**Machine Learning via Financial Word Embedding**

Eghbal Rahimikia1,2, Stefan Zohren3, Ser-Huang Poon1

1 Alliance Manchester Business School; University of Manchester.
2 Oxford-Man Institute of Quantitative Finance; University of Oxford.

**Goal**

Transforming financial textual data to numerical data.

### Why is this challenging?

- Textual data is a high dimensional data.
- Computational feasibility is still a big challenge.

**One Hot Encoding**

Each word is defined by vector of zeros and ones.

**Dictionary**

Group of words with the same sentiment.

**Word embedding**

Each word is defined by a vector with size N from a corpus of text.

**Transforming more information**

Computationally more intensive.

**A word embedding example**

- Data source: Dow Jones Newswires Text News Feed.
- Duration: January 1, 2000, to September 14, 2015.
- Type: All news (viz. financial, political, weather, etc.)
- Pre-processing: Eliminating redundant characters, sentences, and structures.
- Dimension: 300
- Final number of words (tokens): 2,733,035

**FinText: a financial word embedding**

- Developed and trained on The Computational Shared Facility (CSF3), University of Manchester.
- Possible to use it as a stand-alone model or inside of other machine learning models.

**Conclusions**

- *FinText reached the highest portfolio performance with the highest Sharpe ratio.*
- This performance is higher than GPT-3 model. GPT-3 is the most advanced pay-to-use natural language processing model.
- *Focusing on realised volatility forecasting, our results show a statistically significant improvement in forecasting performance for high volatility days.*

---

**Machine Learning via Financial Word Embedding**

**General-purpose word embeddings**

- **GPT**: negative ▼
- **Facebook**: negative ▼
- **FinText**: positive ▼

**Analogy**

- debit:credit :: positive:NULL
- bullish:bearish :: rise:NULL
- apple:iphone :: microsoft:X
- windows:XP :: iphone:NULL
- msft:aapl :: amazon:NULL
- bid:ask :: buy:NULL
- creditor:lend :: debtor:NULL
- rent:short_term :: lease:NULL
- growth_stock:overvalued :: value_stock:NULL
- call_option:put_option :: buy:NULL

**Word embedding**

Google

Facebook

FinText

**Debit to credit is like positive to ?**

**FinText is a winner for financial analogies!**

**If a word embedding works well in finance, it must be able to cluster similar companies.**

### Conclusions

- *FinText reached the highest portfolio performance with the highest Sharpe ratio.*
- This performance is higher than GPT-3 model. GPT-3 is the most advanced pay-to-use natural language processing model.
- *Focusing on realised volatility forecasting, our results show a statistically significant improvement in forecasting performance for high volatility days.*