UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Noviembre 2020, Vol. 13, No. 2, pp. 51–61

Exploración del poder predictivo de datos extraídos de StockTwits respecto a la dirección de variación futura del precio de un activo transado en la Bolsa de Valores de Nueva York

Predictive power exploration of the data extracted from StockTwits over the future price variation direction of a stock traded in the New York Stock Market

Andrés Felipe Rodríguez Perez[a]
andres.rodriguezp@usantotomas.edu.co

Robert Romero[b]
gilromero@usantotomas.edu.co

Resumen

Diariamente se generan grandes volúmenes de información, especialmente en las redes sociales. El uso de esta información como insumo para el estudio del comportamiento de los agentes en el mercado de valores ha venido cobrando fuerza, especialmente en el campo del aprendizaje de máquina. Es por ello que, en este artículo se presenta un estudio de la capacidad predictiva de la información que generan los agentes del mercado en la red social StockTwits sobre la variación de la dirección del precio de una activo transado en la Bolsa de Valores de Nueva York, valiéndose de herramientas de minería de datos y algoritmos de aprendizaje de máquina.

***Palabras clave*:** Bolsa de Valores de Nueva York, Precio del Activo, Minería de Datos, Aprendizaje Automático, Procesamiento del Lenguaje Natural, Análisis de Redes Sociales..

Abstract

High volume of data is generated daily, especially on social networks. The usage of this data as a source in the study of the agent's behavior in the stock market have been gaining interest, specifically in the machine learning field. Hence, in this article; a study about the predictive power of this kind of data over the future price variation direction of a stock is made, using the texts published in the StockTwits social network and machine learning techniques.

***Keywords*:** New York Stock Exchange, Stock Price, Data Mining, Machine Learning, Natural Language Procesing, Social Network Analysis.

[a]Egresado Facultad de Estadística
[b]Docente Facultad de Estadística

1. Introducción

La negociación diaria de activos en la bolsa de valores tiene un factor que la hace compleja e impredecible, la volatilidad. Esto genera mucha incertidumbre y cuando se trata de una decisión de inversión que involucra mucho capital las pérdidas pueden ser millonarias si no se logra un retorno aceptable. Los modelos tradicionales de series de tiempo buscan modelar esta esta incertidumbre usando la información del pasado, ya sea de la serie objetivo o de variables adicionales que pueden estar disponibles o ser estimadas a futuro para complementar el pronóstico, logrando buenos resultados a un bajo costo computacional (Stockinger & Dutter, 1987). A pesar de esto, la aplicación de algoritmos de aprendizaje de máquina en el campo de la predicción de los precios de los activos tranzados en el mercado de valores; se ha venido incrementando en los últimos años gracias a la capacidad de estos modelos para aprender relaciones complejas presentes en los datos y su flexibilidad para incorporar información proveniente de fuentes no estructuradas como las redes sociales.

Los datos generados por los usuarios en las redes sociales se han convertido en un insumo potencial en la determinación de una opinión general o reacción ante un hecho específico y existen trabajos recientes que demuestran que dicho **"sentimiento"** general puede ser tomado como aproximación de las intenciones de un agente en un mercado específico. Asur & Huberman (2010) demostraron con éxito la capacidad predictiva que tiene el sentimiento general en Twitter sobre las ventas de taquilla de un evento, valiéndose de redes neuronales y Feldman (2013) se encargó de discutir diferentes aplicaciones de la clasificación de sentimientos o extracción de información léxica de textos de redes sociales, entre ellas la capacidad de predecir características como los gustos musicales hasta los hábitos civiles de votación.

En el ámbito del mercado de valores existen redes sociales como **StockTwits**, las cual tiene como propósito que los usuarios, inversionistas y aficionados compartan ideas acerca de los precios de las acciones, el comportamiento del mercado y la influencia de los factores microeconómicos y macroeconómicos externos. StockTwits dispone de manera abierta toda esta información a través de una API, por medio de la cual es posible la descarga de los comentarios o textos acerca de un activo específico.

Siguiendo la biografiaría existente y la metodología propuesta por Coyne et al (2017), se propone una estrategia de aprovechamiento de la información de la red social, incluyendo otra manera de identificar usuarios influyentes y la relevancia de los comentarios por su numero de**"likes"**. El objetivo es predecir la dirección del cambio en el precio de apertura y de cierre del activo valiéndose de esta información.

2. Marco Teórico

Caracterizar el comportamiento del mercado de valores es una tarea compleja y los investigadores se han venido centrando en usar la información presente en las redes sociales y noticias para extraer un **"sentimiento"** de mercado y evaluar su correlación con las fluctuaciones de precio de los activos.

Por un lado, Zhang, Fuehres & Gloor (2011) validaron que la polaridad de los textos en Twitter tiene una correlación con el movimiento general del mercado y Oh & Sheng (2011) encontraron que los "micro blogs" o comentarios cortos, debido a su naturaleza de tiempo real y alto volumen tienen poder predictivo sobre movimientos del precio del activo en el futuro. Usando un clasificador de "bolsa de palabras" etiquetaron textos cortos según si expresaban una polaridad positiva o negativa, y el sentimiento promedio fue usado para predecir el cambio en la dirección del precio del activo.

A pesar del escepticismo de Oliveira, Cortez & Areal (2013) y su crítica a los pequeños tamaños de muestra con los que Oh & Sheng obtuvieron sus conclusiones, el poder predictivo encontrado permitió que en la academia se comenzara a plantear el uso del sentimiento del mercado implícito en los datos de redes sociales para robustecer modelos de análisis y predicción.

Asur & Huberman (2010) fueron pioneros en usar información de redes sociales, específicamente la polaridad presente en textos cortos y Tweets; para entrenar redes neuronales y predecir con alta precisión la venta de taquillas, mientras que Tsui (2017) probó varios modelos entre ellos un Clasificador Ingenuo de Bayes, y valiéndose de la polaridad de los textos logró predecir el movimiento del precio de ciertos activos transados en la bolsa de valores.

3. Metodología

Para evaluar la capacidad predictiva de los datos inherentes a los textos cortos o tweets respecto a la volatilidad o cambios en los precios de los activos transados en la Bolsa de Valores de Nueva York, se establece una metodología que considera diferentes fases para finalmente entrenar y evaluar un modelo de aprendizaje de máquina respecto a su capacidad de predecir cambios en la dirección de la variación del precio de un activo, las cuales comprenden la extracción de la información, pre procesamiento, polaridad de los textos extraídos y extracción de variables complementarias y modelado de la información.

3.1. Extracción de la información

La información de Tweets o textos cortos que se usará en este trabajo proviene de la red social StockTwits, especializada en la interacción social respecto a temas de inversión y bolsa de valores. A través de su API gratuita, la página provee

un ambiente limitado para descargar 30 textos cortos sobre una acción específica identificada por un ticket único ($AAPL por ejemplo) usando peticiones de protocolo HTTP, además del texto; la respuesta incluye datos acerca del usuario que lo publicó, la polaridad del texto y permite cierta trazabilidad de las interacciones que ha generado ese texto en términos de los likes, retweets y número de seguidores del usuario (StockTwits, 2020). Valiéndose de lo anterior, se diseño un script que se ejecutó día a día durante 3 meses con una frecuencia de 3 ejecuciones diarias en intervalos de tiempo iguales para extraer textos o Tweets acerca de 3 diferentes acciones, obteniendo un total de 18.900 textos publicados entre el 11 de octubre y el 4 de diciembre de 2018.

Por otro lado, los datos relacionados a los precios de apertura y cierre de la acción son extraídos a través de la plataforma Quandl. Esta plataforma especializada en proveer datos relacionados con indicadores macroeconómicos y mercados financieros provee una API a través de la cual se puede acceder de forma masiva a datos históricos de los precios de los activos negociados en las diferentes bolsas del mundo, algunos de ellos de manera gratuita y otros mediante subscripción.

Tabla 1: Datos extraidos y procesados de la API de StockTwits

Stocktwits API	
Variable	Descripción
Text	Texto Corto o Tweet
creation_date	Fecha de Creación
username	Nombre del Usuario Que Posteo
followers	Número de Seguidores del Usuario
following	Número de Usuarios Seguidos por el Usuario
likes	Likes que Tiene el Texto
symbol	Acción Sobre la Que Habla el Texto

Elaboración propia.

Tabla 2: Datos extraidos y procesados de la API de Quandl

Quandl API	
Variable	Descripción
Date	Fecha Correspondiente
Open	Variación Porcentual Día a Día del Precio de Apertura
Close	Variación Porcentual Día a Día del Precio de Cierre

Elaboración propia.

3.2. Pre procesamiento

Debido a que los datos obtenidos de StockTwits no vienen organizados en forma de tabla, es necesario realizar un tratamiento a la información descargada para obtener un set de datos que pueda ser modelado. La primera variable sobre la que se realiza una limpieza exhaustiva es el texto, ya que este tiene contenido que no provee información relevante para el modelado de los datos. Elementos como links, emojis, signos de puntuación, conectores, stop words y errores ortográficos son eliminados para luego llevar a todas las palabras resultantes a su raíz mediante lematización. Además de la limpieza del texto, es necesario eliminar los textos carentes de polaridad, que en este caso son los que la plataforma etiqueta como "None". Una vez realizada dicha limpieza se obtienen textos correctamente etiquetados (en este caso el API los etiqueta como "Bullish" o alcista y "Bearish" o bajista) junto con información relacionada a la fecha de creación, nombre del usuario que lo publicó, número de seguidores, número de personas que sigue el usuario, número de likes del texto y la acción sobre la que este hace referencia.

Tabla 3: Proporción de los textos de cada acción según su polaridad

Acción	**Polaridad**	**Frecuencia**	**Porcentaje**
AAPL	Bearish	856	13,6 %
	Bullish	1.866	29,6 %
	None	3.578	56,8 %
IBM	Bearish	368	5,9 %
	Bullish	1.881	30,1 %
	None	3.991	64,0 %
MSFT	Bearish	399	6,3 %
	Bullish	1.976	31,1 %
	None	3.985	62,7 %

Elaboración propia.

3.3. Modelado

Con los datos pre procesados, se procede la parametrización de las variables para así construir un modelo entrenado que pueda ser evaluado y así medir en términos de su rendimiento, si la información extraída tiene poder predictivo sobre el comportamiento del precio de una acción transada en la bolsa de Valores de Nueva York. Previo a dicho entrenamiento, se realiza una transformación de los datos obtenidos de StockTwits para aprovechar al máximo la información que viene acompañada del texto, buscando seguir la metodología propuesta por Coyne et al (2017) en lo que respecta a la identificación de "Usuarios Inteligentes" activos en el momento de la publicación figura-1-.

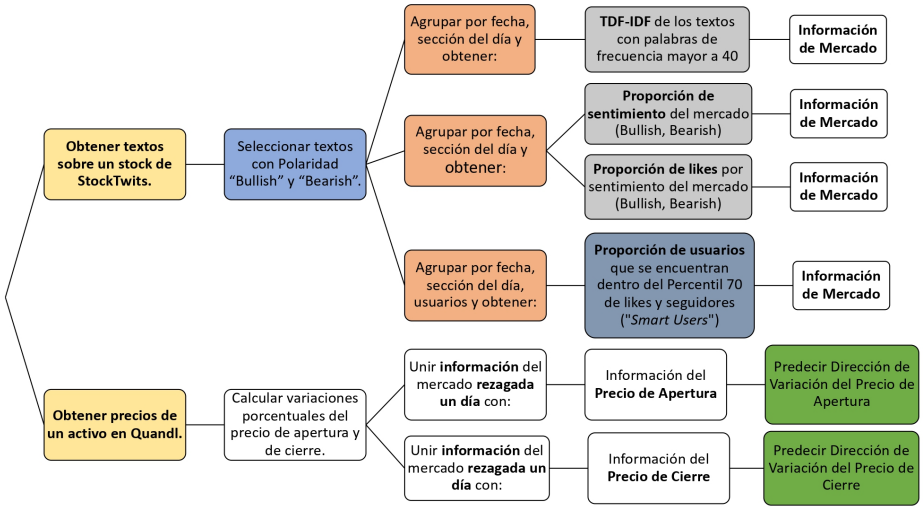

Figura 1: Diagrama de metodología de predicción del la dirección de variación del precio. Elaboración Propia.

Además de la identificación de los "Usuarios Inteligentes" se propone usar la polaridad presente en los textos para calcular una aproximación al sentimiento de mercado usando la proporción de textos "Alcistas" y "Bajistas" junto con la proporción de likes que reciben dichos textos según la polaridad.

Con la información extraída y parametrizada, finalmente se establece el proceso de entrenamiento y selección de diferentes modelos para tres acciones diferentes (Apple Inc 'AAPL', International Business Machines Corporation 'IBM' y Microsoft 'MSFT'). Los modelos para entrenar (red neuronal, un modelo de "Bosques Aleatorios" (Random Forest) y un modelo de Gradiente Extremo basado en árboles "XGBoost") serán sometidos a un proceso de selección de hiperparámetros mediante validación cruzada de 5 particiones aleatorias de los datos de entrenamiento (5-Fold Cross Validation) y con los mejores hiperparámetros en cada caso se seleccionará el que tenga mejor rendimiento en términos de la precisión, falsos positivos y falsos negativos.

4. Resultados

Como se puede observar en la tabla -4- , las redes neuronales obtienen los mejores resultados en dos de los tres activos analizados, siendo Random Forest el mejor modelo para la acción de Apple Inc. Estos resultados son obtenidos con los datos descargados entre el 18 de noviembre y el 4 de diciembre, dejando dos semanas y 4 días para la validación de los modelos.

Tabla 4: Rendimiento general de los predictores escogidos

		AAPL	IBM	MSFT
Random Forest	Precisión	77,3 %	36,4 %	59,1 %
	Sensibilidad	81,3 %	33,3 %	46,7 %
	Especificidad	66,7 %	42,9 %	85,7 %
XGBoost	Precisión	72,7 %	40,9 %	63,6 %
	Sensibilidad	73,7 %	37,5 %	50,0 %
	Especificidad	66,7 %	50,0 %	80,0 %
Red Neuronal	Precisión	59,1 %	59,1 %	72,7 %
	Sensibilidad	68,8 %	46,7 %	62,5 %
	Especificidad	33,3 %	85,7 %	78,6 %

Elaboración propia.

Para las acciones de Apple Inc. el hiperparámetro que permitió las tasas más altas de precisión fue el de una profundidad máxima de 6 para los árboles aleatorios, obteniendo las siguientes curvas de validación:

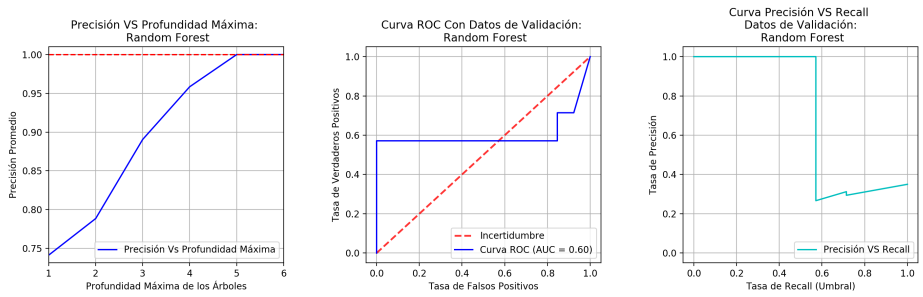

Figura 2: Validación predictor Random Forest para la dirección de variación del precio de la acción de Apple Inc., Elaboración Propia.

En este caso la curva ROC se mantiene aceptable para las tasas de verdaderos positivos hasta un 60 %, sin embargo, a partir de allí la relación verdaderos positivos sobre falsos positivos se invierte; situación que se refleja en la precisión versus el recall dado que para un umbral de decisión superior al 60 % el predictor pierde estabilidad en la decisión.

En cuanto a las acciones de IBM, la selección de los hiperparámetros del mejor modelo se hizo respecto a la tasa de aprendizaje y la penalización L2 de los pesos de la red usando una arquitectura de dos capas ocultas con 6 y 2 neuronas en cada una.

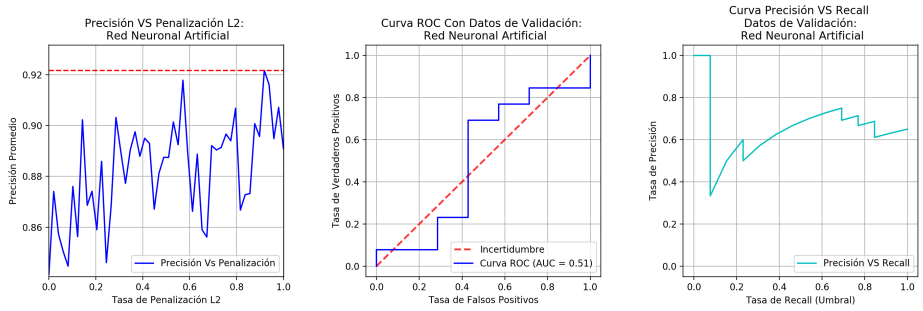

Figura 3: Validación predictor de Redes Neuronales Artificiales para la dirección de variación del precio de la acción de IBM., Elaboración Propia.

En este caso la penalización L2 que generó los mejores resultados fue de 0.9183, entrenando así un modelo que generó una curva ROC muy cercana a la línea de incertidumbre total lo cual explica la baja precisión, sensitividad y especificidad.

Finalmente, para las acciones de Microsoft; el modelo que generó los mejores resultados fue la red neuronal. De la misma forma en que se hizo búsqueda del mejor hiperparámetro para las acciones de IBM, se iteró sobre la tasa de penalización L2 y se seleccionó la que maximizara la precisión promedio.

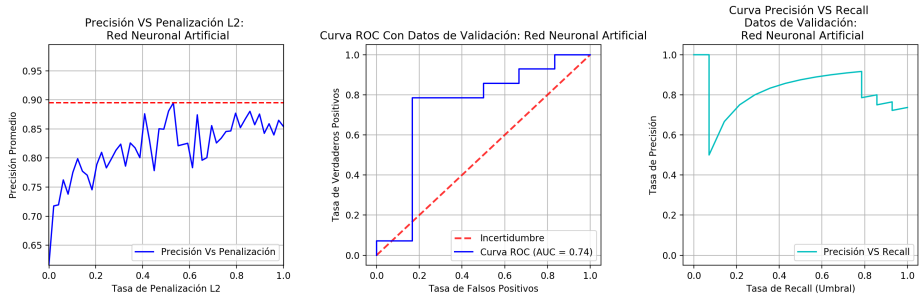

Figura 4: Validación predictor de Red Neuronal para la dirección de variación del precio de la acción de Microsoft, Elaboración Propia.

Para este activo, la mejor red neuronal se estructuró con dos capas ocultas de 6 y 2 neuronas cada una y una tasa de penalización L2 de 0.5306 sobre los pesos de la red. En comparación con los demás modelos, este presenta un comportamiento más estable y una gran parte de la curva ROC se encuentra por encima de la línea de incertidumbre lo que se resume en tasas de precisión, sensibilidad y especificidad por encima del 62 %.

5. Conclusiones

En general se observa que los datos usados no permiten a los predictores seleccionados aprender lo suficiente para generar predicciones confiables y estables en el mediano plazo. Sin embargo, es importante tener en cuenta que para esta investigación se tuvo una limitación en cuanto a la obtención de información y se trabajó únicamente con un intervalo de tiempo de cerca de 2 meses y 2 semanas.

Teniendo lo anterior en mente es destacable que, a pesar de la poca información disponible para el desarrollo de este trabajo; los modelos muestran que existe cierto poder predictivo en estos datos y es probable que al ser complementados con más información llegar a fortalecer la calidad del predictor que se decida entrenar para resolver este problema específico.

Recibido: 15/06/2020
Aceptado: 12/09/2020

Referencias

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01 (pp. 492-499). IEEE Computer Society.

Balaji, S. N., Paul, P. V., & Saravanan, R. (2017, April). Survey on sentiment analysis based stock prediction using big data analytics. In 2017 Innovations in Power and Advanced Computing Technologies (i-PACT) (pp. 1-5). IEEE.

Beysolow II, T. Applied Natural Language Processing with Python.

Copestake, A. (2005). Natural language processing. Lecture Notes, Computer Laboratory, University of Cambridge.

Coyne, S., Madiraju, P., & Coelho, J. (2017, November). Forecasting Stock Prices Using Social Media Analysis. In Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence & Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017 IEEE 15th Intl (pp. 1031-1038). IEEE.

Dale, R., Moisl, H., & Somers, H. (2000). Handbook of natural language processing. CRC Press.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.

Friedman, J. H., & Popescu, B. E. (2003). Importance sampled learning ensembles. Journal of Machine Learning Research, 94305.

Hebb, D. O. (1949). The organization of behavior.

Izenman, A. J. (2008). Modern multivariate statistical techniques. Regression, classification and manifold learning.

Jurafsky, D. (2000). Speech and language processing: An introduction to natural language processing. Computational linguistics, and speech recognition.

Kimoto, T., Asakawa, K., Yoda, M., & Takeoka, M. (1990, June). Stock market prediction system with modular neural networks. In Neural Networks, 1990., 1990 IJCNN International Joint Conference on (pp. 1-6). IEEE.

Kooijman, J. F. (2014). Stock market prediction using social media data and finding the covariance of the LASSO.

LeBaron, B. (2006). Agent-based computational finance. Handbook of computational economics, 2, 1187-1233.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133.

Marquez, L. (2000). Machine learning and natural language processing. In Complementary documentation for the conference "Aprendizaje automático aplicado al procesamiento del lenguaje natural", Soria.

McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based language studies: An advanced resource book. Taylor & Francis.

Nausheen, S., Kumar, A., & Amrutha, K. K. SURVEY ON SENTIMENT ANALYSIS OF STOCK MARKET. ISO 690

Oh, C., & Sheng, O. (2011, December). Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement. In Icis (pp. 1-19).

Oliveira, N., Cortez, P., & Areal, N. (2013, September). On the predictability of stock market behavior using stocktwits sentiment and posting volume. In Portuguese Conference on Artificial Intelligence (pp. 355-365). Springer, Berlin, Heidelberg.

Preethi, G., & Santhi, B. (2012). STOCK MARKET FORECASTING TECHNIQUES: A SURVEY. Journal of Theoretical & Applied Information Technology, 46(1).

Samanidou, E., Zschischang, E., Stauffer, D., & Lux, T. (2007). Agent-based models of financial markets. Reports on Progress in Physics, 70(3), 409.

Stockinger, N., & Dutter, R. (1987). Robust time series analysis: A survey. Kybernetika, 23(7), 1-3.

Theeramunkong, T., Kongkachandra, R., & Supnithi, T. Advances in Natural Language Processing, Intelligent Informatics and Smart Technology.

Tsui, D. (2017). Predicting Stock Price Movement Using Social Media Analysis. Stanford University, Technical Report.

Werbos, P.J. (1974). Beyond regression: new tools for prediction and analysis in the behavioral sciences, Ph.D. dissertation, Harvard University.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". Procedia-Social and Behavioral Sciences, 26, 55-62.