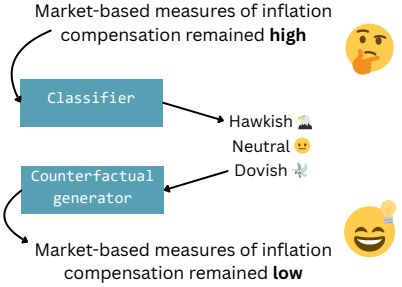
Natural Language Counterfactual Explanations in Financial Text Classification: A Comparison of Generators and Evaluation Metrics

Karol Dobiczek (karol.dobiczek@gmail.com)**, Patrick Altmeyer** (p.altmeyer@tudelft.nl)**, Cynthia C. S. Liem** (c.c.s.liem@tudelft.nl)
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands

Motivation

Large language models are increasingly used in specialized fields such as analyzing **monetary policy communications**, which allow central banks to **communicate their policy stance to markets clearly**. Thus is a difficult task considering the highly nuanced nature of these texts.

Counterfactual explanations (CE) help explain machine learning model classifications by perturbing the original input to yield desired predictions. Various CE methods exist for LLM classifiers, but most are **trained and evaluated on generic tasks** and datasets. In addition, their evaluations often rely on **imprecise metrics**.



We use three generators representing distinct techniques:

1. **LLM-assisted generation**: *Polyjuice* (Wu et al., 2021) uses another LLM as a surrogate model to produce CEs.
2. **Latent perturbation and decoding**: *PPLM* (Dathathri et al., 2019) uses the latent representation of the factual sentence and perturbs it to generate a CE embedding that gets decoded back to text.
3. **Sequential generation**: *RELITC* (Betti et al., 2023) masks a part of the input text and then fills it with new tokens.

Experiments

Data:

- Federal Open Market Committee transcripts split into 3 classes: *dovish*, *hawkish*, *neutral*.

Metrics:

- Quantitative: ex. perplexity, implausibility, edit distances (semantic, embedding), flip rate.
- Qualitative: fluency and plausibility.

Human evaluations:

- Large-scale, crowdsourced focusing on fluency.
- Small-scale, involving field experts, focusing on plausibility and fluency. We ask experts to use the following, more robust definitions:

*A **fluent** segment is one that is grammatically well-formed; contains correct spellings; adheres to the common use of terms, titles and names; contains properly capitalized letters; and is intuitively acceptable. Unfinished sentences also impact the fluency of a segment.*

*A **plausible** counterfactual segment adheres well to samples seen in the real data distribution, and the target sentiment of the target class. The changes made to the factual, considering the meaning and context of the edited words, should also fit the target domain.*

Results

Main findings:

- Methods that apply minimal changes (RELITC) create counterfactuals that are more fluent than those that focus solely on CE validity.
- Domain experts comment that incorrect use of domain-specific terms can **diminish the plausibility** of the explanations.
- Using CE generators that do not use specialist words (Polyjuice) might be preferable in specialist domains, suggesting that faithfulness can be as important as plausibility.
- Most **quantitative metrics do not reflect the generators' desiderata well**.

Main **recommendations for future work**:

- Evaluating CE methods on specialist texts.
- Designing new methods with such texts in mind.
- Involving end users in evaluating CE generators.

Generator	Perplexity ↓	Perpl. ratio	Edit dist. ↓	Tree dist. ↓	Emb. dist. ↓	Implausib. ↓	Faithful. ↑	Succ. rate ↑
Polyjuice	90.98 (172.1)	1.80 (4.6)	0.31 (0.3)	19.67 (24.0)	**20.32** (3.7)	33.64 (4.6)	0.18 (0.4)	0.34 (0.5)
PPLM	**36.97** (16.9)	**0.78** (0.5)	0.69 (0.5)	36.94 (10.3)	20.88 (3.7)	**32.18** (4.0)	0.34 (0.6)	0.51 (0.5)
RELITC	100.94 (125.2)	1.67 (1.2)	**0.14** (0.1)	**10.72** (12.2)	21.96 (3.9)	33.30 (3.9)	**0.54** (0.6)	**0.74** (0.4)

Table 1. Averages and standard deviations of the quantitative metrics calculated for counterfactual explanations of texts in the test set. A perfect result for perplexity ratio is thought to be 1.

Table 2. Results of the human annotation of the counterfactuals using the qualitative metrics. Each counterfactual receives five ratings. Experts evaluated a subset of five samples, we show the fluency scores the non-experts give on the same set of samples.

	Annotators			
	Non-expert	N-exp. 5 CE	Expert	
Generator	Fluency	Fluency	Fluency	Plausibility
Polyjuice	3.40 (0.9)	3.44 (0.7)	3.45 (0.9)	**2.45** (0.7)
PPLM	2.86 (0.7)	2.48 (0.5)	2.26 (0.5)	1.83 (0.3)
RELITC	**3.43** (0.8)	**3.96** (0.5)	**3.90** (0.6)	2.12 (0.3)