Summary of *Uncovering Information: Can AI Tell Us Where to Look?*

Anna Costello, Bradford Levy, Valeri Nikolaev (2024)

2025.04.18 LiZiming

1. **What are the research questions?**
   - How to measure surprise in narrative disclosures?
   - Where new information is most prevalent?
   - When surprise is released?
   - Can explain market reaction?
2. **Why are the research questions interesting?**
   - Stakeholders need quickly and easily to find relevant information, but information buried in sea of unstructured narrative disclosure released by firm.
   - Prior studies mostly focus on quantitative information, but narrative information is highly multidimensional and challenging to find surprise.
3. **What is the paper's contribution?**
   - Contribute to literature using output of LLMs to predict returns.
     - Prior literature: LLM transform a source document into another representation (Lopez-Lira and Tang, 2023; Kim et al., 2023; Chen et al., 2022).
     - Extend: preserve source and overlay measure of information eliminating output hallucination and lookahead bias in predictability.
   - Contribute to literature on costly information processing.
     - Prior literature: measure informativeness as similarity of text across annual and quarterly reports (Cohen et al., 2020).
     - Extend: develop a method for forming priors and identifying new information.
4. **What hypotheses are tested in the paper?**
   - H1: The message is more informative, if a highly unlikely event occurs.
   - H2: If prices impound all publicly available information, and measure captures this information, then it can explain short window price reactions.

   a) Do these hypotheses follow from and answer the research questions?
   - Yes.

   b) Do these hypotheses follow from theory? Explain logic of the hypotheses.
   - It follows the information theory and Shannon's model of communication.
5. **Sample: comment on the appropriateness of the sample selection procedures.**
   - Extracting narrative disclosure from BeanCounter corpus, the plain text versions of all filings accepted by SEC EDGAR system from 1996 to 2023.
6. **Comment on the appropriateness of variable definition and measurement.**
   - Information of the i-th token is causal in the sense that the probability of a given token depends only the context on the left side of the token.
7. **Comment on the appropriateness of the regress/predict model specification.**

- Use cross-sectional pre-trained model and Firm and time-specific models.

8. **What difficulties arise in drawing inferences from the empirical work?**
   - How to improve the interpretability of large language models.
9. **Describe at least one publishable and feasible extension of this research.**
   - Can LLMs measure new information in data and tables?
   - Has information in filings been released through other channels?