Modelo de Sentiment Analysis para la clasificación de noticias en tiempo real en el Mercado de Valores de Buenos Aires

Juan Pablo Braña, Alejandra M.J. Litterio, Cristina Camós, y Alejandro Fernández

Centro de Altos Estudios en Tecnología Informática- Facultad de Tecnología Informática - U.A.I
Av. Montes de Oca 745 - (C1270AAH) Ciudad Autónoma de Buenos Aires, República Argentina

{juan.brana, alejandra.litterio, cristina.camos, alejandroa.fernandez}@uai.edu.ar

Resumen

El proyecto de investigación en curso tiene como propósito mostrar que el monitoreo automático de noticias en tiempo real mediante algoritmos basados en Machine Learning puede servir como herramienta para la toma de decisiones de compra y venta de instrumentos financieros en el Mercado de Valores de Buenos Aires. Con este fin, recolectamos, analizamos y clasificamos opiniones extraídas de Twitter aplicando principios y técnicas de Sentiment Analysis relacionando aquellas noticias que generan un impacto directo sobre las acciones del mercado y aquellas que no lo hacen. Asimismo hemos diseñado un Lexicón de términos económicos-financiero en español que nos permite asignar una etiqueta de polaridad "positiva" o "negativa" al corpus seleccionado. Basados en estas consideraciones, hemos obtenido resultados con buenos índices de precisión.

Palabras clave: Aprendizaje por computador, minería de textos, minería de datos, redes sociales, automatización de compra y venta de instrumentos financieros.

Contexto

El presente proyecto de investigación, cuyo inicio es en abril de 2014, se desarrolla en el Centro de Altos Estudios en Tecnología Informática (CAETI) dependiente de la Facultad de Tecnología Informática de la Universidad Abierta Interamericana (UAI). El proyecto aquí presentado se enmarca dentro de la línea de Algoritmos y Software. Es financiado y evaluado por la Secretaría de Investigación de la Universidad. Cuenta con la participación de docentes y alumnos de la Maestría en Tecnología Informática, la Licenciatura en Matemática y de la Diplomatura en Análisis de Datos para Negocios, Finanzas e Investigación de Mercados.

Introducción

Durante los últimos 10 años se ha abordado el problema de clasificar noticias como "Positivas", "Negativas" o "Neutras" mediante algoritmos basados en Machine Learning. Las aplicaciones de las soluciones a este problemas son múltiples; por ejemplo la predicción de resultados electorales, y la ejecución de transacciones automáticas en los mercados financieros. Muchos de estos algoritmos ya se encuentran en plena producción (Bollen et al., 2010).

Cabe resaltar que las distintas aproximaciones a este tema se han realizado con corpus en idioma inglés estudiando las Bolsas importantes del mundo como las de EEUU, Londres o Alemania. Es así que trabajos recientes han sido publicados en la Bolsa de Alemania donde los autores proponen un modelo basado en Análisis de Componentes Principales y sobre el mismo construyen un modelo basado en indicadores de Sentiment Analysis relacionando aquellas noticias que generan un impacto directo sobre el precio de las acciones y aquellas que no lo hacen (Finter et al., 2010). Estos mismos trabajos se llevaron a cabo en mercados mucho más pequeños como por ejemplo en el de Croacia, donde los autores proponen varios algoritmos basados en Machine Learning para clasificar noticias en "Positivas" o "Negativas" (Željko et al., 2010). Estas noticias también son analizadas desde Twitter, estudiando detalladamente cuáles son las palabras más relevantes de los tweets seleccionados que tienen una correlación con el movimiento del precio de las acciones (Agić et al., 2010; Agarwal et al., 2011; Devitt & Ahmad, 2007).

En cuanto a las técnicas desarrolladas hasta el momento podemos mencionar aquellas que abarcan la utilización de diccionarios hasta algoritmos de detección de patrones con técnicas de Machine Learning y modelos basados en eventos semánticos (Feldman et al., 2011). Por otra parte, las fuentes de datos que son consultadas para generar los modelos suelen ser noticias tomadas de publicaciones específicas, aunque los estudios más recientes y modernos se enfocan a cuentas de Twitter que son las que esparcen una noticia más velozmente (Rao & Srivastava, 2011; Chen & Lazer, 2011; Bollen et al., 2011; Brown, 2012). De la misma manera, y tomando siempre a Twitter como fuente de información, otros autores se preguntan cuáles son las características de las noticias que puedan predecir el movimiento en el precio de las acciones de manera más efectiva y eficiente (Zhang, 2013).

Finalmente, y basados en las fuentes literarias mencionadas, un aspecto central que no podemos ignorar es que al día de la fecha se cuenta con muy poco material en español en finanzas, siendo éste el caso del Mercado de Valores de Buenos Aires, lo cual nos lleva a plantear la necesidad de crear un lexicón en idioma español y desarrollar algoritmos capaces de interpretar dicho idioma. Del mismo modo, se intentará delinear los aspectos sintácticos y una primera aproximación a los aspectos semánticos de los algoritmos en cuestión.

Líneas de Investigación, Desarrollo e Innovación

El objetivo es mostrar que el monitoreo automático y en tiempo real de noticias financieras, en este caso tomaremos como fuente cuentas de la comunidad de expertos en finanzas de Twitter, puede servir como herramienta para la toma de decisiones de compra y venta de instrumentos financieros en el Mercado de Valores de Buenos Aires. Para ello se desarrollan modelos teóricos pertenecientes a las áreas de Minería de Datos, Lingüística Computacional y Finanzas Cuantitativas.

De manera más específica desarrollamos un lexicón inédito que incluye términos financieros, políticos, legales y aquellos términos con carga emocional encuadrados en el contexto del Mercado Bursátil de Buenos Aires.

Luego, entrenamos diversos algoritmos para reconocer la carga Positiva o Negativa de cada tweet y con ello creamos un Índice de Sentimiento para el MERVAL en general, y cada uno de los títulos que componen el panel de acciones líderes en particular.

Correlacionamos estos indicadores con los movimientos alcistas y bajistas en diferentes ventanas de tiempo, esto es, pretendemos mostrar que un indicador de sentimiento positivo se correlaciona directamente con movimientos bursátiles alcistas, de la misma manera que un indicador de sentimiento negativo lo hace con movimientos bajistas. De este modo, se espera que los mismos brinden herramientas eficientes para la toma de decisiones en la compra y venta de acciones.

Resultados y Objetivos

Desde el enfoque lingüístico se ha construido un corpus compuesto por tweets en español para análisis de sentimiento, y diseñado un lexicón basado en semántica especializada en finanzas que reflejan una apreciación o juicio valorativo. El mismo incluye una selección de las palabras más críticas en este contexto para determinar la carga de sentimiento de un texto (enunciado) con contenido económico-financiero. Las mismas han sido clasificadas manualmente asignándoles un peso que refleja su carga negativa/positiva. Cabe mencionar que, el lexicón compuesto de 3023 palabras, se encuentra en una etapa preliminar, el cual será ampliado y modificado de acuerdo con los requerimientos y evolución del proyecto. Por otra parte, debemos aclarar que, en lo que respecta al análisis de textos —entiéndase por ello el discurso económico-financiero que incluye noticias, opiniones, comentarios en las redes sociales, más específicamente en Twitter—, aún nos encontramos en la fase de desarrollo y entrenamiento de algoritmos a nivel semántico, es decir de la unidad léxica que se "activa" en una situación de comunicación determinada, esto es del "valor de la palabra" al "valor del término" (Ciapuscio et al., 2009). Además, es preciso recordar que, con el propósito de llevar a cabo unos de los objetivos específicos, y siguiendo nuestra línea de investigación, nos enfocaremos

en el tratamiento y comportamiento semántico de los términos en relación con la interfaz sintáctica, en trabajos futuros. En otras palabras, la relación de las palabras dentro de una oración, el esquema organizativo de sus "componentes" avanzando en nuestro análisis hacia lo que técnicamente se define como la interfaz "sintaxis-semántica" (Giammatteo & Albano, 2006, 2009).

Considerando las técnicas de Machine Learning, se han desarrollado y evaluado tres algoritmos de clasificación: Random Forest, Naive Bayes, y Support Vector Machines (Zaki & Wagner, 2014; Srivastava & Sahami, 2009). Los mismos toman como entrada el lexicón y un conjunto de tweets de entrenamiento los cuales han sido clasificados manualmente por expertos. En pruebas de "cross validation" sobre el dataset clasificado a mano, el algoritmo basado en Random Forest (Zaki & Wagner, 2014) es el que ha mostrado mejor resultado, como se observa en la siguiente tabla:

ALGORITMO	Precisión en clasificación de tweets Positivos	Precisión en clasificación de tweets Negativos
Random Forest	82,42%	84,23%
Naive Bayes	72,83%	83,52%
Support Vector Machines	61,21%	81,10%

En la actualidad, la evaluación de los algoritmos intenta mostrar que el enfoque de un lexicón específico para el español y las finanzas produce resultados significativamente mejores que la estrategia más simple de utilizar la traducción automática de un lexicón genérico del idioma inglés.

Finalmente, se ha desarrollado un conjunto de herramientas que permiten la obtención y almacenamiento de tweets (en función de términos, hashtags, autores y usuarios referenciados), el armado de datasets de entrenamiento para los algoritmos de aprendizaje por computador, y la edición del lexicón. Las mismas reducen considerablemente el esfuerzo de las tareas rutinarias de la presente investigación.

Formación de Recursos Humanos

El equipo del proyecto, multidisciplinario, se compone principalmente de docentes de la Licenciatura en Matemática, la Diplomatura en Análisis de Datos para Negocios, Finanzas e Investigación de Mercado, y la Maestría en Tecnología Informática así como expertos del área de la lingüística y las finanzas. Durante su primer año el proyecto ha involucrado alumnos avanzados de la Maestría en Tecnología Informática, quienes llevan a cabo su pasantía de investigación al tiempo que identifican temas en los que puedan desarrollar su tesis. Al momento, de los cinco alumnos que participan del proyecto, uno ha enfocado su tema de investigación en la intersección entre la Ingeniería de Software y el modelado semántico de información y se espera la concluya a fines de 2015.

Referencias

Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R. (2011). Sentiment Analysis of Twitter Data. Portland, USA: Association for Computational Linguistics.

Agić, Ž., Ljubešić, N., & Tadić, M. (2010). Towards sentiment analysis of financial texts in Croatian. Bull market, 143(45), 69.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

Brown, E. (2012). Will twitter make you a better investor? A look at sentiment, user reputation and their effect on the stock market. Proceedings of SAIS.

Chen, R., & Lazer, M. (2011). Sentiment analysis of twitter feeds for the prediction of stock market movement.

Ciapuscio, E.; Adelstein, A; Brandani, L.; Ferrari, L.; Gallardo, S.; Kornfeld, S.; Kuguel, I. & Resnik, G. (2009). De la palabra al texto. Estudios lingüísticos del español. Buenos Aires: Eudeba.

Devitt A. & Ahmad, K. (2007). Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. Prague, Czech Republic: Association for Computational Linguistics

Feldman, R., Rosenfeld, B., Bar-Haim, R., & Fresko, M. (2011). The stock sonar—sentiment analysis of stocks based on a hybrid approach. In Twenty-Third IAAI Conference.

Finter, P.; Alexandra Niessen-Ruenzi & Stefan Ruenzi (2010). The Impact of Investor Sentiment on the German Stock Market.

Giammatteo, M. & Albano, H. (2006). ¿Cómo se clasifican las palabras? Buenos Aires: Littera Ediciones.

Giammatteo, M & Albano, H. (2009). Lengua, Léxico, Gramática y Texto. Un enfoque para su enseñanza basado en estrategias múltiples. Buenos Aires: Biblos.

Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) (pp. 119-123). IEEE Computer Society.

Srivastava, N. & Sahami, M. (Eds) (2009). Text Mining: Classification, Clustering and Applications. USA: Chapman & Hall/CRC- Data Mining and Knowledge Discovery Series.

Zaki, M. & M. Wagner (2014). Data Mining and Analysis. Fundamental Concepts and Algorithms. U.S.A: Cambridge University Press.

Željko, A.; Nikola Ljubešić & Marko Tadić (2010). Towards Sentiment Analysis of Financial Texts in Croatian. In Proceedings of the Seventh International Conference on Language Resources and Evaluation. May, 19-20, 2010. European Language Resources Association.

Zhang, L. (2013). Sentiment Analysis on Twitter with stock price and significant keyword correlation (Doctoral dissertation).