EFFECTIVE HEADLINES OF NEWSPAPER ARTICLES IN A DIGITAL ENVIRONMENT

Jeffrey Kuiken, Anne Schuth, Martijn Spitters and **Maarten Marx**

We study the effect on click-through rates of applying textual and stylistic features often related to clickbait to headlines of newspaper articles which can be bought in a digital environment. Having a dataset consisting of triples—original headline, rewritten headline, CTR, where CTR is the click-through rate of the rewritten headline in a newsletter from the online kiosk Blendle—we can directly measure whether these "clickbait features" do what they are believed to do: entice readers to click on them. The main findings are as follows. First, the data shows that editors of Blendle indeed often use clickbait features when rewriting headlines. Second, most, but not all, of the clickbait features lead to a statistically significant increase in the number of clicks. Third, predicting the effectiveness of a headline only on the basis of its clickbait features is not possible. The data on which this article is based is publicly available online.

KEYWORDS clickbait; click-through rate; digital newsletters; effectiveness; headlines; newspaper articles

Introduction

The way people consume newspaper articles is changing: more and more newspaper articles are consumed on the internet rather than from physical newspapers. People used to buy a newspaper, read it from cover to cover while scanning headlines, and reading articles that they thought were interesting (Holmqvist et al. 2003). However, increasingly more people are reading individual articles online, outside of their original publication. Often, a person reads this article because it was shared on social media or some other internet platform (Baresch et al. 2011; Hermida et al. 2012).

With this change, the function of the headline of a news article changed as well. Previously, the primary function of a headline was to give the reader, who was scanning the newspaper, a clear understanding of what the article was about (Van Dijk 1988). But since many headlines are not read within the context of a newspaper anymore, the function of the headline has shifted. The headline, being one of the primary ways to attract the readers' attention, should above all make the reader curious as to what the article is about, so that it *lures* the reader into opening the article (Chen, Conroy, and Rubin 2015).

Therefore, it is important to have a good understanding of the characteristics of an effective headline. In this paper, we answer the following research question:

Digital Journalism, 2017
http://dx.doi.org/10.1080/21670811.2017.1279978
© 2017 Informa UK Limited, trading as Taylor & Francis Group

RQ1: What are the characteristics of an effective headline of a news article in a digital environment?

We answer this question by means of an analysis of data from Blendle,[1] an online kiosk. Blendle is a Dutch[2] online service that combines articles from all major Dutch newspapers and magazines. Users can buy individual articles without having to purchase the full publication. There are two ways in which a user interacts with Blendle. Blendle offers an application for all major mobile platforms and is also available as a website in which users can freely browse through all articles and publications. This can be viewed as a more or less direct translation of the traditional kiosk to an online setting. Users of Blendle can also subscribe to a daily newsletter. This newsletter contains links to approximately 12 articles which are selected by Blendle's editors. In the newsletter, each article is represented by a headline, a short introduction, and sometimes a picture from the article. This headline and introduction are created by Blendle's editors, based on the original headline and article. These rewritten headlines are the focus of this research.

We thus have a unique dataset consisting of triples: original headline, rewritten headline, and click-through rate (CTR). We will then abstract away from these concrete instances by representing each headline as a list of values on a set of textual features, which are all related to the clickbait phenomenon (e.g., whether the headline is phrased as a question). Then we compare the features of the original headlines to those of the rewritten headlines using statistical tests. Features which often change are then hypothesized to have a positive effect of the CTR.

Specifically, we look at three research subquestions:

RQ1a: How are headlines rewritten by the editorial team of Blendle?

Because the rewritten headlines are created for a digital environment, they are often vastly different from the original headlines, which were meant for physical newspapers. In the early days of Blendle, the original headlines were used in the newsletter. However, once Blendle started rewriting the headlines, CTRs increased drastically, giving reason to assume that the rewritten headlines are a better fit for a digital environment than the original headlines. Since then, the editorial team of Blendle has further developed their intuition of how to (re)write headlines for a digital environment by continuously analyzing how well their headlines perform. By analyzing *how* the rewritten headlines differ from the originals, we can gain a better understanding of the differences between headlines written for a newspaper and for online, and, by extension, of what the editorial team thinks is a good headline for a digital environment.

RQ1b: What features of a headline have a significant impact on the performance of that headline?

Using the CTR of newsletter items (controlling for the rank of a headline in the newsletter), we define the *performance* of headlines and test whether headlines with a certain feature perform significantly better than headlines without that feature. This is how we test the hypotheses that were formed as a result of RQ1.

RQ1c: Is it possible to effectively predict the performance of a headline using the discovered features?

The first two subquestions are primarily aimed at answering RQ1, and the third subquestion is aimed at finding a practical use for the newly found knowledge.

This paper is different from previous work in the following ways: while the clickbait phenomenon itself has been described in research before, its actual impact on online journalism has not. Furthermore, most research on newspaper headlines has been of a qualitative nature, or small-scale quantitative research. It used to be challenging, if not impossible, to gather data regarding the effectiveness of headlines. As mentioned before, it has only recently become possible to obtain the data needed to carry out a large-scale quantitative study on this matter. As far as we know, our work is the first to research the impact of clickbait on online journalism by performing a large-scale quantitative analysis.

Even though the analysis in this study was performed with Dutch headlines and Dutch Blendle users only, we believe that the results are more generally applicable to other languages and cultures as well. Dutch is, like German and English, a West Germanic language (Gray and Atkinson 2003), and shows many similarities to those languages. More importantly, we have no reason to assume that Dutch users of Blendle are outliers from a cultural point of view. This is further backed by an early and explorative look at the data of Blendle users outside the Netherlands.

Our main findings are: the rewritten headlines differ vastly from their original counterparts. These differences are significant for all the features that were tested. Some of those features, including the use of personal and possessive pronouns, negative words, and questions were statistically proven to have a significant impact on the performance of a headline. However, models to predict the performance of a headline did not have a high accuracy; none of the trained classification and regression models achieved significantly better results than the baseline. This suggests that it is a difficult task to train a model for this purpose.

The paper is organized as follows. The next section reviews the existing literature regarding headlines of newspaper articles as well as literature on the clickbait phenomenon. Then the methodology is discussed, including a detailed description of the data used, and the final section presents the findings for all three subquestions.

The data on which this article is based is available online.[3].

Related Work

In the literature, the headline of a newspaper article has primarily been given two distinct functions. The first is to summarize the article it belongs to (Van Dijk 1988). It can do so by either being an abstract of the full article, or by highlighting the main point of that article (Bell 1991; Nir 1993). Dor (2003) calls headlines *relevance optimizers*, based on the relevance theory of Sperber and Wilson (1986). Dor (2003) notes that headlines “are designed to optimize the relevance of their stories for their readers.” (696) He argues that headlines require a balance between being short and clear, and being an information-rich summary of the article. An optimum between those two goals has to be found.

The second function of a headline is a more pragmatic one (Iarovici and Amel 1989). It is the function to attract the attention of the readers and to provoke them to read the article (Bell 1991; Nir 1993). Ifantidou (2009) showed that readers actually preferred headlines that are creative, even if that makes a headline longer, more confusing or less informative; he states that “readers seem to value headlines for what they are, i.e. underinformative, creative, yet autonomous texts.” (717).

On the internet, there is much more competition between news sources for the readers' attention (Chen, Conroy, and Rubin 2015). As more and more readers of news articles come from social media networks such as Facebook and Twitter (Mitchell and Page 2015), the need for a good headline that delivers the most *clicks* grows. This often leads to a vague headline that induces curiosity, which is then used to *lure* readers into clicking on the headline. This phenomenon is known as *clickbait*.

A single, concise, definition of clickbait cannot be found in the literature. However, many of the techniques often used for clickbait have been described and investigated. Therefore, clickbait might be best considered as an umbrella term, used to describe many different techniques, all with the common goal of attracting attention and invoking curiosity to get the reader to click on a headline.

Simplification, spectacularization, negativity, and provoking content are characteristics that are often related to clickbait (Blom and Hansen 2015; Rowe 2011; Tenenboim and Cohen 2015). Another stylistic feature that is used in many clickbait headlines is *forward referencing* (Blom and Hansen 2015), which is referring to something that is mentioned in the article. Often signal words like "this," "why" or "what" are used for forward referencing. The use of questions (Lai and Farbrot 2014; Tenenboim and Cohen 2015) and numbers (Safran 2013) have also been linked to clickbait headlines.

On the internet, it is much easier to track user interaction and behavior (Atterer, Wnuk, and Schmidt 2006) than with physical newspapers. The role of these metrics for online journalism is significant, as they have become determining factors in news production. Decisions are less based on *instinct* and more on actual data (Anderson 2011; MacGregor 2007). Lee and Lewis (2012) have shown a statistically significant influence of these data on the decisions that editors make.

Headline writing is influenced by the availability of more data as well. Both Dick (2011) and Tandoc (2014) have shown that, in attempting to attract more readers to their stories, editors and journalists have been changing the way they write headlines for their articles, by using words, phrases, and stylistic techniques that are known to perform well and attract more clicks.

Methodology

In this section, we describe the used data and the list of features used to describe headlines. We define the crucial notion of the *performance* of a headline, and we present the methods used for answering our three research subquestions.

Description of the Data

The dataset used for this analysis consisted of 1836 items originating from all 202 newsletters of Blendle which were sent between August 2015 and April 2016. Five articles appeared in more than one newsletter, and were filtered out so that each unique article appeared only once. For three newsletter items, no information on the original article was available. These headlines were also left out.The remaining 1828 headlines were used for the comparison between the original and rewritten headlines. For 189 articles, the data proved to be incomplete, which made it impossible to reliably calculate the performance of those articles. These articles were not used for analyzing headline performance. The headlines, however, were still used to measure and describe the headline characteristics.

Each newsletter consists of a number of items (with each item referring to one article) that originate from a pool of articles that were selected by the editors of Blendle. Each subscriber of the newsletter receives a personalized version of the newsletter. The selection of articles that appeared in the newsletter and the order in which they appear are based on the interests of the user.

For each newsletter, the following information was available:

1. General metadata, such as the time the newsletter was sent and the total number of recipients.
2. For each recipient, whether or not they opened the newsletter.
3. For each recipient, all the items included in the newsletter, in the order they appeared.
4. For each recipient, the items they clicked on.

For each newsletter item, the following information was available:

1. The headline, as it appeared in the newsletter, which was rewritten by Blendle's editorial team. From now on this will be called the *rewritten headline*.
2. The newsletters in which the item appeared.
3. The headline as it appeared in the original publication and in the various applications of Blendle. From now on this will be called the *original headline*.

With this information, both the original and rewritten headline were available for each newsletter item, as well as enough information to calculate the CTR of that item. This CTR is used to calculate the performance of a headline.

To illustrate the differences between the original and the rewritten headline, Table 1 contains four of these pairs. We give the original Dutch headlines together with their English translations. This small selection demonstrates how much a headline can change, as seen in examples 1 and 4. Sometimes, however, the changes are smaller, as example 2 demonstrates. Here only a few words are replaced, most notably the addition of the signal word "*wat*" (what) as the first word of the headline.[4] The addition of a signal word can be seen in examples 3 and 4 as well, with the addition of "*waarom*" (why). The addition of "*wat*" and "*waarom*" can be seen as forward referencing. Another noteworthy change is the addition of a name and age ("*Boyan (21)*") in the rewritten headline of example 1. The rewritten headline of example 3 is more provoking and negative than its original counterpart. As seen in the literature review, many of these observed changes are linked to the clickbait phenomenon.

Headline features. For each headline we extracted a set of features. These features are listed and described in Table 2. The features were selected based on existing literature regarding clickbait. For each newsletter item, these features were computed for both the original and rewritten headline.

The features of each headline were automatically extracted using self-written Python scripts. The *Pattern* library was used for natural language processing, such as part-of-speech tagging (De Smedt and Daelemans 2012). Furthermore, the *Dutch language subjectivity lexicon* by Jijkoun and Hofmann (2008) was used to identify sentimental words.

TABLE 1

A sample of four original–rewritten headline pairs (both English translations and the original headlines are given for each pair)

Example	Original headline	Rewritten headline
1	"Historically, more things did work out than did not" *"In de geschiedenis zijn meer dingen wel gelukt dan niet"*	Boyan (21) developed from student with an idea to executive of a big enterprise *Van scholier met een idee groeide Boyan (21) uit tot topman van een miljoenenbedrijf*
2	Donald Trump still can learn quite a lot from Frank Underwood *Donald Trump kan nog heel wat leren van Frank Underwood*	What Donald Trump still can learn from Frank Underwood *Wat Donald Trump nog kan leren van Frank Underwood*
3	How predictable is the weather? *Hoe voorspelbaar is het weer?*	Why is the weather forecast so often wrong? *Waarom zit het weerbericht er zo vaak naast?*
4	A hero does not think, he does *Een held denkt niet, die doet*	Why two Dutch marines received high awards for their exceptional courage *Waarom twee Nederlandse mariniers hoge onderscheidingen kregen voor uitzonderlijke moed*

TABLE 2

The features that were extracted from headlines

Feature	Description
Number of characters	The number of characters in the headline, including spaces and punctuation
Number of words	The number of words in the headline
Average word length	The length of all the words divided by the number of words
Number of sentences	The number of sentences in the headline
Number of sentimental words	The number of sentimental words in the headline, split up by positive and negative word counts. This uses the *Dutch language subjectivity lexicon* by Jijkoun and Hofmann (2008)
Readability score	The *Flesch Reading Ease* score of the headline. This is a number between 0.0 and 1.0, where a higher score means easier to understand (Kincaid et al. 1975)
Containing question	Whether or not the headline contains a question
Containing quote	Whether or not the headline contains a quote. A difference is being made between a full quote where the quote spans the entire headline, and a partial quote where the quote only spans part of the headline.
Containing signal words	The number of occurrences for each of the following signal words: *daarom, deze, dit, hierdoor, hierom, hoe, waarom, wanneer, welke, wie*, and *zo* [directly translated as respectively *hence, this, this, therefore, therefore, how, why, when, which, who*, and *like that*]
Containing pronouns	The number of personal and possessive pronouns that are used. This is split up into first, second, or third person, and into singular and plural
Containing number	Whether or not the headline contains a number. Only numerical representations (like "52") are counted, written out numbers (like "fifty-two") were not
First word type	The word type of the first word in the headline. The *Pattern* Python library (De Smedt and Daelemans 2012) was used for word tagging. *Pattern* uses the *Page Treebank tagset* (Marcus, Marcinkiewicz, and Santorini 1993)

Headline performance. For this research, we say that the effectiveness of a headline is related to the number of people that clicked on a headline. In general, and for the Blendle newsletter in particular, the aim of a headline is to provoke readers to click on the headline and read the article. In other words, the more people click, the more people were triggered by the headline, and the more effective it was. Therefore, we will base the performance of a headline on the CTR of its newsletter item. The CTR is the number of people that clicked on an item divided by the number of people that have seen that item. The CTR is a commonly used metric to evaluate performance on the internet (e.g. König, Gamon, and Wu 2009; Richardson, Dominowska, and Ragno 2007).

However, the performance of a headline cannot be equal to the CTR of its newsletter item, because the CTR also depends on the rank of the item in the newsletter and the number of items in the newsletter. In order to compensate for this effect, we define the performance of a headline as the log-ratio of the actual and expected number of clicks.

We first define the average CTR of items at a specific position *pos* in a newsletter of length *len*:

$$CTR(pos, len) = \frac{Clicks(pos, len)}{Newsletters(len)}.$$

Here *Clicks*(*pos*, *len*) is the total number of clicks on items at position *pos* in a newsletter of length *len* and *Newsletters*(*len*) is the number of opened newsletters of length *len*. We can lift this notion to the level of a newsletter of length *l* by simply adding the average CTRs of all positions:

$$CTR(newsletter) = CTR(1, l) + CTR(2, l) + \ldots + CTR(l, l).$$

Given a headline *h*, let *Occ*(*h*) be the set of all newsletters containing h. The position of headline *h* in a newsletter $n \in Occ(h)$ is denoted by *pos*(*h*, *n*).

We can now define the number of expected clicks ($Clicks_{exp}$) and the number of actual clicks ($Clicks_{act}$) of a headline *h*:

$$Clicks_{exp}(h) = \sum_{n \in Occ(h)} \frac{CTR(pos(h, n), len(n))}{CTR(n)}$$

$$Clicks_{act}(h) = \sum_{n \in Occ(h)} \frac{Clicks(h, n)}{CTR(n)}$$

where *Clicks*(*h*, *n*) is the total number of clicks on headline *h* occurring in newsletter *n*. Finally, we define the performance of a headline *h* as

$$Performance(h) = log_2(\frac{Clicks_{act}(h) + 1}{Clicks_{exp}(h) + 1}).$$

To avoid cases where the logarithm is not defined, which would happen when an article has had no clicks at all, add-one smoothing is used.

See Figure 1 for the distribution of the performance scores of all headlines in the dataset. Most headlines have a performance between −1 and +1, and all headlines have a performance in the range of −5 to +5.

Methods

We describe the methods used to answer the three stated subquestions.

RQ1a: Rewriting of the headlines

For each pair of an original and a rewritten headline, the features from Table 2 of both headlines are obtained and compared to each other. The changes that are observed can be described, plotted, and discussed, resulting in a good overview of the changes that are being made by the editorial team.

To analyze the headlines statistically, a null-hypothesis was formed for each feature, assuming equal averages of the group of original and the group of rewritten headlines. For the *number of characters*, *number of words*, *average word length*, *number of sentences*, *number of sentimental words*, and the *readability score*, a two-tailed dependent sample *t*-test was used, with $\alpha = 0.05$. For the *containing question*, *containing quote*, *containing signal words*, *containing pronouns*, *containing number*, and *first word type* features, the rewritten headlines are checked against the original headlines using a one-way chi-square test. The observed frequencies of the rewritten headlines are compared with the "expected" frequencies according to the original headlines. For each chi-square test, only categories with more than five observations were included in the test due to restrictions of the chi-square test. Again, $\alpha = 0.05$.

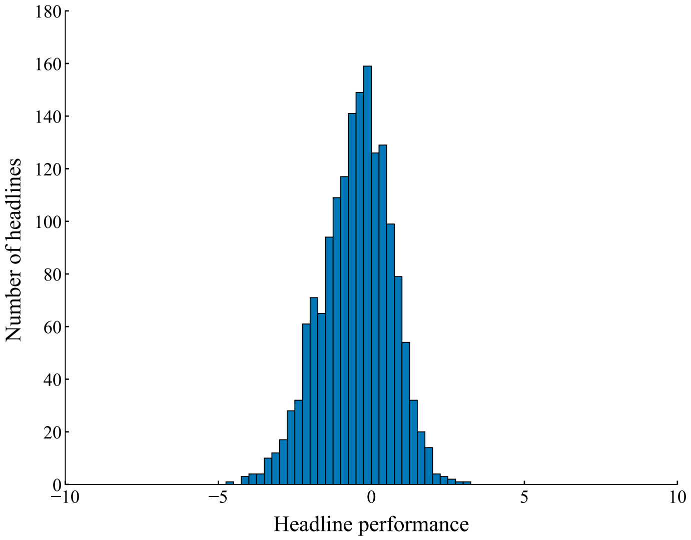

FIGURE 1
Distribution of the performance scores of all headlines in the dataset

RQ1b: Impact on the performance of a headline

The hypotheses that were formed as a result of the first subquestion were tested using the data on the rewritten headlines, combined with their performance. From now on, only data on the rewritten headlines was used. Each hypothesis was formed so that two groups could be formed and compared. For example, a group of headlines that contain a question and a group of headlines that do not contain a question.

Using the Kolmogorov–Smirnov test, the hypothesis that the performance scores are normally distributed (see Figure 1) had to be rejected ($p = 2.07 \cdot 10^{-26}$). Therefore, the nonparametric Mann–Whitney *U* test was used to test the hypotheses on significance. For each hypothesis the α level was set to 0.05.

The Mann–Whitney *U* test linearly orders all samples, from both groups (headlines with and headlines without a certain feature). In this particular case, the headlines are ordered by their performance score. It then tests whether samples from the two groups are equally distributed over this ordering.

RQ1c: Predicting performance of a headline

In order to train a model to predict the performance of a headline based on its features, various machine learning techniques can be used. This paper contains an initial exploration on the possibility of training such a model, and therefore several promising techniques will briefly be tried and evaluated.

First of all, there is the distinction between classification and regression. With classification, the aim of a model is to predict the label of, in this case, a headline. These labels could be "performing slightly better than expected," "performing much worse than expected," and so on. These classes are based on the performance score of a headline. We have trained several classifiers using three different binnings of the performance scores (with two, three, and four classes), but none of them performed significantly better than the baseline model which simply assigns all headlines the majority class.

The most natural model for predicting performance, a numeric score, is regression. For evaluation we use root-mean-square error (RMSE). The RMSE gives a relatively high weight to large errors, and therefore the RMSE is particularly useful when large errors are undesired. The RMSE is defined as

$$RMSE \equiv^{def} \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2}$$

with y_i being the actual value of item *i*, and $\hat{y}_i$ being the predicted value for that item.

The RMSE of the trained models is compared to the baseline obtained by the model which assigns the average performance to each headline (and thus its RMSE is equal to the standard deviation of the performance).

Gradient Boosting (Friedman 2001), Random Forest Generation (Breiman 2001), and Adaptive Boosting (Freund, Schapire, and Abe 1999), which are all supervised learning techniques, have been used to train both regression and classification models. These three techniques are all *ensemble learning decision tree* methods, which work well with a variety of different kinds of data, and that suit the type of data that is used to train the models in this research (Liaw and Wiener 2002).

To find the best parameters for the models trained by Gradient Boosting, a grid search has been performed. To prevent overfitting, the grid search was performed on a dedicated subset of the data, which attributed for 15 percent of all available data. The data in this subset was only used for the grid search, and not for any further training or testing.

For both classification and regression, all headlines were represented as a feature vector. The features with data on an interval scale could be used as they were, and so could binary nominative features (containing question: yes or no). Here, "yes" was converted to a value of 1 and "no" was converted to a value of 0. Nominative features with more than one option, such as containing quote (no, full or partial) were converted into separate binary nominative features. So, for the containing quote feature, there were three corresponding features in the feature vector: containing no quote, containing full quote, and containing partial quote, all with 1 and 0 (yes and no) as possible values. This conversion was needed due to restrictions of the used machine learning implementation.

All models were evaluated using five-fold cross-validation. Using five-fold cross-validation, the training data is split up into five equally distributed *folds*. Then, the model is trained five times on four folds and evaluated on the remaining fold. The fold that is used for evaluation is different each time. For each of the five tests the evaluation metric is calculated separately, and the mean of those separate evaluation metrics is then used as the overall metric.

The *Scikit-learn* Python library (Pedregosa et al. 2011) was used to prepare data, and to train and evaluate the models.

Evaluation

This section lists the results and findings for our three research subquestions.

RQ1a: Rewriting of the Headlines

All features listed in Table 2 show a statistically significant difference between the original and the rewritten headline. Tables 3 and 4 present the basic statistics for most features.

Based on these findings and input from editors and other employees at Blendle, 11 hypotheses were formed, which are listed in Table 5. Some hypotheses, like H2, include a threshold: a value that acts as a limit for the two groups of headlines. These thresholds were chosen after inspecting the results of the analysis, and by looking at how the values for a certain feature are distributed. In the next subsection we will test these hypotheses.

RQ1b: Impact on the Performance of a Headline

All the hypotheses listed in Table 5 are formed as H_1 hypotheses. For each hypothesis an H_0 equivalent was created. For example, for hypothesis H1, the H_0 hypothesis would be "Longer headlines (>50 characters) are *not* preferred over shorter headlines (≤50 characters)."

As described earlier, all hypotheses were tested using the Mann–Whitney U test. In Table 5, the hypothesis number is bold for all the H_1 hypotheses that could be

TABLE 3
Mean, median, and standard deviation for each feature, for both the rewritten and original headlines

Feature	Mean		Median		SD	
	Original	Rewritten	Original	Rewritten	Original	Rewritten
Number of characters	35.57	74.81	34.00	73.00	15.36	24.96
Number of words	6.11	12.39	6.00	12.00	2.84	4.18
Average word length	5.19	5.07	4.80	5.00	1.73	0.89
Number of sentences	1.05	1.21	1.00	1.00	0.26	0.46
Number of sentimental words	0.57	0.94	0.00	1.00	0.75	0.96
Readability score	0.48	0.40	0.45	0.40	0.32	0.23

TABLE 4
The number of headlines with a given feature

Feature	Number of headlines Original	Rewritten
Containing question	181	248
Containing quote	302 full, 55 partial	205 full, 299 partial
Containing signal words[a]	184	805
Containing pronouns[a]	621	1168
Containing number	12	246

[A]THE TOTAL NUMBER OF SIGNAL WORDS/PRONOUNS IS GIVEN.

accepted at an α = 0.05 level. In the third column, the p values for all hypotheses are given, and if a hypothesis could be accepted, the change in performance in percentage points is given in the fourth column. This change indicates how much higher or lower the performance is on average when a headline meets the condition of the hypothesis. For example, looking at H9, the conclusion can be made that the average performance of a headline that contains a negative sentimental word is 17 percentage points higher than a headline that does not.

One thing to note is the Δperformance for H3. The hypothesis was that headlines *with* a question are preferred over headlines *without* a question. Although a statistically significant difference was found between headlines that contained and headlines that did not contain a question, the analysis showed that headlines that do not contain a question achieve better results, contradictory to the hypothesis.

Overall, it can be concluded that the following features have a significant positive impact on the performance of a headline: *average word length*, *absence of a question*, *absence of a quote*, *containing a signal word*, *containing personal or possessive pronouns*, *containing sentimental words*, *containing negative sentimental words* and *starting with personal or possessive pronouns*. The performance of headlines that contain one of those features is on average 14–33 percentage points higher than headlines without the feature.

RQ1c: Training a Model

None of the regression models performed significantly better than the baseline, which had an RMSE of 1.13. An analysis of the individual errors for all three regression models

TABLE 5
The list of hypotheses that were formed as a result of the comparison between the original and rewritten headlines

Hypothesis		*p*	ΔPerformance
H1	Longer headlines (>50 characters) are preferred over shorter headlines (≤50 characters)	0.297	–
H2	Headlines with shorter words (≤7 characters per word) are preferred over headlines with longer words (>7 characters per word)	0.024	+33
H3	Headlines that contain a question are preferred over headlines that do not	0.019	–17
H4	Headlines that contain a partial quote are preferred over headlines that do not contain any quote	0.239	–
H5	Headlines that do not contain any quote are preferred over headlines that contain a full quote	0.030	+17
H6	Headlines that contain one or more signal words are preferred over headlines that do not	0.002	+16
H7	Headlines that contain one or more personal or possessive pronouns are preferred over headlines that do not	0.000	+25
H8	Headlines that contain one or more sentimental words are preferred over headlines that do not	0.018	+12
H9	Headlines that contain one or more negative sentimental words are preferred over headlines that do not	0.001	+17
H10	Headlines that contain a number are preferred over headlines that do not	0.202	–
H11	Headlines that start with a personal or possessive pronoun are preferred over headlines that do not	0.002	+20

In the third column the *p* value is given, showing whether the hypothesis can be accepted at $\alpha = 0.05$. If so, the hypothesis number is bold, and the change in performance in percentage points is given in the last column.

showed that the errors cannot be attributed to a specific subset of the headlines. The models would have, at the current state, little practical use. However, for the models that were trained for this research, minimal effort was put in trying to improve them. With hindsight, we can say that we were too optimistic about the predictive value of the headline features just by themselves. The topic and genre of the news article the headline was referring to may have such a large impact on the CTR that our subtle differences in headline features disappear.

Conclusions

For news read on the internet, the headline of a news article has obtained a new function. Nowadays, a headline is often the primary way of getting a potential reader interested in an article. This has led to something known as clickbait. Clickbait can be seen as a specific style of writing, aiming at inducing the curiosity of the reader and to lure that reader into clicking and opening the article. Often attributed to clickbait is the use of questions, numbers, forward referencing, spectacularization, and negativity.

In the light of this new role of headlines and the relatively new clickbait phenomenon, we investigated its impact on the CTR of headlines in an online environment. We did so using data from Blendle, an online news kiosk.

The analysis of 1828 pairs of original and rewritten headlines revealed that the rewritten headlines differ significantly from their original counterparts on all investigated characteristics. Headlines became longer, and they included more signal words, pronouns, sentimental words, quotes, and questions. Often the rewritten headlines are drastically different compared to the original headlines, but other times these changes are much more subtle and sophisticated. Sometimes, just one additional word is added, or the position of two words is swapped.

While these findings are evidence that the headlines include more clickbait elements than their original headlines, this does not necessarily prove that headlines that include these characteristics are actually more effective. To investigate this, a list of hypotheses on the impact of various characteristics were formed and statistically tested. Based on these tests, we can conclude that some, but not all, of the aforementioned characteristics do have a significant impact on the performance of a headline. Short words, the absence of a question, the absence of a quote, the inclusion of signal words, the inclusion of pronouns, and the inclusion of sentimental words are all characteristics of an effective headline.

These results show that many of the stylistic characteristics that are linked to the clickbait phenomenon actually do have a statistically significant impact on the performance of a headline.

We did not take the topic and genre of an article into account, even though this may have a huge impact on which headlines work and which do not. It is quite possible that some features only work for certain kinds of topics or genres. The absence of a *topic* feature might be one of the primary reasons why the models that were trained to predict the performance of a headline did not achieve sufficient results. Finding a way to include this, or other aspects like the genre of an article, could have a big impact on the performance of those models.

Furthermore, we did not look into the reasons of *why* exactly some features do have a significant impact on the performance of a headline, and why other features did not. While we did this on purpose, due to the focus of our research and the nature of the data we used, we recognize that it would certainly be an interesting starting point for further research.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

NOTES

1. See https://blendle.com.
2. Blendle is available in the Netherlands, Germany, and launched in the United States in early 2016. However, only data from the Dutch version is used for this analysis.
3. See https://figshare.com/articles/Headlines_and_Performance_Blendle_Newsletter/3492512
4. "*Wat*" is also present in the original headline, but there it has a different syntactical function.

REFERENCES

Anderson, Chris W. 2011. “Between Creative and Quantified Audiences: Web Metrics and Changing Patterns of Newswork in Local US Newsrooms.” *Journalism* 12 (5): 550–566.

Atterer, Richard, Monika Wnuk, and Albrecht Schmidt. 2006. “Knowing the User’s Every Move: User Activity Tracking for Website Usability Evaluation and Implicit Interaction.” In Proceedings of the 15th international conference on World Wide Web, 203–212, ACM.

Baresch, Brian, Lewis Knight, Dustin Harp, and Carolyn Yaschur. 2011. “Friends Who Choose Your News: An Analysis of Content Links on Facebook.” In *ISOJ: The Official Research Journal of International Symposium on Online Journalism, Austin, TX* 1: 1–24. 2.

Bell, Allan. 1991. *The Language of News Media*. Oxford: Blackwell.

Blom, Jonas Nygaard, and Kenneth Reinecke Hansen. 2015. “Click Bait: Forward-Reference as Lure in Online News Headlines.” *Journal of Pragmatics* 76: 87–100.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Chen, Yimin, Niall J Conroy, and Victoria L Rubin. 2015. “News in an Online World: The Need for an Automatic Crap Detector.” In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, 81, American Society for Information Science.

De Smedt, Tom, and Walter Daelemans. 2012. “Pattern for Python.” *The Journal of Machine Learning Research* 13 (1): 2063–2067.

Dick, Murray. 2011. “Search Engine Optimisation in UK News Production.” *Journalism Practice* 5 (4): 462–477.

Dor, Daniel. 2003. “On Newspaper Headlines as Relevance Optimizers.” *Journal of Pragmatics* 35 (5): 695–721.

Freund, Yoav, Robert Schapire, and N. Abe. 1999. “A Short Introduction to Boosting.” *Journal-Japanese Society for Artificial Intelligence* 14 (771–780): 1612.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics*: 1189–1232.

Gray, Russell D., and Quentin D. Atkinson. 2003. “Language-Tree Divergence times Support the Anatolian Theory of Indo-European Origin.” *Nature* 426 (6965): 435–439.

Hermida, Alfred, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. “Share, like, Recommend: Decoding the Social Media News Consumer.” *Journalism Studies* 13 (5–6): 815–824.

Holmqvist, Kenneth, Jana Holsanova, Maria Barthelson, and Daniel Lundqvist. 2003. “Reading or Scanning? A Study of Newspaper and Net Paper Reading.” In *The Mind’s Eye: Cognitive and Applied Aspects of Eye Movement Research*, edited by Jana Hyönä, Ralph Radach, and Heiner Deubel, 657–670. Amsterdam: Elsevier Science.

Iarovici, Edith, and Rodica Amel. 1989. “The Strategy of the Headline.” *Semiotica* 77 (4): 441–460.

Ifantidou, Elly. 2009. “Newspaper Headlines and Relevance: Ad Hoc Concepts in Ad Hoc Contexts.” *Journal of Pragmatics* 41 (4): 699–720.

Jijkoun, Valentin, and Katja Hofmann. 2008. “Task-Based Evaluation Report: Building a Dutch Subjectivity Lexicon.” *ILPS-ISLA, University of Amsterdam* 2–11.

Kincaid, J. Peter, 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*. DTIC Document. Technical report.

König, Arnd Christian, Michael Gamon, and Qiang Wu. 2009. “Click-through Prediction for News Queries.” In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 347–354, ACM.

Lai, Linda, and Audun Farbrot. 2014. “What Makes You Click? The Effect of Question Headlines on Readership in Computer-Mediated Communication.” *Social Influence* 9 (4): 289–299.

Lee, Angela M., and Seth C. Lewis. 2012. “Audience Preference and Editorial Judgment: A Study of Time-Lagged Influence in Online News.” In 13th International Symposium on Online Journalism, Austin, TX.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by RandomForest.” *R News* 2 (3): 18–22.

MacGregor, Phil. 2007. “Tracking the Online Audience: Metric Data Start a Subtle Revolution.” *Journalism Studies* 8 (2): 280–298.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. “Building a Large Annotated Corpus of English: The Penn Treebank.” *Computational Linguistics* 19 (2): 313–330.

Mitchell, Amy, and Dana Page. 2015. *State of the News Media 2015*. Pew Research Center. http://www.journalism.org/2015/04/29/state-of-the-news-media-2015

Nir, Raphael. 1993. “A Discourse Analysis of News Headlines.” *Hebrew Linguistics* 37: 23–31.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *The Journal of Machine Learning Research* 12: 2825–2830.

Richardson, Matthew, Ewa Dominowska, and Robert Ragno. 2007. “Predicting Clicks: Estimating the Click-through Rate for New Ads.” In Proceedings of the 16th international conference on World Wide Web, 521–530, ACM.

Rowe, David. 2011. “Obituary for the Newspaper? Tracking the Tabloid.” *Journalism* 12 (4): 449–466.

Safran, Nathan. 2013. *5 Data Insights into the Headlines Readers Click*. https://moz.com/blog/5-data-insights-into-the-headlines-readers-click

Sperber, Dan, and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. vol. 142. Oxford: Blackwell.

Tandoc, Edson C. 2014. “Journalism is Twerking? How Web Analytics is Changing the Process of Gatekeeping.” *New Media & Society* 16 (4): 559–575.

Tenenboim, Ori, and Akiba A. Cohen. 2015. “What Prompts Users to Click and Comment: A Longitudinal Study of Online News.” *Journalism* 16 (2): 198–217.

Van Dijk, T. A. 1988. *News as Discourse*. Hillsdale: Lawrence Erlbaum Associates.

Jeffrey Kuiken, Informatics Institute, University of Amsterdam, The Netherlands. Email: uva@jeffreykuiken.nl

Anne Schuth, Blendle, The Netherlands. Email: anneschuth@blendle.com

Martijn Spitters, Blendle, The Netherlands. Email: martijn@blendle.com

Maarten Marx, (author to whom correspondence should be addressed), Informatics Institute, University of Amsterdam, The Netherlands. Email: maartenmarx@uva.nl