Stock Movement Prediction from Tweets and Historical Prices

ACL 2018 - Yumo Xu, Shay B. Cohen

Alumno: Juan Antonio López Rivera

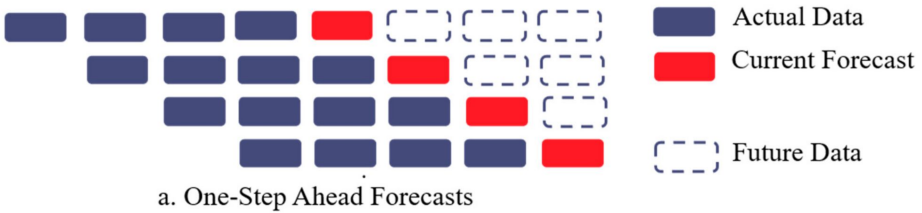
StockNet

Objetivo

Predecir el movimiento de cierta acción en el día $\mathcal{D}$

Usando datos históricos de Twitter y precio de la acción entre los días

$$[\mathcal{D} - \Delta\mathcal{D}, \mathcal{D} - 1]$$



a. One-Step Ahead Forecasts

StockNet

Datos de entrada

Entre las fechas 01/01/2014 a 01/01/2016 se recolectaron datos de 88 compañías.

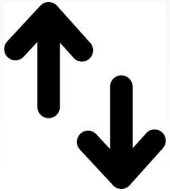
Para cada compañía se tiene:

- Tweets que mencionan esa acción
- Precio histórico

Twitter API

Salida

- 0: Bajará la acción
- 1: Subirá la acción



0

Recolección de datos

Historical Price Dataset

Twitter Dataset

Recolección de datos

- 88 compañías de alto volumen de movimientos
  - Top 10 con mayor capital en cada categoría
- El movimiento debe ser significativo, ie. ≤-0.5% o bien >0.55%
  - Descarta 38.72% de los datos, quedando 26,614
  - Balancea las clases: 49.78% y 50.22%

Subconjuntos de datos

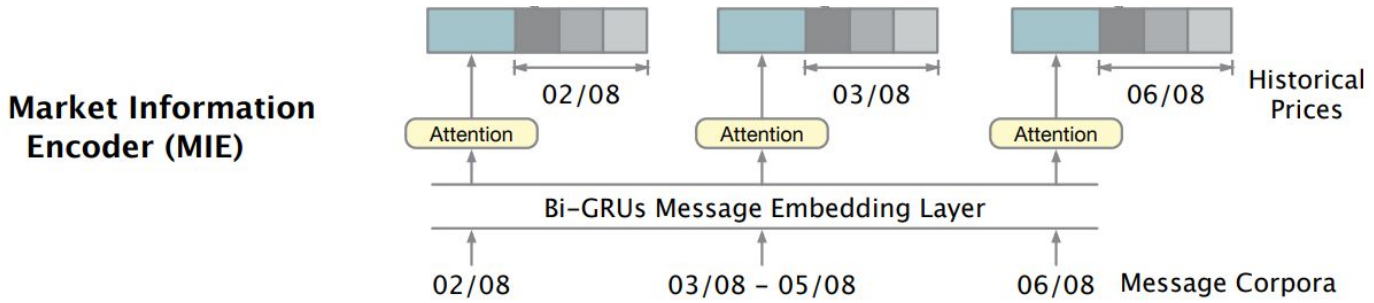
- **Entrenamiento**: 20,399 movimientos entre 01/01/2014 y 01/08/2015
- **Validación**: 2,555 movimientos entre 01/08/2015 y 01/10/2015
- **Prueba**: 3,720 movimientos entre 01/10/2015 y 01/01/2016

1

Codificación de datos

Market Information Encoder (MIE)

- Codifica información de redes sociales y precio histórico para formar la **entrada X**
- Cada día corresponde a una entrada **xt = [ct , pt]** donde
  - **ct**: embedding del corpus
  - **pt**: vector de precio histórico



Market Information Encoder (MIE)

- El vector de precio histórico pt se conforma de 3 valores:
  - Precio al cierre del día
  - Precio más alto del día
  - Precio más bajo del día

$$\tilde{p}_t = \left[\tilde{p}_t^c, \tilde{p}_t^h, \tilde{p}_t^l\right]$$

- Finalmente se normalizan usando el cierre del día anterior:

$$p_t = \tilde{p}_t / \tilde{p}_{t-1}^c - 1$$

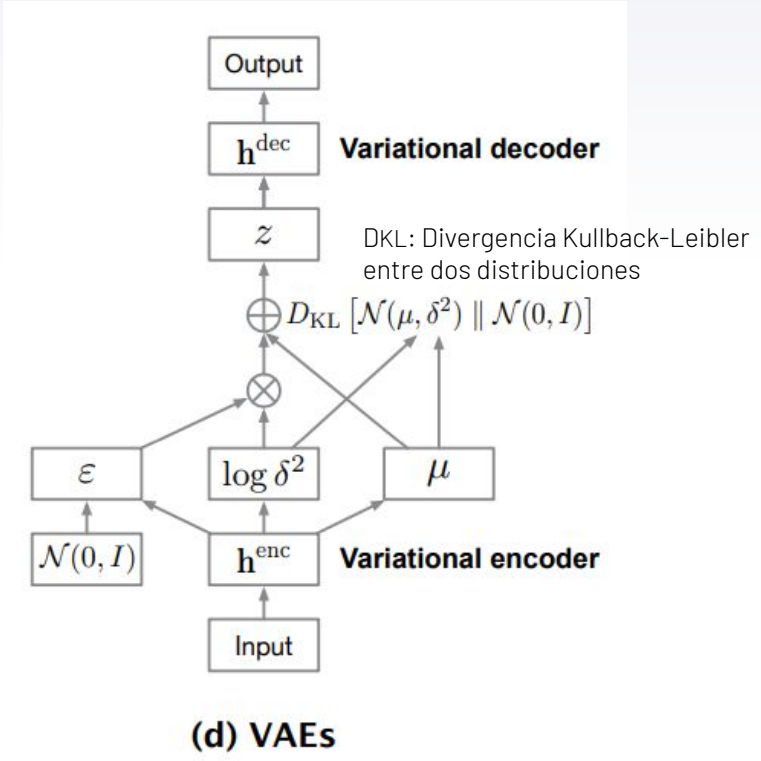
Variational Movement Decoder (VMD)

- El valor **Z** se conoce como ***latent driven factor***, se aproxima así (cuando no se conoce ***y***) en cada instante de tiempo *zt*:

$$z_t = \mu_t + \delta_t \odot \epsilon$$

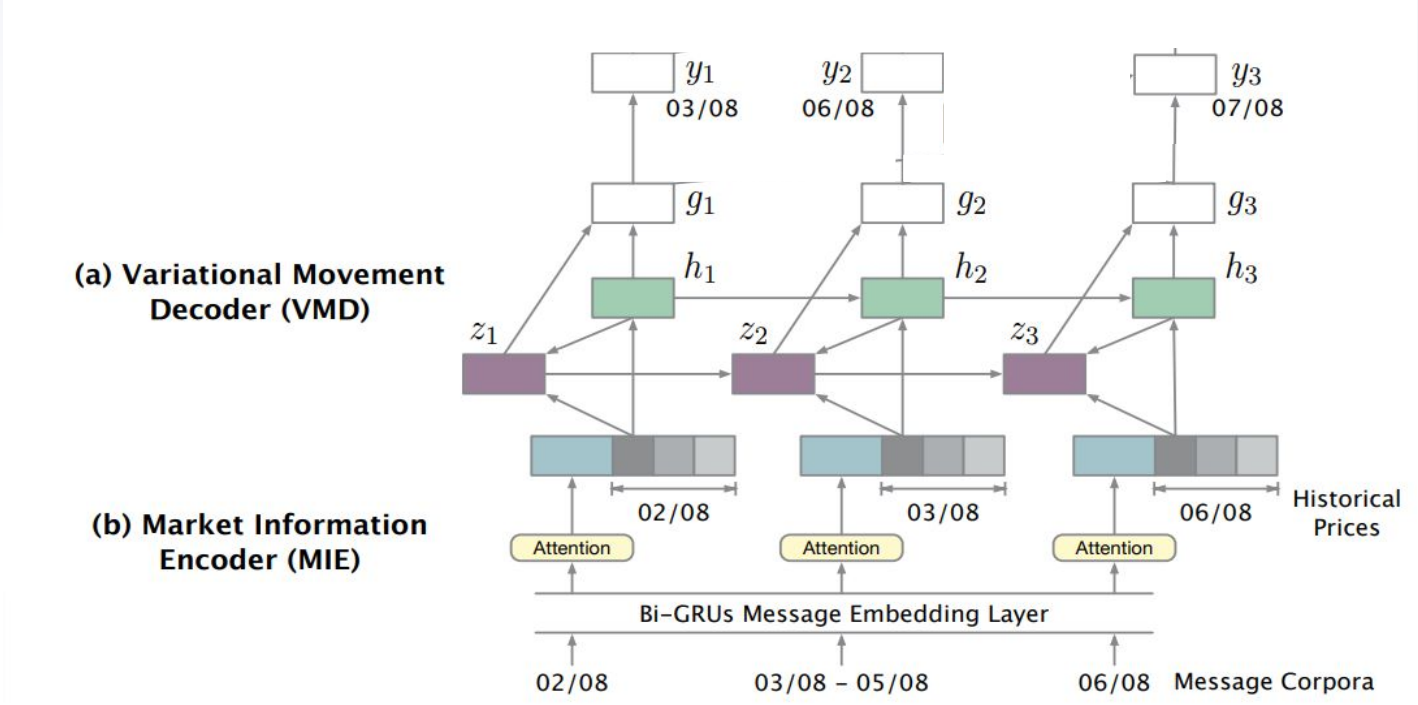
$$\mu'_t = W^{\theta}_{o,\mu} h^{z'}_t + b^{\theta}_{\mu}$$

$$\log \delta'^2_t = W^{\theta}_{o,\delta} h^{z'}_t + b^{\theta}_{\delta}$$



(d) VAEs

Variational Movement Decoder (VMD)

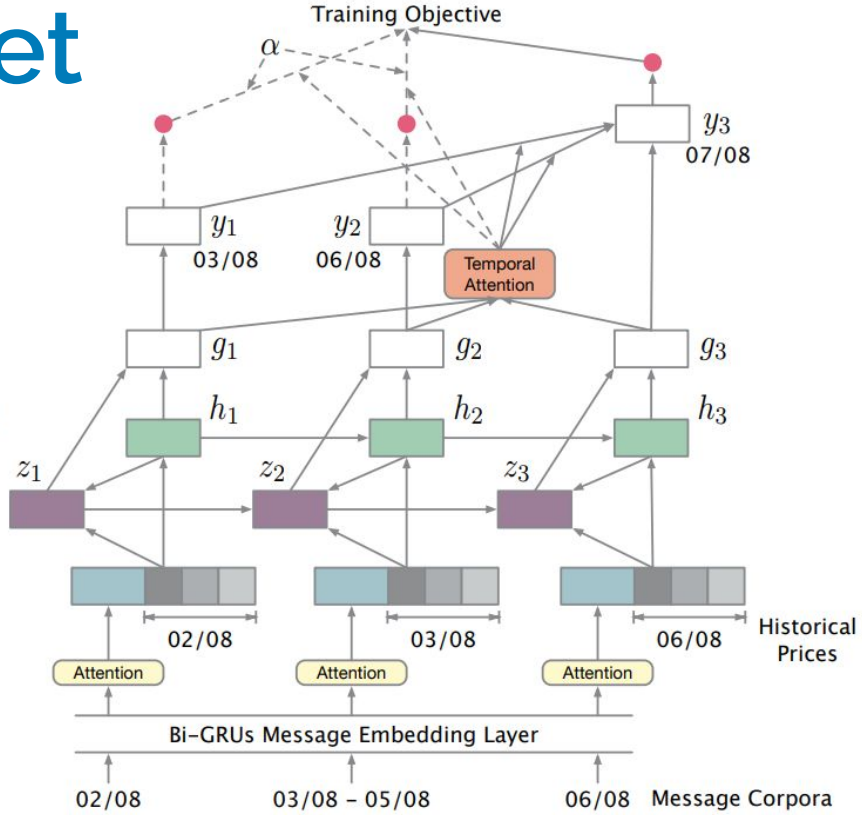


StockNet

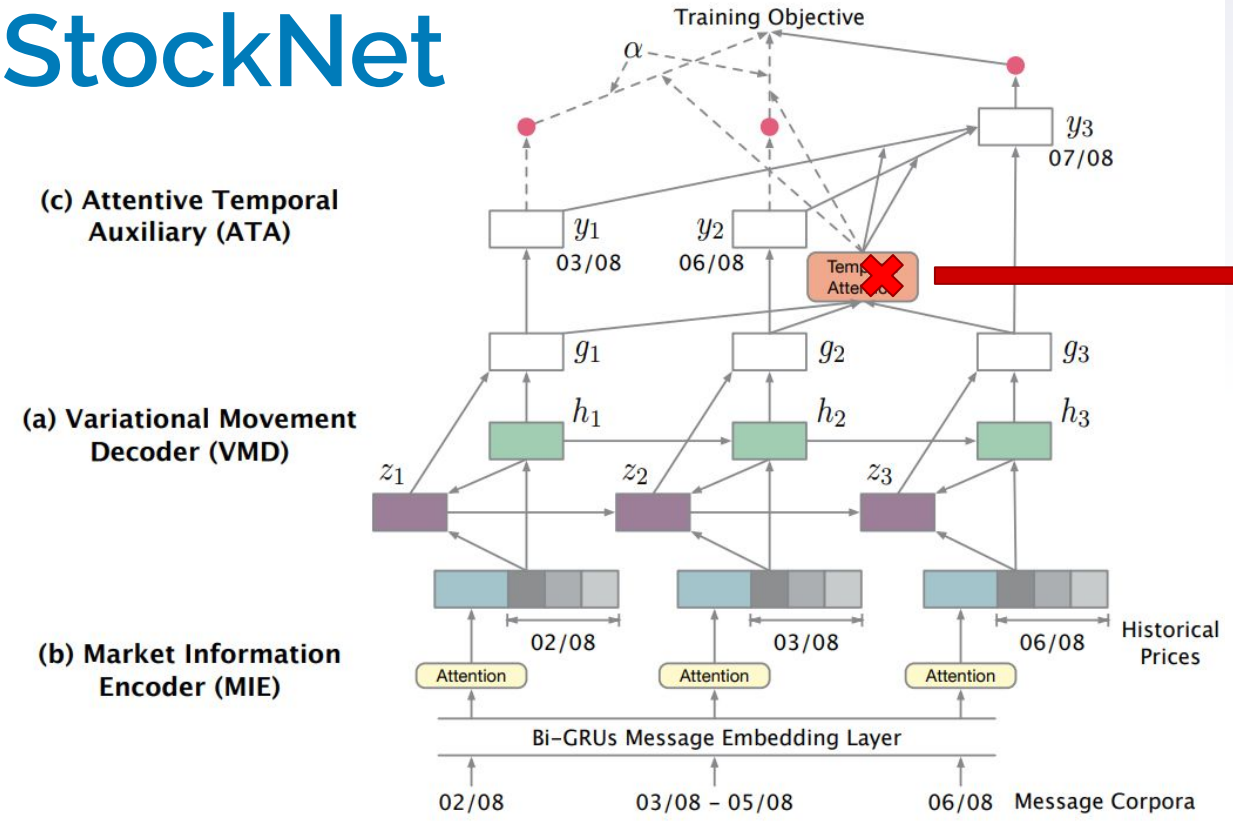
(c) Attentive Temporal Auxiliary (ATA)

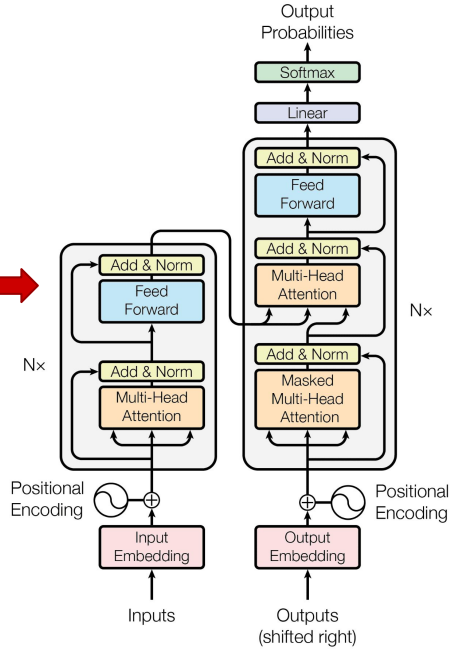
(a) Variational Movement Decoder (VMD)

(b) Market Information Encoder (MIE)



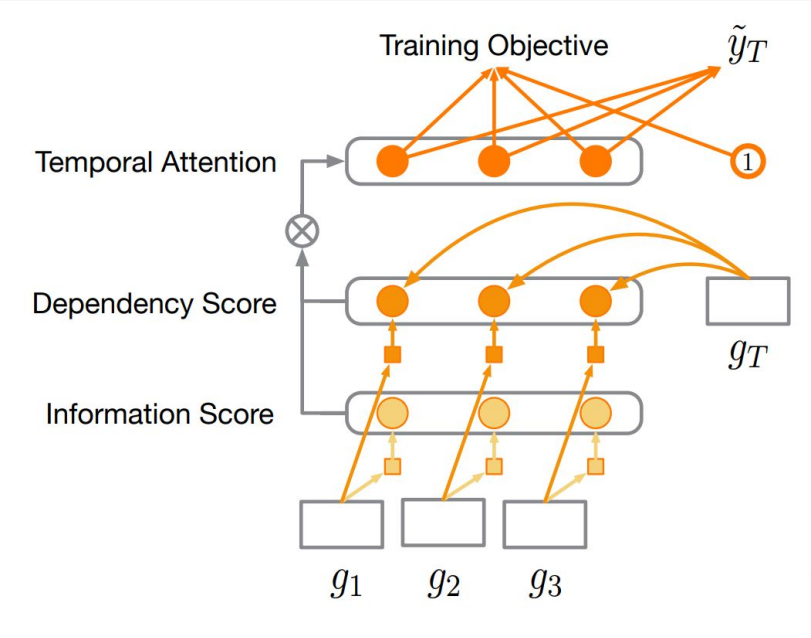
StockNet

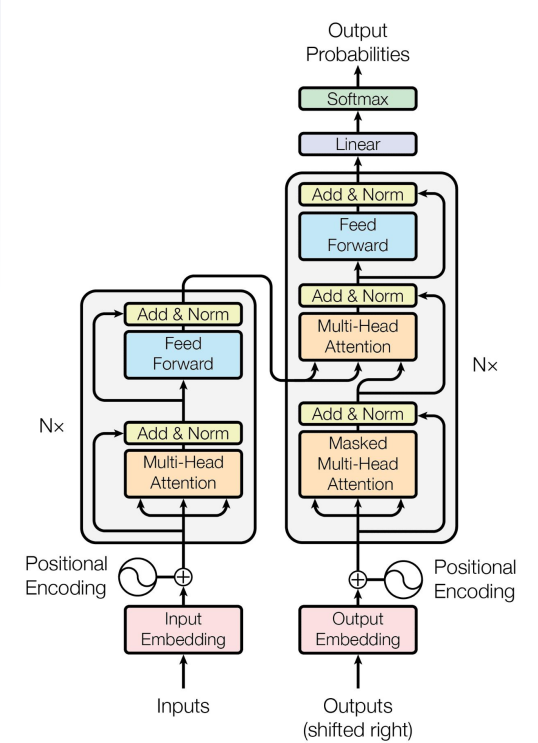




(d) Transformer

Attentive Temporal Auxiliary (ATA)





Parámetros

- Trading days: 5
- Batch Size: 32
- Max. tweets per day: 40
- Max. tokens per tweet: 30
- Word embedding size: 50

- Num. of heads: 8
- Dense Layer: 50
- Key dim (Q,K): 16
- Value dim (V): 16

- Message Embedding Layer: 100
- Variational Movement Decoder: 150
- Adam optimizer
- Learning rate: 0.001
- Input dropout rate: 0.3

5 Resultados

Ejecución

Época	StockNet Original		MultiHeadAttention		Transformer	
	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
1	4.236096	0.594355	3.507428	0.504032	3.778165	0.481855
2	1.935737	0.573387	1.596338	0.506855	1.612628	0.482056
3	1.374547	0.555847	1.045187	0.496774	1.048171	0.486089
4	1.21081	0.542944	0.885669	0.492944	0.890253	0.517944
5	1.139567	0.54254	0.823049	0.487298	0.824322	0.517137
6	1.116311	0.534677	0.789848	0.501008	0.792913	0.517339
7	1.089512	0.551008	0.771544	0.516532	0.778887	0.517137
8	1.074829	0.55121	0.762309	0.518347	0.764101	0.516935
9	1.074176	0.537702	0.756732	0.517944	0.754988	0.517541
10	1.063321	0.546371	0.749721	0.517339	0.749138	0.517541
11	1.057531	0.548387	0.750486	0.517339	0.747233	0.517137
12	1.050444	0.568347	0.750638	0.517137	0.743672	0.518548
13	1.053398	0.556855	0.761624	0.517742	0.742073	0.517137
14	1.042594	0.58246	0.743492	0.516734	0.746247	0.517541
15	1.04298	0.577823	0.739399	0.516734	0.750495	0.517541
Mejor época	1.042594	0.58246	0.762309	0.518347	0.743672	0.518548
Evaluación	1.1229074	0.582301	0.736546278	0.574405	0.731089175	0.574405

Resultados previos

StockNet-Transformer Accuracy = 57.44

Baseline models	Acc.	MCC	StockNet variations	Acc.	MCC
RAND	50.89	-0.002266	TECHNICALANALYST	54.96	0.016456
ARIMA (Brown, 2004)	51.39	-0.020588	FUNDAMENTALANALYST	58.23	0.071704
RANDFOREST (Pagolu et al., 2016)	53.08	0.012929	INDEPENDENTANALYST	57.54	0.036610
TSLDA (Nguyen and Shirai, 2015)	54.07	**0.065382**	DISCRIMINATIVEANALYST	56.15	0.056493
HAN (Hu et al., 2018)	**57.64**	0.051800	HEDGEFUNDANALYST	**58.23**	**0.080796**

- **RAND**: Random guesses
- **ARIMA**: Analysis with only historical prices
- **RANDFOREST**: Random forest with word2vec
- **TSLDA**: Generative model that learns topics and sentiments
- **HAN**: State of the art DNN with hierarchical attention

- **Technical Analyst**: only historical prices
- **Fundamental Analyst**: only tweet information
- **Independent Analyst**: without ATA
- **Discriminative Analyst**: without Z ($z_t = \mu_t$)
- **Hedge Fund Analyst: full STOCKNET**

Posibles modificaciones

- Aumentar número de épocas
- Reducir learning rate
- Reducir batch size
- Reducir dropout
- Modificar *alpha*

¡Gracias por su atención!