

Newsletter-Factory: A Thematic Newsletter Generation Tool for Curating Business Insights

Siddharth Tumre, Alok Kumar, Ajay Phade, Nihar Riswadkar and Sangameshwar Patil

TCS Research, Pune, India

Abstract

Keeping track of the evolving business and technology landscape is essential for enterprises to remain relevant, competitive, and successful in today's dynamic marketplace. Gaining insights into how business narratives unfold is crucial for strategic decision-making, risk assessment, and market analysis. For enterprise executives, filtering through vast amounts of news content to identify what is relevant and useful can be extremely time-consuming as well as act as a distraction from other core tasks for their role. Newsletters provide a solution by curating the most relevant information, reducing information overload, and delivering value to targeted readers. Creating periodic newsletters that deliver value to the target audience and achieve business objectives require a combination of domain knowledge, business expertise, and technology awareness across various disciplines—making it an effort-intensive and expensive proposition for enterprises. In this paper, we demonstrate Newsletter-Factory, a tool that can create thematic and customizable newsletters based on enterprise requirements. It efficiently processes huge amounts of data from multiple sources and leverages near-duplicate detection techniques to eliminate redundancy. The tool categorizes news articles into high-level themes and also generates fine-grained sub-labels enabling comprehensive understanding of emerging trends and key narratives. By automating the newsletter creation process, Newsletter-Factory alleviates the pain and cost of creating enterprise newsletters targeted towards business users and employees.

Keywords

News Analytics, Enterprise Information Dissemination, Thematic Story based Newsletter Generation

1. Introduction

A newsletter is a periodic publication, usually distributed via email, that contains curated content on a specific topic. Typically, it targets a specific audience of subscribers who are interested in the topic of the newsletter. Newsletters are a versatile communication tool for enterprises as well. They can be used for internal communication with focused subsets of employees as well as for engaging customers and external stakeholders for marketing, brand building etc. When executed effectively, newsletters can contribute significantly to the success and growth of an enterprise.

Newsletters are a powerful tool for building a compelling story over time. By delivering consistent content, they allow a narrative to develop gradually, keeping readers engaged and invested. Personalized stories tailored to a user's experience create a deeper connection, making the content more relatable and impactful. Additionally, newsletters are an effective way to strengthen brand storytelling, giving companies a platform to share their values, mission, and vision in an ongoing dialogue. They also offer a unique opportunity to highlight changes and progress over time, whether it's through product updates, company milestones, or evolving narratives, making the storytelling more dynamic and meaningful.

Enterprises often operate in diverse domains or industries and utilize various technologies to support their operations. Thematic newsletters focusing on different domains and technologies enable employees to stay updated on the latest trends, best practices, and advancements within their respective fields. This knowledge sharing fosters innovation, enhances skill development, and ensures that employees have access to the information they need to perform their roles effectively. By highlighting developments in various industry sectors and geographic regions, newsletters provide valuable market insights and trends to leadership and sales teams. This information helps them understand the evolving needs

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): *Proceedings of the Text2Story'25 Workshop, Lucca (Italy), 10-April-2025*

✉ firstname.lastname@tcs.com



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0

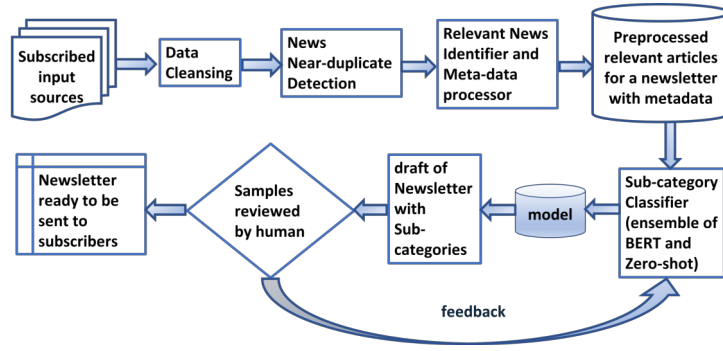


Figure 1: High-level block diagram of key components in Newsletter-Factory

and preferences of customers, anticipate market shifts, and identify new opportunities for growth and expansion.

Creating high-quality newsletters to track business events and industry trends is an effort-intensive and expensive proposition for enterprises. It requires a combination of domain knowledge, technological capabilities, and strategic thinking. Finding relevant and valuable content can be time-consuming. Sorting through vast amounts of information to identify what’s important and interesting to the subscriber audience requires careful curation. Many enterprises have dedicated teams responsible for creating internal and external communications, including newsletters. These teams understand the company’s branding, messaging, and goals intimately. Such expertise is built over years of experience and the time and effort required is significant.

From another aspect, industry trends evolve rapidly, and enterprises need to stay up-to-date with the latest developments. Also, the sheer volume of business relevant news and information is becoming overwhelming. As a result, the demand for industry or technology specific newsletters that curate the input content and tailor it to the interests and needs of subscribers has been increasing [1]. Further, for newer technologies (e.g., quantum computing, blockchain, space and satellite technology), there is scarce availability of manpower who have the requisite domain knowledge and understand the nuances. These developments not only increase the cost and effort required, but also pose challenges in creating newsletters on new emerging topics and themes.

In this paper, we present Newsletter-Factory, a tool that creates industry and technology specific newsletters for enterprise users. The tool helps to automate creation of existing newsletters within an enterprise as well as generation of fresh newsletters on-demand from scratch based on emerging requirements. Such requirements can arise from enterprise executives such as the senior leadership as well as the customer facing teams such as the sales, pre-sales teams. Newsletter-Factory improves the agility and responsiveness of the enterprise communication team which is in-charge of serving such information needs and deliver timely insights. Newsletter-Factory tool provides feedback feature to enterprise users which aids model re-training and performance improvement of classifiers to identify relevant sub-categories for each newsletter. The tool automates various tasks in the information management pipeline of newsletter generation process. The Newsletter-Factory tool currently supports news articles published in English, with plans to expand support for multiple languages in the future.

Rest of the paper has been organized as: In Sec. 2, we give the Newsletter-Factory tool and the various components in it. We highlight the key information processing challenges related to handling near-duplicate news articles. We also improve the data management of ingested news corpus by the enriching it with meta-data as well as sub-categories within a newsletter. Sec. 3 provides experimental evaluation. Sec. 4 gives a brief overview of related work. We conclude and discuss future work in Sec. 5.

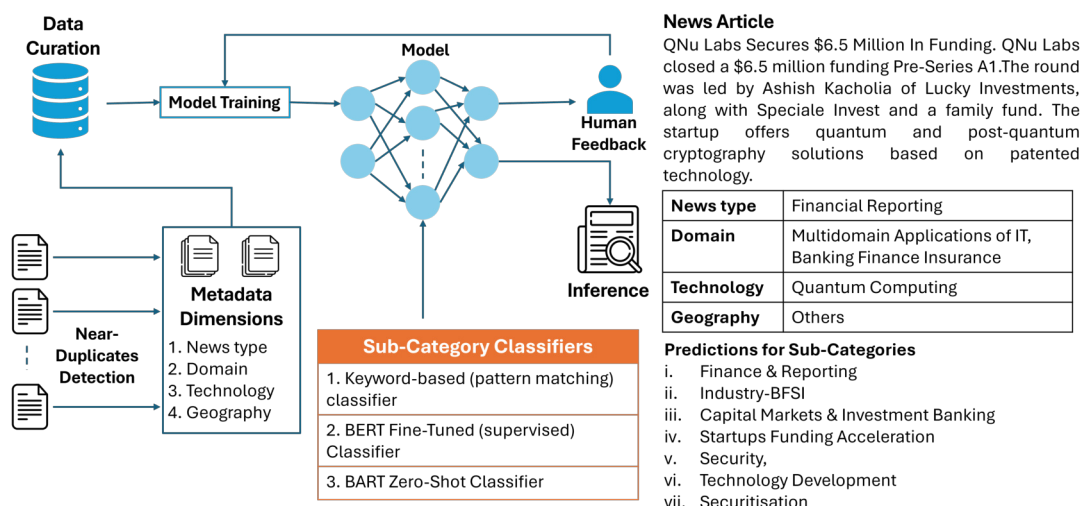


Figure 2: Illustrative example of sub-category classification step for enrichment of raw news data

2. Overview of the Newsletter-Factory Tool

The key steps and components in the Newsletter-Factory (NLF) tool are shown in Figure 1. The first step of data ingestion periodically gathers recently published news data from the vendor and news aggregators. After parsing input data from various vendor specific formats into a standardized format, the NLF tool addresses the key challenges of (i) handling near-duplicate news articles, (ii) identifying relevant articles for the theme of newsletter and avoiding irrelevant news for its subscribers, (iii) improving and organizing the semi-structured news content by attaching informative sub-categories to each news, (iv) model improvement employing human-in-the-loop for sub-category classification. We provide details of each component below.

2.1. Near-duplicate detection and handling

As the vendors aggregate news from various sources, different versions of the same underlying news tend to get reported by various news sources. NLF can detect and cluster near-duplicate articles using two techniques. First technique uses Locality Sensitive Hashing (LSH) based machine learning techniques [2] to cluster similar news articles. The LSH based technique uses MinHash for the task of near duplicate detection. It is based on the Shingling method proposed by Broder et al. [3]. Second technique utilizes a more recent approach that utilizes transformers based model embeddings (all-mpnet-base-v2¹) augmented with metadata and followed by community detection [4, 5, 6] to cluster near-duplicate news articles.

Storage efficiency is critical for a newsletter generation system as it has to deal with extensive content archives. Efficient storage solutions can significantly reduce costs. Efficient near duplicate detection component has helped the Newsletter Factory tool in – (i) saving the computational cost of processing similar articles, (ii) optimizing data storage and improving the efficiency by saving only the analytical processing results for the representative article for a cluster of near duplicate news articles, and (iii) improving the end-user experience and engagement by avoiding redundant content as well as increasing the diversity of news covered in the newsletters.

2.2. Metadata (Newsletter theme) classification

At this step, the tool tries to capture the high-level theme of a news article across various dimensions such as *domain*, *technology*, *newstype* and *geography*. We utilize an ensemble of pattern-based categorizers

¹<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

and machine learning classifiers to tag each news article with appropriate metadata. This metadata helps to identify the relevant news articles that match the theme of the target newsletter (e.g., a healthcare domain focused newsletter or a telecom domain newsletter focusing on advances in 5G technology). After this step, the set of relevant news articles for a specific newsletter are ready in the database of the preprocessed news articles.

For the illustrative example in Figure 2, the Meta-data processor tags sample news-article with the Newstype, Domain, Technology, Geography related metadata. This metadata helps to identify relevance of the news article for a specific newsletter. For instance, the news shown in the above figure would be considered as relevant for a Quantum Computing newsletter. Many existing newsletters (including some of the manually created newsletters) stop at this stage and send out the relevant news within a specific period to the subscribers. Newsletter-Factory enriches this relevant news further by attaching relevant sub-categories to them.

2.3. Subcategory Classification for Each News Article

Figure 2 shows an illustration of the sub-categories attached to a news article. These sub-categories help to further filter and focus on a subset from the relevant articles. The sub-category classifier is an ensemble of different classifiers (i) A high-precision, keyword-based pattern matching classifier, (ii) Efficient Adapter Fine-tuned BERT classifier, (iii) BART-based zero shot classifier. The top predictions generated by the Keyword-based (pattern matching) classifier, along with those from BERT and BART models with confidence above a certain probability threshold, are considered correct predictions. We provide a more detailed description of these components below.

Keyword-based (pattern matching) classifier: This is a simple, yet high precision classifier that looks for repeating phrases within a news article, to perform classification. Every classification label is assigned with a relevant list of keywords and phrases. For every label, we filter out all the stopwords and find the top-10 most frequent unigrams and bigrams. From this most frequent n-grams we select keywords that are representative of the classification label. Further we would like to incorporate advanced keyword extraction techniques like RAKE [7], YAKE [8], TopicCoRank [9]. The news article is searched for all such keywords and phrases to assign a score for the classification label based on the number of occurrences, position of keyword/phrase and length of news article (e.g., for the label Artificial Intelligence we search for AI, Computer Vision, Deep Learning etc.). The labels with score greater than the predetermined threshold are tagged to the news article. The feedback from the user can further help to identify keywords that are relevant to the classification label.

Efficient adapter fine-tuned BERT classifier: This component utilizes a supervised learning paradigm, making it beneficial when training data is available. Given the wide variety of newsletter themes, fine-tuning an entire BERT model for each theme is computationally intensive and resource heavy. Consequently, finding a more efficient method for fine-tuning or adapting models is essential for managing diverse newsletter topics effectively. Recently, adapter-based methods [10, 11, 12] have gained popularity for parameter efficient fine-tuning by adding extra trainable parameters into the model architecture. The rest of the models' pre-trained weights are fixed, reducing the trainable parameters drastically (approximately $\sim 98\%$). For fine-tuning the BERT model, we used labels that cannot be effectively captured by the Keyword-based (pattern matching) classifier and have a sufficient number of training examples. For labels with limited training data, we fallback to BART's zero-shot classification to ensure reliable predictions.

BART zero-shot classifier: BART proposed by [13] is sequence-to-sequence model developed for tasks like text generation, summarization, and translation. For zero-shot classification, the problem is framed as a Natural Language Inference (NLI) task, as suggested in [14]. In this approach, the news article is treated as the premise, and the candidate labels are treated as hypotheses (e.g., "This news article talks about Quantum technology"). The probability of entailment is then assigned to

Newsletter-Factory

Newsletter

EntitySearch

Profile

User Management

About

Logout

Selection Criteria:

Start Date: 01-Jan-2023

End Date: 05-Sep-2024

Newsletter: Quantum Computing

Filters:

NewsType

Domain

Technology

Geography

Total number of entries: 34

Give Feedback/Edit Labels

10 entries per page

Search:

Date	Sub-category	Title	Source/Deep Link
21-Jun-2024	Academia	Physicists propose time crystal-based circuit board to reduce quantum computing errors	Link to the article
17-Jun-2024	Industry-Healthcare/Life Science	Breakthrough: Chinese scientists develop world-leading quantum computing thermometer	Link to the article
16-Jun-2024	Industry-Healthcare/Life Science	Quantum computing set to transform medicine with faster drug discovery	Link to the article
13-Jun-2024	Security	How Quantum Computing and Machine Learning are Shaping Cybersecurity	Link to the article
07-Jun-2024	Industry-Healthcare/Life Science Security	Decoding the Impact of Quantum computing on Data Security: What CIOs Need to Know	Link to the article
07-Jun-2024	Security	IDEMIA Leads Consortium for Quantum-Safe Data Protection Solutions	Link to the article
28-May-2024	Academia	IBM, Quantum Computing Partners Research Sustainable Energy Opportunities	Link to the article

Download as PDF/HTML

Email to Subscribers

Figure 3: A sample Quantum Computing newsletter created using the Newsletter-Factory tool.

each candidate label. *bart-large-mnli* checkpoint (<https://huggingface.co/facebook/bart-large-mnli>) has demonstrated competitive zero-shot classification performance, often matching that of supervised models, and is widely employed for classification task.

2.4. Human-in-the-loop for Quality Check and Model improvement

The tool allows inspection of the generated draft version of newsletter by human experts. The expert can provide feedback to confirm correct predictions, offer corrections for wrong predictions (false-positive cases) and also add missing sub-category labels (false-negative cases). The data collated from the feedback is used to improve the model for future iterations. In the next version of the NLF tool, we plan to incorporate active learning mechanism to optimize the use of human supervision in this feedback loop.

3. Experimental Details

We experiment on a dataset of nine enterprise newsletters on diverse topics, namely (i) Banking Finance Insurance, (ii) Communications, Media, and Information Services, (iii) Education, (iv) Energy and Resource, (v) Healthcare, (vi) Life Science, (vii) Manufacturing, (viii) Quantum Computing, (ix) Travel and Logistics.

For our experimental setup we have used bert-base-uncased (<https://huggingface.co/google-bert/bert-base-uncased>) from the transformers [15] library. To incorporate adapter-based fine-tuning we have used the adapters (<https://github.com/adaptor-hub/adapters>) [16] library. We have compared BERT fine-tuning (full) with different adapter architectures [10, 11, 12]. The hyper-parameters batch size (8, 16), learning rates (1e-5, 2e-5, 5e-5) and epochs (5, 10, 20) were used. Fine-tuning adapters for longer epochs show comparable results to that of BERT all parameter fine-tuning.

Dataset distribution and the accuracy of the adapter-based BERT classifiers is shown in Table 1. Human experts were asked to give feedback to each of the nine newsletters generated by the Newsletter Factory tool. Table 1 shows that our tool is able to classify sub-categories in newsletters quite efficiently.

Table 1

Dataset Distribution (train, eval, and feedback samples) and Performance of BERT-FT (Fine-Tuned) and BERT-FT w/ Feedback in Sub-Category Classification (Accuracy) for each newsletter theme.

Newsletter	# Train	# Eval	# Feed-back	# Sub categ. Labels	# FT Labels	BERT-FT (%)	BERT-FT w/ Feedback (%)
Banking Finance Insurance	2225	2125	272	34	19	75.38	76.04
Communications, Media, and Information Services	869	869	110	8	6	60.75	62.37
Education	172	153	190	7	4	75.81	75.81
Energy and Resource	967	947	179	11	9	94.19	94.19
Healthcare	512	489	305	7	4	78.32	80.77
Life Science	461	441	169	7	3	64.39	64.62
Manufacturing	552	550	80	10	5	89.45	89.81
Quantum Computing	351	315	96	18	8	60.31	62.85
Travel and Logistics	135	110	164	5	4	91.81	91.81

Our models improve after incorporating feedback for most newsletters. However, user feedback may introduce biases or noise, limiting the learning and improvement of some models. Additionally, the fixed evaluation set does not reflect the evolving nature of business news, underscoring the need for a more adaptive evaluation approach.

The Newsletter-Factory tool prototype has been demonstrated and used on pilot basis by the newsletter creator team and domain experts. They have found it extremely useful in their actual day-to-day work of newsletter generation. Figure 3 displays an intuitive user interface of Newsletter-Factory showcasing a sample Quantum Computing newsletter. A user study was conducted to evaluate the ease of use and usability of various Newsletter-Factory features compared to traditional email or spreadsheet-based newsletter publication. The results show that 92.5% of users prefer Newsletter-Factory for searching, filtering, navigating, interactivity, and publishing newsletters. The study also highlights that Newsletter-Factory excels in creating customized, on-demand newsletters and browse through near-duplicate articles. Automated sub-category classification reduces the time and effort required from human experts. Additionally, the dynamic interface allows for interactive querying of data to generate newsletters, and the search feature helps users easily filter news articles based on their needs. User feedback played a vital role in refining the model’s accuracy, ultimately leading to more precise and relevant newsletters.

4. Related Work

Automated newsletter generation for enterprises has received relatively limited attention from the research community in spite of growing need and utility [1]. The work by Zhao [17] describes geography-targeted automated newsletter generation to focus on giving more importance to local news. User specified location filtering is used to increase the geography-specific relevance of news. The system developed by [18] leverages Spring Boot (Java) and RESTful APIs to dynamically generated email templates. Obando [19] focuses on energy sector newsletter generation by extracting key events from news based on user preference. Exploration in the direction of news robots and their perception by end-users was carried out by [20]. Though this is not directly focused on newsletter generation, it explores the broader theme of automated journalism. The relatively less work for newsletter generation and the lack of focused work for enterprise users makes Newsletter-Factory a practical and useful tool that applies relevant algorithms in information management and helps in automation of newsletter generation process in the enterprise setting.

5. Conclusion and Future Work

Creating periodic newsletters that deliver value to the target audience and achieve business objectives requires a combination of domain knowledge, business expertise, and technology awareness across various disciplines. This ends up being an effort-intensive and expensive proposition for enterprises. We presented Newsletter-Factory to create thematic and on-demand newsletters based on requirements from enterprise users. While creating periodic newsletters can be a valuable investment for enterprises in terms of engaging with their audience, driving brand awareness, and generating business leads, it requires allocation of knowledgeable resources. This entails significant effort and cost. Newsletter-Factory tool helps to automate the information processing pipeline to create newsletters. The Newsletter-Factory tool prototype has been demonstrated and used on pilot basis by the newsletter creator team and domain experts. They have found it extremely useful in their actual day-to-day work of newsletter generation. Further, qualitative feedback from actual business executives who consume the newsletters' content has been found to be very positive. As part of future work, we plan to incorporate active learning mechanism [21] to optimize the use of human supervision in this feedback loop. Creating more nuanced, application focused (e.g., industrial safety [22], legal issues [23]) newsletters as well as enabling temporal QA [24] on newsletters are possible areas of extension. We also plan to incorporate multilingual support such as Hindi [25] and other languages in Newsletter-Factory tool.

References

- [1] A. Jack, Editorial email newsletters: the medium is not the only message, Editorial email newsletters: The medium is not the only message (2016).
- [2] S. Rodier, D. Carter, Online near-duplicate detection of news articles, in: Proc. of the Twelfth LREC, 2020, pp. 1242–1249.
- [3] A. Z. Broder, Identifying and filtering near-duplicate documents, in: Annual symposium on combinatorial pattern matching, Springer, 2000, pp. 1–10.
- [4] A. Kumar, S. Tumre, S. Patil, Benchmarking near-duplicate detection in the era of pay-walled news, in: Companion Proceedings of the ACM Web Conference 2025, WWW '25, Association for Computing Machinery, 2025. doi:10.1145/3701716.3715303.
- [5] S. Tumre, S. Patil, A. Kumar, Improved near-duplicate detection for aggregated and paywalled news-feeds, in: 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'25), 2025.
- [6] S. Patil, Domain-specific noisy query correction using linguistic network community detection, in: Companion Proceedings of the Web Conference (WWW'20), 2020, pp. 126–127.
- [7] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, Text mining: applications and theory (2010) 1–20.
- [8] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, Information Sciences 509 (2020) 257–289.
- [9] A. Bougouin, F. Boudin, B. Daille, Keyphrase annotation with graph co-ranking, 2016. URL: <https://arxiv.org/abs/1611.02007>. arXiv:1611.02007.
- [10] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, I. Gurevych, AdapterFusion: Non-destructive task composition for transfer learning, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, ACL, Online, 2021.
- [11] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, CoRR abs/1902.00751 (2019).
- [12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021).
- [13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation,

- and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2020.
- [14] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proc. of the 2019 Conference on EMNLP and the 9th IJCNLP, ACL, Hong Kong, China, 2019.
 - [15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proc. of the 2020 Conference on EMNLP: System Demonstrations, ACL, Online, 2020.
 - [16] C. Poth, H. Sterz, I. Paul, S. Purkayastha, L. Engländer, T. Imhof, I. Vulić, S. Ruder, I. Gurevych, J. Pfeiffer, Adapters: A unified library for parameter-efficient and modular transfer learning, in: Y. Feng, E. Lefever (Eds.), Proc. of the 2023 Conference on EMNLP: System Demonstrations, ACL, Singapore, 2023.
 - [17] L. Zhao, Automated Local Newsletter Generator, Ph.D. thesis, University of Bristol, 2009.
 - [18] M. Animasaun, Real time operation of newsletter generation (2018).
 - [19] S. A. OBANDO MAYORAL, Automated document tagging and newsletter generation using natural language processing and machine learning (2018).
 - [20] C. Oh, J. Choi, S. Lee, S. Park, D. Kim, J. Song, D. Kim, J. Lee, B. Suh, Understanding user perception of automated news generation system, in: Proc. of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–13.
 - [21] S. Patil, Active learning based weak supervision for textual survey response classification, in: Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14–20, 2015, Proceedings, Part II 16, Springer, 2015, pp. 309–320.
 - [22] S. Patil, S. Koundanya, S. Kumbhar, A. Kumar, Improving industrial safety by auto-generating case-specific preventive recommendations, in: Third Workshop on NLP for Positive Impact, co-located with EMNLP’24, 2024.
 - [23] A. Gupta, D. Verma, S. Pawar, S. Patil, S. Hingmire, G. K. Palshikar, P. Bhattacharyya, Identifying participant mentions and resolving their coreferences in legal court judgements, in: Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11–14, 2018, Proceedings 21, Springer, 2018, pp. 153–162.
 - [24] H. Bedi, S. Patil, G. Palshikar, Temporal question generation from history text, in: Proceedings of the 18th international conference on natural language processing (ICON), 2021, pp. 408–413.
 - [25] S. Hingmire, N. Ramrakhiyani, A. K. Singh, S. Patil, G. Palshikar, P. Bhattacharyya, V. Varma, Extracting message sequence charts from Hindi narrative text, in: Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events, co-located with ACL’20, 2020, pp. 87–96.