Predicting User Views in Online News

Daniel Hardt
Copenhagen Business School
dh.itm@cbs.dk

Owen Rambow
Columbia University
rambow@ccls.columbia.edu

Abstract

We analyze user viewing behavior on an online news site. We collect data from 64,000 news articles, and use text features to predict frequency of user views. We compare predictiveness of the headline and "teaser" (viewed before clicking) and the body (viewed after clicking). Both are predictive of clicking behavior, with the full article text being most predictive.

1 Introduction

With so much news being consumed online, there is great interest in the way this news is consumed – what articles do users click on, and why? The data generated in online news consumption constitutes a rich resource for the exploration of news content and its relation to user opinions and behaviors. There are undoubtedly a wide variety of factors that influence reading behavior at online news sights, including the visual presentation of the web site. But certainly the language seen by the user plays a central role.

In this paper we experiment with a dataset from the online news site of Jyllands-Posten, a major Danish newspaper.[1] The data consists both of user logs and news articles. We attempt to predict viewing behavior from the text of articles. We also look at the difference in predictiveness of the text the user sees before clicking, i.e., the headline and the teaser, vs. the body of the article, which the user only sees after clicking, vs. the complete text of the article,

The first question we address is whether a simple lexical representation of articles is predictive of viewer behavior. We investigate bag of words, word vectors, and article length. A second question we investigate is the relative predictiveness of the headline and teaser, which are displayed before clicking, and the body of the article, which is of course only seen after the decision to view.

We explore these questions because we see them as relevant to a fundamental issue in today's media landscape: to what extent are news consumers manipulated by "clickbait", as opposed to making informed decisions about what news to consume? While the term clickbait is difficult to define, we see it as highlighting a potential difference between the promise of a headline or teaser compared to the actual nature of the article being pointed to. The work discussed in this paper is part of an effort (see for example (Blom and Hansen, 2015)) to use large amounts of data and computational methods to understand "clickbait".

2 Data

Our dataset consists of news articles and user logs from the online portal of the Danish daily, Jyllands-Posten. User logs have been maintained since July 2015. An entry in the user logs is created each time a user clicks on a new article on the site. An entry includes the time of the click and the page ID of the article, as well as a user ID if the user is registered on the site. It also includes additional information, including the referring page – the page the user was viewing when they clicked on the article. We collected all online articles published since July 2015, a total of 64,401 articles. The log file includes a total of 213,972,804 article views.

Articles are linked from two types of pages on the Jyllands-Posten website: the start page, and specialized pages for different subject matters (domestic, international, culture, sports, local, etc.).

Jyllands-Posten is a mainstream Danish daily paper, covering all major news topics,with a somewhat right-of center slant. While we have not an-

[1] http://jyllands-posten.dk

alyzed the distribution of topics covered, table 1 gives the most frequent unigrams in the training data, with stopwords manually removed. This listing reveals a focus on immigrants and other perceived external threats to Denmark.

1136	Denmark
652	Danish
633	refugees
618	EU
580	Aarhus (home town of paper)
552	USA
535	Danish
472	Løkke (prime minister)
444	killed
422	Danish
419	DF (Danish anti-immigrant party)
367	police
361	Syria
339	victory
339	death
331	satire
330	Trump
325	children
323	dead
316	Danish
315	Turkey
306	Europe
301	Russia
300	Islamic
286	attack

Table 1: Most frequent words (translated from Danish, note that inflectional variants in Danish of the word *Dansk* 'Danish' result in *Danish* appearing multiple times)

The articles consist of three distinct parts:

- The **headline** of the article, which is the text always displayed as the clickable link to the article on the referring page. It is also repeated on the article page.
- On the article page, there is typically a phrase or short sentence displayed below the headline, called the **teaser**. On the referring page, the teaser is sometimes omitted. We do not have information on whether the teaser was present or not on the referring page.
- The **body** is the text of the actual article, which is only visible after the user has clicked on the headline text.

The text data (headline, teaser, and body) is divided into training and development data, as described in Table 2. (We have a held out test set which we will use in future publications.)

Dataset	**Articles**	**Words**
Train	55,061	25,745,832
Development	9,351	4,134,432

Table 2: Text Data: Articles

The average number of views is 3,337, and the median number of views is 795. See Table 3 for the two most viewed headline/teaser combinations, and Table 4 for a headline/teaser with a median number of views (translations from Danish by the authors). There are evident differences between the high and median examples: the highly viewed example deals with material of immediate relevance to many readers. The top example concerns a garden snail that has preoccupied Danish gardeners for years, and promises a new solution. The second concerns a beloved Danish TV Christmas program, in which some off-color language was clearly visible during the children's program. The language used is also more conversational, informal and extreme. By contrast, the median example is purely informative.

H	Watch the unfortunate mistake in TV 2's family Christmas calendar
T	An attentive viewer caught the writing on the board, which the children probably should not see.
H	See the surprising solution in the fight against the killer snail
T	Nature guide in Herning has made a groundbreaking discovery that benefits all garden owners.

Table 3: Headline (H)/Teaser(T) for the articles with the most views (671,480 and 334,820, respectively)

3 The Task: Predicting Clicks Based on Text

Our task is to predict which articles get the most user views. We bin the articles by numbers of clicks into 2, 3, and 4 bins. This defines three different classification tasks: is the article in the

H	International agreement: Elections in East Ukraine this summer
T	The goal is to hold local elections in Donetsk and Lugansk before August. Germany and Ukraine are skeptical.

Table 4: Headline (H)/Teaser(T) for an article with median views (795 views)

top 50% of clicks, in the top 33.3% of clicks, in the top 25% of clicks? We use different parts of the article text to make the prediction. Specifically, we ask how much each of the text elements (headline, teaser, body) contributes to our ability to predict the highly clicked articles. Our working hypothesis is that the headline on its own, or the headline with the teaser, should have higher predictive power than the article alone. This is because the user sees only the headline (and perhaps the teaser) before making the decision to click and read the article. We investigate the following combinations of text elements, to see which provides the most predictive power:

- Headline only: the reader definitely sees this before clicking.
- Headline and teaser: in most cases, the user also sees a teaser before clicking.
- Body only: the reader does not see the body before clicking.
- Full article (headline, teaser, body): the reader sees all this information together only after clicking.

We experiment with the following classifiers, all using the sklearn package: Support Vector Machines with a linear kernel, Logistic Regression (logreg), Random Forests. For all classifiers, we use the same set of features. For the initial experiments we report in this workshop paper, we use the following set of lexical features:

- Bag of Words (BoW): We construct a bag of words from each article represented as a vector whose size is that of the vocabulary. We experiment with three values: a count of occurrences, a weighted count (term frequency), and tf-idf values.
- Word Vectors (vec): We also use word vector features for each word in each article (Mikolov et al., 2013a,b). These vectors were created using the Python gensim package, using all of the training data. We then form the mean of the word vectors for all words in the text component we are interested in (headline, teaser, or body).
- Text length (wc): the length in words.

4 Results

We found consistently that logistic regression outperforms the other classifiers; we therefore only present results using logreg. Furthermore, we found that term frequency and tf-idf consistently perform about equally, and both outperform simple counts; thus, we report only results using term frequency. These results are shown in Tables 5, 6, and 7 for the top 50%, top 33.3% and top 25% classification tasks, respectively. We provide accuracy results and f-measure results, but we take the f-measure results as the relevant result. The baselines are always choosing the top-clicked category.

We observe that the models consistently beat the baselines (both on accuracy and f-measure). The text features thus are, in general, predictive of users' viewing behavior. Furthermore, we observe across the three tasks that the performance increases from using only the headline to using headline and teaser to using only the body to using the whole article. Put differently, more text is better for this prediction task, contrary to our hypothesis that the body would not contribute predictive power as it is unseen at click time.

In terms of our features, we were surprised to see that the wc (text length) and vec (word vectors) features do not appear to have much effect. While the results for different feature combinations vary somewhat, we do not see variations greater than 0.7% (and usually much less) in the 12 separate experiments (3 tasks and 4 data sources). The one exception is using the body for finding the top 33.3% of clicked articles (Table 6), where the combination of bag of words and word count leads to a drop of 3% over the other feature combinations. We take this to be noise rather than an interesting result.

5 Discussion

Our initial hypothesis was that article body would not be as predictive as headline and particularly

		Accuracy		F-measure				
							Always-H Bl	
Data Source	Feats	Acc	Bl	Recall	Precision	F-m	Prec	F-m
Headline	bow	0.612	0.513	0.856	0.583	0.694	0.513	0.678
Headline	bow, wc	0.611	0.513	0.856	0.582	0.693	0.513	0.678
Headline	bow, vec, wc	0.612	0.513	0.855	0.583	0.693	0.513	0.678
HeadlineTeaser	bow, wc	0.630	0.513	0.847	0.599	0.701	0.513	0.678
HeadlineTeaser	bow, vec, wc	0.629	0.513	0.847	0.598	0.701	0.513	0.678
HeadlineTeaser	bow	0.627	0.513	0.850	0.596	0.700	0.513	0.678
Body	bow, wc, vec	0.652	0.513	0.907	0.607	0.727	0.513	0.678
Body	bow, wc	0.640	0.513	0.92	0.597	0.724	0.513	0.678
Body	bow	0.650	0.513	0.889	0.609	0.722	0.513	0.678
HeadlineTeaserBody	bow, wc, vec	0.664	0.513	0.891	0.620	0.731	0.513	0.678
HeadlineTeaserBody	bow	0.670	0.513	0.875	0.627	0.731	0.513	0.678
HeadlineTeaserBody	bow, wc	0.662	0.513	0.895	0.618	0.731	0.513	0.678

Table 5: Results for finding the top-clicked 50% of articles using logistic regression

		Accuracy		F-measure				
							Always-H Bl	
Data Source	Feats	Acc	Bl	Recall	Precision	F-m	Prec	F-m
Headline	bow, wc	0.470	0.355	0.743	0.450	0.560	0.337	0.504
Headline	bow, vec, wc	0.469	0.355	0.740	0.451	0.560	0.337	0.504
Headline	bow	0.467	0.355	0.739	0.448	0.558	0.337	0.504
HeadlineTeaser	bow, vec, wc	0.480	0.355	0.751	0.471	0.579	0.337	0.504
HeadlineTeaser	bow, wc	0.479	0.355	0.752	0.470	0.578	0.337	0.504
HeadlineTeaser	bow	0.474	0.355	0.755	0.464	0.575	0.337	0.504
Body	bow	0.498	0.355	0.793	0.484	0.601	0.337	0.504
Body	bow, wc, vec	0.499	0.355	0.860	0.458	0.597	0.337	0.504
Body	bow, wc	0.446	0.355	0.939	0.407	0.568	0.337	0.504
HeadlineTeaserBody	bow	0.517	0.355	0.813	0.504	0.622	0.337	0.504
HeadlineTeaserBody	bow, vec, wc							
HeadlineTeaserBody	bow, wc							

Table 6: Results for finding the top-clicked 33.3% of articles using logistic regression (some numbers missing for uninteresting reasons)

		Accuracy		F-measure				
							Always-H Bl	
Data Source	Feats	Acc	Bl	Recall	Precision	F-m	Prec	F-m
Headline	bow, wc	0.363	0.271	0.673	0.357	0.466	0.242	0.390
Headline	bow, vec, wc	0.363	0.271	0.672	0.355	0.465	0.242	0.390
Headline	bow	0.361	0.271	0.665	0.351	0.46	0.242	0.390
HeadlineTeaser	bow, wc	0.370	0.271	0.659	0.368	0.473	0.242	0.390
HeadlineTeaser	bow, vec, wc	0.371	0.271	0.659	0.368	0.472	0.242	0.390
HeadlineTeaser	bow	0.369	0.271	0.662	0.363	0.469	0.242	0.390
Body	bow, wc, vec	0.424	0.271	0.757	0.401	0.525	0.242	0.390
Body	bow, wc	0.419	0.271	0.755	0.399	0.522	0.242	0.390
Body	bow	0.401	0.271	0.763	0.392	0.518	0.242	0.390
HeadlineTeaserBody	bow, wc	0.421	0.271	0.760	0.406	0.529	0.242	0.390
HeadlineTeaserBody	bow, wc, vec	0.421	0.271	0.761	0.406	0.529	0.242	0.390
HeadlineTeaserBody	bow	0.42	0.271	0.765	0.404	0.529	0.242	0.390

Table 7: Results for finding the top-clicked 25% of articles using logistic regression

teaser, since teaser is presumably constructed to induce clicking behaviors, while the article text itself is not visible to the user at the time a clicking decision is made. Thus we find it quite surprising that body is more predictive than headline and teaser, and the model combining headline, teaser and body is the best.

How can it be that the body is more predictive than the text the user actually sees when deciding to click? Here we offer some hypotheses. First, we note that some clicks are the result of social media referrals (this information is present in our log data). In these cases, it makes sense that body data is predictive, since presumably the referrer read the article before making the referral. Second, it is possible that the headline on its own gives readers a lot of semantic information which we are not capturing with our features, but which the whole article does provide. So human readers can "imagine" the article before they read it and implicitly base their behavior on their expectation.

In general, although the bow features are consistently predictive, there is little or no improvement from the vec and wc features. We expected that wc (text length) might be relevant in some ways: for example, that short, punchy teasers might tend to be more effective. No such effect has been observed however. The vec (word embeddings) feature was used to compute an average vector for the entire text. Computing an average of word vectors has been shown effective in other document classification tasks (Alkhreyf and Rambow, 2017). However, clearly such a vector loses a lot of information about a text, and more fine-grained modeling is needed.

6 Plans for Future Work

This work lays the foundation for multi-faceted investigations of news data, language, and user behavior and preferences. We have extracted aggregate totals of article views from the user logs. This dataset, which includes logs of all user behavior since 2015, has rich potential for further data mining. For example, the logs include the referring page for each user view. We intend to produce separate models for views resulting from social media referrals. Our hypothesis is that the body of the article is (even) more predictive in these cases, since the decision to view is, indirectly, based on a reading of the body of the article. We also intend to mine the logs to divide users into different classes based on their reading behavior. In addition, we plan to examine further our use of word embeddings, to explore ways in which they could be better exploited for prediction of views. We will also experiment with topic modeling.

Ultimately, we seek to shed some light on basic questions about online news. In particular, we would like to characterize the nature of different text types in headlines, teasers and article bodies, and in the process to use NLP techniques to help explore the difference between clickbait and genuine journalistic quality.

Acknowledgments

We thank A. Michele Colombo and Ha Le Hgoc for help with the data and experiments. We also thank Jyllands-Posten for giving us access to the data.

References

Sakhar Alkhreyf and Owen Rambow. 2017. Work hard, play hard: Email classification on the Avocado and Enron corpora. In *Proceedings of Textgraphs-11, ACL Workshop*.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics* 76:87 – 100.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Hlt-naacl*. volume 13, pages 746–751.