Information Processing and Management 60 (2023) 103398

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Content still matters. A machine learning model for predicting news longevity from textual and context features

Krzysztof Rybinski [a,b,c,d]

[a] Department of Business and International Relations, Vistula University,Warsaw, Poland
[b] European Humanities University, Vilnius, Lithuania
[c] Vistula University, Stoklosy 3, 02-787 Warsaw, Poland
[d] European Humanities University, Saviciaus 17, 01-127 Vilnius, Lithuania

ARTICLE INFO

Keywords:
News longevity
Sentiment analysis
Emotion detection
Topic modelling
Natural language processing

ABSTRACT

There is an ongoing debate about what is more important in the modern online media newsroom, whether it is the news content and worthiness, or the audience clicks. Using a dataset of over one million articles from five countries (Belarus, Kazakhstan, Poland, Russia, and Ukraine) and a novel machine learning methodology, I demonstrate that the content of news articles has a significant impact on their lifespan. My findings show that articles with positive sentiment tend to be displayed longer, and that high fear emotion scores can extend the lifespan of news articles in autocratic regimes, and the impact is substantial in magnitude. This paper proposes four new methods for improving information management methodology: a flexible version of Latent Dirichlet Allocation (LDA), a technique for performing relative sentiment analysis, a method for determining semantic similarity between a news article and a newspaper's dominant narrative, and a novel approach to unsupervised model validation based on inter-feature consistency.

1. Introduction

The emergence and dominance of online media as a source of information has challenged the traditional strategy of news publication and management. The gatekeeping practices of the newsroom that focused on newsworthiness, media reputation or breadth of coverage were confronted with the availability of timely and very detailed data on audience preferences. A major research question has emerged. What is more important in the modern media newsroom, is it the news content and worthiness, or what the audience wants? This paper provides a new dataset and a novel machine learning methodology to show that the news content remains highly relevant and influences the lifespan of the news.

The professional culture of traditional journalism tends to resist the incorporation of audience data in news posting decisions. However, the increasing availability of such data has led to a shift in the role of reporters and news editors, who are becoming more reliant on audience metrics in their decision-making processes. Previous research has also demonstrated the influence of audience engagement on news placement in online newspapers, with studies showing that audience clicks affect the positioning of subsequent news articles (Lee, Lewis, & Powers, 2014) and that the most viewed articles are more likely to shape subsequent reporting (Welbers, van Atteveldt, Kleinnijenhuis, Ruigrok, & Schaper, 2016).

Editors, however, consider that using news popularity ratings to select news stories conflicts with their professional norms, which should value newsworthiness rather than popularity (Welbers, van Atteveldt, Kleinnijenhuis, & Ruigrok, 2018; Welbers, van Atteveldt,

E-mail address: rybinski@rybinski.eu.

https://doi.org/10.1016/j.ipm.2023.103398
Received 4 January 2023; Received in revised form 15 April 2023; Accepted 20 April 2023
Available online 25 April 2023

Kleinnijenhuis, Ruigrok, & Schaper, 2016). This view continues to shape the gatekeeping policies of newsrooms. Nicholls and Bright (2019) use a corpus of nearly 40,000 articles in the British media and show that more than 50% of news production takes the form of story chains. The gap between the news preferences of users and the news choices of editors remains significant (e.g. Boczkowski & Mitchelstein, 2015). A survey of U.S. internet users showed that only about one-third of the content produced by mainstream news media was perceived as worth reading (Lee & Chyi, 2014).

This ongoing debate about the evolution of news production and management practices has inspired many studies of the life cycle of news. Boczkowski (2010) demonstrated that journalists strongly value speed of publishing and routinely observe the news sites of competitors. On average, it takes only 1.5 h for most German online news sites to report the news (Buhl, Günther, & Quandt, 2018). User activity data has been used to explain the longevity of news, meaning how long a news article is displayed on the newspaper's main website. Bright and Nicholls (2014) explore the relationship between the readership of a news story and the likelihood of it being removed from the front page by looking at the "most read" lists that are commonly found on many news websites. They find that an article being "most read" reduced the likelihood of it being removed from the front page by around 25%. A massive amount of literature is devoted to predicting the popularity of news items in social media using user activity metrics and the digital traces of users in general, with more than 100 papers reviewed in two meta-studies (Moniz & Torge, 2019; Rousidis, Koukaras, & Tjortjis, 2020).

The relationship between the activity of users and the production and management of news in online media has been well researched. There is, however, a significant gap in the literature on the relationship between news content and news longevity in the mainstream online media. Examining the lifespan of news offers valuable insights into the duration of news stories and their retention of public attention. Analysis of the news life cycle additionally sheds light on the efficacy of various approaches for circulating news. Furthermore, the persistence of a news item can serve as a gauge of its impact and offer valuable insights into the media's role in shaping public opinion and discourse.

One point that has been widely documented is that negative news is more popular and receives more clicks or views (e.g., Soroka, Fournier, & Nir, 2019; van der Meer, Kroon, Verhoeven, & Jonkman, 2019). Two questions remain to be answered though. Do features of the news narrative affect the news longevity, or is it all about the user clicks? And if they do, what are these features? Finally, one question is addressed for the first time in the literature, which is whether there are any differences in the lifespans of news in democratic and autocratic countries.

This paper addresses these questions, and the main results and contributions are the following:

- A unique corpus is created of over one million articles from 17 newspapers in two languages and five countries. This corpus is suitable for news longevity analysis because the articles were scraped daily between mid-2017 and January 2021, so it is shown when an article appeared on and disappeared from a news website;
- New types of features based on the news narrative are developed. The paper uses dense contextual embeddings and cosine distance to measure the breadth of coverage of a newspaper and the similarity of the thrust of a news article to the main narrative of the issue. It is shown that both types of feature help to predict news longevity.
- How emotions and sentiment polarity affect the longevity of news is analysed. The paper shows for the first time that: (1) positive news lives longer and negative news lives shorter; (2) the emotion "fear" extends the news lifecycle in autocratic countries.
- A new method is formulated for validating the results of zero-shot unsupervised learning. The method, called inter-feature consistency, is used to validate the results presented in this paper.

The remainder of this paper is organised as follows. Section 2 gives a scoping review of the related work, and the research hypotheses are derived from that work. Section 3 describes the data collected. Section 4 presents the novel methodology, which consists of cosine similarity of articles to measure the breadth of coverage, topic modelling, detection of the names of influential politicians, a measure of relative sentiment, emotion detection, and the structure of the logistic regression model. Section 5 then presents the results of the model and Section 6 discusses and validates them including use of the proposed novel validation technique. Section 7 concludes, and highlights the directions of future research.

2. Related work and research objectives

As discussed in the introduction section, many research papers have analysed the gatekeeping practices for news and the longevity of news in online media. To the best of my knowledge, this is the first study to analyse how sentiment polarisation of an online newspaper article, its emotional composition, its topic, and its similarity to the main themes of the daily issue of the newspaper affect the lifespan of the news.

2.1. The similarity between the theme of the news and the main narrative of the daily issue of the newspaper

Although online newspapers use personalisation algorithms to some extent, they cover essential stories on the main page and show them to all users, regardless of their digital trace properties. Experiments simulating human browsing behaviour on online newspaper sites showed little personalisation (Loebbecke, Oberschulte, & Boboschko, 2021), which means we can measure and analyse how diverse or homogenous the articles displayed on the main page are and how their narratives relate to the overall narrative or focus of the daily issue of the paper. There are opposing forces at play here. One is that the editor's traditional priorities, such as the breadth of coverage or the customer trust and loyalty that are secured by the quality of the newspaper, cannot be compromised by short-term financial goals (Bodo, 2019). Counter to this, the traditional gatekeeping theory has been challenged by the emergence of theories

of digital gatekeeping (Wallace, 2018), which propose new possible configurations of gatekeepers, platforms and mechanisms and allow a wide range of cross-channel interactions, including more attention being paid to the audience's short-term preferences, which are important for advertisers.

Many media outlets are also under pressure to cut costs, which encourages churnalism and copy-paste work, and reduces the breadth of coverage and original reporting, as was documented across online Flemish media (Hendrickx, 2020) and for political stories published by most popular Greek news websites (Saridou, Spyridou, & Veglis, 2017). The debate about news diversity is, however, far from decided, as a longitudinal study of 6 million articles from 12 newspapers in four European countries has shown that the diversity of news has slightly increased over time (de Vries, Vliegenthart, & Walgrave, 2022).

Interestingly, the diversity of news has been analysed between newspapers and news websites. To the best of my knowledge there is no study of the news diversity within a single media outlet and how it influences the longevity of the news. The existing literature does not offer any guidance on the relationship between the similarity of the news story to the dominant narrative in the daily issue of the paper, and the life cycle of the news. On the one hand, an article that is similar to the dominant narrative in the newspaper's daily issue is less likely to remain on the website the next day. The dominant narrative, such as the passing of controversial legislation or a surprising rate of inflation, is expected to generate a large supply of competing articles, making each of them less likely to live long. Furthermore, articles that depart from the main narrative of the daily issue provide highly-valued breadth, so their life cycle may be longer. On the other hand, an important article that sets the tone for the daily issue may be cited by other media and generate a lot of views, and so is likely to have a longer lifespan. Such an article may also be amended with comments and new developments, which increases the odds that it will remain longer on the main page. I believe this argument prevails on balance, which is reflected in hypothesis H1a below.

Finally, if an issue discusses many topics, the likelihood of an article staying on the main page for another day is higher than when the newspaper focuses on fewer topics. Hence, our first hypotheses are:

H1a. A high level of similarity of the content of the news to the main focus or narrative of the daily issue extends the news life cycle.

H1b. A high degree of diversity in the topics covered on the main page in a given daily issue raises the odds that articles published that day will remain longer on the main page.

As argued above, I expect that when testing for Hypothesis H1a, I will find more supporting evidence in the case when a slight modification of a news story is allowed than in the case when it is required to be exactly the same the next day.

2.2. The dominant news topic

Research has shown that there is a strong relationship between the topic of a news story and its popularity measured by the number of clicks, views or comments. Stories that garner large numbers of clicks focus on fun topics, livelihood stories, or sensational stories such as death, violence, sex and surprise (Boczkowski & Mitchelstein, 2015). News stories covering politics, economics and controversial topics received the most comments (Tenenboim & Cohen, 2015).

For many online media outlets the frequency of political news stories is higher than the frequency of articles discussing cultural events. This makes the competition for the main page space more intense for political news than for news about cultural events, and so political news may consequently have a shorter life cycle than culture news. I believe that despite receiving a large number of comments, the argument of high frequency of political news prevails, and that articles discussing politics have a shorter life cycle. Many articles cover more than one topic, but in most cases, the dominant topic or topics can be determined using unsupervised machine learning methods. The second hypothesis states:

H2. The dominant topic or topics of the news article affects the likelihood of the article remaining on the main page the next day.

2.3. Positive and negative news sentiment

Exposure to positive and negative news affects human behaviour. An infamous Facebook experiment (Kramer, Guillory, & Hancock, 2014) targeted 689,003 social network users and showed that people reading predominantly positive or negative news experienced positive or negative emotions without being aware of it. When the researchers reduced the positive expressions in the News Feeds of Facebook users, people produced fewer positive posts and more negative ones; when negative expressions were reduced, the opposite pattern occurred. While this experiment raised many moral concerns, it showed that the inclination of news sentiment could profoundly impact the attitudes and actions of consumers of media.

Sentiment analysis has often been used for forecasting the popularity of news. Many research papers have found that negative news receives more coverage than positive news and that it is more popular (e.g., Soroka et al., 2019; van der Meer et al., 2019). Many studies have documented that emotionally charged posts, especially negative ones, travel faster and generate more feedback and shares than neutral ones (e.g., Zhu, Kim, & Park, 2020). The relationship between the negativity of news and the engagement of readers is far from settled, however. Some studies have shown that positive content attracts more attention, generates more feedback, and spreads more widely than negative content (Ferrara & Yang, 2015; Nave, Shifman, & Tenenboim-Weinblatt, 2018). Muddiman, Pond-Cobb, and Matson (2020) conducted two experiments and showed that articles that included multiple types of incivility discouraged people from engaging with online news. However, the empirical evidence about online newspapers favours the negativity effect of news, and so the third hypothesis tested in this paper is as follows:

H3. Negative news sentiment increases the odds of the news remaining on the main page the next day.

2.4. Basic emotions

Detecting emotions from text is a field that has been intensely researched, and where several review papers have been published recently. Acheampong, Nunoo-Mensah, and Chen (2021) provide an in-depth and comprehensive review of state-of-the-art, transformer-based deep learning models applied to emotion detection from text. A general discussion of a wide range of machine learning models applied to emotion extraction is presented in Nandwani and Verma (2021) and Murthy and Kumar (2021). Somewhat unexpectedly given their popularity, the emotion detection models have not been applied to study the dynamics of news. In a rare exception, Gupta and Yang (2019) found that headlines with any highly intense emotions are significantly more popular. They showed that adding emotion scores to the other features of articles improved the predictive power of models forecasting the popularity of news.

It is not plausible to assume that formal newsroom gatekeeping practices include verifying the emotional composition of news, as the basic emotion scores are latent variables. However, emotionally charged articles can live longer because of their greater popularity and this channel may have an impact on the longevity of news. This paper applies advanced deep learning models to extract the basic emotions from the texts of articles and puts forward the following hypothesis:

H4. The intensity of certain emotions influences the likelihood of the article remaining on the main page the next day.

2.5. Differences between democracies and autocracies

To date there have been no studies devoted to identifying the differences in the dynamics of news between democracies and autocratic countries, although Maerz and Schneider (2020) compare public communication in 27 countries that include both regime types. This is surprising given that autocracies do not have a free media, and so their news gatekeeping practices serve different objectives to those in democracies where the media are free. While the gatekeeping process in autocracies is very complex, it is plausible to assume that articles featuring the ruling autocrats will stay longer on the main page.

In autocratic countries, those in power often use fear as a tool to control the population and maintain their grip on power. This can manifest in various ways in the media, such as through the dissemination of propaganda that promotes fear of the consequences of dissent or opposition to the ruling regime. Political propaganda in autocratic countries is not just rhetoric aimed at weakening a political foe. It is a larger force intended to increase fear and anxiety and divide society against itself (LibertiesEU, 2021). One of the main pillars of propaganda is the inducing of fear and hatred in the audience (Imedashvili, 2022). Fear control is also a technique widely used by certain interest groups to generate and spread uncertainty amongst people in order to create an atmosphere in which their goals are easily reachable (Marco, 2013). Thus, narratives that promote fear are the "dark side" of legitimation of autocratic regimes (Joshua, 2021). The news corpus contains newspapers from democratic Poland and Ukraine and from autocratic Belarus, Kazakhstan and Russia, making it possible to put forward the following two hypotheses:

H5. News featuring influential domestic politicians will have a longer life cycle in autocratic countries than in democracies.

H6. Negative emotions, especially fear, extend the news lifecycle in autocracies.

Table 1
. Newspapers and portals included in the analysis and their Alexia ranks.

Country	Newspaper site	Number of articles	Percentage of articles featured on more than one day (same/almost the same) *	Alexia local rank
Russia	iz.ru	43,782	1.58 / 1.80	45
Russia	kommersant.ru	46,070	27.2 / 28.6	44
Russia	novayagazeta.ru	29,357	48.8 / 51.7	459
Russia	vedomosti.ru	27,797	49.2 / 51.3	288
Kazakhstan	informburo.kz	29,375	51.2 / 53.2	119
Kazakhstan	nur.kz	67,350	26.4 / 26.6	6
Kazakhstan	tengrinews.kz	44,285	10.4 / 12.8	34
Kazakhstan	zakon.kz	109,442	13.4 / 14.0	30
Belarus	bdg.by	33,447	49.7 / 81.4	746
Belarus	belgazeta.by	21,995	52.8 / 87.2	11,041
Belarus	sb.by	83,685	20.4 / 29.5	79
Ukraine	kp.ua	194,792	37.8 / 64.6	860
Ukraine	segodnya.ua	45,835	4.88 / 9.42	256
Ukraine	vesti.ua	90,559	20.7 / 34.7	1096
Poland	gazeta.pl	53,321	9.9 / 10.4	14
Poland	rp.pl	49,587	16.6 / 17.2	167
Poland	wpolityce.pl	76,625	31.7 / 32.9	105

Notes. Alexia's rank, 90-day average, was checked on 17 February 2021. The total number of articles is 1047,304. * Allowing slight modification of the article's text as explained in the methodology section.

3. Data description

The data were collected daily starting from June 2017 for Poland and Kazakhstan, and January 2018 for Russia, Belarus and Ukraine, and running up to January 2021. Table 1 presents the data sources, with the number of articles scraped, the percentage of articles still available the next day on the main page, and the local Alexia ranks. The articles were scraped in the afternoon each day, but the time was not exactly the same each day, so when the same or almost the same article is found the next day on the main newspaper website, it means that the minimum length of time that the article was featured for was between 20 and 28 h.

The websites, except those from Kazakhstan and some from Belarus, are the online platforms of popular or influential newspapers that run paper editions. The corpus covers a broad spectrum of newspapers across many dimensions, taking in liberal and conservative sources, outlets controlled by the state and those controlled by oligarchs, and pro-government and anti-government papers. This diversity in the newspapers allows for testing of whether the estimates of news longevity are robust to different gatekeeping practices.

The dataset for replicating the results is available on Zenodo at the following link: https://doi.org/10.5281/zenodo.6591100.

4. Methodology

All the features extracted from the texts of the articles are listed in Table 3. I use contextual dense embeddings to calculate features from cosine similarity. Topic modelling is conducted within the bag-of-words framework. Sentiment analysis uses a dictionary approach, and emotion detection uses a zero-shot classification with a deep neural network. I apply state-of-the-art natural language processing models with hundreds of millions of parameters trained using self-supervised methods on billions of words. Such models have a deep understanding of the semantic structures of language, understand the different meanings of the exact words depending on the context, and achieve superhuman accuracy in many language tasks (Wang et al., 2019). All the steps are described in detail below and shown in Fig. 1. Sentiment analysis, text mining, and topic modelling were conducted in R, while the embedding calculations, emotion detection and regressions were run in Python. In Appendix C of the Supplementary material, I present a list of references to papers that provide mathematical formulas for other methods employed in this research.

4.1. Tokenisation, embeddings and the cosine similarity of articles

Some studies have analysed the similarity of news articles, but they use either human assessment of similarity (Hendrickx, 2020; Timmerman & Bronselaer, 2022 for detecting news accuracy), human-operated anti-plagiarism software (Saridou et al., 2017), or bag-of-words models and analysis of the co-occurrence of selected words (de Vries et al., 2022; Welbers et al., 2016). While the cosine similarity of texts has recently become a popular approach for studying language properties (Alsudasis, Alotaibi, & Alomary, 2022) or recommendation engines (Agrawal, Roy, & Mitra, 2021), to the best of my knowledge this is the first attempt to apply dense embeddings and cosine distance to measure the similarity of news articles.

Although cosine similarity measures based on dense contextual embeddings have numerous advantages, such as their ability to capture the semantic meaning of words and their context, they are not without limitations. They are computationally intensive and require large amounts of textual data, and they may be sensitive to pre-processing steps such as tokenization. Additionally, they may struggle to handle words that are not present in the vocabulary used during training. The LaBSE model described below has been trained on a vast amount of data, including 1.7 billion sentences in Russian (ranked #2 amongst covered languages) and 0.35 billion sentences in Polish (ranked #10). It also features extensive vocabularies in both languages. Therefore, the limitations associated with cosine similarity measures discussed above do not apply to the dense contextual embeddings generated by this model.

The first step is to tokenise the articles and calculate their embeddings. Words are converted into numerical tokens so that each sentence becomes a vector of numerical tokens. Each tokenised sentence is then converted into a unique embedding that represents that sentence, giving a vector of 768 real numbers. These steps are performed using Google's LaBSE model[1] (Feng, Yang, Cer, Arivazhagan, & Wang, 2020), which can compute sentence embeddings for 109 languages, including Russian and Polish. Finally, I calculate the element-wise average for all the sentence embeddings in the article to obtain the unique embedding vector for each article. These article embeddings are then used to calculate the cosine similarity between articles.

If the same article is published the next day, the cosine similarity of the two articles will be one. However, if slight modifications were made, such as adding a sentence or modifying the text, the cosine similarity between the articles will be close to, but below, one. The cosine similarity is between zero and one, with similar articles having higher cosine similarity than articles that discuss different topics.

Examples of cosine similarity calculated for LaBSE embeddings for short texts are presented below with English translations of texts in Polish.

Text1 = "Russia located 100 thousand troops near the border with Ukraine. Experts predict that invasion may take place within one month."
Text2 = "Russia located 100 thousand troops close to the Ukrainian border. Experts predict that invasion may take place within one month."

[1] The language-agnostic BERT Sentence Embedding model, part of the BERT family of models. The section on emotion detection discusses briefly how these models are trained.

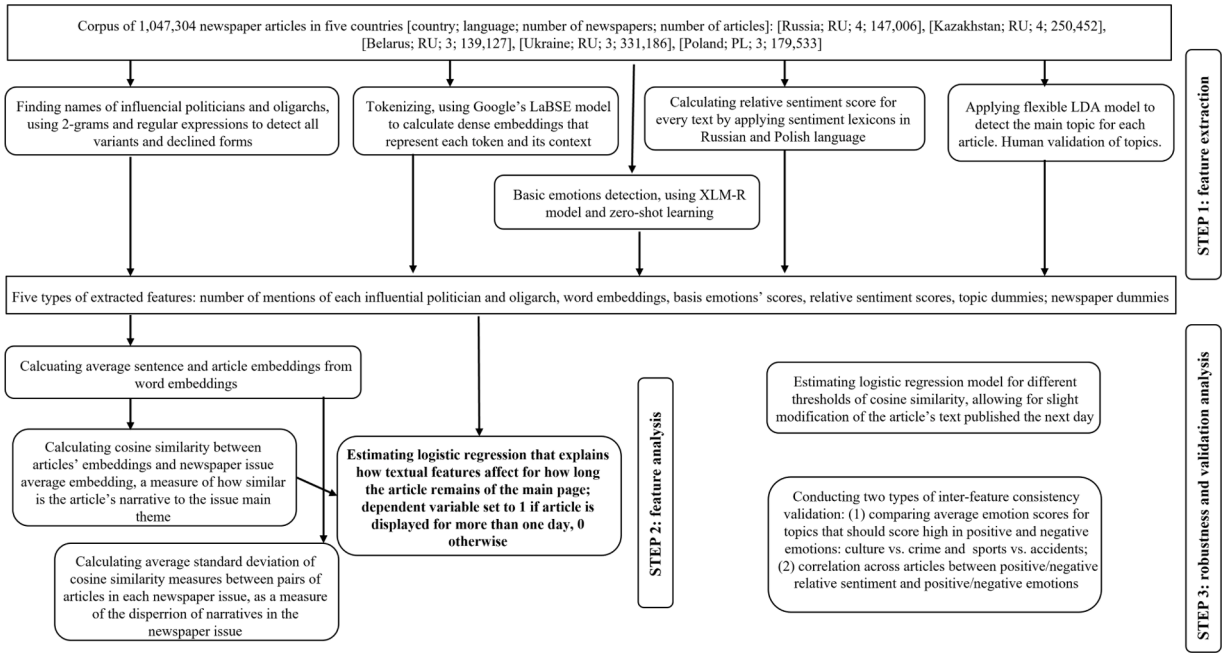

Fig. 1. Schematic overview of the methodology. The main goal is to verify whether the news textual features can explain the news lifespan.

Text3 = "Russia located 100 thousand troops near the border with Belarus. Experts predict that invasion may take place within one month."
Text4 = "Russia halted natural gas exports. Experts predict that natural gas prices will continue to rise."

$$\text{cosine\_similarity}(\text{Text1}, \text{Text2}) = 0.9939$$
$$\text{cosine\_similarity}(\text{Text1}, \text{Text3}) = 0.9826$$
$$\text{cosine\_similarity}(\text{Text1}, \text{Text4}) = 0.6699$$

The cosine similarity is also used to define the variables that are used in the logistic regressions (Table 3).

4.2. Topic modelling

I identify the topic of each article by applying flexible Latent Dirichlet Allocation (LDA). The standard LDA model (Blei, Ng, & Jordan, 2003) requires a fixed number of topics to be selected and assigns each article to a given single topic. In this paper, I use the flexible version of the model applied in Charemza, Makarova, and Rybinski (2022) to a large corpus of Russian newspaper texts.

In the standard topic modelling analysis, an a priori decision is made on the number of latent topics K, and then the analysis is conducted for various K to determine which value is the best choice. In the flexible LDA model, the maximum number of latent topics is decided, and in our case, we set $K = 30$ as suggested in Charemza et al. (2022). During the human validation phase, researchers assign single names to the topics like 'politics' or multiple names like 'politics_finance', depending on the set of words describing each of the 30 latent topics. Finally, the topic dummy variables are created. When the latent topic was assigned to more than one category, such as politics_finance, all the documents in this category will have the dummy variable 'politics' and the dummy variable 'finance' set to one. Our analysis gave 21 human-determined categories of politics (POL); legislation and legal affairs (LEG); economy and industrial sectors (ECO); finance (FIN); military, war and national security (MIL); protests (PRO); crime (CRI); international affairs and foreign countries (INT); technology (TECH); culture (CUL); sport (SPO); education (EDU); family (FAM); transport (TRA); regional issues and housing (REG); health and the Covid-19 pandemic (HEA); media (MED); accidents (ACC); religion (REL); the Soviet Union (USSR); and articles with no topic determined (MISC).

The standard LDA model requires a single topic classification for each text, say, politics, even if the keywords for that latent topic include some that a human researcher would classify as finance. The heatmap analysis by Kim, Ahn, and Jung (2019) shows that this occurs in the data. The proposed flexible LDA with human validation classifies such texts into two topic categories, POL and FIN, because the latent topic name is politics_finance. This approach works well when a corpus contains very different newspapers, such as broadsheets and tabloids, and newspapers from several countries, as ours does. The complete list of LDA-generated words in the original languages used to name the latent topics for each newspaper is available upon request from the author. Appendix A in the Supplementary material gives examples with English translations.

4.3. Text mining for the names of influential domestic politicians

The proposed model verifies how mentions of influential domestic politicians affect the likelihood of the article being displayed the next day. Because the countries analysed have both democratic and autocratic power structures, the criteria for identifying influential politicians were:

- Current President and Prime Minister.
- Former Presidents and former Prime Ministers if a large part of their terms in office overlapped with the period analysed.
- The leaders of the party or parties in the ruling government.
- Heads or first deputy heads of a presidential administration who hold high positions in the influence rankings.
- Family members of influential politicians who hold important political or business positions.
- Heads of the national security agency or committee and long-serving ministers of defence, security or foreign affairs with high positions in the influence rankings.
- Oligarchs in autocratic countries with high positions in the influence rankings.

4.4. Relative sentiment

I use two sentiment lexicons with sentiment polarity manually annotated by human researchers, applying Dziob, Piasecki, and Rudnicka (2019) for Polish and Loukachevitch and Levchik (2016) for Russian. The Polish lexicon uses five levels of polarity, with −2 (very negative), −1 (moderately negative), 0 (neutral), +1 (moderately positive) and +2 (very positive), while the Russian lexicon has three levels of −1 (negative), 0 (neutral) or +1 (positive) .[2]

To apply the lexicon, I check whether each word in an article appears in the lexicon, then add its sentiment polarity to the sentiment score for the article. Words that are not in the lexicon are ignored. Finally, the sentiment score for the article is divided by the number of words in it.[3]

This algorithm is a standard way of applying sentiment analysis using human-generated lexicons[4] to big text data. However, it may not be appropriate for cross-country and multiple language research. Each country, language, or even newspaper may have its own idiosyncrasies, resulting in higher or lower average sentiment scores, or country-specific sentiment effects. I control for such factors by calculating the relative sentiment for each article A in a newspaper or portal P in the corpus. The relative sentiment is the difference between the sentiment score of article A and the average sentiment of all the articles in the newspaper or portal P ($sent_i$ in Table 3).

4.5. Emotion detection

I use the NLP model based on deep neural networks with transformer infrastructure[5] to extract six basic emotions from the newspaper articles. The model is XLM-R, as proposed in Conneau et al. (2019). XLM-R is a multilingual version of the RoBERTa model (Liu et al., 2019). 'Ro' means robustly optimised, 'a' means approach, and 'BERT' stands for Bidirectional Encoder Representations from Transformers.

In simplified terms, Conneau et al. (2019) train the model with a training corpus of text in 100 languages, including Russian and Polish, from the two-terabyte CommonCrawl Corpus. The text is first tokenised or split into words that are converted into numbers or tokens. These numbers are then converted into embeddings, which are numerical vectors with a length of 1024,[6] which are fed into the neural network. Then 15 per cent of the tokens are randomly masked, and the neural network is trained to predict the masked tokens. This is called self-supervised learning. Each model prediction is compared with the masked token, the loss function is computed, and backpropagation is used to recalculate the model weights from the loss function value.

From the training, the model learns contextual embeddings for words, so words that have similar meanings also have similar embedding vectors. The expression for the word embeddings: "king" – "man" + "woman", for example, will give the embedding representing the word "queen". The model learns the language representation for 100 languages in this way. It can be used for many

[2] Various methods have been proposed in the literature for defining polarity scales for sentiment lexicons. In their review, Khoo et al. (2018) discuss six such practices, while Ribeiro et al. (2016) examine 24 English-language lexicons. These scales may comprise discrete levels of polarity, ranging from two to five or more, as well as different continuous scales. The evaluation metrics used for these lexicons (such as accuracy and F1 scores) may vary when calculated using labeled datasets. Consequently, the validation of research hypothesis H3 (i.e., negative news has a longer lifespan) may depend on the choice of lexicon and its polarity scale. Nonetheless, as demonstrated in the results section, the estimates of the polarity effect on news longevity are highly significant across countries ($p < 0.0001$). This suggests that the results are robust to differences in lexicon choice and scoring methodology.

[3] This allows the sentiment scores of articles of different lengths to be compared.

[4] It is good practice to validate the sentiment scores obtained by drawing a random sample of articles and asking human readers to rate the sentiment of these articles. The correlation between the human and machine scores should be positive and significant. Such a validation analysis was conducted for all the sentiment lexicons and for the newspaper articles in Rybinski (2018). There were 17 human readers in Russian and eight in Polish, and the correlations were in the range of 0.44-0.93.

[5] See Acheampong et al. (2021) for a discussion on detecting text emotion using BERT-based models.

[6] There are two versions of the model. The large version uses an embedding vector of length 1024, and the base version has length 768. I use the large version of the model from the huggingface transformers library, available at https://huggingface.co/vicgalle/xlm-roberta-large-xnli-anli

language tasks, including zero-shot classification. Zero-shot means that the model can perform text classification for a selected set of descriptors, which in our case are the six basic emotions (Ekman, 1992) of sadness, joy, disgust, fear, anger, and surprise, without needing additional training on a domain-specific corpus. For newspapers in Russian and Polish, we name the emotions in Russian and Polish. The model calculates the probabilities of the text representing each of the emotions. These probabilities are used in the logistic regression analysis. To fit the model into the GPU memory (Nvidia QUADRO RTX 5000, 16GB), it read at most the first 35 sentences from each article and at most 600 characters in each sentence.

Table 2 provides examples of texts translated from Russian to English, and the emotion probabilities calculated by the model.

4.6. Variables used in logistic regression models

The dependent variable next_day_i takes two values and is 1 if the i^{th} article was available the next day, and 0 if it was not.[7] There are two ways to calculate next_day_i. In the first approach, the next-day version of the article must be identical to the original so that the cosine similarity is 1. The second version allows small modifications and considers the article to be the same when the cosine similarity measure is 0.99 or higher.

4.7. Logistic regression models

The purpose of the logistic regression models is to assess the influence of various contextual features of news on the probability of the news remaining on the main website for more than one day. The estimation results are used to test the research hypotheses H1 - H6. The explained variable next_day is binomial, so 0 means the same news was not featured the next day and 1 means it was featured the next day, where the definition of sameness uses cosine similarity, and therefore the binomial logistic regression specified in Eq. (1) is estimated. The logistic regression models were estimated using the maximum likelihood method. The Logit function from the Python statsmodels library was used, with the following parameters: method='newton', cov_type='hc0', which allows for the computation of robust standard errors.

$$\begin{aligned} &P(\text{next\_day}_i = 1) = 1/(1+\exp(-(\alpha + \\ &\beta_1 \text{cosine\_sim}_i + \beta_2 \text{cosine\_std}_i + \beta_3 \text{issue\_size}_i + \sum_m \gamma_m \text{top}_{m,i} + \\ &\delta \text{sent}_i + \zeta \text{dip}_{k,i} + \sum_{p(k)} \chi_{p(k)} \text{news}_{p(k),i} + \\ &\theta_1 \text{anger}_i + \theta_2 \text{fear}_i + \theta_3 \text{love}_i + \theta_4 \text{joy}_i + \theta_5 \text{sadness}_i + \varepsilon_i))) \end{aligned} \tag{1}$$

The binomial logistic regression model in Eq. (1) is estimated separately for each country k ϵ {RU, KZ, BEL, UKR, PL}, for all countries together and for two definitions of the next_day variable. The interpretation of the model parameters is presented in Table 4 and the expected sign of the estimated parameters is explained.

5. Results

This section discusses which factors increase or reduce the odds of a news story being shown the next day. When an article is similar to the main theme of the daily issue of the paper in the four post-soviet countries, the likelihood increases of it remaining on the website the next day (*cosine_sim*, b_{RU}=1.503, p_{RU}<0.0001, b_{KZ}=0.136, p_{KZ}=0.026, b_{BEL}=0.322, p_{BEL}=0.012, b_{UKR}=0.475, p_{UKR}<0.0001) ,[8] but that likelihood decreases in Poland (b_{PL}=−0.495, p_{PL}<0.0001). Hypothesis H1a was positively validated in four of the countries. As expected, I find more substantial evidence supporting Hypothesis H1a in cases where a slight modification of the news contents is allowed. The estimates for the four post-soviet countries are much higher for slightly modified news than for exact news (similar vs. same news story: b_{RU}=1.851 vs. b_{RU}=1.503, b_{KZ}=0.289 vs. b_{KZ}=0.136, b_{BEL}=1.474 vs. b_{BEL}=0.322, b_{UKR}=1.049 vs. b_{UKR}=0.475) while in Poland, the absolute value of the negative estimate is higher (b_{PL}=−0.427 vs. b_{PL}=−0.495).

When the contents of the issue of the paper are very diverse (high *cosine_std*), the likelihood that an article will also be shown the next day increases in Poland, Russia and Kazakhstan (b_{PL}=15.097, p_{PL}<0.0001, b_{RU}=2.485, p_{RU}<0.0001, b_{KZ}=5.039, p_{KZ}<0.0001). In Belarus and Ukraine, the odds shrink (b_{BEL}=−0.629, p_{BEL}=0.087, b_{UKR}=−7.403, p_{UKR}<0.0001). Hypothesis H1b is consequently supported in three countries. The same country pattern is found for the issue size (*issue_size*).

In the vast majority of cases, the topic variables are significant, so Hypothesis H2 is supported. There is considerable heterogeneity across countries for the sign and the magnitude of the topic effect. As expected, an article that discusses politics (POL) is less likely to be available the next day in all the countries (b_{PL}=−0.114, p_{PL}<0.0001, b_{RU}=−0.384, p_{RU}<0.0001, b_{KZ}=−0.397, p_{KZ}<0.0001,

[7] As discussed in the Data description section, the next day variable is set to 1 when the news article appeared on the main newspaper's website for 20 to 28 hours.

[8] Parameter estimates (b) and p-values (p) are reported with country subscripts: PL – Poland, RU – Russia, KZ – Kazakhstan, BEL – Belarus, UKR – Ukraine. Results in the text are reported for the identical text published the next day, unless explicitly specified otherwise; Table 5 also shows the results (b, p) for slightly modified texts published the next day (cosine similarity >=0.99).

Table 2
Sample sentences and emotion scores calculated by the XLM-R model

	sadness	joy	disgust	fear	anger	surprise
The worker was furious because of low wages.	0.00	0.00	0.22	0.08	**0.64**	0.07
The foreigner was very much in love.	0.01	**0.53**	0.01	0.01	0.01	0.43
He was alone in the woods, surrounded by hungry wolves.	0.02	0.08	0.02	0.06	0.01	**0.81**
Jane was terrified as she felt unprepared for the upcoming exam.	0.05	0.00	0.17	**0.51**	0.01	0.27
Jim roommates disliked his habit to eat old food that smelled badly.	0.02	0.02	0.38	0.01	0.04	**0.53**
Many people were crying during her funeral.	**0.94**	0.00	0.05	0.01	0.00	0.00

Note. The model was applied to short texts in Russian. The table shows their Google translation into English. The maximum score for the row is in bold.

Table 3
. Independent variables used in the logistic regression models. The dependent variable is next_day$_i$ described in Section 4.6.

Name	Description	Method of calculation
$cosine\_sim_i$	A measure of how similar or different the i^{th} article is to the average narrative of the newspaper issue that published the i^{th} article	Cosine similarity between the i^{th} article embedding, and the elementwise average of embeddings of all the articles published in the same issue.
$cosine\_std_i$	A measure of the dispersion of narratives of a newspaper issue that includes the i^{th} article.	Standard deviation of the cosine similarity measures between all pairs of articles in the newspaper issue that published the i^{th} article.
$issue\_size_i$	Size of the newspaper issue that published the i^{th} article	Number of articles scraped from the website on the day when the i^{th} article was published
$sent_i$	Relative sentiment score of the i^{th} article	Described in section on relative sentiment
$top_{m,i}$	Main topic group of the i^{th} article, $m \in$ {POL, LEG, FIN, ECO, MIL, PRO, CRI, INT, TECH, CUL, SPO, EDU, FAM, TRA, REG, HEA, MED, ACC, REL, USSR, MISC}. 21 dummy variables, only 20 used in the regression models as MISC was dropped to avoid singularity.	Flexible LDA algorithm with max=30 topics. Described in section on topic modelling.
$dip_{k,i}$	Intensity of the presence of influential domestic politicians from country k in the i^{th} article, $k \in$ {RU, KZ, BEL, UKR, PL}. Five variables.	If N is equal to the number of occurrences of the names of influential domestic politicians from country k in the i^{th} article, $dip_{k,i} = 0$ if $N = 0$, $dip_{k,i} = 1 + \log(N)$ if $N>0$.
$emo_{n,i}$	Emotion score of the i^{th} article, $n \in$ {sadness, joy, disgust, fear, anger, and surprise}. Six variables, only five used in the regression models as surprise was dropped, because the emotion scores sum to one.	Described in section on emotion detection.
$news_{p,i}$	Dummy variable to measure differences in newspaper policies for how long articles are kept on the main page of portal p in country k, $k \in$ {RU, KZ, BEL, UKR, PL}. Two or three dummy variables depending on the country, as one portal dummy is dropped to avoid singularity.	Dummy variables set to one when the i^{th} article was published by portal p. Defined separately for each country, $p = p(k)$

Table 4
. Expected signs and interpretation of the estimated parameters in Eq. (1)

Parameter	Hypothesis/ Expected sign	Interpretation
β_1	H1a/Positive	An article similar to the dominant newspaper's daily issue narrative is more likely to remain on the website's main page the next day.
β_2	H1b/Positive	If an issue discusses many topics, the likelihood of an article staying on the main page for another day is higher than when the issue focuses on fewer topics.
β_3	-/Positive	It is more likely that an article will stay on the main page for longer if the main page has space for many articles.
Δ	H3/Negative	The empirical evidence concerning online newspapers is tilted toward the news negativity effect. Negative news attracts more attention, so it is more likely to live longer.
Γ	H2/Positive or Negative	The supply of news and gatekeeping practices depend on the news topic; hence the topic dummies should significantly impact the news life cycle, with different signs. The topic effects can be country-specific, but political news is expected to have a shorter lifespan and cultural news longer.
Z	H5/Positive or Negative	When influential politicians are mentioned in the article in an autocratic country, the durability of the article is extended. In democracies, the opposite relationship should hold, as politicians often speak to the media.
$\chi_{p(k)}$	-/Positive or Negative	Newspapers have various policies for the main web page contents, so we expect the newspaper/portal dummies to be significant and have different signs.
$\theta_1 - \theta_5$	H4 & H6/ Mostly positive	A high score for certain emotions is likely to extend the article's life beyond one day. A high score of negative emotions, especially fear, extends the news lifespan in autocratic countries.

$b_{BEL}=-0.230$, $p_{BEL}<0.0001$, $b_{UKR}=-0.117$, $p_{UKR}<0.0001$). The same applies to international relations (INT, $b_{PL}=-0.104$, $p_{PL}<0.0001$, $b_{RU}=-0.441$, $p_{RU}<0.0001$, $b_{KZ}=-0.090$, $p_{KZ}=0.0008$, $b_{BEL}=-0.094$, $p_{BEL}=0.0002$; INT topic was not identified in Ukraine), which indicates that foreign issues are not very popular amongst readers or that the gatekeeping practices attach higher priority to domestic news. As postulated, articles discussing cultural events have a longer lifespan in four countries ($b_{PL}=0.290$, $p_{PL}<0.0001$, $b_{RU}=0.082$ $p_{RU}=0.0067$ for similar articles, $b_{BEL}=0.140$, $p_{BEL}<0.0001$, $b_{UKR}=0.076$, $p_{UKR}=0.0003$), and the effect is insignificant in Kazakhstan ($b_{KZ}=-0.016$, $p_{KZ}=0.516$). For all other topics, the signs of the estimates are country-specific.

Table 5 shows, fairly unexpectedly, that articles with higher sentiment than the newspaper average (*sent*) are more likely to remain on the main page the next day, meaning negative news has a shorter lifecycle. This effect is most substantial in Russia (b_{RU}=4.717 p_{RU}<0.0001), followed by Belarus and Kazakhstan (b_{BEL}=1.687, p_{BEL}<0.0001, b_{KZ}=1.646, p_{KZ}<0.0001), but is found in all five countries (b_{PL}=0.591, p_{PL}<0.0001, b_{UKR}=0.272, p_{UKR}=0.0016 for similar articles). Hence the hypothesis H3 of news negativity bias is not supported.

There are interesting patterns for emotions. When the dominant emotion is fear, the odds of the article remaining on the website rise in Russia, Kazakhstan, and Belarus (b_{RU}=1.117, p_{RU}<0.0001, b_{KZ}=1.290, p_{KZ}<0.0001, b_{BEL}=0.215, p_{BEL}=0.0273), but not in Poland and Ukraine (b_{PL}=0.288, p_{PL}=0.110, b_{UKR}=−0.110, p_{UKR}=0.450), where the media enjoy more freedom. Hypothesis H6, postulating that negative emotions extend the news lifecycle in autocracies, is confirmed for the emotion "fear". The emotion patterns are similar in Russia and Belarus, with sadness ((b_{RU}=−3.610, p_{RU}<0.0001, b_{BEL}=−0.278, p_{BEL}=0.0230) and disgust (b_{RU}=−0.273, p_{RU}=0.354, b_{BEL}=−0.328, p_{BEL}=0.0064) reducing the odds and fear raising them, though the disgust effect is not significant in Russia. Kazakhstan, where the media are also not free, exhibits a somewhat different pattern, with sadness (b_{KZ}=0.977, p_{KZ}<0.0001), fear, and disgust (b_{KZ}=1.030, p_{KZ}<0.0001) raising the odds. Ukraine is the only country where a positive emotion, joy, raises the likelihood of news being shown for more than one day (b_{UKR}=0.343, p_{UKR}<0.0001). There is weak evidence that joy reduces this likelihood in Poland (b_{PL}=−0.260, p_{PL}=0.049). Poland differs from the other four countries in that emotions do not help predict whether an article will remain on the website the next day. So Hypothesis H4 is supported in the four post-soviet countries, but not in Poland.

When an article mentions influential domestic politicians (*dip*), it is less likely to stay on the website for another day (b_{PL}=−0.019, p_{PL}=0.052, b_{RU}=−0.278, p_{RU}<0.0001, b_{BEL}=−0.038, p_{BEL}=0.002, b_{UKR}=−0.055, p_{UKR}<0.0001), except in Kazakhstan (b_{KZ}=0.035 p_{KZ}=0.006). This suggests that the Kazakhstani media experience the most interference from politicians or that their internal gatekeeping practices are heavily biased towards news that mentions politicians having a long lifecycle. Hypothesis H5 is supported in one of the three autocracies analysed.

There are large differences for the newspaper dummies (*news_**) in each country, which reflects the different policies of newspapers for online content management and the different structures of their websites.

The results are robust to the definition of article similarity. The value of the estimates and their significance (p-value) is similar when the article the next day is identical (est_100, p_100) or when there are some text modifications (est_99, p_99).

6. Discussion

Many research papers have documented the negative news bias, where negative news receives more coverage than positive news and generates more audience feedback, as supported by studies by van der Meer et al. (2019), Soroka et al. (2019), and Zhu et al. (2020). The bias towards negative news is often established through audience metrics analysis, such as views, shares, clicks, comments, and likes, and partially through gatekeeping decisions on what news to publish on the main news website. Although these metrics have become a crucial factor in determining newsworthiness, news organisations still regard themselves as custodians of journalistic values, including speed, accuracy, objectivity, and impartiality (Walters, 2022). As a result, the decision to remove news from the main website, and news longevity, depends on these values as much, if not more than, on audience metrics. Therefore, the negative news popularity bias is measured by audience metrics, and the positive news longevity bias results from the complex interplay between audience activity and journalistic values.

One possible explanation for the finding that news with a positive sentiment polarity exhibits longer longevity could be related to the psychological effects of positive emotions. The literature suggests that experiencing positive emotions, including when reading a positively inclined text, can lead to a range of beneficial outcomes, such as improved well-being, better health, and increased resilience (Fredrickson, 2001). News gatekeepers may be aware of this effect and may want to offset the negative news popularity bias with the positive news longevity bias.

In addition, positive news may be perceived as more trustworthy or credible than negative news. Research has shown that negative news can induce scepticism and increase the likelihood of people questioning the source or validity of the news. In contrast, positive news may be seen as more reliable and objective, leading readers to trust and engage with it more. These findings may support traditional gatekeeping practices in newsrooms, which prioritise newsworthiness and media reputation over catering to audience preferences.

Soroka and Krupnikov (2021) argue that despite the seemingly increasing negative tone of media coverage, the prevalence of positive news is likely to increase for three reasons: (1) valence-based asymmetries vary over time, (2) valence-based asymmetries vary across individuals, and (3) technology facilitates diverse news platforms catering to diverse preferences.

The impact of positive sentiment bias on news longevity is most pronounced in Russia, Belarus, and Kazakhstan, where the media environment is controlled by autocratic regimes. This finding is consistent with the results of Gentzkow and Shapiro (2006), who found that biases are less severe when consumers receive independent evidence on the true state of the world and that competition between independently owned news outlets can reduce bias.

As previously demonstrated, some estimated logistic regression parameters display varying values and signs across individual countries. In addition, Eq. (1) was estimated for the entire dataset, with the resulting log-odds presented in Table 1 of Appendix B in the Supplementary material. The positive impact of sentiment on news longevity is significant across all countries, with an odds ratio of 3.1. While the impact of emotions on the news life cycle varies by country, the all-country regressions illustrate which effects remain prevalent when all five countries are taken into account. To avoid singularity, the surprise emotion was omitted from the analysis, and the results should be interpreted in relation to its impact on news longevity. Overall, the presence of joy, disgust, and fear in the news text have been found to increase the longevity of news stories.

Table 5
The results of the estimated logistic regression models

	Poland (N = 155,973)				Russia (N = 130,698)				Kazakhstan (N = 222,773)				Belarus (N = 122,806)				Ukraine (N = 289,451)			
	est_100	p_100	est_99	p_99	est_100	p_100	est_99	p_99	est_100	p_100	est_99	p_99	est_100	p_100	est_99	p_99	est_100	p_100	est_99	p_99
Const	−2.572	0.000	−2.536	0.000	−2.289	0.000	−2.418	0.000	−2.572	0.000	−2.784	0.000	−0.990	0.000	−1.207	0.000	−3.058	0.000	−2.785	0.000
cosine_sim	−0.495	0.000	−0.427	0.002	1.503	0.000	1.851	0.000	0.136	0.026	0.289	0.000	0.322	0.012	1.474	0.000	0.475	0.000	1.049	0.000
cosine_std	15.097	0.000	14.673	0.000	2.485	0.000	2.697	0.000	5.039	0.000	5.300	0.000	−0.629	0.087	−1.452	0.002	−7.403	0.000	−8.761	0.000
issue_size	0.010	0.000	0.010	0.000	0.023	0.000	0.022	0.000	0.000	0.870	0.001	0.001	−0.003	0.000	−0.003	0.000	−0.001	0.000	−0.002	0.000
POL	−0.114	0.000	−0.120	0.000	−0.384	0.000	−0.402	0.000	−0.397	0.000	−0.392	0.000	−0.230	0.000	−0.301	0.000	−0.117	0.000	−0.283	0.000
LEG	−0.286	0.000	−0.267	0.000	−0.460	0.000	−0.511	0.000	−0.531	0.000	−0.487	0.000	0.178	0.001	0.099	0.051	−0.170	0.000	−0.271	0.000
FIN	0.261	0.000	0.247	0.000	−0.212	0.000	−0.223	0.000	−0.090	0.001	−0.093	0.001	0.031	0.264	0.282	0.000	0.277	0.000	0.500	0.000
ECO	0.350	0.000	0.347	0.000	−0.082	0.005	−0.129	0.000	−0.240	0.000	−0.246	0.000	−0.085	0.000	−0.240	0.000	0.379	0.000	0.472	0.000
MIL	−0.017	0.603	−0.029	0.378	−0.383	0.000	−0.373	0.000	−0.043	0.324	−0.014	0.750	0.299	0.000	0.260	0.000	−0.211	0.000	−0.364	0.000
PRO	−0.041	0.351	−0.064	0.149	−0.670	0.000	−0.679	0.000					0.145	0.000	0.165	0.003	0.225	0.000	0.158	0.000
CRI	0.026	0.451	0.012	0.711	−0.413	0.000	−0.442	0.000	−0.438	0.000	−0.431	0.000	−0.205	0.000	−0.258	0.000	0.054	0.007	−0.055	0.003
INT	−0.104	0.000	−0.104	0.000	−0.441	0.000	−0.473	0.000	−0.090	0.001	−0.105	0.000	−0.094	0.000	−0.029	0.319				
TECH	0.968	0.000	0.975	0.000	0.331	0.000	0.316	0.000	0.123	0.000	0.124	0.000	−0.062	0.195	0.109	0.092	−0.134	0.009	−0.173	0.000
CUL	0.290	0.000	0.312	0.000	0.043	0.154	0.082	0.007	−0.016	0.516	−0.009	0.718	0.140	0.000	0.224	0.000	0.076	0.000	0.000	0.990
SPO	−0.506	0.000	−0.537	0.000	−0.254	0.000	−0.285	0.000	0.011	0.666	0.001	0.955	−0.067	0.032	−0.108	0.001	0.540	0.000	0.820	0.000
EDU	−0.302	0.000	−0.291	0.000	0.137	0.249	0.119	0.303	−0.073	0.054	−0.077	0.040	−0.309	0.000	−0.286	0.000	0.245	0.000	0.365	0.000
FAM	−0.570	0.000	−0.568	0.000	0.326	0.000	0.353	0.000	0.331	0.000	0.327	0.000	−0.143	0.000	−0.080	0.038	0.126	0.000	0.114	0.000
TRA	−0.425	0.000	−0.422	0.000	0.003	0.938	0.009	0.824					0.025	0.606	0.038	0.481	0.296	0.000	0.527	0.000
REG	0.729	0.000	0.738	0.000	0.590	0.000	0.562	0.000	−0.187	0.000	−0.167	0.000	0.025	0.515	−0.148	0.000	0.237	0.000	0.454	0.000
HEA	−0.014	0.617	0.122	0.000	−0.667	0.000	−0.667	0.000	−0.227	0.000	−0.224	0.000	−0.031	0.347	−0.133	0.000	0.249	0.000	0.300	0.000
MED	0.047	0.257	0.027	0.513	0.179	0.009	0.191	0.005	0.462	0.000	0.450	0.000	0.282	0.000	0.166	0.000	−0.151	0.000	−0.203	0.000
ACC	0.212	0.000	0.215	0.000	−0.093	0.190	−0.152	0.028	−0.312	0.000	−0.272	0.000	−0.340	0.000	−0.289	0.000	0.091	0.000	0.008	0.706
REL	0.068	0.121	0.035	0.417	0.095	0.362	0.161	0.130	−0.304	0.003	−0.316	0.003					−0.241	0.000	−0.233	0.000
USSR	−0.221	0.000	−0.233	0.000	0.116	0.197	0.179	0.051	0.933	0.000	0.839	0.000								
Sent	0.591	0.000	0.424	0.000	4.717	0.000	4.737	0.000	1.646	0.000	1.578	0.000	1.687	0.000	1.428	0.000	0.062	0.490	0.272	0.002
Dip	−0.019	0.052	−0.005	0.619	−0.278	0.000	−0.257	0.000	0.035	0.006	0.052	0.000	−0.038	0.002	−0.057	0.000	−0.055	0.000	−0.056	0.000
lnews_1	−0.961	0.000	−0.998	0.000	−1.169	0.000	−1.200	0.000	0.558	0.000	0.575	0.000	1.048	0.000	1.801	0.000	2.605	0.000	3.107	0.000
lnews_2	−0.572	0.000	−0.610	0.000	−4.213	0.000	−4.166	0.000	−0.401	0.000	−0.142	0.000	1.103	0.000	2.175	0.000	1.906	0.000	2.061	0.000
lnews_3					0.046	0.047	0.029	0.214	1.428	0.000	1.540	0.000								
Sadness	−0.282	0.217	−0.025	0.912	−3.610	0.000	−3.592	0.000	0.977	0.000	0.906	0.000	−0.278	0.023	−0.337	0.002	0.026	0.869	0.154	0.316
Joy	−0.260	0.049	−0.035	0.784	0.151	0.574	0.162	0.542	0.051	0.636	0.050	0.634	0.159	0.008	0.032	0.562	0.343	0.000	0.187	0.036
Disgust	0.308	0.372	0.577	0.089	−0.273	0.354	−0.374	0.199	1.030	0.000	1.070	0.000	−0.328	0.006	−0.424	0.000	1.192	0.000	0.458	0.004
Fear	0.288	0.110	0.118	0.509	1.117	0.000	0.969	0.000	1.290	0.000	1.195	0.000	0.215	0.027	0.197	0.026	−0.110	0.450	−0.090	0.508
Anger	−0.094	0.663	−0.265	0.214	0.175	0.353	0.152	0.413	0.016	0.876	−0.028	0.783	−0.292	0.022	0.119	0.250	−0.135	0.214	0.580	0.000
Pseudo R^2	0.111		0.1076		0.2385		0.2455		0.1031		0.0989		0.0897		0.2306		0.0777		0.1491	

Notes. Two models are estimated for each country. In the first, labelled 100, the next-day version of the article is identical to the original, so the cosine similarity is 1. The second, labelled 99, allows slight modifications and considers the article the same when the cosine similarity is 0.99 or higher. "est_" is the estimate, and "p_" is the p-value from robust standard errors. The number of observations in each country is lower than the number of scraped articles in Table 1 because newspaper issues that were not followed by a next-day issue were removed from the data, reducing the total number of observations from 1047,304 to 921,701. The correlation matrix for independent variables has been inspected for each country regression and the was no evidence of multicollinearity.

6.1. Validation of the results

An important contribution of this paper concerns the role of emotions in the news lifecycle. One question arises immediately though. The results of the zero-shot emotion extraction were not human-validated, so how sure can we be that the machine learning classification is appropriate?

In recent years, significant progress has been made in deep learning techniques, particularly the emergence of transformer-based neural networks and the development of Large Language Models (LLMs) with billions of parameters, trained on billions of words and sentences. These advances have revolutionised the field of machine learning. The "ChatGPT shock" has brought this paradigm shift to the attention of researchers who were not previously familiar with these models. Prior to this, machine learning models were thought to be error-prone, with results requiring human validation. However, LLMs can now identify intricate semantic relationships amongst words that surpass human comprehension. Nonetheless, various new techniques for result validation have been proposed in this paper.

Human validation of textual emotion scores is problematic for two reasons. First, it is improbable that human coders would be able to estimate the six basic emotion scores for texts, especially for longer ones, and that their classification would be consistent across coders. Second, it would be very costly for a large corpus of one million articles, as coding only a tiny percentage of texts requires thousands of articles to be read. So I propose another method that I call 'inter-feature consistency'. As the name suggests, this approach uses several types of features extracted from the same text corpus. Suppose we are certain about the validity of one feature type because it was previously human-validated. We can calculate the relationship between this feature type and the other type that was not validated. A human researcher then compares this relationship to common sense expectations or established results from the literature.

I calculate several consistency measures for basic emotions and other features extracted from news texts. In the first approach, presented in Fig. 2, I measure the difference in emotion scores for articles with different dominant topics. I choose pairs of topics that intuitively should have different scores for the positive emotion "joy" and the negative emotions "fear" and "anger". These topic pairs are (culture - crime) and (sports - accidents), with the first element in each pair expected to have a higher average score for the positive emotion and a lower score for the negative ones. Fig. 2 shows the differences between the two topics. As expected, articles in all countries discussing cultural events exhibit a higher joy score and lower fear and anger emotion scores than articles discussing crime. For the sports and accidents pair, the relationship is analogical for the joy and fear emotions, but the anger score is lower for articles about sports only in Poland and marginally in Russia. However, while we have clear interpretations and expectations for the fear emotion, for anger, these are less evident. Texts discussing sports events may reveal anger if national or local teams or athletes lose when they were expected to win. Overall, the zero-shot emotion detection gives sensible results for the selected pairs of topics.

A second example of the inter-feature consistency validation is shown in Fig. 3. I present the signs of Pearson correlation scores between the relative sentiment feature and the three selected emotions. As expected, articles with a high joy emotion tend to have a more positive inclination, and articles emanating fear tend to have a predominantly negative inclination. The expected negative correlation is confirmed for anger for four of the countries but not for Kazakhstan. So out of fifteen tests, fourteen give the expected correlation signs.

The inter-feature consistency tests depicted in Figs. 2 and 3 demonstrate that zero-shot classification of emotions is aligned with theoretical expectations and previous research pertaining to the associations between different textual data attributes. Of course, the inter-feature consistency tests conducted here cannot replace a proper validation, but they indicate that the zero-shot emotion detection model produces sensible results.

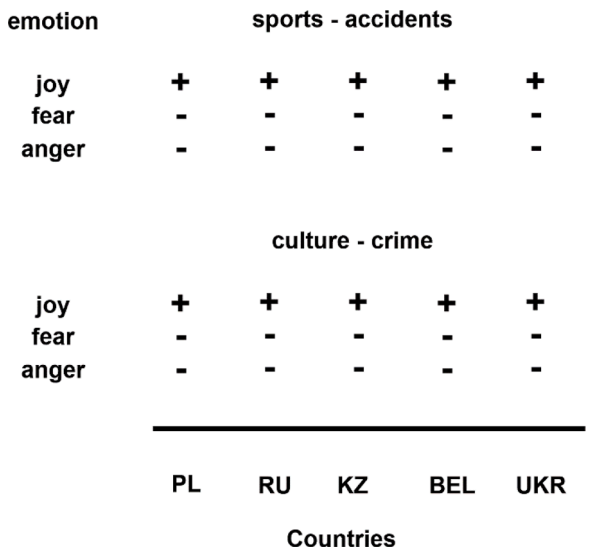

Fig 2. The difference between average emotion scores for articles with the dominant topic culture minus sports and crime minus accidents.

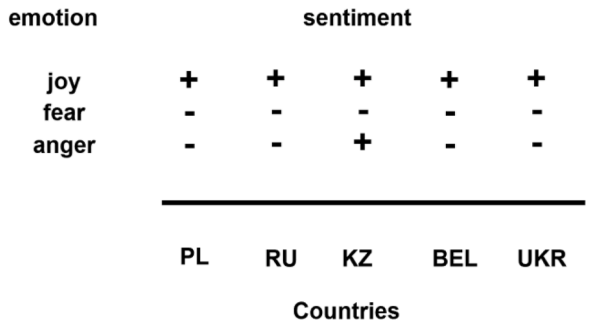

Fig 3. The sign of the Pearson correlation between relative sentiment and emotion score

7. Conclusions and future work

It is popularly held that online media coverage favours negative news. Many research papers have found that negative news receives more coverage than positive news and is more popular. I showed that bad news, meaning articles with a negative relative sentiment, is less likely to remain on the main page the next day. This effect is strong and present in all the five countries analysed. The shorter lifespan of negative news seems partly to compensate for its coverage and popularity bias.

The second contribution of this paper concerns the role of emotions in the news lifecycle. Gupta and Yang (2019) showed that headlines with high emotional intensity were more popular, but, as shown here, this does not hold for the news lifecycle. The estimated statistically significant parameters mostly have different signs and are country-specific, except for the emotion "fear". A high fear score extends the news lifecycle, but only in autocratic regimes, and the impact is substantial in magnitude.

The third contribution of this paper is methodological. When comparative media analysis involves several newspapers, countries, and languages, the relative sentiment measure presented here should be applied, rather than absolute sentiment, which can be affected by the idiosyncratic newspaper, country, or language effects. For the first time in the literature, I applied transformer-based contextual embeddings to assess the semantic similarity between the news texts and their relationship to the main narrative of the daily issue of the newspaper. The features based on semantic similarity are statistically significant in each country's regression, but the estimates often have different signs across countries. I showed that it is possible to extract six basic emotions from news texts, and I proposed an inter-feature consistency method to validate the results of the zero-shot unsupervised learning. While this method cannot replace the proper validation effort, it does give an indication or a partial answer as to whether the unsupervised model generates sensible results.

Finally, the paper integrates many different natural language processing approaches into one modelling framework that is suitable for analysing multilingual text corpora. I applied state-of-the-art NLP models that exceed human skills in many language tasks, combined them with the traditional text mining models used in social sciences, and applied them to a big data news corpus in the Polish and Russian languages.

This study presents the first documentation of the shorter lifecycle of negative news in the literature. Further research is necessary to confirm and expand upon these findings, such as by applying the proposed model to other countries, languages, and newspapers. Additionally, a more detailed analysis is needed to understand the mechanisms behind the observed effect of high fear emotion scores on the extended news lifecycle in autocratic countries.

Other interesting directions for further research include the following: incorporating the time dimension to determine whether the news lifecycle depends on the news that has been recently published; examining more complex relationships between articles in the same issue through the application of textual entailment methods (Yavary, Sajedi, & Abadeh, 2019) and graph analysis; and using different, more complex emotion classifications.

Funding details

The author received no financial support for the research, authorship and/or publication of this article.

Author statement

The sole author of the paper has performed all the work.

Declaration of Competing Interest

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

I have shared my data in Zenodo open repository: https://doi.org/10.5281/zenodo.6591100

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2023.103398.

References

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review, 54*, 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

Agrawal, S., Roy, D., & Mitra, M. (2021). Tag embedding based personalized point of interest recommendation system. *Information Processing & Management*, (6), 58. https://doi.org/10.1016/j.ipm.2021.102690

Alsudasis, A., Alotaibi, W., & Alomary, F. (2022). Similarities between Arabic dialects: Investigating geographical proximity. *Information Processing & Management*, (1), 59. https://doi.org/10.1016/j.ipm.2021.102770

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Boczkowski, P. J. (2010). *News at work: Imitation in an age of information abundance*. Chicago: University of Chicago Press.

Boczkowski, P. J., & Mitchelstein, E. (2015). *The news gap: When the information preferences of the media and the public diverge*. Cambridge, MA: MIT Press.

Bodó, B. (2019). Selling news to audiences – A qualitative inquiry into the emerging logics of algorithmic news personalization in European Quality News Media. *Digital Journalism, 7*(8), 1054–1075. https://doi.org/10.1080/21670811.2019.1624185

Bright, J., & Nicholls, T. (2014). The life and death of political news: Measuring the impact of the audience agenda using online data. *Social Science Computer Review, 32*(2), 170–181. https://doi.org/10.1177/0894439313506845

Buhl, F., Günther, E., & Quandt, T. (2018). Observing the dynamics of the online news ecosystem. *Journalism Studies, 19*(1), 79–104. https://doi.org/10.1080/1461670X.2016.1168711

Charemza, W., Makarova, S., & Rybinski, K. (2022). Economic uncertainty and natural language processing; The case of Russia. *Economic Analysis and Policy, 73*, 546–562. https://doi.org/10.1016/j.eap.2021.11.011

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., & Guzmán, F. et al. (2019). Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116. 10.48550/arXiv.1911.02116.

de Vries, E., Vliegenthart, R., & Walgrave, S. (2022). Telling a different story: A longitudinal investigation of news diversity in four countries. *Journalism Studies*. https://doi.org/10.1080/1461670X.2022.2111323

Dziob, A., Piasecki, M., & Rudnicka, E. (2019). plWordNet 4.1–a linguistically motivated, corpus-based bilingual resource. In *Proceedings of the 10th Global WordNet Conference: July 23-27, 2019* (pp. 353–362). Wroclaw Oficyna Wydawnicza Politechniki Wroclawskiej.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT Sentence Embedding. arXiv:2007.01852. 10.48550/arXiv.2007.01852.

Ferrara, E., & Yang, Z. (2015). Measuring emotional contagion in social media. *PloS one, 10*(11), Article e0142390. https://doi.org/10.1371/journal.pone.0142390

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions. *The American Psychologist, 56*(3), 218–226. https://doi.org/10.1037//0003-066x.56.3.218

Gentzkow, M., & Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy, 114*(2), 280–316. https://doi.org/10.1086/499414

Gupta, R. K., & Yang, Y. (2019). Predicting and understanding news social popularity with emotional salience features. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)* (pp. 139–147). Association for Computing Machinery. https://doi.org/10.1145/3343031.3351048.

Hendrickx, J. (2020). Trying to survive while eroding news diversity: legacy news media's catch-22. *Journalism Studies, 21*(5), 598–614. https://doi.org/10.1080/1461670X.2019.1694430

Imedashvili, N. (2022). Captured emotions" - Russian propaganda. *Georgian Foundation for Strategic and International Studies*. November 7, 2022 https://gfsis.org.ge/blog/view/1512.

Josua, M. (2021). The legitimation of repression in autocracies. *Oxford Research Encyclopedia of Politics*. https://doi.org/10.1093/acrefore/9780190228637.013.1988

Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science, 44*(4), 491–511. https://doi.org/10.1177/0165551517703514

Kim, H., Ahn, S.-. J., & Jung, W.-. S. (2019). Horizon scanning in policy research database with a probabilistic topic model. *Technological Forecasting and Social Change, 146*, 588–594. https://doi.org/10.1016/j.techfore.2018.02.007

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788–8790. https://doi.org/10.1073/pnas.132004011

Lee, A. M., & Chyi, H. I. (2014). When newsworthy is not noteworthy. Examining the value of news from the audience's perspective. *Journalism Studies, 15*(6), 807–820. https://doi.org/10.1080/1461670X.2013.841369

Lee, A. M., Lewis, S. C., & Powers, M. (2014). Audience clicks and news placement: A study of time-lagged influence in online journalism. *Communication Research, 41*(4), 505–530. https://doi.org/10.1177/0093650212467031

LibertiesEU. (2021). Modern political propaganda: definition, examples & how to spot it. *LibertiesEU*. November 24, 2021 https://www.liberties.eu/en/stories/political-propaganda/43850.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., & Chen, D. et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. 10.48550/arXiv.1907.11692.

Loebbecke, C., Oberschulte, F., & Boboschko, I. (2021). Mass Media Deploying Digital Personalization: An Empirical Investigation. *International Journal on Media Management, 23*(3–4), 176–203. https://doi.org/10.1080/14241277.2022.2038605

Loukachevitch, N., & Levchik, A. (2016). Creating a general Russian sentiment lexicon'. In *Proceedings of Language Resources and Evaluation Conference LREC-2016* (pp. 1171–1176). https://www.aclweb.org/anthology/L16-1186 Accessed 15 March 2021.

Maerz, S. F., & Schneider, C. Q. (2020). Comparing public communication in democracies and autocracies: Automated text analyses of speeches by heads of government. *Quality & Quantity, 54*, 517–545. https://doi.org/10.1007/s11135-019-00885-7

Marko, D. (2013). Fear control in media discourse. *Southeastern Europe, 37*(2), 200–219. https://doi.org/10.1163/18763332-03702005

Moniz, N., & Torgo, L. (2019). A review on web content popularity prediction: Issues and open challenges. *Online Social Networks and Media, 12*, 1–20. https://doi.org/10.1016/j.osnem.2019.05.002

Muddiman, A., Pond-Cobb, J., & Matson, J. E. (2020). Negativity bias or backlash: Interaction with civil and uncivil online political news content. *Communication Research, 47*(6), 815–837. https://journals.sagepub.com/doi/10.1177/0093650216685625.

Murthy, A., & Kumar, K. M. A. (2021). A review of different approaches for detecting emotion from text. In *IOP Conference Series: Materials Science and Engineering*. https://doi.org/10.1088/1757-899X/1110/1/012009, 1110 012009.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Networking Analysis and Mining 11, 81*. https://doi.org/10.1007/s13278-021-00776-6

Nave, N. N., Shifman, L., & Tenenboim-Weinblatt, K. (2018). Talking it personally: Features of successful political posts on facebook. *Social Media + Society, 4*, 1–12. https://doi.org/10.1177/2056305118784771

Nicholls, T., & Bright, J. (2019). Understanding news story chains using information retrieval and network clustering techniques. *Communication Methods and Measures, 1*, 43–59. https://doi.org/10.1080/19312458.2018.1536972. *13*.

Ribeiro, F. N., Araujo, M., Goncalves, P., Goncalves, M. A., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science, 5*(23). https://doi.org/10.1140/epjds/s13688-016-0085-1

Rousidis, D., Koukaras, P., & Tjortjis, C. (2020). Social media prediction: A literature review. *Multimedia Tools and applications, 79*, 6279–6311. https://doi.org/10.1007/s11042-019-08291-9

Rybinski, K. (2018). Political sentiment analysis of press freedom. *Studia Medioznawcze, 3*(74), 31–48.

Saridou, T., Spyridou, L.-P., & Veglis, A. (2017). Churnalism on the rise? *Digital Journalism, 5*(8), 1006–1024. https://doi.org/10.1080/21670811.2017.1342209

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences of the United States of America, 116*(38), 18888–18892. https://doi.org/10.1073/pnas.1908369116

Soroka, S., & Krupnikov, Y. (2021). *The increasing viability of good news (Elements in politics and communication)*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108982375

Tenenboim, O., & Cohen, A. A. (2015). What prompts users to click and comment: A longitudinal study of online news. *Journalism, 16*(2), 198–217. https://doi.org/10.1177/1464884913513996

Timmerman, Y., & Bronselaer, A. (2022). Automated monitoring of online news accuracy with change classification models. *Information Processing & Management*, (6), 59. https://doi.org/10.1016/j.ipm.2022.103105

van der Meer, T. G. L. A., Kroon, A. C., Verhoeven, P., & Jonkman, J. (2019). Mediatization and the disproportionate attention to negative news. *Journalism Studies, 20*(6), 783–803. https://doi.org/10.1080/1461670X.2018.1423632

Wallace, J. (2018). Modelling contemporary gatekeeping. *Digital Journalism, 6*(3), 274–293. https://doi.org/10.1080/21670811.2017.1343648

Walters, P. (2022). Reclaiming control: How journalists embrace social media logics while defending journalistic values. *Digital Journalism, 10*(9), 1482–1501. https://doi.org/10.1080/21670811.2021.1942113

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., & Hill, F. et al. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv*: 1905.00537. 10.48550/arXiv.1905.00537.

Welbers, K., van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2018). A gatekeeper among gatekeepers. *Journalism Studies, 19*(3), 315–333. https://doi.org/10.1080/1461670X.2016.1190663

Welbers, K., van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., & Schaper, J. (2016). News selection criteria in the digital age: Professional norms versus online audience metrics. *Journalism, 17*(8), 1037–1053. https://doi.org/10.1177/1464884915595474

Yavary, A., Sajedi, H., & Abadeh, M. S. (2019). Information verification improvement by textual entailment methods. *SN Applied Sciences, 1*(1048). https://doi.org/10.1007/s42452-019-1073-4

Zhu, X., Kim, Y., & Park, H. (2020). Do messages spread widely also diffuse fast? Examining the effects of message characteristics on information diffusion. *Computers in Human Behavior, 103*, 37–47. https://doi.org/10.1016/j.chb.2019.09.006