If it Bleeds, it Leads:
Attention and Negativity in Online News

Markus Dertwinkel-Kalt[a], Johannes Muenster[b], and Dainis Zegners[c]

[a]Frankfurt School of Finance & Management
[b]University of Cologne
[c]Rotterdam School of Management, Erasmus University

February 24, 2019

Abstract

What drives the demand for news? Building on an attention-based theory of news consumption, we hypothesize that readers pay more attention and are thus more likely to read news articles with either more polarized or more negative headlines. This gives news sites the incentive to select news stories with more negative or more polarized content and to position those articles more prominently on their websites. Using a large-scale dataset on 360,000 articles published in 2017 by the largest German online news sites and a Machine Learning approach to estimate the sentiment of articles, we find that while negativity is the main driver of determining the positioning and number of views of articles, polarity is the main driver of the selection of news articles that news sites publish. Our results imply a potentially serious problem for societies where democratic processes rely on an informed electorate.

Keywords: Media Bias, Online News, Machine Learning, Big Data

> Often, the media will focus mostly on the negative and superficial; perhaps this is because media people believe that is what people want and where the money is.
>
> Helmut Schmidt[1]

1 Introduction

What drives the demand for news? Standard economic theory suggests that people require information to make informed decisions. Accordingly, media should have an incentive for truthful reporting (Coase, 1974; Besley and Burgess, 2002; Djankov et al., 2003). Empirical studies, however, have shown that people often request biased reporting, for instance, consumers prefer like-minded news (Gentzkow and Shapiro, 2010; Raymond and Taylor, 2013). The literature in psychology has proposed *confirmation bias* and one's tendency to avoid *cognitive dissonance* (i.e., the tendency to avoid information that conflicts with ones own previously taken actions) as explanations for the demand for like-minded news. These constituents of news demand, however, are not exhaustive and may not even capture the main drivers of demand for news.

As Simon (1971) put it, "a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it." Thus, in order to be perceived, news has to grab the readers' attention (Boik et al., 2016). But what grabs readers' attention and how does this impact the supply of news? In this paper, we seek to answer these questions combing rigorous modeling with a large-scale empirical study on the determinants of supply and demand for news. Building on research by Vosoughi et al. (2018) showing that fake news spreads quicker than the truth as it tends to be more *surprising*, we propose an attention-based theory of the news industry. News stories that grab readers' attention are more read and spread more quickly in the population, which gives the news-media the incentive to supply more attention-grabbing news. If readers do not seek the truth but whatever is surprising and grabs their attention, fake news even has a competitive advantage: Not being bound to the truth it is easier to win the race for consumers' attention. This poses a serious problem for modern societies where democratic processes rely on an informed electorate. To fight fake news, it is important to understand what aspects render (fake) news popular and what aspects render truthful news unsuccessful.

We propose a simple model with one news supplier and many readers who request attention-grabbing news. In line with the literature that we delineate in the following, we assume that the headline of a news-article determines whether it is read or clicked on. In order to determine *what* features grab a consumer's attention and thereby enhance the demand for the corresponding article, we build on insights from fields such as linguistics, where features of style, wording, and sentiment have been pinned down to increase the popularity of news-articles (Molek-Kozakowska, 2013; Galtung and Ruge, 1965; Piotrkowicz et al., 2017). The stronger the *polarity* of a headline — that is, the more sentiment-charged and negative language is used in a headline — the larger the

[1] In the preface to Haagerup (2015).

corresponding article's popularity (Reis et al., 2015; Piotrkowicz et al., 2017).[2] From an entirety of news that can be reported, the news supplier selects *which* of these to report and *how* to report them. All news has a *natural polarity* that is defined by the news' factual content. The article itself reflects the true polarity, but the supplier can choose which articles to publish and where to position articles in its outlet. The higher the positioning of the article in the news outlet or on the news website the higher the probability that the consumer notices it.

Our model makes a number of predictions that we test empirically: (1) The stronger the negativity or polarity of a headline, the higher the demand for an article; (2) the higher the positioning of an article on news site, the larger the demand for an article; (3) the stronger the negativity or polarity of the headline of a news article, the higher its position on a news site (4) articles with a negative or polarized headlines are more likely to be published by news sites.

In order to test the predictions from our model, we combine supply-side data in the form of 360,000 articles published in 2017 by German online news-portals with demand-side data on which articles consumers actually read. Additionally, our data allows us to examine which articles news sites actually publish, as we have also collected 110,000 articles by press agencies such as *dpa* and Agence France.

Our results show that news sites select to publish articles that are based on news report with stronger polarity, that is more sentiment-charged language, and select to position articles with more negative headlines at a higher position on their websites. Readers, on the other hand, are more likely to visit articles with more negative headlines, even when we control for the underlying news event. Taken together, our results support the hypothesis that attention is an important determinant of both news supply and news consumption.

2 Model

Suppose there are N news portals and J readers. Each news portal i has access to a number $N_i \in \mathbb{N}$ of articles.[3] Each of these articles has a headline with sentiment $s \in \mathbb{R}$ that is drawn from a distribution that is symmetric around zero.[4] Each news portal maximizes its profits by deciding which articles to print and how to rank them. Each, $\pi_i(\cdot)$, is an increasing function in the number of overall clicks, X, that is $\pi_i'(X) > 0$. Publishing each article comes at cost $\delta > 0$.

Each reader j randomly selects one news portal i. Each consumer has an *attention span* of n_j: this number gives the first n_j articles that the reader considers.[5] First, each reader randomly

[2] The polarity captures how extreme a message is, and more extreme events can be assumed to be more surprising: For instance the headline "From a Rwandan Dump to the Halls of Harvard" has a strong polarity, alludes to a surprising event and was found to attract high click rates, while "Home Sales Around the Region" does not (Piotrkowicz et al., 2017).

[3] For each news portal this number N_i represents a draw from some distribution with support over $\mathbb{N}$.

[4] All of our predictions carry over to the more general case that the distribution has at least as much probability mass on $\mathbb{R}_{\leq 0}$ as on $\mathbb{R}_{\geq 0}$.

[5] This *attention filter* of only considering the top n_j entries has been used in the theoretical literature on limited attention (e.g. Salant and Rubinstein, 2008; Masatlioglu et al., 2012) and has been supported by empirical research (e.g. Hotchkiss et al., 2004; Dulleck et al., 2008).

selects one article among the considered articles. The reader will then click on this selected article with probability $p \in [0, 1]$. This probability p is assumed to be a function of the sentiment s of the selected article's headline. We have two different hypothesis about how s affects p: (i) $p(s) < 0$ and (ii) $p(|s|) > 0$. By (i), readers are more likely to click on an article the more negative the sentiment of the article's headline is; by (ii), readers are more likely to click on an article the stronger the polarity of the article's headline is, where polarity is measured by the absolute value of the sentiment. We thereby do not only assume that the sentiment affects demand, but we also assume that it is only the sentiment of the headline but not the body of the respective article that affects demand. We regard this as plausible as prior to clicking on an article the headline is the main or the only part of an article that the reader can see. That negativity raises demand has been supported by various studies in political science (e.g. Trussler and Soroka, 2014) that find that more news are negative, and this assumption has been microfounded by McCluskey et al. (2015). We are not aware of any study that tests for polarity, but it can be derived from recent research on the *contrast effect* and salience (e.g. Bordalo et al., 2013) applied to news sentiment: accordingly an article receives more attention the larger the contrast between the sentiment s and a reference sentiment value 0, that is, the larger $|s|$, is. We are also not aware of any study that distinguishes between the sentiment of an article and the sentiment of its headline—which we regard as an important distinction—and for any study that tries to control for the news value of an article and that therefore tries to identify the causal effect of sentiment on demand. This simple model gives rise to the following predictions.

Prediction 1 *The stronger the negativity [negativity] of a headline, the higher the demand for an article.*

As we have two countervailing hypotheses about what attracts consumer demand—negativity or polarity—we test for both and use in our subsequent analysis whatever sentiment measure has more predictive power for readers' demand. We control for news value by determining whether two articles build on the same article by a news agency: if this is the case we feel safe to assume that the news value in the two articles is the same, which allows us to learn about the causal effect of sentiment on demand.

Prediction 2 *The higher the positioning of an article (i.e., the lower its rank), the higher the demand for an article (as measured by the number of clicks).*

As every consumer j considers only the first n_j articles and as the n_j varies across readers, the likelihood to be considered (and therefore also be clicked on) should, *ceteris paribus*, monotonically decrease in the article's rank on the newsportal. We can also test this prediction using the rank on *Google News* or *Google Search.*

Prediction 3 *The stronger the negativity [polarity] of the headline of a news article the higher its positioning.*

A news portal profits from an article only if readers click on it: thus, each news portal will reserve its best spots—that is, its highest positions—to those articles that have the highest likelihood to be clicked on: the particularly negative (or polarized) articles.

Prediction 4 *Articles with a negative (or polarized) headline are more likely to be published.*

As negative (or polarized) articles attract more demand, printing them is more profitable for the news portal.

3 Methodology and Data

In this section, we describe the methodology that we use to estimate the sentiment of articles and the datasets that we use in our analysis.

3.1 Sentiment Estimation Using Machine Learning

Instead of relying on a pre-defined dictionary method to estimate the sentiment of headlines and articles in our dataset, we estimate our own algorithm using a machine learning method. For the estimation of the algorithm, we obtained data from *Media Tenor*, a German-Swiss media consultancy that gathers information on how companies and political actors are presented in the press. The data we obtained includes 80,104 articles from 2012 - 2015 published in 7 large German news outlets.[6] Data from *Media Tenor* has been used in previous research on media bias or the impact of media consumption, e.g. in Friebel and Heinz (2014); Dewenter et al. (2016); Beckmann et al. (2017).

For each article in the *Media Tenor* dataset, we observe the headline, the main subjects mentioned in the article, and how each subject is presented in the article. Articles are coded by human coders, who rate whether a subject is presented in an article "positively", "neutrally" or "negatively." We assume that the sentiment of an article is determined by the average over how the subjects are presented, and thus compute the sentiment of an article by taking the mean over subjects, coding "positive", "neutral" and "negative' as +1, 0, -1 respectively. While this is a strong assumption, we can later show that our algorithm is better at predicting the sentiment of headlines than using out-of-the-box dictionary methods.

After computing the sentiment of each article, we split the *Media Tenor* dataset into a training set consisting of 70,000 headlines that we will use to estimate our sentiment algorithm, and a validation set consisting of 10,104 headlines that will use to judge the performance of the algorithm. Additionally, we will use 300 separate headlines that were coded by our research assistants to validate our algorithm.

To predict the sentiment of an article, we use a regression-based approach, estimating the following model:

[6] The news outlets are *Deutschlandfunk*, *ARD Tagesschau*, *ZDF heute*, *ZDF heute journal*, *Bild am Sonntag*, *ARD Tagesthemen*, , *Bild-Zeitung*, *Der Spiegel*, *Focus*

$$s_i = \beta_0 + \beta_1 w_{1i} + ... + \beta_j w_{2i} + ... + \beta_n w_{ni}, \tag{1}$$

where s_i is the sentiment of article i; $w_{1i}, w_{2i}, ..., w_{ni}$ are dummy variables indicating whether word $j = 1, ..., n$ occurs in the headline of article i, and $\beta_0, \beta_1, ..., \beta_n$ are coefficients measuring the impact the occurrence of word j in a headline has on the sentiment of article i.

We also reduce all words to there stems, e.g. "attackers", "attacked" become "attack" after stemming, remove all stopwords such as "the", "is", "at", and "on."; and remove all numbers. Additionally, we do not consider only single words but also all 2-gram and 3-gram phrases such as "heavy attacks" and "more heavy attacks" that occur in the headlines in our dataset. As the headlines in our dataset include more than 900,000 unique single words, 2-grams and 3-grams, and for computational reasons we can not include all of them, we narrow our estimation to those 3,400 that are either the most common or that are more likely to occur in more positive or more negative headlines. The results of preliminary tests indicated that including more words only lead to marginal improvements in the accuracy of the algorithm.[7]

We estimate the coefficients β_j using Lasso-regression (Tibshirani, 1996), which in contrast to a normal regression (OLS) avoids overfitting our model to the training data. The resulting regression model estimated on the training data has a mean error of 0.446 when predicting the sentiment of the headlines on the validation dataset. Additionally, we use 300 separate headlines that were coded by research assistants to validate the algorithm. The algorithm predicts for 61% of the headlines the same sentiment as a human coder. In comparison, using a out-of-the-box dictionary method, for which we used a translated to German version of the Lexicoder Sentiment Dictionary (Young and Soroka, 2012; Proksch et al., 2018), we were only able to predict 55 % of headlines correctly, while using the *IBM Watson* (`www.ibm.com/watson`) sentiment estimation API were able to predict the sentiment correctly in 59 % cases. Different human coders agreed in 70 % cases the same sentiment for a headline, which can thus be considered the upper boundary of accuracy that any algorithm could achieve when trying to predict which sentiment a human would attach to a headline.

Figure 1 shows the most common positive and negative words that were identified by our algorithm.

3.2 Datasets

We combine several data sources in our project to examine the different stages of the vertical and horizontal structure of the online news industry, as well as the demand side. Our data captures these dimension for Germany in 2017 using data on articles published by 14 of the most-read online news platforms in Germany.

[7] This is a common approach, see e.g. Gentzkow and Shapiro (2010); Gentzkow et al. (2017)

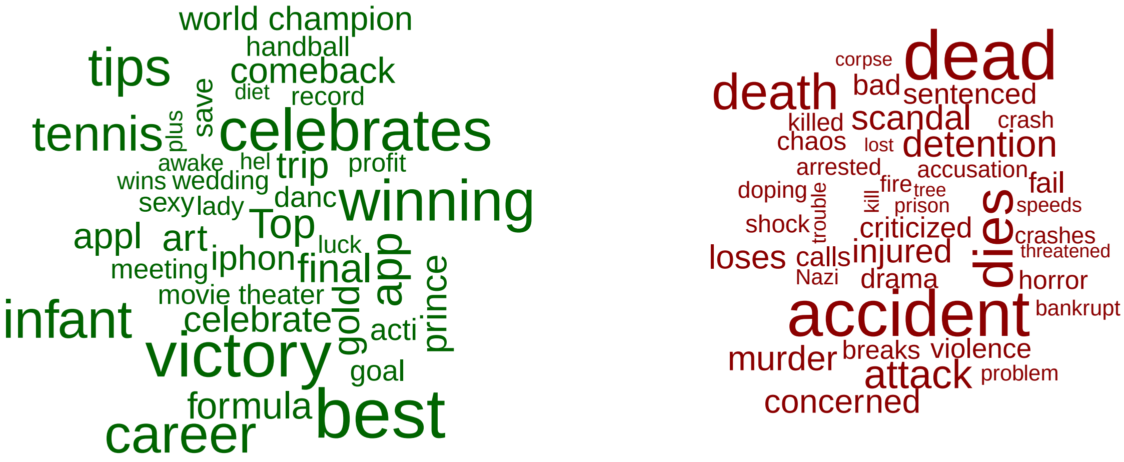

Figure 1: Word clouds of most frequent positive words (left) and negative words (right)

Nielsen NetViews Panel

To gather information on which articles internet users read, we have obtained data from the *Nielsen Netview Online Panel.*[8] The panel tracks the online behavior of 200,000 individuals worldwide including the websites they visit, down to the individual sub-domains. *Nielsen Germany* has granted us access to their 2017 panel of 16,000 individuals that are representative of the German population. Our dataset includes all unique URLs on the 14 platforms that have been visited at least 5 times by participants of the panel in 2017.[9] For each URL in our dataset, we observe the number of times it has been visited in 2017 by participants of the panel. Although our source data contain contains the number of monthly views per articles, we aggregate the views per articles , as the majority of views occur in the first month an articles is published. This dataset consists of 54,478 unique URLs of news articles.

News Agency Reports

We collected via the *LexisNexis* database all articles published in 2017 by the three leading German-language news agencies *Deutsche Presse-Agentur (dpa)*, *Associated Press (ap) Germany* and *Agence France-Presse (afp) Germany*. These articles are typically not released to the general public but only to paid subscribers. They can be thought of as an input to the news that are reported to readers on online platforms, as they capture news events and stories that news platforms can choose to report on.

[8] See `http://en-us.nielsen.com/sitelets/cls/digital/online-netview.html` (accessed April 26, 2018).

[9] Due to privacy concerns, we were not granted access to URLs that were visited less than 5 times.

Table 1: Summary Statistics per News Agency

News Agency	N	Mean(Sentiment Headline)	Mean(Sentiment Body)
Deutsche Presse Agentur (dpa)	38,706	-0.20	-0.34
Agence France Presse (afp) German	47,678	-0.31	-1.31
Associate Press (ap) German	23,919	-0.29	-1.58

Our dataset includes for each news agency article the date it has been published, the headline and the main body of the article. Additionally, we estimate the sentiment of each headline and the text in the body using the methodology described in section 3.1, and compute the number of characters in the headline and body of each article. Table 1 shows summary statistics for the 110,303 news agency articles in our dataset.

News Articles

We have collected data on (most) news articles published in 2017 by 14 of the most-read online news platforms in Germany. Our goal is to capture the news that are supplied online to consumers. We went through the following steps to obtain the data: We first collected the URLs of all news articles that are listed by *Google News* (`news.google.com`) and *Google Search* (`www.google.com`). To do this, we queried *Google News* and *Google Search* repeatedly, each time narrowing the search down to one of the 14 news-sites and a particular day in 2017. Each query returned a unique list of all articles that were published by the specified news-site on that particular day in 2017.[10] Additional to the URL of each article, we recorded the search rank of an article both on *Google News* and *Google Search* among those articles that were published on the same day on the same news-site. This dataset includes 333,616 unique URLs of news articles.

In a next step, we merge this dataset with the unique URLs in the dataset provided by Nielsen. This combined dataset consists of 361,106 unique URLs, of which 26,988 URLs are contained both in the *Google* and the *Nielsen* dataset.

Finally, we download and parse each of the URLs to obtain the headline, the body of the article, and the date it has been published on the news site.

Position of Articles on News Site

An important determinant for the number of times an article is viewed is the position of the article on a news-site, i.e. whether it is published more towards the top of a landing page or more towards the bottom. To obtain data on where news sites position a specific article, we collected multiple snapshots of each day in 2017 of each landing page of the news websites in our dataset via the *Internet Wayback Machine* (`https://archive.org/web`). After downloading each snapshot from the *Wayback Machine*, we extract from the HTML code of each landing page all hyperlinks, and

[10] We used a blank *Amazon Cloud* instance in this process to not contaminate our queries with previous browsing and search behavior that might be stored in Cookies, as Google likely personalizes results based on previous user behavior.

Table 2: Description - Main Variables and Controls

Variable	Description
Rank *Google News*	Rank of an article on *Google News* within the articles published on the same News website on a particular day.
Rank *Google Search*	Rank of an article on *Google Search* within the articles published on the same News website on a particular day.
Rank News Site	Rank of an article on the News website on the date the article was published. To compute the rank, we collected multiple daily snapshots of each news site via the Internet Wayback Machine (`https://archive.org/web`). The final rank of an article is computed as the highest rank an article reached within the available snapshots of a news website we collected.
N Views	The number of times that unique participants of the *Nielsen Netview Online Panel* visited the url of an article in 2017.
Sentiment Headline	Sentiment score of the headline of an article. A higher score implies more positive sentiment.
Sentiment Body	Sentiment score of the text in the body of an article. A higher score implies more positive sentiment.

Table 3: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Rank Google News	241,008	30.09	21.18	1.00	13.00	27.00	44.00	140.00
Rank Google Search	254,247	32.11	23.09	1.00	14.00	28.00	46.00	201.00
Rank News Site	133,280	37.72	22.16	1.00	21.00	32.00	52.00	161.00
N Views	54,478	11.73	11.02	5.00	6.00	8.00	13.00	316.00
Sentiment Headline	361,106	−0.15	0.23	−1.98	−0.24	−0.14	−0.06	2.00
Sentiment Body	361,106	−0.44	3.17	−201.67	−1.26	−0.25	0.74	163.19

filter these hyperlinks such that we only retain the URLs of news articles in our dataset. This gives us an ordered list of URLs of news articles for each snapshot, which we match with dataset of URLs or news-articles, recording the highest position that a specific news article has reached on the landing page of a news-site. In the end, we are able to observe the rank of 133,280 news articles in our dataset.[11]

Tables 3 and 4 show summary statistics of our final dataset of news-articles.

4 Results

So far, we have conducted a preliminary analysis that lends support to our four predictions.

[11] The missing data is due to the fact that although for the most popular news sites the *Wayback Machine* collects hourly snapshots, for less popular websites in collects substantially fewer snaphots. For example, we are able to observe the rank at least once for 26,290 out of 31,674 articles from the popular news site *Spiegel Online*, whereas for the far less popular news site *express.de*, we are only able to observe 2,458 of 20,289 articles in our dataset at least once.

Table 4: Summary Statistics by News-Website

News Site	N	Mean(N Views)	Mean(Sentiment Headline)	Mean(Sentiment Body)
augsburger-allgemeine.de	26,920	6.364	-0.126	0.285
bild.de	38,151	12.726	-0.147	-0.228
express.de	20,289	6.637	-0.161	-0.125
faz.net	28,466	6.530	-0.146	-0.327
focus.de	28,976	9.150	-0.155	-0.243
rp-online.de	24,565	6.272	-0.134	0.147
spiegel.de	31,674	8.999	-0.155	-0.888
stern.de	18,600	7.268	-0.128	-0.777
sueddeutsche.de	32,272	6.716	-0.148	-0.283
t-online.de	32,593	18.494	-0.156	-0.630
tagesschau.de	8,997	8.152	-0.177	-1.255
welt.de	31,227	7.740	-0.161	-0.284
zdf.de	13,230	16.577	-0.133	-2.618
zeit.de	25,146	8.336	-0.145	-0.609

4.1 Position and Demand for Articles

According to Prediction 2, the higher the position of an article (i.e. the lower its rank), the higher the demand for an article as measured by the number of views. Figure 2 shows the average number of views plotted against the rank of an article on a news site. As predicted, there is a strong relationship between an article's rank and the number of views, with articles that have a lower rank, i.e. articles that are positioned higher on a news site, being viewed more often.

The regression results in table 5 confirm this result. Looking at column (1), we see that articles that are being positioned one positioned lower receive 0.7 percent less views (p-value < 0.01). In columns (2) and (3), we examine the impact of being higher placed on *Google News* (column 2) and *Google Search* (column 3). The results show that being placed one rank lower on *Google News*, is associated with 0.2 percent less views (p-value < 0.01). The rank on *Google Search* does not have a statistically significant impact on the number of views. Including all three variables in a single regression in column (4) gives the same result.

As we were not able to find ranks of all articles on their respective news sites, on *Google News* and *Google Search*, but still want to retain such articles in our regression model, we use in column (5) the following approach: We include for all three ranks a dummy variable indicating whether the respective ranks are available in our data, and substitute zero for missing observations of the rank variables. The rank variables thereby can be interpreted as an interaction between whether the rank is available and the actual rank in case it is available in our data. This method of dealing with missing values is discussed in (Gelman and Hill, 2006). The results in column (6) indicate that articles for which a rank is available on a news sites, on *Google News* and *Google Search*, receive a substantially higher number of views (p-values < 0.01). The impact of higher ranks remains qualitatively the same as in the previous regressions.

To conclude, we find conclusive evidence supporting prediction 1, showing that articles with a higher position on a news site and on *Google News* receive a higher number ov views, but not

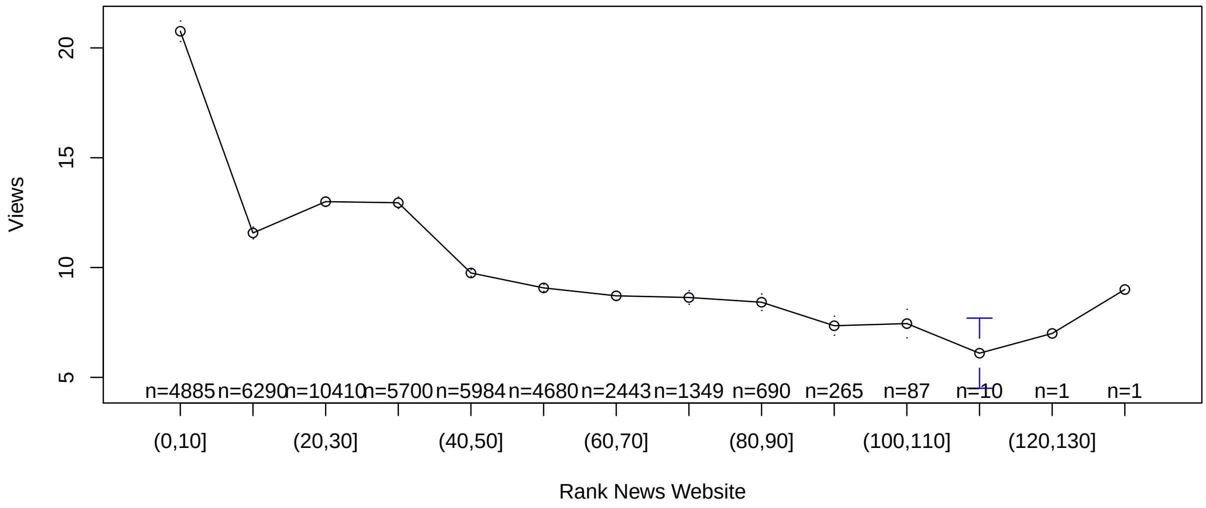

Figure 2: Impact of articles' position on number of views

Note: Error bars show 95 percent confidence intervals. If a confidence interval is not shown, it is too narrow to be visible on the plot.

Table 5: Impact of Articles' Rank on Number of Views

News Agency	N	Mean(Sentiment Headline)	Mean(Sentiment Body)
Deutsche Presse Agentur (dpa)	38, 706	-0.20	-0.34
Agence France Presse (afp) German	47, 678	-0.31	-1.31
Associate Press (ap) German	23, 919	-0.29	-1.58

Note: This table shows result from OLS regressions with news site and date fixed-effects. Each observation is a single article. Standard errors are clustered on the level of a news site.

articles that are higher positioned on *Google Search*.

4.2 Position of Articles with Negative or Polarized Headlines

According to Prediction 3, the stronger the polarity or negativity of a news article, the higher its position. Finding evidence supporting this prediction would indicate that news sites have a strategic incentive to place articles with more negative or polarized headlines on higher positions, as these articles receive more views, which increases the advertising revenues of news sites.

In order to test Prediction 3, we examine not only the relationship between the polarity or negativity of an article's headline and its position on a news site, but also its position on *Google News* and *Google Search*. As *Google* has also an advertising-based business model, it should face the same incentives to position articles that have a higher likelihood of being clicked on higher in its search results.

Figure 3 plots the three relevant relationships in our data. As can be seen in the upper panel of figure 3 plotting the relationship between the sentiment of the headline of an article against its rank on *Google News*, *Google News* seems to rank articles with an intermediate sentiment lower than articles with either a more positive or more negative sentiment. This is evidence in favor of the polarity of headlines driving the decision by Google (and its algorithm) how to rank articles. Looking at the same relationship in case of *Google Search* in the middle panel of figure 3, we observe the opposite relationship: Articles with headlines with intermediate sentiment seem to have lower average ranks than articles with either more positive or more negative headlines, which means that articles with intermediate sentiment are shown higher up in the search results of *Google Search*. We have currently no theoretical explanation why we observe the opposite of prediciton 3 in our data. Possibly, visitors of Google's search engine are more interested in the informational content of news articles as opposed to visitors of *Google News*, which gives Google the incentive to rank more neutral and thus maybe more objective articles higher on *Google Search*.

The lower panel of figure 3 shows the relationship between the sentiment of articles' headlines and their rank on a news site. In this case, there seems to be a positive relationship between headlines sentiment and articles' rank, indicating that articles with a more negative sentiment are positioned higher up on news sites. This would indicate that news sites' decisions where to place articles are determined by the negativity of headlines rather than polarity.

The results of the linear regression models in table 6, including date of publishing and news site fixed effects, confirm the relationships observed in figure 3. In case of the regression using the rank of an article on *Google News* as a dependent variable in columns (1) and (2). Although including only a linear effect of the sentiment in column (1) is also significant, when including both a linear and a square term for the sentiment in column (2), only the square term is statistically significant (p-value < 0.01), confirming an U-type relationship between the sentiment of the headline of an article and its rank on *Google News*, supporting the prediction that polarity of headlines is an important driver of where articles are positioned on *Google News*.

Looking at columns (3) and (4), where we use an article's rank on *Google Search* as the dependent variable, we again observe the opposite relationship: In column (4), only the square term of the sentiment has a statistically significant impact (p-value < 0.01). As the sign of the effect is negative, this indicates an inverse U-relationship, which implies that articles with polar headlines have higher average ranks and are thus placed lower in *Google Search*'s results.

In columns (5) and (6), where we use an article's rank on news sites as a dependent variable, only the linear term of headline's sentiment is statistically significant (p-value < 0.01), indicating that in this case only the negativity of sentiment is a driver of where articles are placed on news sites. The positive sign of the coefficient of the sentiment of the headline indicates that articles with a more positive sentiment of their headline have a higher average rank, meaning they are placed lower on news sites.

To sum up, we find support for prediction 3 that polarity and negativity drives news sites' and Google's decisions where to place articles on their websites.

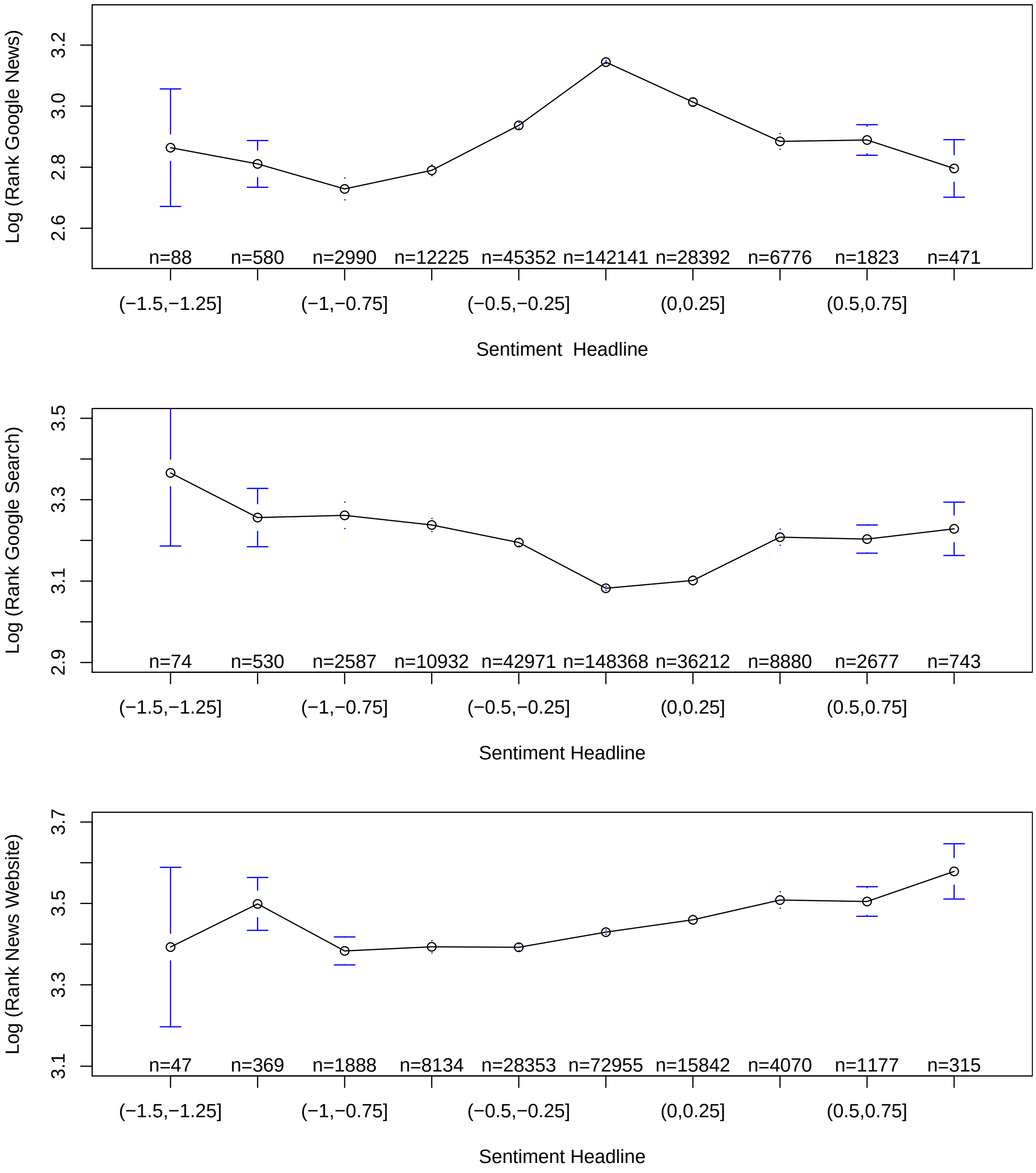

Figure 3: Impact of headlines' sentiment on its rank on Google News (upper panel), *Google Search* (middle panel) and news sites (lower panel).

Note: Error bar show 95 percent confidence intervals. If a confidence interval is not shown, it is too narrow to be visible on the plot.

Table 6: Impact of Articles' Sentiment on Rank

	Dependent variable:					
	Log(Rank Google News)		Log(Rank Google Search)		Log(Rank News Website)	
	(1)	(2)	(3)	(4)	(5)	(6)
Sentiment Headline	0.192***	0.001	−0.113	−0.052	0.117***	0.110***
	(0.044)	(0.034)	(0.070)	(0.052)	(0.043)	(0.036)
Sentiment Headline (Squared)		−0.608***		0.287**		−0.023
		(0.129)		(0.114)		(0.037)
Observations	241,008	241,008	254,247	254,247	133,280	133,280
R^2	0.121	0.127	0.081	0.082	0.567	0.568

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Note: This table shows results from OLS regressions with news-site and date fixed-effects. Each observation is a single article. Standard errors are clustered on the level of a news site.

4.3 Demand for Articles with Negative or Polarized Headlines

According to Prediction 1, the stronger the polarization/negativity of a headline, the higher the demand for an article. We proceed to test this prediction in two steps: First, we look at the simple relationship between the polarity and negativity of headlines and the number of times articles have been viewed. As by looking at this relationship we are not able to distinguish whether it is the polarity or negativity of a headline or the polarity or negativity of the underlying news content which is driving demand, in a next step we control for the underlying news content by using fixed effects on the level of news agency reports that articles are based on and the sentiment of the body of the articles and news agency reports.

Figure 4 plots the raw relationship between the sentiment of the headlines of articles and the number of views of articles. It is already visible in this graph that there might be a negative relationship between the sentiment of headlines and the number of views, indicating that articles with more negative headlines receive a larger number of views.

To control for factors such as the impact of articles' ranking, we run a series of regression models. Table 7 shows the results. In all specifications, the coefficient of the linear term of the sentiment of the headline of an article has a negative and statistically significant effect (p-value < 0.01) on the number of views of an article. In columns (2), (4) and (6), we also include the square term for the sentiment of the headline, but it is not statistically significant in all three cases. In columns (3) and (6), we include the ranks of articles on either the news site or *Google Search* and *Google News* into the regressions. Although as found in the subsection 4.1 and 4.2 these variables have an impact on the number of views, controlling for these variables does not change the direction of the main effect of the sentiment of an article's headline.

One concern with this analysis is, however, that it is not the negativity of a headline, but the negativity of the underlying news issue that drives the demand. For example, when readers click on the headline "Schock in the region after the death of six young people," an article on a tragic traffic

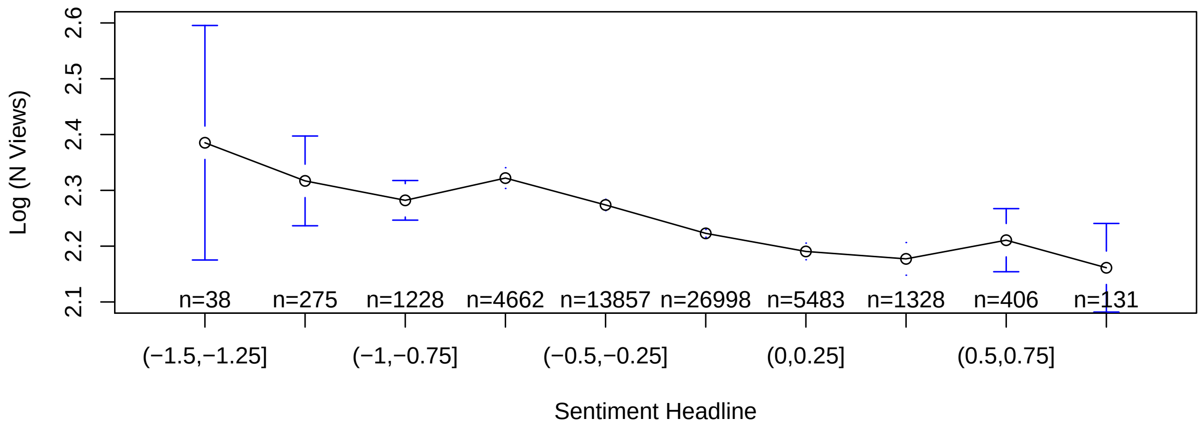

Figure 4: Relationship between the sentiment of the headline of an article and its number of views

Table 7: Impact of Articles' Sentiment on Views

	Dependent variable:					
	Log(N Views)					
	(1)	(2)	(3)	(4)	(5)	(6)
Sentiment Headline	-0.114^{***}	-0.085^{*}	-0.144^{***}	-0.159^{***}	-0.100^{***}	-0.079^{*}
	(0.038)	(0.046)	(0.036)	(0.049)	(0.029)	(0.047)
Sentiment Headline (Squared)		0.075		−0.036		0.051
		(0.065)		(0.034)		(0.065)
Rank on News Site			-0.007^{**}	-0.007^{**}		
			(0.003)	(0.003)		
Rank on Google News			-0.002^{**}	-0.002^{**}		
			(0.001)	(0.001)		
Rank on Google Search			0.00002	0.00003		
			(0.0002)	(0.0002)		
Rank on News site Available (Dummy)					0.462^{***}	0.462^{***}
					(0.107)	(0.107)
Rank on News Site (0 for not available)					-0.007^{***}	-0.007^{***}
					(0.002)	(0.002)
Rank on Google News Available (Dummy)					0.103^{***}	0.103^{***}
					(0.026)	(0.026)
Rank on Google News (0 for not available)					-0.001^{**}	-0.001^{*}
					(0.001)	(0.001)
Rank on Google Search Available (Dummy)					0.086^{***}	0.084^{***}
					(0.019)	(0.018)
Rank on Google Search (0 for not available)					0.0001	0.0001
					(0.0002)	(0.0002)
Observations	54,478	54,478	15,480	15,480	54,478	54,478
R^2	0.193	0.193	0.274	0.274	0.252	0.252

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Note: This table shows results from OLS regressions with news-site and date fixed-effects. Each observation is a single article. Standard errors are clustered on the level of a news site.

accident published on the news site of the Bavarian regional newspaper “Augsburger Allgemeine”, it might not be the specific choice of words by the newspaper (e.g. “schock” and “death”) that drives the observed correlation of the sentiment of the headline and the number of views of an article, but the unobserved negativity of the news event, in this case the traffic accident that the article is based on, that impacts both the number of views and the negativity of the headline.

In order to control for the unobserved negativity or polarity of the underlying news event, we use the following approach: By using the text of the body of each article and the text of the press agency report that we have collected, we can identify for every news article published by one of the news portals whether it is based on a report by a news agency and if yes on which specific news agency report. To do this, we compute the text similarity between the body of each news article and the body of every news agency report that was published on the same day or one or two days prior to the news article. To compute text similarity, we use cosine similarity, a very common and established measure in computational linguistics (Gomaa and Fahmy, 2013), where the upper bound is 1, indicating that two articles use the exact same words with the same frequencies. This measure has been used in the Economics literature for example to compute product differentiation based on text-descriptions of products (Gomaa and Fahmy, 2013) or to estimate the impact of Wikipedia on the dissemination of ideas and concepts in published natural science papers (Thompson and Hanley, 2018). In the next step, we identify for news articles the news agency report that has the highest text similarity. Examining a subset of articles ourselves, we judged that above a value of 0.8, the news articles and the most similar news agency report are likely to be on the same news issue.[12] For example, for the article mentioned above, we identified a corresponding news agency report with a cosine similarity of 0.86 published on the same day by *Deutsche Presse Agentur (dpa)* with the headline “Young people die in an accident caused by fog” which describes the same accident in the region of Bavaria as the news site’s article. Overall, we are able to identify for 71,279 of 361,000 news articles a corresponding news agency report that describes the same news event.

We use the identified news agency reports in two ways. As we observe several instances of news articles by different portals being published based on the same news agency report, we can include news agency report fixed effects, thus identifying the effect of headlines’ sentiment on views only by using variation within articles that are based on the same news agency report. As all articles that are based on the same news agency report are based on the same underlying news event, we can be more certain that we measure the effect of the sentiment of headlines and not the effect of the negativity or polarity of the underlying news event. Additionally, we can also use the sentiment of the body of an article, the sentiment of the headline and the body of the agency report as a control in a regression. If the sentiment of the headline of an article has still a statistically significant effect on the number of views of an article in such a regression, we can also be more certain that we measure the impact of the sentiment of the headline.

Table 8 shows the regression results. In columns (1) to (4), we include news agency report fixed effects, while in column (5) we include the sentiment of the body of an article, and the corresponding

[12] We plan to validate this cut-of value with the help of human coders in future revisions of our paper.

news agency report. The sentiment of the headline has a negative effect on the number of views of an article in all regressions, although it is not statistically significant in the case of only including articles were we observe the rank on a news site, *Google News* and *Google Search* in column (2), and in case of a regression where we also cluster standard errors on the news site level in column (4).

To sum-up, we find strong evidence that the negativity of headlines positively impacts the number of views that articles receive, and some tentative evidence that the effect is partly driven by the actual choice of words by the news portal and not the negativity of the underlying news story.

4.4 Selection of Negative or Polarized Articles

According to Prediction 4, either more negative articles or more polarized articles are more likely to be published by news sites. To test this prediction, we use the fact that as described in the previous section, we can identify for all news articles whether they are based on a news agency report. Thereby we can examine whether news agency reports that have either a more negative or more polarized sentiment are more likely to be used as an input to news articles by news sites.

Figure shows 5 the relationship between the sentiment of the headline of a news agency report and the probability of being picked up by at least one news portal (upper panel) and the average number of times a news agency report is picked up by different news sites (lower panel). In both cases, visually the relationship is indicative of a U-shape relationship, which would imply that polarity and not negativity of news agency reports is driving news sites decisions whether to publish articles based on them. Table 5 showing the results of four regression models confirms this initial impression. While the sentiment of headlines of news agency reports does not have a statistically significant effect when only included as a linear effect in columns (1) and (3), it has a strong and statistically significant effect when included as a square effect ($p-$values < 0.01) in columns (2) and (4).

To sum up, we find strong evidence that the polarity of agency reports is driving the selection of articles by news sites.

5 Conclusion

Our results show that news sites select to publish articles that are based on news reports with stronger polarity, that is more sentiment-charged language, and select to position articles with more negative headlines at a higher position on their websites. Readers, on the other hand, are more likely to visit articles with more negative headlines, even when we control for the underlying news event. Taken together, our results support the hypothesis that attention is an important determinant of both news supply and news consumption.

Table 8: Impact of Headlines' Sentiment on Views - News Agency Fixed Effects and Sentiment of underlying News Issues

	Dependent variable:				
	Log(N Views)				
	(1)	(2)	(3)	(4)	(5)
Sentiment Headline	-0.079^{**}	-0.097	-0.067^{*}	-0.067	-0.071^{***}
	(0.037)	(0.135)	(0.035)	(0.046)	(0.019)
Rank on News Site		-0.008^{***}			
		(0.002)			
Rank on Google News		0.003			
		(0.002)			
Rank on Google Search		-0.002			
		(0.001)			
Sentiment Body					-0.002^{**}
					(0.001)
Sentiment Headline (News Agency)					-0.036^{**}
					(0.017)
Sentiment Body (News Agency)					0.001
					(0.004)
Rank on News site Available (Dummy)			0.549^{***}	0.549^{***}	0.497^{***}
			(0.021)	(0.074)	(0.053)
Rank on News Site (0 for not available)			-0.009^{***}	-0.009^{***}	-0.009^{***}
			(0.001)	(0.002)	(0.001)
Rank on Google News Available (Dummy)			0.029	0.029	0.072^{***}
			(0.022)	(0.032)	(0.010)
Rank on Google News (0 for not available)			0.0003	0.0003	0.001
			(0.001)	(0.001)	(0.001)
Rank on Google Search Available (Dummy)			0.136^{***}	0.136^{***}	0.111^{***}
			(0.029)	(0.042)	(0.031)
Rank on Google Search (0 for not available)			-0.0004	-0.0004	-0.0003
			(0.001)	(0.001)	(0.0004)
News Agency Article FE	*Yes*	*Yes*	*Yes*	*Yes*	*No*
Clustered Stand. Errors	*No*	*No*	*No*	*Yes*	*Yes*
Observations	19,509	4,517	19,509	19,509	19,509
R^2	0.701	0.910	0.731	0.731	0.283

*p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

Note: This table shows result from OLS regressions with news-site and date fixed-effects. Each observation is a single article. Standard errors that are clustered are clustered on the news-site level.

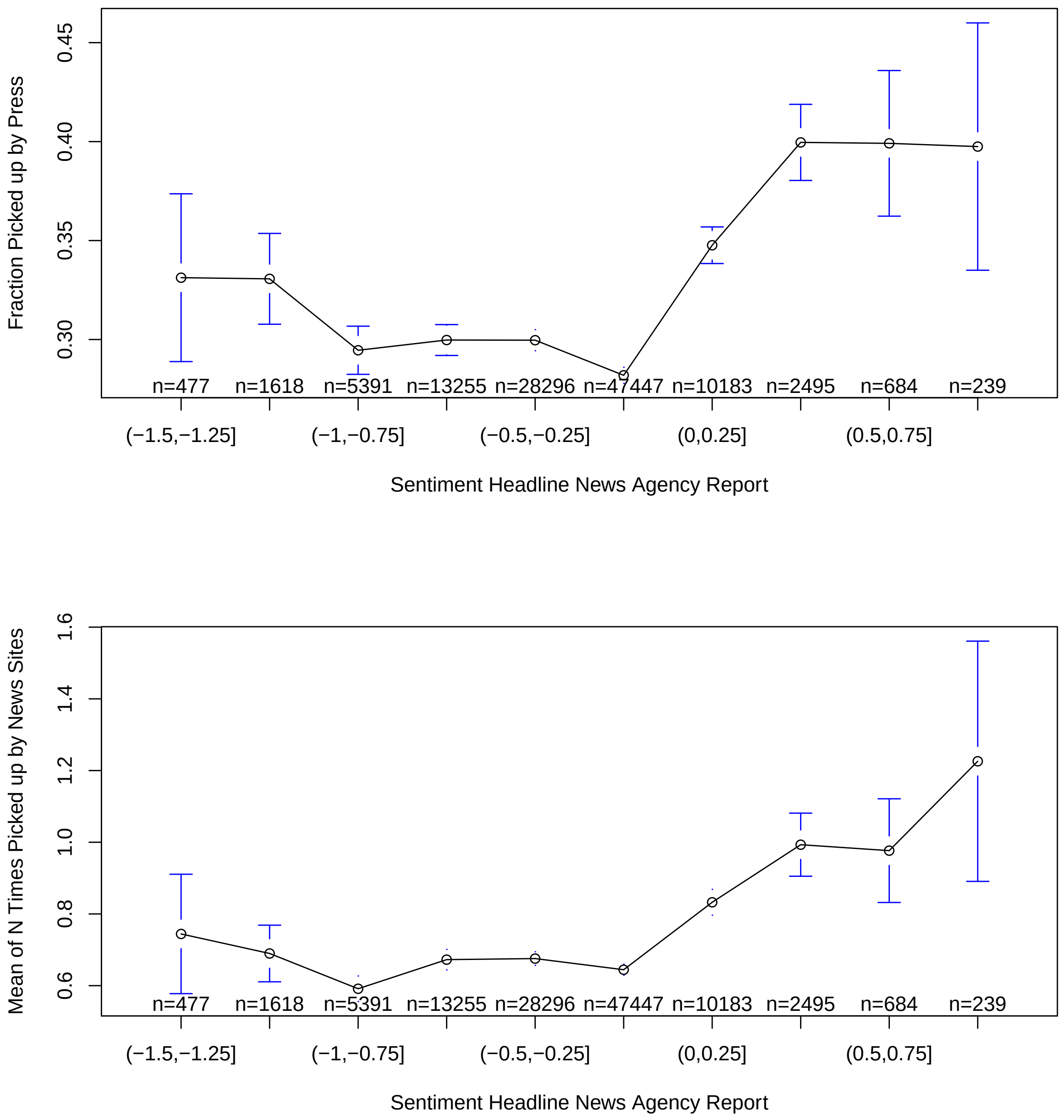

Figure 5: Probability that a given news agency report is picked up by at least one news site (upper panel) and number of different news sites that a given news agency report is picked up by (lower panel)

Table 9: Impact of Sentiment of News Agency Report 5

	Dependent variable:			
	picked.up.by.press		n.picked.up.by.press	
	logistic		*OLS*	
	(1)	(2)	(3)	(4)
Sentiment Headline	-0.009	0.182^{***}	0.024	0.152^{***}
	(0.022)	(0.031)	(0.017)	(0.024)
Sentiment Headline (Squared)		0.300^{***}		0.198^{***}
		(0.034)		(0.027)
Published by Deutsche Presse Agentur (dpa)	-0.933^{***}	-0.915^{***}	-0.342^{***}	-0.332^{***}
	(0.021)	(0.021)	(0.013)	(0.013)
Published by Associate Press (ap)	0.428^{***}	0.437^{***}	0.460^{***}	0.465^{***}
	(0.015)	(0.015)	(0.012)	(0.012)
Constant	-0.851^{***}	-0.856^{***}	0.603^{***}	0.600^{***}
	(0.012)	(0.012)	(0.009)	(0.009)
Observations	110,303	110,303	110,303	110,303
R^2			0.032	0.033
Akaike Inf. Crit.	129,966.800	129,893.500		

*p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

Note: Each observation is a single article. Date fixed effects are not included in these regressions.

References

Beckmann, K. B., Dewenter, R., and Thomas, T. (2017). "Can news draw blood? the impact of media coverage on the number and severity of terror attacks." *Peace Economics, Peace Science and Public Policy*, *23*(1).

Besley, T., and Burgess, R. (2002). "The political economy of government responsiveness: Theory and evidence from India." *Quarterly Journal of Economics*, *117*(4), 1415–1451.

Boik, A., Greenstein, S., and Prince, J. (2016). "The empirical economics of online attention." *National Bureau of Economic Research, Working Paper No. 22427.*

Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). "Salience and consumer choice." *Journal of Political Economy*, *121*(5), 803–843.

Coase, R. H. (1974). "The market for goods and the market for ideas." *American Economic Review*, *64*(2), 384–391.

Dewenter, R., Heimeshoff, U., and Thomas, T. (2016). *Media coverage and car manufacturers' sales.* 215, DICE Discussion Paper.

Djankov, S., McLiesh, C., Nenova, T., and Shleifer, A. (2003). "Who owns the media?" *Journal of Law and Economics*, *46*(2), 341–382.

Dulleck, U., Hackl, F., Weiss, B., and Winter-Ebmer, R. (2008). "Buying online: Sequential decision making by shopbot visitors."

Friebel, G., and Heinz, M. (2014). "Media slant against foreign owners: Downsizing." *Journal of Public Economics*, *120*, 97–106.

Galtung, J., and Ruge, M. H. (1965). "The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers." *Journal of Peace Research*, *2*(1), 64–90.

Gelman, A., and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). "Text as data." *NBER Working Paper.*

Gentzkow, M., and Shapiro, J. M. (2010). "What drives media slant? Evidence from us daily newspapers." *Econometrica*, *78*(1), 35–71.

Gomaa, W. H., and Fahmy, A. A. (2013). "A survey of text similarity approaches." *International Journal of Computer Applications*, *68*(13), 13–18.

Haagerup, U. (2015). *Constructive news: Why negativity destroys the media and democracy-and how to improve journalism of tomorrow.* InnoVatio Publishing AG.

Hotchkiss, G., Jensen, S., Jasra, M., and Wilson, D. (2004). "The role of search in business to business buying decisions a summary of research conducted." *Enquiro White Paper.*

Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). "Revealed attention." *American Economic Review*, *102*(5), 2183–2205.

McCluskey, J. J., Swinnen, J., and Vandemoortele, T. (2015). "You get what you want: A note on the economics of bad news." *Information Economics and Policy*, *30*, 1–5.

Molek-Kozakowska, K. (2013). "Towards a pragma-linguistic framework for the study of sensationalism in news headlines." *Discourse & Communication*, *7*(2), 173–197.

Piotrkowicz, A., Dimitrova, V., Otterbacher, J., and Markert, K. (2017). "The impact of news values and linguistic style on the popularity of headlines on Twitter and Facebook." In *Proceedings of the Second International Workshop on News and Public Opinion (ICWSM NECO 2017)*, Association for the Advancement of Artificial Intelligence.

Proksch, S.-O., Lowe, W., Wäckerle, J., and Soroka, S. (2018). "Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches." *Legislative Studies Quarterly.*

Raymond, C., and Taylor, S. (2013). ""Tell all the truth, but tell it slant": Testing models of media bias." *Working paper.*

Reis, J., Benevenuto, F., de Melo, P. V., Prates, R., Kwak, H., and An, J. (2015). "Breaking the news: First impressions matter on online news." In *Proceedings of the 9th International AAAI Conference on Web-Blogs and Social Media.*

Salant, Y., and Rubinstein, A. (2008). "(a, f): choice with frames." *The Review of Economic Studies*, *75*(4), 1287–1296.

Simon, H. A. (1971). "Designing organizations for an information rich world." In M. Greenberger (Ed.), *Computers, communications, and the public interest*, 37–72, Baltimore.

Thompson, N., and Hanley, D. (2018). "Science is shaped by wikipedia: evidence from a randomized control trial." *Working Paper.*

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Trussler, M., and Soroka, S. (2014). "Consumer demand for cynical and negative news frames." *The International Journal of Press/Politics*, *19*(3), 360–379.

Vosoughi, S., Roy, D., and Aral, S. (2018). "The spread of true and false news online." *Science*, *359*(6380), 1146–1151.

Young, L., and Soroka, S. (2012). "Affective news: The automated coding of sentiment in political texts." *Political Communication*, *29*(2), 205–231.