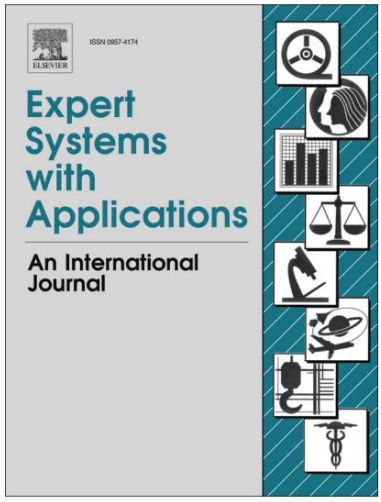
manuscrito aceptado



Revisar

Minería de texto para la predicción del mercado: una revisión sistemática

Arman Khadjeh Nassirtoussi, Teh Ying Wah, Saeed Reza Aghabozorgi, David Ngo Chek Ling

IIP: S0957-4174(14)00345-5
DOI: http://dx.doi.org/10.1016/j.eswa.2014.06.009 ESWA
Referencia: 9376

Aparecer en: Sistemas Expertos con Aplicaciones

Cite este artículo como: Nassirtoussi, AK, Wah, TY, Aghabozorgi, SR, Ling, DNC, Text Mining for Market Prediction: A Systematic Review, Expert Systems with Applications (2014), doi: http://dx.doi. org/10.1016/j.eswa. 2014.06.009

Minería de texto para la predicción del mercado: una revisión sistemática

Arman Khadjeh Nassirtoussi1*, Teh Ying Wah2 , Saeed Reza Aghabozorgi3 y David Ngo Chek Ling4

[1,2 y 3] Departamento de Ciencias de la Información, Facultad de Ciencias de la Computación y Tecnología de la Información, Universidad de Malasia, 50603 Kuala Lumpur, Malasia

[4] Investigación y títulos superiores, Sunway University, No 5, Jalan University, Bandar Sunway, 46150 Petaling Jaya, Selangor DE, Malasia

* Autor correspondiente. Correo electrónico: armankhnt@gmail.com

Resumen

La calidad de la interpretación del sentimiento en el rumor en línea en las redes sociales y la
Las noticias en línea pueden determinar la previsibilidad de los mercados financieros y causar grandes ganancias o pérdidas.
Es por eso que varios investigadores han puesto toda su atención en los diferentes aspectos de este
problema últimamente. Sin embargo, no existe un marco teórico y técnico completo para
aproximando el problema a lo mejor de nuestro conocimiento. Creemos que la falta existente de tal claridad
sobre el tema se debe a su naturaleza interdisciplinaria que implica en su esencia tanto la economía del comportamiento
temas, así como la inteligencia artificial. Profundizamos en el carácter interdisciplinar y
contribuir a la formación de un marco claro de discusión. Repasamos los trabajos relacionados que están
sobre la predicción del mercado basada en la minería de texto en línea y producir una imagen del genérico
componentes que todos tienen. Nosotros, además, comparamos cada sistema con el resto e identificamos
sus principales factores diferenciadores. Nuestro análisis comparativo de los sistemas se amplía a la
fundamentos teóricos y técnicos detrás de cada uno. Este trabajo debería ayudar a la comunidad investigadora a
estructurar este campo emergente e identificar los aspectos exactos que requieren más investigación y son de
significado especial.

Palabras clave

Análisis de sentimiento en línea; minería de texto en redes sociales; Análisis de Sentimiento de Noticias; Mercado de divisas
Predicción; Predicción de acciones basada en noticias

1. Introducción

El bienestar de las sociedades modernas de hoy depende de sus economías de mercado. En el corazón de cualquier economía de mercado, residen los mercados financieros con sus equilibrios de oferta y demanda. Por lo tanto Es fundamental estudiar los mercados y conocer sus movimientos. Entendiendo los movimientos del mercado principalmente facilita a uno con la capacidad de predecir movimientos futuros. Capacidad de predecir en un mercado. economía es igual a poder generar riqueza evitando pérdidas financieras y haciendo ganancias. Sin embargo, la naturaleza de los mercados es tal que son extremadamente difíciles de predecir, en todo caso.

En general, las medidas predictivas se dividen en análisis técnicos o fundamentales. Ellos son diferenciados en función de sus datos de entrada, con datos históricos de mercado que se utilizarán para el primero y cualquier otro tipo de información o noticias sobre el país, sociedad, empresa, etc. para estos últimos. La mayoría de la investigación en el pasado se ha hecho sobre enfoques de análisis técnico, principalmente debido a la disponibilidad de datos de mercado históricos cuantitativos y el deseo general entre los comerciantes de información técnica Métodos cuantitativos. Los datos fundamentales son más difíciles de usar como entrada, especialmente cuando son desestructurado Los datos fundamentales pueden provenir de fuentes estructuradas y numéricas como macro datos económicos o informes financieros regulares de bancos y gobiernos. Incluso este aspecto de los datos fundamentales rara vez se han investigado; pero en ocasiones se ha demostrado que es de valor predictivo como en los trabajos de Chatrath, Miao, Ramchander y Villupuram (2014), Khadjeh Nassirtoussi, Ying Wah y Ngo Chek Ling (2011) y Fasanghari y Montazer (2010).

Sin embargo, los datos fundamentales disponibles en texto no estructurado son la investigación más desafiante. aspecto y por lo tanto es el foco de este trabajo. Algunos ejemplos serían los datos fundamentales disponible en línea en la información textual en las redes sociales, noticias, blogs, foros, etc.

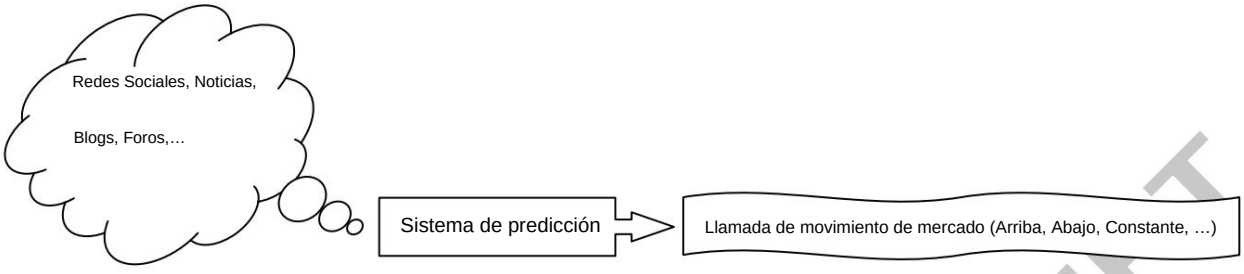


Figura 1 Sistema de predicción sentimental de texto en línea

En este trabajo se realiza una revisión sistemática de los trabajos de investigación pasados con aportes significativos a la Se ha llevado a cabo el tema de la predicción del mercado basada en texto en línea, lo que lleva a la aclaración de la la investigación de vanguardia actual y sus posibles direcciones futuras. Las principales contribuciones de este trabajo en resumen son:

1- Se revisan los conceptos económicos fundamentales y de informática/ciencias de la información pertinentes y se aclara cómo se vinculan con las soluciones actualmente propuestas para este problema de investigación.

2- Se ha revisado la literatura pasada más significativa con énfasis en la más avanzada piezas de trabajo

3- Se identifican los principales factores diferenciadores entre los trabajos actuales, y se utilizan para comparar y contrastar las soluciones disponibles.

4- Se hacen observaciones sobre las áreas con falta de investigación que pueden constituir posibles oportunidades de trabajo futuro.

El resto del artículo se estructura de la siguiente manera. La sección 2 proporciona información sobre la interdisciplinariedad naturaleza del problema de investigación en cuestión y define los conceptos fundamentales necesarios para un comprensión de la literatura. La sección 3 presenta la revisión de los principales trabajos disponibles. Sección 4 hace sugerencias para futuras investigaciones. Y la Sección 5 concluye este trabajo.

2. Una revisión de los conceptos fundamentales de trasfondo interdisciplinario

Esencialmente, el objetivo de esta investigación es la utilización de modelos computacionales y inteligencia artificial para identificar posibles relaciones entre la información textual y la economía. Ese es el problema de investigación.

Para abordar adecuadamente este problema de investigación, se deben considerar al menos tres campos de estudio separados. a saber: Lingüística (para comprender la naturaleza del lenguaje), Aprendizaje automático (para permitir modelado computacional y reconocimiento de patrones), economía del comportamiento (para establecer sentido).

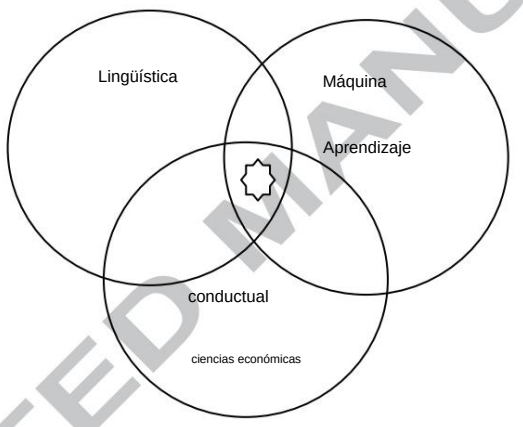


Figura 2 Interdisciplinariedad entre lingüística, aprendizaje automático y economía del comportamiento

La premisa principal está alineada con hallazgos recientes de principios económicos conductuales mediante los cuales las condiciones del mercado son productos del comportamiento humano involucrado (Bikas, Jurevičienė, Dubinskas, & Novickytė, 2013; Tomer, 2007).

Nuestra investigación ha identificado los siguientes temas de fondo como esenciales para desarrollar un sólido comprensión de este problema de investigación:

2.1. Hipótesis del Mercado Eficiente (EMH)

La idea de que los mercados son completamente aleatorios y no son predecibles tiene sus raíces en la eficiencia hipótesis del mercado (Fama, 1965) que afirma que los mercados financieros son "informativamente eficientes"

y que, en consecuencia, uno no puede lograr constantemente rendimientos superiores a los rendimientos promedio del mercado sobre una base ajustada al riesgo, dada la información disponible en el momento de realizar la inversión. Sin embargo, se encuentra que esta hipótesis absoluta no es del todo exacta y el propio Fama la revisa. para incluir 3 niveles de eficiencia como fuerte, semifuerte y débil (Fama, 1970). Esto indica que hay muchos mercados donde la previsibilidad es plausible y viable y dichos mercados se denominan como "débilmente eficiente". La eficiencia del mercado está correlacionada con la disponibilidad de información y un mercado es sólo "fuertemente eficiente" cuando toda la información está completamente disponible, lo que en realidad rara vez es el caso. Por lo tanto, admitió que su teoría es más fuerte en ciertos mercados donde la información es abierta, amplia e instantáneamente disponible para todos los participantes y se debilita cuando tal suposición no puede celebrarse concretamente en un mercado.

2.2. Comportamiento-economía

Los economistas cognitivos y conductuales ven el precio como un valor puramente percibido en lugar de un derivado del costo de producción. Los medios de comunicación no informan únicamente sobre el estado del mercado, sino que activamente crear un impacto en la dinámica del mercado en función de las noticias que publican (Robertson, Geva, & Wolff, 2006; Wisniewski y Lambe, 2013).

La interpretación de la gente de la misma información varía mucho. Los participantes del mercado tienen conocimientos sesgos como exceso de confianza, reacción exagerada, sesgo representativo, sesgo de información y varios otros errores humanos predecibles en el razonamiento y procesamiento de la información (Friesen & Weller, 2006).

Las finanzas conductuales y la teoría del sentimiento de los inversores han establecido firmemente que el comportamiento de los inversores puede ser moldeado por si se sienten optimistas (alcistas) o pesimistas (bajistas) sobre el mercado futuro valores (Bollen & Huina, 2011).

2.3. Hipótesis del Mercado Adaptativo (AMH)

El dilema de la eficiencia del mercado en cuanto a su grado y aplicabilidad a diferentes mercados es sigue siendo un tema de investigación vibrante y en curso con resultados muy contradictorios. Por cada papel

produciendo evidencia empírica que apoya la eficiencia del mercado, un artículo contradictorio quizás pueda
que establece empíricamente la ineficiencia del mercado (Majumder, 2013). Por lo tanto, algunos años
Hace una investigación ha producido una contra-teoría con el nombre de Hipótesis de los Mercados Adaptativos en un
esfuerzo por conciliar la hipótesis de los Mercados Eficientes con las Finanzas del Comportamiento (Lo, 2005). (Urquhart y
Hudson, 2013) han llevado a cabo una investigación empírica exhaustiva sobre los Mercados Adaptativos
Hipótesis (AMH) en tres de los mercados bursátiles más consolidados del mundo; Estados Unidos, Reino Unido y
Mercados japoneses utilizando datos de muy largo plazo. Su investigación proporciona evidencia de mercados adaptables,
con retornos pasando por periodos de independencia y dependencia aunque la magnitud de
la dependencia varía considerablemente. Por lo tanto, la dependencia lineal de los rendimientos de las acciones varía con el tiempo.
pero la dependencia no lineal es fuerte en todo momento. Sus resultados generales sugieren que la AMH proporciona
una mejor descripción del comportamiento de los rendimientos de las acciones que la Hipótesis del Mercado Eficiente.

2.4. Previsibilidad de los mercados

Cuando los mercados son débilmente eficientes, entonces debe ser posible predecir su comportamiento o al menos
determinar criterios con impacto predictivo sobre los mismos. Aunque, la naturaleza de los mercados es tal que
una vez que dicha información está disponible, la absorben y se ajustan y, por lo tanto, se vuelven
eficientes frente a los criterios predictivos iniciales y, por lo tanto, haciéndolos obsoletos. Información
La absorción por los mercados y el logro de nuevos equilibrios ocurren constantemente en los mercados y algunos
los investigadores han profundizado en la modelización de su dinámica y parámetros en circunstancias especiales
(García & Uroševiÿ, 2013). No obstante, la existencia de burbujas económicas especulativas indica
que los participantes del mercado operan sobre suposiciones irracionales y emocionales y no pagan
suficiente atención al valor subyacente real. La investigación de (Potì & Siddique, 2013) indica
existencia de previsibilidad en el Mercado de Divisas (FOREX) también. Aunque los mercados en el real
mundo puede no ser absolutamente eficiente, la investigación muestra que algunos son más eficientes que otros,
como se mencionó pueden ser fuertes, semifuertes y débiles en términos de eficiencia (Fama, 1970). Tiene
Se ha demostrado que los mercados en las economías emergentes como Malasia, Tailandia, Indonesia y

Filipinas tiende a ser significativamente menos eficiente en comparación con las economías desarrolladas y por lo tanto, las medidas predictivas como las reglas técnicas de negociación parecen tener más poder (H. Yu, Nartea, Gan, y Yao, 2013). Además, las variantes a corto plazo de las reglas técnicas de negociación tienen mejores capacidad predictiva que las variantes a largo plazo, además de que los mercados se vuelven más eficientes desde el punto de vista de la información con el tiempo (H. Yu, et al., 2013). Por lo tanto, existe la necesidad de revisar continuamente el nivel de eficiencia de mercados emergentes económicamente dinámicos y de rápido crecimiento (H. Yu, et al., 2013).

2.5. Análisis fundamental vs. técnico

Aquellos que están convencidos de que los mercados pueden predecirse, al menos hasta cierto punto, están segmentados en dos campamentos. Aquellos que creen que los movimientos históricos del mercado están obligados a repetirse están conocidos como analistas técnicos. Simplemente creen que hay patrones visuales en un gráfico de mercado que un ojo experimentado puede detectar. Con base en esta creencia, se nombran muchos de los movimientos del gráfico que constituye la base del análisis técnico. En un nivel superior, los analistas técnicos tratan de detectar tales sutiles modelos matemáticos mediante el uso de potencia de cálculo y técnicas de reconocimiento de patrones. Aunque las técnicas de análisis técnico son las más difundidas entre muchos de los brokers del mercado y participantes, para una mente científica con un punto de vista holístico, el análisis técnico por sí solo no puede parecer muy atractivo. Especialmente porque la mayoría de los análisis técnicos no respaldan ninguna de sus observaciones. nada más que afirmar que existen patrones. Hacen muy poco o nada para averiguar el razón detrás de la existencia de patrones. Algunas de las técnicas comunes en el análisis técnico son la reglas de promedio móvil, reglas de fuerza relativa, reglas de filtro y reglas de ruptura del rango comercial. en un estudio reciente, la efectividad y las limitaciones de estas reglas fueron puestas a prueba nuevamente y fue demostró que en muchos casos y contextos estas reglas no tienen mucho poder predictivo (H. Yu, et al., 2013). Sin embargo, existen y continúan existiendo muchos métodos sofisticados de predicción financiera. esfuerzos de modelado basados en varios tipos o combinaciones de algoritmos de aprendizaje automático como neural (Anastasakis & Mort, 2009; Ghazali, Hussain, & Liatsis, 2011; Sermpinis, Laws, Karathanasopoulos y Dunis, 2012; Vanstone & Finnie, 2010), lógica difusa (Bahrepour, Akbarzadeh-T.,

Yaghoobi, & Naghibi-S., 2011), Regresión de vectores de soporte (S.-C. Huang, Chuang, Wu, & Lai, 2010; Premanode & Toumazou, 2013), programación de redes genéticas basadas en reglas (Mabu, Hirasawa, Obayashi y Kuremoto, 2013).

Sin embargo, existe una segunda escuela de pensamiento conocida como análisis fundamental, que parece estar más prometedor En el análisis fundamental, los analistas observan los datos fundamentales que están disponibles para ellos. de diferentes fuentes y hacer suposiciones basadas en eso. Podemos llegar a por lo menos 5 principales fuentes de datos fundamentales: 1- los datos financieros de una empresa como datos en su balance o datos financieros sobre una moneda en el mercado FOREX, 2- Datos financieros sobre un mercado como su índice, 3- Datos financieros sobre las actividades del gobierno y los bancos, 4- Circunstancias políticas, 5- Circunstancias geográficas y meteorológicas como desastres naturales o no naturales. Sin embargo, determinar el valor fundamental subyacente de cualquier activo puede ser un desafío y con muchos incertidumbre (Kaltwasser, 2010). Por lo tanto, la automatización del análisis fundamental es bastante rara. Fasanghari y Montazer (2010) diseñan un sistema experto difuso para carteras bursátiles recomendación que toma como entrada algunos de los fundamentos de la empresa a través de métricas numéricas.

Sin embargo, la mayoría de los participantes del mercado intentan estar atentos tanto a los datos técnicos como a los fundamentales. Los datos fundamentales suelen ser de naturaleza no estructurada y sigue siendo un desafío hacer que la mejor uso de él de manera eficiente a través de la informática. El reto de la investigación aquí es hacer frente a este datos no estructurados. Un enfoque reciente que está surgiendo para facilitar tales temas datos no estructurados y extraer datos estructurados de ellos es el desarrollo de búsqueda especializada motores como este buscador semántico de noticias financieras (Lupiani-Ruiz, et al., 2011). De todos modos, eso Sigue siendo un desafío extraer significado de manera confiable del texto y de un motor de búsqueda como lo anterior se limita a extraer los datos numéricos disponibles en los textos relevantes. Un estudio reciente muestra el impacto de las noticias de EE. UU., las noticias del Reino Unido y las noticias holandesas en tres bancos holandeses durante el período financiero crisis de 2007-2009 (Kleinnijenhuis, 2013). Este estudio específico continúa explorando los pánicos del mercado desde

una perspectiva del periodismo financiero y teorías de la comunicación específicamente en una era algorítmica y comercio de frecuencias.

2.6. Comercio algorítmico

El comercio algorítmico se refiere a los mecanismos predictivos de los agentes comerciales robóticos inteligentes que son participando activamente en cada momento de las operaciones del mercado. La velocidad de tal toma de decisiones ha ha aumentado drásticamente recientemente y se ha creado el nuevo término comercio de frecuencia. Comercio de frecuencia ha sido más popular en el mercado de valores y Evans, Pappas y Xhafa (2013) están entre los primeros que están utilizando redes neuronales artificiales y algoritmos genéticos para construir un modelo de comercio algorítmico para especulación cambiaria intradiaria.

2.7. Sentimiento y Análisis Emocional

Se trata de detectar el sentimiento emocional preservado en el texto a través de técnicas semánticas especializadas. análisis para una variedad de propósitos, por ejemplo, para medir la calidad de la recepción del mercado para un nuevo producto y los comentarios generales de los clientes o para estimar la popularidad de un producto o marca (Ghiassi, Skinner, & Zimbra, 2013; Mostafa, 2013) entre personas. Hay un cuerpo de investigación que es centrada en el análisis de sentimiento o la denominada “minería de opinión” (Balahur, Steinberger, Goot, Pouliquen y Kabadjov, 2009; Cambria, Schuller, Yunqing y Havasi, 2013; Hsinchun y Zimbra, 2010). Se basa principalmente en la identificación de palabras positivas y negativas y en el procesamiento de textos con el fin de clasificando su postura emocional como positiva o negativa. Un ejemplo de tal esfuerzo de análisis de sentimiento es el trabajo de Maks y Vossen (2012) que presenta un modelo de léxico para el análisis de sentimiento profundo y minería de opinión. Un concepto similar se puede utilizar en otras áreas de investigación, como la detección de emociones. en notas de suicidio como las realizadas por Desmet y Hoste (2013).

Sin embargo, también se puede explorar un análisis del sentimiento emocional del texto de la noticia para el mercado. predicción. Schumaker, Zhang, Huang y Chen (2012) han tratado de evaluar el sentimiento en artículos de noticias financieras en relación con el mercado de valores en su investigación, pero no ha sido completamente

exitoso. Un ejemplo reciente más exitoso de esto sería (L.-C. Yu, Wu, Chang y Chu, 2013),
donde se propuso un modelo de entropía contextual para expandir un conjunto de palabras semilla al descubrir
palabras de emociones similares y sus intensidades correspondientes de artículos de noticias del mercado de valores en línea.
Esto se logró calculando la similitud entre las palabras semilla y las palabras candidatas.
de sus distribuciones contextuales utilizando una medida de entropía. Una vez que las palabras semilla han sido
expandido, tanto las palabras semilla como las palabras expandidas se utilizan para clasificar el sentimiento de la noticia
artículos. Sus resultados experimentales muestran que el uso de las palabras de emoción expandida mejoró
rendimiento de clasificación, que se mejoró aún más al incorporar sus correspondientes
intensidades que hicieron que los resultados de precisión variaran del 52% al 91,5% al variar la diferencia de
niveles de intensidad de clases positivas y negativas de (-0.5 a 0.5) a >9.5 respectivamente. Tambien es
interesante notar que el análisis emocional del texto no tiene que ser mero basado en la positividad
negatividad y puede hacerse en otras dimensiones o en multidimensiones (Ortigosa-Hernández, et al.
al., 2012). Una investigación reciente de Loia y Senatore (2014) presenta un marco para
extraer las emociones y los sentimientos expresados en los datos textuales. los sentimientos son
expresado por una polaridad positiva o negativa. Las emociones se basan en la concepción de Minsky de
emociones que consta de cuatro dimensiones afectivas (Agrado, Atención, Sensibilidad y
Aptitud). Cada dimensión tiene seis niveles de activación, llamados niveles sénticos. Cada nivel representa un
estado de ánimo emocional y puede ser más o menos intenso, dependiendo de la posición en el
dimensión correspondiente. Otro desarrollo interesante en el análisis de sentimientos es un esfuerzo por ir
desde el sentimiento de fragmentos de texto hasta características o aspectos específicos que están relacionados con un concepto o
producto; Kontopoulos, Berberidis, Dergiades y Bassiliades (2013) proponen un modelo más eficiente
análisis de sentimientos de las publicaciones de Twitter, en el que las publicaciones no se caracterizan simplemente por un sentimiento
puntuación, sino que reciben una calificación de sentimiento por cada noción distinta en ellos que está habilitada por el
ayuda de una ontología. W. Li y Xu (2014) adoptan otro ángulo al buscar características que son
"significativo" para las emociones en lugar de simplemente elegir palabras con un alto grado de concurrencia y
por lo tanto, cree una clasificación de emociones basada en texto utilizando la extracción de la causa de la emoción.

3. Revisión de los principales trabajos disponibles

A pesar de la existencia de múltiples sistemas en esta área de investigación, no hemos encontrado ninguno dedicado y análisis comparativo completo y revisión de los sistemas disponibles. nikfarjam, Emadzadeh y Muthaiyah (2010) han hecho un esfuerzo en forma de un documento de conferencia que proporciona un resumen aproximado bajo el título de "Enfoques de minería de texto para la predicción del mercado de valores". hagenau, Liebmann y Neumann (2013) también han presentado una tabla comparativa básica en su literatura revisión a la que se ha hecho referencia en algunas partes de este trabajo. En esta sección estamos llenando este vacío al repasando los principales sistemas que se han desarrollado en la última década.

3.1. La visión general genérica

Todos estos sistemas tienen algunos de los componentes que se muestran en la figura 3. En un extremo, el texto se alimenta como entrada al sistema y en el otro extremo se generan algunos valores predictivos de mercado como salida.

En las siguientes secciones, se examina más de cerca cada uno de los componentes representados, sus funciones y Fundamento teórico.

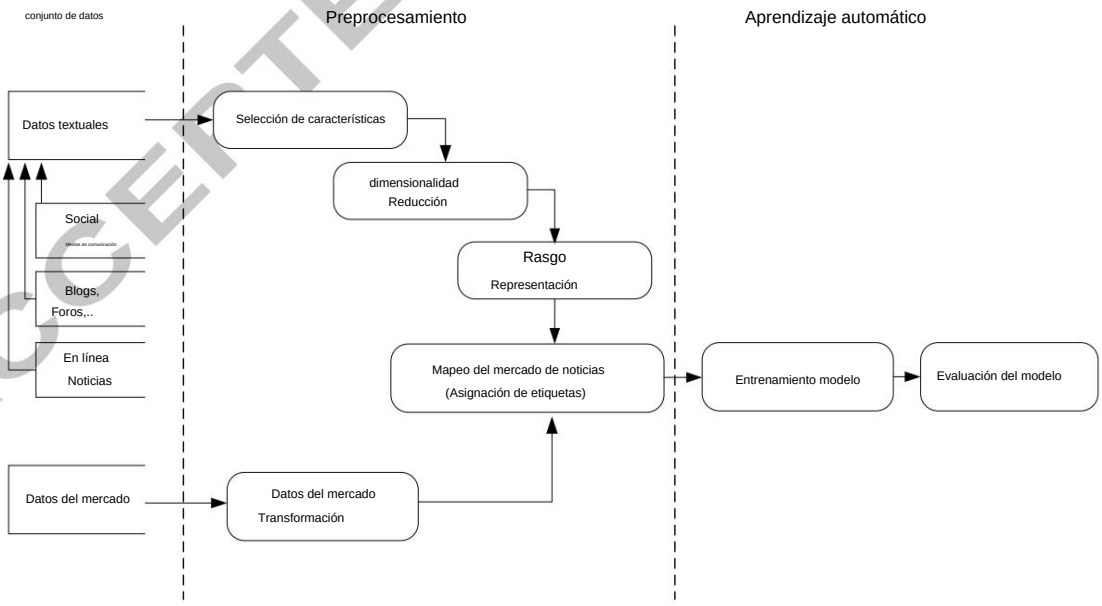


Figura 3 Diagrama de componentes comunes del sistema genérico

3.2. Conjunto de datos de entrada

Todos los sistemas toman al menos dos fuentes de datos como entrada, a saber, los datos textuales del sitio web en línea.
los recursos y los datos del mercado.

3.2.1. Datos textuales

La entrada de texto puede tener varias fuentes y tipos de contenido, como se muestra en la tabla 1.
La mayoría de las fuentes utilizadas son importantes sitios web de noticias como The Wall Street Journal (Werner & Myrray
Z., 2004), Financial Times (Wuthrich, et al., 1998), Reuters (Pui Cheong Fung, Xu Yu y Wai, 2003),
Dow Jones, Bloomberg (Chatrath, et al., 2014; Jin, et al., 2013), Forbes (Rachlin, Last, Albergue, &
Kandel, 2007), así como Yahoo! Finanzas (Schumaker, et al., 2012). El tipo de noticia es
noticias generales o noticias financieras especiales. La mayoría de los sistemas están utilizando noticias financieras tal como son.
se considera que tiene menos ruido en comparación con las noticias generales. Lo que se está extrayendo aquí es la noticia.
texto o el titular de la noticia. Los titulares de noticias se usan ocasionalmente y se argumenta que son más directos.
al punto y, por lo tanto, de menos ruido causado por el texto detallado (C.-J. Huang, Liao, Yang, Chang, & Luo,
2010; Peramunetilleke & Wong, 2002). Menos investigadores han buscado en fuentes menos formales de
información textual, por ejemplo, Das y Chen (2007) han examinado el texto en los tableros de mensajes en línea en
su trabajo. Y más recientemente, Y. Yu, Duan y Cao (2013) han analizado el contenido textual de
las redes sociales como twitter y blogs. Algunos investigadores se centran únicamente en Twitter y lo utilizan para
predicción del mercado y el análisis del estado de ánimo del público de manera más eficiente (Bollen & Huina, 2011; Vu, Chang, Ha, &
Coller, 2012). Una tercera clase de fuente textual para los sistemas han sido los informes anuales de la empresa,
comunicados de prensa y divulgaciones corporativas. Observamos que una diferencia acerca de esta clase de
información sobre las empresas es la naturaleza de su oportunidad, por lo que los informes y divulgaciones regulares
tener horarios preestablecidos. Se sospecha que el tiempo preestablecido tiene un impacto en la capacidad de predicción o
naturaleza que puede haber sido causada por reacciones anticipatorias entre los participantes del mercado (Chatrath,
et al., 2014; C.-J. Huang, et al., 2010). Por lo tanto, hemos incluido una columna para esta información en
la mesa. También es importante recordar que algunos informes textuales pueden tener estructura o semi-

formatos estructurados como comunicados macroeconómicos que provienen de gobiernos o centrales bancos sobre las tasas de desempleo o el Producto Interno Bruto (PIB). Chatrath, et al. (2014) tienen usó tales datos estructurados para predecir saltos en el mercado de divisas (FOREX). Es más, con respecto a la frecuencia de publicación de información, una observación interesante que se ha hecho en la investigación reciente es que el aumento de la frecuencia de los comunicados de prensa puede causar una disminución en la información comercio y, por lo tanto, el grado de asimetría de la información es menor para las empresas con noticias más frecuentes (Sankaraguruswamy, Shen y Yamada, 2013). Esto plantea un punto interesante por el cual el los comerciantes desinformados aumentan y, por lo tanto, juegan un papel importante en el mercado con demasiada frecuencia comunicados de prensa Aunque puede parecer contrario a la intuición, ya que uno puede esperar un comercio más informado para ocurrir bajo tales circunstancias.

Referencia	Tipo de texto	Fuente de texto	No. de ítems Preprogramado	No estructurado	
Wuthrich, et al. (1998)	noticias generales	The Wall Street Journal, Financial Times, Reuters, dow jones , Bloomberg	No dado	No	Sí
Peramunetillek e y Wong (2002)	Noticias financieras	HFDF93 vía www.olsen.ch	40 titulares por hora	No	Sí
Pui Cheong Fung, et al. (2003)	Empresa noticias	Mercado Reuters 3000 Extra	600.000	No	Sí
Werner y Myrray Z. (2004)	Publicaciones de mensajes	yahoo! Finanzas, Raging Bull, Wall Street Journal	1.5 millones mensajes	No	Sí
Mittermayer (2004)	Noticias financieras	No mencionado	6,602	No	Sí
Das y Chen (2007)	Publicaciones de mensajes	Pizarrón de mensajes	145,110 mensajes	No	Sí
Soni, van Eck y Kaymak (2007)	Noticias financieras	FT Intelligence (Financial Tiempos de servicio en línea)	3493	No	Sí
Zhai, Hsu y Halgamuge (2007)	Sector del mercado noticias	financiera australiana Revisar	148 directo noticias de la empresa y 68 los indirectos	No	Sí
Rachlin, et al. (2007)	Noticias financieras	Forbes.com, today.reuters.com	No mencionado	No	Sí
Paul C. Tetlock, Saar Tsechansky y Macskassy (2008)	Noticias financieras	Wall Street Journal, Dow Jones News Service de Base de datos de noticias Factiva.	350.000 historias	No	Sí
Mahajan, Dey y Haque	Noticias financieras	No mencionado	700 noticias artículos	No	Sí

(2008)					
mayordomo y Keselj (2009)	Anual informes	sitios web de la empresa	No mencionado	Sí	Sí
Schumaker y Chen (2009)	Noticias financieras	Yahoo Finanzas	2800	No	Sí
F. Li (2010)	Presentaciones corporativas	de la gerencia Sección de discusión y análisis de las presentaciones 10-K y 10-Q de la SEC Edgar Sitio web	13 millones delantero buscando declaraciones en 140.000 Presentaciones 10-Q y K	Sí (empresa anual reporte)	Sí
C.-J. Huang, et al. (2010)	Noticias financieras	Periódicos electrónicos líderes en Taiwán	12.830 titulares	No	Sí
Groth y Müntermann (2011)	Ad hoc locutores t	Divulgaciones corporativas	423 revelaciones	No	Sí
Schumaker, et al. (2012)	Noticias financieras	Yahoo! Finanzas	2802	No	Sí
Lugmayr y Gossen (2012)	Corredor boletines	Corredores	No disponible	Sí	Sí
Y. Yu, et al. (2013)	Diario convencional y sociales medios de comunicación	Blogs, foros, noticias y microblogs (p. ej., Twitter)	52,746 mensajes	No	Sí
Hagenau y otros (2013)	Corporativo locutores t y financiero noticias	DGAP, EuroAdhoc	10870 y 3478 respectivamente	No	Sí
Jin, et al. (2013)	General news	Bloomberg	Chatrath, et al. 361,782	No	Sí
(2014)	Macroeconomía noticias ic	Bloomberg	No mencionado	Sí	No
Bolén y Huina (2011)	tuits	Gorjeo	9,853,498	No	Sí
Vu, et al. (2012)	Tweets	Gorjeo	5,001,460	No	Sí

Tabla 1 Comparación de la entrada de texto para diferentes sistemas

3.2.2. Datos del mercado

La otra fuente de datos de entrada para los sistemas proviene de los valores numéricos en los mercados financieros en forma de puntos de precios o índices. Estos datos se utilizan principalmente con el propósito de entrenar el aprendizaje automático. algoritmos y ocasionalmente se utiliza con fines de predicción por lo que se introduce en la máquina algoritmo de aprendizaje como una variable independiente para una función, este tema se discutirá en un artículo posterior sección. En la tabla 2, se proporcionan detalles cruciales de dichos datos de mercado. En primer lugar, hay una diferenciación entre las bolsas de valores y el mercado de divisas (FOREX). Investigaciones anteriores han sido centrado principalmente en la predicción del mercado de valores, ya sea en forma de un índice bursátil como el Dow Jones

Promedio industrial (Bollen & Huina, 2011; Werner & Myrray Z., 2004; Wuthrich, et al., 1998), EE. UU.
Índice NASDAQ (Rachlin, et al., 2007), Índice Morgan Stanley High-Tech (MSH) (Das & Chen, 2007),
el Indian Sensex Index (Mahajan, et al., 2008), S&P 500 (Schumaker & Chen, 2009) o la bolsa
precio de una empresa específica como BHP Billiton Ltd. (BHP.AX) (Zhai, et al., 2007) o como Apple, Google,
Microsoft y Amazon en Vu, et al. (2012) o un grupo de empresas (Hagenau, et al., 2013). Él
El mercado FOREX solo se ha abordado ocasionalmente en aproximadamente el diez por ciento de los trabajos revisados; me gusta
en el trabajo de Peramunetilleke y Wong (2002) y más recientemente en los trabajos de Chatrath, et al.
(2014) y Jin, et al. (2013). Vale la pena considerar que los niveles de eficiencia de los mercados FOREX
alrededor del mundo varían (Wang, Xie, & Han, 2012), y por lo tanto, debería ser posible encontrar menos
pares de divisas eficientes que son propensos a la previsibilidad.

Además, casi todos los tipos de pronóstico en cualquiera de las medidas de mercado anteriores son categóricos con
valores discretos como Up, Down y Steady. Hay muy pocas investigaciones que hayan explorado
un enfoque basado en la regresión lineal (Jin, et al., 2013; Schumaker, et al., 2012; Paul C. Tetlock, et
al., 2008).

Además, se compara el calendario predictivo de cada obra. El tiempo desde el punto de
comunicado de prensa a una observación de impacto en el mercado puede variar de segundos a días, semanas o meses. Él
la predicción de impacto de segundo o milisegundo es el elemento que alimenta toda una industria con el nombre de
micro-comercio en el que equipos especiales y proximidad a la fuente de noticias, así como al mercado
los sistemas informáticos son críticos; un trabajo sobre tendencias comerciales ha explicado bien tales transacciones cuantitativas
(Chordia, Roll y Subrahmanyam, 2011). Otro nombre para el mismo concepto es alta frecuencia.
Comercio que es explorado en detalle por Chordia, Goyal, Lehmann y Saar (2013). otro parecido
El término es comercio de baja latencia que se amplía en el trabajo de Hasbrouck y Saar (2013). Sin embargo,
investigaciones anteriores han indicado que sesenta minutos deben considerarse como un tiempo de convergencia de mercado razonable
a la eficiencia (Chordia, Roll y Subrahmanyam, 2005). La convergencia del mercado se refiere al período
durante el cual un mercado reacciona a la información que está disponible y se vuelve eficiente al

reflejándolo completamente. La disponibilidad de información y los canales de distribución son críticos aquí y la investigación en el tiempo de convergencia de la eficiencia del mercado está en curso (Reboredo, Rivera-Castro, Miranda, & García Rubio, 2013). La mayoría de las obras comparadas en la tabla 2 tienen una franja horaria diaria seguida de una intradiaria marcos de tiempo que están en el rango de 5 minutos (Chatrath, et al., 2014), 15 minutos (Werner & Myrray Z., 2004), 20 minutos (Schumaker & Chen, 2009) a 1, 2 o 3 horas (Peramunetilleke & Wang, 2002). Los períodos de experimentación también se contrastan, siendo el más corto solo 5 días en el caso de trabajo de Peramunetilleke y Wong (2002) y hasta múltiples años con el más largo a los 24 años de 1980 a 2004 por PC Tetlock (2007) seguido de 14 años de 1997 a 2011 en la obra de Hagenau, et al. (2013), 13 años de 1994 a 2007 en el trabajo de F. Li (2010) con este último buscando en un plazo anual y los primeros en plazos diarios. La mayoría restante de las obras asumir un período de experimentación con una duración de varios meses como se detalla en la tabla 2.

Referencia	Índice de mercado de mercado		Periodo de tiempo	Período	Tipo de pronóstico
Wuthrich, et al. (1998)	Acciones Dow Jones	Promedio Industrial, el Nlkkei 225, el Tiempos financieros 100, el Hang Seng y el Singapur Estrecheces	Diario	6 de diciembre de 1997 al 6 de marzo 1998	Categórico: Arriba, constante, abajo
Peramunetilleke y Wong (2002)	Tipo de cambio FOREX (USD DEM, USD-JPY)		Intradía (1, 2 o 3 horas)	22 a 27 septiembre de 1993	Categórico: Arriba, constante, abajo
Pui Cheong Fung, et al. (2003)	Acciones 33 acciones de la	cuelgue seng	Diariamente (sin retrasos y retrasos variables)	1 de octubre de 2002 al 30 de abril de 2003	Categórico: Subida, Gota
Werner y Myrray Z. (2004)	Acciones Dow Jones	Promedio industrial y el Dow Jones Internet	Intradía(15-min, 1 hora y 1 día)	Año 2000	Categórico: Comprar, Vender y mantener mensajes
Mittermayer (2004)	Cepo	Precios de las acciones	Diario	1 ene al 31 diciembre de 2002	Categórico: buenas noticias, malas noticias, no mudanzas
Eso y Chen (2007)	Almacena 24 sectores tecnológicos en	el morgan stanley Alta tecnología	Diario	julio y agosto de 2001	Índice de sentimiento agregado
Soni, et al. (2007)	Cepo	11 empresas de petróleo y gas	Diario	1 de enero de 1995 al 15 de mayo de 2006	Categórico: Positivo negativo
Zhai, et al. (2007)	Cepo	BHP Billiton Ltda. de Australia Bolsa	Diario	1 de marzo de 2005 al 31 de mayo de 2006	Categórico: Arriba, Abajo
Rachlin, et al.	Cepo	5 acciones de EE. UU.	Dias	7 de febrero a 7	Categórico: Arriba,

(2007)		NASDAQ		mayo de 2006	Ligero hacia arriba, esperado, levemente hacia abajo, hacia abajo
Paul C. Tetlock, et al. (2008)	Cepo	S&P 500 individuales empresas y sus flujos de efectivo futuros	Diario	1980 a 2004	Regresión de una medida llamada neg.
Mahajan, et al. (2008)	Cepo	Sensex	Diario	5 ago al 8 Abr	Categórico
Mayordomo y Keselj (2009)	Cepo	Deriva del mercado de 1 año	Anual	2003 a 2008	Categórico: Índice S&P 500 de rendimiento superior o inferior durante el próximo año
Schumaker y Chen (2009)	Acciones	S&P 500 acciones	Intradía (20 min)	26 oct al 28 noviembre de 2005	Categórico: Numérico discreto
F. Li (2010)	Cepo	(1) Índice (2) Resultados y flujos de caja trimestrales (3) Rendimientos de las acciones	Anualmente (Trimestral con 3 trimestres ficticios)	1994 a 2007	Categórico basado en el tono: Positivo, Negativo, Neutro, Incierto
C.-J. Huang, et al. (2010)	Acciones	Acciones de Taiwán Intercambio Financiero Precio	Diario	junio a noviembre 2005	Solo asignación de grado de importancia
Groth y Müntermann (2011)	Acciones	Exposición anormal al riesgo ARISKÿ (Thomson Reuters Marca de alcance de datos Historia)	Intradía (volatilidad durante el ÿ=15 y 30 min)	1 de agosto de 2003 al 31 de julio 2005	Categórico: Positivo negativo
Schumaker, et al. (2012)	Acciones	S&P 500	Intradía (20min)	26 oct al 28 noviembre de 2005	Regresión
Lugmayr y Gossen (2012)	Acciones	DAX 30 Índice de rendimiento	Intradía (3 o 4 veces al día)	No disponible	Categórico: Sentimiento [-1, 1]; Tendencia (oso, toro, Neutral); Fuerza de la tendencia (en %)
Y. Yu, et al. (2013)	Acciones	Rentabilidades anormales y acumulativo retornos anormales de 824 empresas	Diario	1 de julio al 30 septiembre de 2011	Categórico: Positivo o Negativo
Hagenau y otros (2013)	Acciones	Empresa específica	Diario	1997 a 2011	Categórico: Positivo o Negativo
Jin, et al. (2013)	Tipo de cambio	FOREX	Diario	1 ene al 31 diciembre de 2012	Regresión
Chattrath, et al. (2014)	Tipo de cambio de divisas		Intradía(5min)	Ene 2005 a diciembre de 2010	Categóricas: Saltos Positivos o Negativos
Bollen y Huiña (2011)	Valores	DJIA	Diario	28 febrero al 19 diciembre de 2008	Análisis de regresión
Sierra, et al. (2012)	Inventario	Precios de las acciones (en NASDAQ para AAPL, GOOG, MSFT, AMZN)	Diario	1 de abril de 2011 al 31 de mayo de 2011, prueba en línea 8 de septiembre al 26 de septiembre de 2012	Categórico: arriba y abajo

Tabla 2 Los datos del mercado de entrada, el período de tiempo del experimento, la duración y el tipo de pronóstico

3.3. Preprocesamiento

Una vez que los datos de entrada están disponibles, se deben preparar para que puedan incorporarse a un sistema de aprendizaje automático.
algoritmo. Esto para los datos textuales significa transformar el texto no estructurado en un representante
formato que está estructurado y puede ser procesado por la máquina. En minería de datos en general y texto
la minería específicamente, la fase de preprocesamiento tiene un impacto significativo en los resultados generales (Uysal &
Gunal, 2014). Hay al menos tres subprocesos o aspectos del preprocesamiento que hemos
contrastados en los trabajos revisados, a saber: Rasgo-Selección, Dimensionalidad-Reducción, Rasgo
Representación

3.3.1 Selección de características

La decisión sobre las características a través de las cuales se representará un fragmento de texto es crucial porque desde
de una entrada de representación incorrecta no se puede esperar más que una salida sin sentido. En
En la tabla 3 se enumera el tipo de selección de características del texto para cada una de las obras.

En la mayor parte de la literatura se han utilizado las técnicas más básicas cuando se trata de minería de texto.
problemas de predicción basados en el mercado a los que se refiere Kleinnijenhuis (2013) también en consonancia con nuestra
recomendaciones. La técnica más común es la llamada "bolsa de palabras", que consiste esencialmente en romper
el texto hasta sus palabras y considerando cada una de ellas como una característica. Como se presenta en la tabla 3, alrededor
tres cuartas partes de las obras se basan en esta técnica básica de selección de características en la que el orden
y la co-ocurrencia de palabras son completamente ignoradas. Schumaker, et al. (2012) y Schumaker y
Chen (2009) ha explorado otras dos técnicas, a saber, los sintagmas nominales y las entidades nombradas. En el
primero, las palabras con una parte del discurso del sustantivo se identifican con la ayuda de un léxico y luego
usando reglas sintácticas en las partes circundantes del discurso, se detectan frases nominales. En este último, un
Se añade un sistema de categorías en el que se organizan los sustantivos o los sintagmas nominales. Ellos usaron el tan
llamado marco MUC-7 de clasificación de entidades, donde las categorías incluyen fecha, ubicación, dinero,
organización, porcentaje, persona y tiempo. Sin embargo, no tuvieron ningún éxito en mejorar
sus sintagmas nominales resultan a través de esta categorización adicional en su técnica de entidades nombradas. Sobre el
Por otro lado, Vu, et al. (2012) mejoran con éxito los resultados mediante la creación de un reconocimiento de entidad nombrada

(NER) para identificar si un Tweet contiene entidades nombradas relacionadas con sus objetivos
empresas basadas en un modelo lineal de campos aleatorios condicionales (CRF). Otro de uso menos frecuente.
pero una técnica interesante es la llamada técnica Latent Dirichlet Allocation (LDA) utilizada por Jin, et al.
Alabama. (2013) así como Mahajan, et al. (2008) en la tabla 3 para categorizar palabras en conceptos y usar el
conceptos representativos como las características seleccionadas. Algunas de las otras obras utilizan una técnica
llamados n-gramas (Butler & Kešelj, 2009; Hagenau, et al., 2013). Un n-grama es una secuencia contigua de
n elementos que suelen ser palabras de una determinada secuencia de texto. Sin embargo, la secuencia de palabras y
las estructuras sintácticas podrían ser esencialmente más avanzadas. Un ejemplo de tal configuración podría ser el uso
de N-gramas Sintácticos (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2013).
Sin embargo, la inclusión de tales funciones puede dar lugar a problemas de dependencia del idioma que deben
tratarse (Kim, et al., 2014). Técnicas de combinación de características léxicas y semánticas para abreviaturas.
también se puede mejorar la clasificación del texto (Yang, Li, Ding y Li, 2013).

La selección de características es un paso estándar en la fase de preprocesamiento de la minería de datos y hay muchas
otros enfoques que se pueden considerar para la selección de características textuales, siendo los algoritmos genéticos uno
de ellos como se describe en detalle en el trabajo de Tsai, Eberle y Chu (2013). En otra obra Hormiga
Colony Optimization se ha utilizado con éxito para la selección de características textuales (Aghdam, Ghasem
Aghaee y Basiri, 2009). La selección de características para la clasificación de texto también se puede hacer en función de un filtro
basado en el enfoque probabilístico (Uysal & Gunal, 2012). Feng, Guo, Jing y Hao (2012) proponen una
modelo probabilístico generativo, describiendo categorías por distribuciones, manejando la selección de características
problema mediante la introducción de un vector latente de exclusión/inclusión binaria, que se actualiza a través de un eficiente
Búsqueda de metrópolis. Se ha demostrado que el preprocesamiento delicado tiene un impacto significativo en textos similares
problemas mineros (Haddi, Liu, & Shi, 2013; Uysal & Gunal, 2014).

3.3.2. Reducción de dimensionalidad

Tener un número limitado de características es extremadamente importante ya que el aumento en el número de las características que pueden ocurrir fácilmente en la selección de características en el texto pueden hacer que la clasificación o el agrupamiento problema extremadamente difícil de resolver al disminuir la eficiencia de la mayoría de los algoritmos de aprendizaje, este situación es ampliamente conocida como la maldición de la dimensionalidad (Pestov, 2013). En la tabla 3 debajo Dimensionalidad-Reducción se señala el enfoque adoptado por cada una de las obras. Zhai, et al. (2007) hace esto eligiendo los 30 conceptos principales con los pesos más altos como características en lugar de todos conceptos disponibles. Mittermayer (2004) hace esto filtrando los 1000 términos principales de todos los términos. Sin embargo, el enfoque más común es establecer un límite mínimo de ocurrencia y reducir los términos seleccionando las que alcanzan un número de ocurrencias (Butler & Kešelj, 2009; Schumaker & Chen, 2009). El siguiente enfoque común es usar un diccionario predefinido de algún tipo para reemplazarlos con un nombre de categoría o valor. Algunos de estos diccionarios están elaborados especialmente por un experto en el mercado como el utilizado por Wuthrich, et al. (1998) o Peramunetilleke y Wong (2002) o son más específico de un campo específico como la psicología en el caso de Harvard-IV-4 que se ha utilizado en la trabajo de Paul C. Tetlock, et al. (2008). Y otras veces son diccionarios de uso bastante general como el WordNet Thesaurus utilizado por Zhai, et al. (2007). A veces se crea un diccionario o diccionario de sinónimos dinámicamente basado en el corpus de texto utilizando una herramienta de extracción de términos (Soni, et al., 2007). otro conjunto de Las actividades que suelen constituir el mínimo de dimensionalidad-reducción son: características derivadas, conversión a minúsculas, eliminación de puntuación y eliminación de números, direcciones de páginas web y palabras vacías. Estos pasos se dan casi siempre y en algunas obras son los únicos pasos que se dan como en el trabajo de Pui Cheong Fung, et al. (2003).

La reducción de funciones se puede mejorar de varias maneras, pero la investigación actual aún no ha profundizado en él tanto como se observa en los trabajos reseñados. Berka y Vajteršic (2013) introducen una método detallado para la reducción de la dimensionalidad del texto, basado en un vector paralelo de términos raros reemplazo.

3.3.3. Característica-Representación

Después de determinar el número mínimo de funciones, cada función debe estar representada por un valor numérico para que pueda ser procesado por algoritmos de aprendizaje automático. Por lo tanto, el título "Característica Representación" en la tabla 3 se utiliza para la columna en la que el tipo de valor numérico que se asociado con cada característica se compara para todos los trabajos revisados. Este valor numérico asignado actúa como una partitura o un peso. Hay al menos 5 tipos que son muy populares, a saber, Ganancia de información (IG), estadísticas de chi-cuadrado (CHI), frecuencia de documentos (DF), precisión equilibrada (Acc2) y Frecuencia de Término-Frecuencia de Documento Inversa (TF-IDF). Una comparación de estos cinco métricas junto con propuestas para algunas métricas nuevas se pueden encontrar en el trabajo de Taÿcÿ y Güngör (2013).

La técnica más básica es una representación booleana o binaria en la que dos valores como 0 y 1 representar la ausencia o presencia de una característica, por ejemplo, una palabra en el caso de una técnica de bolsa de palabras como en estos trabajos (Mahajan, et al., 2008; Schumaker, et al., 2012; Wuthrich, et al., 1998). El siguiente La técnica más común es la frecuencia de término-frecuencia de documento inversa o TF-IDF (Groth & Muntermann, 2011; Hagenau, et al., 2013; Peramunetilleke y Wong, 2002; Pui Cheong Fung, et al., 2003). El valor de TF-IDF aumenta proporcionalmente al número de veces que aparece una palabra en el documento, pero está compensado por la frecuencia de la palabra en el corpus, para equilibrar el general popularidad de algunas palabras. También existen otras medidas similares que se utilizan ocasionalmente como la Métrica de Discriminación de categoría de frecuencia de término (TF-CDF) que se deriva de Frecuencia de categoría (CF) y demuestra ser más efectivo que TF-IDF en este trabajo (Peramunetilleke & Wong, 2002).

En general, en la minería de texto, la reducción de funciones mejorada (reducción de dimensionalidad) y la función la ponderación (representación de características) puede tener un impacto significativo en la clasificación final del texto eficiencia (Shi, He, Liu, Zhang y Song, 2011).

Referencia	Selección de características	Reducción de dimensionalidad	Rasgo Representación
Wuthrich, et al.	Bolsa de palabras	Diccionarios predefinidos (secuencias de palabras)	Binario

(1998)		por un experto)	
Peramunetillek e y Wong (2002)	Bolsa de palabras	Conjunto de registros de palabras clave	Booleano, TF-IDF, TF-CDF
Pui Cheong Fung, et al. (2003)	Bolsa de palabras	Stemming, conversión a minúsculas, eliminación de puntuación, números, direcciones de páginas web y palabras vacías	TF-FDI
Werner y Myrray Z. (2004)	Bolsa de palabras	Criterio de información mínima (primeras 1000 palabras)	Binario
Mittermayer (2004)	Bolsa de palabras	Selección de 1000 términos	TF-FDI
Das y Chen (2007)	bolsa de palabras, trillizos	Diccionarios predefinidos	Diferente discreto valores para cada uno clasificador
Soni, et al. (2007)	Visualización	Tesauro elaborado con la herramienta de extracción de términos de NJ van Eck	Visual coordenadas
Zhai, et al. (2007)	Bolsa de palabras	WordNet Thesaurus (eliminación de palabras vacías, etiquetado POS, conceptos de nivel superior a través de WordNet). Los 30 conceptos principales.	Binario, TF-IDF
Rachlin, et al. (2007)	Bolsa de palabras, valores financieros de uso común	Lista de palabras clave más influyentes (Extracción automática)	TF, booleano, Extractor salida de software
Paul C. Tetlock, et al. (2008)	Bolsa de palabras para palabras negativas	Diccionario predefinido. Diccionario psicosocial Harvard-IV-4.	Frecuencia dividida por el número total de palabras
Mahajan, et al. (2008)	Dirichlet latente Asignación (LDA)	Extracción de veinticinco temas	Binario
mayordomo y Keselj (2009)	N-Grams de caracteres, tres puntajes de legibilidad, rendimiento del año pasado	Ocurrencia mínima por documento.	Frecuencia del n-grama en un perfil
Schumaker y Chen (2009)	Bolsa de palabras, frases nominales, entidades nombradas	Ocurrencia mínima por documento	Binario
F. Li (2010)	Bolsa de palabras, tono y contenido	Diccionarios predefinidos	Binario, valor de diccionario
C.-J. Huang, et al. (2010)	Simultáneo términos, pares ordenados	sustitución de sinónimos	Ponderado en función de la relación de subida/ bajada del índice
Groth y Müntermann (2011)	Bolsa de palabras	Métodos de puntuación de características usando ambos Métricas de ganancia de información y chi-cuadrado	TF-FDI
Schumaker, et al. (2012)	OpinionFinder tono general y polaridad	Ocurrencia mínima por documento	Binario
Lugmayr y Gossen (2012)	Bolsa de palabras	derivación	Valor de sentimiento
Y. Yu, et al. (2013)	Bolsa de palabras	No mencionado	Binario
Hagenau y otros (2013)	bolsa de palabras, Sintagmas nominales, Combinaciones de palabras, N gramos	Frecuencia para noticias, enfoque Chi2 y separación binormal (BNS) para selección de características basada en retroalimentación exógena, diccionario.	TF-FDI

Jin, et al. (2013)	Asignación latente de Dirichlet (LDA)	Extracción de temas, identificación de temas principales mediante la alineación manual de artículos de noticias con fluctuaciones monetarias,	Distribución temática de cada artículo
Chattrath, et al. (2014)	Datos estructurados	Datos estructurados	Datos estructurados
Bollen y Huiña (2011)	Por OpinionFinder	Por OpinionFinder	Por OpinionFinder
Vú, et al. (2012)	Número total diario de positivos o negativos en Twitter Sentimiento Tool (TST) y un léxico de emoticonos. Media diaria de Pointwise Mutual Information (PMI) para valores alcistas-bajistas predefinidos palabras ancla	Palabras clave predefinidas relacionadas con la empresa, Reconocimiento de entidad nombrada basado en lineal Condicional Campos aleatorios (CRF)	número real para Neg_Pos diario y Alcista_Bajista

Tabla 3 Preprocesamiento: Selección de funciones, Reducción de funciones, Representación de funciones

3.4. Aprendizaje automático

Una vez que se completa el preprocesamiento y el texto se transforma en una serie de características con un Representación numérica, se pueden utilizar algoritmos de aprendizaje automático. A continuación un breve Se presenta un resumen de estos algoritmos, así como una comparación de algunos de los otros detalles. tecnicismos en los diseños de los sistemas revisados.

3.4.1. Algoritmos de aprendizaje automático

En esta sección, se intenta proporcionar un resumen de los algoritmos de aprendizaje automático utilizados en el trabajos revisados. Se observa que comparar tales algoritmos no es fácil y está lleno de trampas (Salzberg, 1997). Por lo tanto, el objetivo principal es informar lo que se utiliza; para que ayude a entender lo que falta y puede ser posible para futuras investigaciones. Casi todos los algoritmos de aprendizaje automático que se han se utilizan los algoritmos de clasificación que se enumeran en la Tabla 4. Básicamente, los sistemas utilizan los datos de entrada aprender a clasificar una salida generalmente en términos del movimiento del mercado en clases como UP, ABAJO y CONSTANTE. Sin embargo, también hay un grupo de trabajos que utilizan el análisis de regresión para hacer predicciones y no clasificar.

La Tabla 4 clasifica los trabajos revisados en función de su algoritmo utilizado en 6 clases:

A) Máquina de vectores de soporte (SVM)

B) Algoritmos de regresión

C) Naïve Bayes

D) Reglas o árboles de decisión

E) Algoritmos Combinatorios

F) Experimentos de algoritmos múltiples

A) Máquina de vectores de soporte (SVM): Esta sección en la tabla 4 contiene la clase de algoritmos que la gran mayoría de los trabajos revisados están utilizando (Mittermayer, 2004; Pui Cheong Fung, et al., 2003; Schumaker y Chen, 2009; Soni, et al., 2007; Werner y Myrray Z., 2004). SVM es un clasificador lineal binario no probabilístico utilizado para el aprendizaje supervisado. La idea principal de las SVM es encontrar un hiperplano que separe dos clases con un margen máximo. El entrenamiento El problema en SVM se puede representar como un problema de optimización de programación cuadrática. UN implementación común de SVM que se utiliza en muchos trabajos (Mittermayer, 2004; Pui Cheong Fung, et al., 2003) es SVM Light (T. Joachims, 1999). SVM Light es una implementación de un alumno de SVM que aborda el problema de las tareas grandes (T. Joachims, 1999). Él los algoritmos de optimización utilizados en SVM Light se describen en T. Joachims (2002). Otro La implementación más utilizada de SVM es LIBSVM (Chang & Lin, 2011). implementos LIBSVM un algoritmo de tipo SMO (Sequential Minimal Optimization) propuesto en un artículo por Fan, Chen, y Lin (2005). Soni, et al. (2007) está utilizando esta implementación para la predicción del precio de las acciones movimientos Aparte de la implementación, la entrada a la SVM es bastante única en el trabajo de Soni, et al. (2007). Toman las coordenadas de las noticias en forma visualizada document-map como características. Un documento-mapa es un espacio de baja dimensión en el que cada

la noticia se posiciona sobre la media ponderada de las coordenadas de los conceptos que aparecen en la noticia. Las SVM se pueden extender a clasificadores no lineales aplicando kernel mapeo (truco del núcleo). Como resultado de aplicar el mapeo del núcleo, la clasificación original problema se transforma en un espacio dimensional superior. SVM que representan lineal clasificadores en este espacio de alta dimensión pueden corresponder a clasificadores no lineales en el espacio de características originales (Burges, 1998). La función kernel utilizada puede influir en la rendimiento del clasificador. Zhai, et al. (2007) están utilizando SVM con un kernel RBF gaussiano y un núcleo polinomial.

B) Algoritmos de regresión: toman diferentes formas en los esfuerzos de investigación como se enumeran en la tabla 4. Un enfoque es la regresión del vector de soporte (SVR), que es una variación basada en la regresión de SVM (Drucker, Burges, Kaufman, Smola y Vapnik, 1997). Es utilizado por Hagenau, et al. (2013) y Schumaker, et al. (2012).

Hagenau, et al. (2013) utilizan principalmente SVM, pero también evalúan la capacidad de predecir la valor discreto del rendimiento de las acciones usando SVR. Predicen rendimientos y calculan el R2 (coeficiente de correlación al cuadrado) entre la rentabilidad prevista y la realmente observada. Él optimización detrás de la SVR es muy similar a la SVM, pero en lugar de una medida binaria (es decir, positivo o negativo), se entrena sobre los rendimientos realmente observados. Mientras que una medida binaria puede ser 'verdadero' o 'falso', esta medida da más peso a las mayores desviaciones entre rendimientos reales y previstos que a los más pequeños. Como las ganancias o pérdidas son mayores con mayores desviaciones, esta medida captura mejor los rendimientos comerciales reales que se realizarán (Hagenau, et al., 2013).

Schumaker, et al. (2012) eligen implementar la Optimización Mínima Secuencial SVR (Platt, 1999) funcionan a través de Weka (Witten & Frank, 2005). Esta función permite discreta predicción numérica en lugar de clasificación. Seleccionan un núcleo lineal y una cruz de diez veces. validación.

A veces se utilizan directamente modelos de regresión lineal (Chatrath, et al., 2014; Jin, et al., 2013; Paul C. Tetlock, et al., 2008). Paul C. Tetlock, et al. (2008) utilizan MCO (Mínimos cuadrados ordinarios) método para estimar los parámetros desconocidos en el modelo de regresión lineal. Ellos usan dos variables dependientes diferentes (rendimientos crudos y anormales al día siguiente) retrocedidos en diferentes medidas de palabras negativas. Su resultado principal es que las palabras negativas en los términos específicos de la empresa las noticias predicen sólidamente rendimientos ligeramente más bajos en el siguiente día de negociación.

Chatrath, et al. (2014) utilizan una regresión multivariante paso a paso en un Probit (unidad de probabilidad) modelo. El propósito del modelo es estimar la probabilidad de que una observación con las características particulares caerán en una categoría específica. En este caso se trata de averiguar el probabilidad de comunicados de prensa que resulten en saltos. Jin, et al. (2013) aplican agrupación de temas y utilice diccionarios de opiniones personalizados para descubrir tendencias de opiniones analizar oraciones relevantes. Un modelo de regresión lineal estima el peso de cada tema y hace pronósticos de divisas.

C) Naïve Bayes: Es el siguiente algoritmo utilizado en un grupo de trabajos de la tabla 4. Probablemente sea el algoritmo de clasificación más antiguo (Lewis, 1998). Pero todavía es muy popular y se usa entre muchos de los trabajos (Groth & Muntermann, 2011; F. Li, 2010; Werner & Myrray Z., 2004; Wuthrich, et al., 1998). Se basa en el Teorema de Bayes y se llama ingenuo porque es basado en la suposición ingenua de la independencia completa entre las características del texto. Eso se diferencia de enfoques como k-Nearest Neighbors (k-NN), Artificial Neural Redes (ANN) o Máquina de vectores de soporte (SVM) en el sentido de que se basa en probabilidades (de un característica perteneciente a una determinada categoría) mientras que los otros enfoques mencionados interpretan espacialmente la matriz de características del documento.

Y. Yu, et al. (2013) aplican el algoritmo Naïve Bayes (NB) para realizar un análisis de sentimiento para Examinar el efecto de múltiples fuentes de medios sociales junto con el efecto de los medios convencionales. medios e investigar su importancia relativa y su interrelación. F. Li (2010)

utiliza Naïve Bayes para examinar el contenido de la información de las declaraciones prospectivas en la sección Discusión y análisis de gestión de los archivos de la empresa. Él usa el Niave Módulo Bayes en el lenguaje de programación Perl para realizar el cálculo.

D) Reglas y árboles de decisión: Es el siguiente grupo de algoritmos utilizados en la literatura como se indica en la tabla 4. Algunos de los investigadores han hecho un esfuerzo para crear una clasificación basada en reglas (C.-J. Huang, et al., 2010; Peramunetilleke & Wong, 2002; Vu, et al., 2012).

Peramunetilleke y Wong (2002) utilizan un conjunto de palabras clave proporcionadas por un dominio financiero experto . El clasificador que expresa la correlación entre las palabras clave y una de las los resultados es un conjunto de reglas. Cada uno de los tres conjuntos de reglas (DOLLAR_UP, DOLLAR_STEADY, DOLLAR_DOWN) produce una probabilidad que indica la probabilidad de que ocurra el evento respectivo en relación con las palabras clave disponibles.

C.-J. Huang, et al. (2010) observan que la combinación de dos o más palabras clave en un El titular de las noticias financieras podría jugar un papel crucial en el próximo día de negociación. Así aplicaron algoritmo de reglas de asociación ponderadas para detectar los términos compuestos importantes en las noticias titulares

Rachlin, et al. (2007) utilizan un algoritmo de inducción de árboles de decisión, que no supone independencia de atributos. El algoritmo es C4.5 desarrollado por Quinlan (Quinlan, 1993). Este El algoritmo produce un conjunto de reglas de predicción de tendencias. Además, muestran el efecto de la combinación entre datos numéricos y textuales. Vú, et al. (2012) también utilizan la decisión C4.5 árbol para el problema de clasificación binaria de texto para predecir los cambios diarios hacia arriba y hacia abajo en los precios de las acciones.

Para los categorizadores de reglas de decisión, las reglas se componen de palabras y las palabras tienen significado, las reglas mismas pueden ser perspicaces. Más que simplemente intentar asignar una etiqueta, un conjunto de las reglas de decisión pueden resumir cómo tomar decisiones. Por ejemplo, las reglas pueden sugerir una

patrón de palabras que se encuentran en los cables de noticias antes de la subida del precio de una acción. Lo malo de las reglas es que pueden ser menos predictivos si el concepto subyacente es complejo (Weiss, Indurkhya, & Zhang, 2010). Aunque las reglas de decisión pueden ser soluciones particularmente satisfactorias para la minería de textos, los procedimientos para encontrarlos son más complicados que otros métodos (Weiss, et al., 2010).

Los árboles de decisión son reglas de decisión especiales que se organizan en una estructura de árbol. Una decisión El árbol divide el espacio del documento en regiones que no se superponen en sus hojas y las predicciones se realizan en cada hoja (Weiss, et al., 2010).

E) Algoritmos combinatorios: En la tabla 4, se hace referencia a una clase de algoritmos que son compuesto por una serie de algoritmos de aprendizaje automático apilados o agrupados. das y Chen (2007) han combinado varios algoritmos de clasificación mediante una votación sistema para extraer el sentimiento de los inversores. Los algoritmos son, a saber, Naive Classifier, Vector Clasificador de distancia, Clasificador basado en discriminantes, Clasificador de frase adjetivo-adverbio, Clasificador Bayesiano. Los niveles de precisión resultan ser similares a los clasificadores Bayes ampliamente utilizados, pero los falsos positivos son más bajos y la precisión del sentimiento es más alta.

Mahajan, et al. (2008) identifican y caracterizan los principales eventos que impactan en el mercado utilizando un Mecanismo de extracción de temas basado en la asignación de Dirichlet latente (LDA). Entonces un clasificador apilado se utiliza, que es un clasificador entrenable que combina las predicciones de múltiples clasificadores a través de un procedimiento de votación generalizado. El paso de votación es un problema de clasificación separado. Ellos usar un árbol de decisión basado en la ganancia de información para manejar atributos numéricos en junto con un SVM con kernel sigmoide para diseñar el clasificador apilado. La media la precisión del sistema de clasificación es del 60%.

Butler y Kešelj (2009) proponen 2 métodos y luego los combinan para lograr el mejor actuación. El primer método se basa en perfiles de n-gramas de caracteres, que se generan

para cada informe anual de la empresa, y luego etiquetados según el Common N-Gram (CNG)
clasificación. El segundo método combina puntajes de legibilidad con entradas de rendimiento y
luego los envía a una máquina de vectores de soporte (SVM) para su clasificación. El combinado
La versión está configurada para tomar decisiones solo cuando los modelos estén de acuerdo.

Bollen y Huina (2011) implementan un modelo de red neuronal difusa autoorganizada (SOFNN) para
probar la hipótesis de que incluir mediciones del estado de ánimo del público puede mejorar la precisión de
Modelos de predicción del Dow Jones Industrial Average (DJIA). Una red neuronal difusa es un aprendizaje
máquina que encuentra los parámetros de un sistema difuso (es decir, conjuntos difusos, reglas difusas) por
explotando técnicas de aproximación a partir de redes neuronales. Por lo tanto, se clasifica como un
algoritmo combinatorio de la tabla 4.

F) Experimentos multi-algoritmo: Es otra clase de trabajos en la tabla 4, por lo que el mismo
Los experimentos se llevan a cabo utilizando varios algoritmos diferentes.

Wuthrich, et al. (1998) es uno de los primeros trabajos de investigación en esta área. no se apilan
múltiples algoritmos juntos para formar un algoritmo más grande. Sin embargo, realizan su
experimentos utilizando múltiples algoritmos y comparar los resultados.

Werner y Myrray Z. (2004) también realizan todas las pruebas utilizando dos algoritmos, Naïve Bayes y
SVM. Además, Groth y Muntermann (2011) emplean Naïve Bayes, k-Nearest
Vecino (k-NN), Redes Neuronales Artificiales (ANN) y SVM para detectar patrones en
los datos textuales que podrían explicar la mayor exposición al riesgo en los mercados bursátiles.

En conclusión, SVM se ha utilizado extensa y exitosamente como una clasificación textual y
enfoque de aprendizaje de sentimientos, mientras que otros enfoques como las redes neuronales artificiales (ANN), K
vecinos más cercanos (k-NN) rara vez se han considerado en la literatura de minería de texto para el mercado
predicción. Esto también es confirmado por Moraes, Valiati y Gavião Neto (2013). su investigacion
presenta una comparación empírica entre SVM y ANN con respecto al sentimiento a nivel de documento

análisis. Sus resultados muestran que ANN puede producir resultados superiores o al menos comparables a los de SVM. Dichos resultados pueden proporcionar motivos para investigar el uso de otros algoritmos además del actual. mayormente utilizado SVM. CH Li, Yang y Park (2012) demuestran un alto rendimiento de k-NN para texto categorización, así como Jiang, Pang, Wu y Kuang (2012). Tan, Wang y Wu (2011) afirman que para la categorización de documentos, el clasificador centroide funciona ligeramente mejor que el clasificador SVM y lo supera en tiempo de ejecución también. Gradojevic y Gençay (2013) presentan una rara investigación que utiliza lógica difusa para mejorar las señales técnicas de negociación con éxito en los tipos de cambio EUR-USD pero difusa la lógica rara vez se usa en los trabajos revisados para la predicción de marcadores basada en minería de texto. Sólo Bollen y Huina (2011) lo utilizan en combinación con redes neuronales para diseñar un sistema neuronal difuso autoorganizado. red (SOFNN) con éxito. Sin embargo, Loia y Senatore (2014) muestran que puede ser muy útil para modelado de emociones en general. Exploración de tales algoritmos poco investigados en el contexto de la predicción del mercado puede conducir a nuevos conocimientos que pueden ser de interés para futuros investigadores.

Para comprender mejor los sistemas revisados en los que los algoritmos de aprendizaje automático han utilizado, un número adicional de propiedades del sistema se han revisado en la tabla 4 que son explicado en las siguientes secciones.

3.4.2. Entrenamiento versus volumen de prueba y muestreo

En esta columna, en la tabla 4, se resumen dos aspectos si se dispusiera de la información. En primer lugar, el volumen de los ejemplos que se utilizaron para el entrenamiento frente al volumen utilizado para la prueba; alrededor de 70 o 80 por ciento para entrenamiento vs. 30 o 20 por ciento para pruebas parece ser la norma. En segundo lugar, si el el muestreo para entrenamiento y prueba era de un tipo especial; lo que aquí interesa especialmente es saber si Se ha seguido un muestreo lineal ya que, en esencia, las muestras se encuentran en una serie temporal. Algunos de los los trabajos han mencionado claramente el tipo de muestreo como Estratificado (Groth & Muntermann, 2011) mientras que la mayoría de los demás, sorprendentemente, no han mencionado nada en absoluto.

3.4.3. Ventana deslizante

El objetivo general de los sistemas revisados es predecir el movimiento del mercado en un tiempo futuro
ventana (ventana de predicción) basada en el aprendizaje obtenido en una ventana de tiempo pasada (ventana de entrenamiento)
donde se entrenan los algoritmos de aprendizaje automático y se reconocen los patrones.

Por ejemplo, un sistema puede aprender patrones en función de los datos disponibles durante varios días (entrenamiento
ventana) para predecir el movimiento del mercado en un nuevo día (ventana de predicción). La longitud
y la ubicación de la ventana de entrenamiento en la línea de tiempo puede tener dos formatos posibles: fijo o deslizante.
Si la ventana de entrenamiento es fija, el sistema aprende en función de los datos disponibles desde el punto 'A' hasta el punto
'B' en la línea de tiempo y esos 2 puntos son fijos; por ejemplo, desde la fecha 'A' hasta la fecha 'B'. En semejante
escenario el aprendizaje resultante de la ventana de entrenamiento se aplica a la ventana de predicción
independientemente de en qué parte de la línea de tiempo se encuentre la ventana de predicción. Puede ser justo después de la
ventana de entrenamiento o puede estar más lejos en el futuro con una distancia de la ventana de entrenamiento. Está
obvio que si hay una gran distancia entre la ventana de entrenamiento y la ventana de predicción, el
el aprendizaje capturado en el algoritmo de aprendizaje automático puede no estar actualizado y, por lo tanto, ser preciso
suficiente porque la información disponible en el gap no se utiliza para la formación.

Por lo tanto, se introduce un segundo formato para resolver el problema anterior mediante el cual toda la formación
ventana o un lado de ella (el lado al final) es capaz de deslizarse dinámicamente hasta el punto donde
se inicia la ventana de predicción. En otras palabras, si la ventana de entrenamiento comienza en el punto 'A' y termina en
punto 'B' y la ventana de predicción comienza en el punto 'C' y termina en el punto 'D'. En la ventana corredera
formato, el sistema siempre se asegura de que el punto 'B' siempre esté justo antes y adyacente al punto 'C'. Este
El enfoque se denomina simplemente en este trabajo como “ventana deslizante”. Los trabajos revisados que hacen
poseen una ventana corredera como propiedad del diseño de su sistema están identificados y marcados en la tabla 4
debajo de una columna con el mismo nombre.

Aunque, intuitivamente, parece necesario implementar una ventana deslizante, hay muy pocos de los
trabajos revisados que realmente tienen (Butler & Kešelj, 2009; Jin, et al., 2013; Peramunetilleke &

Wang, 2002; Paul C. Tetlock, et al., 2008; Wuthrich, et al., 1998). Este parece ser un aspecto que puede recibir más atención en los sistemas futuros.

3.4.4. Semántica y Sintaxis

En el Procesamiento del Lenguaje Natural (PNL) se investigan atentamente dos aspectos del lenguaje: La semántica y sintaxis. En pocas palabras: la semántica se ocupa del significado de las palabras y la sintaxis se ocupa de su orden. y posicionamiento relativo o agrupación. En esta sección: En primer lugar, se analiza más de cerca el significado de cada uno de ellos y algunos de los trabajos de investigación recientes relacionados. En segundo lugar, se informa si y cómo se han observado en los trabajos de minería de textos revisados para la predicción del mercado.

Abordar la semántica es un tema importante y los esfuerzos de investigación se dedican a ello en una serie de frentes Es importante desarrollar ontologías especializadas para contextos específicos como las finanzas; Lupiani Ruíz, et al. (2011) presentan un buscador semántico de noticias financieras basado en la Web Semántica tecnologías El motor de búsqueda está acompañado por una herramienta de población de ontologías que ayuda a mantener actualizada la ontología financiera. Además, la semántica se puede incluir en el diseño de esquemas de ponderación de características; Luo, Chen y Xiong (2011) proponen un nuevo esquema de ponderación de términos mediante explotando la semántica de las categorías y los términos de indexación. Específicamente, la semántica de las categorías están representados por los sentidos de los términos que aparecen en las etiquetas de categoría, así como la interpretación de ellos por WordNet (Miller, 1995). Además, en su trabajo el peso de un término está correlacionado con su semántica similitud con una categoría. WordNet (Miller, 1995) proporciona una red semántica que vincula los sentidos de palabras entre sí. Los principales tipos de relaciones entre los synsets de WordNet son los super relaciones de subordinación que son la hiperonimia y la hiponimia. Otras relaciones son la meronimia y la holonimia (Loia & Senatore, 2014). Es fundamental facilitar las relaciones semánticas de los términos para obtener un resultado satisfactorio en la categorización del texto; CH Li, et al. (2012) muestran un alto rendimiento del texto categorización en la que las relaciones semánticas de los términos se basan en dos tipos de tesauros, un corpus Se buscaron tesauros basados en (CBT) y WordNet (WN). Cuando una combinación de CBT y WN fue utilizados, obtuvieron el mayor nivel de desempeño en la categorización del texto.

La sintaxis también es muy importante y su adecuada observación y utilización junto con la semántica (o a veces en lugar de ello) puede mejorar la precisión de la clasificación textual; Kim, et al. (2014) proponen un kernel novedoso, llamado kernel semántico independiente del lenguaje (LIS), que es capaz de calcular de manera efectiva la similitud entre documentos de texto corto sin usar etiquetas gramaticales y bases de datos léxicas.

A partir de los resultados del experimento en conjuntos de datos en inglés y coreano, se muestra que el núcleo LIS tiene mejor rendimiento que varios núcleos existentes. Este es esencialmente un patrón basado en la sintaxis.

método de extracción Es interesante notar que hay varios enfoques para este tipo basado en la sintaxis.

métodos de reconocimiento de patrones: en el método de aparición de palabras, un patrón se considera como una palabra que aparece en un documento (Thorsten Joachims, 1998). En el método de secuencia de palabras, consiste en un conjunto de palabras consecutivas que aparecen en un documento (Lodhi, Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002). En el método del árbol de análisis, que se basa en la estructura sintáctica, se crea un patrón extraído del árbol de un documento considerando no sólo la ocurrencia de la palabra sino también la secuencia de palabras en el documento (Collins & Duffy, 2001).

Duric y Song (2012) proponen un conjunto de nuevos esquemas de selección de características que utilizan un Contenido y Modelo de sintaxis para aprender automáticamente un conjunto de características en un documento de revisión separando el entidades que están siendo revisadas de las expresiones subjetivas que describen esas entidades en términos de polaridades. Los resultados obtenidos al utilizar estas características en un clasificador de máxima entropía son competitivo con los enfoques de aprendizaje automático de última generación (Duric & Song, 2012). Tema modelos como Latent Dirichlet Allocation (LDA) son modelos generativos que permiten que los documentos sean explicado por temas no observados (latentes). El Modelo Oculto de Markov LDA (HMM-LDA) (Griffiths, Steyvers, Blei, & Tenenbaum, 2005) es un modelo de tópicos que modela simultáneamente tópicos y sintácticas estructuras en una colección de documentos. La idea detrás del modelo es que una palabra típica puede jugar diferentes roles Puede ser parte del contenido y servir con un propósito semántico (tópico) o puede utilizarse como parte de la estructura gramatical (sintáctica). También se puede utilizar en ambos contextos (Duric & Canción, 2012). HMM-LDA modela este comportamiento induciendo clases sintácticas para cada palabra basadas en cómo aparecen juntos en una oración usando un Modelo Oculto de Markov. Cada palabra se asigna a un

clase sintáctica, pero una clase está reservada para las palabras semánticas. Las palabras de esta clase se comportan como lo haría en un modelo de tema LDA regular, participando en diferentes temas y teniendo ciertas probabilidades de aparecer en un documento (Duric & Song, 2012).

En la tabla 4, hay una columna dedicada a cada uno de estos aspectos, a saber: Semántica y Sintaxis. Alrededor de la mitad de los sistemas utilizan algún aspecto semántico en su enfoque de minería de texto, que es generalmente se hace usando un diccionario o diccionario de sinónimos y categorizando las palabras según su significado pero ninguno avanza en las direcciones señaladas anteriormente. Además, muy pocos trabajos han hecho una esfuerzo por incluir la sintaxis, es decir, el orden y el papel de las palabras. Estos enfoques básicos y algo indirectos son sintagmas nominales (Schumaker & Chen, 2009), combinaciones de palabras y n-gramas (Hagenau, et al., 2013) y aparición simultánea de palabras (C.-J. Huang, et al., 2010) y los “trillizos” que consisten en un adjetivo o adverbio y las dos palabras inmediatamente posteriores o anteriores (Das & Chen, 2007). Algunos trabajos como Vu, et al. (2012) incluyen el etiquetado de parte del discurso (POS) como una forma de atención a la sintaxis. Loia y Senatore (2014) logran un análisis de sentimiento a nivel de frase teniendo en cuenta cuenta cuatro categorías sintácticas, a saber: sustantivos, verbos, adverbios y adjetivos. La necesidad de profundidad El análisis sintáctico para el análisis de sentimiento a nivel de frase ha sido investigado por Kanayama y Nasukawa (2008).

3.4.5. Combinar noticias y datos técnicos o señales

Es posible pasar datos técnicos o señales junto con las características del texto en la clasificación algoritmo como variables independientes adicionales. Ejemplos de datos técnicos podrían ser un precio o un índice nivel en un momento dado. Las señales técnicas son los resultados de algoritmos técnicos o reglas como la promedio móvil, reglas de fuerza relativa, reglas de filtro y reglas de ruptura del rango comercial. pocos de los Los investigadores han aprovechado estos aportes adicionales como se indica en la tabla 4 (Butler & Kešelj, 2009; Hagenau, et al., 2013; Rachlin, et al., 2007; Schumaker y Chen, 2009; Schumaker, et al., 2012; Zhai, et al., 2007).

3.4.6. software usado

Por último, es interesante observar cuáles son algunas de las aplicaciones comunes de terceros que son utilizados para la implementación de los sistemas. En esta columna de la tabla 4 el lector puede ver los nombres de las piezas de software que fueron utilizadas por los trabajos revisados como parte de su preprocesamiento o aprendizaje automático. En su mayoría son diccionarios (F. Li, 2010; Paul C. Tetlock, et al., 2008), clasificación implementaciones de algoritmos (Butler & Kešelj, 2009; Werner & Myrray Z., 2004), extracción de conceptos y paquetes de combinación de palabras (Das & Chen, 2007; Rachlin, et al., 2007) o valor de sentimiento proveedores (Schumaker, et al., 2012).

Referencia	Algoritmo Tipo	Algoritmo Detalles	Entrenamiento versus volumen de prueba y muestreo	Corredizo Ventana	Sema -nticas	Noticias de sintaxis y tecnología datos	Software
Pui Cheong Fung, et al. (2003)	MVS	SVM-Luz	Primeros 6 meses consecutivos vs. el mes pasado	No	No	No	No mencionado
Mittermayer (2004)		SVM-Luz	200 frente a 6002 ejemplos 80 %	No	No	No	No	NoticiasGATOS
Soni, et al. (2007)		SVM con núcleo lineal estándar	frente a 20 %	No	Sí	No	No	Paquete LibSVM
Zhai, et al. (2007)		SVM con FBR gaussiano núcleo y núcleo polinomial	primeros 12 meses contra el dos meses restantes	No	Sí	No	Sí	No mencionado
schumaker y Chen (2009)		MVS	No mencionado No		Sí	Sí	Sí	Arizona Extractor de texto (AzTeK) y texto AZFin.
Lugmayr y Gossen (2012)		MVS	No mencionado No		Sí	No	Sí	SentiPalabra Red
Hagenau y otros (2013)	Regresión Algoritmos	SVM con un núcleo lineal, RVS	No mencionado No		Sí	Sí	Sí	No mencionado
Schumaker, et al. (2012)		RVS	No mencionado No		Sí	No	Sí	OpiniónFin la
Jin, et al. (2013)		Lineal Modelo de regresión	Día anterior frente a un día determinado (2 semanas para regresión)	Sí	Sí	No	No	divisas pronosticador, Loughran mcdonald dic. financiera, dic. AFINN.
Chattrath, et al. (2014)		paso a paso multivariante Regresión	No aplica No		No	No	No	No mencionado

		Modelo						
pablo c Tetlock, et al. (2008)		MCO Regresión	30 y 3 días hábiles antes de un anuncio de ganancias	Sí	Sí	No	No	Harvard IV 4 psicosocial al diccionario
Y. Yu, et al. (2013)	Naïve Bayes Naïve Bayes		No mencionado No		Sí	No	No	Abierto fuente Natural Idioma Caja de herramientas (NLTK)
F. Li (2010)		Naïve Bayes y basado en diccionario	30,000 al azar contra sí mismo y el descanso.	No	No	No	No	Dicción, General Investigador, el Lingüístico Consulta, Palabra Recuento (LIWC).
Peramunetilleke y Wong (2002)	Decisión Reglas o Árboles	Clasificador de reglas	22 sept 12:00 a 27 sept 9:00 vs. 9:00 a 10:00 el 27 septiembre	Sí	Sí	No	No	No mencionado
C.-J. Huang, et al. (2010)		asociación ponderada normas	2005 junio a 2005 octubre vs. 2005 noviembre	No	Sí	Sí	No	No mencionado
Rachlin, et al. (2007)		C4.5 Decisión Árbol	No mencionado.	No	No	No	Sí	Extractor Paquete de software
Vu, et al. (2012)		C4.5 Decisión Árbol	Entrenado por características del día anterior	Sí	sí Sí		No	FRC++ caja de herramientas, Manguera, TST, CMU TPV etiquetas, AltaVista
Eso y Chen (2007)	Combinacional Algoritmos	Combinación de diferentes clasificadores	1000 contra el resto	No	Sí	Sí	No	General investigador
Mahajan, et al. (2008)		clasificador apilado	05 de agosto - 07 de diciembre frente a enero 08 – 08 de abril	No	Sí	No	No	No mencionado
mayordomo y Keselj (2009)		distancia GNC medida & SVM y combinado	año x contra años x ÿ 1 y x ÿ 2. y todos los vectores representación s vs. año de prueba particular	Sí	No	No	Sí	Perl n-grama módulo Texto :: Ngra milisegundo desarrollado por Keselj. LIBSVM
Bollen y Huiña (2011)		red neuronal difusa autoorganizada (SOFNN)	28 febrero al 28 noviembre vs. 1 a 19 de diciembre de 2008	No	N/A/A		No	GPOMS, OpiniónFi honor
Wuthrich, et al. (1998)	Multi algoritmo	k-NN, ANN, naïve Bayes,	últimos 100 Días de entrenamiento	Sí	Sí	No	No	No mencionado

	experimentos basados en reglas		pronosticar 1 dia					
Werner y Myrray Z. (2004)		bayesiano ingenuo, MVS	1,000 mensajes vs. el resto	No	No	No	No	paquete arcoiris
Groth y Müntermann (2011)		Naïve Bayes, k-NN, ANN, MVS	Cruz estratificada validaciones	No	No	No	No	No mencionado

Tabla 4 Algoritmos de clasificación y otros aspectos de aprendizaje automático

3.5. Hallazgos de los trabajos revisados

En la tabla 5 se han repasado los mecanismos de evaluación y los nuevos hallazgos de cada pieza de investigación.

La mayoría de los trabajos están presentando una matriz de confusión o partes de la misma para presentar sus resultados. Y calcular la exactitud, recuperación o precisión y, a veces, la medida F, siendo la exactitud la más común. La precisión en la mayoría de los casos se reporta en el rango de 50 a 70 por ciento, mientras que argumentando mejores resultados que el azar, que se estima en un 50 por ciento (Butler & Kešelj, 2009; F. Li, 2010; Mahajan, et al., 2008; Schumaker & Chen, 2009; Schumaker, et al., 2012; Zhai, et al., 2007). It es un enfoque de evaluación común y los resultados superiores al 55% se han considerado dignos de informe en otras partes de la literatura también (Garcke, Gerstner y Griebel, 2013). Sin embargo, lo que hace más de los resultados cuestionables es que la mayoría de ellos sorprendentemente no han examinado o informado si los datos de su experimento están desequilibrados o no. Como esto es importante en la minería de datos (Duman, Ekinci, & Tanrÿverdi, 2012; Thammasiri, Delen, Meesad y Kasap, 2014) se ha colocado una columna adicional en la tabla 5 para verificar esto. Entre los trabajos revisados solo Soni, et al. (2007), Mittermayer (2004) y Peramunetilleke y Wong (2002) han prestado cierta atención a este tema en sus trabajos. Está También es crucial tener en cuenta si un conjunto de datos desequilibrado con clases desequilibradas se encuentra especialmente con un alta dimensionalidad en el espacio de características, ideando una selección de características adecuada que pueda lidiar tanto con los datos desequilibrados como con la alta dimensionalidad se vuelve crítico. Selección de características para datos desequilibrados de alta dimensión se amplifica en detalle en el trabajo de Yin, Ge, Xiao, Wang y Quan (2013). Liu, Loh y Sun (2009) abordan el problema de los datos textuales desequilibrados utilizando un simple esquema de ponderación de términos basado en la probabilidad para distinguir mejor los documentos en categorías menores.

Smales (2012) examina la relación entre el desequilibrio de pedidos y las noticias macroeconómicas en el contexto del mercado de futuros de tasas de interés de Australia e identificar nueve principales macroeconómicos anuncios con impacto en el desequilibrio de pedidos.

Otro enfoque de evaluación popular además del anterior para la mitad de los trabajos revisados es el ensamblaje de una estrategia o motor comercial (Groth & Muntermann, 2011; Hagenau, et al., 2013; C.-J. Huang, et al., 2010; Mittermayer, 2004; Pui Cheong Fung, et al., 2003; Rachlin, et al., 2007; Schumaker & Chen, 2009; Schumaker, et al., 2012; Paul C. Tetlock, et al., 2008; Zhai, et al., 2007). A través del cual se simula un periodo de negociación y se miden las ganancias para evaluar la viabilidad del sistema.

En general, los investigadores utilizan mecanismos de evaluación y datos experimentales que varían ampliamente y esto hace inalcanzable una comparación objetiva en términos de niveles concretos de efectividad.

Referencia	Recomendaciones	Estrategia comercial	datos equilibrados
Wuthrich, et al. (1998)	Ftse 42%, Nky 47%, Dow 40% Hsi 53% y Sti 40%. Sí		No mencionado
Peramunetilleke y Wong (2002)	Mejor que la oportunidad	No	Sí
Pui Cheong Fung, et al. (2003)	La ganancia acumulada de monitorear múltiples series de tiempo es casi el doble que la de monitorear una sola serie de tiempo.	Sí	No mencionado
Werner y Myrray Z. (2004)	Evidencia de que los mensajes de acciones ayudan a predecir la volatilidad del mercado, pero no los rendimientos de las acciones.	No	No
Mittermayer (2004)	Beneficio promedio 11% en comparación con el beneficio promedio por comerciante aleatorio 0%	Sí	Sí
Das y Chen (2007)	La regresión tiene poco poder explicativo	No	No mencionado
Soni, et al. (2007)	Tasa de aciertos del clasificador: 56,2 % en comparación con el 47,5 % del clasificador ingenuo y el 49,1 % de la bolsa de palabras SVM	No	Sí (más o menos)
Zhai, et al. (2007)	Precio 58,8%, Noticias directas 62,5%, Noticias indirectas 50,0%, Noticias combinadas 64,7%, Price & News 70,1% Beneficio: por Price & News 5,1% en 2 meses y por Price y News solo alrededor de la mitad cada uno	Sí	No mencionado
Rachlin, et al. (2007)	No se puede mejorar la precisión predictiva del análisis numérico. Precisión del 82,4 % para análisis combinado textual y numérico, del 80,6 % para análisis textual y del 83,3 % solo numérico.	Sí	No mencionado

Paul C. Tetlock, et al. (2008)	1) la fracción de palabras negativas en noticias específicas de la empresa pronostica bajas ganancias de la empresa; 2) los precios de las acciones de las empresas reaccionan brevemente a la información incrustada en palabras negativas; y 3) la previsibilidad de las ganancias y el rendimiento de las palabras negativas es mayor para las historias que se centran en los fundamentos.	Sí	Irrelevante
Mahajan, et al. (2008)	Precisión 60%	No	No mencionado
mayordomo y Keselj (2009)	Primer método: 55 % y 59 % para la precisión de gramos de caracteres y gramos de palabras, respectivamente, aún superior a la cartera de referencia. Segundo método: la precisión general y la precisión de rendimiento superior fueron 62,81% y 67,80% respectivamente.	No	No mencionado
Schumaker y Chen (2009)	Precisión direccional 57,1 % , Retorno 2.06% , Cercanía 0,04261 Precisión de	Sí	No mencionado
F. Li (2010)	tono 67 % y contenido 63 % con naïve Bayes y menos del 50 % con diccionario.	No	No
C.-J. Huang, et al. (2010)	Precisión de predicción y tasa de recuperación de hasta 85,2689 % y 75,3782 % en promedio, respectivamente.	Sí	No mencionado
Groth y Müntermann (2011)	Precisión (ligeramente) por encima del 75 % de referencia equivalente de adivinación.	Sí	No
Schumaker, et al. (2012)	Los artículos objetivos tenían un desempeño deficiente en precisión direccional versus línea de base. Los artículos neutrales tuvieron rendimientos comerciales más bajos en comparación con la línea de base. Los artículos subjetivos se comportaron mejor con una precisión direccional del 59,0 % y un rendimiento comercial del 3,30 %. La polaridad tuvo un desempeño deficiente en comparación con la línea de base.	Sí	No mencionado.
Lugmayr y Gossen (2012)	En curso	No	No mencionado
Y. Yu, et al. (2013)	Polaridad con una precisión del 79 % y una medida F de 0,86 en el equipo de prueba. Solo el número total de cuentas de redes sociales tiene una relación positiva significativa con el riesgo, pero no con el rendimiento. Se muestra que el término de interacción tiene una relación marginalmente negativa con el rendimiento, pero una relación significativa altamente negativa con el riesgo.	No	No mencionado
Hagenau y otros (2013)	Selección de características basada en comentarios combinada con 2-combinaciones de palabras lograron precisiones de hasta el 76%	Sí	No
Jin, et al. (2013)	Precisión en torno a 0,28 de media.	No	No mencionado
Chatrath, et al. (2014)	(a) los saltos son un buen indicador de la llegada de noticias en mercados de divisas; (b) hay una reacción sistemática de los precios de las divisas a las sorpresas económicas; y (c) los precios responden rápidamente dentro de los 5 minutos del comunicado de prensa	No	No mencionado
Bollen y Huiña (2011)	El estado de ánimo 'Calma' tuvo la relación de causalidad de Granger más alta con el DJIA para intervalos de tiempo que oscilaron entre dos y seis días (valores de $p < 0,05$). Las otras cuatro dimensiones del estado de ánimo de GPOMS y OpinionFinder no tuvieron una correlación significativa con los cambios en el mercado de valores.	No	No mencionado

Vú, et al. (2012)	Combinación del movimiento de precios de días anteriores, Las características Bullish/bearish y Pos_Neg crean un modelo superior en las 4 empresas con precisiones de: 82,93 %, 80,49 %, 75,61 % y 75,00 % y para la prueba en línea como: 76,92 %, 76,92 %, 69,23 % y 84,62 %	NO	No mencionado

Tabla 5 Hallazgos de los trabajos revisados, existencia de una estrategia comercial y datos balanceados

4. Sugerencias para trabajos futuros

Los mecanismos de predicción del mercado basados en la minería de textos en línea están emergiendo para ser investigados. utilizando rigurosamente el pico radical de la potencia de procesamiento computacional y la velocidad de la red en el tiempos recientes. Prevemos que esta tendencia continuará. Esta investigación ayuda a poner en perspectiva el papel de reacciones humanas a los eventos en la formación de los mercados y puede conducir a una mejor comprensión de eficiencias de mercado y convergencia a través de la absorción de información. En resumen, este trabajo identifica los a continuación como áreas o aspectos que necesitan futuras investigaciones y avances:

A. Semántica: Como se discutió en la sección 3.4.4, los avances de las técnicas en semántica son crucial para el problema de clasificación de texto en cuestión, ya que los investigadores de minería de texto ya han mostrado. Sin embargo, tales avances aún no han entrado en el campo del mercado.
Minería predictiva de texto. Desarrollo de ontologías especializadas mediante la creación de nuevas o la personalización de diccionarios actuales como WordNet requiere más atención. Mucho de trabajos de investigación actuales todavía se centran demasiado en los métodos de aparición de palabras y rara vez incluso usa WordNet. Además, las relaciones semánticas pueden investigarse con diferentes objetivos, desde la definición de esquemas de ponderación para la representación de características hasta la semántica compresión o abstracción para la reducción de características. Probablemente desde el texto predictivo del mercado la minería en sí es un campo emergente, los investigadores aún no se han sumergido en la semántica mejora para el contexto predictivo del mercado específicamente.

B. Sintaxis: las técnicas de análisis sintáctico probablemente han recibido menos atención que los semánticos, como también se discutió en la sección 3.4.4. Técnicas más avanzadas basadas en la sintaxis como el uso de árboles de análisis para el reconocimiento de patrones en el texto puede mejorar la calidad del texto-

la minería de manera significativa. Este aspecto requiere también la atención de futuros investigadores, comenzando con el intento de transferir parte del aprendizaje en otras áreas de minería de texto como revisiones clasificación en minería de texto predictiva de mercado.

C. Sentimiento: el análisis de sentimientos y emociones ha ganado una importancia significativa en el campo. de minería de texto debido al interés de gobiernos y empresas multinacionales por mantener un dedo en el pulso del estado de ánimo público para ganar elecciones en el caso de los primeros o simplemente sorprender a sus clientes por la cantidad de información acerca de sus preferencias por este último. Curiosamente, la predicción del mercado está muy relacionada con el estado de ánimo del público o del mercado. participantes según lo establecido por la economía del comportamiento. Sin embargo, en el caso del análisis de sentimiento con respecto a un producto la anticipación de lo que implica un texto está lejos más sencillo que en el caso de la predicción del mercado. No hay secretos en cuanto a si una revisión del producto implica emociones positivas o negativas al respecto. Sin embargo, incluso el mejores comerciantes e inversores nunca pueden estar completamente seguros de qué reacción del mercado esperar como resultado de una noticia-texto. Por lo tanto, hay mucho espacio para la predicción del mercado. investigación de sentimientos para futuras investigaciones.

D. Componente de minería de textos, fuente textual o especialización en el mercado de aplicaciones: este trabajo tiene aprendió que los trabajos de investigación actuales sobre la minería de texto predictiva del mercado son bastante holísticos con sistemas únicos de extremo a extremo. Sin embargo, en el futuro, el proceso de minería de textos debería ser desglosado en sus componentes críticos como selección de características, representación de características y reducción de funciones y cada uno de ellos debe investigarse específicamente para el especialista contexto de predicción del mercado; se han hecho algunas sugerencias específicas para cada componente en los apartados 3.3.1, 3.3.2 y 3.3.3 de este trabajo. Además, la minería de texto predictiva del mercado también puede volverse aún más especializado centrándose en una fuente específica de texto, por ejemplo, un medio de comunicación social o fuente de noticias o función de texto como titulares de noticias frente a cuerpos de noticias, etc. Además, existe la necesidad de una investigación especializada en cada tipo de mercado financiero (acciones,

bonos, materias primas, dinero, futuros, derivados, seguros, divisas) o en cada ubicación geográfica localización.

E. Algoritmos de Machine Learning: En la sección 3.4 se ha explicado detalladamente cómo SVM y Naïve Bayes son muy favorecidos por los investigadores, probablemente debido a su sencillez, mientras que muchos otros algoritmos o técnicas de aprendizaje automático como las redes neuronales artificiales (ANN), K-vecinos más cercanos (k-NN), lógica difusa, etc. muestran potenciales muy prometedores para clasificación textual y análisis de sentimientos en otras partes de la literatura, pero aún no han experimentado en el contexto de la minería de texto predictiva del mercado o son significativamente poco investigado en esta etapa.

F. Integración de señales técnicas: a pesar de su popularidad práctica entre los comerciantes del mercado, señales técnicas que son las salidas de algoritmos técnicos o reglas como el movimiento el promedio, las reglas de fuerza relativa, las reglas de filtro y las reglas de desglose del rango comercial, son casi siempre fuera de la investigación que se basa en la minería de texto como se señala en la sección 3.4.5 de este trabajo. Puede deberse al hecho de que los investigadores que están a favor de la predicción basada en minería de texto son probablemente para enfoques de análisis fundamental y, por lo tanto, se oponen a los enfoques de análisis técnico como se explica en la sección 2.5. Sin embargo, es lógico concebible que los modelos híbridos basados en la suma de lo mejor de los dos mundos (técnica y fundamental) debe producir resultados aún mejores y esto debe ser considerado más vigorosamente en futuras investigaciones.

G. Relación con la investigación en economía del comportamiento: Como se ha señalado en el apartado 2 de este trabajo, existe un carácter interdisciplinario sustancial en este campo de investigación, especialmente entre economía e informática. Profundizar la comprensión de la economía es crucial para el futuro investigadores Actualmente, la economía y especialmente las teorías de la economía del comportamiento están referidas en la literatura solo superficialmente y solo para establecer que el estado de ánimo público tiene un impacto en mercados. Sin embargo, un estudio más profundo de la economía del comportamiento debería revelar principios y

aprendizaje cuya implementación paralela en la minería de texto por informáticos puede conducir a verdaderos avances. Esta dirección, aunque algo inmadura o vaga en esta etapa, es muy alentado por los investigadores de este trabajo.

H. Disponibilidad y calidad de conjuntos de datos experimentales: uno de los principales desafíos observados es la falta de disponibilidad de conjuntos de datos altamente estandarizados que contienen asignaciones de texto en los mercados por ciertos periodos de tiempo que los investigadores pueden usar para la asimilación de sus esfuerzos de experimentación y evaluación. En el trabajo disponible, la mayoría de los investigadores han intentaron acumular sus propios conjuntos de datos. Naturalmente, esto ha resultado en una fragmentación formatos y contenidos de conjuntos de datos y una falta de observación adecuada para las características críticas en conjuntos de datos. Por ejemplo, como se indica en la tabla 5, la mayoría de los trabajos no han observado si sus los datos experimentales están desequilibrados de alguna manera que puedan haber favorecido sus resultados. Futuro Se alienta a los investigadores a estandarizar y publicar conjuntos de datos para la experimentación en minería de texto predictiva del mercado. Actualmente, los conjuntos de datos estándar predominantes que circulan en trabajos de minería de textos son críticas de películas que no son apropiadas para este trabajo. Wu y Tan (2011) presenta una investigación intrigante sobre la transferencia del dominio del sentimiento conocimiento de un dominio a otro mediante la construcción de un marco entre ellos que actúa como un puente entre el dominio de origen y el dominio de destino. Esto puede inspirar algunas nuevas pensamientos en esta área también.

I. Métodos de evaluación: al igual que los conjuntos de datos experimentales, los métodos de evaluación en su conjunto son muy subjetivos. La mayoría de los investigadores todavía están comparando sus resultados con las probabilidades del azar y no tanto con las obras de los demás. Como se indicó en la sección 3.5, los investigadores son generalmente utilizando mecanismos de evaluación que varían ampliamente; que hacen un objetivo la evaluación comparativa del desempeño es virtualmente imposible. Por lo tanto, los futuros investigadores podrían centrarse en tales iniciativas de estandarización como objetivos principales de su investigación en este

El campo como la minería de texto predictiva del mercado está aquí para quedarse.

5. Conclusión

Se han revisado los principales sistemas de predicción de mercado basados en minería de texto en línea y se han de los vacíos predominantes que existen dentro de ellos han sido identificados. La revisión se llevó a cabo el tres aspectos principales, a saber: el preprocesamiento, el aprendizaje automático y el mecanismo de evaluación; con cada uno desglosándose en múltiples sub-discusiones. Se cree que es el primer esfuerzo para proporcionar una revisión integral desde un punto de vista holístico e interdisciplinario. Este trabajo pretendía lograr: En primer lugar, facilitar la integración de actividades de investigación de diferentes campos sobre el tema. de predicción de mercado basada en minería de texto en línea; En segundo lugar, la provisión de un marco de estudio para aislar el problema o diferentes aspectos del mismo con el fin de aclarar el camino para seguir mejorando; En tercer lugar, presentación de sugerencias direccionales y teóricas para futuras investigaciones.

Los avances en el campo de la minería de texto predictiva del mercado pueden tener las siguientes implicaciones en particular entre muchos:

1- Los bancos e instituciones financieras de inversión, así como las casas de bolsa que están invirtiendo y el comercio en los mercados financieros puede utilizar sistemas especializados de análisis y predicción de tendencias del mercado que se desarrollan utilizando los conocimientos adquiridos en esfuerzos de investigación de minería de texto específicos como este trabajo. La existencia de tales sistemas inteligentes para esas instituciones ayuda a hacer mejores decisiones financieras que conducen a considerables rendimientos financieros de sus inversiones y evitación de pérdidas severas.

2- En la economía global actual, es necesario obtener conocimientos aún más sofisticados sobre los mercados financieros. necesarios, porque la falta de ellos, como hemos presenciado recientemente durante el ejercicio económico de 2008, crisis, puede afectar negativamente los medios de subsistencia de millones de personas en todo el mundo. Por lo tanto, se vuelve imperativo continuar la investigación en el campo del texto predictivo del mercado. la minería como una solución viable que puede generar un grado mucho mayor de confianza en comprensión de los movimientos del mercado basada en la obtención de conocimientos sobre la psicología humana en un

nivel macro a través de la minería de texto de los recursos textuales, ahora ampliamente disponibles, en el Internet a un ritmo prácticamente en tiempo real.

3- Con la asombrosa cantidad de información textual disponible en línea, sobre todos los aspectos de todos los temas imaginables, la necesidad de desarrollar rápidamente sistemas especializados de minería de textos surge Dichos sistemas están altamente dirigidos a un área de aplicación específica y un cierto tipo de texto entre demasiadas alternativas posibles. Un enfoque en la minería de texto predictiva del mercado n esta investigación ayuda a la formación de este campo emergente como un campo reconocible e independiente campo en el que se puede profundizar enérgicamente y no solo a la sombra de la minería de texto general investigar. La formación de tal campo de investigación independiente para el texto predictivo del mercado. la minería, distinta del análisis de sentimiento de revisión de productos o similar, es una implicación esperanzadora de este trabajo.

Se espera que este trabajo ayude a otros investigadores a poner en perspectiva las diversas ideas en este campo más convenientemente y ser capaz de tomar decisiones estratégicas por adelantado en el diseño del futuro sistemas

Agradecimientos

Se agradece el apoyo de la Universidad de Malaya en la producción de este documento.

Referencias

Aghdam, MH, Ghasem-Aghaee, N. y Basiri, ME (2009). Selección de características de texto utilizando la optimización de colonias de hormigas. Sistemas Expertos con Aplicaciones, 36, 6843-6853.

Anastasakis, L. y Mort, N. (2009). Pronóstico del tipo de cambio utilizando un método combinado paramétrico y enfoque de modelado no paramétrico de autoorganización. Sistema experto Appl., 36, 12001-12011.

Bahrepour, M., Akbarzadeh-T., M.-R., Yaghoobi, M. y Naghibi-S., M.-B. (2011). Un pedido adaptativo Serie temporal difusa con aplicación a FOREX. Sistema experto Appl., 38, 475-485.

Balahur, A., Steinberger, R., Goot, E. vd, Pouliquen, B. y Kabadjov, M. (2009). Minería de opinión sobre citas periodísticas. En Actas de la Conferencia internacional conjunta IEEE/WIC/ACM de 2009 sobre inteligencia web y tecnología de agentes inteligentes - Volumen 03 (págs. 523-526): IEEE Computer Society.

Berka, T. y Vajteršic, M. (2013). Reemplazo paralelo de vectores de términos raros: rápido y efectivo Reducción de la dimensionalidad del texto. Revista de Computación Paralela y Distribuida, 73, 341-351.

Bikas, E., Jureviÿienÿ, D., Dubinskas, P. y Novickytÿ, L. (2013). Finanzas conductuales: las tendencias de aparición y desarrollo. Procedia - Ciencias Sociales y del Comportamiento, 82, 870-876.

Bollen, J. y Huina, M. (2011). Estado de ánimo de Twitter como predictor del mercado de valores. Informática, 44, 91-94.

Burges, CC (1998). Un tutorial sobre máquinas de vectores de soporte para el reconocimiento de patrones. Procesamiento de datos y Knowledge Discovery, 2, 121-167.

Butler, M. y Kešelj, V. (2009). Pronóstico financiero mediante el análisis de caracteres N-Gram y Puntuaciones de legibilidad de los informes anuales. En Y. Gao & N. Japkowicz (Eds.), Avances en Inteligencia Artificial (Vol. 5549, pp. 39-51): Springer Berlin Heidelberg.

Cambria, E., Schuller, B., Yunqing, X. y Havasi, C. (2013). Nuevos caminos en la minería de opinión y Análisis de los sentimientos. Sistemas Inteligentes, IEEE, 28, 15-21.

Chang, C.-C. y Lin, C.-J. (2011). LIBSVM: una biblioteca para máquinas de vectores de soporte. ACM Trans. Intel. sist. Tecnología, 2, 1-27.

Chatrath, A., Miao, H., Ramchander, S. y Villupuram, S. (2014). Saltos de divisas, cojumps y el papel de las noticias macro. Revista de Dinero y Finanzas Internacionales, 40, 42-62.

Chordia, T., Goyal, A., Lehmann, BN y Saar, G. (2013). Negociación de alta frecuencia. diario de finanzas Mercados.

Chordia, T., Roll, R. y Subrahmanyam, A. (2005). Evidencia sobre la velocidad de convergencia a la eficiencia del mercado. Revista de Economía Financiera, 76, 271-292.

Chordia, T., Roll, R. y Subrahmanyam, A. (2011). Tendencias recientes en la actividad comercial y la calidad del mercado. Revista de Economía Financiera, 101, 243-263.

Collins, M. y Duffy, N. (2001). Núcleos de convolución para lenguaje natural. En (págs. 625-632): MIT Presionar.

Das, SR y Chen, MI (2007). yahoo! para Amazon: Extracción de sentimientos de Small Talk en el Web. Gestionar. Sci., 53, 1375-1388.

Desmet, B. y Hoste, V. (2013). Detección de emociones en notas de suicidio. Sistemas Expertos con Aplicaciones, 40, 6351-6358.

Drucker, H., Burges, CJC, Kaufman, L., Smola, A. y Vapnik, V. (1997). Admite máquinas de regresión vectorial. En AVANCES EN SISTEMAS DE PROCESAMIENTO DE INFORMACIÓN NEURAL (págs. 155-161): MIT Press.

Duman, E., Ekinci, Y. y Tanrÿverdi, A. (2012). Comparación de clasificadores alternativos para la base de datos marketing: el caso de los conjuntos de datos desequilibrados. Sistemas Expertos con Aplicaciones, 39, 48-53.

Duric, A. y Song, F. (2012). Selección de características para el análisis de opiniones basado en el contenido y la sintaxis modelos Sistemas de soporte de decisiones, 53, 704-711.

Evans, C., Pappas, K. y Xhafa, F. (2013). Utilizando redes neuronales artificiales y algoritmos genéticos para construir un modelo de comercio algorítmico para la especulación de divisas intradía. Modelado matemático e informático, 58, 1249-1266.

Fama, EF (1965). Paseos aleatorios en los precios del mercado de valores. Revista de analistas financieros, 21, 55-59.

Fama, EF (1970). Mercados de capital eficientes: una revisión de la teoría y el trabajo empírico. El Diario de Finanzas, 25, 383-417.

Fan, R.-E., Chen, PH-H. y Lin, C.-J. (2005). Selección de conjuntos de trabajo utilizando información de segundo orden para entrenar máquinas de vectores de soporte. J. Mach. Aprender. Res., 6, 1889-1918.

Fasanghari, M. y Montazer, GA (2010). Diseño e implementación de un sistema experto difuso para la recomendación de cartera de la Bolsa de Valores de Teherán. Sistemas expertos con aplicaciones, 37, 6138-6147.

Feng, G., Guo, J., Jing, B.-Y. y Hao, L. (2012). Un paradigma bayesiano de selección de características para texto clasificación. Gestión y procesamiento de la información, 48, 283-302.

Friesen, G. y Weller, PA (2006). Cuantificación de los sesgos cognitivos en las previsiones de ganancias de los analistas. Revista de Mercados Financieros, 9, 333-365.

García, D. y Uroševiÿ, B. (2013). Ruido y agregación de información en grandes mercados. Revista de Mercados Financieros, 16, 526-549.

Garcke, J., Gerstner, T. y Griebel, M. (2013). Pronóstico del tipo de cambio intradía mediante Cuadrículas dispersas. En J. Garcke & M. Griebel (Eds.), Sparse Grids and Applications (Vol. 88, pp. 81-105): Springer Berlín Heidelberg.

Ghazali, R., Hussain, AJ y Liatsis, P. (2011). Red neuronal polinomial de cresta dinámica: Pronosticar las señales comerciales univariadas no estacionarias y estacionarias. Sistema experto Appl., 38, 3765-3776.

Ghiassi, M., Skinner, J. y Zimbra, D. (2013). Análisis de sentimiento de marca de Twitter: un sistema híbrido que utiliza análisis de n-gramas y una red neuronal artificial dinámica. Sistemas expertos con aplicaciones, 40, 6266-6282.

Gradojevic, N. y Gençay, R. (2013). Lógica difusa, incertidumbre comercial y negociación técnica. Revista de Banca y Finanzas, 37, 578-586.

Griffiths, TL, Steyvers, M., Blei, DM y Tenenbaum, JB (2005). Integración de temas y sintaxis. En (págs. 537-544): MIT Press.

Groth, SS y Muntermann, J. (2011). Un enfoque de gestión del riesgo de mercado intradía basado en analisis textual. Sistemas de soporte de decisiones, 50, 680-691.

Haddi, E., Liu, X. y Shi, Y. (2013). El papel del preprocesamiento de texto en el análisis de sentimiento. procedimiento Informática, 17, 26-32.

Hagenau, M., Liebmann, M. y Neumann, D. (2013). Lectura automatizada de noticias: Predicción del precio de las acciones basada en noticias financieras utilizando funciones de captura de contexto. Sistemas de soporte de decisiones, 55, 685-697.

Hasbrouck, J. y Saar, G. (2013). Comercio de baja latencia. Revista de Mercados Financieros.

Hsinchun, C. y Zimbra, D. (2010). IA y Minería de Opinión. Sistemas Inteligentes, IEEE, 25, 74-80.

Huang, C.-J., Liao, J.-J., Yang, D.-X., Chang, T.-Y. y Luo, Y.-C. (2010). Realización de una noticia. agente de difusión basado en reglas de asociación ponderada y técnicas de minería de textos. Sistema experto Appl., 37, 6409-6413.

Huang, SC-C., Chuang, P.-J., Wu, CF-F. y Lai, H.-J. (2010). Regresiones de vectores de soporte basadas en el caos para pronóstico del tipo de cambio. Sistema experto Appl., 37, 8590-8598.

Jiang, S., Pang, G., Wu, M. y Kuang, L. (2012). Un algoritmo K-vecino más cercano mejorado para la categorización de texto. Sistemas Expertos con Aplicaciones, 39, 1503-1509.

Jin, F., Self, N., Saraf, P., Butler, P., Wang, W. y Ramakrishnan, N. (2013). Forex-foreteller: modelado de tendencias de divisas utilizando artículos de noticias. En Actas de la 19.ª conferencia internacional ACM SIGKDD sobre descubrimiento de conocimiento y minería de datos (págs. 1470-1473). Chicago, Illinois, Estados Unidos: ACM.

Joaquín, T. (1998). Categorización de texto con máquinas de vectores de soporte: aprendizaje con muchas características relevantes. En C. Nédellec & C. Rouveirol (Eds.), Machine Learning: ECML-98 (Vol. 1398, pp. 137-142): Springer Berlín Heidelberg.

Joaquín, T. (1999). Hacer práctico el aprendizaje {SVM} a gran escala. En B. Schölkopf, C. Burges & A. Smola (Eds.), (págs. 169-184). Cambridge, MA: MIT Press.

Joaquín, T. (2002). Aprendiendo a clasificar texto utilizando máquinas de vectores de soporte: métodos, teoría y Algoritmos: Kluwer/Springer.

Kaltwasser, PR (2010). Incertidumbre sobre los fundamentos y el comportamiento de manada en el mercado FOREX. Physica A: Mecánica estadística y sus aplicaciones, 389, 1215-1222.

Kanayama, H. y Nasukawa, T. (2008). Análisis de demanda textual: detección de deseos y necesidades de los usuarios a partir de opiniones. En Actas de la 22ª Conferencia Internacional sobre Lingüística Computacional - Volumen 1 (págs. 409-416). Manchester, Reino Unido: Asociación de Lingüística Computacional.

Khadjeh Nassirtoussi, A., Ying Wah, T. y Ngo Chek Ling, D. (2011). Una novedosa predicción de FOREX metodología basada en datos fundamentales. Revista africana de gestión empresarial, 5, 8322-8330.

Kim, K., Chung, B.-s., Choi, Y., Lee, S., Jung, J.-Y. y Park, J. (2014). Núcleos semánticos independientes del idioma para la clasificación de textos breves. Sistemas Expertos con Aplicaciones, 41, 735-743.

Kleinnijenhuis, J., Schultz, F., Oegema, D. y Atteveldt, WH van. (2013). Noticias financieras y pánico del mercado en la era de los algoritmos comerciales de alta frecuencia. Periodismo, 14.

Kontopoulos, E., Berberidis, C., Dergiades, T. y Bassiliades, N. (2013). Sentimiento basado en ontología análisis de publicaciones en twitter. Sistemas expertos con aplicaciones, 40, 4065-4074.

Lewis, D. (1998). Ingenuo (Bayes) a los cuarenta: El supuesto de independencia en la recuperación de información. Cª. Nédellec & C. Rouveirol (Eds.), Machine Learning: ECML-98 (Vol. 1398, pp. 4-15): Springer Berlin Heidelberg.

Li, CH, Yang, JC y Park, SC (2012). Algoritmos de categorización de texto utilizando enfoques semánticos, Tesauro basado en corpus y WordNet. Sistemas Expertos con Aplicaciones, 39, 765-772.

Li, F. (2010). El contenido de la información de las declaraciones prospectivas en los registros corporativos: un enfoque de aprendizaje automático bayesiano ingenuo. Revista de Investigación Contable, 48, 1049-1102.

Li, W. y Xu, H. (2014). Clasificación de emociones basada en texto usando extracción de causa de emoción. Experto Sistemas con Aplicaciones, 41, 1742-1749.

Liu, Y., Loh, HT y Sun, A. (2009). Clasificación de texto desequilibrada: un enfoque de ponderación de términos. Experto Sistemas con Aplicaciones, 36, 690-701.

Lo, AW (2005). Reconciliación de mercados eficientes con finanzas conductuales: los mercados adaptables Hipótesis. Revista de Consultoría de Inversiones.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. y Watkins, C. (2002). Clasificación de texto usando núcleos de cuerda. J. Mach. Aprender. Res., 2, 419-444.

Loia, V. y Senatore, S. (2014). Un análisis sentico de orientación difusa para capturar la emoción humana en Contenido basado en la web. Sistemas basados en el conocimiento, 58, 75-85.

Lugmayr, A. y Gossen, G. (2012). Evaluación de Métodos y Técnicas de Lenguaje Basado Análisis de Sentimiento para la Bolsa de Valores DAX 30 – Un Primer Concepto de un Indicador de Sentimiento "LUGO". En A. Lugmayr, T. Risse, B. Stockleben, J. Kaario, B. Pogorelc & E. Serral Asensio (Eds.), SAME 2012 – 5th International Workshop on Semantic Ambient Media Experience.

Luo, Q., Chen, E. y Xiong, H. (2011). Un esquema de ponderación de términos semánticos para la categorización de texto. Sistemas Expertos con Aplicaciones, 38, 12708-12716.

Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J. T., & Camón-Herrero, J. B. (2011). Financial news semantic search engine. Sistemas expertos con aplicaciones, 38, 15565-15572.

Mabu, S., Hirasawa, K., Obayashi, M. y Kuremoto, T. (2013). Mecanismo mejorado de toma de decisiones de programación de redes genéticas basadas en reglas para crear señales de negociación de acciones. Sistemas Expertos con Aplicaciones, 40, 6311-6320.

Mahajan, A., Dey, L. y Haque, SM (2008). Noticias Financieras Mineras para Grandes Eventos y sus Impactos en el Mercado. En Inteligencia Web y Tecnología de Agente Inteligente, 2008. WI-IAT '08. Conferencia internacional IEEE/WIC/ACM sobre (Vol. 1, págs. 423-426).

Majumder, D. (2013). Hacia un mercado de valores eficiente: evidencia empírica del mercado indio. Revista de modelado de políticas, 35, 572-587.

Maks, I. y Vossen, P. (2012). Un modelo de léxico para análisis profundo de sentimientos y aplicaciones de minería de opiniones. Sistemas de soporte de decisiones, 53, 680-688.

Miller, GA (1995). WordNet: una base de datos léxica para inglés. común ACM, 38, 39-41.

Mittermayer, MA (2004). Pronóstico de tendencias de precios de acciones intradía con técnicas de minería de texto. En System Sciences, 2004. Actas de la 37.ª Conferencia Internacional Anual de Hawái sobre (págs. 10 págs.).

Moraes, R., Valiati, JF y Gavião Neto, WP (2013). Clasificación de sentimiento a nivel de documento: una comparación empírica entre SVM y ANN. Sistemas Expertos con Aplicaciones, 40, 621-633.

Mostafá, MM (2013). Más que palabras: minería de texto de las redes sociales para los sentimientos de marca del consumidor. Sistemas Expertos con Aplicaciones, 40, 4241-4251.

Nikfarjam, A., Emadzadeh, E. y Muthaiyah, S. (2010). Enfoques de minería de texto para el mercado de valores predicción. En Ingeniería Informática y Automatización (ICCAE), 2010 La 2ª Conferencia Internacional sobre (Vol. 4, pp. 256-260).

Ortigosa-Hernández, J., Rodríguez, JD, Alzate, L., Lucania, M., Inza, I., & Lozano, JA (2012). Aproximación al análisis de sentimientos mediante el uso del aprendizaje semisupervisado de clasificadores multidimensionales. Neurocomputación, 92, 98-115.

Peramunetilleke, D. y Wong, RK (2002). Pronóstico del tipo de cambio de divisas a partir de titulares de noticias. agosto computar ciencia Comun., 24, 131-139.

Pestov, V. (2013). ¿El clasificador -NN en dimensiones altas está afectado por la maldición de la dimensionalidad? Informática y Matemáticas con Aplicaciones, 65, 1427-1437.

Platt, JC (1999). Entrenamiento rápido de máquinas de vectores de soporte usando optimización mínima secuencial. En Avances en métodos kernel (págs. 185-208): MIT Press.

Potì, V. y Siddique, A. (2013). ¿Qué impulsa la previsibilidad de la moneda? Revista de dinero internacional y Finanzas, 36, 86-106.

Premanode, B. y Toumazou, C. (2013). Mejora de la predicción de los tipos de cambio utilizando Diferencial EMD. Sistemas Expertos con Aplicaciones, 40, 377-384.

Pui Cheong Fung, G., Xu Yu, J. y Wai, L. (2003). Predicción bursátil: integración del enfoque de minería de texto utilizando noticias en tiempo real. En Inteligencia Computacional para Ingeniería Financiera, 2003. Actas. 2003 Conferencia Internacional IEEE sobre (págs. 395-402).

Quinlan, JR (1993). C4.5: programas para aprendizaje automático: Morgan Kaufmann Publishers Inc.

Rachlin, G., Last, M., Alberg, D. y Kandel, A. (2007). ADMIRAL: una empresa financiera basada en la minería de datos Sistema de comercio. En Inteligencia Computacional y Minería de Datos, 2007. CIDM 2007. Simposio sobre IEEE (págs. 720-725).

Reboredo, JC, Rivera-Castro, MA, Miranda, JGV, & García-Rubio, R. (2013). ¿Qué tan rápido se ajustan los precios de las acciones a la eficiencia del mercado? Evidencia de un análisis de fluctuación sin tendencia. Physica A: Mecánica estadística y sus aplicaciones, 392, 1631-1637.

Robertson, C., Geva, S. y Wolff, R. (2006). ¿Qué tipos de eventos proporcionan la evidencia más sólida de que el mercado de valores se ve afectado por noticias específicas de la empresa? En Actas de la quinta conferencia de Australasia sobre minería y análisis de datos - Volumen 61 (págs. 145-153). Sídney, Australia: Sociedad Australiana de Computación, Inc.

Salzberg, S. (1997). Sobre la comparación de clasificadores: errores a evitar y un enfoque recomendado. Datos Minería y Descubrimiento del Conocimiento, 1, 317-328.

Sankaraguruswamy, S., Shen, J. y Yamada, T. (2013). La relación entre la frecuencia de los comunicados de prensa y la asimetría de la información: el papel del comercio desinformado. Revista de Banca y Finanzas, 37, 4134-4143.

Schumaker, RP y Chen, H. (2009). Análisis textual de la predicción del mercado de valores utilizando la ruptura noticias financieras: El sistema de texto AZFin. ACM Trans. información Sist., 27, 1-19.

Schumaker, RP, Zhang, Y., Huang, C.-N. y Chen, H. (2012). Evaluación del sentimiento en artículos de noticias financieras. Sistemas de Soporte a la Decisión.

Sermpinis, G., Laws, J., Karathanasopoulos, A. y Dunis, CL (2012). Pronóstico y negociación del tipo de cambio EUR/ USD con Gene Expression y Psi Sigma Neural Networks. Sistemas Expertos con Aplicaciones, 39, 8865-8877.

Shi, K., He, J., Liu, H.-t., Zhang, N.-t., y Song, W.-t. (2011). Método eficiente de clasificación de texto basado en reducción de términos mejorada y ponderación de términos. The Journal of China Universities of Posts and Telecommunications, 18, Suplemento 1, 131-135.

Sidorov, G., Velásquez, F., Stamatatos, E., Gelbukh, A. y Chanona-Hernández, L. (2013). N gramos sintácticos como funciones de aprendizaje automático para el procesamiento del lenguaje natural. Sistemas Expertos con Aplicaciones.

Smales, LA (2012). Desequilibrio de órdenes, rendimientos del mercado y noticias macroeconómicas: Evidencia del mercado de futuros de tasas de interés de Australia. Investigación en Negocios y Finanzas Internacionales, 26, 410-427.

Soni, A., van Eck, NJ y Kaymak, U. (2007). Predicción de movimientos de precios de acciones basada en información de mapas conceptuales. En Computational Intelligence in Multicriteria Decision Making, Simposio sobre IEEE (págs. 205-211).

Tan, S., Wang, Y. y Wu, G. (2011). Adaptación del clasificador centroide para la categorización de documentos. Experto Sistemas con Aplicaciones, 38, 10264-10273.

Taÿcÿ, ÿ. y Güngör, T. (2013). Comparación de políticas de selección de características de texto y uso de un marco adaptativo. Sistemas Expertos con Aplicaciones, 40, 4871-4886.

Tetlock, PC (2007). Dar contenido al sentimiento de los inversores: el papel de los medios en el mercado de valores. Diario de Finanzas, 62, 1139-1168.

Tetlock, PC, Saar-Tsechansky, M. y Macskassy, S. (2008). Más que palabras: cuantificación del lenguaje para medir los fundamentos de las empresas. Diario de Finanzas, 63, 1437-1467.

Thammasiri, D., Delen, D., Meesad, P. y Kasap, N. (2014). Una evaluación crítica del problema de distribución de clases desequilibrada: el caso de la predicción de la deserción de estudiantes de primer año. Sistemas Expertos con Aplicaciones, 41, 321-330.

Tomer, JF (2007). ¿Qué es la economía del comportamiento? The Journal of Socio-Economics, 36, 463-479.

Tsai, CF-F., Eberle, W. y Chu, C.-Y. (2013). Algoritmos genéticos en la selección de características e instancias. Sistemas basados en el conocimiento, 39, 240-247.

Urquhart, A. y Hudson, R. (2013). ¿Mercados eficientes o adaptables? Evidencia de los principales mercados bursátiles utilizando datos históricos de muy largo plazo. Revista Internacional de Análisis Financiero, 28, 130-142.

Uysal, AK y Gunal, S. (2012). Un novedoso método de selección de características probabilísticas para la clasificación de textos. Sistemas basados en el conocimiento, 36, 226-235.

Uysal, AK y Gunal, S. (2014). El impacto del preprocesamiento en la clasificación del texto. Información Procesamiento y Gestión, 50, 104-112.

Vanstone, B. y Finnie, G. (2010). Mejorar el rendimiento comercial del mercado de valores con ANN. Experto Sistemas con Aplicaciones, 37, 6602-6610.

Vu, TT, Chang, S., Ha, QT y Collier, N. (2012). Un experimento en la integración de funciones de sentimiento para la predicción de acciones tecnológicas en Twitter. En Actas del taller sobre extracción de información y análisis de entidades en datos de redes sociales (págs. 23-38). Mumbai, India: El Comité Organizador de COLING 2012.

Wang, G.-J., Xie, C. y Han, F. (2012). Análisis de entropía aproximada multiescala de divisas Eficiencia de los mercados. Ingeniería de Sistemas Procedia, 3, 201-208.

Weiss, SM, Indurkhya, N. y Zhang, T. (2010). Fundamentos de la Minería Predictiva de Textos.

Werner, A. y Myrray Z., F. (2004). ¿Todo lo que se habla es solo ruido? El contenido de información de los tableros de mensajes de acciones de Internet. Diario de Finanzas, 1259--1294.

Wisniewski, TP y Lambe, B. (2013). El papel de los medios en la contracción del crédito: El caso del sector bancario. Revista de Comportamiento y Organización Económica, 85, 163-175.

Witten, IH y Frank, E. (2005). Minería de datos: herramientas y técnicas prácticas de aprendizaje automático, Segunda edición (Serie de Morgan Kaufmann en sistemas de gestión de datos): Morgan Kaufmann Publishers Inc.

Wu, Q. y Tan, S. (2011). Un marco de dos etapas para la clasificación de sentimientos entre dominios. Experto Sistemas con Aplicaciones, 38, 14269-14275.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K. y Zhang, J. (1998). Stock diario pronóstico del mercado a partir de datos web textuales. En Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on (Vol. 3, pp. 2720-2725 vol.2723).

Yang, L., Li, C., Ding, Q. y Li, L. (2013). Combinación de características léxicas y semánticas para la clasificación de textos breves. Procedia Informática, 22, 78-86.

Yin, L., Ge, Y., Xiao, K., Wang, X. y Quan, X. (2013). Selección de características para alta dimensión datos desequilibrados. Neurocomputación, 105, 3-11.

Yu, H., Nartea, GV, Gan, C. y Yao, LJ (2013). Capacidad predictiva y rentabilidad de reglas comerciales técnicas simples: evidencia reciente de los mercados bursátiles del sudeste asiático. Revista Internacional de Economía y Finanzas, 25, 356-371.

Yu, L.-C., Wu, J.-L., Chang, P.-C. y Chu, H.-S. (2013). Uso de un modelo de entropía contextual para expandir las palabras de emoción y su intensidad para la clasificación de sentimiento de las noticias del mercado de valores.
Sistemas Basados en el Conocimiento.

Yu, Y., Duan, W. y Cao, Q. (2013). El impacto de los medios sociales y convencionales en el valor de las acciones de las empresas: un enfoque de análisis de sentimientos. Sistemas de Soporte a la Decisión.

Zhai, Y., Hsu, A. y Halgamuge, SK (2007). Combinación de indicadores técnicos y de noticias en la predicción diaria de tendencias de precios de acciones. En Actas del cuarto simposio internacional sobre redes neuronales: avances en redes neuronales, parte III (págs. 1087-1096). Nanjing, China: Springer Verlag.

- Revisión de conceptos esenciales para la predicción de mercados basada en minería de texto en línea.
- Revisión de trabajos de vanguardia en la literatura.
- Identificación de los principales factores diferenciadores entre las soluciones disponibles.
- Observaciones sobre posibles oportunidades de trabajo futuro.