Understanding Reader Backtracking Behavior in Online News Articles

Uzi Smadja
Technion - Israel Institute of Technology
Haifa, Israel
uzi.smadja@gmail.com

Max Grusky
Cornell Tech, Cornell University
New York, NY, USA
grusky@cs.cornell.edu

Yoav Artzi
Cornell Tech, Cornell University
New York, NY, USA
yoav@cs.cornell.edu

Mor Naaman
Cornell Tech, Cornell University
New York, NY, USA
mor@jacobs.cornell.edu

ABSTRACT

Rich engagement data can shed light on how people interact with online content and how such interactions may be determined by the content of the page. In this work, we investigate a specific type of interaction, *backtracking*, which refers to the action of scrolling back in a browser while reading an online news article. We leverage a dataset of close to 700K instances of more than 15K readers interacting with online news articles, in order to characterize and predict backtracking behavior. We first define different types of backtracking actions. We then show that "full" backtracks, where the readers eventually return to the spot at which they left the text, can be predicted by using features that were previously shown to relate to text readability. This finding highlights the relationship between backtracking and readability and suggests that backtracking could help assess readability of content at scale.

KEYWORDS

Backtracking, User Engagement, Readability, Online News

ACM Reference Format:
Uzi Smadja, Max Grusky, Yoav Artzi, and Mor Naaman. 2019. Understanding Reader Backtracking Behavior in Online News Articles. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313571

1 INTRODUCTION

Monitoring the interaction of users with Web pages at a large scale has contributed to a better understanding of reading behaviors [19] and to connecting patterns of readers' engagement to the textual content of articles [7]. At the same time, textual features have been used to model important aspects of writing, such as how readable [5] or engaging [21] a specific document may be. The latter, however, does not take into account the empirical behavior of readers, nor does it take into account specific attributes of the publication audience.

In this work, we use page instrumentation data[1] collected during reading sessions of online news articles in order to measure and study a reading behavior we refer to as *backtracking*. This under-explored signal of engagement occurs when readers scroll back or upward (see Figure 1) within an online article. A reader may backtrack, for instance, to re-read a previous section of the text. This definition of backtracking diverges from how backtracking was defined in lab studies of readability and text comprehension. For example, studies of *saccades* [31] — short, rapid eye movements between two close fixation points — refer to backtracking as regressive saccades, when the reader goes back a few words in the text. In contrast, our notion of online backtracking, while not as fine grained, presents multiple advantages. Such signals can be collected at scale by simple instrumentation of a Web page, allowing for regular observation of reading behavior in naturalistic settings over long periods of time.

Studying fine-grained signals of engagement such as backtracking at a large scale has the potential to greatly inform our understanding of reading and text. Recent work has examined similar signals to understand reading behavior in relation to textual content. In particular, Lagun and Lalmas [19] use within-page measurements of engagement, such as dwell time, to categorize reading sessions depending on the time spent in different areas of the page. Grusky et al. [8] use similar data to show that reading speeds (i.e., words-per-minute reading rates) are consistent with the ones observed in lab studies. Finally, recent work has begun to explore the modeling of user behavior using content-based features. Grinberg [7] has used accumulated information gain in the article to predict the farthest point read by users on the page.

In this work, we focus on backtracking behavior and its potential to reflect the qualities of the text. To this effect, we examine 1.4M sessions in which about 25K users read and engage with 8,000 online news articles across two major online publications. Because readers can backtrack for a variety of reasons, we define, explore, and measure the prevalence of three different forms of backtracking within articles. To better understand the connection between backtracking and article text, we develop text-based models that predict which articles experience high and low frequency of backtracking actions. We find that the signal we characterize as *full backtracking*– where a reader, after backtracking, eventually returns to the spot

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313571

[1]The anonymized instrumentation data was graciously provided by Chartbeat for research purposes as detailed in Section 3.

at which they stopped reading–is related to properties of the text such as coreference and lexical features. Our results point towards the potential of backtracking as a new audience-driven approach to predict text comprehension at scale.

2 RELATED WORK

User engagement has been studied in depth since the early days of the Web. Signals have been gathered at page level (e.g. click-through rate [16]) or within pages, and in different settings, either transparently via instrumented pages or in controlled user studies. Such signals have then been leveraged for multiple tasks, such as understanding interaction with news articles [19], search satisfaction in browsers via mouse movements [14], or in mobile devices via touch actions [10]. Within the large body of work on this topic, we focus on work pertaining to reading behavior observed via eye-tracking devices, to readability, and to measuring the relationship between engagement and content.

Eye-tracking devices have been used in lab studies in order to examine readers' behavior. Most related to our work is the study of *saccades*, which are defined as a rapid movement of the eye between fixation points [31]. Saccades occur most often when scanning a text or trying to spot a specific piece of information while reading. Regressive saccades consist of going back by a few words, or a line at most, and represent about 10-15% of saccade events. They have also been referred to as backtracking, but differ from our definition of online backtracking as they operate at a much smaller scale. Eye tracking has also been used to explore the processing complexity of text when reading [17]. Other studies [22–24, 32] have incorporated eye-tracking and psychology in qualitative experiments and showed that good readers backtrack efficiently, while poor readers sometimes struggle to find the location that caused the difficulty, under the conjecture that comprehension failures are the cause of the regressive fixation. In contrast, our backtracking signal uses scrolling as a proxy for eye movement on the news article and is less granular, but requires no equipment and can be observed at scale.

Reading behavior has been studied extensively through the analysis and prediction of text readability. Traditional readability measures are typically derived from simple ratios of text properties, such as the average number of words per sentence, or the frequency of difficult words [3]. We take advantage of commonly used reading metrics, such as the Automated Readability Index (ARI) [34], which is defined as a function of the average character-per-word and word-per-sentence ratios. Other readability measures that we explore count the number of syllables per word [6, 26] or compute a ratio of multi-syllable words and the number of words and sentences [9, 18] to assess reading ease or grade level difficulty. Other reading indices include the Dale-Chall formula [4] that leverages a predefined list of "easy" words, with every other word being considered "difficult," or the nonlinear SMOG index [25], which uses syllable counting in order to identify challenging polysyllable words. Readability studies have also explored more complex language models [1, 29] considering grammatical structures [12, 13] and methods such as entity-grid representation [30, 33]. Feng et al. [5] have compared between different features for readability assessment, including coreferences, lexical and discourse features. These readability studies focus on assessing readability by grade level, using for instance manually-tagged datasets for grades 2-5 [28], or are targeted to non-native speakers, who are trying to learn a new language [29]. This makes these approaches not applicable to our domain of online news articles, for which most readers are past grade-level reading and are to the most part native speakers. For such an analysis, we lack large-scale manually tagged corpora, thus adopting the backtracking signal instead.

Similar to our work, which does not rely on manually-tagged corpora, other work has taken advantage of the textual content of Web pages to understand user behavior. In particular, Grusky et al. [8] and Lagun and Lalmas [19] have examined ways of measuring, validating, and modeling how users engage with online news articles. Similarly, Grinberg [7] has looked at semantic information gain and the development of ideas within the body of articles along with their topics in order to predict different types of engagement on news articles. We extend this existing line of work by focusing on a new signal of engagement, backtracking, which captures different aspects of reading behavior.

3 UNDERSTANDING BACKTRACKING

We first define backtracking events and identify the different types of backtracking seen in the data. We also describe backtracking patterns at page level, user level, and session level.

Backtracking takes place when readers break their sequential reading of an online document and scroll up, returning to a previous point in the text. One possible interpretation of such an event in the context of online news is that the user interrupted their sequential reading in order to re-read portions of the document, for example to find earlier references, to names, or to get additional clarity on content that may impact their understanding of the rest of the article [22–24]. Backtracking events may also involve other tasks the reader may perform on the page, such as, navigation or response to ad placement [36]. We explore several definitions that capture different aspects of backtracking.

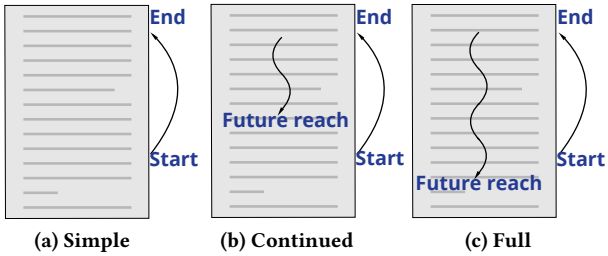

Figure 1: Different types of backtracking on Web pages

Defining backtracking (BT) events. A backtracking event occurs during a page-read event and is represented by an origin and target point defined as the tuple:

$$b = (start, end, reach)$$

Figure 1 presents the reading pattern for three different types of backtracking events. The first two values, *start* and *end*, represent the actual backtracking action and indicate a location in the article (hence, $start > end$). We also define *reach* as the furthest location

Table 1: P_{20} dataset size and number of backtracking events

	Users	Pages	Sessions	Simple BT	Continued BT	Full BT
LONG	15,949	3,968	694,728	57,897	28,840	246,787
SHRT	11,117	3,957	763,335	136,153	49,211	127,421

reached by the user following the current backtracking event. Capturing the *reach* value for each backtracking action helps interpret the context of the action by noting whether the user leaves the page, keeps scrolling or eventually returns to the *start* point. We define three types of backtracking events:

- Simple backtrack: Reader r scrolls back on page p but does not keep scrolling afterwards. Based on our previous definition, for the tuple $(start, end, reach)$, $reach \leq end$. This is the most basic backtracking signal.
- Continued backtrack: Reader r scrolls back, and proceeds to scroll forward afterwards but does not return to the *start* location ($start > reach > end$).
- Full backtrack: Reader r scrolls back, and eventually returns all the way to the backtrack starting point ($reach \geq start$).

In order to better capture intentional backtracks and meaningful movements of readers, we only consider backtracking of at least 100 pixels in range (consisting of about a dozen lines of text), where the user spends at least 15 seconds on the page after completing the backtrack.

Dataset description. Our dataset was provided by *Chartbeat* and includes data from two popular news websites from 2015. The first website is a popular news and entertainment magazine with long-form content, which we will refer to as LONG. The second, a popular short-format news website, will be referred to as SHRT. The raw data consists of "pings," which are defined as a record of viewport information [8] associated with a single user reading session on a single page. Pings are collected every 15 seconds, or at longer time intervals if a user is inactive. Every ping contains information about the visited page in the browser by a specific user, including height, width, length of the news article and the current position of the user on the page, all measured in pixels. We keep only data pertaining to English articles and originating from desktop (non-mobile) sessions. We consider only sessions that are longer than 30 seconds, as they are more likely to be indicative of actual reading behavior.

We refer to the session where a single reader r is reading a single page p as a *read event*. For most of the analysis below, we are using a subset of the dataset with pages p that have at least 20 reading sessions which we refer to as P_{20}. We chose to use P_{20} in order to ensure we have sufficient signal of behavior for each page, and reduce the effect of outliers (We perform sensitivity analysis in Section 4 to show how this choice effects performance in a prediction task.) The number of users, pages sessions and backtracking events performed in P_{20} are shown in Table 1.

Table 1 presents statistics on the different types of backtracking for pages that have passed the initial filtering P_{20}. We see that for roughly the same number of pages and reading sessions, LONG has almost twice as many full backtracking events, but less than half the number of simple backtracking events. There are two key factors that may explain this divergence. First, on the LONG site, pages are typically longer (5,000 pixels on average) than on SHRT (3,000 pixels on average). The longer the page, the more chances it will experience a backtracking action. In addition, we observed that on SHRT, a large proportion of simple backtracks end at the top of the page. This is likely a result of a static the navigation bar at the top of the page on SHRT website, causing the user to return to the top of the page in order to navigate to another area in the site.

As shown in Table 1, the P_{20} the dataset contains 1.4M reading sessions, but far fewer backtracking events. The majority of reading sessions have no backtracking event: 78% of the sessions in the LONG dataset, and 88% of the sessions on SHRT do not include any full backtracking event.

Backtracking patterns. Figure 2 shows the typical sizes of backtracking actions using a cumulative distribution function (CDF) of sizes for the three types of backtracking, measured in pixels. Given backtracking event $b \in B$, where B is the set of all backtracking events, let $\text{size}(b) = \text{start}(b) - \text{end}(b)$. Given any size x (x-axis), the CDF is $F(x) = P(\text{size}(b) \leq x)$ for all $b \in B$ (y-axis). As the figure shows, even though the pages on LONG are significantly longer than on SHRT, the typical size of full backtracking is smaller on LONG. For example, about 80% of the full backtracks (top curve in each figure) are under 550 pixels in size in LONG, while the same is true for only 75% of the backtracks on SHRT (top curve on both figures).

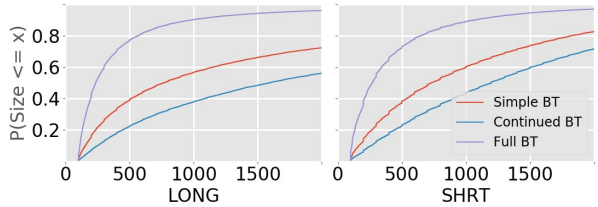

Figure 2: Backtracking sizes (in pixels) CDF

When readers backtrack, do they backtrack more than once? Figure 3 shows the number of backtrack events for sessions where there was at least one backtrack, for each type of backtracking event. For example, the right-most pane shows that 15% of LONG (red) sessions have exactly one full backtracking event. The rest of the sessions, that account for about 7% of the total number of sessions, contain two or more full backtracking events.

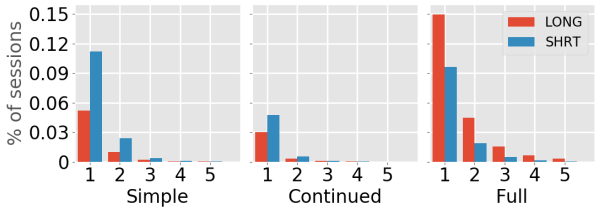

Figure 3: Number of backtracking events occurring across all sessions

We expect some pages to have more backtracking events than others, and we indeed verified that the variance of pages is much

greater than would have been be expected if the backtracking was a random event, e.g., modeled as a Bernoulli variable. We consider backtracking behavior to be close in nature to a Bernoulli, as we can see that backtracking more than once is rare and as such it is better to ask whether or not any backtracking event occurred. Figure 4 shows the page-level backtracking distribution for articles in P_{20}. The x-axis shows the backtracking-per-session average for each page, which is the number of backtracking events divided by number of reading sessions per page. The y-axis shows the number of pages with this average in each bin. For example, the right-most pane shows that, on LONG, very few pages have an average of more than 0.6 backtracking events per session. We see that for full backtracking, there is a significant difference in both variance and mean of the backtracking event distributions of the two domains.

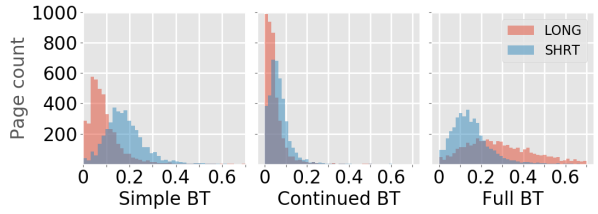

Figure 4: Are some pages more prone to backtracking? Pages' mean number of backtracking events per reading session

Are readers different in their tendency to perform backtracking? To answer this question we examine a slightly different subset of the data, U_{20}, comprised of reading sessions by users that have at least 20 reading sessions. Figure 5 shows the histogram of users' mean number of backtracks per session. Again, we see that the observed variance is much greater than if users backtracked at random, as a Bernoulli variable. The x-axis represents the average number of backtracking events per session, and the y-axis is the number of users within that bin. For example, half of LONG's readers (red) perform a full backtrack (right pane) once every three reading session on average. Compared to the page data (Figure 4), both the simple and full backtracking figures are more right-skewed, for both domains. For example, for the SHRT users (blue), there are many users that rarely perform any full backtracking.

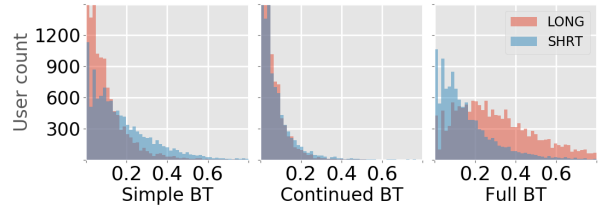

Figure 5: Do some users backtrack more frequently? Users' mean number of backtracking events per reading session

4 PREDICTING BACKTRACKING

How well can backtracking events be predicted, and what is the contributions of different features in predicting backtracking events? In this section we explore this question, also touching on the difference between types of backtracking events. The descriptive analysis suggests that LONG pages generate higher numbers of full backtracking events, which are more likely to relate to readability (as we also show below). Considering this finding, the analysis in this section is using only the LONG articles (line 1 in Table 1).

We define a binary classification task where we predict whether a page is likely to have a high or low number of backtracks per reading session. Based on the hypothesis that backtracking occurs when readers re-read a piece of text that they did not fully understand, we suggest that backtracking is an indicator of readability issues. We therefore examine several readability-related measures for texts as part of our prediction task, using a set of features inspired by previous work on readability by Feng et al. [5]. We consider four families of features at the article level:

Lexical features. These features have been shown to have strong predictive power in readability tasks, and can capture many different aspects of the text. The features include several established readability measures mentioned earlier, like the Dale-Chall score, and some simple text properties, such as average sentence length and the number of syllables per word. These features are computationally cheap to generate and do not require deep computational analysis.

Entity-density features. Entity-density features have been shown to be useful in readability prediction [5] mostly due to the fact that new concepts are often introduced in a narrative by entities (e.g. names, locations) and might impose additional burden on readers' working memory. In order to collect entities, we used spaCy[2] [15].

Part-of-speech (POS) density features. POS-based grammatical features are widely used for various tasks, including readability. In particular, [5] showed that noun-based POS features generate the highest classification accuracy in predicting correct readability grade level. In order to generate POS features, we use the spaCy POS tagger [15].

Coreference chain features. Coreference features were first studied in [5] for readability modeling, though they did not perform as well as two previous families of features in the earlier work. Our conjecture is that these features may help in representing the cognitive load on the user and should directly affect backtracking. We extract these features using the *End-to-end Neural Coreference Resolution* library[3] based on work by Lee et al. [20].

Table 2 provides more details about the features used for each family.

4.1 Defining the backtracking prediction task

Since our goal is to predict whether or not an article will have a high or low number of backtracking per reading session, we divide the training data into three equally sized sections, representing *high backtracking*, *medium backtracking*, and *low backtracking* based on the distribution shown in Figure 4. By excluding the third of the pages with *medium backtracking* and focusing on the equally sized *high backtracking* and *low backtracking* pages, we end up with

[2]https://spacy.io/
[3]https://github.com/kentonl/e2e-coref

Table 2: List of features computed for every article.

	Lexical Features
Dale-Chall score	Readability formula based on frequency of difficult words
ARI	Readability formula based on the relative length of words and sentences
Avg. syllables	Average syllables per word
SMOG grade	Readability formula based on the number of polysyllables (3 or more syllables)
Flesch-Kincaid score	Readability formula based on the ratio between sentences, words and syllables
Images	Number of images in article
Words	Number of words in article
Sentences	Number of sentences in article
	Entity-density Features
Entities	Number of named entities in article
Unique entities	Number of unique entities in article
	Coreference Features
Coreference chains	Number of coreference chains in article
Avg. coreference chain span	Average distance (in words) between words in the same chain
Avg. coreference chain size	Average number of words per chain
	POS Density Features (Nouns)
Percentage of nouns	Percentage of nouns in article (including entities) out of all tokens
Percentage of unique nouns	Percentage of unique nouns in article (including entities) out of all tokens.

balanced classes and a binary class label $y \in \{0, 1\}$, for two thirds of our training examples. We perform the same procedure for the development and test sets. We tested in preliminary experiments a set of classifiers such as SVM, logistic regression, and decision trees, we found that random forests performed best across all sets of features. Therefore, we use random forests for our analysis.

4.2 Prediction based on textual features

We examine how well the prediction performs with the full set of features, and assess the importance of the different types of features. From the P_{20} dataset, we keep only articles from LONG and split the samples into training, development and held out test sets with $75\% - 10\% - 15\%$ proportions. After filtering the *medium backtracking* pages, this dataset consists of 1,991-256-404 pages and 324,708-38,927-71,364 user sessions for training, development, and test sets. The allocation of pages to training, development and test set was determined by a random hash of the page URL. With a similar setup, but on a different subject, we perform an analysis akin to [2], including feature ablation. As mentioned before, the data for the prediction task is balanced and thus the expectation for a random prediction model is 50%. However, if we consider using the length (in pixels) of the page as the only prediction feature, the baseline to surpass is roughly 75%, in terms of accuracy.

The results, achieved on the held-out test set (and thus only computed after the work was complete) are shown in Table 3. The best performance was achieved using a random forest model[4] for full backtracking events, with an accuracy of 83.8%. We performed feature ablation by choosing a subset of features for the prediction task based on Table 2. For the feature ablation procedure, we associate each feature type with the length of the page, our control variable. We use length because, as mentioned above, longer pages provide more opportunity for backtracking, and length alone been shown to have significant predictive power for backtracking.

[4]hyper-parameters: maximum depth:10, minimum sample split:4

Table 3: Feature ablation for full backtracking prediction

Features	Accuracy	F1	AUC
All features	0.838	0.839	0.901
Lexical + length	0.824	0.826	0.889
Coreferences + length	0.802	0.806	0.861
Nouns + length	0.786	0.786	0.856
Entities + length	0.785	0.788	0.841
Length only	0.747	0.739	0.809

We notice that the lexical features hold the most predictive power, followed by coreference features. The strong performance of the coreference features, which characterize the average length and span of coreference chains in the text, is especially illuminating. These features represent a single facet of the textual context, compared to the set of different features included in the "lexical" set

When applied to the other two forms of backtracking, the complete model with all features can predict simple and continued backtracking with accuracy of 55% and 60% respectively, compared to the 85% accuracy of full backtracking prediction, the reduced performance of these models suggests that textual features are less useful in predicting simple and continued backtracking events. This follows our hypothesis that these two forms of backtracking are driven to other aspects of user actions, such as scrolling for inter-page navigation purposes.

Another aspect of the prediction task is the amount of information we have on every page. As mentioned earlier, pages with fewer than 20 views were discarded from our prediction task. We wish to know if requiring more information (a larger number of sessions) per page could improve the prediction task. We performed the training, test and prediction task on different subsets P_i of the full dataset, using a different minimum session count i for pages ($i = 1, 5, 10, 15, \ldots, 50$). The results are shown in Figure 6. The x-axis represents the minimum number of sessions per page (i), and the y-axis represents the average prediction accuracy, calculated in the same manner as defined earlier. For example, the prediction accuracy in P_{10} with the full model (red line, full circle markers) is over 80% (close to P_{20} performance), while length-only prediction on the same dataset (green line, empty circles) is 73.7%. We see that there is an increase in performance with a higher threshold for the minimum number of sessions per page, despite using fewer data points as i grows. Looking at the different sets of features, we see that all sets of features increase with a higher threshold of minimum views, but the improvement curve flattens out after $i = 20$. While we expect larger thresholds to yield better performance when more data is available, Figure 6 shows that it is sufficient to use 20 sessions per page to get enough signal for a prediction task.

4.3 Controlling for Users and Topics

We show above that the textual features of an article could help predict pages with high versus low levels of full backtracking. However,

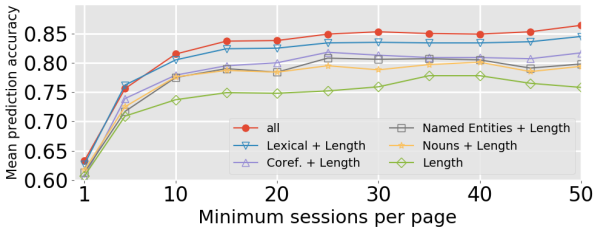

Figure 6: Prediction accuracy of different minimum sessions

we do not rule out the possibility that the text-based features are a function of the topic of the article (e.g., "health" or "sports"), or that the reader's choice of topics and articles to read is the one causing the observed differences in backtracking. For example, it could be that sports articles are easier to read or that users who choose to read sports articles tend to backtrack more often. In this section, we extend the analysis above to control for topics and users.

Topic vector. Article topics have been shown to have an effect on a user's reading behavior [7, 19]. Do topics have an effect on backtracking as well, and do they explain away the impact of any of the textual features? To examine this question, we used the set topics associated with every article as provided by LONG. We use the 35 most representative topics by removing from the dataset those that are either too rare or too frequent. From the selected set of topics, we produce a binary topic vector $T_i \in \{0,1\}^{35}$ for every page. We then add this vector as a new feature to our previously described backtracking prediction model. We do not observe any significant changes in the results, reaching +-1% in terms of accuracy, F1, and AUC compared to the results of the original model, as shown in Table 3. We also verified that the topic vector on its own is not a strong predictor of backtracking. Using the length of a page as control, a length + topic vector prediction task reaches 0.755, 0.747, 0.815 in terms accuracy, F1, and AUC — i.e. does not performs any better than a prediction based on length alone (as shown in Table 3).

Types of users. To effectively account for the effect of users choice of articles on backtracking we perform a similar prediction model to the one above, this time using the reduced-dimensionality user-page matrix as a feature in the prediction. Kallus et al [27] have shown that the use of matrix factorization to infer confounders can reduce the bias caused by measurement noise. Assuming specific users visit specific pages, we wish to learn the latent factors capturing different types of users, and further explore their significance on backtracking behavior. To this effect, we have built a binary page-user matrix $X \in \mathbb{R}^{mxn}$, where m, n represent the number of pages and users respectively from Table 1 and where X_{ij} is assigned the value 1 if user j visited page i and the value 0 otherwise.

In order to discover latent factors or low-dimensional embedding, we performed the matrix factorization $X \approx \hat{X} = UV$ or $\hat{X}_{ij} = U_i^T V_i$. Matrix $U \in \mathbb{R}^{mxk}$ holds the pages latent factors, while $V \in \mathbb{R}^{kxn}$ holds the users' latent factors. We take as a simplifying assumption that the difference between X and $\hat{X}$ is normally distributed, which allows us to use the mean-square error as our loss function.

We note as Ω the set of all observed (i, j) sessions. In the current settings, all sessions are observed, and we know for a fact whether or not a user has visited a given page. However, in order to better generalize our latent factors and account for the noise in our observed matrix, we hold out part of the of indices and perform cross-validation. This method is commonly used in collaborative filtering methods [35] in order to avoid overfitting. We add normally distributed regularization parameters to our model. This results into a new loss function, as stated below:

$$J = \sum_{(i,j)\in\Omega} (X_{ij} - U_i^T V_j)^2 + \lambda(\|U\|_F + \|V\|_F) \tag{1}$$

We use the alternative least squares [11] algorithm to minimize our loss function in an iterative way. This algorithm alternatively sets one matrix (U or V), and estimates the other by minimizing the loss function via derivation. This step is followed by adding the factor below to the latent factors' U, V matrices, iteratively until convergence:

$$\begin{cases} U_i = \left(\sum\limits_{j\in\Omega_i} v_j v_j^T + \lambda I \right)^{-1} \sum\limits_{j\in\Omega_i} (X_{ij} - \mu)\, v_j \\ V_j = \left(\sum\limits_{i\in\Omega_j} u_i u_i^T + \lambda I \right)^{-1} \sum\limits_{i\in\Omega_j} (X_{ij} - \mu)\, u_i \end{cases} \tag{2}$$

where Ω_i is the set of all users that visited page i, and Ω_j is the set of all pages that user j visited.

Similarly to the topics vector, adding the latent factors of pages to the prediction model showed a slight degradation in the prediction result on the held-out set, reaching 81.2%, 81.3%, 86.7% for accuracy, F1 score and AUC. In addition, the low-dimensional factors (with length of page as control) perform similar to our length-only baseline and reach 75.3%, 75.2%, 79.2% for accuracy, F1 score and AUC respectively.

5 CONCLUSION AND FUTURE WORK

This paper examines an under-explored signal, backtracking, and what it can teach us about the content of online news articles. We define and explore different types of backtracking actions on Web-based articles, showing that different news sites exhibit different backtracking patterns. We also demonstrate that full backtracking events can be accurately predicted with textual features related to readability assessment. By controlling for readers and topics, we discover that backtracking behavior is driven primarily by the textual content of an article. Our findings open a promising opportunity towards revisiting readability using large-scale web signals.

Our work has several limitations, related to data size and scope. Future work can extend the current work by analyzing a larger news corpora, or different content types, such as e-books. While our work only examined content from desktop interactions, understanding backtracking behavior across both desktop and mobile devices could expose different types of behaviors and perhaps more detailed understanding of backtracking. Finally, future work can leverage the insights about language's impact on readability and backtracking behavior. Such lines of work include supporting authors and editors in crafting content and helping struggling readers navigate text with language features known to hinder reading, such as coreferences. In both cases, the impact of these interventions could be measured at scale using the backtracking signal.

6 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1840751, and by Yahoo! Research through the Connected Experiences Lab at Cornell Tech. We thank the Chartbeat data science team for generosity with their time and their data.

REFERENCES

[1] Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability Assessment for Text Simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications (IUNLPBEA '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–9.

[2] Justin Cheng, Lada Adamic, Alex Dow, Jon Kleinberg, and Jure Lescovec. 2014. Can Cascades be Predicted?. In *Proceedings of the 23rd international conference on World wide web (WWW '14)*. ACM, New York, NY, USA, 925–936.

[3] William H, DuBay. 2007. The Classic Readability Studies. *Online Submission* (2007).

[4] Jeanne S Chall, Edgar Dale. 1948. A formula for predicting readability: Instructions. *Educational research bulletin* (1948), 37–54.

[5] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 276–284.

[6] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.

[7] Nir Grinberg. 2018. Identifying Modes of User Engagement with Online News and Their Relationship to Information Gain in Text. In *Proceedings of the 2018 Web Conference (WWW '18)*. ACM, New York, NY, USA.

[8] Max Grusky, Jeiran Jahani, Josh Schwartz, Dan Valente, Yoav Artzi, and Mor Naaman. 2017. Modeling Sub-Document Attention Using Viewport Time. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 6475–6480.

[9] Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication* 6, 2 (1969), 3–13.

[10] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. 2013. Mining Touch Interaction Data on Mobile Devices to Predict Web Search Result Relevance. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)*. ACM, New York, NY, USA, 153–162.

[11] Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. 2015. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research* (2015), 3367–3402.

[12] Michael Heilman, Kevyn Collins-Thompson, James P. Callan, and Maxine Eskénazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *HLT-NAACL*.

[13] Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (EANL '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 71–79.

[14] Rosie Jones Henry A. Feild, James Allan. 2010. Predicting Searcher Frustration. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM, New York, NY, USA, 34–41.

[15] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

[16] Thorsten Joachims and Laura Granka. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of the 28 International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. ACM, New York, NY, USA.

[17] Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

[18] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Technical Report. Naval Technical Training Command Millington TN Research Branch.

[19] Dmitry Lagun and Mounia Lalmas. 2016. Understanding User Attention and Engagement in Online News Reading. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. ACM, New York, NY, USA, 113–122.

[20] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

[21] Annie Louis and Ani Nenkova. 2013. What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *TACL* 1 (2013), 341–352.

[22] Keith Rayner, Lyn Frazier. 1982. Making and Correcting Errors during Sentence Comprehension: Eye Movements in the Analysis of Structurally Ambiguous Sentences. *Cognitive Psychology* (1982).

[23] Patricia A. Carpenter, Marcel Adam Just. 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological review* (1980).

[24] Pekka Niemi, Marja Vauras, Jukka Hyona. 1992. Comprehending coherent and incoherent texts: evidence from eye movement patterns and recall performance. *Journal of Research in reading* (1992).

[25] G. Harry, Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading* 12, 8 (1969), 639–646.

[26] Ta Lin Liau, Meri Colema. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60, 2 (1975), 283.

[27] Madeleine Udell, Nathan Kallus, Xiaojie Mao. [n. d.]. Causal Inference with Noisy and Missing Covariates via Matrix Factorization. In *arXiv:1806.00811*.

[28] Sarah E. Petersen and Mari Ostendorf. 2009. A Machine Learning Approach to Reading Level Assessment. *Comput. Speech Lang.* 23, 1 (Jan. 2009), 89–106.

[29] Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. *Int. J. Comput. Linguistics Appl.* 7 (2016), 143–159.

[30] Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 186–195.

[31] Keith Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin* (1998).

[32] Keith Rayner, Kathryn H. Chace, Timothy J. Slattery, and Jane Ashby. 2006. Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading* (2006).

[33] Mirella Lapata, Regina Barzilay. 2008. Modeling Local Coherence: An Entity-based Approach. *Comput. Linguist.* 34, 1 (March 2008), 1–34.

[34] Edgar A Smith and RJ Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)* (1967), 1–14.

[35] Chris Volinsky, Yehuda Koren, Robert Bell. 1982. Matrix Factorization Techniques for Recommender Systems. *Computer* (1982), 30–37.

[36] Chris Volinsky, Yehuda Koren, Robert Bell. 2009. Eye movements when viewing advertisements. *Journal of frontiers in psychology* 43 (2009). Issue 8.