NU:BRIEF – A Privacy-aware Newsletter Personalization Engine for Publishers

Ernesto Diaz-Aviles
recsyslabs
University College Dublin
Ireland
ernesto@recsyslabs.com

Claudia Orellana-Rodriguez
recsyslabs
University College Dublin
Ireland
claudia@recsyslabs.com

Igor Brigadir
recsyslabs
University College Dublin
Ireland
igor@recsyslabs.com

Reshma Narayanan Kutty
recsyslabs
University College Dublin
Ireland
reshma@recsyslabs.com

ABSTRACT

Newsletters have (re-) emerged as a powerful tool for publishers to engage with their readers directly and more effectively. Despite the diversity in their audiences, publishers' newsletters remain largely a one-size-fits-all offering, which is suboptimal. In this paper, we present NU:BRIEF, a web application for publishers that enables them to personalize their newsletters without harvesting personal data. Personalized newsletters build a habit and become a great conversion tool for publishers, providing an alternative readers-generated revenue model to a declining ad/clickbait-centered business model.

Demo: https://demo.nubrief.com/md03PaAJSwXMegL5BbKpQlA rK3elb3hDUglcHodx4gE=/
Explainer video: https://www.youtube.com/watch?v=AUZGuyP JYH4

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Newsletter Personalization, AI, Federated Learning, ML, NLP, Privacy, Personalized Ranking

1 INTRODUCTION

Newsletters are an attractive way for publishers to quickly build an audience and convert people into paying subscribers within weeks. For publishers, newsletters have become a viable revenue stream and an alternative to a declining click-bait business model.

For example, the New York Times has about 15 million subscribers across its 71 newsletters compared to 6.69 million digital-only subscriptions. In total, readers opened more than 3.6 billion newsletter emails from the publisher in 2020. The New York Times reports that digital revenue overtook print revenue, and digital subscriptions was its largest revenue stream. [6, 13]

However, the accelerated growth observed by major outlets such as the New York Times requires large amounts of resources for newsletter development, such as, an interdisciplinary team of technical staff, editorial specialists, and project managers. For small to medium publishers or independent writers with a limited budget, it is nearly impossible to tap on resources like these to maintain a diverse choice of newsletters. Their option is to settle on the better-than-nothing alternative to send the same newsletter to their diverse audience. This is suboptimal and clearly not on a growth path as the one reported by major media publishers.

Small to mid size publishers, however, do not have the legacy of a heavy machinery of old-line publishers, but have a better chance to innovate faster and start operating more like a digital product and technology company. One can observe the emergence of Neo-Media players serving and delighting audiences neglected by old-media that forgot their readers in exchange for minimal commission on click-bait ads served alongside their content. One key characteristic of these new players is that they are more respectful of reader privacy as a consequence of the relationship and trust they are looking to establish with their readers, who are increasingly privacy-conscious. In addition, publishers do not want the risk and liability of collecting and storing personal information given the intense scrutiny from data regulators.

NU:BRIEF, the tool we introduce in this work, enables publishers to monetize quality journalism by personalizing the newsletter experience for readers without the risk of collecting personal identifiable information, thereby moving away from a volatile advertising business. NU:BRIEF uses AI based on Machine Learning to automatically segment the publishers' audience into cohorts based on
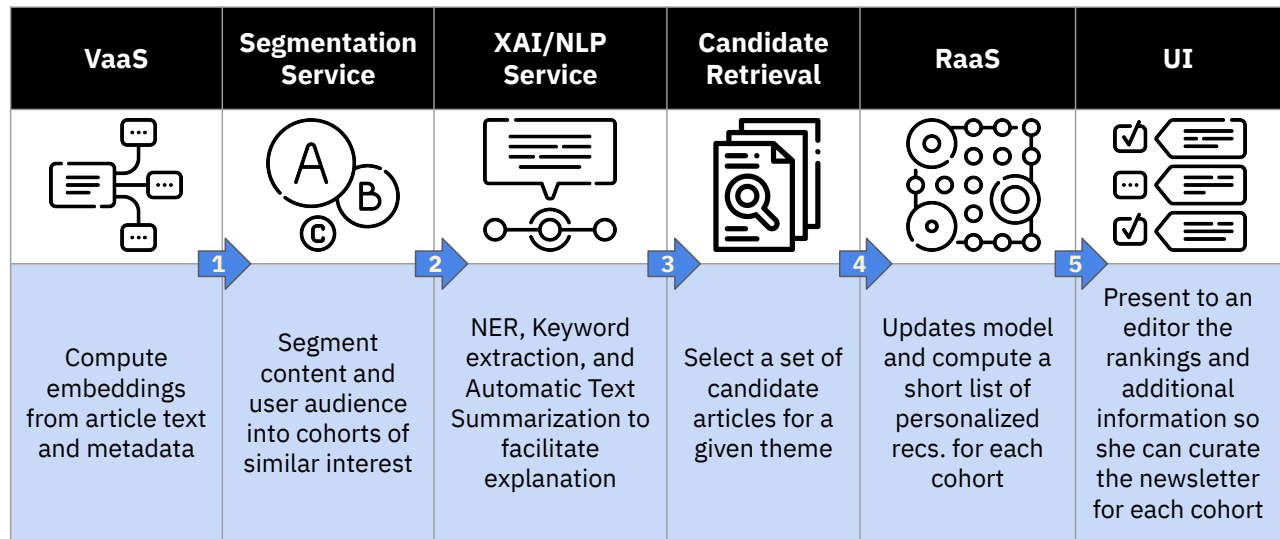
VaaS	Segmentation Service	XAI/NLP Service	Candidate Retrieval	RaaS	UI
Compute embeddings from article text and metadata	Segment content and user audience into cohorts of similar interest	NER, Keyword extraction, and Automatic Text Summarization to facilitate explanation	Select a set of candidate articles for a given theme	Updates model and compute a short list of personalized recs. for each cohort	Present to an editor the rankings and additional information so she can curate the newsletter for each cohort

Figure 1: NU:BRIEF pipeline overview.[1]

their interests and to produce a newsletter for each cohort with interesting articles tailored to the users' taste in each of these groups, without harvesting personal identifiable information or selling it to third parties.

In the rest of the paper, we introduce user stories to give context on how publishers are currently using NU:BRIEF to achieve their goals, we provide a technical overview of the system implementation, and present preliminary results on the online tests conducted.

USER STORIES

Let us introduce two user stories to illustrate the scenarios in which NU:BRIEF can assist publishers to personalize their newsletters.

User Story 1. "As an editor, I want to be able to select a short list of interesting articles from a set of candidates about a theme of interest, so I can write an introduction about the theme and include my article selection in our weekly newsletter."

User Story 2. "As an editor, I want to be able to rank (with the most relevant first) the articles we wrote this week to better inform and make our audience happy, so I can include them in our weekly newsletter."

2 NU:BRIEF OVERVIEW

Figure 1 shows NU:BRIEF's main pipeline. We detail each component as follows.

Vector as a Service (VaaS). VaaS is responsible for computing vector representations, i.e., embeddings, from the text content of articles and its metadata. These embeddings are latent features extracted automatically using pre-trained NLP large language models based on transformers [3, 7, 12]. The standard flow is that articles from publishers using NU:BRIEF are initially processed offline during publisher onboarding. The articles are also indexed and the

---

[1]Icon illustrations designed by Freepik https://www.freepik.com.

embedding is stored. New articles are processed periodically during the day, e.g., every 15m, to keep a cache system and index up to date.

Segmentation Service. This component is responsible for discovering, initially from the content, a set of cohorts of interest. The rationale is that publishers write articles for a diverse audience and the topics of their production reflect the diversity of their target readers. We use k-means clustering [9] to this end and the embeddings output by the VaaS component as input. We have found that for publishers currently using NU:BRIEF a typical number of clusters (cohorts) k ranges between 4 and 6, which is determined based on clustering metrics such as silhouette coefficient and a qualitative assessment.

XAI/NLP Service. The goal of this component is to enrich the articles with additional information that can help editors, and readers, understand why a particular article is recommended. To this end, it extracts named entities, keywords, and automatic summaries from the articles. NU:BRIEF also computes embeddings using VaaS for these additional text elements and caches them to be used in later stages of the pipeline.

NU:BRIEF computes two automatic summaries, one *abstractive* and one *extractive* using transformers-based NLP models [8, 15]. In abstractive summarization the machine summarizes in its 'own' words what the article is about using language generation techniques, that is, the output concise text might not appear exactly in the input article. In contrast, extractive summarization identifies salient information in the article, which is then extracted and grouped together to form a concise summary. Both summaries are presented to the editor in the user interface.

Candidate Retrieval Service. This service is responsible for indexing articles, their metadata, and embeddings. It is implemented as a search engine that assists, in production, the selection of a subset of candidate articles relevant to a specific theme or context.

(a) Define a "theme".   (b) Select a time range.   (c) Recommendations for each cohort.
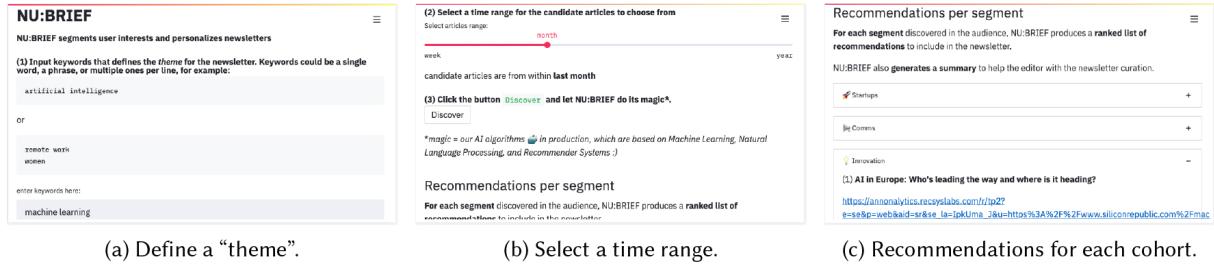
Figure 2: `NU:BRIEF` User Interface. Personalized newsletters in three basic steps.

These candidate articles are the ones to be re-ranked by the RaaS component to compute the personalized recommendations. This component is critical during inference since the cardinality of the candidates subset is significantly smaller than the whole set of articles indexed.

Privacy-aware Recommender as a Service (RaaS). One of our design principles for NU:BRIEF is to develop a recommender system service that is private-by-default, compliant with privacy regulations, and offers a high quality personalization. As part of this goal, NU:BRIEF computes an aggregate representation of a user's taste (*taste vector*) based on their history interactions (e.g., clicks) with the recommended content. Differential privacy [5], e.g., data perturbation, is applied in this step for additional privacy guarantees.

The core recommender system engine uses a hybrid approach of content-based [1], matrix factorization collaborative filtering [4, 10] and k-nearest neighbors (k-NN) [14]. In our experience, deep learning architectures have served us well in production for learning representations from text and in the XAI pipeline, e.g., automatic summarization, but we have found that relatively simple recommender system algorithms, e.g., embedding based models and dot product operations on those embeddings to compute scores, have proven very practical to deploy in production and have performed well in live tests. In addition, offline evaluations we have conducted show that these methods have not been inferior to neural-based models for recommendation, which is in line with recent studies, e.g., [2, 11]. We are currently experimenting with Deep and Cross Network [17] architectures and with attention mechanisms [16] to improve model interpretability guided by high-level journalistic features.

As a side product of our recommender technology, we generate privacy-preserving analytics which can inform how cohorts of users interact with content in different contexts and scenarios, without singling out or exposing any individual.

User Interface. NU:BRIEF user interface is shown in Figure 2. It is designed so that the newsletter editor fulfills the user stories presented in Section 1. There are three basic steps: (i) define newsletter theme using phrases or keywords, (ii) specify a time range for candidate articles to consider, and (iii) receive personalized recommendations per cohort of interest. NU:BRIEF versions in production also offer an export to html functionality for publishers so they can include the selected articles recommended into their newsletter distribution system.

3 EVALUATION

Over the first half of 2021, we have conducted a pilot test in a private beta with small-to-medium publishers whose content covers topics ranging from technology, STEM, startups, local news, and food recipes. These publishers serve a combined audience of more than 43K subscribers for their newsletters. In total, NU:BRIEF pipeline has processed 135K documents published by the pilot publishers.

The main lessons learned from A/B tests conducted using NU:BRIEF are as follows. (i) Personalized rankings per cohort computed by NU:BRIEF perform at the same level as the ones curated by a human editor. That is, newsletters produced based on NU:BRIEF rankings achieve opening rates and CTR that are equivalents to the ones achieved by the ones curated entirely manually by a journalist. (ii) Publishers report that time to curate a newsletter drops from one hour to 10 minutes when using NU:BRIEF (80% time savings). (iii) The conversion rate from newsletter readers to paying subscribers increases with the personalization options offered by NU:BRIEF.

4 CONCLUSION

In this work, we have given an overview of our approach, NU:BRIEF, towards building a private-by-default, personalized newsletter recommender engine. NU:BRIEF produces a ranked list of recommendations per segment (cohort) of common interest to assist editors in newsletter curation. As private-by-default recommender systems do not exist off-the-shelf, our work will allow publishers to provide personalized newsletters without violating privacy regulations, and allow users to find content they enjoy without sacrificing their private data.

REFERENCES

[1] John S. Breese, David Heckerman, and Carl Kadie. 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (Madison, Wisconsin) *(UAI'98)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 43–52.
[2] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 101–109. https://doi.org/10.1145/3298689.3347058

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[4] Ernesto Diaz-Aviles, Lucas Drumond, Lars Schmidt-Thieme, and Wolfgang Nejdl. 2012. Real-Time Top-n Recommendation in Social Streams. In *Proceedings of the Sixth ACM Conference on Recommender Systems* (Dublin, Ireland) *(RecSys '12)*. Association for Computing Machinery, New York, NY, USA, 59–66. https://doi.org/10.1145/2365952.2365968

[5] Cynthia Dwork and Aaron Roth. 2014. *The Algorithmic Foundations of Differential Privacy*. NOW, Lange Geer 44 A 2611 PW, Delft, Zuid-Holland, Netherlands. https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

[6] Edmund Lee. The New York Times. 2021. The New York Times Tops 7.8 Million Subscribers as Growth Slows. https://www.nytimes.com/2021/05/05/business/media/nyt-new-york-times-earnings-q1-2021.html. Accessed: 2021-07-10.

[7] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. arXiv:1909.10351 [cs.CL]

[8] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. arXiv:1908.08345 [cs.CL]

[9] S. Lloyd. 2006. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theor.* 28, 2 (Sept. 2006), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[10] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) *(UAI '09)*. AUAI Press, Arlington, Virginia, USA, 452–461.

[11] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. 2020. Neural Collaborative Filtering vs. Matrix Factorization Revisited. In *Fourteenth ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 240–248. https://doi.org/10.1145/3383313.3412488

[12] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL]

[13] Sara Guaglione. Digiday. 2021. The New York Times aims to convert newsletter readers into paid subscribers as The Morning newsletter tops 1 billion opens. https://digiday.com/media/the-new-york-times-aims-to-convert-newsletter-readers-into-paid-subscribers/. Accessed: 2021-07-10.

[14] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (Hong Kong, Hong Kong) *(WWW '01)*. Association for Computing Machinery, New York, NY, USA, 285–295. https://doi.org/10.1145/371920.372071

[15] Sam Shleifer and Alexander M. Rush. 2020. Pre-trained Summarization Distillation. arXiv:2010.13002 [cs.CL]

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[17] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. DCN V2: Improved Deep and Cross Network and Practical Lessons for Web-Scale Learning to Rank Systems. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1785–1797. https://doi.org/10.1145/3442381.3450078