Characterization and Early Detection of Evergreen News Articles

Yiming Liao[1], Shuguang Wang[2], Eui-Hong (Sam) Han[3,⋆], Jongwuk Lee[4], and Dongwon Lee[1] ✉

[1] The Pennsylvania State University, USA
{yiming,dongwon}@psu.edu
[2] The Washington Post, USA
shuguang.wang@washpost.com
[3] Marriott International, USA
mmmshan@gmail.com
[4] Sungkyunkwan University, Korea
jongwuklee@skku.edu

Abstract. Although the majority of news articles are only viewed for days or weeks, there are a small fraction of news articles that are read across years, thus named as ***evergreen*** news articles. Because evergreen articles maintain a timeless quality and are consistently of interests to the public, understanding their characteristics better has huge implications for news outlets and platforms yet there are few studies that have explicitly investigated on evergreen articles. Addressing this gap, in this paper, we first propose a flexible *parameterized* definition of evergreen articles to capture their long-term high traffic patterns. Using a real dataset from the Washington Post, then, we unearth several distinctive characteristics of evergreen articles and build an early prediction model with encouraging results. Although less than 1% of news articles were identified as evergreen, our model achieves 0.961 in ROC AUC and 0.172 in PR AUC in 10-fold cross validation.

Keywords: News articles · Long-term popularity · Evergreen

1 Introduction

Articles that consistently gain traffic over time, named as ***evergreen*** articles, are of importance to newsrooms because they signal a topic of lasting interests to readers. News outlets and platforms want to continue to serve such articles to new readers over time in many ways–e.g., re-promoting through social media channels or linking next to regular news. Evergreen articles can also provide authoritative and reliable information during recurring events–e.g., solar eclipses or seasonal flu.

Journalists at one of top-10 US daily newspapers by circulations, the **Washington Post** (referred to as *WaPo* in the following), whom we interviewed in

⋆ This work was finished when Eui-Hong (Sam) Han worked at the Washington Post.

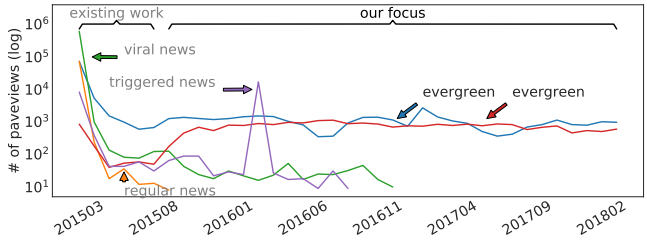


Fig. 1. Examples of news articles' pageview patterns.

2018, currently rely on their memory for keeping track of evergreen news articles. Some newsroom editors in WaPo manually maintain a short list of evergreen articles for reference while others try to identify evergreens by their traffic history and manual confirmation. While the traffic pattern of evergreen articles needs to be thoroughly understood based on a time series analysis, at present, journalists do not have a good or agreed-upon definition of evergreen articles in terms of time series of traffic performance patterns, nor easy-to-use query tools to explore time series data. Furthermore, even after identifying potential candidate evergreen articles, journalists face a daunting challenge of manually reviewing a large number of candidate articles.

Even though we only interviewed journalists at WaPo for this study, we believe that the utility of evergreen articles and challenges associated with identifying them are universal across different news organizations. We believe that the characterization and automatic identification of evergreen articles, especially at early stages, are important tasks with research challenges and practical benefits.

Prior work includes popularity analysis conducted in various domains, such as news articles [9, 21, 16], videos [12, 5], shared images [7, 24] and online series [3]. As illustrated in Figure 1, we can categorize news articles into different types per their temporal traffic patterns. For example, viral news gain significant attention when publishing, though are only viewed for a few days just like regular news, while triggered news refer to those articles that are revisited due to some occurring events. Because the lifespans of the majority of news articles are found to be short [21, 9], existing studies mainly focused on the short-term popularity (i.e., ***trending***) prediction. However, some noticed that certain topics of news articles have a longer life cycle [11, 17], but inferring the popularity of such long-term popular contents (i.e., ***evergreen***) turned out to be very difficult [19]. To our best knowledge, none of prior works have systematically examined on the characteristics of evergreen articles and their automatic identification.

To remedy this gap, this paper starts with the first research question **(RQ1)**: *how can we reliably identify evergreen news articles at a large scale?* Journalists' judgment on evergreen articles currently relies on the manual inspection on the content of articles. Due to the large number of articles being published daily, however, it is infeasible for journalists to qualitatively check the traffic data to select potential candidate articles for further manual review. Therefore, to auto-select quality evergreen candidates for journalists, we propose to define

an evergreen in terms of its traffic aligned with journalists' judgment. Having obtained a highly qualified dataset of evergreen news articles using traffic-based definition, then, we investigate on **RQ2:** *what are the characteristics of evergreen news articles?* and explore possible factors correlating their long-term popularity. Instead of taking years to monitor articles' long-term traffic pattern, finally, we move to tackle **RQ3:** *can we identify evergreen news articles at early stages?* and recommend potential evergreen articles published in recent months to journalists.

In answering these research questions, we make the following main contributions:

1. To our best knowledge, this paper is the first work to explicitly study news articles' long-term popularity.
2. We formally propose a parameterized definition of an evergreen article with respect to the article's historical page view data, as validated by journalists.
3. We analyze possible factors correlating news articles' long-term popularity.
4. Based on our analysis, we build machine learning models for the early detection of evergreen news articles and report promising results from empirical experiments as well as real deployment in WaPo.

2 Related work

Since predicting news articles' popularity has long been considered as an important research area, there are a lot of existing efforts to this problem. However, unlike other online content, such as videos and photos, the life spans of news articles tend to be shorter and their view counts often decrease faster [19, 9]. As a consequence, the majority of previous studies lies in the scope of *short-term* news articles' popularity analysis. In this section, we review related works and contrast them to our work under two categories: (1) news articles' popularity prediction; (2) long-term popularity prediction for other online content.

Predicting Popularity of News. As one of the first investigations on the popularity prediction of online content, [19] discovers that news stories exhibit a much shorter lifespan than videos and usually become outdated after hours. Besides inferring future popularity by historical time series, recent works take advantage of news articles' content information in prediction [9, 14]. To better understand user participation in news' propagation, researchers are also interested in predicting the number of users' comments [22, 21, 16, 23], number of votes [19, 16, 10] and number of tweets [1]. As noted earlier, due to the short lifespan nature of news, almost all of these existing works focus on predicting news articles' popularity within a short time. The most related work to ours is [6], where they define *long shelf life* news articles–i.e., those taking at least 80 hours to reach 60 percent of their total page views in the lifetime. Despite targeting at analyzing long shelf life news articles, however, [6] fails to consider the temporal dynamics of news articles' long-term popularity, thus cannot capture *evergreen* news articles, which consistently attract high traffic over their lifetime across many years. In this paper, taking temporal dynamics into consideration, we

propose a reliable measurement to capture long-term popular news articles and perform a systematic analysis on them.

Predicting Long-Term Popularity. Although few works in news domain study long-term popularity prediction, there are some works focusing on long-term popularity prediction of other online content, such as videos [5, 20] and paper publications [18]. To model popularity evolution, viewing content propagation as a stochastic process is one of the most common methods. For example, [18] adopts the reinforced Poisson process in paper citation network and [26] models the cascading in social network as a Hawkes process. For complex systems such as video platforms, where content propagation is difficult to model, researchers turn to time series approach and feature driven approach. In time series approaches, often, early popularity series are used for future predictions. For instance, [5] assigns varying weights to videos' historical popularity series via a multiple-linear regression model to predict future popularity. Due to the dependency on the historical popularity, time series approaches usually require an extended period of observations and suffer from the so-called *cold-start* problem. Feature driven approaches are proposed to address such issues. Diverse types of potential features that may impact popularity are incorporated into the prediction, such as text features [4], author features [15] and meta features [12]. Similar to videos, news articles have various traffic sources, including search engine and social media, which makes the propagation also hard to model. Hence, this paper adopts the feature driven approach with historical popularity series to predict news articles' long-term popularity.

Compared with other online content, news stories generally are more time sensitive. Only a very small fraction of published news articles exhibit long-term popularity patterns over many years, which makes the problem more challenging. Understanding why these small number of articles are consistently of interest to the public will benefit both journalists and news sites. To our best knowledge, our work is the first to systematically define evergreen articles, to examine their characteristics, and to predict them at early stages.

3 RQ1: Defining Evergreens with Traffic

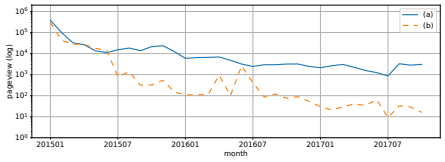
This section first introduces the dataset used in the later analysis and experiments. Then, we answer *RQ1: how can we reliably identify evergreen news articles at a large scale?* by proposing a parameterized definition of evergreen news articles in terms of traffic data.

3.1 The Washington Post News Article Dataset

Our dataset contains more than 400,000 news articles published by WaPo from January 2012 to June 2017. For each article, we dumped its monthly page view data from its publication date to March 2018. The *median* page view of each article's i-th month after initial publication is shown in Figure 3. Because the

(a) evergreen

(b) non-evergreen



(c) two news articles' monthly page views

Fig. 2. Articles with similar initial traffic data show different long-term popularity pattern. For example, article (a) consistently receives high traffic in long term, while (b)'s monthly page views drop dramatically over time.

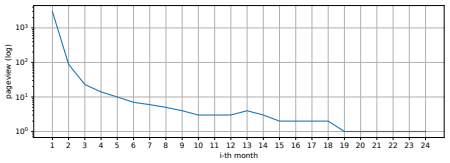


Fig. 3. Median page views of news articles published from January 2012 to December 2015

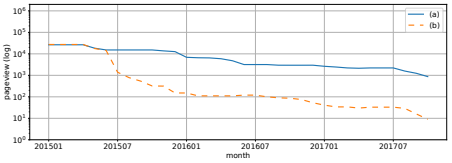


Fig. 4. Median filtered page views of two example articles from Figure 2(c).

distribution of news articles' page views in i-th month is highly skewed, where most page views are zero, we present median page views, instead of average page views here. Note that the articles' monthly page views in Figure 3 drop faster than log-linear and down to around 20 at 3rd month already, which is consistent with previous observation [19] and confirms the short lifespan of news articles.

3.2 Definition of Evergreen Articles

Consulting journalists at WaPo, we set the observation time of each article up to 2 years after its first publication, which is significantly longer than the period studied in prior studies. Specifically, we denote the monthly page view series of an article in its first 24 months after publication as $PV = (pv_1, pv_2, ..., pv_{24})$, where pv_i is the page view number in i_{th} month after its publication. As shown in Figure 3, the popularity of news articles drops dramatically and down to tens at 3rd month. Since most news articles become outdated after 3rd month, we view the first 3-month traffic data as a news article's *initial traffic* and use it to define trending articles. More specifically,

Definition 1 (Trending Article) *During the observation period, we sort news articles by their total page views in the first 3 months,* $\sum_{i=1}^{3} pv_i$*, in descending order and consider the first* k *ranked articles as* **top-k trending** *articles.*

As a motivating example to capture evergreens, in Figure 2, we present two news articles published in January 2015 and their monthly page views till October 2017. Note that, despite receiving similar page views in the first a few months, these two articles exhibit radically different long-term traffic patterns. Article (a)

maintains high traffic long after its publication and receives more than 5,000 page views a year later, while the traffic of article (b) drops dramatically to around 100 after a year. Clearly, article (a) is consistently of more interest to readers, thus fits the definition of an evergreen.

Although most news articles have a short lifespan similar to article (b) in Figure 2 and exhibit a fast decaying traffic pattern after the initial publication, we still observe occasional peaks long after their publication. Interviewed with domain experts at WaPo, we found out these peaks mainly resulted from 1) journalists at WaPo promoted these articles; 2) these articles were associated with occurring events. These sudden traffic peaks are usually caused by stochastic events[1] and very difficult to predict. Accordingly, to better capture long-term popular news articles, we propose to use *median filters* to remove those sudden peaks. In processing page view series via median filtering with window size γ, each month contain median value in a γ-size window around the corresponding month (e.g. $\hat{pv}_i = median([pv_{i-\frac{\gamma}{2}}, ..., pv_{i+\frac{\gamma}{2}}])$). We denote the smoothed page view series as $\hat{PV} = (\hat{pv}_1, \hat{pv}_2, ..., \hat{pv}_{24})$, and present examples of smoothed traffic pattern in Figure 4.

In addition, because of news articles' trending nature, initial traffic of news articles generally are significantly higher than the rest, up to several orders of magnitude. Therefore, we ignore the first 3-month traffic when identifying evergreen news articles. Traffic patterns of evergreen articles should not decrease too fast. To get smooth traffic series, we adopt accumulated traffic series and use the following *normalized* metric to measure a traffic pattern,

Definition 2 (Accumulated Traffic Ratio (ATR)) *Denote an article's total page view number from 4th month to i_{th} month after its publication as $N_i = \sum_{j=4}^{i} \hat{pv}_j$, where $4 \leq i \leq 24$. Then, an article's ATR for i_{th} month is defined as $ATR_i = \frac{N_i}{N_{24}}$, where $4 \leq i \leq 24$.*

When an article has constant traffic, its ATR starts low but will increase linearly. At the same time, since trending articles' traffic mostly falls in the first a few months, their ATR starts high but increases slowly. If we look at the area under ATR for an evergreen article vs. a trending article, the evergreen article will have smaller area[2]. In addition, to ensure the quality of candidate evergreen articles, journalists at WaPo require a few hundred pageviews per month for each article.

With these observations, we propose the following parameterized operational definition of an evergreen news article.

Definition 3 ((α,β,γ)-Evergreen Article) *Ignoring the initial first 3-month traffic, we denote the monthly page view time series of an article x during the*

[1] Even though some events occur more regularly than others, such as seasonal festivals and holidays, and thus might be predictable, the impact of events on news articles' long-term popularity is beyond the scope of this paper. We leave it for future study.

[2] Because the median filtering is employed to smooth page view series, occasional traffic peaks are removed and will not affect the area.

remaining observation period as $PV = (pv_4, pv_5, ..., pv_{24})$*. Then, first, we use: (1) the median filter with a window size* γ *months to smooth the time series* PV *as* $\hat{PV} = (\hat{pv}_4, \hat{pv}_5, ..., \hat{pv}_{24})$*. If the smoothed time series* $\hat{PV}$ *satisfies (2) average monthly page view at least* α *such that:* $\frac{1}{21}\sum_{i=4}^{24} \hat{pv}_i \geq \alpha$*, and (3) normalized area under ATR at most* β *such that:* $\frac{1}{21}\sum_{i=4}^{24} ATR_i \leq \beta$*, then, the article* x *is referred to as an* (α, β, γ)**-evergreen** *article.*

Note that α guarantees the minimum monthly page views, β controls the decaying rate of an article's page views, and γ is used to remove the sudden page view peaks caused by unpredictable events. Although *median filters* help remove sudden traffic peaks, smoothing page view series will lose some information on series dynamics, and the loss will increase with the increase of γ. With the observation that most sudden traffic peaks only last for 1 or 2 months, a smooth window of 5-month should remove most sudden traffic peaks. As such, empirically, we set $\gamma = 5$ (months), and explore different α and β values in Section 3.3.

α	250-500	500-1000	>1000
Positive Rate%	3.33%	10.00%	25.00%

(a) β=1.0, γ=5(months)

β	0.0-0.6	0.6-0.7	0.7-1.0
Positive Rate%	18.33%	11.67%	6.67%

(b) α=250, γ=5(months)

Table 1. Journalist validation with different α and β

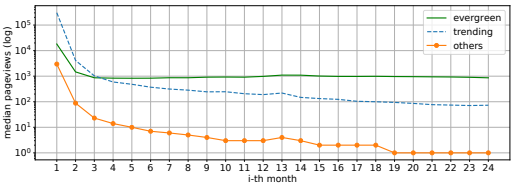


Fig. 5. Median page view comparison between evergreen and trending news articles

3.3 Tuning α and β

To guarantee each article to have at least a 2-year traffic history, we use news articles published from Jan. 2012 to Dec. 2015 for analysis, and reserve the remaining articles published from Jan. 2016 to Jun. 2017 for testing purpose. Ignoring articles with zero first 3-month page views, which are mainly caused by traffic tracking errors, we have 250,642 news articles in the training set.

To further validate the effect of α and β in capturing evergreens, we consulted domain experts at WaPo to manually label a few samples from each criterion[3]. More specifically, we sample 60 articles from each α with β=1.0 and each β with $\alpha \geq 250$, then mix them up for labeling. As expected, Table 1 shows the agreement of journalists on our definition that an article with larger α and lower β is more likely to be evergreen. Considering the rarity of evergreen articles, our definition is confirmed to filter out most non-evergreen articles and produce highly qualified evergreen datasets. Weighing both the quality and size of the dataset, we adopt (α=500, β=0.6, γ=5) as the criterion to finally obtain 1,322 evergreens out of 250,642 new articles in WaPo, a mere 0.5 % of all news articles, and use them as the base evergreen articles for further analysis and experiments.

[3] Labeling details are similar to Newsroom Editor's Evaluation in Section 5 E4.

Comparison with trending articles To make a fair comparison, from January 2012 to December 2015, we select the same number of 1,322 trending articles, of which less than 10% are overlapped with evergreen articles. Then, the median page view comparison between evergreen and trending articles is shown in Figure 5. As expected, trending articles initially attract significantly higher traffic than evergreen articles, but quickly fade away from users' attention. On the contrary, evergreen articles are consistently of interest to the public, and generally obtain almost one order of magnitude higher monthly page views than trending articles after one year of publication.

4 RQ2: Characterizing Evergreens

Having identified evergreen news articles from traffic data, we turn to *RQ2: what are the characteristics of evergreen news articles?* In this section, we focus on characterizing evergreen news articles, and examine on possible factors correlating their long-term popularity.

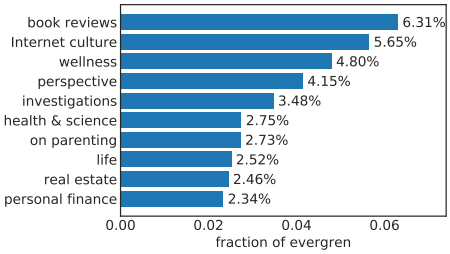


Fig. 6. Top 10 categories with the highest evergreen ratios

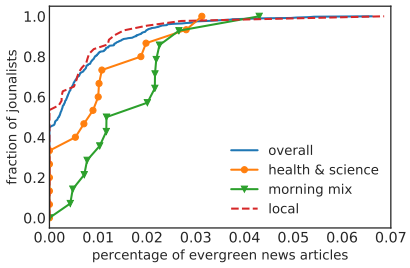
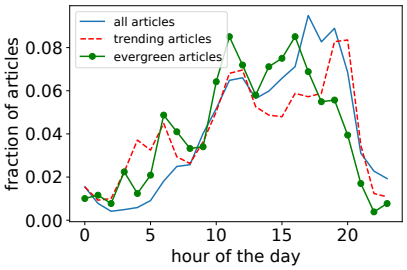
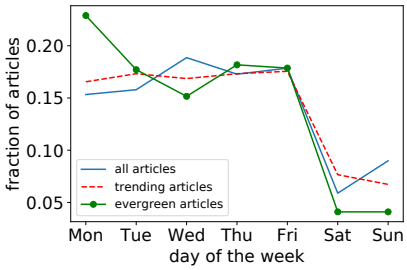


Fig. 7. CDF of fraction of evergreens in each journalist's publication.



(a)



(b)

Fig. 8. (a) fraction of articles published in each hour of the day; (b) fraction of articles published on each day of the week.

Category Previous studies show categories that are manually assigned by journalists play an important role in identifying trending articles, where some categories tend to generate more viral articles than others [9]. Therefore, we investigate here if there exists a similar relationship between categories and

evergreens. Excluding categories with less than 100 articles, we sort the remaining 127 categories by their evergreen ratio (i.e., a fraction of evergreen articles over all articles in a category) and present the top 10 categories in Figure 6. Unsurprisingly, *book reviews*, *perspective* and *investigations* are less time sensitive and convey higher proportion of evergreen articles. In addition, people pay continuous attention to Internet culture, health and lifestyle related topics. Among categories conveying fewer evergreen articles, besides *politics* and *sports*, *movies* and *music* are less intuitive ones, where only 1 or 2 articles out of thousands meet our criterion. One possible explanation is, though movies and music are less time-sensitive and so do their reviews, they are more likely to be affected by popular culture.

Topic Although categories are manually assigned by editors and describe the main genres of articles, we found the manually assigned categories are both noisy and coarse-grained. For example, some categories such as *perspective* and *local* are broad and could include articles related to education and real estate, which are more likely to be evergreen than others in the same category. Therefore, to extract more fine-grained topics and better understand articles in terms of content, we utilize topic modeling techniques. More specifically, we train a 1,000-topic noun only topic model using the LightLDA [13, 25], and compare topic distributions between evergreen and non-evergreen articles. Based on article contents, *wellness*, *health*, *housing*, *research studies* and *parenting* are the most sustainable and lasting topics, which are consistent with a few top evergreen categories. When digging into other top categories, we discover that articles about history, family and relationship/friend are more likely to be evergreen in *book reviews*, while evergreen articles in *Internet culture* talk more about diseases, research studies, relationship/friend and parenting than non-evergreen. In addition, history, education, parenting and family are among the most evergreen topics in *perspectives*. As expected, article contents go beyond categories and provide more informative topics, indicating that categories are not enough and we should consider article contents in evergreen detection.

Publication Time Next, we explore the relationship between news articles' long-term popularity and their publication time. Figure 8 shows the fraction of articles published in each hour of the day and on each day of the week. Interestingly, trending articles are more evenly distributed across the hours and days, even well represented at odd times such as those from midnight to 8 am or over weekends. Although most news articles are published in the afternoon, evergreen news articles are mostly published in the middle of the day. More interestingly, few evergreen news articles are published on weekends, while Monday conveys the most evergreen articles. Unlike trending articles, which are time sensitive, evergreen articles are less urgent, so journalists may spend more time on editing or polishing over the weekends and publish them on Monday. Since the distribution of evergreen news articles' publication times does show distinctly different patterns from regular articles, we include publication times of articles as an important feature in learning.

Journalist "Who writes an article" is another important meta data to consider. In this section, we examine the role of journalists on articles' long-term popularity. There are over 12,000 journalists in total in our dataset. After removing the journalists who have written less than 100 articles, we have 479 journalists who have written a total of 187,769 articles. Consistent with the ratio of the entire dataset, it turns out that 870 articles among 187,769 articles (i.e., 0.5%) are (α, β, γ)-evergreen articles. Around 50% of the journalists who wrote more than 100 articles published at least one evergreen news article. In Figure 7, we present cumulative distribution function (CDF) of fraction of evergreen news articles in each journalist's publication. For each fraction X, CDF shows the proportion of journalists with the fraction of evergreen publication less than or equal to X. This figure illustrates that, although most journalists have written a small proportion of evergreen news articles, there are indeed a few journalists good at publishing long-term popular news articles. In order to exclude the possible effect of category on journalists' publication, we select a few categories, including *health & science*, *morning mix* and *local*, and check CDF of fraction of evergreens per journalist in exactly the same category[4]. As shown in Figure 7, even for articles in the same categories, we still observe similar CDF of percentage of evergreen articles per journalist, indicating that, besides categories, writing styles, such as wordings and article structures, should also matter in evergreen production. As such, journalist information can give some hints on the early detection of evergreens.

5 RQ3: Predicting Evergreens Early

Taking insights from last section, we intend to answer *RQ3: can we predict evergreen articles at an early stage?* and attempt to build an accurate machine learning model to unearth early a small fraction of evergreen articles among many non-evergreen articles. To better validate our learned model, we conduct the following experiments:

E1. We first show the results of stratified 10-fold cross validation on the training set (Jan. 2012–Dec. 2015).
E2. Using the best setting from 10-fold cross validation, then, we train a model on the training set (Jan. 2012–Dec. 2015) and test the model on the new articles published later (Jan. 2016–Apr. 2016).
E3. In addition to classification measurements, we present temporal page view patterns of predicted evergreen news articles published in subsequent months (Jan. 2016–Dec. 2016).
E4. Finally, we consult newsroom editors in WaPo to manually check and validate the quality of predicted evergreen articles.

[4] Although each journalist usually publishes in various categories, here we only consider journalists' articles in each selected category and exclude journalists having < 100 articles in each selected category.

Set-Up. For a given news article, after observing its initial traffic in the first three months, we predict whether it will be an evergreen or ephemeral story. Based on the analysis in the last section, we propose to exploit three feature sets in our model as follows:

1. **Traffic features**: As indicated in prior works, the degree of popularity at the early stages are strongly related to that of future popularity. Thus, we extract the traffic features from the first 3-month traffic data, including the number of monthly page views, the difference of page views of two consecutive months, and the decreasing rate between two consecutive months. Note that, when defining evergreen news articles, we ignored their first 3-month page views and labeled articles only by their later popularity patterns.
2. **Content features**: Topic analysis demonstrates that content features, including keywords and topics, are also important clues in detecting evergreens. Therefore, we exploit word embedding and bag-of-words trick to extract content features from news articles. More specifically, we train an unsupervised FastText [2] model on all news articles with 200-dimension feature vectors.
3. **Meta features**: Based on the observations in the last section, we selectively include meta features such as news articles' category, publication time, and journalist information. There are two challenges to encode categories and journalists information as features: 1) both categories and journalists are numerous in our dataset (e.g. our dataset contain more than 120,000 journalists), *one-hot encoding* (i.e., encoding each category or journalist as an isolate feature dimension) will easily cause overfitting; 2) categorical values of category and journalist are not fixed and vary with time, for example, many journalists come and go. To tackle these issues and obtain more compact meta features for each article, we propose a simple but effective approach to encode category and journalist information: use the average FastText embedding of all prior published articles in the same category or by the same journalist as its meta feature. As a result, similar categories are close to each other in our category feature space. For example, *opinion* is close to *postpartisan*, while *lifestyle* is close to *entertainment*. Likewise, journalists with similar writing styles are close to one another. Finally, we include publication hour of the day and day of the week as categorical features.

Evaluation Metrics. As described in the Section 3.3, we use (α=500, β=0.6, γ=5) as the criterion that yields 1,322 articles out of the total 250,642 news articles as evergreen news articles. Note that this is a binary classification problem with a significantly skewed class distribution ratio of 0.5 : 99.5. Because of the highly imbalanced dataset, we choose to compare models with both Area Under Receiver Operating Characteristic Curve (ROC AUC) and Area Under Precision Recall Curve (P-R AUC), where random baselines are 0.5 and 0.005 respectively.

Moreover, in real settings, journalists expect top-K potential evergreen candidates and want to manually review them to determine true evergreens. Therefore, Precision@K is also provided to measure the performance of top-K predictions. Since only 0.5% articles meet our criterion, the perfect results of Precision@K

will be 1.0 when K $\leq$ 0.5%. We set K to 0.1%, 0.2%, 0.5% and 1.0%. Considering averagely there are around 5,000 articles per month, these percentages correspond to top 5, 10, 25 and 50 predictions for each month respectively.

We experimented with three learning models–i.e., Logistic Regression, Random Forest, and Gradient Boost Decision Tree (GBDT)–using all features via stratified 10-fold cross validation. As the model learned using LightGBM package [8] consistently performed the best, in the following experiments, we only report the prediction results using the GBDT as the main learning model.

Table 2. 10-fold cross validation

Metric	initial traffic	content feature	traffic + meta	traffic + content	content + meta	traffic + content + meta
Precision@0.1%	0.1240	0.1440	0.2720	0.2840	0.2120	**0.3480**
Precision@0.2%	0.1400	0.1320	0.2560	0.2780	0.1800	**0.2900**
Precision@0.5%	0.1296	0.1040	0.1912	0.2312	0.1304	**0.2360**
Precision@1.0%	0.0940	0.0892	0.1580	0.1868	0.1032	**0.1952**
ROC AUC (0.5000)	0.9302	0.8752	0.9485	0.9601	0.8851	**0.9608**
P-R AUC (∼0.005)	0.0763	0.0533	0.1292	0.1606	0.0705	**0.1718**

E1. Cross Validation. The stratified 10-fold cross validation results of different feature combinations are conducted on the exactly same 10 folds and given in Table 2. The results show that first, using only the first 3-month page view series, it is difficult to identify evergreen news articles. However, adding content and meta features gives a significant improvement on all metrics. When utilizing all features, we achieve the best performance, where the P-R AUC improves from 0.0763 to 0.1718 and Precision@0.5% increases from 0.1296 to 0.2360. More interestingly, pre-publication prediction, which only considers content and meta features, achieves very promising results of having 0.1304 accuracy in top-0.5% prediction. Moreover, when adding traffic features, we observe the boosted improvements in P-R AUC and Precision@K, implying that news article's initial traffic and contents contribute to its long-term popularity in different aspects and both of them are indispensable in the early prediction of evergreens. A possible explanation is that news articles with high initial traffic enjoy high visibility, thus evergreen articles in this group are more likely to be shared or archived by users. For evergreen articles with limited initial traffic, though lacking enough visibility to users in the beginning, because of their timeless quality and interesting topics, they still gain high traffic in the long term via other channels like search engine or re-promotion.

Table 3. Time-Split experiment

Prec@0.1%	Prec@0.2%	Prec@0.5%	Prec@1.0%	ROC AUC	P-R AUC
0.2941	0.2464	0.2326	0.1884	0.9399	0.1534

E2. Time-Split Classification. In real applications, early detection models aim to predict evergreen articles from newly published news articles. Therefore, this time-split experiment is designed to examine how general a learned model is across time. Using the best setting from 10-fold cross validation, we *train* a model on the articles published between Jan. 2012 and Dec. 2015 and *test* it on articles published from Jan. 2016 to Apr. 2016. In the test set, 254 articles ($\sim$ 0.7%) out of 34,509 are identified as *(500, 0.6, 5)-evergreen*. The result is presented in Table 3. Comparable with 10-fold cross-validation, our model with all of traffic, content, and meta features achieves 0.2941 top-0.1% accuracy and 0.1534 P-R AUC. Note that, since data distribution often varies over time, a

time-split experiment is a more challenging setting than cross-validation. Even for evergreen news articles, popularity still varies over time. Thus, the performance gap between cross validation and time-split experiment is due to the changes in data distributions.

E3. Time-Split Traffic Evaluation. Because classification measurements are too strict to distinguish news articles with slightly different temporal patterns, we propose to examine temporal page view trajectories of predicted evergreen news articles, which gives more intuition on articles' long-term popularity and serves as more valuable indicators in production. Similar to E2, we train a model on the articles published from Jan. 2012 to Dec. 2015, predict evergreen news articles in each month from Jan. 2016 to Dec. 2016 and monitor their page view trajectories from the publication month till Mar. 2018.

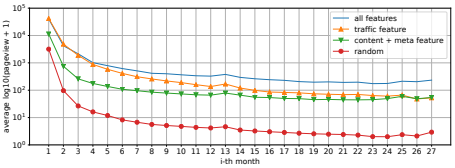


Fig. 9. Page view patterns of top predicted evergreen articles under different feature groups

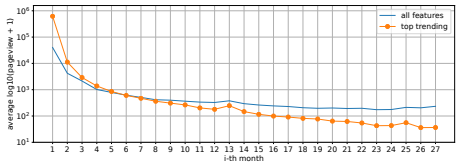


Fig. 10. Page view pattern comparison between top predicted evergreen articles with top trending articles

Aimed at comparing different feature groups, we present the traffic patterns of top predicted evergreen articles with different feature groups in Figure 9. More specifically, for each model, news articles are ranked by the predicted probabilities among their publication month, and we choose the top 100 ranked articles in each month as potential evergreen articles. Consequently, 1,200 articles out of ∼100,000 in total are selected in each group. Overall, news articles selected from all feature groups exhibit a long-term popularity pattern, while combining all features obtains the highest monthly page views in the long run. For each month starting from 4th month, we conduct t-test and find that articles predicted with all features attract significantly more page views than other article groups at a p-value < 0.0001. Additionally, although articles predicted with traffic features show dramatically higher initial traffic, they display similar page view trajectories to articles predicted with pre-publication features long after publication.

As shown in Figure 5, trending news articles, which enjoy significantly higher visibility when just being published, display much higher popularity than most news articles, and thereby should be a strong baseline in long-term popularity prediction. Therefore, in Figure 10, we present page view pattern comparison between top predicted evergreen articles and top trending articles. Top trending articles consist of the top 100 articles with highest initial traffic in each month. As expected, top trending articles receive much higher traffic in the first few months, while top predicted evergreen articles demonstrate a more stable traffic pattern and consistently gain more page views one year later. For each month starting from 9th month, articles predicted with all features attract significantly more page views than trending articles at a p-value < 0.0001 under t-test.

E4. Newsroom Editor's Evaluation. In order to help newsroom editors at WaPo, we plan to deploy the early detection model and recommend recently published potential evergreen news articles to them. Although we have presented promising performances in classification experiments and indeed observed long-term popularity patterns of predicted evergreens in the time-split evaluation, quality check by journalists themselves is indispensable before actual deployment. Therefore, we include newsroom editors' evaluation in this section, presenting both pre-deployment evaluation and product evaluation.

Identifying evergreen news articles at early stages is challenging to both machines as well as journalists. Predicting whether a news article will be long-term popular requires domain expertise in several areas, including news editing, social media promotion, and search engine optimization (SEO). As only a few domain experts are qualified for this task, we consult an audience development team in WaPo, who is responsible for optimizing news articles and conducting social promotions, to manually check the quality of evergreen predictions in recent months. More specifically, there are three major use cases for evergreen news articles at present in the newsroom:

1. Regularly re-promote evergreens (e.g. parenting guide) on social media.
2. Worth updating. Readers may constantly go over some evergreen news articles (e.g. mass shooting statistics) through web favorites or search engines. To ensure readers get the most recent information, editors should keep updating those articles with new content at times.
3. Embedded as linking URLs to related context (e.g. disease information) and references when editing new articles, which could both motivate readers to be more engaged on the website and serve as SEO purpose.

Based on these three use cases, for each news article in the evaluation, if its value is confirmed in any use case, it will be labeled as a true evergreen.

E4.1 Pre-deployment Evaluation. In the evaluation, for each month from Jan. 2017 to Jun. 2017, we recommended the top 10 potential evergreen news articles predicted by our model to the team at WaPo. Through manually reviewing these 60 news articles, 65% of predictions, 39 in total, were labeled as true evergreens. Considering that only tens of news articles out of thousands published per month could be evergreen articles, 65% top-10 accuracy in monthly prediction is encouraging and indicates that our model is practical enough to be deployed in production.

E4.2 Product Evaluation. Based on the promising results from pre-deployment evaluation, we deployed the evergreen prediction system at WaPo. The system regularly updates evergreen prediction model and sends out weekly recommendations to the newsroom. On every Monday, editors receive the top 5 potential evergreens from articles published during last week and provide feedback in Slack. An example of weekly recommendation is shown in Figure 11, where 3 out of 5 articles are verified as true evergreens. The system was deployed at WaPo in July 2018 and achieved $\sim 43\%$ accuracy in the weekly recommendation.

More importantly, Both 65% of top-10 accuracy in monthly prediction and 43% of top-5 accuracy in weekly prediction are achieved without any human labels, validating that our parameterized definition of evergreen is reasonable and can produce highly qualified evergreen datasets.

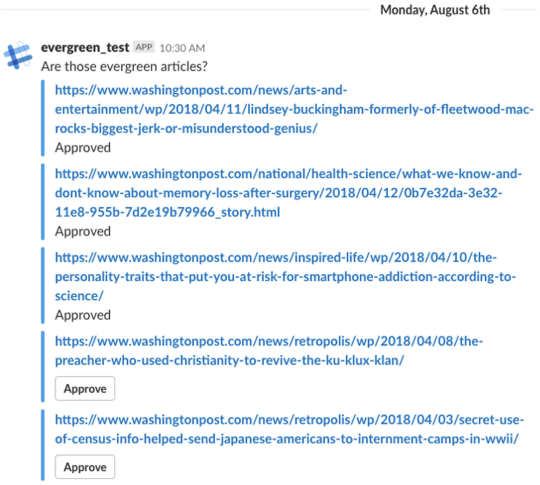


Fig. 11. An example of weekly recommendation

6 Conclusion

This paper presents a study on characterizing and early detecting evergreen news articles. Firstly, taking temporal dynamics into consideration, we proposed a parameterized definition to capture evergreen news articles, which is validated by journalists (**RQ1**). Next, we conducted a quantitative analysis to shed light on evergreen news articles' long-term popularity, and discovered that evergreen articles are closely correlated with several features such as categories, topics, publication time, and journalists (**RQ2**). Finally, based on the insights from our analysis, we built early prediction models on evergreen identification (**RQ3**). Throughout extensive experiments, we have validated that our models gain promising results in early detection of evergreen news articles, and shown that the predicted evergreen news articles indeed exhibit a long-term high traffic pattern. More importantly, verified by journalists, our proposed model achieves encouraging performance in production.

Acknowledgments

We truly appreciate Everdeen Mason, Sophie Ho, Greg Barber and their journalist team at the Washington Post for the valuable feedback and support in both problem formulation and manual evaluation.

References

1. Bandari, R., Asur, S., Huberman, B.A.: The pulse of news in social media: Forecasting popularity. ICWSM pp. 26–33 (2012)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. TACL **5**, 135–146 (2017)
3. Chang, B., Zhu, H., Ge, Y., Chen, E., Xiong, H., Tan, C.: Predicting the popularity of online serials with autoregressive models. In: CIKM. pp. 1339–1348. ACM (2014)
4. Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted? In: WWW. pp. 925–936. ACM (2014)
5. Pinto, H., Almeida, J.M., Gonçalves, M.A.: Using early view patterns to predict the popularity of youtube videos. In: WSDM. pp. 365–374. ACM (2013)

6. Elhenfnawy, W., Wright, J., Kallepalli, K., Racheal, K., Gupta, A., Parimi, R., Shah, P., Li, Y.: What differentiates news articles with short and long shelf lives? a case study on news articles at bloomberg. com. In: BDCloud-SocialCom-SustainCom. pp. 131–136. IEEE (2016)
7. Gelli, F., Uricchio, T., Bertini, M., Del Bimbo, A., Chang, S.F.: Image popularity prediction in social media using sentiment and context features. In: MM. pp. 907–910. ACM (2015)
8. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: NIPS. pp. 3149–3157 (2017)
9. Keneshloo, Y., Wang, S., Han, E.H., Ramakrishnan, N.: Predicting the popularity of news articles. In: SDM. pp. 441–449. SIAM (2016)
10. Lerman, K., Hogg, T.: Using a model of social dynamics to predict popularity of news. In: WWW. pp. 621–630. ACM (2010)
11. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: SIGKDD. pp. 497–506. ACM (2009)
12. Ma, C., Yan, Z., Chen, C.W.: Larm: A lifetime aware regression model for predicting youtube video popularity. In: CIKM. pp. 467–476. ACM (2017)
13. Martin, F., Johnson, M.: More efficient topic modelling through a noun only approach. In: ALTA Workshop. pp. 111–115 (2015)
14. Marujo, L., Bugalho, M., Neto, J.P.d.S., Gershman, A., Carbonell, J.: Hourly traffic prediction of news stories. arXiv preprint arXiv:1306.4608 (2013)
15. Mishra, S., Rizoiu, M.A., Xie, L.: Feature driven and point process approaches for popularity prediction. In: CIKM. pp. 1069–1078. ACM (2016)
16. Rizos, G., Papadopoulos, S., Kompatsiaris, Y.: Predicting news popularity by mining online discussions. In: WWW. pp. 737–742. ACM (2016)
17. Setty, V., Anand, A., Mishra, A., Anand, A.: Modeling event importance for ranking daily news events. In: WSDM. pp. 231–240. ACM (2017)
18. Shen, H.W., Wang, D., Song, C., Barabási, A.L.: Modeling and predicting popularity dynamics via reinforced poisson processes. In: AAAI. pp. 291–297 (2014)
19. Szabo, G., Huberman, B.A.: Predicting the popularity of online content. Communications of the ACM **53**(8), 80–88 (2010)
20. Tan, Z., Wang, Y., Zhang, Y., Zhou, J.: A novel time series approach for predicting the long-term popularity of online videos. IEEE Transactions on Broadcasting **62**(2), 436–445 (2016)
21. Tatar, A., Antoniadis, P., De Amorim, M.D., Fdida, S.: From popularity prediction to ranking online news. Social Network Analysis and Mining **4**(1), 174 (2014)
22. Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., de Amorim, M.D., Fdida, S.: Predicting the popularity of online articles based on user comments. In: WIMS. p. 67. ACM (2011)
23. Tsagkias, M., Weerkamp, W., De Rijke, M.: Predicting the volume of comments on online news stories. In: CIKM. pp. 1765–1768. ACM (2009)
24. Wu, B., Mei, T., Cheng, W.H., Zhang, Y., et al.: Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In: AAAI. pp. 272–278 (2016)
25. Yuan, J., Gao, F., Ho, Q., Dai, W., Wei, J., Zheng, X., Xing, E.P., Liu, T.Y., Ma, W.Y.: Lightlda: Big topic models on modest computer clusters. In: WWW. pp. 1351–1361. ACM (2015)
26. Zhao, Q., Erdogdu, M.A., He, H.Y., Rajaraman, A., Leskovec, J.: Seismic: A self-exciting point process model for predicting tweet popularity. In: SIGKDD. pp. 1513–1522. ACM (2015)

Appendix 1 (α,β,γ)-Evergreen

We present page view trajectories of the articles filtered by different parameters in Figure 12. As shown, controlling average monthly page view number α is insufficient to filter out evergreens, while adding β and γ help.

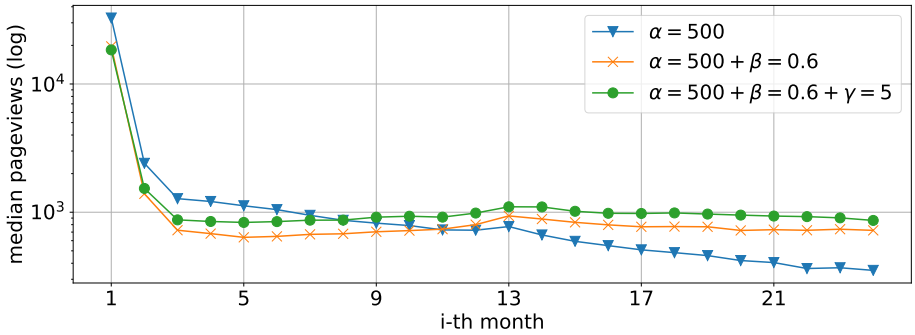


Fig. 12. Effect of α,β and γ on the dataset

Appendix 2 Topic modeling

Sorted by the *evergreen topic ratio* (i.e., $\frac{\text{topic proportion in evergreens}}{\text{topic proportion in non-evergreens}}$) of each topic's proportion in evergreen articles to its in non-evergreen articles, the top ranked evergreen topics are shown in Table 4.

Table 4. Top 10 evergreen topics

Rank	Top 10 words in each topic	Evergreen Topic Ratio
#1	health, diet, study, sugar, food, fat, disease, weight, calorie, body	9.13
#2	doctor, patient, hospital, health, treatment, physician, care, medicine, surgery, illness	6.05
#3	paint, wood, material, water, wall, tile, piece, brick, surface, floor	5.65
#4	brain, memory, welch, trauma, mind, body, love, emotion, activity, people	5.61
#5	scientist, human, bone, dinosaur, animal, specie, researcher, fossil, creature, study	5.06
#6	study, researcher, research, university, author, behavior, journal, finding, professor, effect	4.61
#7	kid, parent, child, school, teen, adult, parenting, mom, son, family	4.29
#8	illness, flu, symptom, strain, allergy, nose, people, fever, disease, health	4.11
#9	hour, night, sleep, time, day, bed, minute, holmes, morning, schedule	4.00
#10	study, percent, datum, research, rate, poverty, researcher, income, factor, effect	3.96

Appendix 3 Sentiment analysis

While sentiment features are widely used in trending news prediction, we find it less helpful in evergreen prediction. More specifically, we utilize Vader sentiment

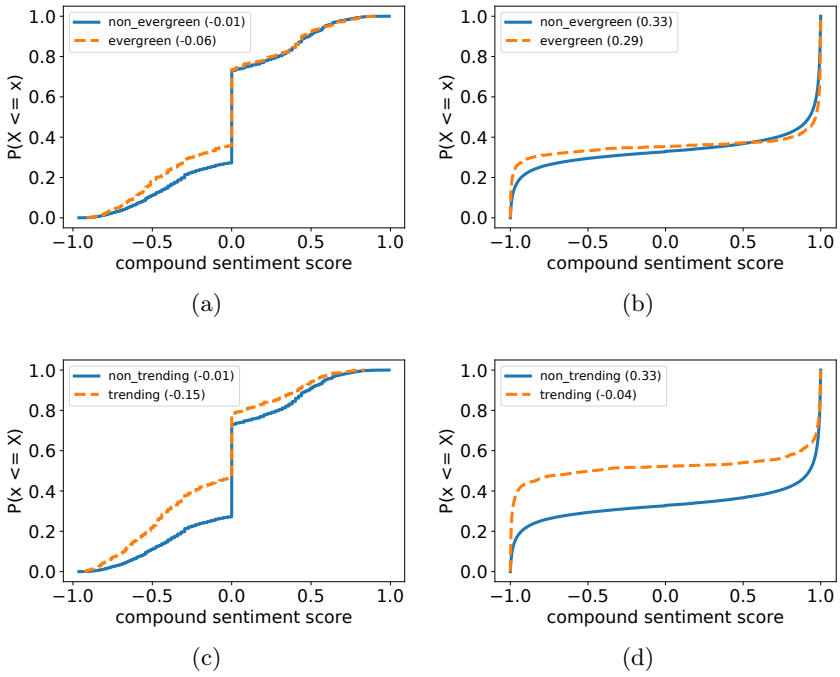


Fig. 13. Both (a) and (b) show CDF of evergreen articles' compound sentiment scores on title and full content individually; Both (c) and (d) present CDF of trending articles' title and full content compound sentiment scores correspondingly. Average compound sentiment scores are included in the parenthesis. The cliffs in (a) and (c) indicate that most titles are neutral.

analyzer [2] to examine articles' sentiment and present CDF of the compound sentiment scores for both full content and title in Figure 13. Compound sentiment score ranges from -1 to 1, indicating the most extreme negative to the most extreme positive. These figures reveal that most news articles at WaPo carry neutral titles and positive contents, while evergreen news articles show more clear polarity in contents. Interestingly, trending articles present much more distinctive patterns than evergreen articles in sentiment, where they convey slightly more negative titles and contain a higher percentage of negative articles. One possible explanation is that breaking and thus viral news may include reports about accidents, disasters, or surprising events, that often have negative tones in their coverage. Although our finding is contrary to the previous study [1], which discovers that positive articles are more likely to be viral than negative ones, sentiment features still contribute to identifying trending articles. However, in the case of evergreen detection, the importance of sentiment-based features seems less significant.

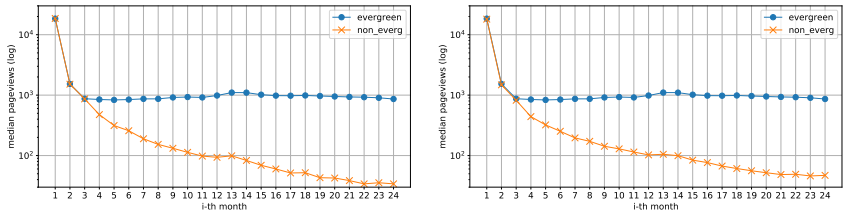
Appendix 4 Category feature

As described in the paper, we encode meta feature with article content and here we include Table 5 to show that similar categories share similar feature vectors.

Table 5. Examples of the most relevant categories within our feature space under Euclidean distance

categories	opinion	politics	health & science	travel	sports	lifestyle
Top 10	opinions	powerpost	to your health	going out guide	colleges	express
	postpartisan	in the loop	storyline	kidspost	early lead	kidspost
	right turn	post politics	wonkblog	lifestyle	maryland terrapins	entertainment
	outlook	virginia politics	innovations	express	allmetsports	arts and entertainment
	posteverything	2chambers	tripping	post local	fancy stats	comic riffs
	letters to the editor	d.c. politics	outlook	in sight	d.c. sports bog	post local
	the post's view	national	animalia	real estate	d.c. united/soccer	blogs & columns
	in theory	opinions	posteverything	home & garden	washington wizards	going out guide
	on leadership	fact checker	letters to the editor	social issues	wizards/nba	style

Appendix 5 Importance of content feature



(a) (b)

Fig. 14. (a) Paired setting with comparable initial traffic; (b) Paired setting with comparable initial traffic + same category

To further validate the effectiveness of article content on evergreen detection, we conduct two paired setting experiments, and match each evergreen article with 1) another non-evergreen article that gets the most similar first 3-month traffic (under euclidean distance) and 2) another non-evergreen article that gets the most similar first 3-month traffic in the same category. The monthly median page view comparison between evergreen and non-evergreen articles in each pair set is shown in Figure 14 (a) and (b) respectively. As expected, in each pair set, evergreen and non-evergreen articles demonstrate similar initial traffic but divergent page view trajectories in the following months.

In the prediction, we divide article pairs into 10 folds. For each round, we leave 1 fold out as test set and mix the rest 9 folds with remaining non-evergreen articles as training set. Using the model proposed in the paper, we measure the percentage of pairs whose evergreen article gets higher probability than non-evergreen article, and present the results in Table 6. Although articles with similar initial traffic (and from the same category), our model achieves $\sim 70\%$

accuracy in distinguishing evergreen and non-evergreen articles. The results not only validated the importance of content features in evergreen detection, but also prove the inadequacy of only traffic and category features

Table 6. paired setting experiment (10 folds average)

Metric	comparable initial traffic	comparable initial traffic + same category
Accuracy	74.36% ± 2.26%	68.47% ± 3.09%

Appendix 6 Case study

In table 7, we present a few evergreen articles that are successfully detected by our deployed model. Despite the encouraging performance of our system, there are still some false positives. When manually reviewing those false positives, we discover that most articles still carry evergreen topics. For example, two articles entitled *What marijuana legalization did to car accident rates* and *Large study supports 'weekend warrior' approach to lifetime fitness* are about studies but labeled as non-evergreens, because journalists think "this is a study that's very specific, it will get outdated very quickly" and "its basis on a study makes it easily and quickly outdated". Another example is *America will only end racism when it stops being racist*, which is "pegged to a specific news event (Charleston shooter), rather than as a comprehensive look at the cause of racism". Apparently, beyond topics, article analysis at semantic levels is one of the future directions for evergreen detection. In addition, with our system running, we are collecting more highly qualified true and false positive evergreens to share with the publishing community.

Table 7. Examples of confirmed evergreen identification

Headlines
The slow, secret death of the electric guitar. And why you should care.
A baby girl. A baffling disease. And the only way to help her is to hurt her.
Does playing on artificial turf pose a health risk for your child?
Why can 12-year-olds still get married in the United States?
Potatoes get a bad rap. They dont deserve it.
How to inexpensively repair a crumbling retaining wall at your home?
'Life or death for black travelers': How fear led to 'The Negro Motorist Green-Book'
What We Know And Dont Know About Memory Loss After Surgery?

References

1. Berger, J., Milkman, K.L.: What makes online content viral? Journal of marketing research **49**(2), 192–205 (2012)
2. Gilbert, C.H.E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: ICWSM (2014)