Prioritizing Original News on Facebook

Xiuyan Ni, Shujian Bu, Lucas Adams, Igor L. Markov
{xni,shujian,lucasadams,imarkov}@fb.com
Facebook Inc.

ABSTRACT

This work outlines how we prioritize *original* news, a critical indicator of news quality. By examining the landscape and lifecycle of news posts on our social media platform, we identify challenges of building and deploying an originality score. We pursue an approach based on normalized PageRank values and three-step clustering, and refresh the score on an hourly basis to capture the dynamics of online news. We describe a near real-time system architecture, evaluate our methodology, and deploy it to production. Our empirical results validate individual components and show that prioritizing original news increases user engagement with news and improves proprietary cumulative metrics.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; *Data mining*;
• **Computing methodologies** → Machine learning.

KEYWORDS

News; News Feed; Personalization; Originality; PageRank; Clustering

ACM Reference Format:
Xiuyan Ni, Shujian Bu, Lucas Adams, Igor L. Markov. 2021. Prioritizing Original News on Facebook. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3459637.3481943

1 INTRODUCTION

Large amounts of news are published online every day, and many people now primarily consume news online [21]. News quality affects how people consume news and which platforms they prefer. Therefore, faithfully capturing news quality by a score promises significant benefits to both users and platforms. Among various aspects of news quality, we focus on *originality*, which can be contrasted with duplicates, slightly edited text, and coverage that references original news. Producing original news is laborious and requires expertise, but such efforts initiate the typical news cycle and drive the entire news industry. Original news inform people around the world, from breaking-news articles, eye-witness reports and critical updates at the time of crisis, to in-depth investigative reports that uncover new facts and data. Hence, prioritizing original news on Facebook is in everyone's long-term interest.

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11.
https://doi.org/10.1145/3459637.3481943

In this work, we first explore the landscape of online news, using the Facebook platform as an example. To enable a quantitative approach, we tabulate the spectrum of news originality from *completely unoriginal* to *highly original* news. Our static analysis suggests that highly original news are rare, despite a large inventory which needs to be indexed and processed to accurately identify the original ones. We also explore the dynamics of the news lifecycle on Facebook and find that news posts typically attain the greatest exposure in the first couple of hours, followed by a long tail. Therefore, if an originality score is used to improve News Feed ranking, it needs to be computed promptly.

Given two challenges — search quality at scale and fast response — we build a near real-time system and construct a synthesized signal for news originality. News articles that cover the same news event are clustered together based on specialized BERT embeddings [7], which are finetuned on pairwise-labeled data (same subject or different subjects). After evaluating several clustering algorithms against human-labeled pairwise data, we settle on a two-stage clustering algorithm that is both effective and highly scalable to large datasets. To adequately capture news dynamics, our system performs incremental updates on an hourly basis.

We concluded that content alone is insufficient to judge news originality, but behavioral signals such as citations of prior posts can also be used. Integrity considerations are particularly important, given the high incentives to game online news distribution. To de-bias our algorithms, we filter out news articles produced within patterns of nefarious activity. We first evaluate the performance of our originality signal offline against ratings by professional journalists. Online evaluation is based on an A/B test where we additionally monitor the impact on news article ranking. The signal is incorporated in the News Feed ranking system.

Our contributions include:

- We examine the news originality landscape and the dynamics of the news lifecycle, then propose a quantitative approach to news ecosystem quality. We categorize the news originality level by the effort spent in generating news content.
- We propose a methodology and architect a near real-time system that processes individual news articles at a large scale. Using the PageRank algorithm and three-step clustering, it calculates a synthetic score to estimate news originality. PageRank normalization within clusters is particularly novel. The method can be applied to other news serving systems.
- To facilitate live-data analysis of perceived news quality and of news quality scores, we develop quantitative and qualitative methods. These methods can zoom in on individual news articles and their distribution, and also measure entire news ecosystems. Such analyses can help both news publishers and consumers, which now depend on online news.

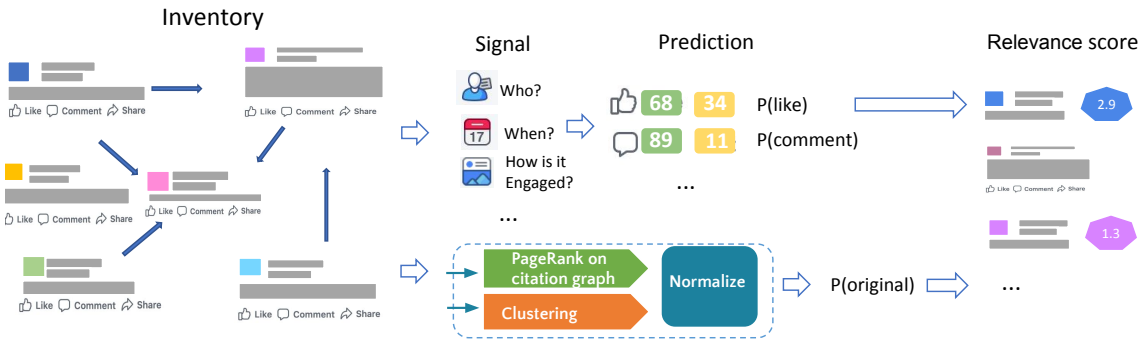

Figure 1: News Feed ranking at Facebook

2 CONTEXT AND RELATED WORK

2.1 News Feed Ranking

Given that we deploy our news originality signal in a social network, we review the basics of News Feed ranking. Related ranking formulations have been studied both in academia and industry [6, 11, 26], with many publications in the information retrieval community [5, 10, 13, 21, 24, 27, 28]. In 2018, Nuzzle announced a ranking system for news sources called NuzzleRank (https://nuzzel.com/rank) that integrates various signals, including publisher authority information, into a single score to rank news sources.

Ye and Skiena [26] built an automated ranking system called MediaRank to rank news sources. They applied the PageRank algorithms on news reporting citation to rank news sources and proved that PageRank values are positively related to reporting quality measured by peer reputation and so on. Zhang et al. [27] introduced a set of signals for indicating the credibility of news collected from expert annotators. They grouped their indicators into two categories. The first group contains content indicators determined by the articles themselves — mentions of organizations, studies, etc. Context indicators in the other group require analysis of external sources, such as author reputation and/or recognition by peers in terms of the PageRank algorithm, as in [3, 4].

Facebook's News Feed ranks not only news content, but also events from users' social graph. Ranking objectives optimize metrics of user engagement and long-term user satisfaction, with additional considerations for communities (friends and family, etc) and News Feed integrity (e.g., to discourage clickbait and prevent unlawful activities). A user sees in their News Feed fresh updates from their friends, groups they joined, and pages they followed. News Feed ranking can be roughly divided into four stages: inventory, signals, prediction, and relevance scores [16, 17]. Once a piece of content is posted, numerous signals are extracted — text, publication time, engagement counts, etc. The signals collected are used in prediction models for the probability of each action that a user may take for each piece of content in the inventory (should they see it), e.g., the $P(Comment)$ model predicts the likelihood that a user will comment on the update, while the $P(Like)$ model predicts the likelihood that a user will like the content. At the last stage, these predictions are aggregated into a ranking score for each piece of content (Figure 1).

As seen above, News Feed ranking routinely incorporates many signals, and if some content is not promoted by a particular signal, it can be promoted by a different signal.

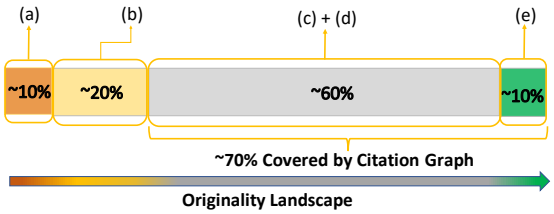

Figure 2: News originality by bucket: (a) completely unoriginal; (b) highly unoriginal; (c) somewhat unoriginal; (d) potentially original but lacking peer recognition; (e) recognized as original by peers. For each bucket, we show estimated total views received by all news articles.

2.2 The Page Rank algorithm

The PageRank algorithm was originally developed at Google to rank Web pages and sites to improve search results [2–4, 18, 26]. Mathematically, it is a random-walk based algorithm to rank vertices in a graph. A Web page with many incoming links from large-weight web pages, has a greater weight. Page weights are propagated from each Web page to pages it links to. In the news domain, the work by Del Corso et al. [6] introduced a related graph-based ranking algorithm where each node represents a news source, focusing on authoritative news sources and interesting news events.

2.3 Representing news articles

When estimating article originality, it is important to check how similar two articles are. Such checks are commonly implemented with cosine similarity on vector embeddings. Prior work uses BERT (Bidirectional Encoder Representations from Transformers) embeddings [7], which achieved state-of-art results in many natural language processing tasks across different applications [14, 19]. Original BERT models were DNNs pre-trained on the BooksCorpus [29] and the English Wikipedia, but can be specialized via transfer learning by adding one additional layer to the neural net. For example, a multilingual BERT implementation[1] was trained on top 100 languages with Wikipedia data[2] to represent each news article based on its title. Using only titles conveniently neglects changes in article bodies, but emphasizes adequate handling of synonyms, rare words, and equivalent phrases — BERT excels at these. BERT is capable of handling previously unseen words by breaking them

[1]https://github.com/google-research/bert/blob/master/multilingual.md
[2]https://meta.wikimedia.org/wiki/List_of_Wikipedias

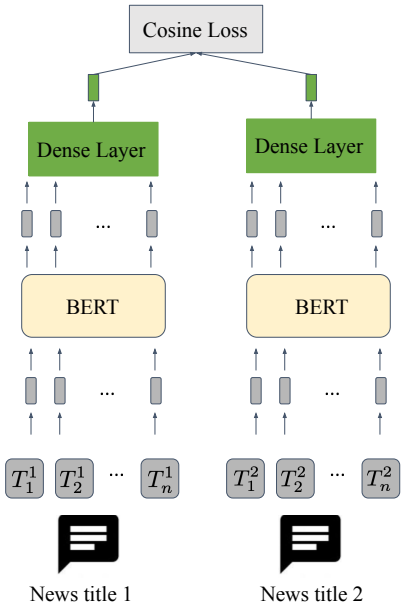

Figure 3: Estimaiting sentence similarity using pre-trained BERT networks [20]. The shared MLP layer is trained.

down into subword fragments. It can also be updated on a regular basis to handle emerging keywords such as "COVID". BERT can be specialized to a given use case by adding one MLP layer and training it respective labeled data. In this work, we perform semantic similarity estimation using a Siamese-twins network with two copies of BERT [20], shown in Figure 3.

3 PROBLEM ANALYSIS

Here we examine the news originality landscape and motivate our algorithmic contributions. Then we investigate the lifecycle of news stories on social media platforms. Understanding the news lifecycle is critical to deploying the originality signal.

3.1 The landscape of news originality

To facilitate a quantitative approach to news originality, we introduce the following content buckets:

a) *completely unoriginal*, scraped or spun content with no editorial effort
b) *highly unoriginal*, with very low editorial effort
c) *somewhat unoriginal*, may be editorially produced but heavily cite other content without original reporting or analysis
d) *potentially original but lacking peer recognition*
e) *recognized as original by peers*: breaking news, eyewitness reports, exclusive scoop, investigative reporting, etc

Scraped content is copied from other sources without editorial efforts. *Spun content* is taken from a post or a Web page, and posted with only minor modifications by humans or machines (see examples in Table 1). Common methods include paraphrasing, replacing words, and reordering paragraphs. By automating the *spinning* of existing content one can quickly produce a large amount of content without scraping. Scraped and spun content can outnumber original content and undermine its value, which warrants removal or limited distribution compared to original content.

Highly unoriginal articles are produced by low-effort text changes. We find most of the news articles actually fall into the third bucket - *somewhat unoriginal*. These articles may provide useful information, but do not require much effort to produce.

Potentially original but lacking peer recognition — this bucket includes content that does not fit in earlier buckets and so may be original, but for various reasons does not receive peer recognition throughout the news cycle. Opinion pieces that receive little support often fall into this category. Thus, citation signals alone cannot distinguish between this bucket and unoriginal articles.

The *highly original* news are produced with significant effort to fact-check information and produce clear narratives, high-quality writing and visuals. Thoughtful and original news content is usually cited heavily by industry peers and contributes to the reputation of individual content creators. Due to the effort and expertise required, the original news content are scarce. Prioritizing the distribution of original content can help it reach greater audiences and benefits both the readers and the news industry in the long run.

In general, it is difficult to judge each article for originality in isolation because this would require careful analysis of contents with the understanding of current events. Particularly challenging would be to distinguish rumors and fake news from reasonable reporting. Therefore, we draw additional insights from the news citation graph and the dynamics of online news. The special cases of scraped and spun content are handled by dedicated systems that are based on text hashing and fingerprinting, as well as text similarity metrics. In practice, such content does not appear in users' News Feed inventory and is therefore not treated in our work.

3.2 The dynamics of online news

News content published on the Internet can be easily indexed and archived, but it is often assumed that social media platforms favor fresh news. That's why news reporters strive to break a new story. To re-examine this conventional wisdom, we explore a large volume of news articles shared on Facebook and track the dynamics of user engagement metrics. We also visualize the lifecycle of typical online news stories and check the impact of adding valuable information days after the original publication. As it turns out, the same pattern

Table 1: Examples of spun content. Publisher 1 posted original articles, while Publisher 2 replaced isolated words, phrases, and sentences in articles from Publisher 1.

Publisher 1 - Original	Publisher 2 - Spun
Israel grants Rashida Tlaib West **Bank visit** on humanitarian grounds	Israel grants Rashida Tlaib West **Financial Institution go to** humanitarian grounds
Israel's **interior** minister on Friday said	Israel's **inside** minister on Friday said
Pod **Foods gets** VC backing to reinvent grocery distribution	Pod **Meals will get** VC backing to reinvent grocery distribution

persists across different news categories — world and local news, politics and entertainment news.

Figure 4 illustrates how quickly users lose interest in a particular story. On September 27, 2019 Disney and Sony reached a deal for Spiderman movies, announcing that Spiderman would stay in the Marvel Universe. One publisher reported the story first. Around 772 websites covered the news on the exact same day. On the second day, the engagement metrics of this story dropped significantly and eventually vanished on September 29, in just 3 days.

Figure 5 shows that adding information at a later time does not help gain traffic. On November 11, 2019 the Ebola vaccine by Johnson & Johnson was approved. Our inventory showed that 17 websites published 34 related articles on that day, and user engagement metrics hit a peak. The news was first reported by a publisher who focuses on life science and medicine, which gained most traffic. Two days later, on November 13, the World Health Organization officially approved the vaccine. Many mainstream publishers covered this news, and we observed an inventory increase. However, this did not stimulate another engagement peak: traffic was mostly flat and almost vanished after seven days.

Our data analysis suggests that ranking interventions can only be effective early in the lifecycle of a news story. This poses architecture and implementation challenges for both signal computation and ranking deployment. Therefore, we only focus on news articles published within the last seven days.

4 ESTIMATING NEWS ORIGINALITY

In this section, we first develop necessary infrastructure — the citation graph — and then introduce our technique for estimating news originality, based on the insights from Section 3.2.

4.1 The news citation graph

Credible news sources are often explicitly cited in follow-up news articles by various news providers, and such citations are important indicators of news source quality. Therefore, we introduce the news citation graph with edges between news articles. Figure 6 shows examples of news article citations. The top article in Figure 6 is excerpted from a news article reported by Publisher 1. This article cites multiple sources, and Figure 6 shows one of them: a news article from a another publisher, which cites another article published earlier by the same publisher. If a publisher breaks an important story, many authors tend to explicitly cite such original news. Qualitatively speaking, when a news article is disproportionately cited (compared to similar articles) by its peers, this indicates higher journalistic credibility.[3] A similar observation is used in Web ranking, where the link graph is traditionally defined at the domain level and is much coarser than our news citation graph [26]. The PageRank algorithm precomputes domain scores and uses them to rank different pages matching a query [18]. In our context, not having a query makes it difficult to compare news articles by score — a highly cited article on a niche topic would lose out to a mediocre article on a popular topic. Document-level citation graphs can be built for academic papers which itemize their references and use reference numbers in citations, but the same problem remains — numerous citations often reflect the size of the research community rather than paper quality [23]. Moreover, academic papers remain relevant for many years and can be ranked offline, whereas news articles become stale in hours or days. Building a real-time system calls for faster, more efficient techniques. Compared to simple metrics that work for academic papers [23], we need to be more careful about potential abuse (Section 5.1).

To operationalize the ideas above and address their apparent shortcomings, we extract information from the news ecosystem in global snapshots and index all notation by time t. In particular, $\mathcal{V}$ is the set of all news articles at time t, and $v \in \mathcal{V}$ denotes an

[3] Some high-quality news analysis material appears later and does not get cited much, but this is relatively rare, and such material can be promoted by other ranking signals.

Figure 4: The lifecycle of a Spiderman story

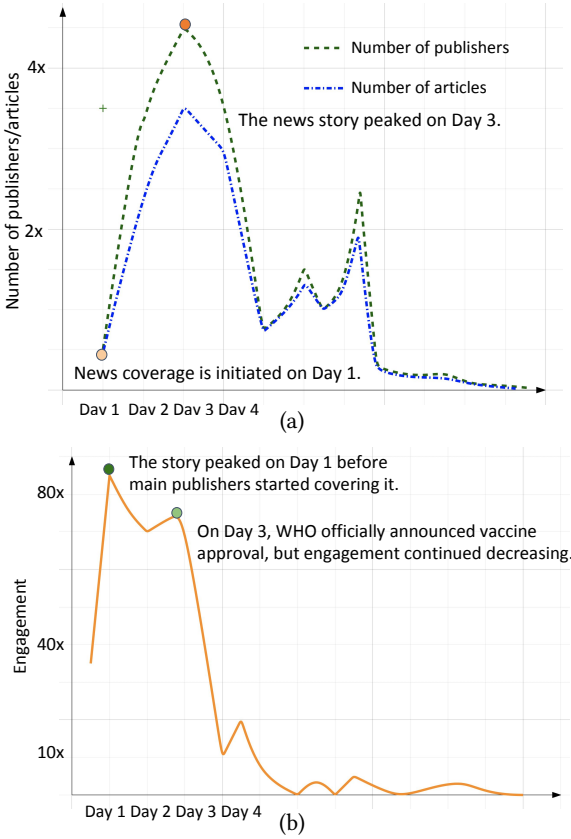

Figure 5: The lifecycle of the J&J Ebola vaccine story

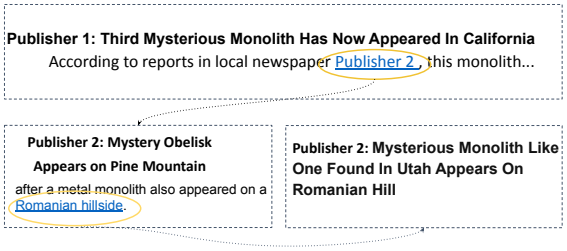

Figure 6: Citations in news articles. The top snippet cites an article by another publisher. The cited article cites another article from the same publisher.

individual article. We cluster such articles by news event or news story (Section 5.2), denoting individual clusters $C \subset \mathcal{V}$. When a news article v cites another article u, we represent this by a directed edge $e_{v,u} \in \mathcal{E}$, where $\mathcal{E}$ is the set of edges in the citation graph. Furthermore, we say that $e_{v,u}$ is v's outbound edge and u's inbound edge. Using these directed edges in the citation graph, we can compute the PageRank values of individual vertices (Appendix 2.2 and Section 5.1) by iteratively applying the following formula on every vertex in the graph in a topological order:

$$n_v = \frac{1-d}{|\mathcal{N}|} + d \sum_{u \in \mathcal{B}_v} \frac{n_u}{|\mathcal{B}_v|}, \tag{1}$$

where n_v is the PageRank of article v (initialized to 1) at time t, $\mathcal{B}_v$ denotes the set of adjacent vertices (neighbors) of vertex v, $|\mathcal{B}_v|$ is the number of neighbors of v, and d is a (constant) *damping factor*, usually set to 0.85. The latter parameter dampens the propagation of weights through multiple edges.

4.2 From citations to news originality

Intuitively, *news originality* refers to the process by which news content is created as well as the quality of news content. However, capturing these notions computationally appears challenging, especially when the content creation process remains opaque. Professional journalists and rates often find isolated text insufficient to rate originality and need additional context. Useful context includes ongoing news events and how much coverage they enjoyed, and also how a given news article is perceived by peers in the news ecosystem. A major precept in our work is that direct content analysis is neither sufficient nor necessary, whereas adequate context may provide sufficient signals to estimate originality.

To capture the context of individual news articles, we construct a *news citation graph* for the entire news inventory at a fixed-time snapshot. Peer recognition of each article is evaluated using the PageRank algorithm on this graph. An original piece of news could be cited by different publishers; it could also be a local news story cited by a major publisher with many subsequent citations — both cases are captured adequately by PageRank. Here we emphasize the use of global PageRank values not restricted to particular news events. That is because quality articles often cite out-of-topic background material and may be cited under later news events.

We try to capture news ecosystem dynamics and emulate how professional raters or journalists estimate news originality level. To this end, PageRank values cannot be compared across topics

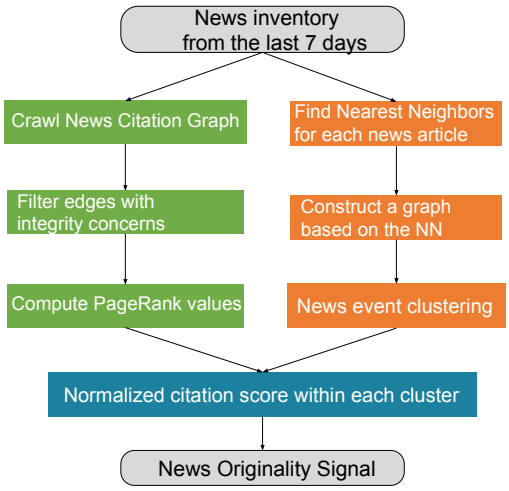

Figure 7: The workflow of our methodology.

and news events that differ greatly by the total amount of news coverage. For a given news event or news story, we consider the entire news coverage as a cluster. An insight in our work is that articles with the highest global PageRank values within each news-event cluster are most likely to be original. Therefore, we compute news originality estimates by normalizing global PageRank scores n_v within each cluster C_v as follows (see notation in Section 4.1).

$$s_v = \left(\frac{n_v^p}{\sum_{u \in C_v} n_u^p} \right)^{\frac{1}{p}}, \; p \in (0, \infty) \tag{2}$$

where C_v is the cluster of article v, and the p constant defaults to $p = 1.0$. Increasing p would favor articles with higher n_v values.

Figure 7 outlines how we estimate news originality. This approach cannot evaluate a newly published article for originality until it is recognized by peers with citations, which introduces an inherent delay and requires a near real-time system to deliver originality scores early enough in the news cycle.

When using originality scores s_v in News Feed ranking, we first convert them into P(original) $\in (0, 1]$ as follows

$$\text{P(original)} = \frac{\max(s_v, \theta) - \theta}{1 - \theta}. \tag{3}$$

Here $\theta \in (0, 1]$ is the promotion threshold, i.e., only contents with $s_i^t > \theta$ can be promoted. Then, we add P(original) to the relevance score as a second-order term:

$$\begin{aligned} \text{Relevance} = \; & \alpha_1 \cdot \text{P(comment)} + \alpha_2 \cdot \text{P(share)} + \alpha_3 \cdot \text{P(like)} \\ & + \cdots + \alpha_n \cdot \text{P(click)} \cdot \text{P(original)}. \end{aligned} \tag{4}$$

Needless to say, the proposed originality signal is just one component of News Feed ranking that elevates content recognized by peers as original. Other signals elevate other types of content.

5 IMPLEMENTATION AND SCALING

Our preliminary investigation found that news articles highly cited by other articles tend to exhibit a higher level of originality. Therefore, we first build a citation graph of all news articles published in a seven-day window. Then, we calculate global PageRank values

for individual articles, cluster news articles by news event/story in a scalable way, and normalize PageRank values within each cluster.

5.1 Integrity considerations for citations

We index all the news articles shared on the platform by leveraging the Facebook Crawler tool[4]. The Facebook Crawler tool crawls the HTML of an app or website that was shared on Facebook via copying and pasting the link or by a Facebook social plugin. There are other open-source crawlers that serve the same purpose. Common Crawl[5] is a well-maintained open repository of web crawl data that can be accessed and analyzed by anyone.

We limit the creation time of news articles in the graph to be within a seven-day moving window. After parsing the HTML, we traverse the output to get all <a> tags, which define hyperlinks to other Web pages. Hyperlinks specified in the <a> tag may point to the same Web page, but differ in URL query parameters. We resolve those URLs to Canonical URLs[6] and assign each news citation graph node a unique ID. If the cited Web page is also a recent news article, we establish an edge between the two vertices. Based on this news citation graph, we compute PageRank for each news article.

The raw citation graph is vulnerable to *link farming*, as per Du et al. [8]. That is, the graph may be manipulated by changing interconnected link structure of pages to add many inbound edges to a target page. To counter such manipulation, we disregard several types of citations before applying the PageRank algorithm (Appendix 2.2). As seen in Figure 6, one typical example is *self-linking* edges in $\mathcal{G}^t$ that cite an article published by the same publisher. Some Web sites link their articles to Web sites without real content but with auto-redirect to phishing sites or simply return to the citing article. These integrity filters mitigate the risk of manipulation. A filtered citation graph snapshot at each hour typically contains 300K–500K edges. The news articles that are not cited and not citing others are excluded when computing PageRank values.

The original PageRank calculations (Appendix 2.2) work well with graphs that exhibit cycles, created when popular Web pages are revised to link to pages published later. Unlike the Web link graph, our news citation graph mostly contains links to past content since news posts on social networks are typically not revised. PageRank calculations simplify significantly on acyclic graphs and require a single linear-time graph traversal. However, in practice our citation graph contains enough cycles to question such simplifications.

5.2 News event clustering

A large variety of clustering techniques are available in the literature and software packages that pursue different goals and satisfy different constraints. In our work, the challenge is to assemble a near real-time pipeline and find news clusters consistent with human perception, and then validate this performance. As explained in Section 4, we normalize PageRank scores for individual news articles using PageRank scores of other articles in the same cluster. Intuitively, an important national news event and a local breaking news might carry similar amount of originality, but original articles in a larger cluster get more citations and higher PageRank scores. In addition to cluster normalization, computational scalability is also important — on an uneventful day, our inventory snapshot contains 2M-3M articles, and we strive to process them in minutes.

We estimate the *topical similarity* of articles based on their titles, noting that articles with identical titles may have different PageRank scores. We first lowercase article titles, remove punctuation and hash the titles to assemble duplicates into mini-clusters. For each unique title, we calculate a vector embedding based on the powerful and adaptable BERT DNN (Appendix 2.3). Not only BERT handles synonyms and equivalent phrases well, but it also supports transfer learning. To this end, we use a Siamese-twins network architecture shown in Figure 3, previously proposed for semantic similarity estimation [20]. The two article titles are processed by the two constituent BERT models, which we implement in PyTorch using HuggingFace transformers [25]. An additional MLP layer on top of BERT is a 128-dimensional fully connected (FC) layer with *tanh* activation. In Figure 3, T_i represent the i^{th} token in input sentences With the BERT network weights fixed, the top level is trained on labeled article pairs using the cosine embedding loss function $\mathcal{L}$

$$\mathcal{L}(x_1, x_2, y) = \begin{cases} 1 - \cos(x_1, x_2) & \text{if } y = 1 \\ \max(0, \cos(x_1, x_2) - \text{ margin }) & \text{if } y = -1 \end{cases} \quad (5)$$

where x_1 and x_2 represent the two input sentences respectively. $y = 1$ means the two sentences are same news event, while $y = -1$ means the two sentences are about completely different news event.

BERT-based vector embeddings optimized to capture title similarity by cosine similarity support vector-based clustering algorithms. Algorithm choices are driven by both quality and scalability, which we need to ensure frequent refresh of the news originality signal (in the context of Section 3.2). Clustering algorithms based on K-Nearest-Neighbors (KNN) are a natural starting point, but specifying K is not straightforward and for any given K such algorithms risk producing inconsistent results in our application. Therefore, our three-step clustering in Figure 8 combines text hashing and KNN with greedy local search. Topical clusters often contain just a few different titles, while national news receive up to thousands citations per article.

The set of unique article vectors is converted into an undirected KNN graph $\mathcal{G}$. For each vector, we find its $K = 5000$ nearest neighbors based on *cosine similarity* (1 - *cosine distance*) and use cosine similarity for edge weights between adjacent vertices v_i^t and v_j^t. Lightweight edges are ignored, and subgraphs are defined by connected components of the resulting graph. Reasonable weight

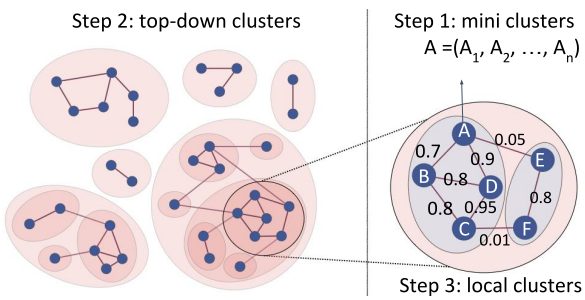

Figure 8: Three-step clustering

[4]https://developers.facebook.com/docs/sharing/webmasters/crawler
[5]https://commoncrawl.org/
[6]https://developers.facebook.com/docs/sharing/webmasters/getting-started/versioned-link

thresholds are found with a form of binary search guided by a subgraph size target. See details in Algorithm 1.

Algorithm 1: Split a graph into subgraphs with target size

Input: Weighted graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, subgraph target size t, optimization threshold ϵ, $\ell = 0.0$, $h = 1.0$
Output: A set of subgraphs $\mathcal{S}$ of approximately target size t
Function `findSubgraphs`$(\mathcal{G}, \epsilon, \ell, h)$:
  $\mathcal{S} = \emptyset$
  while $h - \ell > \epsilon$ **do**
    $m = \frac{\ell+h}{2}$
    $\mathcal{G}' = \mathcal{G}$ without edges of weight < m
    $\mathcal{C} =$ `connectedComponents`$(\mathcal{G}')$
    foreach $c \in \mathcal{C}$ **do**
      if $|c| > b$ **then**
        Remove vertices in c and their incident edges from $\mathcal{G}$
        $\mathcal{S} = \mathcal{S} \cup$ `findSubgraphs`(c, ϵ, l, m)
      end
    end
    $h = m$
  end
  $\mathcal{G}' = \mathcal{G}$ without edges of weight < m
  $\mathcal{S} = \mathcal{S} \cup$ `connectedComponents`$(\mathcal{G}')$
  return $\mathcal{S}$
end

An investigation of typical outputs of Algorithm 1 suggested that clusters were generally reasonable, but local news and events with low coverage were not handled well. To remedy this deficiency, we form local clusters using greedy optimization to maximize the total edge weight w_c inside clusters. We impart a default negative weight ω to pairs of vertices within a top-down cluster that are not connected by edges (not nearest neighbors). The smaller the ω, the harder it is to create subclusters. For details, see Algorithm 2.

Example 5.1. Figure 8 illustrates local clusters in a subgraph: $\{A, B, C, D\}$ and $\{E, F\}$. Suppose $\omega = -0.1$. Then the total edge weight in cluster 1 is $w_1 = 0.7 + 0.8 + 0.8 + 0.9 + 0.95 - 0.1 = 4.15$ (no edge between A and C), and in cluster 2 $w_2 = 0.8$. Although A and E are connected, the edge weight is so low that adding E would not increase the total weight of cluster 1. The same reasoning applies to F. Therefore local clustering produces two clusters.

5.3 Scalability

Building and processing the KNN graph with $K = 5000$ nearest neighbors per vertex is a major performance bottleneck. On a typical day, all news articles from the last week fit in the RAM of a single server and can be processed reasonably quickly. However, this architecture is insufficiently scalable for the following reasons.

- *Potential surges* of the news inventory during the election season, the New Year's Eve, etc.
- *Near real-time processing* benefits from additional compute resources (lower processing latency via using multiple servers).

Algorithm 2: Greedy local clustering

Input: Weighted graph $\boldsymbol{g}$, negative weight ω for missing edges, number R of independent randomized passes
Output: An integer c_v for each vertex v (cluster assignment)
repeat R **times**
  Randomize the order of vertices in $\boldsymbol{g}$ Initialize each vertex v in its own cluster $\boldsymbol{c}_v$
  foreach $v \in \boldsymbol{g}$ **do**
    foreach $u \in \mathcal{B}_v$ **do**
      Try moving v from cluster c_v to cluster c_u Add up internal weights for c_u and c_v Record u with the highest sum of weights seen
    end
    Move v to maximize the sum of weights of c_v and c_u
  end
  if $\sum w_c$ *increased* **then**
    repeat **foreach** $v \in \boldsymbol{g}$
  else
    Record the solution with the highest $\sum w_c$ seen
  end
until

- *Need for scaling to larger content inventory.* The challenge we are solving and our methods are fairly general, so can be applied to other social-network platforms that value originality. Now or in the future, such platforms may enjoy a much larger scale of content inventory.

The overall design described in Section 5.2 naturally supports distributed processing to ensure greater overall scalability and robustness to surges. In fact, this is why Algorithm 1 performs *balanced* partitioning. Our implementation supports distributed clustering as well. We found that the upper bound on single-server capacity is an important parameter — individual servers must receive a sufficient amount of work to justify distributed processing, but the data must fit into available RAM. Between the implied lower and upper bounds, there is a transition point where one can reduce the amount of computation at the cost of greater processing latency.

6 EVALUATION AND DEPLOYMENT

Before deploying our news originality signal to production at Facebook, we evaluate its functional components individually, evaluate the entire signal with the help of professional raters, then embed the signal into News Feed ranking and explore examples to check that everything works as expected. The production deployment is evaluated with an A/B test on live data for a limited subset of users before it is enabled for the main group of users.

6.1 Evaluation of embeddings and clustering

In our rating flow, we ask professional raters to review pairs of news articles. The raters assign a similarity level to each pair of articles: *different subjects*, *different subject but some common contents*, *same subjects with different aspects*, and *same subjects* (the four levels are explained in Table 2). For training, we collect 100K pairs of randomly sampled English news titles, using 40% for finetuning, 10% for validation, and 50% for test. Separately, we collect another

Table 2: Guidelines for rating the similarity of article pairs

Score	Rating	Criteria
0.0	different subjects	the two articles cover completely different subjects
1.0	different subjects / some commonality	the two articles cover different subject but with share some content
2.0	same subject / different aspects	the two articles cover the same subject but report different aspects of the same story
3.0	same subject	the two articles cover the same subjects

10K pairs of news articles to evaluate clustering performance. To sample likely-positive examples, we take some number of closest neighbors in terms of document embeddings and/or text similarity. Likely-negative samples are drawn from further-away neighbors that are sufficiently close to make the labeling task nontrivial.

To compare our vector embeddings with FastText [12] and Pytorch-BigGraph [15] embeddings, we represent similarity levels numerically by 0.0, 1.0, 2.0, 3.0 during training following Table 2. During evaluation, we binarize model scores at thresholds 0.5, 1.5 and 2.5, then use ROC AUC as the evaluation metric. For example, AUC@0.5 considers article pairs with cosine similarity ≥ 0.5. Table 3 describes the performance of our BERTPairwise model, which consistently outperforms pre-trained state-of-art embeddings.

To evaluate our news-event clustering vs. human labels, we randomly sample 10K pairs of news articles in English from the candidate pool and send the pairs to professional annotators, along with guidelines in Table2. Then, we apply the clustering algorithms to the entire candidate pool. For each sampled pair, if the two articles appear the same cluster, the predicted label is positive, otherwise — negative. The clustering algorithm is evaluated by precision and recall, then compared with two well-known algorithms in Table 4. `DBSCAN` (density-based spatial clustering of applications with noise)[9, 22] is a highly scalable density-based algorithm. The `Louvain` algorithm [1] is one of the fastest and best-known community detection algorithms for large networks.

6.2 Evaluation by professional raters

To assess the accuracy of our citations score signal, we sample the most viewed news articles identified as original, and the most viewed article not identified as original from the most viewed news

Table 3: The pairwise embedding vs. FastText [12] and Pytorch-BigGraph [15] embeddings

Model	AUC @ 0.5 (%)	AUC @ 1.5	AUC @ 2.5
FastText	80.20	83.89	89.66
BigGraph	82.95	84.87	89.61
BERTPairwise	83.66	88.67	96.13

Table 4: The performance of three-stage clustering with DBSCAN [9] and the Louvain algorithm [1]

Algorithm	Precision	Recall
DBSCAN	43.07	73.04
Louvain	81.01	47.57
Stage 1 + Louvain	81.85	32.63
three-stage clustering	**83.73**	**45.33**

Table 5: Originality rating guidelines for human raters

Score	Rating	Criteria
1.0	unoriginal	borrows most of the content and language from other sources or is extremely thin / low information overall, and anything that is not properly syndicated.
2.0	possibly/somewhat unoriginal	rewords borrowed content with its own language, but >70% is borrowed OR properly syndicated
3.0	fully original	is not a syndicated republishing, little to no content is borrowed

domains over a seven-day period. Our professional raters have many years of news-industry experience and follow a deliberate process to ensure fair judgement for each article they rate on a three-point scale of news originality (Table 5). For the rating 3.0, our predicted labels match these results 90% of the time. In other words, our signal attains 90% accuracy in identifying original news.

6.3 An illustrative example

Besides the quantitative evaluation, we also performed qualitative case studies. Here we describe one example that illustrates how our system works. On January 26, 2020, an article *n* about the death of Kobe Bryant in a Calabasas helicopter crash was first reported by the publisher TMZ[7]. In just 10 minutes, many publishers covered this story and cited TMZ. Over 200 articles fell into this news-event cluster, and the original story by TMZ ranked the highest. For such events, users would see news articles posted by the newspages they follow and shared by their friends. If the original news article is in a users' feed inventory, it gets prioritized. Note that our originality signal is only one component in the ranking formula. Users with preferences for certain publishers or strong affinity with their friends continue seeing articles shared by those actors.

6.4 Production deployment and evaluation

The originality signal is intended for the relevance score calculation (see Figure 1 and Equation 4) to increase the distribution of original news articles. To ensure its availability early in the news cycle, it is

[7]TMZ: https://www.tmz.com

Table 6: User engagement lift in promoting original news

Originality threshold	Increase in num. views (%)
0.4	15.36
0.5	14.72
0.6	14.30
0.7	13.83
0.8	13.38

recalculated from scratch on an hourly basis. Building the news citation graph and news clusters takes only a few minutes, but system bottlenecks are observed in our current crawling infrastructure and in generating vector embeddings. In practice, it takes time for the original articles to get cited, but running the workflow more often could find and promote original articles earlier. Such improvements are likely with further infrastructure optimization.

Before making proposed changes to News Feed ranking at Facebook, we consulted with the academic and publishing communities and performed careful empirical evaluation. In particular, we ran an A/B test on live data for several weeks, where the control group used prior production ranking rules and a small test group used revised ranking rules. To estimate impact, we computed the increase in view counts at different thresholds (Table 6) and found the percentages stable across different thresholds. We have not observed statistically significant deteriorations in our proprietary metrics during our A/B test or after the subsequent full product launch. We have been tracking a goal metric called News Ecosystem Quality score. It is a synthetic score that combines several proprietary metrics such as clickbait prevalence. We observed a statistically significant score increase of **0.41%** in our experiment. After additional checks and consultations, our signal was enabled for English-language content within Facebook's News Feed ranking system for most users in June 2020.[8]

7 CONCLUSIONS AND PERSPECTIVES

In this paper, we introduce a strategy to prioritize original news in social networks. This strategy computes PageRank scores of news articles and estimates originality by normalizing PageRank scores for each news event. Equation 2 is a particularly novel contribution.

We deployed the originality signal to personalized Facebook News Feed, which compiles articles from sources followed by the user and user's friends. When multiple articles are available in a user's inventory, we promote the more original ones. While subtle, such changes influence what the community sees. As part of our work, we performed conceptual, qualitative and quantitative evaluation to confirm that our techniques positively impact the news ecosystem. In particular, the exposure of original content has grown, and users received more content they liked. Over a longer timeframe, these developments should encourage publishers to invest more in original content.

[8] https://about.fb.com/news/2020/06/prioritizing-original-news-reporting-on-facebook/

REFERENCES

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *J. Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), 10008.

[2] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* 30 (1998), 107–117. http://www-db.stanford.edu/~backrub/google.html

[3] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2015. Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems* 80 (2015), 56–71.

[4] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proc. WWW*. ACM, Perth, Australia, 963–972.

[5] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. 2012. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *Proc. WSDM*. ACM, Washington, USA, 153–162.

[6] Gianna M Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proc. WWW*. ACM, Chiba, Japan, 97–106.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *In Proc. 17th NAACL*. ACL, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Ye Du, Yaoyun Shi, and Xin Zhao. 2007. Using spam farm to boost PageRank. In *Proc. Intl. Workshop on Adversarial Information Retrieval on the Web*. ACM, Banff Alberta, Canada, 29–36.

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Proc. KDD*, Vol. 96. AAAI Press, Portland, Oregon, USA, 226–231.

[10] Robert Gwadera and Fabio Crestani. 2009. Mining and ranking streams of news stories using cross-stream sequential patterns. In *Proc. ACM Conference on Information and Knowledge Management*. ACM, Hong Kong, China, 1709–1712.

[11] Yang Hu, Mingjing Li, Zhiwei Li, and Wei-ying Ma. 2006. Discovering authoritative news sources and top news stories. In *Asia Information Retrieval Symposium*. Springer, Beijing, China, 230–243.

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models.

[13] Nattiya Kanhabua, Roi Blanco, and Michael Matthews. 2011. Ranking related news predictions. In *Proc. Intl. ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Beijing, China, 755–764.

[14] Jinhyuk Lee et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[15] Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. 2019. Pytorch-biggraph: A large-scale graph embedding system. *Proc. ML Sys Conference* 1 (2019), 12 pages.

[16] Adam Mosseri. 2018. News Feed Ranking in Three Minutes Flat. https://newsroom.fb.com/news/2018/05/inside-feed-news-feed-ranking/

[17] Xiuyan Ni and other. 2019. Feature Selection for Facebook Feed Ranking System via a Group-Sparsity-Regularized Training Algorithm. In *Proc. 28th ACM Intl. CIKM*. ACM, Beijing, China, 2085–2088.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. *The PageRank citation ranking: Bringing order to the Web.* Technical Report. Stanford InfoLab.

[19] Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Trans. ACL* 7 (2019), 249–266.

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of EMNLP*. ACL, Hong Kong, China, 3982–3992. https://arxiv.org/abs/1908.10084

[21] J. Reis and all. 2015. Breaking the news: First impressions matter on online news.

[22] E. Schubert and all. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans. on Database Sys. (TODS)* 42, 3 (2017), 1–21.

[23] Sotaro Shibayama and Jian Wang. 2020. Measuring originality in science. *Scientometrics* 122, 1 (2020), 409–427.

[24] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias De Amorim, and Serge Fdida. 2014. From popularity prediction to ranking online news. *Social Network Analysis and Mining* 4, 1 (2014), 174.

[25] Thomas Wolf et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proc. of EMNLP*. ACL, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[26] Junting Ye and Steven Skiena. 2019. MediaRank: Computational ranking of online news sources. In *Proc. KDD*. ACM, Anchorage, AK, USA, 2469–2477.

[27] A. X. Zhang et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Proc. WWW*. ACM, Lyon, France, 603–612.

[28] Guanjie Zheng et al. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proc. WWW*. ACM, Lyon, France, 167–176.

[29] Yukun Zhu and all. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proc. IEEE Intl Conf. Computer Vision*. IEEE Computer Society, Santiago, Chile, 19–27.