CLEF 2017 NewsREEL Overview: A Stream-based Recommender Task for Evaluation and Education

Andreas Lommatzsch[1(✉)], Benjamin Kille[1], Frank Hopfgartner[2], Martha Larson[3], Torben Brodt[4], Jonas Seiler[4], and Özlem Özgöbek[5]

[1] TU Berlin, Berlin, Germany
{benjamin.kille,andreas.lommatzsch}@dai-labor.de
[2] University of Glasgow, Glasgow, UK
frank.hopfgartner@glasgow.ac.uk
[3] Radboud University and TU Delft, Netherlands
m.a.larson@tudelft.nl
[4] Plista GmbH, Berlin, Germany
{torben.brodt,jonas.seiler}@plista.com
[5] NTNU, Trondheim, Norway
{ozlem.ozgobek}@ntnu.no

Abstract. News recommender systems provide users with access to news stories that they find interesting and relevant. As other online, stream-based recommender systems, they face particular challenges, including limited information on users' preferences and also rapidly fluctuating item collections. In addition, technical aspects, such as response time and scalability, must be considered. Both algorithmic and technical considerations shape working requirements for real-world recommender systems in businesses. NewsREEL represents a unique opportunity to evaluate recommendation algorithms and for students to experience realistic conditions and to enlarge their skill sets. The NewsREEL Challenge requires participants to conduct data-driven experiments in NewsREEL Replay as well as deploy their best models into NewsREEL Live's 'living lab'. This paper presents NewsREEL 2017 and also provides insights into the effectiveness of NewsREEL to support the goals of instructors teaching recommender systems to students. We discuss the experiences of NewsREEL participants as well as those of instructors teaching recommender systems to students, and in this way, we showcase NewsREEL's ability to support the education of future data scientists.

Keywords: Recommender systems · News · Evaluation · Living lab · Stream-based recommender

1 Introduction

Many recommender systems operate in large-scale, highly dynamic environments. Thousands of users must simultaneously receive suggestions fitting their individual preferences. In order to serve these users, system designs have to fulfill multiple requirements. These requirements include accurate predictions, reliability, responsiveness, and maintainability among others. Unfortunately, a majority of university curricula falls short of providing students with the necessary background to design systems that address these

requirements. As a result, students have to pick up the skills necessary to design, maintain, and optimize recommender systems outside of the classroom, for example, on the job. Students complete their degrees with a lopsided profile. Some students may have proficient skills to accurately compute relevance estimates, but struggle to implement scalable systems. Conversely, some students may be excellent programmers, but lack the knowledge of statistics needed to develop better models.

Providing students with the comprehensive background needed to develop and deploy recommender systems is a challenging problem. Universities may be aware of the issue, but are still unable to address it effectively. For example, when teaching courses that focus on Recommender Systems, the lack of resources to conduct multi-criteria evaluation represents a major reason for universities to focus on simplified aspects only. The NewsREEL challenge is well-suited to allow educators to move beyond current restrictions. With NewsREEL 2017, we provide the required resources such that students can experience authentic conditions. They can deepen their skills in many ways and prepare for a career as system engineer, data scientist, or business analyst. All these roles command an understanding for a variety of aspects related to the performance of intelligent systems.

The world of news offers a use case that is widely familiar. Nearly everyone is a consumer of news in some form. Societies in the information age demand a continuous influx of information pieces, steadily provided by busy journalists. News recommender systems filter the information flow for news readers. They automatically select a subset of articles in pursuit of the goal of high user engagement with the recommendations.

Participants in the CLEF NewsREEL (News REcommendation Evaluation Lab) challenge face a complex environment. They have to define a strategy to produce accurate recommendations. At the same time, they have to deploy the strategy onto a server accessible to recommendation requests. They have to maintain the server and assure reliability even at times with plenty of simultaneous requests. In addition, the rapidly changing sets of users and items demand the recommendation algorithms to respond quickly to updates. This environment forces participants to develop systems that perform well with respect to real-world multi-criteria requirements.

In 2017, NewsREEL features a set of changes compared to the former editions. We have released a renewed data set, which covers more recent events from February 2016. We continue our cooperation with plista, a company offering personalization and targeted advertising services. Plista has revised the Open Recommendation Platform (ORP) in 2016. The revised platform operates with a RESTful API which facilitates automating administrative processes such as starting and stopping recommendation services. Finally, we have published a new evaluator for the offline evaluation along with a tutorial. Completing the tutorial took some time, which meant that it was not available until some time after the start of NewsREEL in October 2016.

The purpose of this paper is to introduce the NewsREEL 2017 challenge, and to discuss its ability to foster education and provide students with skills necessary to succeed in industry. We give a general overview of how NewsREEL provides an opportunity for multi-dimensional benchmarking in stream-based scenarios. We then turn to the discussion of the potential of NewsREEL in higher education. We report the results of a survey of past participants that gives us insight into how the challenge has contributed to the

development of the participants' skills. We also discuss the contribution of NewsREEL to education from an instructor's point of view.

The remainder of this paper conveys the following parts. First, Section 2 reviews previous work on news recommender systems and resources used for teaching information access systems. Section 3 introduces the tasks defined for NewsREEL 2017 and presents the main results. Section 4 looks at the challenge from different perspectives. We highlight results from a participant survey and discuss experiences of using NewsREEL as practical course work. Finally, Section 5 concludes the paper and anticipates future directions for research on news recommender systems.

2 Related Work

NewsREEL, as mentioned in the introduction, is now in its fourth year. While its principles and practices of benchmarking recommender systems have been covered in the overview papers of the previous years, in this paper we focus on the use of NewsREEL as a resource for teaching and learning. In this section, we first briefly introduce the importance of stream-based recommendation in an industry context in Section 2.1. In Section 2.2 we then discuss NewsREEL in the context of higher education.

2.1 Evaluation of News Recommendation Systems

News Recommender Systems, a type of information access system, facilitate finding relevant news articles (*cf.* [4]). The evaluation of recommender systems represents a challenging endeavor. Unlike for information retrieval systems, a consistent notion of relevance has not been established. Shani and Gunawardana [25] distinguish three evaluation methodologies: offline experiments, user studies, and online evaluation. In NewsREEL, our focus is on the static environments of offline experimentation and dynamic environments of online evaluation.

Benchmarking in Static Environments A myriad of offline experiments emerged from academic research on recommender systems. Data sets facilitate repeating experiments under identical conditions. Initially, large-scale data sets focused on movie ratings (*cf.* [3, 10]). In 2013, *plista* released a data set specifically for news recommender systems [14]. Subsequently, multiple updates of the data set have been released in scope of CLEF NewsREEL [13, 18, 15]. Li *et al.* [19] model news recommendation as contextual bandit problem. They define an evaluation procedure yielding valid results in offline settings. Similarly, Joachims *et al.* [12] apply counterfactual reasoning to logs of a news recommender system. They show how estimating propensity scores yields meaningful insights even if the ranking strategy had not been applied to collect the data.

Benchmarking in Dynamic Environments Online evaluation has been established as the preferred mode for industrial applications. Das *et al.* [6] describe how Google's news aggregator presents news stories in a personalized fashion. The system combines MinHash clustering, probabilistic latent semantic indexing, and covisitation counts to

estimate how relevant stories are to a given user. Garcin *et al.* [9] present the case of a Swiss news publisher. They devise a system using context trees to capture changing preferences. Finally, we mention the literature documenting CLEF NewsREEL, which offers participants the opportunity to evaluate their ideas on an industrial news recommender system (*cf.* [13, 17, 16]).

2.2 Education

In 2011, the Royal Academy of Engineering announced that teaching STEM is vital for the UK given the large numbers of industries that "depend on engineering knowledge and skills and [all] are signaling increasing demand and experiencing a scarcity of supply of suitable qualified young people" [1]. This suggests that students of STEM programs need to gain a technical skillset that will allow them to thrive in industry.

Addressing this, many computer science courses consist of lectures and accompanying lab sessions in which students are required to implement pieces of software to better understand the techniques taught in the course. This approach is based on the idea of deep learning outlined by Fry *et al.* [8] where students aim to "gain maximum meaning from their studying", *i.e.*, they are learning by doing. Similarly, Barr [2] highlights the importance of interactive learning environments for the development of valuable skills such as problem solving and the ability to communicate. Smart and Csapo [26] argue that such learning-by-doing activities can engage students, hence supporting active learning.

As outlined by Efthimiadis *et al.* [7], leading educators follow a very similar format when teaching information retrieval. For example, Mizzaro [21] first teaches the theoretical foundations of the subject in the lectures and then asks students to develop a search engine using open source software components. Lopez-Garcia and Cacheda [20] refer to this teaching methodology as "technical-oriented IR methodology" consisting of theoretical lectures and practical work.

Although this technical-oriented teaching method has been introduced in good faith to familiarize students with the theoretical foundations as well as appropriate technical skill sets, Hopfgartner *et al.* [11] argue that the technical challenges addressed in these courses are often too limited and therefore do not support the students in gaining the more advanced skill sets required to thrive in our technology-oriented economy. Hoping to address this shortcoming, they suggest incorporating realistic and complex challenges that model real-world problems faced in industrial settings. In this paper, we provide a preliminary analysis of the potentials of NewsREEL for student learning.

3 News Recommendation Scenario

Recommender systems reduce a large collection of items, news articles in our case, to a manageable subset. Early recommender systems addressed the reduction problem by optimizing a single criterion on a fixed data set. More recently, researchers have pointed out that multiple criteria affect recommendations' perception. Castells *et al.* [5] introduce novelty and diversity as additional criteria. Ribeiro *et al.* [23] emphasize that multiple criteria can be considered when learning suited reduction strategies. Said *et al.* [24]

elaborate on the need to consider non-functional criteria. NewsREEL's scenario encompasses multiple criteria in two tasks.

3.1 NewsREEL Live

Participants deploy their recommendation algorithms into a *living lab* environment. The environment consists of three major parts: recommendation services, communication platform, and publishers' webservers. Participants contribute the recommendation services. They run these on their own systems connected to the communication platform via HTTP. Alternatively, plista offered virtual servers to participants who could not afford their own servers or were located far away. Increased network latency puts server located far off at a disadvantage. The communication platform orchestrates messages and monitors performances and issues. Publishers' webservers interface with visitors and initiate recommendation requests. We consider four basic types of messages. Recommendation requests arise from visitors reading news articles. The communication platform receives recommendation requests, forwards them to available recommendation services, and randomly selects a valid list of recommendations to return to the publisher. The publisher displays the recommendations to the reader. The second type of message serves to keep the article collection up to date. Whenever publishers add new articles or update existing ones, they inform the communication platform, which subsequently forwards the information to all connected recommendation services. The third type of message concerns actions of readers. Readers can access news articles, thus generating impressions. Alternatively, they can click on recommendations thus creating clicks. The communication platform recognizes these events and forwards the information to all connected recommendation services. Simultaneously, it keeps track of click events to measure to what degree individual recommendation services succeed. Finally, the fourth type of message represents errors, *i.e.* cases of failure. Errors occur if the communication fails or an invalid list of recommendation is produced. Recommendation lists are invalid if they include invalid or too few articles. Table 1 summarizes the results of NewsREEL Live. Eighteen algorithms provided recommendations in addition to the baseline. The evaluation period lasted from 24 April to 7 May, 2017, excluding 28 April due to technical difficulties with the logging. Participants collected up to 1268 clicks with a maximum of $81\,245$ impressions. We observe click through rates of up to $2.71\,\%$. Click through rates describe the proportion of supplied recommendations which users subsequently clicked.

3.2 NewsREEL Replay

Participants receive a large-scale data set comprising messages similar to those exchanged in the living lab environment. Each message has a timestamp assigned. Thus, participants can replay the sequence of messages creating conditions similar to the living lab. In contrast to NewsREEL Live, NewsREEL Replay allows participants to issue each request to multiple recommendation algorithms. This enables them to compare algorithms in a repeatable fashion. The spectrum of algorithms include relatively simplistic methods, for example based on popularity or freshness, content-based filtering,

Table 1. Observations from the algorithms run in NewsREEL Live from 24 April–7 May, 2017 except 28 April, 2017

Recommender	Clicks	Impressions	CTR
baseline	726	62052	0.0117
5	58	3708	0.0156
9	879	77723	0.0113
21	817	61524	0.0133
33	166	23023	0.0072
34	600	49830	0.0120
35	810	68768	0.0118
45	813	79120	0.0103
55	2	349	0.0057
56	747	60814	0.0123
59	764	75535	0.0101
61	875	63950	0.0137
62	813	59227	0.0137
63	925	68582	0.0135
64	1139	72601	0.0157
65	1268	81245	0.0156
66	896	42786	0.0209
67	6	816	0.0074
70	12	443	0.0271

and collaborative filtering. Participants can measure predictive accuracy as well as scalability. For instance, they may increase the rate at which requests arrive and compare how many requests various algorithms process within a specified time. An evaluator has been made available to participants. It takes chronologically ordered messages and sends them to a recommendation service. Subsequently, the evaluator checks whether any of the recommended articles appear in the session's future impressions. In addition, the evaluator records the time elapsing until the recommendations arrive. The evaluator produces click rates and a response time distributions based on the records. The response time distribution enables us to assess how quickly and reliably an algorithm generates recommendations. Supplying recommendations quickly is necessary for publishers to include them as the webpage is loaded. The more comfortably the average response time ranges below the permitted limit, the less likely the recommender will fail to supply recommendations. Comparing algorithms' response time distributions, we expect to find differences with respect to all major characteristics describing distributions. These characteristics include the average, dispersion, and skewness. For instance, recommendation algorithms may exhibit slightly higher average response rates yet less dispersion.

3.3 Summary

NewsREEL allows participants to experience realistic conditions as they evaluate recommendation algorithms. They may use the data set to establish repeatable results. Con-

versely, the living lab setting highlights the necessity to pay attention to non-functional aspects such as response time limitations. NewsREEL is particularly interesting for researchers in academia and students. Researchers get access to actual users unavailable in the majority of evaluation initiatives. Students can develop many skills required for future careers in industry. These skills are difficult to obtain via toy examples on static data sets prevalent at university courses.

4 NewsREEL for Learning & Teaching

In [11], Hopfgartner *et al.* argue that the technical skills that are taught in STEM courses at higher education institutes often are limited in scope and therefore do not support students in learning the more advanced skill sets that are required by industry nowadays. They therefore suggest to incorporate realistic and complex challenges that model real-world problems faced in industrial settings in the teaching curriculum. More specifically, focusing on the recommender systems domain, they hypothesize that campaigns such as NewsREEL can be employed to teach students the skills required by modern data scientists. Following through on this line of thinking, this section addresses this hypothesis from two directions. In Section 4.1, we present the results of a survey that was sent out to anyone who had signed up for the lab since it became part of CLEF. Our main motivation for this study was to gain further insights on who is interested in the campaign, and to understand *if* and *how* they benefited from NewsREEL. Section 4.2 presents these challenges and opportunities from the perspective of a course instructor who embedded NewsREEL as a case study in a Data Science course. Section 4.3 concludes this part.

4.1 Participant's Perspective

In order to identify and assess the potentials of NewsREEL as an innovative tool to learn new technical skills that are in high demand in industry, we sent out an online survey to everyone who had registered for any of the NewsREEL tasks since CLEF 2015. The survey consisted of three main parts: With the first part, we aimed to better understand the demographics of our participants. The second part focused then on gathering information about the participants' technical skill set with an emphasis on recommender systems. In the last part, we focused on learning about the participants' motivation to register for NewsREEL and on gathering feedback about what they experienced while participating in the campaign. The participants' responses are summarized and discussed in the remainder of this section.

Demographics The survey was sent by email to 160 people who had registered for NewsREEL in the past. It was completed by ten subjects, nine male and one female (ca. six per cent response rate). Although ten respondents were enough to provide us with valuable insight, ideally, we would have liked to have heard back from more of the registrants who we had contacted. Here, we comment briefly on why the expectation of a higher response rate was probably unrealistic. First of all, it is important to know that a significant number of participants decided to register for multiple, if not even all labs

that were organized as part of CLEF. We realize that it is unlikely that these registrants really had the intention to participate in all labs. Moreover, various participants registered with an email address that suggests that they are students at a higher education institution. While these individuals might have participated in any of the tasks, *e.g.*, as part of their training or teaching, they may have graduated by now and are no longer interested in academic work. In fact, a few emails that were send out to the registrants bounced since the email addresses no longer existed. While we had alternative email addresses for a few of those students, there remained four registrants whom we could not reach.

Our ten respondents were a diverse group. While two participants stated that they are between 18–25 years old and two others indicated that they are between 26–30 years of age, six participants reported that they are in their thirties. Ninety per cent of participants stated they either already hold a postgraduate degree (*i.e.*, MSc or PhD) or that they currently study towards such a degree. Only one participant stated that he has no academic degree. When asked what they currently study, computer science and related degrees were named. Four participants stated that they are currently employed in a university teaching position, two described their job position as programmer or developer. We conclude from this that the group of our participants consists of students, academics, and IT professionals.

Technical Skills In order to better understand the technical skill sets of our participants, we focused in the second part of our survey on aspects related to the implementation and operation of recommender systems.

In the first question, we asked whether our participants had any experience in setting up a recommender system (*e.g.*, as part of their studies or job) before registering for NewsREEL. While 50% confirmed that they did have prior experience, the other half did not have any experience. When asked to outline their experience further, participants reported that they had performed offline evaluation using publicly available datasets, that they had developed such system as part of their thesis, or that they had worked on recommender systems for study and research purposes. One participant indicated that he was involved in the implementation of a commercial enterprise search and recommender system.

In the next question, we asked participants to select from a list of applications and frameworks that they had used before. Multiple answers were possible. The most common answers were Mahout and Idomaar with 40% each, followed by Lenskit and MyMediaLite with 10% each. Three users indicated that they had used none of these.

Next, we wanted to learn more about the datasets that they have used in the past when implementing a recommender system by asking them to choose from a list of the most commonly used datasets. Seventy per cent of participants selected the plista dataset that is used in the NewsREEL challenge. Half of the participants have experimented with the MovieLens dataset. This is hardly surprising given the dominant usage of this dataset for research purposes in the past. Other options that were chosen (by 10% each) include the Million Song Dataset, the Netflix Prize, MovieTweetings dataset, datasets provided as part of the ACM RecSys Challenge, and datasets shared on Kaggle.

Finally, we wanted to learn more about the challenges that the participants faced while developing a recommender system. To this end, we asked them in an open question to outline their experiences. Answers included issues such as evaluation, scalability and responsiveness, data size, lack of information about user and historical data, and finding the right tradeoff between the quality of recommendations and the performance of the algorithms.

Feedback on NewsREEL In the final part of the survey, we explicitly asked the students about their experience in participating in NewsREEL.

First, we asked the participants to indicate the year in which they registered for CLEF NewsREEL. Multiple answers were possible since participants could also register for more than one iteration of NewsREEL in the past few years. We were pleased to see that participants from all three iterations provided feedback in this survey. Two subjects had participated since 2015, six subjects had participated in NewsREEL'16, and five subjects had registered for the most recent iteration. One subject indicated that he did not remember the year in which he registered.

Next, we wanted to learn which task the participants were most interested in. Although the description of the individual tasks have evolved slightly throughout the past three years, all tasks can broadly be categorized as online or offline evaluation tasks. Again, we were happy to see that participants stated interest in both tasks: 80% expressed interest in the online evaluation task, referred to as Online Evaluation in a Living Lab or NewsREEL Live, and 60% were interested in the offline task using the plista dataset, also referred to as NewsREEL Replay. When asked whether they managed to participate in NewsREEL, only one participant stated that he did not participate. When asked for the reason, he stated that he "did not solve it on time".

One of the main motivations for us for developing this survey was to understand what motivated participants most to register for NewsREEL. Our main assumption is that participants participated because they would like to develop a new technical skill set. In order to gain more insights into this issue, we explicitly asked the participants to indicate on a five-point Likert scale whether they are "looking for experience both with software engineering and recommender systems algorithms". A vast majority of 70% either fully agreed or agreed with this statement. Moreover, we asked participants to select from a list the technical aspects that motivated them most to participate. The most commonly chosen answer, selected by 80% of participants, is the challenge of providing recommendations in real-time, followed by the possibility to benchmark recommenders in a large-scale setting. Half of the participants stated that they were attracted by the challenging scenario of news recommendation, and by the opportunity of getting access to a new dataset.

Further, we asked the participants to rate on a five-point Likert scale whether "NewsREEL allowed [them] to acquire new skills relevant for [their] career". An overwhelming majority of 80% either fully agreed or agreed with this statement. When asked to indicate which skills they have learned while participating, the most commonly chosen answers included stream processing, real-time processing, providing recommendations, software development, and data analysis. This supports the hypothesis put forward by

Hopfgartner *et al.* [11] that NewsREEL provides the opportunity to acquire new skill sets that are in high demand by industry.

Different from traditional evaluation campaigns that follow the Cranfield evaluation paradigm, one of the challenges that participants of NewsREEL face is the increased complexity that comes with setting up and connecting with the Open Recommendation Platform. Given this complexity, we assume that it might be easier to work on NewsREEL as a team, thus splitting the workload. While the majority of 70% stated that they worked alone or mostly alone but with input from others, three participants stated that they worked as a team. When asked to elaborate on the advantages and disadvantages of this, they argued that "evaluating many approaches in a team is more efficient and usually yields better results" and that it is "a lot of work for a single person". Although these statements support our assumption we were surprised to learn that the majority of participants worked on their own to solve the NewsREEL challenge. We aim to address this issue by developing a communication channel among participants that would allow them to collaborate with each other further.

An obvious choice for such channel would be to use the CLEF conference as a venue that allows participants to engage and network with each other. Interestingly, only 30% stated that the possibility to present their work at an Academic conference motivated them to participate in NewsREEL. Addressing this issue further, we also asked participants to rate on a five-point Likert scale how much they agree with the statement that "the possibility to network with other researchers of [their] area at the CLEF conference is a motivating factor". Here, we received very mixed responses, suggesting that the academic networking component is of lesser interest to the participants than the opportunity to acquire new technical skills.

4.2 An Instructor's Perspective

After having reported on people's motivation to participate in NewsREEL, in this section we now provide insights and discuss lessons learned from the perspective of a university instructor who embedded NewsREEL in their teaching.

Course Details In the 2016/17 semester, NewsREEL was used as a learning & teaching resource of the Web Intelligence course at Norwegian University of Science and Technology (NTNU). The course targets postgraduate students of the Master and PhD programs of the Computer Science Department. The course objective is to train students to use semantic technologies and open linked data to analyze unstructured content as well as teaching the theoretical foundations of building recommender systems. The course consists of theoretical classes, guest lectures from industrial partners, practical group project assignment and exercise classes.

The course's group project seeks to have students solve a practical problem. They ought to apply methods covered in the lectures to a problem they might encounter in an industrial setting. In previous years, students were assigned different tasks. In the Spring 2017 edition, all students have been tasked to take part in CLEF NewsREEL Replay. The forty students were assigned to eleven groups with three to five members each. All students were in the first year of their master studies except three PhD students.

The allowed time to complete the assignment was set to 18 January to 7 April, 2017. Students had to conduct the experiment, present their findings, and write a delivery report in order to pass the assignment.

Why CLEF NewsREEL? News domain in recommender systems exhibits specific properties and challenges compared with other recommender systems domains such as movies or music [4, 22]. Articles being textual objects support content-based filtering techniques. The data set provided for NewsREEL Replay includes articles' textual content along with interaction data. The latter can be used to apply collaborative filtering techniques. These characteristics facilitate applying the theoretical models taught in the course. These models include recommender systems, text analytics/natural language processing, open linked data, and further semantic technologies. The data set's scale lets students experience conditions in which processing all data on a single computer becomes infeasible. These conditions are expected to become increasingly prevalent as systems keep producing increasing amounts of data. Finally, NewsREEL Live offers students the opportunity to evaluate their ideas with real user feedback. Participation in NewsREEL Live was left as a voluntary exercise for students curious on how well their ideas would perform with real user feedback.

In short, assigning the CLEF NewsREEL challenge to students as a group project naturally fits to the content of the course and the intended learning outcomes of the group project.

Course Assessment NTNU employs multiple tools to assert high quality education. At the beginning of the semester, a student group is voluntarily selected as "reference group" to represent students' interests during the course. The group's task is to give feedback to the teachers. Three meetings between the reference group and the teachers took place. Two meetings took place during the semester and another meeting at the end of the course. In addition, the course organizers held a review session in the course of which all students presented their results and gave feedback about their experiences.

Results and Lessons Learned Two teaching assistants were available to help students with questions and problems throughout the semester. Exercise classes took place on a bi-weekly basis and provided a forum for discussion. Additionally, an online discussion group was added to the official NTNU e-learning platform. At the end of the semester, the reference group stated that more teaching assistants with more experience in CLEF NewsREEL challenge would be helpful. When we consider the high number of questions through e-mails, messages on the discussion platform and face to face meetings, previous hands on experience with CLEF is definitely recommended for the teaching assistants. A lot of the questions from students were about updated framework, documentation and data set inconsistencies and, practical problems about setting up and running the framework, which requires up to date experience on the challenge to answer the questions. This could be related to the fact that NewsREEL Replay does not require using a particular set of tools. Instead, participants are free to use whatever technology they like.

In NewsREEL challenge there are no restrictions for the choice of technology. But in order to make the challenge more efficient within the course context, some restrictions provided by the instructors could be useful and prevent students to get lost in the very many choices of technologies available.

As the representative of all the students in the classroom, the reference group stated at the end of the semester that the group project topic was exciting and up to date, and that they gained a lot of experience about recommender systems technology. They also stated that it was fun to work with a practical project where they can apply what they have learned in the class theoretically. As a downside of the group project, they stated that they have spent much more time than expected to set up and run the framework before they can start implementing their recommender system algorithm. However, with the experience and existing implementations from this year's students, this problem can be minimized in the coming years.

Naturally, there are no textbook clear, step by step instructions for the framework and challenge. And this is quite different from what most of the students get used to. So students should be clearly told what they will deal with during the CLEF challenge and what prior knowledge is expected. Even though we mentioned these, we have observed some confusion within students as a result of dealing with a real challenge.

One of the main goals of this course was to encourage the students to collaborate with each other, hence helping them to develop graduate attributes. Throughout the group project, we observed that students collaborated not only within their own group but also between the groups.

Even though only a few groups could come up with some evaluation results, getting the best results from the challenge was not the main goal in this course's group project assignment. As stated above, the main goal of this assignment was to teach the students about applying theoretical knowledge in practice as well as challenge them with real-world implementation problems and encourage them to work in teams. As a result, we believe that we have reached to the learning goals of this group project.

4.3 Summary

In this section, we have discussed the opportunities that NewsREEL brings for higher education. In particular, we studied the hypothesis presented in [11] that participants of NewsREEL can gain important skills that are of importance in the labor market. For this, we first approached all past and current registrants of NewsREEL to better understand their motivation for registering for the lab. The survey results suggest that the participants did indeed acquire new technical skills while participating. Moreover, we presented a reflection of an instructor who embedded NewsREEL in their teaching. We argue that these insights can serve as guidelines for Academics who might consider using NewsREEL as a tool for learning & teaching as well.

5 Conclusion and Outlook

This paper has discussed the NewsREEL 2017 news recommendation challenge with a particular emphasis on the contribution it makes to education in the area of recommender systems. Our point of departure has been the observation that universities often

fall short of teaching students the full range of skills necessary in order to develop and deploy effective recommender system algorithms. In order to be effective, real-world recommender systems must not only maintain high prediction performance, but must also fulfill requirements of scalability, availability, and response time. The opportunities to evaluate recommender systems offered by NewsREEL allow students to gain first-hand experience in addressing the challenges faced in the types of stream-based scenarios typical for today's large online recommender systems.

The NewsREEL Live task offers a living lab environment in which participants can test their stream-based recommendation algorithms online. The NewsREEL Replay task allows more detailed examination of algorithms by supporting the replay of the information (items, requests, interactions) in the stream. Taken together students have exercised a full spectrum of skills from algorithm design, to implementation, to performance analysis with respect to multiple criteria.

We reported information collected from a survey of past registrants of the NewsREEL task. We found that there was great interest in the opportunity for online recommendation, and also for access to real-world data. The survey revealed that past participants indeed acquired skills in multiple areas important for recommender systems.

In the future, we would like to invest explicit effort into bringing people interested in NewsREEL together in order to solve the problem as a team. This will help to lighten the load on any individual participants. For students, support of people with previous experience was identified as being important. Here, explicit attention to bringing the right people together to address NewsREEL together could be particularly important.

We close by noting that here we have focused on the education aspects of NewsREEL: what past participants have learned, and how NewsREEL supports teachers in the university setting. Moving forward, we are interested in gaining further insight into industry perspectives. In particular, we want to understand whether the skills learned by NewsREEL participants prove valuable in practice "on the job" in an industry setting.

References

1. F. Banks and D. Barlex. *Teaching STEM in the Secondary School*, chapter Enabling the 'E' in STEM. Routledge, 2014.
2. M. Barr. Video games can develop graduate skills in higher education students: A randomised trial. *Computers & Education*, 2017. Accepted for publication.
3. J. Bennett, S. Lanning, et al. The Netflix Prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, NY, USA, 2007.
4. D. Billsus and M. J. Pazzani. Adaptive News Access. *The Adaptive Web*, pages 550–570, 2007.
5. P. Castells, N. J. Hurley, and S. Vargas. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook*, pages 881–918. Springer, 2015.
6. A. Das, M. Datar, A. Garg, and S. Rajaram. Google News Personalization - Scalable Online Collaborative Filtering. In *WWW*, pages 271–280, New York, New York, USA, 2007. ACM.
7. E. Efthimis, J. M. Fernandez-Luna, J. F. Huete, and A. MacFarlane, editors. *Teaching and Learning in Information Retrieval*. Springer Verlag, 2011.
8. H. Fry, S. Ketteridge, and S. Marshall. *A Handbook for Teaching and Learning in Higher Education*, chapter Understanding Student Learning. Routledge, 2003.

9. F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and Online Evaluation of News Recommender Systems at swissinfo.ch. *RecSys*, pages 169–176, 2014.
10. F. M. Harper and J. A. Konstan. The Movielens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.
11. F. Hopfgartner, A. Lommatzsch, B. Kille, M. Larson, T. Brodt, P. Cremonesi, and A. Karatzoglou. The potentials of recommender systems challenges for student learning. In *Proceedings of CiML'16: Challenges in Machine Learning: Gaming and Education*, 10 2016.
12. T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased Learning-to-Rank with Biased Feedback. In *the Tenth ACM International Conference*, pages 781–789, New York, New York, USA, 2017. ACM Press.
13. B. Kille, T. Brodt, T. Heintz, F. Hopfgartner, A. Lommatzsch, and J. Seiler. NEWSREEL 2014: Summary of the News Recommendation Evaluation Lab. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, pages 790–801, 2014.
14. B. Kille, F. Hopfgartner, T. Brodt, and T. Heintz. The plista dataset. In *International News Recommender Systems Workshop and Challenge*, pages 16–23, New York, New York, USA, Oct. 2013. ACM.
15. B. Kille, A. Lommatzsch, G. G. Gebremeskel, F. Hopfgartner, M. Larson, J. Seiler, D. Malagoli, A. Serény, T. Brodt, and A. P. De Vries. Overview of NewsREEL'16: Multi-dimensional Evaluation of Real-time Stream-recommendation Algorithms. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 311–331. Springer, 2016.
16. B. Kille, A. Lommatzsch, F. Hopfgartner, M. Larson, J. Seiler, D. Malagoli, A. Serény, and T. Brodt. CLEF NewsREEL 2016: Comparing Multi-dimensional Offline and Online Evaluation of News Recommender Systems. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 593–605, 2016.
17. B. Kille, A. Lommatzsch, R. Turrin, A. Serény, M. Larson, T. Brodt, J. Seiler, and F. Hopfgartner. Overview of CLEF NewsREEL 2015: News Recommendation Evaluation Lab. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, 2015.
18. B. Kille, A. Lommatzsch, R. Turrin, A. Serény, M. Larson, T. Brodt, J. Seiler, and F. Hopfgartner. *Stream-Based Recommendations: Online and Offline Evaluation as a Service*, pages 497–517. Springer International Publishing, Cham, 2015.
19. L. Li, R. E. Schapire, W. Chu, J. Langford, and J. Langford. A Contextual-bandit Approach to Personalized News Article Recommendation. In *the 19th international conference*, pages 661–670, New York, New York, USA, 2010. ACM Press.
20. R. Lopez-Garcia and F. Cacheda. *Teaching and Learning in Information Retrieval*, chapter A Technical Approach to Information Retrieval Pedagogy. Springer Verlag, 2011.
21. S. Mizzaro. *Teaching and Learning in Information Retrieval*, chapter Teaching Web Information Retrieval to Computer Science Students: Concrete Approach and Its Analysis. 2011.
22. Ö. Özgöbek, J. A. Gulla, and R. C. Erdur. A Survey on Challenges and Methods in News Recommendation. *WEBIST*, 2014.
23. M. T. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani. Pareto-efficient Hybridization for Multi-objective Recommender Systems. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 19–26. ACM, 2012.
24. A. Said, D. Tikk, K. Stumpf, Y. Shi, M. Larson, and P. Cremonesi. Recommender Systems Evaluation: A 3D Benchmark. In *RUE@ RecSys*, pages 21–23, 2012.
25. G. Shani and A. Gunawardana. Evaluating Recommendation Systems. In *Recommender Systems Handbook*, pages 257–297. Springer US, Boston, MA, Oct. 2010.
26. K. L. Smart and N. Csapo. Learning by doing: Engaging students through learner-centred activities. *Business and Professional Communication Quarterly*, 40:451–457, 2007.