Machine Learning via Financial Word Embedding

Eghbal Rahimikia[1,2], Stefan Zohren[2], Ser-Huang Poon[1]

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

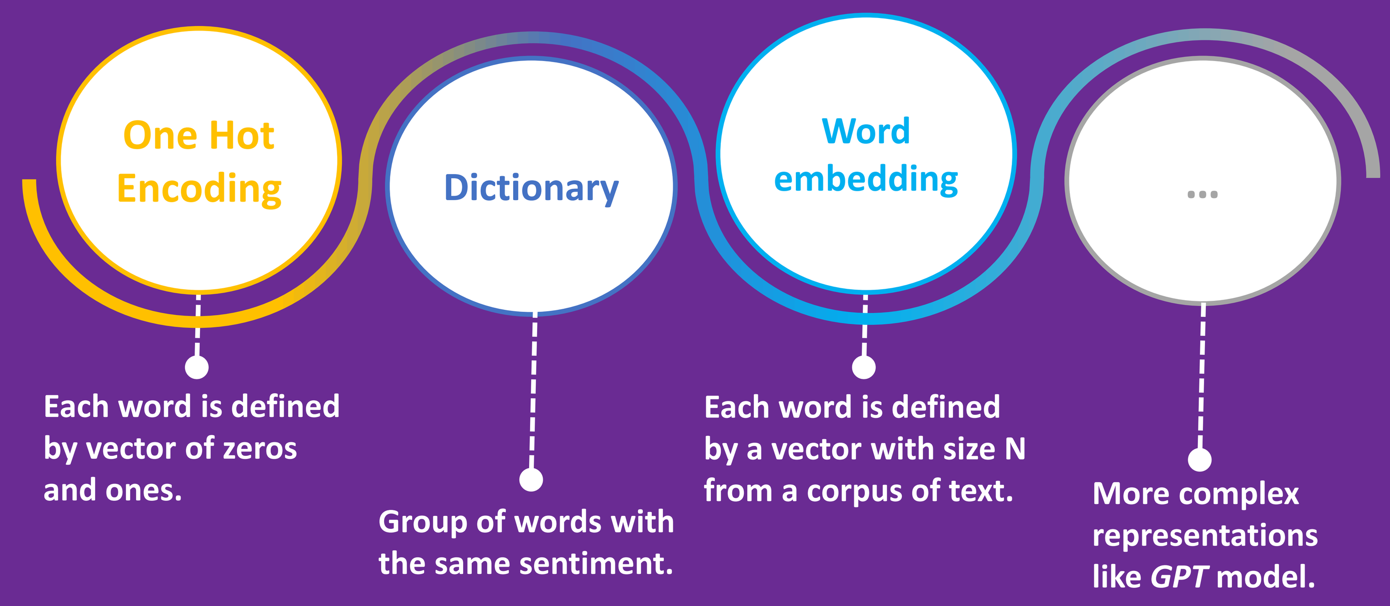
START FROM HERE

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.**



***Word2Vec*, *Negative sampling*, and *GloVe* are among the main algorithms for training this.**

A word embedding example

Dimension	King	Queen	Prince	Man	Woman	Child
Dimension 1 (Royalty)	0.99	0.99	0.95	0.01	0.02	0.01
Dimension 2 (Masculinity)	0.94	0.06	0.02	0.99	0.02	0.49
Dimension 3 (Age)	0.73	0.81	0.15	0.61	0.68	0.09

These values are found using a corpus (a large set of texts).

* For defining each word, 300 dimensions is common in literature.

FinText: a financial word embedding

- **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**

- **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.**

GO UP!

FinText vs. *Google* vs. *Facebook*

Debit to credit is like positive to ?

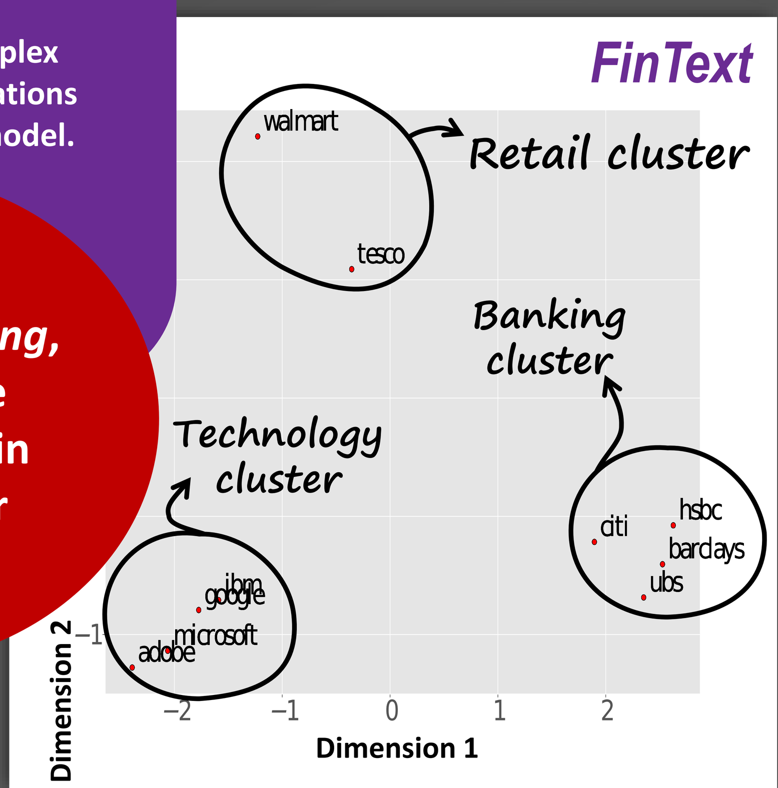
FinText is a winner for financial analogies!

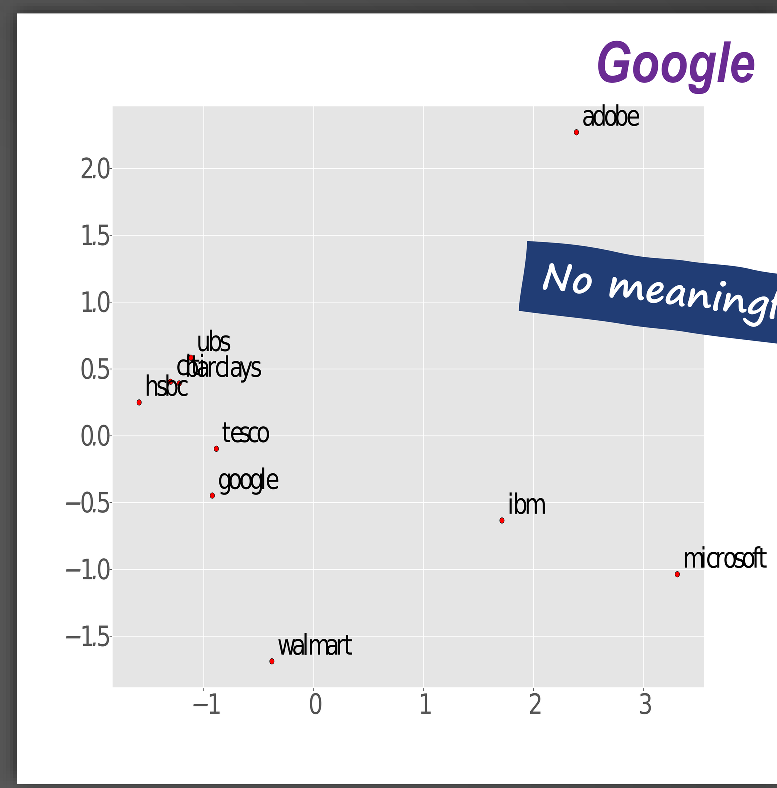
General-purpose word embeddings

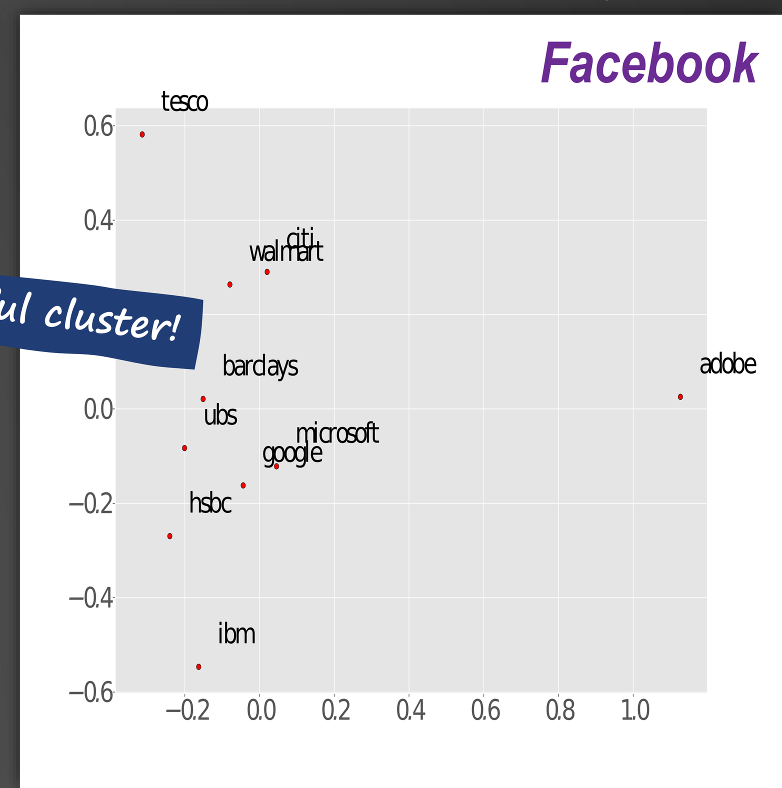
Analogy	Word embedding: Google	Word embedding: Facebook	Word embedding: FinText
debit:credit :: positive:X	positive	negative ✓	negative ✓
bullish:bearish :: rise:X	rises	rises	fall ✓
apple:iphone :: microsoft:X	windows_xp	iphone	windows ✓
us:uk :: djia:X	NONE	NONE	ftse_100 ✓
microsoft:msft :: amazon:X	aapl	hmv	amzn ✓
bid:ask :: buy:X	tell	ask-	sell ✓
creditor:lend :: debtor:X	lends	lends	borrow ✓
rent:short_term :: lease:X	NONE	NONE	long_term ✓
growth_stock:overvalued :: value_stock:X	NONE	NONE	undervalued ✓
us:uk :: nyse:X	nasdaq	hsbc	lse ✓
call_option:put_option :: buy:X	NONE	NONE	sell ✓

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

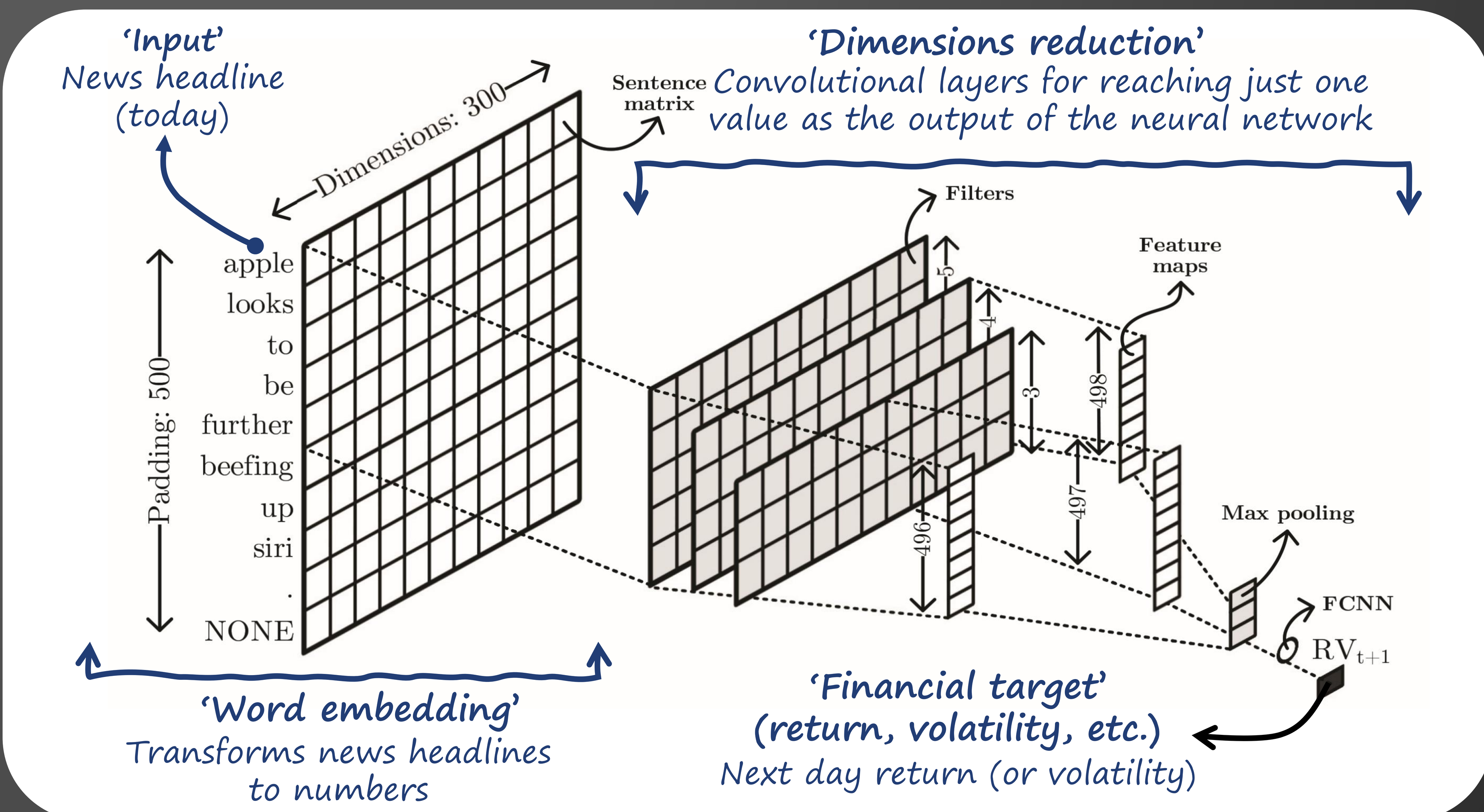
Another challenge...







We can use word embedding inside of a Convolutional Neural Network (CNN) for financial forecasting!



Trading: zero-investment portfolio performance (2005-2018)

Model	Leg	E(R)	STD(R)	Sharpe	DD(R)	Sortino	% of + Ret	Ave P./Ave L.
LM dictionary	L	0.084	0.079	1.064	0.06	1.391	0.548	1.211
	S	-0.035	0.108	-0.324	0.06	-0.587	0.452	0.824
	L-S	0.028	0.046	0.601	0.025	1.13	0.512	1.049
GPT-3	L	0.136	0.082	1.668	0.063	2.173	0.579	1.378
	S	-0.062	0.079	-0.791	0.049	-1.28	0.442	0.793
	L-S	0.039	0.032	1.229	0.02	1.988	0.526	1.108
GPT-J	L	0.094	0.077	1.224	0.057	1.643	0.542	1.181
	S	0.02	0.157	0.129	0.06	0.334	0.439	0.781
	L-S	0.064	0.075	0.859	0.024	2.692	0.516	1.065
$\text{FinText}_{W2V}(100)$	L	0.126	0.083	1.531	0.063	2	0.574	1.345
	S	-0.029	0.094	-0.313	0.051	-0.579	0.448	0.813
	L-S	0.052	0.034	**1.522**	0.019	2.716	0.532	1.138
$\text{FinText}_{FT}(60)$	L	0.098	0.079	1.245	0.059	1.668	0.552	1.235
	S	-0.014	0.087	-0.159	0.052	-0.264	0.444	0.798
	L-S	0.045	0.031	1.479	0.018	2.52	0.526	1.109
$\text{Google}_{W2V}(20)$	L	0.113	0.081	1.401	0.06	1.898	0.569	1.318
	S	-0.021	0.089	-0.236	0.051	-0.412	0.445	0.801
	L-S	0.049	0.032	1.519	0.018	2.736	0.53	1.127
$\text{Facebook}_{FT}(60)$	L	0.079	0.086	0.919	0.069	1.154	0.551	1.227
	S	-0.008	0.081	-0.094	0.045	-0.168	0.454	0.832
	L-S	0.039	0.032	1.21	0.022	1.802	0.519	1.08

* W2V: *Word2Vec*; FT: *FastText*; (X): number of filters in CNN (complexity of model).

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.**

***FinText* is publicly available**

SCAN ME

Rahimikia, E., Zohren, S., & Poon, S. H. (2021). Realised Volatility Forecasting: Machine Learning via Financial Word Embedding. *Available at SSRN 3895272*.