The Linked Data Mining Challenge 2015

Petar Ristoski[1], Heiko Paulheim[1], Vojtěch Svátek[2], and Václav Zeman[2]

[1] University of Mannheim, Germany
Research Group Data and Web Science
{petar.ristoski,heiko}@informatik.uni-mannheim.de
[2] University of Economics
Department of Information and Knowledge Engineering
{svatek,vaclav.zeman}@vse.cz

Abstract. The 2015 edition of the Linked Data Mining Challenge, conducted in conjunction with Know@LOD 2015, has been the third edition of this challenge. This year's dataset collected movie ratings, where the task was to classify well and badly rated movies. The solutions submitted reached an accuracy of almost 95%, which is a clear advancement over the baseline of 60%. However, there is still headroom for improvement, as the majority vote of the three best systems reaches an even higher accuracy.

1 The Linked Data Mining Challenge Overview

This year, the Linked Data Mining Challenge was held for the third time, following past editions co-located with *DMoLD* (at ECML/PKDD) [9] and Know@LOD [8]. As the past editions did not draw too much participation, we sought for feedback after the 2014 edition. That feedback included [8]

- Using a dataset from a popular domain
- Using a standard classification or regression task

Picking up on these issues, we used a dataset for movie rating prediction this year, instead of data from the public procurement and research collaboration domains, as in the past editions. Furthermore, the dataset was built as a standard two-class classification problem with balanced data for both classes.

The rest of this paper is structured as follows. Section 2 discusses the dataset construction and the task to be solved. In section 3, we discuss the entrants to the challenge and their results. We conclude with a short summary and an outlook on future work.

2 Task and Dataset

The 2015 edition of the challenge used a dataset built from movie recommendations, turned into a two-class classification problem.

2.1 Dataset

The task concerns the prediction of a review of movies, i.e.,"good" and "bad". The initial dataset is retrieved from Metacritic.com[3], which offers an average rating of all time reviews for a list of movies[4]. Each movie is linked to DBpedia using the movie's title and the movie's director. The initial dataset contained around 10,000 movies, from which we selected 1,000 movies from the top of the list, and 1,000 movies from the bottom of the list. The ratings were used to divide the movies into classes, i.e., movies with score above 60 are regarded as "good" movies, while movies with score less than 40 are regarded as "bad" movies. For each movie we provide the corresponding DBpedia URI. The mappings can be used to extract semantic features from DBpedia or other LOD repositories to be exploited in the learning approaches proposed in the challenge.

The dataset was split into training and test set using random stratified split 8020 rule, i.e., the training dataset contains 1,600 instances, and the test dataset contains 400 instances. The training dataset, which contains the target variable, was provided to the participants to train predictive models. The test dataset, from which the target label is removed, is used for evaluating the built predictive models.

2.2 Task

The task concerns the prediction of a review of movies, i.e.,"good" and "bad", as a classification task. The performance of the approaches is evaluated with respect to accuracy, calcuated as:

$$Accuracy = \frac{\#true\ positives + \#true\ negatives}{\#true\ positives + \#false\ positives + \#false\ negatives + \#true\ negatives} \quad (1)$$

2.3 Submission

The participants were asked to submit the predicted labels for the instances in the test dataset. The submission were performed through an online submission system. The users could upload their prediction and get the results instantly. Furthermore, the results of all participants were made completely transparent by publishing them on an online real-time leader board (Figure 1). The number of submissions per user was not constrained.

In order to advance the increase of Linked Open Data [7] available as a side-effect of the challenge, we allowed users to also exploit non-LOD data sources, given that they transform the datasets they use to RDF, and provide them publicly.

[3] http://www.metacritic.com/
[4] http://www.metacritic.com/browse/movies/score/metascore/all

Rank	User	Accuracy
1.	Topper	94.75%
2.	Meyer Bossert	92.25%
3.	Emir Munoz	91.75%
4.	Yibo Yao	74.25%
5.	ldmc_baseline	60.25%
6.	ZeroR	49.75%

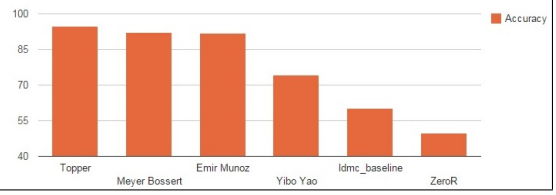

Fig. 1: Participants Results

3 The Linked Data Mining Challenge results

In total, four parties participated in the challenge, three of which finally submitted results and a paper. We compare those results against two baselines.

3.1 Baseline Models

We provide a simple classification model that will serve as a baseline. The model is implemented in the RapidMiner platform, using the Linked Open Data extension [4, 5]. In this process we use the movies DBpedia URI to extract the direct types and categories of each movie. On the resulting dataset we built k-NN classifier (k=3), and applied it on the test set, scoring an accuracy of 60.25%.

In addition, we built the trivial model ZeroR, which simply predicts the majority class. The model achieved an accuracy of 49.75%.

3.2 Participants' Approaches

During the submission period, four approaches participated in the challenge. Finally, three teams completed the challenge, by submitting a solution to the online evaluation system, and describing the used approach in a paper. In the following, we describe and compare the final participant approaches.

Topper. Utilizing the Open Movie Data Base for Predicting the Review Class of Movies [6]
By Johann Schaible, Zeljko Carevic, Oliver Hopt, and Benjamin Zapilko (GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany)

In this approach, the authors use features extracted from the Open Movie Database[5] (OMDB) to build several predictive models, and compare their results. The OMDB database contains many information about the quality of the movie. The authors extract the following features: number of awards, number of nominations, IMDB movie ratings, IMDB number of votes, Rotten Tomatoes Tomatometer, Tomatorating, Tomato User Meter, Tomato User Rating and Tomato number of reviews. Furthermore, aggregation features are used,

[5] http://www.omdbapi.com/

tomatoFreshRatio that is calculated as the quotient of the number of "fresh" Tomato ratings and the Tomatometer, and *tomatoRottenRatio* that is the quotient of the number of "rotten" Tomato ratings and the Tomatometer. The data is converted into RDF, resulting into 36,020 RDF triples.[6]

To build the predictive models, the authors use RapidMiner including the RapidMiner Linked Open Data Extension. Moreover, they build three classifiers, i.e., Naïve Bayes, K-NN, and Decision Trees. To compare the performances of the classifiers, the authors first perform 10-fold cross validation on the training dataset, using different combination of features. The best results are achieved when using the Decision Trees classifier with all features, scoring an accuracy of 97% on the training data. In comparison, the Naïve Bayes classifier scored 95%, and the k-NN classifier only 51% accuracy on the training data. The Decision Trees classifier scored an accuracy of 94.75% on the test dataset, taking the first place in the challenge. The decrease of 2% on the test set is explained by the authors as an overfitting problem. The authors state that the reasons for the bad performances of the k-NN classifier might be that they did not normalize the feature, thus the distance measure might have been dominated by features with large scales.

Furthermore, the authors provide some insights on the relevance of the features for the classification task. For example, the authors observe that user generated critics, such as the IMDB score and the Tomato user rating/meter, provide a 10 to 20 percent lower prediction accuracy than the "official" critics like the Tomatometer. Also, the number of winning or nominated awards provides a decent prediction accuracy as well.

Meyer Bossert. Predicting Metacritic Film Reviews Using Linked Open Data and Semantic Technologies [2]
By Meyer Bossert (Cray Inc., Seattle Washington, USA)

In his approach, the author solves the task of classification by only using SPARQL. To start with, using the Cray Urika GD[7] graph appliance, the author loads the complete DBpedia and Freebase datasets as well as the challenge training and test dataset into a single graph. Next, all irrelevant predicates for the task are removed from the graph. To implement the classification task, a similar approach as the Naïve Bayes method is used, i.e., the author tries to find for each attribute associated with a movie, on average how many times that attribute is associated with a "good" or "bad" movie. Then, the value can be used to surmise with some degree of certainty that the score as determined by taking the average of all attributes will be a good indicator of the likelihood of a film receiving positive or negative reviews. Furthermore, the author makes an assumption that some specific properties, like awards, should get higher weights than the rest of the properties. The code and the data can be found online[8]. This approach scored an accuracy of 92.25%, taking the second place in the challenge.

[6] http://lod.gesis.org/gmovies/
[7] http://www.cray.com/products/analytics/urika-gd
[8] https://github.com/mabossert/LDMC_2015

Furthermore, the author provides some interesting observations about the task. For example, films featured at a film festival are disproportionately well reviewed by critics, however, the experiments showed that there was little correlation between film festivals and good critical reviews despite the fact that the average percentage of good vs. bad films that had properties associated with film festivals was 80.34% for the training dataset. Next, regardless of the film, those that were identified as documentaries received overwhelmingly high praise from critics. Finally, the author observes that it is slightly easier to predict the review class of good movies than bad ones. The hypothesis is that the reason for the imbalance is that good movies tend to have a wide variety of information entered into DBpedia and Freebase while bad movies tend to have less effort put into their documentation.

Emir Munoz. A Linked Data Based Decision Tree Classifier to Review Movies [1]
By Suad Aldarra and Emir Muñoz (Insight Centre for Data Analytics, National University of Ireland, Galway)

In this approach, the authors use several sources to extract useful features to build a decision trees algorithm to predict the class of the movies. The features used for building the predictive model are extracted from multiple Linked Open Data sources, as well as semi-structured information from HTML pages. The features were extracted from five different sources: DBpedia, Freebase, IMDB, OMDB and Metacritic. First, DBpedia is used to extract the categories (i.e. *dcterms:subject*) of the movies, and explore the *owl:sameAs* links to Freebase. From Freebase, personal information about actors and directors are retrieved, such as, genre, nationality, date of birth, IMDB ID, among others. The IMDB ID is used as a link to retrieve features from IMDB: actors, directors and movies awards, movies budget, gross, common languages, countries, and IMDB keywords. The authors use the OMDB API to query for further movies data, including MPAA ratings. Finally, for each movie the authors collected textual critics' reviews from Metacritic website and applied an existing API for sentiment analysis using NLTK[9], which returns either a positive, negative or neutral sentiment label for a given text. In order to reduce the feature space, the authors applied feature aggregation over actors, directors, and critics' reviews. The collected data is transformed into RDF, resulting in 338,140 RDF triples[10].

The authors use the previously extracted features to build a C4.5 decision trees, using the Weka[11] J48 implementation. This approach scored an accuracy of 91.75%, taking the third place in the challenge.

Furthermore, the authors provide a solid analysis of the relevance of the features for the classification task. The sentiment analysis over critics' reviews generate the attributes with higher information gain. Negative critics have an

[9] http://text-processing.com/docs/sentiment.html
[10] https://github.com/emir-munoz/ldmc2015
[11] http://www.cs.waikato.ac.nz/ml/weka/

Table 1: Comparison of the participants approaches.

Approach	Classification methods	Knowledge Source	#Triples	Tools	Score	Rank
Topper	Naïve Bayes, K-NN (k=3), Decision Trees	Open Movie Database	36,020	RapidMiner, LOD extension	94.75%	1
Meyer Bossert	Custom SPARQL	DBpedia and Freebase	3.45 billion	Cray Urika GD	92.25%	2
Emir Munoz	C4.5 Decision Trees	DBpedia, Freebase, IMDB, OMDB and Metacritic	338,140	Weka	91.75%	3

information gain of 0.71886 bits, thus, selected as root of the decision tree. Experiments removing all sentiment features from the training show that accuracy is reduced by ca. 9%. While removing positive or negative does not affect the accuracy severely. That shows the relevance of sentiment analysis-based features for this task, which are directly related to the taste of users.

Movie keywords are the next features with higher information gain, and their analysis provide interesting insights to be considered by writers and directors: (i) bad movies are based on video games, with someone critically bashed, using a taser, pepper spray, or hanged upside down, with dark heroine involved; and (ii) good movies include family relationships, frustration, crying, melancholy, very little dialogue, and some sins with moral ambiguity.

3.3 Meta Learner

We made a few more experiments in order to analyze the agreement of the three submissions, as well as the headroom for improvement.

For the agreement of the three submissions, we computed the Fleiss' kappa score [3], which is 0.757. This means that there is a good, although not perfect agreement of the three approaches about what makes good and bad movies.

To exploit advantages of the three approaches, and mitigate the disadvantages, we analyzed how a majority vote of the three submissions would perform. The accuracy totals at 97%, which is still higher than the best solution submitted. This shows that there is still headroom for improvement by combining the different approaches pursued by the challenge participants.

4 Conclusion

In this paper, we have discussed the task, dataset, and results of the Linked Data Mining Challenge 2015. The submissions show that Linked Open Data

is a useful source of information for data mining, and that it can help to build good predictive models. On the other hand, the experiment with majority voting shows that there is still some headroom for improvement.

One problem to address in future editions is the presence of false predictors. The dataset at hand, originating from MetaCritic, averages several ratings on movies into a final score. Some of the LOD datasets used by the competitors contained a few of those original ratings, which means that they implicitly used parts of the ground truth in their predictive models (which, to a certain extent, explains the high accuracy values). Since all of the participants had access to that information, a fair comparison of approaches is still possible; but in a real-life setting, the predictive model would perform sub-optimally, e.g., when trying to forecast the rating of a *new* movie.

In summary, this year's edition of the Linked Data Mining challenge showed some interesting cutting-edge approaches for using Linked Open Data in data mining. As the dataset is publicly available, it can be used for benchmarking future approaches as well.

Acknowledgements

We thank all participants for their interest in the challenge and their submissions. The preparation of the Linked Data Mining Challenge and of this paper has been partially supported by the by the German Research Foundation (DFG) under grant number PA 2373/1-1 (Mine@LOD), and by long-term institutional support of research activities by the Faculty of Informatics and Statistics, University of Economics, Prague.

References

1. Suad Aldarra and Emir Muñoz. A linked data-based decision tree classifier to review movies. In *4th International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data*, 2015.
2. Meyer Bossert. Predicting metacritic film reviews using linked open data and semantic technologies. In *4th International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data*, 2015.
3. Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 1973.
4. Heiko Paulheim, Petar Ristoski, Evgeny Mitichkin, and Christian Bizer. Data mining with background knowledge from the web. In *RapidMiner World*, 2014.
5. Petar Ristoski, Christian Bizer, and Heiko Paulheim. Mining the web of linked data with rapidminer. In *Semantic Web challenge at ISWC*. 2014.
6. Johann Schaible, Zeljko Carevic, Oliver Hopt, and Benjamin Zapilko. Utilizing the open movie database api for predicting the review class of movies. In *4th International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data*, 2015.
7. Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In *The Semantic Web–ISWC 2014*, pages 245–260. Springer, 2014.

8. Vojtěch Svátek, Jindřich Mynarz, and Heiko Paulheim. The linked data mining challenge 2014: Results and experiences. In *3rd International Workshop on Knowledge Discovery and Data Mining meets Linked Open Data*, 2014.
9. Vojtěch Svátek, Jindřich Mynarz, and Petr Berka. Linked Data Mining Challenge (LDMC) 2013 Summary. In *International Workshop on Data Mining on Linked Data (DMoLD 2013)*, 2013.