Predicción de movimientos de acciones a partir de tweets y precios históricos

Cortés Montes Enrique E., Uc Kantun Orlando de Jesus
Centro de Investigación en Matemáticas. Unidad Monterrey
Email: enrique.cortes@cimat.mx, orlando.uc@cimat.mx

***Resumen*—El siguiente reporte muestra un modelo generativo profundo para la predicción de los precios de acciones en el mercado de valores, explotando tweets y precios históricos. El modelo presenta variables latentes continuas y recurrentes para un mejor tratamiento de la estocasticidad y usa inferencia variacional neuronal para abordar la inferencia posterior. También se ilustra la implementación del modelo, llamado StockNet. Al final se dan unos comentarios de cómo se podría realizar una implementación para datos de México.**

I. Introducción

La predicción del movimiento de acciones ha atraído durante mucho tiempo tanto a inversores como investigadores. Se presenta un modelo para predecir el movimiento del precio de las acciones a partir de tweets y precios históricos.

En el procesamiento del lenguaje natural (NLP), las noticias y las redes sociales son dos recursos de información para la predicción del mercado de valores, y los modelos que utilizan estas fuentes suelen ser discriminatorios. Entre ellos, la investigación clásica se basa en gran medida en ingeniería de características (Schumaker y Chen, 2009; Oliveira et al., 2013). Se utilizan redes neuronales proundas (Le y Mikolov, 2014), y se estudiaron enfoques con estructura de representaciones de eventos (Ding et al., 2014, 2015).

Recientemente, Hu et al. (2018) propusieron analizar la secuencia de noticias directamente desde el texto, con mecanismos de atención jerárquicos, para la predicción de la tendencia de las existencias.

En esencia, la predicción del movimiento de acciones es un problema de series de tiempo. El impacto de la temporalidad entre dos observaciones no se aborda en NLP. Por ejemplo, si una compañía tiene un gran escándalo en el día de trading d_1, generalmente, los precios de sus acciones tenderán a disminuir hasta un día d_2. Si un modelo es capaz de identificar la baja en los precios en el intervalo $[d_1, d_2]$, entonces se pueden desarrollar estrategias de inversión que permita aprovechar ésa situación. Lo anterior viene del hecho de que la información pública, por ejemplo, un escándalo, requiere tiempo para ser absorvida en los movimientos de los precios (Luss y d'Aspremont, 2015).

Se propone el modelo StockNet, un modelo generativo profundo para la predicción del movimiento de las acciones, que toma en consideración la estocasticidad del mercado, la información caótica del mercado (tweets) y la predicción temporalmente dependiente (precios históricos).

En comparación con otros baselines, los experimentos realizados muestran que StockNet logra un mejor desempeño, incorporando datos de Twitter y precios históricos de acciones.

II. Formulación del problema

El objetivo es predecir el movimiento del precio de una acción s en una colección de acciones preseleccionada S en un día de trading objetivo d. Formalmente, usamos la información del mercado M compuesta por tweets, y precios históricos, en la ventana de observación $[d - \Delta d, d - 1]$ donde Δd es un tamaño de lag fijo. Estimamos el movimiento binario donde 1 denota alza y 0 denota baja,

$$y = \begin{cases} 1 & p_d^c > p_{d-1}^c \\ 0 & p_d^c \leq p_{d-1}^c \end{cases}$$

donde p_d^c denota el precio de cierre ajustado ajustado por los movimientos corporativas que afecten el precio de las acciones, es decir, reparto de dividendos, etc.

III. Recolección de Datos

En finanzas, las acciones se clasifican en 9 industrias: Materiales Básicos, Bienes de Consumo, Sanidad, Servicios, Servicios Públicos, Conglomerados, Financiero, Bienes Industriales y Tecnología. Dado el alto volumen comercial de las acciones, éstas tienden a ser más discutidas en Twitter. Se seleccionar 88 acciones diferentes, en el periodo entre 01/01/2014 y 01/01/2016, siendo las 8 acciones de los Conglomerados y las 10 principales acciones (en tamaño de capital) de cada una de las otras 8 industrias.

Hay dos componentes principales en el conjunto de datos:

1. Datos de Twitter,
2. Datos de precios históricos.

Los precios históricos de las 88 acciones fueron tomados de Yahoo Finance.

Los tweets fueron tomados desde la API oficial de twitter y fuero preprocesados a través del módulo NLTK de Python.

IV. Descripción general del modelo

Se supondrá que la predicción del movimiento del día de trading d se puede beneficiar de la predicción de los movimientos de días anteriores, por lo que se realizan predicciones de los movimientos previos en un periodo de observación.

Por ejemplo, como se puede observar en la Figura 2, para un día objetivo 07/08/2015 y un lag de observaciones previas de

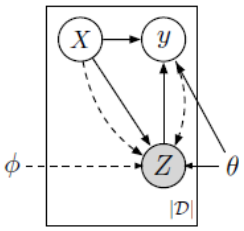

Figura 1: Gráfico del proceso generativo de la información observada del mercado del movimiento de las acciones.

5 días, 03/08/2012 y 06/08/2012 son días de trading elegibles y por lo tanto, también se realizan predicciones con ellos.

No todos los días se consideran días de trading, dado que existen, por ejemplo, fines de semana y días inhábiles, por lo que para mejor organización, se consideran como input los días de trading en lugar de los días calendarios.

Primero se encuentran todos los T días de trading elegibles, en otras palabras, existentes en el intervalo de tiempo $[d - \Delta d + 1, d]$, es decir, indexamos estos días de trading con $t \in [1, T]$ y cada uno de ellos se asigna a un día de trading (absoluto) d_t. Luego se propone una alineación del día de trading: se reordenan los corpus de tweets y los precios históricos, alineándolos con estos días de negociación T. Específicamente, para el t-ésimo día de trading, se reconoce la información del mercado M_t en $[d_{t-1}, dt)$ y los precios históricos p_t en d_{t-1}, para predecir el movimiento y_t en d_t.

En la Figura 1 se muestra el proceso generativo del modelo. Se realiza el encoding de la información del mercado como una variable aleatoria $X = [x_1, \ldots, x_T]$, desde la que se genera un factor latente $Z = [z_1, \ldots, z_T]$ para la predicción. Se busca modelar la distribución de probabilidad condicional $p_\theta(y|X) = \int_Z p_\theta(y, Z|X)$ en lugar de $p_\theta(y_T|X)$.

En la Figura 2 se muestra un ejemplo de una alineación y se observa la arquitectura del modelo propuesto StockNet que, de manera general, tiene como componentes:

1. Market Information Enconder (MIE) que realiza el enconding de los precios históricos y los tweets a X.
2. Variational Movement Decoder (VMD) que infiere a Z con X, y y realiza el decoding de los movimientos de las acciones y a partir de X, Z.
3. Attentive Temporal Auxiliary (ATA) que integra la función de costo temporal a través de un mecanismo de atención para el entrenamiento del modelo.

V. Componentes del modelo

En ésta sección se detallan los componentes del modelo (MIE, VMD, ATA) y la forma en la que se estiman sus parámetros.

V-A. Market Information Encoder

El MIE realiza el enconding de la información de las redes sociales y de los precios de las acciones para mejorar la calidad de la información del mercado, y tiene como salida la información del mercado X, y siendo X el input del VMD.

Cada input temporal se define como:

$$x_t = [c_t, p_t]$$

donde c_t y p_t son el embedding del corpus y el vector de precios históricos, respectivamente.

La estrategia básica para adquirir c_t es primero alimentar los tweets a la Message Embedding Layer para realizar una representación en bajas dimensiones y poder seleccionarlos de acuerdo a su calidad. Específicamente, la capa consiste de una forward GRU y una backward GRU. En el tweet del t-ésimo día denotamos la secuencia de palabras del k-ésimo mensaje como W, donde W, donde $W_{l^*} = s, l^* \in [1, L]$ y su matriz de embeddings de palabras es $E = [e_1; e_2; \ldots; e_L]$. Las GRUs son:

$$\overrightarrow{h}_f = \overrightarrow{\text{GRU}}(e_f, \overrightarrow{h}_{f-1})$$

$$\overleftarrow{h}_b = \overleftarrow{\text{GRU}}(e_b, \overleftarrow{h}_{b+1})$$

$$m = \frac{\overrightarrow{h}_{l^*} + \overleftarrow{h}_{l^*}}{2}$$

donde $f \in [1, \ldots, l^*]$, $b \in [l^*, \ldots, L]$, y donde los valores ocutos $\overrightarrow{h}_{l^*}$ y $\overleftarrow{h}_{l^*}$ sin promediados para obtener el tweet embedding m.

Uniendo todos los tweet embeddings del t-ésimo día se tiene una matriz de tweet embeddings $M_t \in \mathbb{R}^{d_m \times K}$. La calidad de los tweets varía drásticamente, por lo que se realiza una ponderación a través de medidas de inteligencia colectiva, de la forma:

$$u_t = \zeta(w_u^T \tanh(W_{m,u} M_t))$$

donde ζ es la función sigmoide y $W_{m,u}, w_u$ son parámetros del modelo. Entonces, el embedding del corpus es:

$$c_t = M_t u_t^T$$

Con respecto a los precios históricos, en lugar de ingresar el vector de precios $\tilde{p}_t = [\tilde{p}_t^c, \tilde{p}_t^h, \tilde{p}_t^l]$, se toma el precio ajustado y se normaliza con respecto a la observación anterior, de ta forma que:

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

V-B. Variational Movement Decoder

El objetivo del VMD es inferir y decodificar, de manera recurrente, el factor Z y el movimiento y de la información del mercado X.

V-B1. Inferencia: Se utilizan redes neuronales profundas para ajustar las distribuciones latentes, es decir, la distribución a priori $p_0(z_t|z_{<t}, x_{<t})$ y la posterior $p_\theta(z_t|z_{<t}, x_{<t}, y_t)$ y esquivar la complejidad de la estimación a partir de aproximación neuronal y reparametrización.

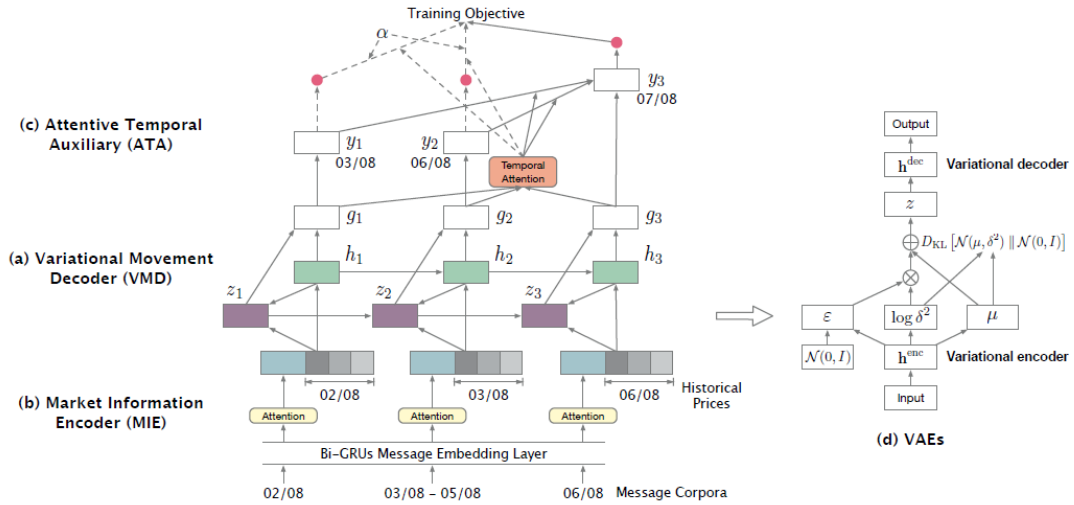

Figura 2: Arquitetura de la red StockNet

A partir de una minimización de la divergencia de Kullback-Leibler se maximiza el siguiente límite inferior recurrente variacional:

$$\mathcal{L}(\theta,\phi;X,y) = \sum_{t=1}^{T} \mathbb{E}_{q_\phi(z_t|z_{<t},x_{<t},y_t)}\{\log p_\theta(y_t|z_{<t},x_{<t}) - D_{KL}[q_\phi(z_t|z_{<t},x_{<t},y_t)||(y_t|z_{<t},x_{<t})]\} \leq \log p_\theta(y|X)$$

V-B2. Decoding: Como en el caso de las series de tiempo, VMD utiliza una RNN con una GRU para extraer las características y realizar el decoding de manera recurrente:

$$h_t^s = \text{GRU}(x_t, h_{t-1}^s)$$

y con la representación oculta:

$$h_t^z = \tanh(W_z^\phi[z_{t-1}, x_t, h_t^s, y_t] + b_z^\phi)$$

donde $W_{z,\mu}^\phi$, $W_{z,\mu}^\delta$, W_z^ϕ son matrices peso y b_μ^ϕ, b_δ^ϕ y b_z^ϕ son sesgos.

Finalmente, se integran características determinísticas y la predicción final está dada por:

$$g_t = \tanh(W_g[x_t, h_t^s, z_t] + b_g)$$

$$\tilde{y}_t = \zeta(W_y g_t + b_y), t < T$$

donde W_g, W_y son matrices peso y b_g y b_y son sesgos. La función softmax ζ determina si la salida es un alza o una baja.

V-C. Attentive Temporal Auxiliary

Dado que se generó una serie de predicciones auxiliares $\tilde{Y}^* = [\tilde{y}_1; \ldots; \tilde{y}_{T-1}]$ se introduce un mecanismo de atención temporal, a través de un *information score* y un *dependency score*:

$$v_i' = w_i^T \tanh(W_{g,i} G*)$$

$$v_d' = g_t^T \tanh(W_{g,d} G*)$$

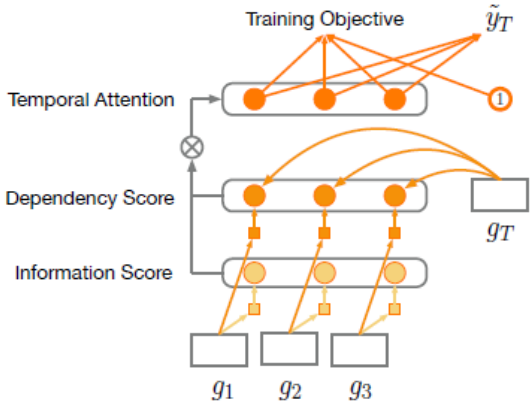

Figura 3: Atención temporal del modelo StockNet.

$$v* = \zeta(v_i' \odot v_d')$$

donde $W_{g,i}$ y $W_{g,d}$ son parámetros del modelo. Las representaciones integradas $G^* = [g_1; \ldots;]$ y g_T son reutilizadas en las representaciones finales de la información temporal del mercado. Los information score v_i' evalúan la calidad de la información de los días de trading, mientras que los dependency score v_d' capturan sus dependencias con respecto a la respuesta. Integrando los dos se obtiene el peso de atención normalizado final v^*, como se puede observar en la Figura 3.

A partir de v^* se construye el vector de peso temporal final $v \in \mathbb{R}^{1\times T}$

$$v = [\alpha v^*, 1]$$

donde 1 es para la predicción principal y se adopta un peso auxiliar $\alpha \in [0, 1]$ para controlar los efectos auxiliares

generales en el entrenamiento del modelo. Finalmente, se escribe la función objetivo $\mathcal{F}$ por recomposición:

$$\mathcal{F}(\theta, \phi; X, y) = \frac{1}{N} \sum_{n}^{N} v^{(n)} f^{(n)}$$

Y se toma la derivada de $\mathcal{F}$ con respecto a todos los parámetros del modelo $\{\theta, \phi\}$ a través de backpropagation para la actualización.

VI. Experimentos

En ésta sección se describen los detalles del experimento y sus resultados.

VI-A. Configuración de entrenamiento

Se utilizó una ventana de observaciones de 5 días y 32 muestras aleatorias por batch. El número máximo de tokens por tweet y el máximo número de tweets en un día de trading son 30 y 40 respectivamente, tomados de manera empírica. El tamaño de embedding de palabras es de 50 para limitar el costo computacional. Se fija el tamaño del Message Embedding Layer en 100 y el del VMD en 150. Se utiliza el optimizador Adam con un learning rate inicial de 0.001. Se utiliza una tasa de dropout de 0.3 para regularización. El ajuste del modelo se implementó en Tensorflow.

VI-B. Métricas de evaluación

Se utilizan la presición y el Coeficiente de correlación de Matthews (MCC) que evita el sesgo de medición debido al sesgo de los datos. Dada la matriz de confusión:

$$\begin{pmatrix} tp & fn \\ fp & tn \end{pmatrix}$$

la métrica MCC se define cómo:

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$$

VI-C. Baselines y modelos propuestos

Se construyeron cinco baselines de diferente tipo:

- `RAND`: Un predictor que aleatoriamente adivina movimiento hacia el alza o a la baja.
- `ARIMA`: Un modelo Autorregresivo Integrado de Medias Móviles ajustado solamente sobre los precios.
- `RandForest`: Un clasificador Random Forest utilizando representacines word2vec de los textos.
- `TSLDA`: Un modelo generativo que aprende conjuntamente temas y sentimientos.
- `HAN`: Una red neuronal del estado del arte con atención jerárquica.

Para hacer un análisis detallado de todos los componentes de StockNet se realizaron cinco variaciones:

- `HedgeFundAnalyst`: La versión completa de StockNet.
- `TechnicalAnalyst`: La versión de StockNet generativa utilizando solamente los precios históricos.
- `FundamentalAnalyst`: La versión de StockNet generativa utilizando solamente la información de los tweets.
- `IndependentAnalyst`: La versión de StockNet generativa sin utilizar predicciones auxiliares.
- `DiscriminativeAnalyst`: La versión de StockNet discriminativa que optimiza directamente la verosimilitud objetivo.

VI-D. Resultados

Una precisión de 56 % es generalmente reportada como un resultado satisfactorio para la predicción del movimiento de acciones.

Se puede observar en la Figura 4 que el modelo completo `HedgeFundAnalyst` tiene la mejor precisión con 58.23 y el mayor MCC con 0.080796, superando a los modelos `TSLDA` y `HAN`.

Resalta el hecho de que el modelo `TechnicalAnalyst` tiene un mejor desempeño que el modelo `ARIMA`, probablemente debido a que el primero aprende de los datos de entrenamiento e incorpora una mayor flexibilidad frente a la no-linealidad, mientras que el segundo tiene cierta sensibilidad ante la estacionariedad.

El desempeño del modelo `FundamentalAnalyst` comparado con el modelo `TechnicalAnalyst` confirma el impacto positivo de los tweets y los precios históricos en la predicción del movimiento de las acciones, respectivamente.

El comportamiento del modelo `HedgeFundAnalyst` sugiere que analizar el estatus del mercado a través factores latentes beneficia la predicción del movimiento de las acciones.

VI-E. Efectos del Temporal Auxiliary

El parámetro α controla los efectos del auxiliar temporal del modelo, por lo que funciona como una especie de parámetro de regularización.

Como se puede observar en la Figura 5 el modelo `HedgeFundAnalyst` tiene un mejor desempeño con $\alpha = 0{,}05$, mientras que el modelo `DiscriminativeAnalyst` tiene un mejor desempeño con $\alpha = 0{,}05$.

Desde la perspectiva del entrenamiento del modelo, el auxiliar temporal ayuda al modelo `HedgeFundAnalyst` a realizar el encoding de información más útil en el factor latente Z.

Se observó los modelos que incluyen el auxiliar temporal tienen un mejor desempeño que aquellos que no lo incluyen, como es el caso del modelo `IndependentAnalyst`.

VII. Comentarios sobre la implementación con datos de México

Para realizar la implementación del modelo StockNet con datos de México primero sería necesario conectar con la API de Twitter que permita descargar los tweets de las empresas mexicanas de las que se desee analizar el movimiento de los precios de sus acciones.

Habría que recolectar la información de los precios de las acciones de interés, que se podría realizar a partir de una plataforma como *mx.investing.com.*

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**

Figura 4: Medidas de precisión y de MCC de los baselines y de las variaciones StockNet.

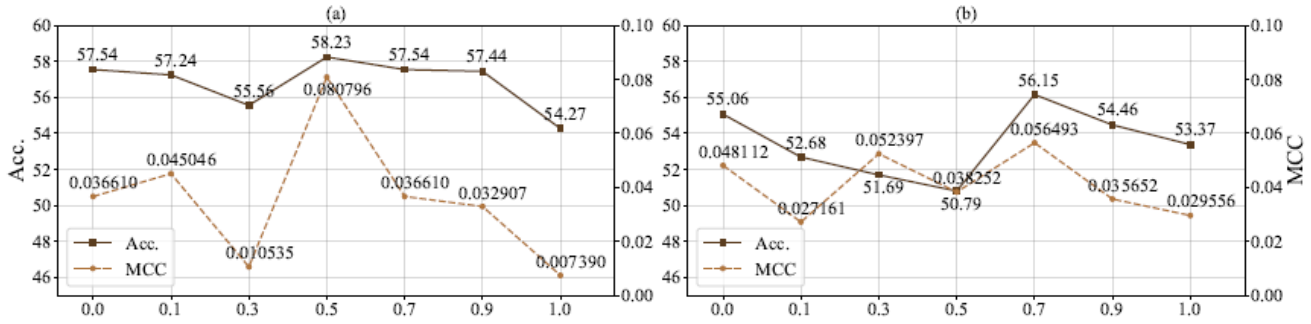

Figura 5: (a) Desempeño de `HedgeFundAnalyst` con diferentes valores de α. (b) Desempeño de `DiscriminativeAnalyst` con diferentes valores de α.

Se tendrían que obtener (o crear) embeddings preentrenados de tweets en español.

Una posible problemática sería la falta de conexión de la población general con el comportamiento de la Bolsa Mexicana de Valores, es decir, existe la posibilidad de que no existan suficientes tweets que permitan alimentar el modelo.

También se tendría que realizar el preprocesamiento de los tweets en español, o bien, considerar como una posibilidad el incluir también tweets en inglés.

Actualmente no existen artículos de investigación que cumplan con las características anteriores.

VIII. CONCLUSIONES

Se demostró la efectividad de los modelos generativos profundos para la predicción del movimiento de acciones a partir de información de redes sociales al introducir StockNet.

Se probó el modelo en un nuevo dataset y se mostró que tiene un mejor desempeño que otros baselines robustos, incluyendo implementaciones de trabajos previos.

Se comentó que las adecuaciones principales para implementar el modelo StockNet con datos de México sería primero generar la base da datos compuesta por los tweets y los precios históricos, así como la creación de embeddings preentrenados de tweets en español.

REFERENCIAS

[1] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014, October). Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1415-1425).

[2] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015, June). Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

[3] Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T. Y. (2018, February). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 261-269).

[4] Luss, R., & d'Aspremont, A. (2015). Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6), 999-1012.

[5] Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).

[6] Oliveira, N., Cortez, P., & Areal, N. (2013, June). Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics* (pp. 1-8).

[7] Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2), 1-19.

[8] Xu, Y., & Cohen, S. B. (2018, July). Stock movement prediction from tweets and historical prices. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1970-1979).