Made it accessible via standard SW technologies
  – Add corresponding LOD links

• The metacritic dataset is not allowed
Submission Procedure

- Online submission system
  - Instant results
  - Real-time leaderboard

Top Scorers

<table>
<thead>
<tr>
<th>Rank</th>
<th>User</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Topper</td>
<td>64.75%</td>
</tr>
<tr>
<td>2.</td>
<td>Meyer Bossett</td>
<td>52.20%</td>
</tr>
<tr>
<td>3.</td>
<td>Emir Munoz</td>
<td>51.75%</td>
</tr>
<tr>
<td>4.</td>
<td>Yibo Yao</td>
<td>44.25%</td>
</tr>
<tr>
<td>5.</td>
<td>idmc_baseline</td>
<td>40.25%</td>
</tr>
<tr>
<td>6.</td>
<td>ZeroR</td>
<td>19.75%</td>
</tr>
</tbody>
</table>
Baseline Models

- K-NN (k=3) classifier
  - Features: DBpedia **direct types** and **categories** (dcterms:subject)
  - Acc=60.25%

- ZeroR classifier
  - Acc=50%
Results

- 4 online submissions
  - All performed above the baseline
- 3 final submissions

<table>
<thead>
<tr>
<th>Rank</th>
<th>User</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Topper</td>
<td>94.75%</td>
</tr>
<tr>
<td>2.</td>
<td>Meyer Bossert</td>
<td>92.25%</td>
</tr>
<tr>
<td>3.</td>
<td>Emir Munoz</td>
<td>91.75%</td>
</tr>
<tr>
<td>4.</td>
<td>Yibo Yao</td>
<td>74.25%</td>
</tr>
<tr>
<td>5.</td>
<td>Idmc_baseline</td>
<td>60.25%</td>
</tr>
<tr>
<td>6.</td>
<td>ZeroR</td>
<td>49.75%</td>
</tr>
</tbody>
</table>
Meta Learner

• Degree of agreement among the three approaches
  – Fleiss’ kappa: $\kappa = 0.757$
Meta Learner

- Degree of agreement among the three approaches
  - Fleiss’ kappa: $\kappa = 0.757$

- Meta learner based on majority vote on the prediction of the three approaches
  - Acc=97%
The Linked Data Mining Challenge 2015

Petar Ristoski, Heiko Paulheim, Vojtěch Svátek, Václav Zeman