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

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

TCS Research, Tata Consultancy Services Ltd., Pune, India.

Text2Story

Introduction

- A newsletter is a periodic publication, usually distributed via email, that contains curated content on a specific topic.
- Newsletters are a powerful tool for building a compelling story over time.
- Gaining insights into how business narratives unfold is crucial for strategic decision-making, risk assessment, and market analysis.
- The demand for industry or technology specific newsletters that tailor to the interests of enterprises is increasing as they have to stay up-to-date with rapidly evolving industry trends.

High Level Block Diagram

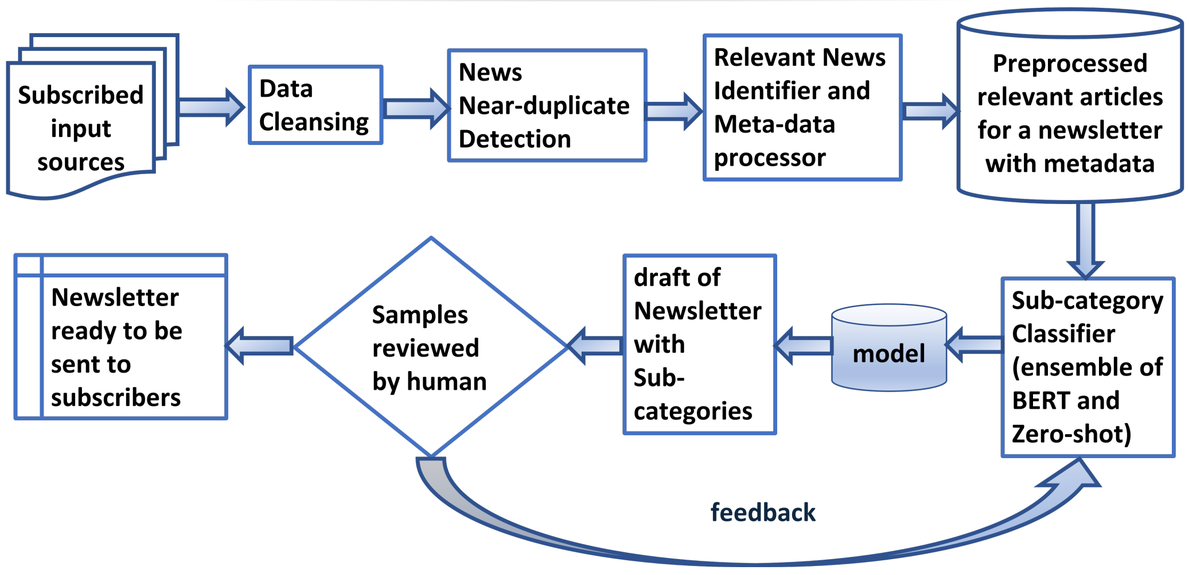
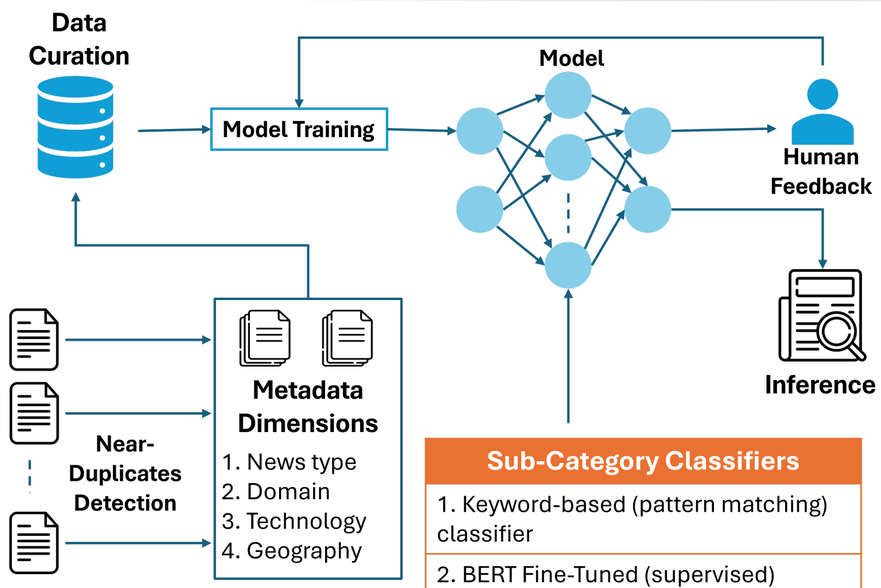


Figure: (1) High-level block diagram of key components in Newsletter-Factory

Overview of the Newsletter-Factory Tool



Sub-Category Classifiers

1. Keyword-based (pattern matching) classifier
2. BERT Fine-Tuned (supervised) Classifier
3. BART Zero-Shot Classifier

News Article

QNu Labs Secures $6.5 Million In Funding. QNu Labs closed a $6.5 million funding Pre-Series A1.The round was led by Ashish Kacholia of Lucky Investments, along with Speciale Invest and a family fund. The startup offers quantum and post-quantum cryptography solutions based on patented technology.

News type	Financial Reporting
Domain	Multidomain Applications of IT, Banking Finance Insurance
Technology	Quantum Computing
Geography	Others

Predictions for Sub-Categories

i. Finance & Reporting
ii. Industry-BFSI
iii. Capital Markets & Investment Banking
iv. Startups Funding Acceleration
v. Security,
vi. Technology Development
vii. Securitisation

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

Newsletter-Factory categorizes news articles into high-level themes and also generates fine-grained sub-labels enabling comprehensive understanding of emerging trends and key narratives.

Near-duplicate detection and handling

- Storage efficiency is critical for a newsletter generation system as it has to deal with extensive content archives.
- We utilize standard Locality Sensitive Hashing (LSH) based MinHash-LSH and transformers based sentence embeddings techniques to cluster near-duplicates.

Metadata (Newsletter theme) classification

- NLF captures the high-level theme of a news article across various dimensions such as domain, technology, newstype and geography. (e.g. refer Figure 2)
- An ensemble of pattern-based categorizers and machine learning classifiers are implemented to tag each news article with appropriate metadata to map them to relevant newsletter.

Subcategory Classification for Each News Article

- Sub-categories further help to filter and focus on a subset from the relevant articles. It is an ensemble of different classifiers.
- A **high-precision, keyword-based** pattern matching classifier.
  - The news article is searched for relevant keywords and phrases.
  - A score is assigned for a 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.).
- Efficient **Adapter Fine-tuned BERT** classifier.
  - This component utilizes a supervised learning paradigm, making it beneficial when training data is available.
  - We utilize parameter efficient fine-tuning by adding extra trainable parameters into the BERT architecture. This significantly reduces the compute and storage required.
- **BART-based zero-shot** classifier.
  - For zero-shot classification, the problem is framed as a Natural Language Inference (NLI) task where the news article is treated as premise and candidate labels as hypotheses.
  - *bart-large-mnli*, a widely used model for classification task is employed as zero-shot classifier.

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

- NLF allows inspection of the generated draft version of newsletter by human experts.
- The expert can provide feedback for correct predictions, wrong predictions and also add missing sub-category labels.
- This feedback is used to improve the model for future iterations.

Experimental Details

Table: 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	# Feedback	# 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

Contact: {firstname.lastname}@tcs.com