Digital Content Profiling Based on User Engagement Features

Pawel Misiorek[1](✉), Michal Ciesielczyk[2], and Bartosz Rzycki[2]

[1] Faculty of Computing and Telecommunications, Institute of Computing Science,
Poznan University of Technology, ul. Piotrowo 2, 61-138 Poznan, Poland
pawel.misiorek@put.poznan.pl

[2] Deep.BI Poland, Warszawa, Poland
{michal.ciesielczyk,bartosz.rzycki}@deep.bi
https://deep.bi

Abstract. Exploring audience engagement with digital media content may lead to many various benefits. In this paper, we study how adding engagement-based features to the article description can influence the efficiency of algorithms aimed at detecting digital media readers' propensity to buy a subscription. Based on the propensity score, the publishers can optimize a decision to display a paywall. Moreover, it is observed that more and more page views are of new or anonymous users. Consequently, the decision concerning the paywall application has to rely only on digital content features. In order to address this application scenario, we propose a novel digital content enrichment framework based on the engagement statistics of users reading a given article. We experimentally evaluate the performance of machine learning algorithms for predicting the propensity to subscribe using the dataset based on events describing the behavior of users exploring the digital news site of the large media publisher. The results of experiments demonstrate that enrichment of article profiles with user engagement features significantly improves prediction models' efficiency.

Keywords: Digital content scoring · Behavioural profiles for articles · Subscription for digital content · Streaming data processing · Big Data

1 Introduction

The digitization in the media industry forces the vast majority of enterprises to rethink and reorganize the revenue model of their organizations. That is why those companies transformed into the digital space to stay profitable and embrace contemporary trends, mainly focusing on the online part of the business. Nevertheless, low barriers to entry into the industry, widespread access to content on similar topics, and reduced attention span among readers make the competition even more fierce. Two primary ways for gaining a competitive advantage emerge for digital media businesses focus on customers or content.

M. Papadaki et al. (Eds.): EMCIS 2022, LNBIP 464, pp. 91–104, 2023.
https://doi.org/10.1007/978-3-031-30694-5_8

In the first case, the organization personalizes the content and strategy based on the user behavior using such tools as cookie tracking, data analysis, dynamic paywall, and dynamic pricing [2,11]. In the second case, attention is paid to analyzing the characteristics of the article and recognizing whether there are any repeating patterns among successful articles in order to provide adequate feedback to content producers - journalists, editors, and the editorial board. The importance of such content analysis methods is increasing in the highly competitive digital media business [1,7]. That is especially relevant considering that companies are striving for more data about users in order to personalize the offer and increase their chances of subscription. However, only a small percentage of readers register on the website and leave contact data. Moreover, it is observed that more and more page views are of new or anonymous users. That is why organizations turn to other ways of increasing the pool of subscribers. Firstly, getting to know the specifics of the created articles enables to optimize the paywall strategy of the organization, including the decision of which articles should be available for the user before displaying the paywall. Secondly, analysis of articles can be a feedback for both the editorial board and authors themselves when it comes to preferences and tastes of the users as well as successful publishing strategy.

The next big step for content analysis is the implementation of artificial intelligence to enhance the business processes of the publishers [1,3,7]. As discussed in this paper, one of the outcomes of such implementation is data-based scoring for the content to better represent its chance of success. Success may be defined differently for every enterprise, data team, and editorial board. In this paper, we focus on success expressed as the situation where an article increases the chances of the user to subscribe. Therefore, adequate machine learning models for Propensity to Subscribe (P2S) are applied to provide a score of the article – how likely it is that the user will subscribe having read that article. The models include variables mentioned above, such as when users read the article, the daily/weekly patterns of interest, behavioral features of the article, number of referred visits (using a link), time of reading and attention, age of the article, and the interaction with a paywall. On top of that, custom variables that help assess the article are added and calculated.

Inspired by the research on modeling user engagement profiles for detection of reader's propensity to subscribe presented in [11], we introduce the content scoring solution based on articles' engagement profiles and aimed to be applied to enhance the dynamic paywall policy. However, unlike in the cases of the analysis of the behavior of readers [11], P2S models for articles are a highly dynamic issue and so the method and approach cannot be copied. That is why, in this paper, we propose the new architecture for building scalable and efficient content scoring solution.

The paper contribution is as follows. We propose the novel content profiling framework being a part of the Deep Glue System [11] responsible for managing and optimizing the access for digital media users. In particular, we describe the article profiles based on comprehensive engagement statistics of users reading this article. Furthermore, we demonstrate how such profiles can be enriched and dynamically updated in real-time and then applied to propensity to subscribe modeling and paywall control. Finally, we experimentally evaluate the performance of machine learning algorithms which utilize the proposed digital content profiles for the application scenario of predicting the propensity to subscribe based on article features.

2 Related Work

The digital content profiling for detecting users' propensity to subscribe is an underexplored research problem [1,2,7]. Many studies concerning the general content scoring problem have already been published, i.e., [10,12], but none of them is focused on digital article scoring aimed at optimizing subscription sales. The most relevant research results on modeling and measuring user engagement with digital articles are presented in [1,7]. Unlike our approach, Carlton et al. [1] study the problem of engagement prediction. Furthermore, they use short video content as their application scenario. On the other hand, the author of [7] analyses user engagement patterns with page views of news articles. Specifically, he investigates the relationship between engagement levels and information gained in the articles' text. In contrast to his research, we are not limited to news articles. Moreover, we are focused on user features closely related to the subscription process, e.g., describing user interactions with a paywall. In [3], Davoudi et al. propose the subscription prediction model using user engagement measures. Additionally, in more recent research presented in [2] the authors propose engagement-based paywall control policies. However, their research is not focused on modeling article profiles and does not investigate the influence of engagement-based profiling on the efficiency of machine learning models.

In the case of research on user profile enrichment techniques, many solutions focus on social media applications [5,9]. Unlike in the case of our method, they are based on processing textual data [9] extracted from social services such as Twitter [5]. Our approach is closer to the research of Tang et al. [13] and Li et al. [8], which propose to build time-agnostic temporal features based on aggregations in a specific time window as some time-forgetting mechanism. However, their studies apply to real-time recommendation systems [13], and streaming service churn prediction [8]. Our research is strictly connected to studies presented in [11], in which the framework for building digital media user profiles using their engagement features has been presented. However, in this paper we introduce the propensity-to-subscribe scoring solution based on articles' engagement profiles, which is aimed to optimize the dynamic paywall mechanism for the case of new or anonymous users.

3 Content Profiling Based on Behavioral Features

In this paper, we introduce the content profiling framework to enrich the page view events with additional engagement-based features of articles. These new features may be seen as the current article profile based on statistics concerning article readers. Specifically, for a given article a and a given timestamp t, the article profile $p(a, t)$ is formally modeled as a sequence of features:

$$p(a, t) = (f_1, \ldots, f_m),$$

where m is the total number of profile features. The new features are generated using events collected in various periods before the time t, which usually corresponds to the timestamp of the enriched event. The details about profile feature types and the description of specific features applied in tests presented in this paper are presented in Tables 1 and 2, respectively.

Table 1. Article profile features.

Feature type	**Description**
Counters	Features counting the number of specific events in a given time window (e.g., today, yesterday, lastN days), e.g., the number of page views, the number of conversions just after reading the article, the number of paywall displays, the number of paywall clicks
Total sums	Sums in a given time window, e.g., total attention time for users reading the article
Averages	Averages of a given feature, e.g., average attention time per user in a given time window
Ratios	Ratios of some features, e.g., ratios to paywall clicks to paywall displays in a given time window, ratio subscribers to users reading the article
Unique values counters	Counters of unique values of a given raw feature, e.g., the number of unique user locations, the number of unique users, the number of unique subscribers
Segment-based features	Features defined by user engagement segments, e.g., numbers of unique users from a given engagement segments (e.g. engaged, perspective, new, fly-by, won-back, etc.)
Percentage features	Percentages of occurrences for some raw feature values, e.g., percentages of visits from a given source (e.g., home, search, social media), percentages of users from a given segment.
Dynamics features	Dynamics of change of a given feature in time, e.g. dynamics of the change in the number of page views between today and yesterday, or today and last week

The proposed article profiles contain the information intended to be helpful when predicting if reading the given article may increase the user's propensity to subscribe for content. In particular, it includes the most recent historical data on article page views, readers' attention time, types of users reading the article, user engagement segments, traffic sources, statistics of paywall displays and clicks, and the number of subscription purchases. Most of the features are aggregation features, including counters of specific events (e.g., the number of subscriptions sold just after reading the article) or total sums of a given original numeric feature (e.g., the number of seconds spent in the system) in the given time window (e.g., today, yesterday or during last 7 days). Additionally, profiles include features based on simple statistics such as the average or percentage of occurrence of some feature values, including segment-based features corresponding to readers from different user groups. Finally, we defined custom features based on predefined formulas involving the current values of original or enriched profile features. Some of them are just simple ratios, and others describe dynamics of given feature change in time, e.g., modeling differences between today and yesterday or between today and last week.

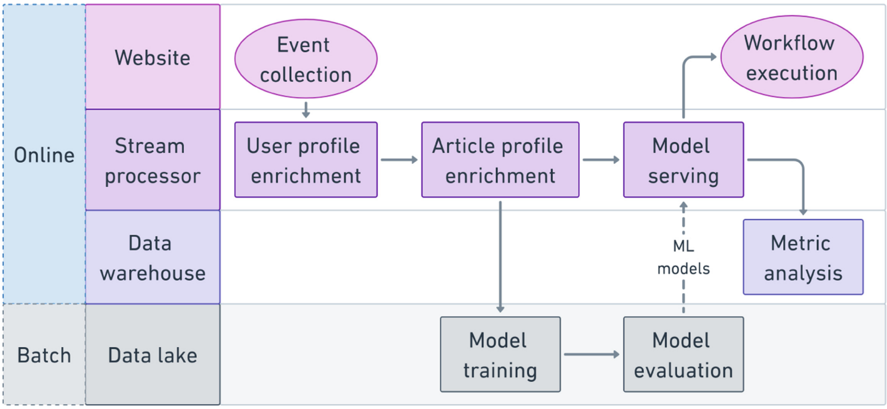

Fig. 1. Deep Glue content scoring architecture diagram.

The overview of the content scoring system architecture is depicted in Fig. 1. Stream of events (describing all user-article interactions) is collected in real-time and stored on a distributed messaging system (Apache Kafka[1]), which is one of the Stream processor components. Each event is enriched on Apache Flink[2] with current engagement features. First, the corresponding user profile, updated on every interaction (based on the solution presented in [11]), is added. Subsequently, the events are enhanced with article profile features (as described in Table 1). Events enriched with engagement features are used to generate predictions controlling the workflow execution for every article on the website. The performance can be monitored online using metrics emitted to a Data warehouse solution. Machine learning model used for serving is trained and evaluated offline, periodically, in batch manner. Models that passed the evaluation are serialized and pushed to the Stream processor environment.

[1] https://kafka.apache.org.
[2] https://flink.apache.org.

4 Experimentation Dataset

We use the unique dataset containing the real data collected based on the traffic on a digital media webpage. It consists of events describing the article views of users exploring the content of a digital site of a large media publisher. Articles published on the website are news, reports or reviews on politics, technology, environment, business, and economics. They have various characteristics including both short news with timely content which are popular for a limited time and then become irrelevant as well as reports or reviews with content which continues to be relevant long time after its publication date.

In this paper we use the data collected during the second half of 2021. The raw data contains around 100M events corresponding to page views of 200K unique articles viewed by 50M unique users. Each event has been dynamically enhanced by features from the most recent article profiles built using available historical information on the engagement of users which read a given article. Then, the enriched samples were used to build the ML models predicting user's propensity to subscribe after reading the article. In order to make our results reproducible, we made our anonymized dataset publicly available[3]. The details of dataset's preprocessing including data cleaning, filtering, engagement-based enhancement, and final dataset's statistics are presented in Sect. 5.

5 Experimentation Setup

In this section, we present the details of the experimentation scenario. The description contains the information on dataset preparation, building machine learning models, and the way of their evaluation.

Dataset Prepartion. The original dataset – introduced in Sect. 4 – consists of 100M events corresponding to article views. Just after its collection, each event was enriched by the most recent article profile available in the profile store. The profiles contain engagement features from the Deep Glue Content Profiling System described in Sect. 3. The outline of article profile features used in experiments presented in this paper is presented in Table 2. The article views are labeled based on the information on the subscription purchase just after reading a given article. Specifically, we selected 2098 new subscription purchases (i.e. non-renewal purchases from newly acquired users) with registered information about the last seen article. Moreover, we excluded all the articles views of users with active subscription from the datasets. Then, due to the fact that the labeled dataset was highly imbalanced, we decided to randomly downsample the events with negative label with the downsampling ratio set experimentally to 0.02. The final data samples were generated synthetically based on the characteristics of data collected. The basic statistics of the datasets are summarized in Table 3.

[3] https://www.kaggle.com/datasets/pawelwmm/contentscoringdataset.

Table 2. Features used in experiments.

Feature Group	**Description**
Raw features (non-profile)	article author, topic, number of days from publication, weekday, day, hour of a view
Attention time/page views (article profile)	attention time today, yesterday, last N days ($N = 7, 30, 60$); average attention time today, yesterday, last N days ($N = 7, 30, 60$); # of page views today, yesterday, last N days ($N = 7, 30, 60$); dynamics of the change in # of page views between today and yesterday, today and last week, and yesterday and last week;
Paywall/conversion (article profile)	# of conversions to subscribed users # of paywall clicks today, yesterday, last N days ($N = 7, 30, 60$); # of paywall displays today, yesterday, last N days ($N = 7, 30, 60$); ratio of paywall clicks to paywall displays today, yesterday, last N days ($N = 7, 30, 60$);
User types (article profile)	# of unique users (which read the article); # of unique subscribers; # unique user locations (city, region, country) of users which read the article; # of unique readers (reader: a user with # of recent views above a threshold); # of unique subscriber readers; ratio of subscribers to users, ratio of readers to users, ratio of subscriber readers to subscribers # of users from a given engagement segments (engaged, perspective, new, fly-by, won-back, etc.) percentage of users from a given engagement segment
Traffic source (article profile)	# of user reading sessions # percentage of visits from a given source: home, search, social media, other

Table 3. Basic statistics of a preprocessed dataset used in experiments.

Total number of samples	106,998
Number of positive samples	2,098
Number of negative samples	104,900
Number of unique articles	19,602
Number of unique article authors	1,868

Experimental Scenarios. We tested the effectiveness of our approach using two experimentation scenarios: (i) a basic off-line scenario assuming 10 repetitions of the experiment based on different random splits to train and testing data, and (ii) an additional real-world scenario assuming efficiency evaluation of models built using historical data. Both scenarios are based on the dataset presented in Sect. 5. The purpose of off-line tests was to obtain the reliable results provided as averages of 10 individual results. Each repetition is based on different random splits on training and test data for a training ratio equal to 0.75. For the real-world scenario, the model was built using the data collected during 20 weeks, and then evaluated during the next 5 week period. By choosing time as a factor for partitioning the data, we could mimic the real-time nature of the target infrastructure. The goal of the real-world experiment was to demonstrate the impact of real-time profile enrichment with time-agnostic behavioral features on the efficiency of propensity-to-subscribe modeling.

5.1 Approaches Under Comparison

To demonstrate the efficiency improvement caused by engagement-based enrichment of article profiles, we compared the following approaches:

- the *baseline* prediction algorithm based on the ML model utilizing the raw features describing the article (see Table 2), i.e., article author, topic, number of days from publication, as well as weekday, day, hour of a page view,
- the *basic profile* prediction algorithm utilizing the content profile enriched by basic engagement-based features based on general counters, i.e., total number of distinct users which read the article, total sum of readers' attention time, total numbers of page views, paywall displays, paywall clicks, and conversions,
- the *full profile* prediction algorithm utilizing each feature of digital content profile introduced in Table 2.

Furthermore, in order to provide some more detailed and insightful discussion, we delivered the additional efficiency comparison for models build using different thematically-grouped parts of engagement-based content profiles. We followed the group definition introduced in Table 2. Specifically, we compared the impact of features related to (i) attention time and page views statistics, (ii) paywall and conversion statistics, (iii) types of users which read the article, and (iv) the traffic source.

We used the CatBoost classifier [4] (implemented using its official library [14]) to build machine learning models compared in this paper. We applied CatBoost with default parameters and predefined *random_state* for our experiments. We indicate all the raw features presented in Table 2 as categorical features. The classifier choice was driven by technological constraints and business needs. Firstly, we were limited to algorithms that did not cause high execution latency, which was crucial to ensure the high-quality real-time performance of the infrastructure. Secondly, since most crucial article basic features, such as the author's name, and the article topic, are categorical, we chose the solution known to handle this kind of data effectively. The efficiency of algorithms was evaluated using the test set by means of standard ML measures [6], i.e., the Area Under the ROC curve (AUC), the average precision (AP), accuracy, balanced accuracy - included due to the fact of dealing with highly imbalanced data, precision, recall, and F1. The evaluation results are presented in Sect. 6.

6 Results

In this section, the results of experiments introduced in Sect. 5 are presented.

6.1 Results of Off-Line Experiments

The results of off-line experiments are presented in Figs. 2, 3, 4 and 5, and then summarized using Tables 4 and 5.

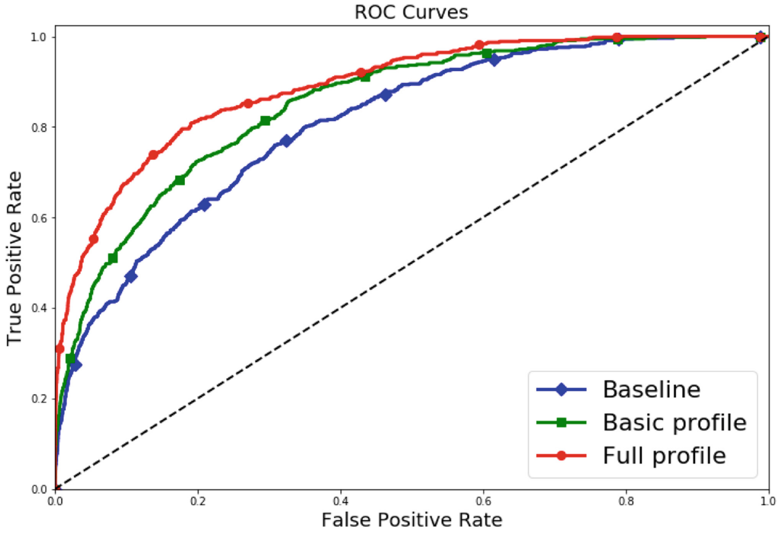

Fig. 2. ROC Curves presenting the impact of article profile enhancement with user engagement features.

Comparing AUC curves (see Fig. 2) proves that the models exploiting enriched data have achieved better efficiency than the models based on raw features describing the articles. We can also observe the quality progress implied by applying full profile features when looking at prediction efficiency through precision-recall curves (see Fig. 3). This observation confirms the importance of adding more specific features, such as counters and averages within the shorter time window concerning events from today or yesterday, and custom features modeling ratios, percentages, or change dynamics. The values of most popular machine learning measures [6] collected in Tables 4 confirm the crucial role

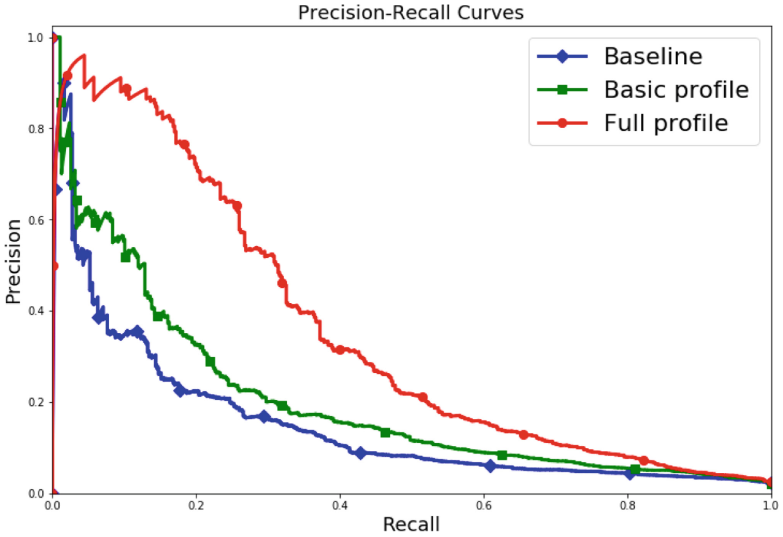

Fig. 3. Precision Recall Curves presenting the impact of article profile enhancement with user engagement features.

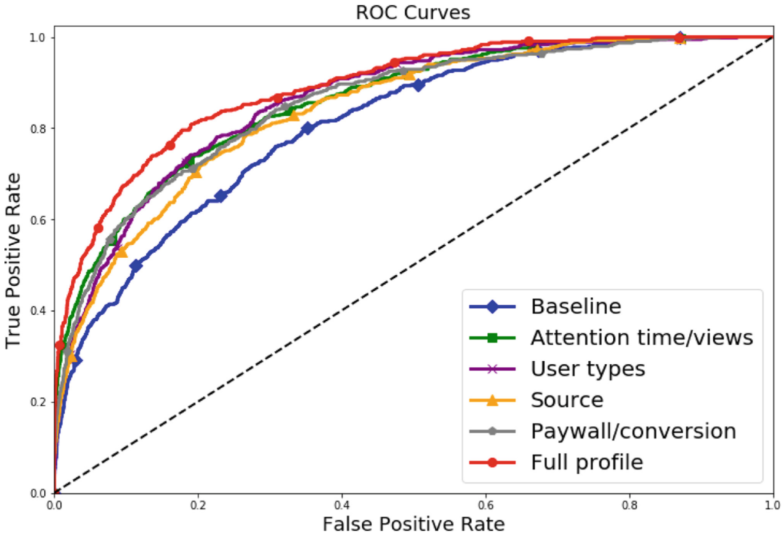

Fig. 4. ROC Curves presenting the impact of various groups of profile features.

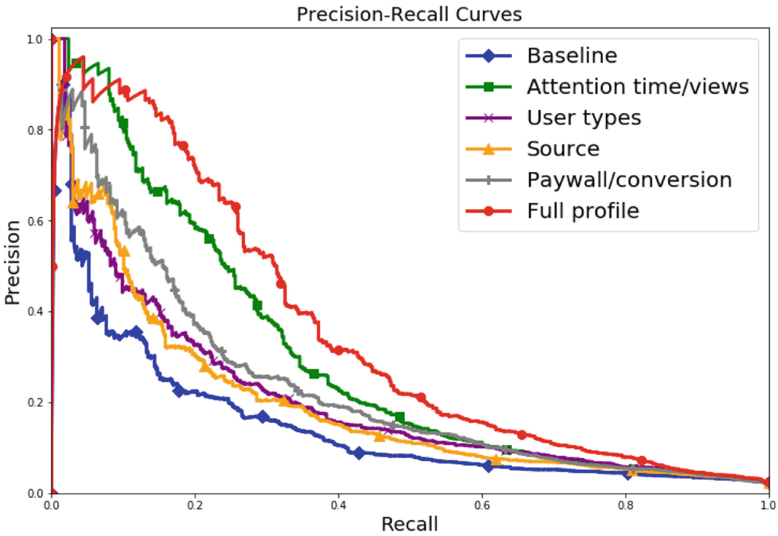

Fig. 5. Precision Recall Curves presenting the impact of various groups of profile features.

Table 4. Results of experiments presented as means and standard deviations for series of 10 experiment iterations (baseline vs user engagement profiles).

Measure	**Baseline**	**Basic Profile**	**Full Profile**
AP	0.1650(±0.0088)	0.2023(±0.0092)	0.3465(±0.0156)
AUC	0.8084(±0.0051)	0.8476(±0.0031)	0.8838(±0.0058)
Accuracy	0.9805(±0.0009)	0.9807(±0.0007)	0.9826(±0.0007)
Balanced acc.	0.5287(±0.0061)	0.5399(±0.0048)	0.5939(±0.0071)
Precision	0.6167(±0.0465)	0.6255(±0.0608)	0.7511(±0.0383)
Recall	0.0581(±0.0123)	0.0808(±0.0099)	0.1890(±0.0144)
F1	0.1059(±0.0207)	0.1426(±0.0146)	0.3013(±0.0159)

Table 5. Results of experiments presented as means and standard deviations for series of 10 experiment iterations (impact of different groups of features).

Measure	**Attention time/views**	**Paywall/ Conversion**	**User types**	**Source**
AP	0.3031(±0.0138)	0.2316(±0.0125)	0.2047(±0.0099)	0.1943(±0.0104)
AUC	0.8523(±0.0065)	0.8474(±0.0050)	0.8517(±0.0041)	0.8402(±0.0044)
Accuracy	0.9821(±0.0009)	0.9811(±0.0008)	0.9808(±0.0008)	0.9807(±0.0008)
Balanced acc.	0.5755(±0.0096)	0.5545(±0.0051)	0.5418(±0.0058)	0.5379(±0.0046)
Precision	0.7486(±0.0309)	0.6575(±0.0430)	0.6346(±0.0524)	0.6273(±0.0462)
Recall	0.1521(±0.0193)	0.1102(±0.0103)	0.0846(±0.0117)	0.0768(±0.0093)
F1	0.2522(±0.0265)	0.1884(±0.0148)	0.1490(±0.0188)	0.1365(±0.0144)

of engagement-based article profiles in the performance of digital subscription propensity models. The efficiency progress is evident when looking at metrics such as average precision, balanced accuracy, and the F1 score. Figure 4–5 and Table 5 deliver the additional efficiency comparison for models build using different thematically-grouped parts of engagement-based content profiles. We have observed that the biggest quality improvement is caused by using features corresponding to the number of article views and their attention time. The results confirm the still high but slightly smaller importance of features describing the paywall displays and clicks and subscriptions bought after reading the article. The impact of user types and traffic sources has appeared as less significant.

6.2 Results for the Real-World Scenario

This section presents the results of tests conducted according to the real-world scenario provided in Sect. 5. This additional experiment aims to check the efficiency of real-time article profile enrichment with time-agnostic behavioral features in the target business scenario involving the online propensity-to-subscribe scoring. This scenario is much harder than the offline scenario based on a random split of train and testing sets since it requires testing the performance using new data, which usually is different from the one used for training. The results are collected in Tables 6 and 7. The biggest performance decrease is observed for baseline models when comparing the metric values with the corresponding results of the offline scenario. In the case of models involving the use of article profiles, this deterioration is much smaller, especially when we limit to basic most-general features. The results confirm that the proposed profile features are more general, time-agnostic, and, therefore, more useful when applying the models in the target online environment. In addition, similar results to those presented were obtained in online tests using a real system, which further reinforces this hypothesis.

Table 6. Results of the experiment for the real-world scenario (baseline vs user engagement profiles).

Measure	**Baseline**	**Basic Profile**	**Full Profile**
AP	0.0769	0.1358	0.2715
AUC	0.7401	0.8543	0.8788
Accuracy	0.9872	0.9871	0.9882
Balanced accuracy	0.5128	0.5184	0.5769
Precision	0.6000	0.4815	0.6750
Recall	0.0258	0.0372	0.1547
F1	0.0495	0.0691	0.2517

Table 7. Results of the experiment for the real-world scenario (impact of different groups of features).

Measure	**Attention time/views**	**Paywall/ Conversion**	**User types**	**Source**
AP	0.2677	0.1731	0.1208	0.1094
AUC	0.8603	0.8363	0.8607	0.8447
Accuracy	0.9884	0.9873	0.9870	0.9871
Balanced accuracy	0.5685	0.5439	0.5141	0.5184
Precision	0.7742	0.5536	0.4545	0.4815
Recall	0.1375	0.0889	0.0287	0.0372
F1	0.2336	0.1531	0.0539	0.0691

7 Conclusions

In this paper, we present that real-time article profile enrichment with time-agnostic features based on users' engagement leads to significantly improved machine learning models to detect the user's propensity to buy. We demonstrate that AI-based accurate targeting of the users interested enough in the offer to pay for access, is ready to become a standard in the digital media industry.

Our findings indicate that to attract more subscribers, the media companies should invest into modeling data-driven features describing article performance. While metrics such as the number of unique users or total number of views are useful in many ways, by themselves are not enough to infer a user's propensity to buy. A fit for purpose ML model can properly take into account multiple factors to predict attractiveness of any particular article in real-time, outperforming any attribution model heuristic that is currently widely used in the industry.

As article attractiveness is usually a highly dynamic concept – especially in the case of news – modeling additional data-driven features describing the most recent readers' engagement has turned out to be an efficient tool for propensity-to-subscribe model improvement. Our profiling method delivers user engagement features that are more context-independent and time-agnostic. Consequently, they are more generalizable and applicable to many scenarios within the industry, especially when the ML models are usually served in an environment that differs from the one where the models are built. Finally, our research opens future work directions, including extending the framework with additional features modeling the content type, such as video, text, or rich-media articles.

Acknowledgments. This work was supported by the Polish National Centre for Research and Development, grant POIR.01.01.01-00-1352/17-00 and by Poznan University of Technology, grant 0311/SBAD/0727.

References

1. Carlton, J., Brown, A., Jay, C., Keane, J.: Using Interaction Data to Predict Engagement with Interactive Media, pp. 1258–1266. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475631
2. Davoudi, H., Rashidi, Z., An, A., Zihayat, M., Edall, G.: Paywall policy learning in digital news media. IEEE Trans. Knowl. Data Eng. **33**(10), 3394–3409 (2021). https://doi.org/10.1109/TKDE.2020.2969419
3. Davoudi, H., Zihayat, M., An, A.: Time-aware subscription prediction model for user acquisition in digital news media. In: Proceedings of the 2017 SIAM International Conference on Data Mining (SDM), pp. 135–143. SIAM (2017). https://doi.org/10.1137/1.9781611974973.16
4. Dorogush, A.V., Gulin, A., Gusev, G., Kazeev, N., Prokhorenkova, L.O., Vorobev, A.: Fighting biases with dynamic boosting. CoRR abs/1706.09516 (2017). http://arxiv.org/abs/1706.09516
5. Fei, Y., Lv, C., Feng, Y., Zhao, D.: Real-time filtering on interest profiles in twitter stream. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, pp. 263–264. JCDL 2016, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2910896.2925462
6. Flach, P.: Machine Learning: The Art and Science of Algorithms That Make Sense of Data. Cambridge University Press, New York, NY, USA (2012)
7. Grinberg, N.: Identifying modes of user engagement with online news and their relationship to information gain in text. In: Proceedings of the 2018 World Wide Web Conference, pp. 1745–1754. WWW '18 (2018). https://doi.org/10.1145/3178876.3186180
8. Li, H., Vu, Q.H., Pham, T.L., Nguyen, T.T., Chen, S., Lee, S.: An ensemble approach to streaming service churn prediction. In: WSDM Cup 2018 Workshop, The 11th ACM International Conference on Web Search and Data Mining, Los Angeles, California, USA, pp. 1–8 (2018). https://wsdm-cup-2018.kkbox.events/
9. Liang, S.: Collaborative, dynamic and diversified user profiling. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 4269–4276 (2019). https://doi.org/10.1609/aaai.v33i01.33014269
10. Madnani, N., Loukina, A., Cahill, A.: A large scale quantitative exploration of modeling strategies for content scoring. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 457–467 (2017). https://doi.org/10.18653/v1/W17-5052
11. Misiorek, P., Warmuz, J., Kaczmarek, D., Ciesielczyk, M.: Modeling user engagement profiles for detection of digital subscription propensity. In: Themistocleous, M., Papadaki, M. (eds.) Information Systems. EMCIS 2021. LNBIP, vol. 437, pp. 55–68. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-95947-0_5
12. Riordan, B., Flor, M., Pugh, R.: How to account for mispellings: quantifying the benefit of character representations in neural content scoring models. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 116–126 (2019). https://doi.org/10.18653/v1/W19-4411
13. Tang, X., Xu, Y., Geva, S.: Integrating time forgetting mechanisms into topic-based user interest profiling. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 3, pp. 1–4 (2013). https://doi.org/10.1109/WI-IAT.2013.132
14. Yandex: Catboost - open-source gradient boosting library. https://catboost.ai/. Accessed October 2022