Sentiment Analysis para la clasificación de noticias financieras en los Mercados Argentinos. Un modelo híbrido de POST enriquecido semánticamente.

Juan Pablo Braña, Alejandra M.J. Litterio 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, alejandroa.fernandez}@uai.edu.ar

Resumen

El proyecto de investigación en curso, que aquí se presenta, propone un modelo híbrido enriquecido semánticamente, en el cual aplicar un *etiquetador morfosintáctico* con el fin de identificar cómo una determinada secuencia de palabras, a partir de una estructura sintáctica, refleja un indicador de sentimiento, esto es, clasificar una cláusula en positivo, negativo o neutro, dentro de un contexto específico, en nuestro caso particular los Mercados Financieros Argentinos. Con el propósito de llevar a cabo este estudio recolectamos, analizamos y clasificamos opiniones extraídas de usuarios de Twitter, comentarios de blogs especializados en finanzas, artículos periodísticos en economía y finanzas –que constituirá nuestro corpora ampliado−, aplicando principios y técnicas de Sentiment Analysis y Machine Learning.

Palabras clave: Sentiment Analysis, Machine Learning, Mercados Financieros Argentinos, Computational Linguistics, POS Tagging.

Contexto

El presente proyecto, cuyo inicio es Marzo 2016, 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). Se enmarca dentro de la línea de Algoritmos y Software y continúa las investigaciones iniciadas en el Proyecto "Modelo de Sentiment Analysis para la clasificación de noticias en tiempo real". 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 y de la Diplomatura en Análisis de Datos para Negocios, Finanzas e Investigación de Mercados.

Introducción

Como hemos observado en nuestro estudio anterior, identificar y extraer información subjetiva de comentarios en las redes sociales (principalmente, Twitter) para el estudio de los mercados financieros constituye, en la actualidad, un tema de radical interés para desarrollar herramientas analíticas que brindan la posibilidad a analistas

financieros de usar estas opiniones para mejorar la precisión de las predicciones del mercado. De la misma manera hemos hecho referencia al surgimiento de nuevas formas de realizar transacciones financieras mediante el *Trading Algorítmico* (AT) o trading automático y el estrecho vínculo con *Sentiment Analysis*, siendo lo más notable de esta tecnología la capacidad de un algoritmo de detectar e interpretar semánticamente una noticia o acontecimiento en el mismo instante que se está produciendo y así conocer su correlación con los movimientos de precios con el Mercado de Valores. Ahora bien, todo esto ha sido explicado considerando aspectos esenciales de la clasificación de textos determinando su tendencia positiva o negativa desde su *estructura o composición semántica contextualmente dependiente*. Así, hemos mostrado cómo estos aspectos semánticos del lenguaje en un contexto o dominio específico son fundamentales para generar modelos de *Machine Learning* con buenos índices de certeza para clasificar tweets relacionados a finanzas en los mercados argentinos, en positivos o negativos.

Por otra parte, creemos que es imprescindible dar continuidad al estudio y realizar el tratamiento de nuestra corpora incorporando el nivel sintáctico, si bien las tecnologías o modelos basados en semántica han revolucionado no sólo la forma en que integra y se le asigna una valoración a la información sino también a los resultados en el análisis de la polaridad.

El modelo de procesamiento de lenguaje natural (PLN) que aquí proponemos consiste en aplicar un *etiquetador morfosintáctico* con el fin de identificar cómo una determinada secuencia de palabras, a partir de una estructura sintáctica, refleja un indicador de sentimiento, esto es, clasificar una cláusula en positivo, negativo o neutro−nos extendemos a textos completos que contienen una carga valorativa como artículos periodísticos en finanzas y mercados (“noticias duras”), comentarios en blogs especializados en finanzas y notas de opinión general asociadas a la temática que estamos analizando. En una primera etapa, evaluaremos los etiquetadores ya existentes y su aplicabilidad a nuestra corpora. En una segunda fase determinaremos si es necesario realizar alguna mejora incorporando estructuras sintagmáticas y paradigmáticas de mayor complejidad a aquellas ya disponibles. Hacia la etapa final de la investigación, propondremos un *modelo híbrido enriquecido semánticamente*, es decir, que nos permita no sólo etiquetar sino también establecer una correspondencia con el *lexicón FSAL* (Financial Spanish Dictionary of Affect) y verificar si existe un avance en cuanto al desempeño (performance) de los algoritmos de clasificación mediante la aplicación de este nuevo modelo.

En la literatura existente, encontramos diversos trabajos que documentan el uso de etiquetadores morfosintácticos (POS Tagger, en inglés) para incrementar el desempeño en la clasificación de sentimiento u opiniones de comentarios de consumidores sobre un producto o reseña literaria (Patel y Chang, 2014; Jagtap and Pawar, 2013, Nicholls and Song, 2009). En estudios comparados de etiquetadores MBT, QTAG y MXPSOT de corpora en español, así como en mejora y adaptación de métodos de etiquetado (Martí et al., 2006). Cabe mencionar, también, *El Grial*, una interfaz computacional que permite la realización de anotaciones morfosintácticas en textos planos en lengua española además de la consulta en forma de base de datos de los corpora allí reunidos (Parodi, 2006; Venegas, 2008).

En el ámbito de mercados y finanzas, y más específicamente en el dominio de las noticias en financieras, se encuentran los trabajos de Devitt y Ahmad (2007) para explicar la relación entre el contenido afectivo de los textos y su impacto en el mercado.

Líneas de Investigación, Desarrollo e Innovación

En la actualidad los Mercados Financieros Argentinos se encuentran en una etapa de plena innovación en cuanto a Trading Algorítmico. Cabe señalar que tanto el ROFEX, el MERVAL como el MAV ya cuentan con las plataformas necesarias para operar con esta nueva tecnología. De modo que, todo un ecosistema que une inversores y analistas financieros con desarrolladores, científicos, investigadores en el ámbito académicos están trabajando de manera colaborativa con el propósito de avanzar en este nuevo escenario.

El presente trabajo se enmarca en la investigación de aplicación de técnicas y algoritmos de minería de datos en bases de datos y pretende dar un enfoque y soporte académico a toda la comunidad sea esta financiera como académica. En este estudio, en particular sentamos las bases para aplicar las herramientas de Sentiment Analysis, teniendo en cuenta que las noticias financieras son uno de los pilares fundamentales a la hora de la toma de decisiones por parte de los inversores.

Resultados y Objetivos

Pregunta Problema: Un modelo combinado de análisis morfosintáctico enriquecido semánticamente y enmarcado en un contexto económico-financiero en los mercados argentinos podría arrojar mejores resultados al momento de determinar la polaridad de un texto, que si consideráramos estos mismos modelos de manera autónoma e independiente del contexto de aplicación.

Objetivo General:

Avanzar en los aspectos sintácticos de Sentiment Analysis, como el etiquetado morfosintáctico (en inglés, Part Of Speech Tagging), brindándole mayor complejidad lingüística a nuestros modelos y una aproximación más formal a la Inteligencia Artificial aplicada a las Finanzas, especialmente en los Mercados Argentinos.

Objetivos Específicos:

a. Adentrarse en los aspectos Sintácticos de la Lingüística Computacional creando modelos de Sentiment Analysis basados en POST (Part of Speech Tagging).

b. Entrenar diversos Algoritmos Estadísticos para la clasificación de Noticias en "Positivas", "Negativas" o "Neutras", de acuerdo a los modelos del punto anterior.

c. Crear señales de compra/venta de instrumentos financieros basados en los resultados obtenidos en nuestros modelos y correlacionar dichos resultados con valores en tiempo real de los mercados argentinos.

d. Desarrollar diferentes Indicadores de Sentimiento y correlacionarlos con diferentes indicadores financieros, como por ejemplo Volatilidad o Volumen.

e. Estudiar las posibilidades de crear un modelo híbrido entre los resultados obtenidos en la etapa anterior del proyecto y los nuevos objetivos, más específicamente un modelo híbrido entre los aspectos Semánticos y Sintácticos del Lenguaje en modelos de Sentiment Analysis.

f. Ampliar el corpus en referencia de la etapa anterior, esto es, aquel que se utilizó en el desarrollo de la investigación 2013-2015. De modo que, no sólo se trabajará

con tweets sino también con artículos periodísticos financieros, comentarios de blogs especializados en finanzas y notas de opinión general relacionadas a la temática propuesta en este proyecto.

Metodología de Trabajo

El enfoque teórico-metodológico del presente trabajo se basa en aplicar herramientas de Natural Language Processing y Machine Learning desarrolladas en los lenguajes de programación R y Python y descripto en los siguientes puntos:

1. Con un corpus constante, es decir, mantenemos el corpus del proyecto anterior en una fase inicial, esto es 800 tweets clasificados en POSITIVOS y NEGATIVOS de manera manual por una comunidad de expertos en finanzas.
2. Se realiza un estudio de diferentes herramientas Open Source de análisis del lenguaje: Freeling, openNLP, NLTK y Stanford Log Linear Part-Of-Speech Tagger para evaluar los resultados que se obtienen en conjunto y con cada uno de ellos.
3. Se crean modelos de Sentiment Analysis basado en Part of Speech Tagging (Etiquetador Morfosintáctico) y los clasificadores Naive Bayes y Random Forest. Los mismos se implementan en la plataforma estadística R y la librería CARET.
4. Se comparan los resultados entre la fase semántica obtenida anteriormente y la fase morfosintáctica de nuestra presente investigación.
5. Se evalúa adecuar las herramientas Open Source utilizadas en base a los resultados obtenidos. Esto es, mejorar manualmente las herramientas analizadas en el punto (2).
6. Se propone la creación de un modelo combinado que incluya los niveles de semántica-morfosintáctica para enriquecer nuestro modelo de Sentiment Analysis aplicado a finanzas.

Resultados Esperados

En una primera instancia esperamos que las clasificaciones arrojadas por nuestros algoritmos, sin un tratamiento previo de las herramientas descriptas en el punto (2), al que se refiere en la sección Metodología de Trabajo, tengan una performance inferior a los resultados obtenidos a nivel semántico en el estudio llevado a cabo durante 2013-2015. Sin embargo, consideramos que estos resultados pueden tener un desempeño superior si perfeccionamos manualmente los sistemas analizadores mencionados.

Por último, en un estadio más avanzado, nos proponemos crear un modelo combinado que muestre un mayor grado de precisión al integrar los niveles de semántica y morfosintaxis, que incrementen el grado de fiabilidad de las anotaciones morfosintácticas además de producir etiquetas en español que sean transparentes y acertadas en su nominación, y que superen el resultado de los modelos aplicados por separado.

Formación de Recursos Humanos

El equipo del proyecto, multidisciplinario, se compone principalmente de docentes de 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 Lingüística y Finanzas. Por su parte, señalamos que el proyecto contará con la participación de 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. Además se cuenta con la colaboración de alumnos de la mencionada Diplomatura.

Referencias

Braña, J.P; Litterio; A.; Camós, C. y Fernández, A. (2015). Modelo de Sentiment Analysis para la clasificación de noticias en tiempo real en el Mercado de Valores de Buenos Aires. En *XVII Workshop de Investigadores en Ciencias de la Computación*, Salta. Disponible en: http://sedici.unlp.edu.ar/handle/10915/45547

Devitt, A. and Ahmad, K. (2007). "Sentiment Polarity Identification in Financial News: A Cohesion-based Approach", *Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic.

Enriquez, F.; Troyano, J.; Cruz, F. y Ortega, J. (2006). Ampliación automática de corpus mediante la colaboración de varios etiquetadores. En *Procesamiento del Lenguaje Natural*, Nro. 37, pp. 6 -11.

Genereux, M., Poibeau, T. and Koppel, M. (2008). "Sentiment analysis using automatically labeled financial news items". *LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology*, Marrakech, Morocco.

Hernández, S., Miranda, S., Jiménez, E., Villaseñor, E., Tellez, S. y Graff, M. (2015). "Minería de opinión en blogs financieros para la predicción de tendencias en mercados bursátiles". *Research in Computing Science*, Vol.92, pp. 101-109.

Jagtap, V. S. and Pawar, K. (2013). "Analysis of different approaches to Sentence-Level Sentiment Classification", *International Journal of Scientific Engineering and Technology*, Vol. 2, pp. 164-170.

Martí, J. et al. (2006). Adaptación del Método de Etiquetado No Supervisado TBL. En *Procesamiento del Lenguaje Natural*, Nro. 37, pp. 3-5.

Morales de Jesús, V. M. (2014). *Utilización de expresiones de actitud para el Análisis de Sentimientos*. Tesis de Licenciatura, Benemérita Universidad Autónoma de Puebla, México.

Nicholls, C. and Song, F. (2009). "Improving sentiment analysis with Part-of-Speech weighting". Department of Computation & Information Science, University of Guelph, Guelph, ON, Canada.

Parodi, G. (2006). "El Grial: Interfaz Computacional Para Anotación e Interrogación de Corpus en Español". En *Revista de Lingüística Teórica y Aplicada*, 44 (2), II Sem., pp. 91-115.

Patel, N. D. et al. (2014). "Selecting Best Features Using Combined Approach in POS Tagging for Sentiment Analysis", *International Journal of Computer Science and Mobile Computing*, Vol.3 Issue.3, pp. 425-430.

Peifeng, L., Qiaoming, Z. and Wei, Z. (2011). "A Dependency Tree Based Approach for Sentence-Level Sentiment Classification". *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD),12th ACIS International Conference.* pp, 166-171.

Venegas, R. (2008). Interfaz Computacional de Apoyo al Análisis Textual: "El Manchador de Textos". En *Revista de Lingüística Teórica y Aplicada*, 46 (2), II Sem., pp. 53-79.