**Proceedings Paper:**

Lommatzsch, A., Kille, B., Hopfgartner, F. orcid.org/0000-0003-0380-6088 et al. (1 more author) (2018) NewsREEL Multimedia at MediaEval 2018: News Recommendation with Image and Text Content. In: Working Notes Proceedings of the MediaEval 2018 Workshop. MediaEval 2018, 29-31 Oct 2018, Cagnes-sur-Mer, France. CEUR-WS , Cagnes-sur-Mer, France .

Lommatzsch, A. et al : NewsREEL Multimedia at MediaEval 2018: News Recommendation with Image and Text Content. Working Notes Proceedings of the MediaEval 2018 Workshop, Cagnes-sur-Mer, France, October 29-31, 2018, CEUR-WS.org, online http://ceur-ws.org/Vol-2283/MediaEval_18_paper_5.pdf.

# NewsREEL Multimedia at MediaEval 2018: News Recommendation with Image and Text Content

Andreas Lommatzsch, Benjamin Kille
TU Berlin, Berlin, Germany
{firstname.lastname}@dai-labor.de

Frank Hopfgartner
University of Sheffield
Sheffield, UK
f.hopfgartner@sheffield.ac.uk

Leif Ramming
plista GmbH
Berlin, Germany
leif.ramming@plista.com

## ABSTRACT

NewsREEL Multimedia premiers 2018 as part of the MediaEval Benchmarking Initiative. The NewsREEL task combines recommendation algorithms with image and text analysis. Participants must predict engagement with news items based on text snippets and annotated images. Several major German news portals have supplied data. The algorithms are evaluated in terms of precision on unknown data. This paper describes the task and the provided data in detail and explains the applied evaluation approach. The algorithms are evaluated based on *Precision* and *Average-Precision* for the top news items.

## KEYWORDS

Multimedia, News, Recommender Systems

## 1 INTRODUCTION

Recommender systems help users to find the most interesting items in huge sets of available items [7, 12]. Traditionally, recommender systems focus on Collaborative Filtering (CF), which makes use of users sharing similar tastes [9]. CF-based approaches rely on users being trackable with a user ID and on the possibility to collect enough user feedback or interaction data. If the majority of users are anonymous and if items have short lifecycles and receive few interactions, the resulting "cold start" issue impedes Collaborative Filtering. A majority of anonymous users and items with short lifecycles induce a permanent cold-start setting [5]. As a result, publishers struggle to apply collaborative filtering in their news recommender systems. Content-based recommendation approaches offer an alternative way to address the problem [11]. Usually, news articles come in the form of text accompanied by an image. Both affect readers' perception. NewsREEL Multimedia tasks participants to predict items' popularity based on text snippets and image features.

The remaining paper is structured as follows: Section 2 describes the NewsREEL task in detail. Section 3 introduces the provided dataset. The evaluation procedure is discussed in Section 4. Section 5 describes the experiment. Section 6 discusses the baseline evaluation results. Finally, Section 7 concludes the paper.

## 2 TASK DESCRIPTION

NewsREEL Multimedia tasks participants to predict news items' popularity from texts and images. The task dataset comprises the news items collected by several German publishers over the course of thirteen weeks. The task focuses on non-personalized recommendation. It measures popularity by counting how many readers engage with items. In other words, participants must compute which articles receive the most impressions. The data include textual features, such as headline and text snippet, visual features extracted from images, and interaction features, derived from web server logs, such as the number of impressions. Participants receive all item-related data for the entire thirteen weeks. The training set covers item access data of the weeks 0–2 and 6–8. Participants must predict items' popularity for weeks 4, 10, 11, and 12 (evaluation set). The popularity data for the weeks 3, 5, and 9 have been excluded to prevent extrapolation of time series. Instead, participants should focus on image and text content. Item IDs and features are available for all weeks. Participants must predict the most popular items for the evaluation weeks as well as the number of impressions for the most popular news items.

## 3 DATA DESCRIPTION

The dataset covers thirteen weeks of four selected publishers, who publish predominantly German articles. We encounter 51 289 images displayed alongside articles during this period. The images distribute unequally with one publisher accounting for 42 003 images. In addition, we provide a total of 1 691 unique labels automatically assigned to images by seven annotators trained on *ImageNet* [4]. The dataset amounts to approximately 8.6 GB in size. We observe a total of about 153 million impressions, 397 million recommendations, and 1.1 million clicks.

The dataset includes the following features for each item:
**1)** item data (ID, item URL, item image URL, timestamp)
**2)** textual features (headline, text snippet; in German)
**3)** image features (up to ten labels per image and a weighting, activation weights of a standard deep learning network encoding the image). The images have been annotated by means of different frameworks (*Keras* [2], *TensorFlow* [1] and existing, pre-trained models (VGG16, VGG19 [13]).
**4)** Items' popularity data (number of views, number of clicks, number of times recommended). The image popularity data are provided only for the training weeks.

In addition to the provided features, participants may compute additional features or integrate data from external sources. Corsini and Larson [3] discuss how to apply image feature extraction for news recommendation. Kille et al. [8] and Gulla et al. [6] describe additional datasets for news recommendation.

The entire dataset has been collected by plista GmbH. Access to the data is subject to a usage agreement with their providers.

*MediaEval '18, 29-31 October 2018, Sophia Antipolis, France*

## 4 EVALUATION AND GROUND TRUTH

News recommender systems determine the most relevant articles. For the NewsREEL challenge, we have computed the number of impressions for the items published on selected news portals. We split the data into a training and a test set. For the test set we have removed the number of impressions. Task participants must predict the unknown number of impressions for items in selected weeks. We consider the number of impressions as a proxy for relevance. The quality of the predictions is computed by comparing the predictions with the true number of impressions (observed in the test weeks). In the evaluation we consider different metrics.

**Precision** measures how precisely participants identify relevant items. We consider two cut-off points. First, we compute the **Precision@n** to check whether participants manage to identify the most popular items. In the challenge, we analyze $n = 10$ and $n = 0.1*$number of items in the bin. Second, we compute the **Average Precision@n** (AP). We define the AP as the mean of the top $n$ precision scores: AP $= \sum_{n=1}^{M} \frac{1}{M} *$ Precision@$n$, where $M$ describes the number of elements in the test set. For computing Precision@$n$ we assume the top $n$ items to be the target. In other words, task participants succeed if they manage to identify the most relevant items. We compute the precision metrics for each publisher separately.

Baseline strategies and the observed evaluation results are discussed in the subsequent section. Baseline strategies and their evaluation results are discussed in [10].

## 5 RUN DESCRIPTION

The data cover thirteen weeks indexed from 0 to 12. Participants receive the content-related features to all items read in this period. The interaction-related features, such as impressions and clicks, remained unavailable for the weeks 3 to 5 as well as 9 to 12. Participants must create a predictor using the data from weeks 0 to 2 and estimate the number of impressions for items in week 4. Likewise, they ought to use the data from weeks 6 to 8 to predict impressions in weeks 10 to 12. We obtain a prediction for each combination of item and week in the specified periods. For each of the weeks we compute 4 metrics: Precision@10, Precision@Top10%, and AP@Top10%. We average those measurements over the weeks to determine the overall score of the submission.

## 6 EVALUATION

We have implemented three baseline algorithms:
**(1) The random recommender** shuffles the itemIDs randomly and assigns each item the average number of impressions for an item at that rank as the prediction.
**(2) The text similarity-based recommender** computes the similarity of each item in the test set with all items in the training set. The similarity computation is done using the cosine similarity of the term vectors provided for the news items. Subsequently, the similarity-weighted average of the number of impressions for all items (similar to the current news item) is computed and used as prediction.
**(3) The image label-based recommender** computes similarity of news items based on the overlap of images labels (having a probability bigger than 0.3). Given an item of the test set, the algorithm estimates the number of impressions by the a similarity-weighted average of the number of impression (similar to the current item in the training set).

**Table 1: Baseline Evaluation Results (portal 13554).**

| | Random | | | Text-based | | | Image-based | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | P@ 10 | P@ 10% | AP@ 10% | P@ 10 | P@ 10% | AP@ 10% | P@ 10 | P@ 10% | AP@ 10% |
| 04 | 0.00 | 0.11 | 0.07 | 0.40 | 0.19 | 0.23 | 0.40 | 0.21 | 0.18 |
| 10 | 0.00 | 0.10 | 0.06 | 0.40 | 0.19 | 0.22 | 0.30 | 0.24 | 0.18 |
| 11 | 0.00 | 0.09 | 0.04 | 0.40 | 0.18 | 0.21 | 0.30 | 0.22 | 0.16 |
| 12 | 0.00 | 0.10 | 0.04 | 0.40 | 0.19 | 0.21 | 0.30 | 0.22 | 0.17 |
| avg. | 0.00 | 0.10 | 0.05 | 0.40 | 0.18 | 0.21 | 0.33 | 0.22 | 0.17 |

**Table 2: Baseline Evaluation Results (portal 39234).**

| | Random | | | Text-based | | | Image-based | | |
|---|---|---|---|---|---|---|---|---|---|
| Week | P@ 10 | P@ 10% | AP@ 10% | P@ 10 | P@ 10% | AP@ 10% | P@ 10 | P@ 10% | AP@ 10% |
| 04 | 0.00 | 0.08 | 0.04 | 0.30 | 0.13 | 0.09 | 0.10 | 0.13 | 0.09 |
| 10 | 0.00 | 0.12 | 0.07 | 0.30 | 0.14 | 0.09 | 0.10 | 0.08 | 0.05 |
| 11 | 0.00 | 0.10 | 0.04 | 0.30 | 0.12 | 0.05 | 0.00 | 0.10 | 0.06 |
| 12 | 0.00 | 0.11 | 0.06 | 0.40 | 0.11 | 0.06 | 0.00 | 0.10 | 0.04 |
| avg. | 0.00 | 0.10 | 0.05 | 0.33 | 0.13 | 0.07 | 0.05 | 0.10 | 0.06 |

The evaluation results for two publishers are listed in Table 1 and 2. The random recommender reaches a very low precision. The image label-based recommender shows a slightly better precision. The text-based recommender outperforms the image label-based recommender. This indicates that text provides more information that image label when computing the popularity in news items. Moreover, we observe big differences between the portals. This indicates that the importance of images depends on the specific news portal. In addition, the variance in the weeks indicates that the dataset is sparse and differences in the user behavior in the considered weeks exist.

## 7 CONCLUSION

NewsREEL Multimedia is a challenging task combining recommendation with text and image analysis. The challenge provides a real-world data set collected by several major German news portals. The evaluation centers on anticipating the most popular articles by their contents. Details on the developed methods and the obtained results are reported in the workshop working notes of the MediaEval workshop.

## REFERENCES


[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association.

[2] F. Chollet et al. Keras. https://keras.io, 2015.

[3] F. Corsini and M. Larson. CLEF NewsREEL 2016: Image-based Recommendation. CLEF 2016: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum, 2016.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[5] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 169–176. ACM, 2014.

[6] J. A. Gulla, L. Zhang, P. Liu, Ö. Özgöbek, and X. Su. The adressa dataset for news recommendation. In *Proceedings of the International Conference on Web Intelligence*, pages 1042–1048. ACM, 2017.

[7] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[8] B. Kille, F. Hopfgartner, T. Brodt, and T. Heintz. The plista dataset. In *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*, pages 16–23. ACM, 2013.

[9] Y. Koren and R. Bell. Advances in collaborative filtering. In *Recommender systems handbook*, pages 77–118. Springer, 2015.

[10] Lommatzsch, Andreas and Kille, Benjamin. Baseline Algorithms for Predicting the Interest in News based on Multimedia-Data. In *Proceedings of the MediaEval Workshop*, 2018.

[11] P. Lops, M. De Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer, 2011.

[12] F. Ricci, L. Rokach, and B. Shapira. Recommender systems: introduction and challenges. In *Recommender systems handbook*, pages 1–34. Springer, 2015.

[13] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations, 2015.