Automating Financial Journalism: Natural Language Generation via Transformers

InHwa Mo

Simon Fraser University / Burnaby, BC, Canada

`imo@sfu.ca`

Abstract

My project aims to deliver timely, accurate, and data-driven financial news through text generation. Utilizing transformer models, it undertakes a data-to-text natural language generation (NLG) task. The system converts real-time financial data into concise, plain English articles, facilitating financial insights and decisions.

1 Problem Statement and Motivation

The complexity of financial data often poses a barrier to understanding financial and macroeconomic trends and making informed decisions, especially for those without the requisite education and those less inclined to interpret charts and graphs. My goal is to produce content that, while potentially less engaging or professionally written compared to human-written news, provides simple, objective updates into financial trends and events, making financial information accessible to a broader audience.

2 Related Work

Much of journalism is already automated. Bloomberg News generates a third of its content with Cyborg, an in-house automation system that takes financial reports and churns out news articles (Sharma et al., 2023). Other areas in which data-to-text NLG has been applied include weather forecasts (Turner et al, 2006), and writing biographies (Lebret et al., 2016). NLP is used widely in finance, however the application of NLP techniques to financial journalism is relatively unexplored and closed.

3 Approach

I plan to use a pre-trained transformer model, a BERT-based model, to generate news articles from financial data. The model will be fine-tuned on a dataset of financial news articles.

4 Experimental setup

Unfortunately of the publically available NLG benchmark datasets, none are financial in nature. I will have to access real-time and historical data from APIs from government and financial institutions such as with the Bank of Canada's Valet API and the US Treasury's FiscalData API. The model will be fine-tuned using PyTorch and the Hugging Face Transformers library. For evaluating text generated output, there are many options including BLEU, but also measures such as Informational Adequacy (Does the text match the information contained in the data?), which seems to be most suitable among a set of proposed NLG benchmarks by Perez-Beltrachini et al. (2017)

5 Timeline and work breakdown

By the project milestone I aim to have a functional prototype that is connected to real-time data, and to have fine-tuned a pre-trained transformer model, and have a preliminary evaluation of the model's output.

- Week 1: Set up data fetching and preprocessing from financial APIs. Choose and prepare the pre-trained transformer model.

- Week 2: Start fine-tuning the model on a preliminary financial news dataset. Generate initial news articles. Conduct a basic evaluation using automated metrics

- Week 3: Refine model based on initial results. Document progress and setup.

6 References

1. Sharma, Gogineni and Ramakrishnan. (2023). Innovations in Neural Data-to-text Generation: A Survey. arXiv preprint arXiv:2207.12571v2

2. Ross Turner, Somayajulu Sripada, Ehud Reiter, and Ian P Davy. 2006. Generating Spatio-

Temporal Descriptions in Pollen Forecasts. In Demonstrations, pages 163–166.

3. Lebret, Rémi and others. (2016). Neural Text Generation from Structured Data with Application to the Biography Domain. arXiv preprint arXiv:1603.07771v3.

4. Laura Perez-Beltrachini and Claire Gardent. 2017. Analysing Data-To-Text Generation Benchmarks. In Proceedings of the 10th International Conference on Natural Language Generation, pages 238–242, Santiago de Compostela, Spain. Association for Computational Linguistics.