quantil

Matemáticas Aplicadas

Predictibilidad de Bolsa con Minería de Texto

Diego Jara

Plan de la presentación

- Motivación y Enfoque de Charla
- Tipos de Datos
- Herramientas Lingüísticas y Estadísticas
- Revisión Literaria de Resultados
- Ideas de Aplicación

Plan de la presentación

- Motivación y Enfoque de Charla
- Tipos de Datos
- Herramientas Lingüísticas y Estadísticas
- Revisión Literaria de Resultados
- Ideas de Aplicación

Motivación

- Estamos trabajando en un modelo de predicción de precios de Bolsa usando tweets
- Un modelo exitoso podría implementarse mediante QUANTIL AUTÓMATA
- Alcance de esta charla: hacer una revisión literaria de predicción de Bolsa usando datos en forma de texto
  - No es una charla matemática
- Se buscan técnicas y resultados obtenidos

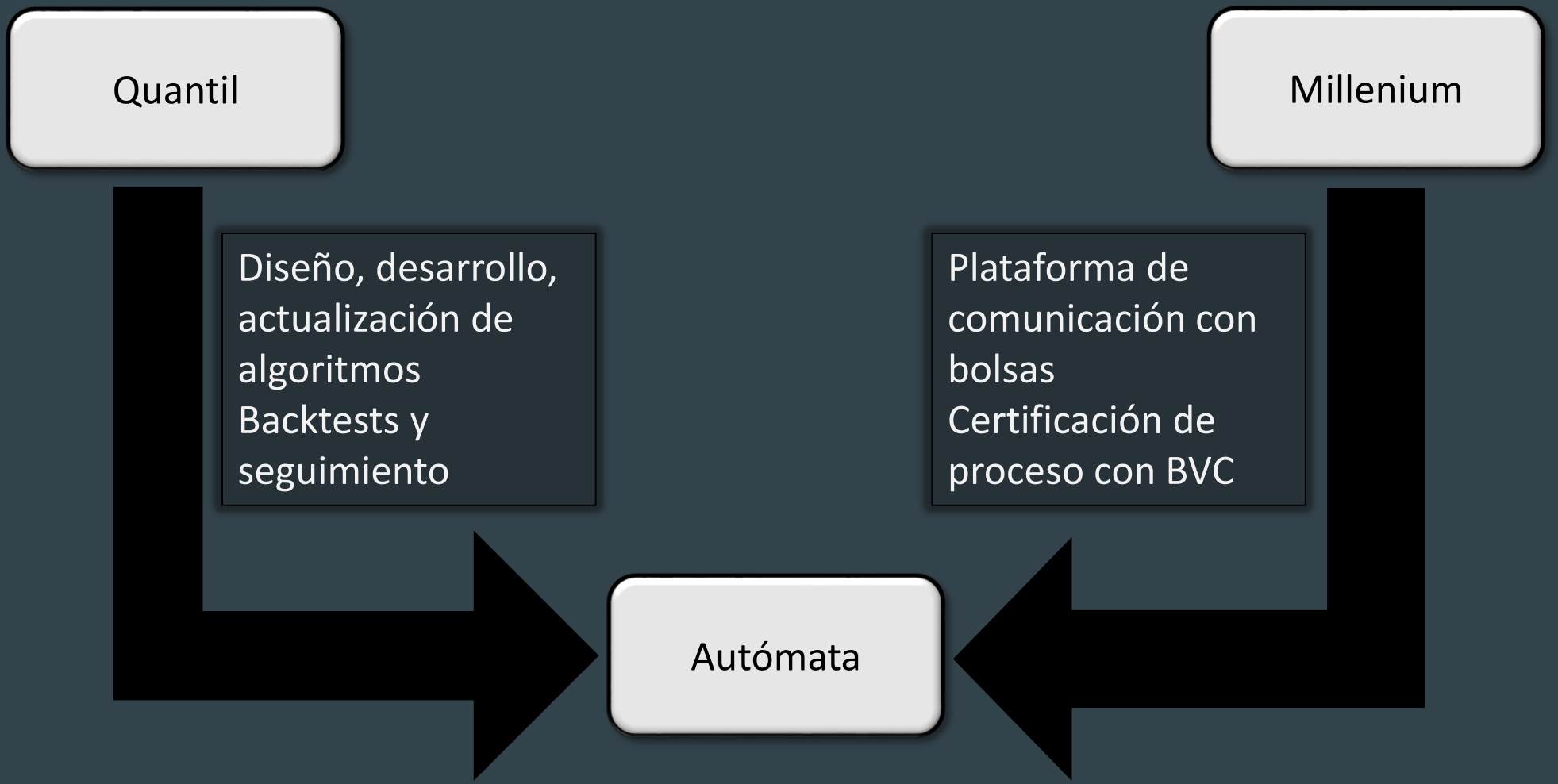
Qué es Quantil Autómata?

Autómata: Dispositivo o conjunto de reglas que realizan un encadenamiento automático y continuo de operaciones capaces de procesar una información de entrada para producir otra de salida. Diccionario de la lengua española © 2005 Espasa-Calpe

Quantil Autómata

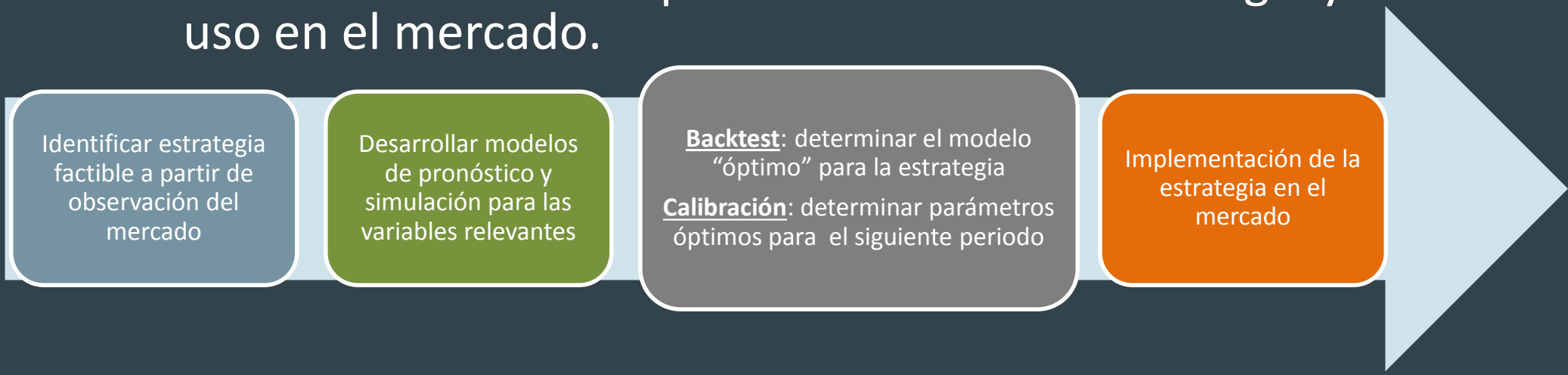
- S.A.S. que busca ofrecer servicios que integran
  - Ambiente tecnológico de comunicación de algoritmos con sistemas transaccionales
  - Algoritmos transaccionales que operen en los mercados de forma automática
- Oferta de Valor: Diseño, Desarrollo, Implementación técnica y tecnológica en manos de socios

Qué es Quantil Autómata?



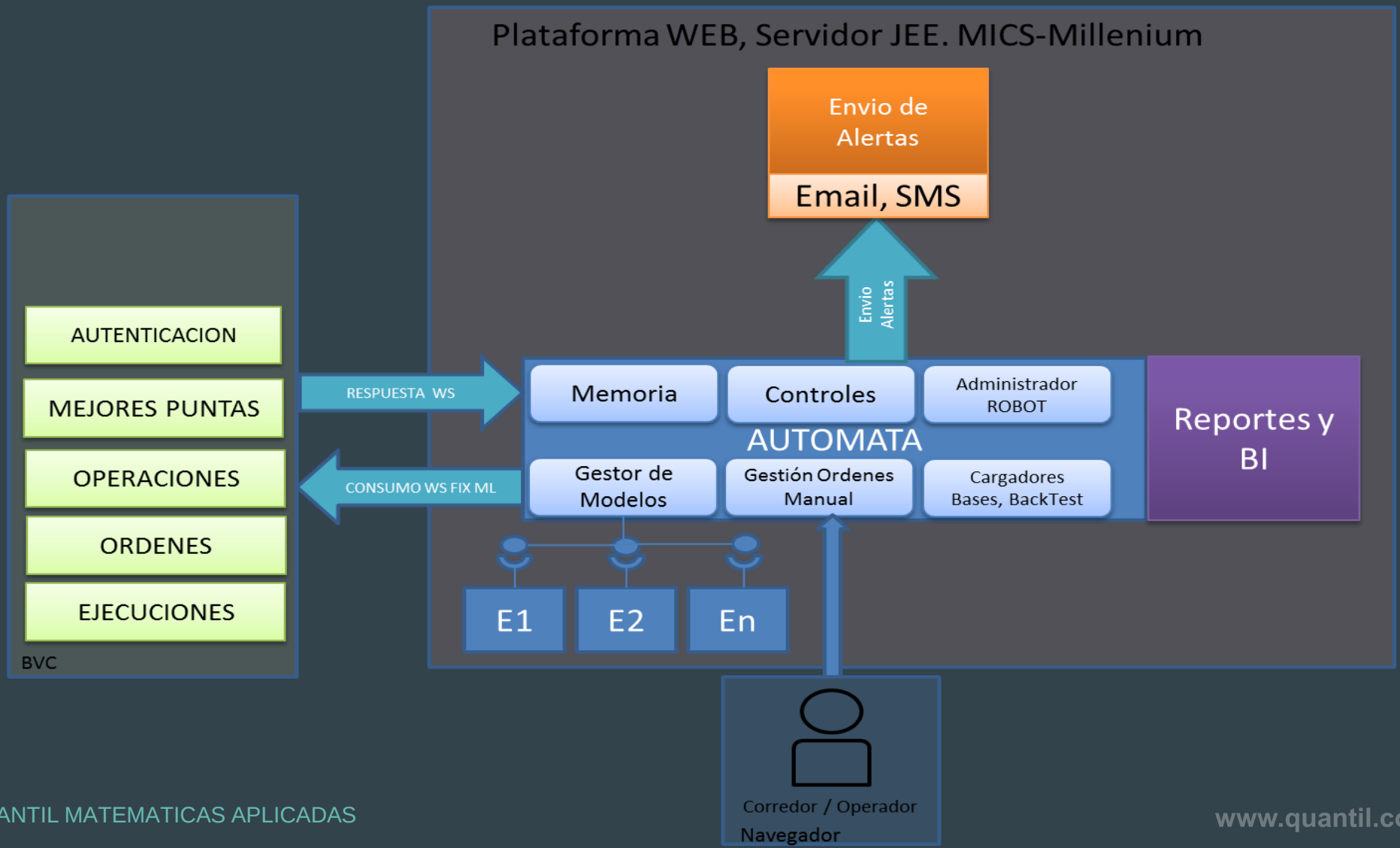
Qué es Quantil Autómata?

- Elementos de desarrollo:
  - Estrategia: conjunto de parámetros y reglas que busca sacar provecho a comportamientos recurrentes del mercado.
  - Modelo: objeto matemático que permite pronosticar (p.ej., mediante simulaciones) la evolución de variables de interés.
  - Algoritmo: implementación computacional de la determinación de los parámetros de una estrategia y de su uso en el mercado.



Qué es Quantil Autómata?

➢ Elementos de implementación



Qué es Quantil Autómata?

- **Mercados de aplicación: en general, sistemas que permitan conexión**
  - Mercado local
    - TES
    - Acciones (incluido iColcap)
    - Derivados estandarizados (futuros: COLCAP, acciones, TES, TRM)
  - Mercado externo
    - ETFs
    - Acciones - ADRs
    - Derivados estandarizados

Qué es Quantil Autómata?

- **Herramientas cuantitativas**:
  - Variables de entrada
    - Promedios móviles
    - Cambios y rezagos
    - Volumen
    - Tiempo
    - Relaciones entre precios de distintos productos
    - ***Redes sociales***
  - Modelos de predicción
    - Reversión a la media
    - Desviación de precios “correctos” (arbitraje estadístico o real)
    - Técnicas de Aprendizaje de máquinas
      - Support Vector Machines
      - Redes neuronales
  - Estrategias de trading
    - Backtests
    - Calibración de parámetros
    - Definición de señales de entrada y salida
    - Agregación eficiente de estrategias

Qué es Quantil Autómata?

Tipos de algoritmos que pueden ofrecerse

- Especulativos: buscan anticipar el movimiento del Mercado para “comprar barato y vender caro”
  - Direccionales
  - Pairs / Switches / Spreads / Mariposas
- Arbitraje: buscan capturar desviaciones de precios justos
  - Derivados vs subyacentes: p.ej., TES, acciones
  - Canastas vs components: p.ej., COLCAP
  - Acciones locales vs. ADRs
- Órdenes de clientes: buscan satisfacer restricciones de ejecución
  - “observar” continuamente el Mercado y ejecutar cuando se llegue a niveles dados: activos individuales o switches
  - Órdenes estilo VWAP
- Promoción de liquidez: buscan satisfacer restricciones de postura de puntas

Tipos de horizonte

- Corto plazo: minutos para convergencia
- Mediano plazo: horas
- Largo plazo, días
- No se busca algo mayor a una o dos semanas

Otros

- **Desarrollados**
  - TendenciaID (TES)
  - SpreadCruces (TES)
  - ArranqueDia (TES)
  - PicosIntradia (TES)
  - Arbitraje ADRs
- **En etapa avanzada de desarrollo**
  - AccionesIntradia: predicción de puntas de bid ask con Support Vector Machines
  - Futuros TES Nocional
  - Futuros Acciones vs. Subyacente
- **En etapa de desarrollo preliminar**
  - VWAP (Volume Weighted Average Price)
  - RezagosTES: aplicación de random forests para identificar patrones
- **Por estudiar y desarrollar**
  - Promoción liquidez
  - Futuro COLCAP
  - Valor Relativo en TES

Plan de la presentación

- Motivación y Enfoque de Charla
- Tipos de Datos
- Herramientas Linguísticas y Estadísticas
- Revisión Literaria de Resultados
- Ideas de Aplicación

Datos usados en papers revisados

- Texto
  - Tweets
    - En ocasiones, filtrando según ciudad de origen del tweet (??)
    - Aplicable en Colombia
  - Mensajes de usuarios de Yahoo Finance
    - En Colombia no hay un equivalente de un site donde la gente comente continuamente (DataIFX pretende algo similar)
  - Comunicados de Agencias de Noticias
    - Aplicable en Colombia
  - Reportes a reguladores (10K)
    - Aplicable en Colombia

Datos usados en papers revisados

- Financieros
  - Acciones, ETFs e índices
    - Cambios diarios (también cierre a apertura y apertura a cierre)
    - Cambios en 20 minutos
    - Retornos diarios, 3 días, semanales, mensuales
      - Absolutos
      - Exceso sobre índices
    - Precios de cierre, apertura, máximo, mínimo diario
    - Volumen diario
    - Volatilidad diaria
  - VIX

Plan de la presentación

- Motivación y Enfoque de Charla
- Tipos de Datos
- **Herramientas Lingüísticas y Estadísticas**
- Revisión Literaria de Resultados
- Ideas de Aplicación

Herramientas usadas en papers revisados

- Lingüísticas
  - Limpieza de texto
    - Normalización → crear parejas (OOV, IV) … las palabras “Out Of Vocabulary” emparejarlas con “In Vocabulary”
      - IV se define con diccionarios (Aspell)
      - Se puede pagar a Amazon Mechanical Turk para emparejar
      - OOV se obtiene de corpus de tweets como las que no están en IV
    - Eliminación de excesos
      - Letras repetidas
      - Reemplazar usuarios particulares (@xxx) por USUARIO
      - Reemplazar weblinks (http://xxx) por LINK
      - Reemplazar palabras de negación por NEGACION
      - Remover Stop Words (palabras “ociosas”) … riesgo de alterar signficado del texto
      - Agrupar palabras por raíces comunes
      - Eliminar palabras muy poco usadas

Herramientas usadas en papers revisados

- Lingüísticas
  - Filtro de mensajes
    - Considerar solo los que tengan ciertos keywords (relacionados con empresas – productos, gente clave, ….)
    - Considerar el contexto de estas keywords

OJO PARA RUBICAM: Named Entity Recognition (NER) system para los keywords: se toman las palabras vecinas (-2, -1, +1, +2), y se usa un modelo lineal de Conditional Random Fields (CRF) para determinar si el keyword está en el contexto de la compañía analizada

Los CRF son modelos de probabilidad condicional de características (de textos) sobre información de entidades

Básicamente se modela la probabilidad de que un texto con un keyword se refiera a la marca buscada, condicionando sobre secuencias de palabras alrededor de la keyword

Herramientas usadas en papers revisados

- Lingüísticas
  - "*Features*" de los mensajes
    - Palabras de forma independiente (Bag of Words)
    - N-gramas
      - Colección ordenada de palabras; frases (problema de dimensinoalidad; cómo definirlas?)
    - Otros símbolos
      - Emoticones
      - Símbolo de exclamación

Herramientas usadas en papers revisados

- Lingüísticas
  - Marcación
    - Rara vez se hace análisis no supervisado para aplicar en Bolsa
    - Bag of Words + Word Count + Diccionario
      - Diccionario general (Harvard, por ejemplo)
      - Diccionario de sentimientos específicos
      - Diccionario Financiero
    - Marcaciones de autores de mensajes
      - P.ej., en Yahoo Finance se escribe un mensaje de una acción, y muchas veces viene con "recomendación" (BUY, SELL, …
    - Marcación manual de humanos buscando calificar un sentimiento
    - Cálculo de variables "estadísticas" agregando por empresa

Bullishness: $$B_t = \ln\left(\frac{1 + M_t^{Positive}}{1 + M_t^{Negative}}\right)$$ Agreement: $$A_t = 1 - \sqrt{1 - \frac{M_t^{Positive} - M_t^{Negative}}{M_t^{Positive} + M_t^{Negative}}}$$

$$\text{Polarity} = \frac{positive - negative}{positive + negative} \qquad \text{RVD1}_t = \frac{M_t^{\text{Revealed Buy}} - M_t^{\text{Revealed Sell}}}{M_t^{\text{Revealed Buy}} + M_t^{\text{Revealed Sell}}} \qquad \text{CLD1}_t = \frac{M_t^{\text{Classified Buy}} - M_t^{\text{Classified Sell}}}{M_t^{\text{Classified Buy}} + M_t^{\text{Classified Sell}}}$$

  - Marcación con el subsecuente movimiento de mercado

Herramientas usadas en papers revisados

- Clasificación/Estadísticas
  - SVMs
    - Es la herramienta más usada en este contexto de Bolsa
  - Árboles de Clasificación
  - Naïve Bayes
    - Hay labels (sentimiento del tweet), y features (palabras, por ejemplo)
    - Suponiendo independencia entre features,

$$\mathrm{P}[\mathrm{label} \mid \mathrm{features}] = \frac{\mathrm{P}[\mathrm{label}]}{\mathrm{P}[\mathrm{features}]} \times \mathrm{P}[\mathrm{f}_1 \mid \mathrm{label}] \times \ldots \times \mathrm{P}[\mathrm{f}_N \mid \mathrm{label}]$$

  - Maximum Entropy (como Naive Bayes, relajando independencia)
  - *Fuzzy* Neural Networks
  - Tradicionales (regresión – OLS, correlación, modelos econométricos, mirar causalidad, por acción y transversal, …)
  - Software disponible
    - OpinionFinder de U Pitt
    - Google-Profile of Mood States
    - CMU POS (Part of Speech) Tagger
    - Twitter Sentiment Tool (TST)
  - Evaluación con medidas típicas
    - Recall: $\frac{VP}{PosReal}$, Precisión: $\frac{VP}{PosModelo}$, F: $\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$

Plan de la presentación

- Motivación y Enfoque de Charla
- Tipos de Datos
- Herramientas Lingüísticas y Estadísticas
- **Revisión Literaria de Resultados**
- Ideas de Aplicación

Metodologías y Resultados en Revisión Literaria Selecta

Vu Chang Ha Collier

- Datos de mensajes – Tweets
  - API de Twitter (solo acceden al 1% del corpus)
  - 5 millones de tweets diarios (abril 1 a mayo 31 de 2011)
  - Determinan keywords de cada empresa con Google Adwords (no me es claro cómo se hace)
  - Solo tweets de NY, LA, SF y Chicago (cómo hacen esto?)
  - Normalización y limpieza profunda de tweets
  - Consideración de productos: eliminación de tweets según contexto (linear CRF)
- Variables de predicción
  - Cambios en la acción en t-3, t-2, t-1
  - Positivo vs Negativo
    - Twitter Sentiment Tool agregando léxico de emoticones
    - Pos_Neg = Cambio en Positivo – Cambio en Negativo (de t-1 a t)
  - Bulish vs Bearish
    - Usan AltaVista para evaluar cercanía de palabras a “bullish” o “bearish”

$$SO(w) = \log_2\left(\frac{\#(wNEAR\text{“}bullish\text{”})\#(\text{“}bearish\text{”})}{\#(wNEAR\text{“}bearish\text{”})\#(\text{“}bullish\text{”})}\right)$$

    - Para cada día suman sentimiento Bullish – sentimiento Bearish
    - Bullish/bearish = Cambio en este sentimiento neto (de t-1 a t)
- Datos de Mercados
  - GOOG, APPL, MSFT, AMZN (solo estas tienen muchos tweets de sus productos)
  - Solo miran 40 días de cambios (+ o -) de cierre a apertura

Metodologías y Resultados en Revisión Literaria Selecta

Vu Chang Ha Collier

- Solo se considera el signo de las variables. No la magnitud
- Usan Árboles de decisión; 10-fold cross validation
- Analizan tema de causalidad: consideran información rezagada 1 día de Pos_Neg
- Resultados

Method	AAPL	GOOG	MSFT	AMZN
Only bullish/bearish	53.66%	58.54%	56.10%	37.50%
Only previous days	51.22%	53.66%	73.73%	37.50%
Only Pos_Neg	73.17%	68.29%	56.10%	71.88%
Bullish/bearish + previous days	63.41%	63.41%	**75.61%**	62.50%
Bullish/bearish + Pos_Neg	73.17%	70.73%	56.10%	71.88%
Pos_Neg + previous days	73.17%	68.29%	70.73%	71.88%
Pos_Neg + bullish/bearish + previous days	**82.93%**	**80.49%**	**75.61%**	**75.00%**

Table 4: Prediction accuracy on each stock

Metodologías y Resultados en Revisión Literaria Selecta

Kim Kim

- Datos de mensajes – Posts en Yahoo Finance
  - Con Python bajan mensajes de 91 boards (un board por cada acción/empresa)
  - 32 millones de mensajes (2005-2010) escritos por 550 mil autores
  - 26% de mensajes recomiendan STRONG BUY, BUY, HOLD, SELL, o STRONG SELL
  - Usando Naive Bayes, entrenan una máquina para tener recomendaciones (usan bag of words, así que cada palabra es un "feature", y la recomendación es el "label")
- Variables de predicción

$$\mathrm{RVD1}_t = \frac{M_t^{\text{Revealed Buy}} - M_t^{\text{Revealed Sell}}}{M_t^{\text{Revealed Buy}} + M_t^{\text{Revealed Sell}}} \qquad \mathrm{RVD2}_t = \ln\left[\frac{1 + M_t^{\text{Revealed Buy}}}{1 + M_t^{\text{Revealed Sell}}}\right]$$

$$\mathrm{CLD1}_t = \frac{M_t^{\text{Classified Buy}} - M_t^{\text{Classified Sell}}}{M_t^{\text{Classified Buy}} + M_t^{\text{Classified Sell}}} \qquad \mathrm{CLD2}_t = \ln\left[\frac{1 + M_t^{\text{Classified Buy}}}{1 + M_t^{\text{Classified Sell}}}\right]$$

  - Agregan para distintos periodos (mes, semana, día)
  - Incluyen volatilidad y volumen
  - Controlan por tamaño de firma y book-to-market ratio
- Resultados
  - Con tests de causalidad de Granger, no encuentran evidencia de que el sentimiento de inversionistas afecte retornos de acciones (individual o agregado)
  - Sí ven evidencia de que el sentimiento de inversionistas se afecta por retornos de acciones

Metodologías y Resultados en Revisión Literaria Selecta

Li Xie Chen Wang Deng

- Datos de mensajes – Noticias
  - Agencia grande de noticias, que relaciona la noticia con símbolo de acción
  - Noticias se agregan por firma, y se crea matriz de frecuencia de palabras (filas = noticias, columnas = palabras)
  - Se usan diccionarios de sentimientos (Harvard IV-4 – HVD – y LMD): Positivo, Negativo, Incierto, Litigioso, Fuerte Modal, Débil Modal
  - Quedan matrices de frecuencia de sentimientos por noticia (filas = noticias, columnas = sentimientos)
- Variables de predicción $\text{Polarity} = \frac{positive - negative}{positive + negative}$
- Datos de Mercados
  - Acciones en Hong Kong Stock Exchange y Hang Seng Index
  - Datos OHLC (Open, High, Low, Close), y "retornos" a partir de éstos (positivo, estable o negativo); 5 años
- Resultados
  - Incluir análisis de sentimiento con diccionarios mejora el simple uso de bag of words
  - Solo Positivo y Negativo no mejora predictibilidad
  - LMD (diccionario financiero) se comporta mejor que HVD (diccionario general)

Metodologías y Resultados en Revisión Literaria Selecta

Loughran McDonald

- Datos de mensajes – Reportes Contables
  - Reportes contables 10k a accionistas
  - 50 mil reportes de 8 mil firmas (1994 a 2008)
  - Miden frecuencia de palabras
  - Excluyen tablas
  - Consideran reportes de empresas en problemas (con litigación importante, fraude, debilidad en control interno, …)
  - Autores desarrollan diccionario financiero LMD para seis sentimientos: Positivo, Negativo, Incierto, Litigioso, Fuerte Modal (para nivel de confianza; cosas como always, highest, must, will), Débil Modal (cosas como could, depending, might, possibly).
- Variables de Análisis
  - Conteo de palabras según diccionarios (el de ellos, y el de Harvard)
  - Se normaliza por recurrencia de términos escogidos (“loss” aparece 1.7 millones de veces y “aggravates” 10)
- Datos de Mercados
  - Retorno de las acciones en un periodo de tres días después de emitido el reporte
- Resultados/Conclusiones
  - 75% de palabras negativas (según diccionario de Harvard) no son negativos en contexto financiero (ej. depreciation, foreign, liability)
  - En regresión hay significancia, pero obviamente advierten en usar estos resultados para predecir el mercado

Metodologías y Resultados en Revisión Literaria Selecta

Loughran McDonald

Comparison of Negative Word Lists Using Filing Period Excess Return Regressions

The dependent variable in each regression is the event period excess return (defined as the firm's buy-and-hold stock return minus the CRSP value-weighted buy-and-hold market index return over the 4-day event window, expressed as a percent). The proportional weights are the word list counts relative to the total number of words appearing in a firm's 10-K. The tf.idf weighted values are defined in equation (1) of the text. See the Appendix for the other variable definitions. Fama-French (1997) industry dummies (based on 48 industries) and a constant are also included in each regression. The coefficients are based on 60 quarterly Fama-MacBeth (1973) regressions with Newey–West standard (1987) errors using one lag. The estimates use a sample of 50,115 10-Ks over 1994 to 2008.

	Proportional Weights		tf.idf Weights	
	(1)	(2)	(3)	(4)
Word Lists				
H4N-Inf (Harvard-IV-4-Neg	−7.422		−0.003	
with inflections)	(−1.35)		(−3.16)	
Fin-Neg (negative)		−19.538		−0.003
		(−2.64)		(−3.11)
Control Variables				
Log(size)	0.123	0.127	0.131	0.132
	(2.87)	(2.93)	(2.96)	(2.97)
Log(book-to-market)	0.279	0.280	0.273	0.277
	(3.35)	(3.45)	(3.37)	(3.41)
Log(share turnover)	−0.284	−0.269	−0.254	−0.255
	(−2.46)	(−2.36)	(−2.32)	(−2.31)
Pre_FFAlpha	−2.500	−3.861	−5.319	−6.081
	(−0.06)	(−0.09)	(−0.12)	(−0.14)
Institutional ownership	0.278	0.261	0.254	0.255
	(0.93)	(0.86)	(0.87)	(0.87)
NASDAQ dummy	0.073	0.073	0.083	0.080
	(0.86)	(0.87)	(0.97)	(0.94)
Average R^2	2.44%	2.52%	2.64%	2.63%

Metodologías y Resultados en Revisión Literaria Selecta

Schumaker Chen

- Usan tres representaciones lingüísticas: Bag of Words, Noun Phrases, Named Entities
- Estudian impacto de noticas a un horizonte de 20 minutos
- Nov 2005 de datos de noticias (Yahoo). Solo se toma una noticia cada 20 minutos.
- SVM, con 10-fold cross validation.
- Tres métricas:
  - Closeness (MSE de diferencia entre predicción y actual)
  - Directional Accuracy (Accuracy de arriba o abajo)
  - Simulated Trading Engine (backtest de estrategia de comprar o vender si predicción es más de 1% de movimiento).
- Se comparan contra una regresión lineal de tendencia del mercado.

Schumaker Chen

MSE Analysis	Regression	Our Model
Bag of Words	0.07253	0.04713
Noun Phrases	0.07257	0.05826
Named Entities	0.07244	0.03346

Table 4. MSE analysis of the data models

Directional Accuracy	Regression	Our Model
Bag of Words	47.6%	49.3%
Noun Phrases	47.6%	50.7%
Named Entities	47.4%	49.2%

Table 5. Direction Accuracy of the data models

Simulated Trading	Regression	Our Model
Bag of Words	-$1,809	$5,111
Noun Phrases	-$1,809	$6,353
Named Entities	-$1,879	$3,893

Table 6. Simulated trading engine on the data models

Metodologías y Resultados en Revisión Literaria Selecta

Ñapa: Metodología Hu Tang Gao Liu

- El contexto no es financiero; desarrollan un esquema para análisis de sentimiento semi-supervisado con tweets
- Construyen matriz X_{mxn} de mensajes y palabras (X_{ij} es frecuencia de palabra j en tweet i)
- Definen un número de sentimientos (pero no definen los sentimientos). P.ej., c=3
- Buscan resolver $\min_{\mathbf{U},\mathbf{H},\mathbf{V}\geq 0} \quad \mathcal{O} = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2,$

$$s.t. \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}.$$

- Teorema: esto es igual a "clusterizar" tweets (U_{mxc}) y palabras (V_{nxc}) usando k-means (con métrica definida como distancia de vectores en R^n)
- H_{ij} resulta ser el número de palabras del cluster j en el cluster de tweets i (normalizado por raíz del tamaño de estos dos clusters)
- Después incorporan emoticones para penalizar por desviaciones de tweets que tienen emoticones similares
- $$\min_{\mathbf{U},\mathbf{H},\mathbf{V}\geq 0} \quad \mathcal{J} = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2 + \lambda_I^u\|\mathbf{G}^u(\mathbf{U} - \mathbf{U}_0)\|_F^2$$
- $$+ \lambda_I^v\|\mathbf{G}^v(\mathbf{V} - \mathbf{V}_0)\|_F^2 + \lambda_C^u Tr(\mathbf{U}^T\mathcal{L}^u\mathbf{U}) + \lambda_C^v Tr(\mathbf{V}^T\mathcal{L}^v\mathbf{V})$$

$$s.t. \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}.$$

U0 es matriz de tweets con sentimiento de emoticones
V0 es matriz de sentimiento de palabras
G corrigen para tweets (o palabras) sin sentimiento reconocido

Metodología Hu Tang Gao Liu

- Expresiones resaltadas son

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\mathbf{U}(i,*)-\mathbf{U}(j,*)\|_2^2\mathbf{W}^u(i,j)$$

- Donde W es una matriz indicadora de pertenecer al mismo cluster (Wij es 1 si los tweets i y j están en el mismo cluster)

Plan de la presentación

- Motivación y Enfoque de Charla
- Tipos de Datos
- Herramientas Lingüísticas y Estadísticas
- Revisión Literaria de Resultados
- Ideas de Aplicación

Ideas para una máquina local

- ✓ Datos financieros: acciones líquidas (Ecopetrol, PREC, PfBancolombia); de pronto TES
- ✓ Mensajes de tweets: ya está QuantTweet recogiendo mensajes
- ✓ Horizonte de corto plazo (revisar impacto a los 20 minutos, una hora); la intención es ver si hay posibilidad de explotar la rapidez de reacción, más que la dirección de mediano plazo
- ✓ Considerar relevancia de tweet (popularidad del twittero, por ejemplo)
- ✓ Para evitar mucha limpieza de datos, de pronto filtrar y usar solo twitteros institucionales
- ✓ Igual, limpiar y normalizar tweets; filtrar por relevancia, estilo Vu Chang Ha Collier
- ✓ Creación de keywords (ya empezamos, pero más difícil de lo que suena)
- ✓ Puede ser intentar evaluar el sentimiento de los tweets y evaluar predictibilidad en el mercado
  - Diccionarios en español!
  - Diccionarios financieros?
  - Estructura lingüística? (estilo Stanford)
- ✓ Puede ser una idea como Hu et. al., donde los emoticones se reemplazan por el subsecuente retorno del mercado

GRACIAS