# Minería de opinión en blogs financieros para la predicción de tendencias en mercados bursátiles

Sergio Hernández, Sabino Miranda-Jiménez, Elio Villaseñor,
Eric Sadit Tellez, Mario Graff

INFOTEC - Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Aguascalientes, México

kertr3@hotmail.com, {sabino.miranda, elio.villasenor, eric.tellez,mario.graff}@infotec.mx

**Resumen.** El análisis de redes sociales para el estudio de mercados financieros se ha vuelto un tema de investigación y desarrollo de herramientas que permite a los agentes financieros usar las opiniones de la gente para aumentar la precisión en las predicciones de mercado. Nuestra investigación se enfoca en la predicción de la tendencia de índices financieros usando la minería de opinión, basado en el análisis de blogs especializados en finanzas para el idioma inglés. Los comentarios vertidos en estos blogs son clasificados en términos de su opinión respecto a la tendencia de mercado (a la alza, estable o a la baja). Se evalúan distintas técnicas de aprendizaje computacional y minería de textos para la clasificación de los comentarios realizados durante un periodo de tres meses. Los resultados obtenidos muestran que este análisis puede ser incorporado como un factor en la toma de decisión de los agentes financieros y mejorar la precisión de sus proyecciones.

**Palabras clave:** minería de opinión, *algorithmic trading*, aprendizaje computacional.

## 1. Introducción

En los últimos años, la minería de opinión se ha usado en diferentes campos de aplicación como reseñas de películas, ranking de productos, prestigio de personalidades (políticos, figuras del medio de espectáculo, etc.), entre otros [1, 2]. En particular, hay un fuerte interés de aplicar los métodos de minería de opinión en el ámbito económico-financiero [3,5] para crear o mejorar los métodos de predicción con base en la información de blogs especializados en el tema, donde participan activamente agentes (*traders*, especuladores, *hedgers*) o líderes de opinión (analistas financieros). Por otro lado, dado que se han creado nuevas formas de realizar transacciones financieras mediante el uso de computadoras con más capacidades y mejor desempeño, surgió el área de aplicación llamada *Algorithmic Trading* (AT), también conocida como *trading* automático.

El AT es el uso de plataformas electrónicas para realizar transacciones financieras con algoritmos en variables que incluyen tiempo, precio y volumen de los índices financieros [4]. Los dos enfoques principales del AT son análisis de alta frecuencia y el análisis cuantitativo. Uno de los intereses es el análisis cuantitativo o también conocido como *Quantitative Trading* (QT). El QT es generalmente usado por instituciones financieras, donde se manejan grandes volúmenes de transacciones que involucran la compra y venta de cientos de miles de instrumentos financieros (acciones, futuros, etc.). El QT utiliza datos históricos que se remontan en el tiempo para crear una proyección a futuro, analiza la estructura, encuentra patrones y tendencias de mercado.

Recientemente, los métodos de QT están incorporando información de medios sociales (Twitter, Facebook, blogs, etc.) para mejorar o crear estrategias financieras y predicciones del mercado bursátil [6]. En el dominio de las finanzas, muchos *bloggers* publican sus opiniones sobre compañías específicas y sobre los mercados financieros en general.

Los medios sociales especializados, en particular los blogs dedicados al tema financiero, tienen la ventaja que sus usuarios están relacionados con este ámbito y es probable que sus opiniones y su conocimiento del mercado bursátil este expresado en dichos mensajes. Por lo que se espera que este conocimiento pueda ser útil para mejorar las predicciones de los algoritmos sobre el comportamiento de los instrumentos financieros.

En los diferentes blogs especializados, se discuten diversos temas como: el del mercado financiero, de las acciones, estrategias de operación, divisas, entre muchos otros; dicha información puede usarse para realizar predicciones del mercado bursátil basadas en el análisis del sentimiento de las opiniones [8, 9]. Por ejemplo, se han realizado estudios sobre la red social Twitter que demostró una conexión entre las relaciones de trabajo de las personas y el sentimiento creado por medio de la red de Twitter referente a las acciones del mercado, donde el sentimiento de la red a su vez influye en el movimiento de ciertas acciones [10].

También, se han realizado varios estudios acerca de la relación entre la búsqueda en línea y el comportamiento de acciones en el mercado [11, 12]. En [13] se encontró que el incremento en la intensidad de la búsqueda de acciones en Japón, incrementa la actividad de operaciones realizadas en dichas acciones. Otro tipo de estudio [14] relacionado con la minería de opinión, se basa en los títulos de noticias financieras y no en el texto completo del documento para la predicción del mercado de divisas.

Otro estudio en Twitter [6] demostró una relación entre el sentimiento y la ganancia en las acciones del mercado, así como el volumen de mensajes con el volumen de transacciones bursátiles realizadas. En [7] se muestra un método de AT que considera el análisis de sentimientos de comentarios en un blog que produce mejores ganancias que los métodos tradicionales de compra-venta, ya que obtuvieron un aumento de 56.28% en la precisión de la tendencia.

Nuestro trabajo se enfoca en la minería de opinión de los comentarios subjetivos vertidos en blogs especializados del ámbito financiero (*Business Insider*, *Econbrowser*, *Dealbreaker*, *Investing*, etc.). Esto involucra la clasificación de dichos comentarios en las siguientes clases: a la alza, estable, a la baja. Además nuestro enfoque incorpora un

lexicón de palabras clasificadas de acuerdo a las clases de tendencia mencionadas. Este lexicón se construyó manualmente.

El resto del artículo está organizado como sigue. En la sección 2, se describe, la metodología de nuestro enfoque. En la sección 3, se presentan los resultados obtenidos. Por último, en la sección 4, se presentan las conclusiones y el trabajo futuro.

## 2. Metodología

El enfoque que se propone es incorporar información de medios sociales (blogs especializados) a los métodos de QT para mejorar sus predicciones. Para lograr esto, se plantea el esquema de la Fig. 1, el cual es un enfoque supervisado tradicional, donde se cuenta con información etiquetada manualmente (corpus de entrenamiento) y se usa como entrada en la etapa de entrenamiento, donde se construye el modelo de predicción. En este artículo, consideramos solamente la información de blogs especializados y, por ahora, no incorporamos la información de los índices financieros dentro del modelo de predicción.

### 2.1. Esquema general

En esta parte del trabajo, se considera únicamente la clasificación de los comentarios que indican si los índices financieros están a la alza, a la baja o se mantienen estables.

Se presenta la construcción del corpus financiero etiquetado por un experto, así como el preprocesamiento de la información, que consiste en la limpieza de datos de los comentarios, y la clasificación de estos por medio de los métodos supervisados propuestos; además, se incorpora conocimiento adicional al modelo de predicción a través de un diccionario de palabras etiquetado manualmente que representan tendencias del mercado financiero.

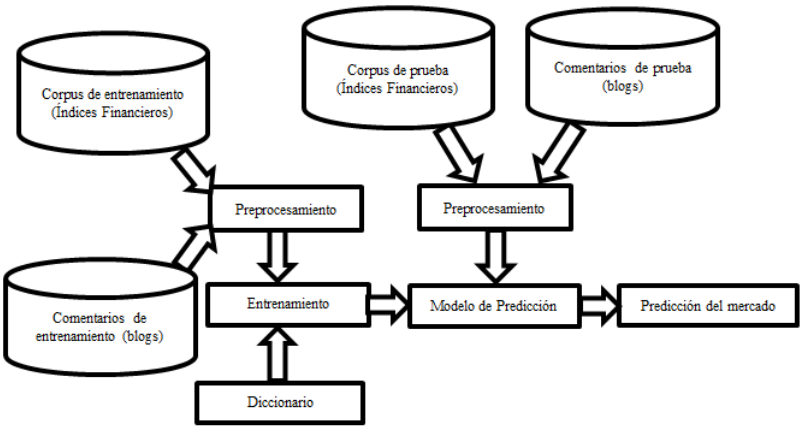


**Fig. 1.** Esquema general de la solución propuesta.

### 2.2. Índices financieros

Un índice financiero es un número abstracto que representa el movimiento en conjunto de varios activos financieros que lo componen (acciones, bonos, monedas, *commodities*, etc.). Cada uno de estos activos tiene un peso relativo dentro del índice, medido según parámetros previamente establecidos a la creación del índice. Entonces, ante un movimiento de precios del activo se moverá el índice y la variación del índice será mayor o menor según el peso del activo. Los Índices más conocidos son los de las bolsas de valores como el NYSE, Nasdaq, Dow jones, SP&500, Merval, Nikkei, etc.

### 2.3. Creación del corpus

La información que se utilizó para la creación del corpus proviene del blog financiero *investing.com* [15] sobre el índice S&P500 [16]. El blog de discusión Investing.com es un portal de internet que proporciona noticias, análisis, datos técnicos sobre los mercados financieros globales donde los usuarios pueden obtener una fuente óptima integral de características innovadoras en diferentes instrumentos financieros.

Para la generación del corpus de entrenamiento de comentarios financieros, se automatizó la descarga de comentarios del blog para el activo 'futuros' referente al índice financiero S&P500. El etiquetado fue realizado por un experto en el tema financiero. Se tomaron 3 meses de comentarios (06-ene-2015 a 03-mar-2015). Se etiquetó cada comentario con alguna de las siguientes clases: a la alza, estable o a la baja; debido a que el precio de cierre del índice S&P500 solo puede tomar alguna de las siguientes posiciones con respecto a su estado anterior: mayor al anterior (alza = 1), Igual al anterior (estable = 0) y menor al anterior (baja = -1). Por ejemplo, en la Tabla 1, el cometario 3 clasificado como 1, contiene palabras como *bull*, que hacen referencia generalmente a una tendencia a la alza del mercado. En el comentario 2, la palabra *short* denota normalmente una tendencia a la baja y se clasifica como -1. En el comentario 4, no hay alguna expresión o palabra significativa que marque una tendencia, por lo cual se clasifica como 0.

Dentro del etiquetado el experto tuvo problemas al clasificar ciertos comentarios debido a la falta de contexto o falta de conocimiento sobre la posición en que se encontraba el mercado al momento de publicarse el comentario; por ejemplo, en el siguiente comentario "*Looks like the 2100 support will be broken today -- get ready*"; no se tiene suficiente contexto para decidir si el mercado tiene una tendencia a la alza o a la baja, por lo que se decidió clasificar este comentario de forma estable (0).

**Tabla 1.** Ejemplos de comentarios etiquetados.

| Num. | Comentario | Clase |
|---|---|---|
| 1 | if that's what you believe you need to read this : http://finance.yahoo.com/news/us-running-room-store-oil-171025359.html | 0 |
| 2 | low risk entry for a scalp short | -1 |
| 3 | end of this bull run is near.. | 1 |
| 4 | Sweet! | 0 |
| 5 | Shorting again.. sorry couldn't resist. Lol | -1 |

El corpus recopilado consta de un total 837 comentarios etiquetados; y contiene 1,884 palabras únicas.

En la Tabla 1 se muestran algunos ejemplos etiquetados por el experto del dominio.

### 2.4. Descripción del diccionario

Nuestro enfoque incorpora el conocimiento de un experto al modelo por medio de un diccionario de palabras clave relacionadas con la alza o la baja en la tendencia del índice financiero. Las palabras clave fueron etiquetadas por el experto con las clases a la alza (1) y a la baja (-1). El diccionario cuenta con 52 palabras/frases únicas usadas comúnmente en los foros especializados en temas financieros. En la Tabla 2, se presentan algunos ejemplos de las palabras del diccionario.

El etiquetado fue añadido al conjunto de entrenamiento para comparar si una metodología sin diccionario o con diccionario obtenía una mejor predicción.

**Tabla 2.** Ejemplo de palabras del diccionario propuesto

| Palabra | Clase | Palabra | Clase | Palabra | Clase | Palabra | Clase |
|---|---|---|---|---|---|---|---|
| *Bear trap* | 1 | *Go north* | 1 | *Heading below* | -1 | *Niagara* | -1 |
| *Bull trap* | -1 | *Going up* | 1 | *overbought* | -1 | *Bearish* | -1 |
| *Oversold* | 1 | *Climbing up* | 1 | *Bottom line* | -1 | *Drown* | -1 |
| *Downtrend* | -1 | *Trend head up* | 1 | *Going down* | -1 | *Bottom* | -1 |
| *Heading above* | 1 | *bullish breakup* | 1 | *downtrend* | -1 | *Rush sell* | -1 |
| *Upper line* | 1 | *Up leg* | 1 | *Bull trap* | -1 | *Break down* | -1 |
| *Bear trap* | 1 | *Up higher* | 1 | *Run down* | -1 | *Head down* | -1 |

### 2.5. Métodos de clasificación

A continuación se presentan las herramientas y métodos para para la solución del problema planteado, los diferentes pasos que se siguieron para procesar el corpus construido; además, se presentan los clasificadores utilizados para la clasificación de los comentarios.

**Preprocesamiento de los datos.** En esta etapa, se realiza una reducción del vocabulario usado por medio del filtrado de las palabras de contenido. Este proceso se aplica a todos los comentarios que se consideran para el proceso de clasificación. Para reducir el vocabulario, usamos un filtrado basado en lematización, y solo usamos las palabras de contenido, es decir, sustantivos, adjetivos, verbos y adverbios; también usamos las interjecciones debido a que contienen un contexto emotivo; también se eliminan las *stop words* (preposiciones, determinantes, etc.).

Para realizar esta etapa, usamos el etiquetador de partes de oración de Stanford para el idioma inglés [17] y usamos solo los lemas de las palabras. Para el filtrado de las palabras usamos las etiquetas definidas por el *Penn Tree Bank Project* para el idioma

inglés [18]. Por ejemplo, en el comentario "*Looks like the bulls returned after lunch*" que indica una tendencia a la alza, es etiquetado como *look*_VBZ *like*_IN *the*_DT *bull*_NNS *return*_VBD *after*_IN *lunch*_NN.

Para realizar el filtrado de palabras, se usa el inicio de las etiquetas de cada palabra, solo consideramos las palabras de contenido como se mencionó. Para el ejemplo, se seleccionaron del comentario los sustantivos (_NN), y verbos (_V); se descartaron las preposiciones (_IN) y los determinantes (_DT), no se considera como parte del mensaje final. De esta forma, el contenido del mensaje resultante sería "*look bull return lunch*". Estas palabras son las que se consideran como características (*features*) para la etapa de entrenamiento de los modelos.

**Modelo de predicción.** En nuestros experimentos usamos dos clasificadores, a saber, *Support Vector Machine* (SVM) [19] y Naive Bayes (NB) [20] que usualmente se aplican en minería de opinión [1, 2]. Se utilizó la librería del software *Python* llamada *sklearn* [21] que implementa ambos clasificadores.

Para el clasificador SVM, se aplicaron 3 diferentes tipos de kernel: el primero, una función de base radial dada por $exp(-\gamma|x - x'|^2)$ con parámetro gamma 1; en el segundo caso, se aplicó un kernel polinomial de grado d=3 expresado con el polinomio de la forma $(\gamma\langle x, x'\rangle + r)^d$ con un coeficiente r=0 y gamma 1; para el último caso, se utilizó un kernel lineal $\langle x, x'\rangle$.

En el caso del clasificador probabilístico bayesiano, su ventaja es que solo se requiere de una pequeña cantidad de datos de entrenamiento para estimar los parámetros (las medias y las varianzas de las variables) para la clasificación.

Para cada uno de los clasificadores, se aplicó una evaluación de *ten-fold cross-validation* con 2 métricas diferentes, que nos dice que tan independientes son nuestros datos de prueba contra los datos de entrenamiento. Se utilizó el 90% de los datos como entrenamiento, usando una función de la misma librería sklearn (*ShuffleSplit*) que toma diferentes muestras del total de datos para obtener el porcentaje requerido. Una vez obtenido el conjunto de entrenamiento se evalúa con el 10% restante de los datos, los resultados fueron promediados sobre las 10 iteraciones realizadas.

## 3. Resultados

Los resultados que se muestran en la Tabla 3, son obtenidos después de realizar la evaluación *ten-fold cross validation*. Se usaron medidas estándares de evaluación para los clasificadores como Precisión y *Recall* así como la medida-F, que es la media harmónica de Precisión y *Recall* [22]. Donde la Precisión se define como el cociente del número de comentarios correctamente clasificados a la clase *X* entre el número de comentarios que se atribuyeron a la clase *X*. *Recall* se define como el cociente del número de comentarios correctamente clasificados a la clase *X* entre el número total de comentarios de la clase *X*.

Se realizaron principalmente dos experimentos para cada uno de los clasificadores, donde se hace uso del diccionario propuesto como conocimiento adicional y sin uso del diccionario. En la Tabla 3, se presentan los porcentajes promedio de los resultados obtenidos con tres diferentes núcleos (kernel) para el clasificador SVM. Se puede observar

que tanto el kernel de función de base radial, como el de base polinomial de grado tres presentan un promedio de precisión del 38.45% sin el uso de diccionario. En el caso del kernel lineal sin diccionario la precisión sube en aproximadamente 17 puntos porcentuales, el promedio final es de 55.12%, por lo que, es el kernel con mejor precisión.

En el caso del clasificador Naive Bayes (NB) sin uso del diccionario, se obtuvo un promedio de precisión del 55.11%, el resultado es aproximado al obtenido por el clasificador SVM con kernel lineal. El mejor resultado obtenido fue tanto del SVM lineal como del NB dando un promedio de precisión final de 57.77% incorporando el diccionario.

**Tabla 3.** Resultados de la validación.

| **Clasificador** | **Kernel** | **Sin diccionario** | | | **Con diccionario** | | |
|---|---|---|---|---|---|---|---|
| | | **Precisión** | ***Recall*** | **Medida-F** | **Precisión** | ***Recall*** | **Medida-F** |
| SVM | Función radial | 38.45 | 35.48 | 18.58 | 34.77 | 33.96 | 17.20 |
| SVM | Polinomio | 38.45 | 35.48 | 18.58 | 32.00 | 33.96 | 17.20 |
| **SVM** | **Lineal** | 55.12 | 51.36 | 51.04 | **57.77** | 56.03 | 55.83 |
| **NB** | – | 55.11 | 48.74 | 48.18 | **57.77** | 51.70 | 50.36 |

## 4. Conclusiones y trabajo futuro

En este trabajo se exploró, el uso de un blog especializado en finanzas como fuente de información para realizar análisis de sentimientos en las tendencias del mercado bursátil; también se presentó la creación de un corpus con más de 800 comentarios etiquetados que representan, según su contenido, la percepción subjetiva del mercado financiero. En el experimento se consideraron los precios de cierre del instrumento futuro en el índice S&P500 para etiquetar los comentarios, puesto que es lo que más prevalece en los comentarios de dicho blog.

También se evaluó el desempeño de dos clasificadores para esta tarea: una máquina de soporte de vectores (SVM) con tres diferentes núcleos y un clasificador bayesiano. Los mejores resultados obtenidos fueron por el clasificador SVM con un kernel lineal y por el clasificador Naive Bayes. En ambos casos, se usó el diccionario de dominio específico de palabras, que denotan una tendencia en el mercado bursátil, como conocimiento adicional de los clasificadores, con lo que se observa que el diccionario ayuda a tener mejores resultados.

Como trabajo futuro, se propone probar más clasificadores para tener una comparativa más amplia al realizar la clasificación de los comentarios, así como extender el diccionario de palabras que denotan alguna tendencia en el mercado bursátil.

Referente a la etiquetación del corpus de comentarios financieros, se identificó que el experto humano tiene dificultad al etiquetar la clase de tendencia en el mercado debido a que no cuenta, en ocasiones, con el suficiente contexto entre el comentario y el

mercado bursátil. En orden para enfrentar este problema se propone, como trabajo futuro tomar datos históricos en un lapso de tiempo referente al instrumento o índice financiero que se desea analizar y presentarlos al etiquetador experto como contexto para que tome una mejor decisión al momento de clasificar el comentario. Además, se considera usar otros blogs orientados a finanzas así como considerar otros índices financieros (NASDAQ, DOW JONES, etc.).

**Agradecimientos.** Este trabajo fue parcialmente apoyado por el gobierno de México (Cátedras Conacyt, SNI).

## Referencias

1. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, vol. 2, pp. 1−135 (2008)
2. Viveros-Jiménez, F., Sidorov, G., Castillo Velásquez, F., Castro- Sánchez, N., Miranda-Jiménez, S., Treviño, A., Gordon, J.: Sondeos automatizados en las redes sociales a través de la minería de opinión. Komputer Sapiens, vol. 2 (2012)
3. Handbook of Financial Data and Risk Information. Eds. Margarita S. BroseMark D. FloodDilip Krishna, Deloitte & Touche, LLPBill Nichols, Cambridge University Press (2014)
4. Lin, T. C.: The New investor. UCLA L. Rev., vol. 60, p. 678 (2012)
5. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., Gordon, J.: Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. LNAI 7630, pp. 1−13 (2012)
6. T. O. Sprenger, A. Tumasjan, P. G. Sandner y I. M.Welpe: Tweets and trades: The information content of stock microblogs. European Financial Management, vol. 20 (5), pp. 926−957 (2014)
7. Z. Chen, X. Du.: Study of stock prediction based on social network. In: International Conference on Social Computing (SocialCom), IEEE, pp. 913−916 (2013)
8. Grayson, M. R., Kwak, M., Choi, A.: Automated platform for aggregation and topical sentiment analysis of news articles, blogs, and other online publications. Issues in Information Systems, vol. 15, no. 2 (2014)
9. Van de Kauter, M. Breesch, D., Hoste, V.: Fine-grained analysis of explicit and implicit sentiment in financial news articles. Expert Systems with Applications, vol. 42, no. 11, pp. 4999−5010 (2015)
10. Zhang, W., Li, C., Ye, Y., Li, W., Ngai, W.: Dynamic business network analysis for correlated stock price movement prediction. Intelligent Systems, IEEE, vol. 30, no. 2, pp. 26−33 (2015)
11. X. Li, H. Xie, L. Chen, J. Wang, X. Deng.: News impact on stock price return via sentiment analysis. Knowledge-Based Systems, vol. 69, pp. 14−23 (2014)
12. Fung, S., Tsai, S.-C.: Stock market-driven investment: new evidence on information, financing and agency effects. Applied Economics, vol. 47, no. 27, pp. 2821−2843 (2015)
13. Takeda, F., Wakao, T.: Google search intensity and its relationship with returns and trading volume of japanese stocks. Pacific-Basin Finance Journal, vol. 27, pp. 1−18 (2014)
14. Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., Ngo, D. C. L.: Text mining of news-headlines for {FOREX} market prediction: A multi-layer dimension reduction algorithm

with semantics and sentiment. Expert Systems with Applications, vol. 42, no. 1, pp. 306−324 (2015)
15. Investing com. http://www.investing.com/indices/us-spx-500-futures-commentary. Consultado el 1 abril de 2015
16. S&P500. http://es.wikipedia.org/wiki/S%26P_500. Consultado el 1 mayo de 2015
17. De Marneffe, M. C., MacCartney, B., Manning, C. D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC, vol. 6, pp. 449−454 (2006)
18. Santorini, B.: Part-of-speech tagging guidelines for the Penn treebank project. (1990)
19. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning, vol. 20, no. 3, pp. 273−297 (1995)
20. Murphy, K. P.: Naive Bayes classifiers. University of British Columbia (2006)
21. Python Scikit-Learn. http://scikit-learn.org/stable. Consultado el 15 de enero de 2015
22. Manning, C., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA (1999)