

Máster Universitario en Tecnologías del sector financiero  
2019-2020

*Trabajo Fin de Máster*

“Predicción de volatilidad de la cotización de empresas del IBEX35 basada en Google Trends.”

D. Álvaro Andrés Suárez Alfonso

Tutor

Dr. Ricardo Martín Manso

Leganés, 2020



Esta obra se encuentra sujeta a la licencia Creative Commons **Reconocimiento – No Comercial – Sin Obra Derivada**

# Resumen

Este trabajo fin de máster aplica técnicas de extracción y minado de datos desde Google trends y Yahoo finance para predecir la volatilidad de los principales índices bursátiles de los mercados financieros.

Este proyecto hace un análisis de comparación entre tendencias de búsqueda en diferentes fuentes de Google y la convergencia de precios de las cotizaciones de un índice bursátil para demostrar una posible correlación entre dichas fuentes de información.

# Abstract

# Palabras clave

Predicción, mercado, volatilidad, S&P500, DAX30, IBEX35, índice bursátil, big data, Yahoo finance, Google trends, data mining.

Índice

[Resumen 2](#_Toc52469894)

[Abstract 3](#_Toc52469895)

[Palabras clave 4](#_Toc52469896)

[1. Introducción 6](#_Toc52469897)

[1.1 Definiciones y acrónimos 6](#_Toc52469898)

[1.1.1 Renta variable 6](#_Toc52469899)

[1.1.1.1 Volatilidad de mercado 6](#_Toc52469900)

[1.1.2 Índice bursátil 7](#_Toc52469901)

[1.1.2.1 Índices de importancia a nivel mundial 7](#_Toc52469902)

[1.1.2.1.1 Ibex 35 7](#_Toc52469903)

[1.1.2.1.2 DAX 30 8](#_Toc52469904)

[1.1.2.1.3 S&P 500 8](#_Toc52469905)

[1.1.3 Análisis de datos 8](#_Toc52469906)

[1.1.3.1 Predicción de mercado 8](#_Toc52469907)

[1.2 Objetivo de la tesis 8](#_Toc52469908)

[2. Estado del arte 8](#_Toc52469909)

[2.1 Motivación 10](#_Toc52469910)

[3. Metodología y ejecución del proyecto 11](#_Toc52469911)

[3.1 Preprocesado de los datos 12](#_Toc52469914)

[3.2 Exploración de datos 15](#_Toc52469915)

[3.3 Modificación y transformación de los datos 17](#_Toc52469916)

[3.4 Clasificación y predicción 19](#_Toc52469917)

[4. Gestión del proyecto 20](#_Toc52469918)

[4.1 Planificación 20](#_Toc52469919)

[4.1.1 WBS 21](#_Toc52469920)

[4.2 Presupuesto 24](#_Toc52469921)

[5. Análisis de resultados 25](#_Toc52469922)

[5.1 Experimento 25](#_Toc52469923)

[5.2 Resultados 27](#_Toc52469924)

[5.2 Análisis de Inferencia Estadística: 37](#_Toc52469928)

[6. Conclusiones 39](#_Toc52469929)

[7. Bibliografía 40](#_Toc52469930)

# Introducción

# Definiciones y acrónimos

# Renta variable

En el mercado bursátil se define la renta variable como un tipo de inversión en el que existe un instrumento como parte de un capital en representación de una empresa. Se define como variable a aquel instrumento del cual no se tiene conocimiento sobre su rentabilidad, es decir, a un plazo futuro se pueden reportar ganancias o pérdidas, por lo cual se considera la renta variable como una renta con riesgo alto.

# Volatilidad de mercado

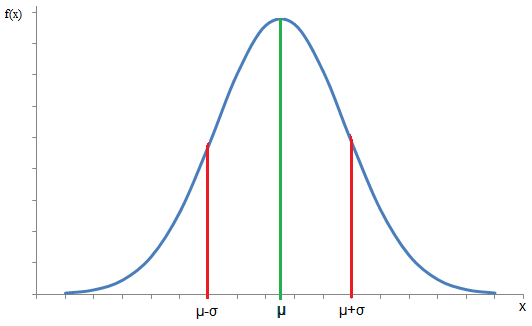
Conocido también como la desviación estándar del precio de un activo, en matemáticas financieras corresponde a la amplitud de frecuencia en que el precio de un activo se mueve, ya sea positiva o negativamente respecto de su valor inicial.

Los agentes de mercado especulan sobre la volatilidad de mercado con el fin de hacer inversiones de riesgo en el que puedan obtener rentabilidad al vender el activo en los picos de precio más altos, u otra forma es comprando a su precio más bajo para posteriormente tomar una posición corta.

En estadística la desviación estándar es conocida por la letra griega σ (sigma) y se calcula de la siguiente forma:

Donde n es el número de muestras, el valor de la muestra i-esima y la media de las muestras.

La siguiente figura es una representación de la desviación estándar en una curva de Gauss:



Una desviación estándar alta indica que las muestras se encuentran alejadas de su media, mientras que una desviación estándar baja señala que la mayoría de los datos están agrupados cerca de su media, acentuando de esta forma la curva de distribución normal como se muestra a continuación.

# Índice bursátil

El índice bursátil concierne a un grupo de acciones clasificadas por una o varias características similares. Por ejemplo, un grupo de acciones puede ser clasificado por su similitud en el índice de capitalización o por que se encuentran en la misma bolsa de mercado. Este es de gran importancia para las gestoras de fondos ya que con estos permiten buscar una cartera que replique dicho índice.

# Índices de importancia a nivel mundial

# Ibex 35

Es el índice bursátil propio de la bolsa española y creado por Bolsas y mercados españoles (BME) en el cómo su posfijo lo indica es un grupo formado por 35 empresas con mayor liquidez que cotizan en la bolsa española. Empresas se calcula en base a la ponderación sobre la capitalización de las sus empresas componentes, estas empresas pertenecen a diversos sectores económicos como energía, telecomunicaciones, finanzas, entre otras.

# DAX 30

También llamado DAX Xetra es el índice bursátil que contiene las 30 compañias de Alemania mas grandes que cotizan en la bolsa de Francfort, fundado el 30 de diciembre de 1987. La envergadura de las empresas aquí listadas está basada en términos de volumen y capitalización de mercado. Es llamado también DAX Xetra debido a que el modo de negociación se hace a través de la plataforma electrónica de la bolsa alemana, Xetra, en el cual se negocia mas del 90% del volumen de las acciones de los stocks del DAX.

# S&P 500

Originalmente llamado como índice Standard & Poor’s 500, es el índice mas relevante de Estados Unidos, y el índice apreciado como el mas representativo de la vitalidad de la economía de mercado como un todo. El índice contiene las 500 empresas mas grandes listadas en las bolsas del NYSE y NASDAQ, los cuales abarcan mas del 80% de la capitalización de mercado en el país. A diferencia del anteriormente mencionado DAX30, el índice S&P500 es ponderado por un comité del S&P Dow Jones indices, el cual cuanta con varios criterios de selección como un mínimo monto de capitalización, el monto de negociación anual, el volumen de acciones negociadas mensualmente, entre otros.

# Análisis de datos

# Predicción de mercado

# Objetivo de la tesis

* Analizar la capacidad de predicción del modelo utilizando las bases de Yahoo finance y Google trends.
* Conocer el impacto que tienen las noticias publicadas en la web en los índices mas representativos en el mercado internacional.
* Comparar la calidad del modelo predictivo de extracción y análisis de datos BigData planteado con un modelo de análisis utilizando machine learning.
* Reforzar el conocimiento de análisis de mercado y sus distintas aristas, así como, las técnicas de análisis de datos utilizadas en el pasado y actualmente.
* Comprender la importancia del análisis cuantitativo de datos, y entender el papel que juega en la capitalización del mercado.

# Estado del arte

La capitalización total de mercado de todas las acciones en el mundo era aproximadamente 70.75 Trillones de Dólares [1]. El IBEX 35 por si solo alcanza un Trillón de Euros[2]. En 2013, el volumen medio diario de ejecuciones en el New York Stock Exchange fue de 169 Billones de Dólares [3]. Estas cifras nos hacen conscientes de las magnitudes que se manejan en el mercado de acciones a nivel mundial y nacional en términos de riqueza, así como la importancia de gestionar los riesgos en estos mercados para la preservación del capital en la economía mundial.

Para una correcta gestión del riesgo se necesitan monitorizar adecuadamente las principales métricas de riesgo, para poder distribuir el capital de forma que el riesgo quede minimizado. Esta práctica tiene origen en la Teoría de Gestión de Carteras del premio Nobel H. Markowitz[4], y sigue siendo la base en gestión eficiente de riesgo en carteras financieras de activos. La repercusión de este trabajo reside en la presentación de una forma de monitorizar cambios en la medida de riesgo más importante, la volatilidad, por medio de cambios extremos en los retornos diarios, de forma anticipada y precisa gracias a la tecnología. Esto tiene implicaciones prácticas tanto en la gestión del riesgo, como medida del sentimiento de mercado y para la toma de decisiones de compra o venta de productos derivados de volatilidad como opciones. El marco de este trabajo es el de predicción de series temporales de activos financieros por medio de análisis de sentimiento de mercado. Más específicamente, el trabajo se enfoca en la predicción de grandes cambios (más de una desviación estándar) en los retornos diarios de los activos financieros.

En la literatura del análisis de sentimiento de mercado encontramos que la gran mayoría de investigaciones se enfocan en identificar la polaridad entre opiniones positivas y negativas para inferir subidas o bajadas del precio de los activos, como en S. Nausheen et al. [5]. Pero también, dentro de la literatura encontramos trabajos enfocados en la medición de atención que se da a un término en específico como en F. Audrino et al.[6]. Este enfoque dentro de la literatura de análisis de sentimiento de mercado, pero dedicado a la medición de atención de ciertos términos es el que se adopta en este trabajo. Ambos enfoques son muy similares en técnicas de procesado de datos y en los posteriores modelos de clasificación o predicción, generalmente con técnicas supervisadas de Machine Learning. En [6], los autores realizan un estudio de regresión para identificar las palabras que mayor impacto tienen en la predicción de la volatilidad realizada (la desviación estándar del precio, para nuestro documento es la volatilidad). Utilizan como medida de atención el número de veces que aparece al día la búsqueda de la palabra referente al stock en Google trends. Este trabajo difiere de este en que nosotros nos enfocamos en la predicción de grandes cambios en los retornos diarios que afectan a la volatilidad cuando se producen grandes cambios en la medición de atención de ciertos términos que nosotros ya conocemos de antemano. Esto hace que nos fijemos en los extremos de esta relación y no en el continuo como hacen estos autores con la regresión y podamos así probar nuestra hipótesis de partida. Otro estudio similar es el de M. Y. Huang et al. [7], en donde los autores analizan las relaciones entre los volúmenes de búsqueda con Google Trends y los movimientos direccionales del SP500. Encuentran que la relación entre cambios en el volumen de búsquedas y cambios direccionales es condicional al sentimiento de mercado implícito en el término buscado. Además, identifican términos consistentes con esta regla anterior y utilizan un modelo para predecir los movimientos en base a estos términos y construyen una estrategia de trading con muy buenos resultados. Nuestro estudio se diferencia de este en que, al igual que en [6], los autores analizan cambios continuos y no eventos extremos de las variables.

En la literatura de sentimiento de mercado encontramos dos ejes principales, el de minería de datos de texto en la web con el que se obtiene la variable independiente y el de modelos de predicción (y en algunos casos de clasificación) de la variable dependiente en cuestión. Para el primer eje, encontramos técnicas como Procesamiento del Lenguaje Natural (NLP), como en WS Chan [8] donde las aplican a diferentes textos de internet o las aplican a blogs, noticias y redes sociales como en I. Vaishali y S. Deshmukh [9], C. Vega y R.A. Albuquerque [10], y C. Jiang et al. [11]. Trabajos más recientes de NLP en sentimiento de mercado incluyen S.A. Phand et al. [12], o N. C. Dang et al. [13], que están enfocados en predicción. Y trabajos como X. Du y K. Tanaka-Ishii [14], donde se utilizan los embeddings adquiridos de artículos y series de precios para optimizar carteras. También encontramos casos como el nuestro en los que se mide la atención del inversor de ciertos términos en los que NLP no es necesario. Tradicionalmente para medir el interés del inversor se utilizaban movimientos extremos del activo, cambios en el volumen de trading o límites de precio. Con el avance de la tecnología se han añadido noticias y titulares, así como ciertas fuentes de internet [7]. En H.S. Moat et al. [12], se investiga el impacto que tiene el número de visitas que en páginas de Wikipedia en el movimiento del precio de la acción. En Z. Da et al. [13], los autores analizan el número de búsquedas en Google de compañías cerca de un evento IPO para predecir el precio de una acción. En N. Vlastakis y R.N. Markellos [14], los autores aproximan la demanda de acciones e índices con las búsquedas semanales en Google Trends y lo usan para analizar la volatilidad. M. Bank et al. [15], muestran la relación entre el volumen de búsquedas en Google Trends y el retorno de las acciones. Se enfocan en los nombres de las compañías para las búsquedas y demuestran una relativamente alta correlación con el retorno positivo del activo. En C. Curme et al. [16], los autores encuentran que la búsqueda en Google trends de términos políticos y financieros está relacionado con una caída del precio del activo en el siguiente día. También observan una caída de retornos en estrategias de trading con compras o ventas basadas en señales de atención de términos de Google Trends. Por otra parte, en T. Preis et al. [17], encuentran que se acumula mucho ruido cuando se utiliza un solo término de atención como señal. Como bien hemos dicho antes, nuestro estudio difiere de los anteriores en que nosotros nos enfocamos en la relación entre cambios extremos de variables y no en relaciones en un continuo de su valor. Como últimas aportaciones a la literatura podemos ver el trabajo de D. Borup et al. [18], donde los autores utilizan búsquedas en Google Trends de términos relacionados con desempleo para mejorar la metodología oficial que se sigue actualmente para predecir los Initial Claims, unos datos de desempleo. Ellos utilizan las tendencias como variables predictoras añadidas en una serie de modelos estándar y son capaces de probar que esta inclusión mejora considerablemente la predicción. Se prueba así que poco a poco estas metodologías son introducidas por organismos oficiales en sus análisis de datos.

# Motivación

Teniendo en cuenta el contexto del marco teórico que enmarca esta investigación, por un lado tenemos el análisis de movimiento de los mercados y sus repercusiones entre si y los grandes inversionistas que buscan capitalización a través de herramientas de análisis como las mencionadas anteriormente para predecir un estado futuro de cualquier indicador financiero que proporcione una ganancia potencial, y por otro lado tenemos la tecnología como herramienta de soporte al análisis de datos que nos permiten realizar operaciones sobre estos datos de manera más rápida y segura y con la cual podemos explotar con mas cantidad de datos históricos con los que se pueden predecir y efectuar simulaciones en el pasado para poder entender y adelantarse al movimiento de cualquier mercado.

Como se ha mencionado previamente el incentivo al realizar esta investigación reside en dar un hallazgo en la capacidad de predicción de nuestro modelo aplicado a los mercados mas importantes del mundo y dar una perspectiva diferente a los hallazgos hechos en el pasado, y que busca monitorizar cambios en la medida de riesgo mas importante, la volatilidad, por medio de cambios en los retornos diarios, de forma anticipada y precisa con la tecnología como soporte.

Se tiene en consideración que para un análisis confiable del modelo se necesita extraer una gran cantidad de datos de dos fuentes distintas en lo que es imprescindible que los diferentes componentes tecnológicos sean capaces de extraer, procesar y verificar los flujos de datos obtenidos, por lo cual se intenta esbozar la arquitectura tecnológica de la forma mas simple posible para que dicha aplicación tenga cabida en otros contextos de análisis de datos futuros.

# Metodología y ejecución del proyecto

El modelo utiliza como datos de entrada la serie de cierre de posiciones de los últimos 10 años de unos de los índices bursátiles más importantes existentes en los mercados financieros, el IBEX35 del mercado español, el S&P 500 índice estadounidense y posiblemente el índice más característico del clima mercantil actual, y el NASDAQ 100, también de origen estadounidense que aglomera los 100 valores de las compañías de tecnología más relevantes en este mismo mercado. La fuente de los datos proviene de [Yahoo Finance](https://es.finance.yahoo.com/) y son extraídos los datos históricos como series temporales con frecuencia diaria.

También se utiliza como datos de entrada una serie temporal de tendencias de búsqueda en Google el cual se extrae a través del servicio de [Google Trends](https://trends.google.es/) y que muestra el conteo de búsquedas que se han hecho diariamente a una palabra determinada.

Teniendo estas dos series temporales se procede a ver un patrón entre la tendencia de búsquedas del índice o una empresa específica del mismo, y la frecuencia de variación en su precio de cierre. Como métrica para valorar la bondad de la predicción se ha establecido una distribución para cada serie en la que la volatilidad forma parte del límite en el que se considera si el cambio en el precio o en la tendencia de búsquedas es atípica.

Para llevar a cabo nuestro trabajo debemos convertir las series continuas de volatilidades y de volumen de búsquedas en series binarias. Este es un método que entra dentro de la literatura de extracción de características, según la cual extraemos de un proceso continuo una característica como es su comportamiento en valores extremos en función del tiempo. Esto nos permite evaluar nuestra hipótesis con técnicas similares a las antes mencionadas sobre dichas nuevas características. En cuanto a los modelos de predicción utilizados en la literatura de sentimiento de mercado y atención encontramos tanto modelos estadísticos y econométricos tradicionales, como modelos de aprendizaje automático tanto de clasificación como de predicción. Para una detallada literatura de modelos de predicción referir a G. Mahalakshmi et al. [18]. No incidimos en los modelos de predicción porque a diferencia de la mayoría de los estudios en la literatura nosotros no utilizamos un modelo para predecir o tomar decisiones ajustando unos parámetros, o una ecuación a los datos, sin embargo probamos la hipótesis que el volumen de búsquedas de ciertos términos conocidos de antemano categorizados en valores extremos tienen una beta de 1 (sin termino independiente) con cambios extremos del mismo orden estadístico de los retornos diarios del activo y su subsecuente impacto en la volatilidad.

Además, ajustamos una regresión logística a las series binarias de tendencias y volatilidad extremos. La finalidad no es para nada predictiva como hemos justificado antes, sino de inferencia estadística para entender mejor la relación entre ambas series, que al ser binarias debemos usar la regresión logística, y con la que podemos, mediante medidas de precisión en muestra y otros parámetros realizar dichas inferencias.



# Preprocesado de los datos

La data de las dos series es extraída mediante un servicio hecho con el framework NodeJS el cual hace uso del paquete de npm google-trends-api y yahoo-finance:

const googleTrends = require('google-trends-api');

const yahooFinance = require('yahoo-finance');

Se desarrolla un api que hace uso de estas dos api’s de npm:

##### Servicio de Google trends:

app.get('/volatilitypred/extractTrends/:trendWord/:startDate/:endDate', (req, res, next) => {

  const keyword = req.params.trendWord;

  const startDate = new Date(req.params.startDate);

  const endDate = new Date(req.params.endDate);

  const optionsObject = {

    keyword: keyword,

    startTime: startDate,

    endTime: endDate

  }

  googleTrends.interestOverTime(optionsObject)

  .then(function(results){

     results = JSON.parse(results).default.timelineData;

     results.forEach(function(result) {

      result.time = result.time

      result.formattedTime = result.formattedTime;

      result.formattedAxisTime = result.formattedAxisTime;

      result.value = result.value[0];

      result.formattedValue = result.formattedValue;

      result.symbol= keyword;

    });

    res.status(201).json({

      gtrendsdata: results

    });

  })

  .catch(function(err){

    console.error('Oh no there was an error', err);

  });

});

El cual responde con datos con la siguiente estructura:

Time: Fecha en formato numerico

Formattedtime: Fecha de consulta de la palabra

FormattedAxisTime: Fecha del registro

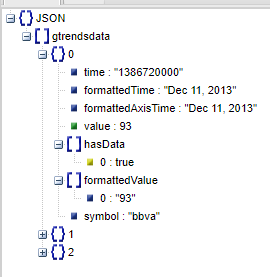
Value: Conteo total de la tendencia (palabra buscada) en registro.

hasData: Si value arroja un resultado mayor a 0

symbol: Palabra buscada

Aquí en la imagen 1 un ejemplo de la respuesta del api rest:

Imagen 2



##### Servicio de Yahoo finance:

app.get('/volatilitypred/extractFinance/:stock/:startDate/:endDate', (req, res, next) => {

  const startDate = new Date(req.params.startDate);

  const endDate = new Date(req.params.endDate);

  const optionsObject = {

    symbol: req.params.stock,

    from: startDate,

    to: endDate

  }

  yahooFinance.historical(optionsObject)

    .then(function(results){

      res.status(201).json({

        stockdata: results.reverse()

      });

  });

});

Los datos provenientes de yahoo cuentan con la siguiente estructura:

Date: Fecha en que se hace el registro.

Open: Precio de apertura en el mercado del activo, bono o valor.

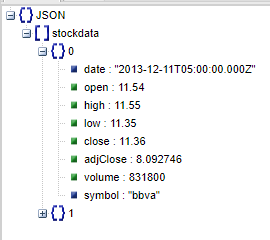
High: Precio más alto alcanzado entre la hora de apertura y la de cierre.

Low: Precio más bajo alcanzado entre la hora de apertura y la de cierre.

Close: Precio de cierre del día.

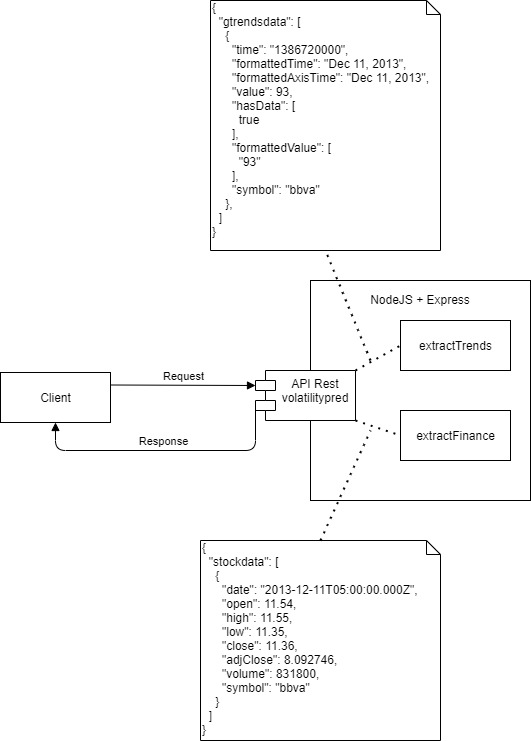
Adj Close: Precio de cierre del día después de pagar dividendos o stock splits. Este en particular va a ser el que incluiremos en la muestra de preprocesado de datos posterior, debido a que es mas preciso a la hora de valorar la subida o la caída de una acción gracias al análisis en los cálculos de los diferentes factores contemplados. Aquí en la imagen 2 un ejemplo de la respuesta del api rest:

Imagen 3



Como se muestra en el siguiente diagrama el api rest está disponible para ser consumida por cualquier cliente a través de una petición http:

Imagen 4

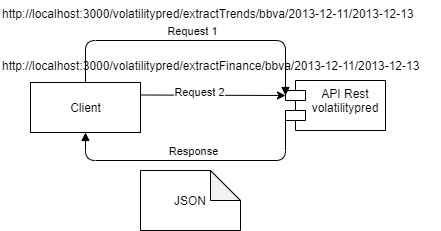


# Exploración de datos

En este apartado y correspondiente a la imagen 3, nos centraremos en el cliente el cual está encargado de hacer la extracción de los datos y pre transformarlos de tal forma que podamos manipularlos fácilmente para verificar la hipótesis.

Se hacen dos llamados al api, uno para extraer Google trends y el otro para yahoo finance. La respuesta se recibe en formato JSON:

Imagen 5



Y desde el cliente efectuamos una transformación por cada serie recibida:

def extract\_data(getDataFromAPI: GetDataFromAPI, start\_date, end\_date):

    response\_trend\_df = getDataFromAPI.request\_transform\_trend\_data(start\_date, end\_date)

    response\_finance\_df = getDataFromAPI.request\_transform\_finance\_data(start\_date, end\_date)

    return response\_trend\_df, response\_finance\_df

def request\_transform\_trend\_data(self, start\_date, end\_date) -> pd.core.frame.DataFrame:

        response\_json = self.request\_service(CONSTANTS.googleTrendsAPIMethodURL + self.trend\_word + '/' + start\_date + '/' + end\_date)

        if len(response\_json['gtrendsdata']) > 0:

            response\_df = pd.DataFrame(response\_json['gtrendsdata'])\

                .drop(['formattedTime', 'formattedAxisTime', 'formattedValue', 'hasData', 'symbol'], axis=1)

            response\_df['time'] = response\_df['time'].apply(self.format\_date\_fromtimestamp)

            response\_df = response\_df.rename(columns={"time": "date"}, errors="raise")

        else:

            response\_df = None

        return response\_df

def request\_transform\_finance\_data(self, start\_date, end\_date):

        response\_json = self.request\_service(CONSTANTS.yahooFinanceAPIMethodURL + self.stock + '/' + start\_date + '/' + end\_date)

        if len(response\_json['stockdata']) > 0:

            response\_df = pd.DataFrame(response\_json['stockdata'])\

                .drop(['symbol', 'open', 'close', 'volume', 'low', 'high'], axis=1)

            response\_df['date'] = response\_df['date'].apply(self.format\_date\_string)

        else:

            response\_df = None

        return response\_df

De esta forma se cargan los datos ya transformados en dataframe con las estructuras siguientes. En la tabla se muestra un ejemplo de la serie temporal extraída de Yahoo Finance:

Tabla 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Adj Close | Volume |
| 28/07/2011 | 9544.299805 | 9707.400391 | 9517.099609 | 9657.000000 | 9656.990234 | 247610000 |
| 29/07/2011 | 9540.299805 | 9679.000000 | 9499.799805 | 9630.700195 | 9630.690430 | 254006000 |
| 01/08/2011 | 9713.000000 | 9752.299805 | 9316.400391 | 9318.200195 | 9318.190430 | 290079000 |
| 02/08/2011 | 9236.099609 | 9342.599609 | 9114.900391 | 9114.900391 | 9114.891602 | 387974000 |
| 03/08/2011 | 9033.700195 | 9297.299805 | 8958.599609 | 9037.700195 | 9037.691406 | 306543000 |
| 04/08/2011 | 9217.700195 | 9219.700195 | 8686.500000 | 8686.500000 | 8686.491211 | 385372000 |

Date: Fecha de registro de la tendencia (palabra buscada).

Value: Conteo total de la tendencia (palabra buscada) en registro.

La siguiente tabla muestra un ejemplo de los datos provenientes de Google trends:

Tabla 2

|  |  |
| --- | --- |
| **date** | **value** |
| 2011-01-02 00:00:00 | 50 |
| 2011-01-03 00:00:00 | 36 |
| 2011-01-04 00:00:00 | 45 |
| 2011-01-05 00:00:00 | 44 |
| 2011-01-09 00:00:00 | 57 |
| 2011-01-10 00:00:00 | 42 |
| 2011-01-11 00:00:00 | 20 |

# Modificación y transformación de los datos

En el marco del desarrollo de la hipótesis se hace un procesado de los datos extraídos de yahoo finance y Google trends, y se hace una poda inicial de las series temporales eliminando los sábados de las dos series, domingos de la serie finance y los viernes de la serie trends quedando con los días lunes a viernes para los datos yahoo finance y domingo a jueves para los de Google trends:

response\_trend\_df = prd.prune\_day\_from\_dataframe(response\_trend\_df, [4, 5])  # Viernes y Sabado

response\_finance\_df = prd.prune\_day\_from\_dataframe(response\_finance\_df, [5, 6])  # Sabado y Domingo

def prune\_day\_from\_dataframe(df, days\_to\_exclude):

    # Quita un día determinado de un time series data frame

    df["weekday"] = pd.to\_datetime(df.date).dt.dayofweek

    for day\_to\_exclude in days\_to\_exclude:

        df = df.loc[df["weekday"] != day\_to\_exclude]

    df = df.reset\_index(drop=True)

    return df

La razón de purgar estos datos es que el mercado de stocks no se mueve en días no hábiles pero las tendencias de búsqueda contienen información de todos los días calendario del año, por lo cual se deben ajustar las dos series entre si eliminando el ruido que genera la serie de Google trends. A tener en cuenta y es que con el siguiente algoritmo se ajustan las dos series también contemplando los días feriados:

def prune\_row\_correspondance\_by\_value(left\_df, right\_df,

                                      column\_name, value, right\_correspondance, date\_column\_name):

    # Selecciono primero los valores que se van a podar en el primer dataframe

    df = left\_df.loc[left\_df[column\_name] == value].copy()

    # Se guarda un array con las fechas que van a servir para podar los datos en el otro array

    df[date\_column\_name] = df[date\_column\_name] + timedelta(days=right\_correspondance)

    date\_np = np.array(df[date\_column\_name])

    # Borro en el df derecho

    for date\_ in date\_np:

        right\_df = right\_df[right\_df[date\_column\_name] != date\_]

    right\_df = right\_df.reset\_index(drop=True)

    # Borro en el df izquierdo

    left\_df = left\_df.loc[left\_df[column\_name] != value]

    left\_df = left\_df.reset\_index(drop=True)

    return left\_df, right\_df

Finalmente, y en consiguiente a la transformación y purga de los datos de las dos series con la ayuda de la librería numpy se obtienen los diferentes datos estadísticos que vamos a usar para binarizar las series, la media y la variación estándar (sigma). D este modo tenemos dos muestras con su correspondiente distribución.

Posterior a la generación de la distribución de las dos muestras y como se menciona previamente para la serie de yahoo finance se tomará como referencia el precio de cierre ajustado debido a que es más preciso a la hora de valorar la subida o la caída de una acción gracias al análisis en los cálculos de los diferentes factores contemplados. Con respecto al precio de cierre se hará un cálculo de los retornos en t con respecto a t-1:

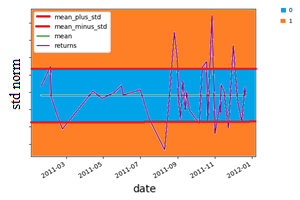
De esta forma se calculan iterativamente los retornos para cada instante t:

Tabla 3

|  |  |  |
| --- | --- | --- |
| **date** | **adjClose** | **returns** |
| 2011-01-03 00:00:00 | 5,881024 | 0 |
| 2011-01-04 00:00:00 | 5,869412 | -0,00197449 |
| 2011-01-05 00:00:00 | 5,74485 | -0,02122223 |
| 2011-01-06 00:00:00 | 5,504013 | -0,04192224 |
| 2011-01-10 00:00:00 | 5,304294 | -0,03628607 |
| 2011-01-11 00:00:00 | 5,410028 | 0,019933661 |

Y con respecto a este retorno se obtiene la media y variación estándar de la muestra. La media será la constante más cercana a cada punto de la distribución y sigma+media y sigma-media serán los limites en los cuales se considera que la mayoría de los datos de retorno se encuentran aglomerados, los puntos de la serie que se encuentren por fuera de estos límites (límite superior y límite inferior) son retornos atípicos por lo cual nos servirá para poder hacer una clasificación binaria en función del tiempo t:

Imagen 6



Representación gráfica de binarización de la serie de volatilidad

# Clasificación y predicción

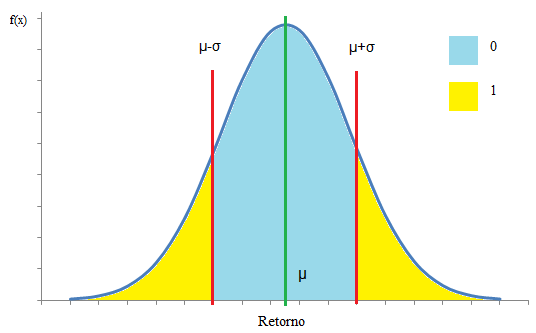
Por lo anterior lo que se encuentre por fuera de los limites sigma+ media y sigma- media en la hipótesis se toma con un cambio brusco en el retorno y por tanto una volatilidad por fuera de la normal.

En el caso de la serie de tendencias se calcula la media y variación estándar con respecto del valor del conteo de búsqueda del stock y se obtiene su distribución. Los retornos del stock en cada que se encuentren por fuera de los límites de su sigma+ y sigma- se considera una señal para el posible cambio atípico en el retorno del stock en un instante futuro, es decir, un valor binario positivo en la serie de tendencias en el instante t, implicaría un aumento o decremento considerable en el retorno del activo que generaría un retorno por fuera de su volatilidad en t+1:

Imagen 7



Imagen 8



Representación gráfica de binarización de la curva de volatilidad en una distribución normal

# Gestión del proyecto

# Planificación

Se divide el proyecto en 7 fases desarrolladas en paralelo, en la fase de planeación se realiza un esquema de las diferentes tareas del proyecto desmenuzando cada uno de los componentes necesarios para cumplir la meta del proyecto. En la fase de requerimientos se plantean los casos de uso de las piezas de la aplicación para la extracción y el procesado de los datos. En el diseño se desarrolla la arquitectura de gobierno de la aplicación. En la fase de documentación se tienen en cuenta la investigación previa al modelo de hipótesis, así como su motivación y explicación del modelo usado. Para la fase de codificación entramos en materia técnica en la que desarrollamos los requerimientos de la aplicación previamente documentados. En fase test se realizan pruebas de extracción y procesado sobre diferentes stocks con diferentes ventanas de tiempo para luego comenzar con la fase de extracción y procesado en la que se corre el modelo para los diferentes índices bursátiles propuestos y así poder hacer el análisis de resultados con las métricas obtenidas a partir de los datos de cada stock:

Tabla 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fase** | **Parte** | **Tarea** | **# Engineers** | **Horas** |
|  |
|  |
|  |
|  |
|  |
| PLAN | 1 | Planeación | 1 | 56,00 |  |
| REQ | 1 | Requerimientos | 1 | 15,00 |  |
| DES | 1 | Diseño | 1 | 48,00 |  |
| DOC | 1 | Documentación | 1 | 168,00 |  |
| CODE | 1 | Codificación | 1 | 112,00 |  |
| T | 1 | Test | 1 | 24,00 |  |
| EP | 1 | Extracción y procesado | 1 | 8,00 |  |
| AR | 1 | Análisis de resultados | 1 | 112,00 |  |
|  |  |  |  | 543,00 |  |

# WBS

Acá utilizamos un work breakdown structure para especificar las subtareas de cada fase del proyecto con su respectivo costo de horas, con un total de una duración de 567 horas, siendo la fase mas costosa la de documentación en la que se involucra la investigación de la literatura y desarrollo del documento investigativo.

Tabla 5

|  |  |  |  |
| --- | --- | --- | --- |
| **Fase** | **Tarea** | **Subtarea** | **Duración** |
| PLAN | Planeación | Introducción al proyecto por parte del tutor | 2,00 |
| Lectura artículos evaluación de métricas para datos desbalanceados | 18,00 |
| Lectura artículos sobre machine learning orientado a procesado de datos con python | 16,00 |
| Delimitar la problemática | 8,00 |
| Planteamiento de objetivos | 12,00 |
| REQ | Requerimientos |  | 15,00 |
| DES | Diseño | Modelo de datos tendencias | 2,00 |
| Modelo de datos finance | 2,00 |
| Backend API extracción de datos | 16,00 |
| Backend preprocesado y procesado de datos | 26,00 |
| Output de datos | 2,00 |
| DOC | Documentación | Recopilación bibliografía | 28,00 |
| Evaluación bibliografía dentro del marco del proyecto | 6,00 |
| Diseño de índice | 8,00 |
| Documentación de definiciones | 8,00 |
| Estado del arte | 60,00 |
| Metodología del proyecto | 24,00 |
| Gestión del proyecto | 8,00 |
| Análisis de resultados | 24,00 |
| Conclusiones | 2,00 |
| CODE | Code | Servidor http | 1,00 |
| Cascaron API de extracción de datos | 4,00 |
| función extractTrend | 16,00 |
| función extractFinance | 16,00 |
| parametrizacion constantes de datos stocks | 4,00 |
| obtener data de API trend | 2,00 |
| obtener data de API finance | 2,00 |
| formatear y transformar data | 20,00 |
| podar data | 16,00 |
| normalizar series | 8,00 |
| binarizar series | 4,00 |
| alinear binarias tendencias y finance | 8,00 |
| extraer datos estadísticos | 6,00 |
| generar reporte de hipótesis | 5,00 |
| T | Test | extracción de datos | 4,00 |
| transformación data trend | 1,00 |
| transformación data finance | 1,00 |
| poda serie trend | 4,00 |
| poda serie finance | 4,00 |
| obtener media de tendencias | 0,50 |
| obtener desviación estándar de tendencias | 0,50 |
| binarizar tendencias a partir de desviación y media | 0,50 |
| obtener media de finance | 0,50 |
| obtener desviación estándar de finance | 0,50 |
| binarizar finance a partir de desviación y media | 0,50 |
| calcular retornos por día en serie finance | 1,00 |
| combinar series tendencias y finance | 1,00 |
| obtener matriz de confusión | 5,00 |
| obtener test de fisher | 4,00 |
| obtener sensibilidad | 2,00 |
| obtener especificidad | 2,00 |
| generar reporte por stock | 8,00 |
| generar reporte aglomerado | 8,00 |
| EP | Extracción y procesado | Ejecución modelo en Índices bursátiles | 0,33 |
| Ejecución modelo en IBEX35 | 0,33 |
| Ejecución modelo en DAX30 | 0,33 |
| Ejecución modelo en S&P500 | 7,00 |
| AR | Análisis de resultados | Aplicación de modelo en índices bursátiles | 8,00 |
| Aplicación de modelo en IBEX35 | 24,00 |
| Aplicación de modelo en DAX30 | 24,00 |
| Aplicación de modelo en S&P500 | 24,00 |
| Análisis de inferencia estadística | 32,00 |
|  |  |  | 566,99 |

# Presupuesto

En la siguiente tabla se involucran los roles participantes y numero de colaboradores por rol. El costo hora/hombre se toma de paginas portal de empleo como Indeed y Glassdoor, y el total de este proyecto es de 9745 €, teniendo en cuenta que el ingeniero de software invirtió la mayor parte del tiempo en el desarrollo del proyecto.

Tabla 6

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Personal** | | | | | |
| **Rol** | **# Colaboradores** | **Horas/hombre** | **Precio/Hora** | **Total** | **Total con impuestos** |
| Product owner | 1 | 96 | 15,00 € | 1.440,00 € | 1.684,80 € |
| Ingeniero de software | 1 | 567 | 12,15 € | 6.889,05 € | 8.060,19 € |
| **Costo total proyecto** |  |  |  |  | **9.744,99 €** |

# Análisis de resultados

Entrando en materia en el análisis de resultados, se entiende de acuerdo al apartado 3.4 (Clasificación y predicción) de este documento que se tienen dos series temporales, la de tendencias y de volatilidad las cuales se binarizan mediante una distribución normal, obteniendo la media y desviación estándar de tendencias y tomando como limites μ + σ y μ – σ, tal que si el tick de la tendencia se encuentra por fuera de estos límites, se marca como 1, sino, es decir, si se encuentra por dentro de μ + σ y μ – σ, se marca como 0. El método es similar para la serie de volatilidad en la cual se obtiene la media y la desviación estándar (volatilidad) de los retornos de la muestra completa, y, si el retorno en el tick t se encuentra por fuera de los limites μ + σ y μ – σ, se marca como 1, si no, el retorno respecto del total de la muestra no sufrió un cambio tal que haya superado por el limite superior o inferior la volatilidad esperada.

# Experimento

Los resultados se basan en la predicción diaria de un aumento de volatilidad en el precio del stock por fuera de la normal en función de la distribución normal de la serie de tendencias. En este caso de estudio se usaron dos series binarias resultantes del procesado de los datos sobre las series de tendencias de Google trends y movimientos de precio de cierre ajustado de yahoo finance extraídos sobre los índices más representativos en los mercados internacionales; IBEX35, DAX30 y S&P500. Adicionalmente y complementando el análisis de los resultados se ha utilizado la serie trends como predictor de la serie de volatilidad, es decir, contando el numero de 0 y 1 presente en volatilidad en t+1 que siguen a una tren en t 0 o 1, se ha obtenido una matriz de confusión con la que se obtiene el conteo de verdaderos positivos (True positive: TP), verdaderos negativos (True negative: TN), falsos positivos (False positive: FP), y falsos negativos (False negative: FN) que nos permite calcular posteriormente la precisión, sensibilidad, y especificidad para esbozar la calidad de clasificación en torno a su predicción. Por tanto, en este apartado se compara la precisión obtenida en la predicción de la serie de volatilidad, usando como predictor la serie trends.

Para contextualizar los resultados en función de los datos de la matriz de confusión a continuación una tabla que enumera las distintas métricas que se utilizaron para su análisis:

Tabla 7

|  |  |  |  |
| --- | --- | --- | --- |
|  | **predicción positiva** | **predicción negativa** |  |
| clase positiva | TP | FN (Error) | Sensibilidad: TP/(TP+FN) |
| clase negativa | FP (Error) | TN | Especificidad: TN/(TN+FP) |
|  | Precisión binaria: TP/(TP+FP) |  |  |

En la tabla 7 se hace un esquema de acuerdo con los valores alojados en la matriz de confusión. Tenemos las clases a predecir como son la positiva, representada por 1 y la negativa representada por 0, y cruzando con las predicciones positivas y negativas para cada clase. Denotamos una diagonal por cada color, el verde significa la diagonal de predicción acertada, y el salmón la diagonal de predicción errada. Junto con esta data tenemos unas métricas que miden la calidad de los resultados; la sensibilidad, que nos dice que tan bien se predijo la clase positiva, la especificidad que nos dice que tan bien se predice la clase negativa, y la precisión que nos informa cuantos de los predichos son correctos.

Cabe aclarar que en este análisis los datos de precisión calculados son dados por la fórmula de precisión ponderada (weighted precision), dado que este tipo de precisión nos da un mejor desempeño debido a que las frecuencias de las dos clases pueden ser muy diferentes y hacer el solo computo de una precisión binaria hace que las dos clases tenga la misma importancia. Por tanto, siempre que nos referimos a precisión, se tiene como referencia al calculo dado por la precisión ponderada, dado por:

Donde y son las precisiones de las dos clases y c1 y c2 son el numero de instancias de cada clase

Los conjuntos de datos usados se establecen en dos grupos para ser analizados, el conjunto de los índices bursátiles y el conjunto de cada índice con sus respectivos stocks:

    ALL\_INDEX\_LIBRARY = {

      '^IBEX': 'IBEX',

      '^GDAXI': 'DAX',

      '^GSPC': 'S&P 500'

    }

    IBEX\_STOCK\_LIBRARY = {

      'ANA.MC': 'Acciona S.A.',

      'ACX.MC': 'Acerinox',

      'ACS.MC': 'ACS Actividades de Construccion y Servicios',

    DAX30\_STOCK\_LIBRARY = {

      'ADS': 'Adidas',

      'ALV': 'Allianz',

      'BAS': 'BASF',

    SP500\_STOCK\_LIBRARY\_1 = {

      'MMM': '3M Company',

      'ABT': 'Abbott Laboratories',

La transformación de las series temporales a las series binarias se genera a partir de la distribución de las series temporales extraídas de Google y Yahoo en la cual cada día se transforma en 1 o 0 dependiendo de las siguientes condiciones:

* Si returns > μ + σ V returns < μ - σ: 1 Sino 0
* Si value > μ + σ V value < μ - σ: 1 Sino 0

El índice con mayor cantidad de data extraída de Google y yahoo entre 2012-01-01 y 2019-12-31 es el IBEX con un tamaño en cada serie de 2031. Las series más balanceadas en cuanto a misma cantidad de ceros y unos son la S&P500 para las tendencias con 1179 0’s y 820 1’s, es decir, una relación de 1.4:1, y el DAX30 para la finance con 1494 0’s y 514 1’s, es decir, una relación de 2.9:1. La precisión en el índice S&P500 como 61,54% la cual parece aportar una pequeña cantidad de información a la hora de predecir eventos de gran volatilidad. La precisión medida en el índice DAX30 es del 61,49% la cual está relativamente cerca a la obtenida en el índice S&P500. Relativo al índice IBEX35 se obtuvo una precisión del 60,3% y una relación de 0’s a 1’s en la serie de tendencias de 1.4:1 y la de finance sobre 3:1, por lo cual podemos ver que el modelo tuvo peor desempeño sobre los índices S&P500 y DAX30 que el IBEX35 siendo este ultimo ligeramente mas rico en datos; tamaño de serie de 2031 contra 2008 para el DAX30 y 1999 en el S&P500.

# Resultados

Se conformaron 4 grupos para el análisis de resultados en los que se encuentran primeramente el grupo de los índices bursátiles:

Tabla 8

|  |  |  |
| --- | --- | --- |
|  | **stock** | **value** |
| richest\_data\_stock | ^IBEX | 4062 |
| best\_precision\_weighted | ^GSPC | 0,615405 |
| best\_recall\_weighted | ^GDAXI | 0,559761 |
| most\_balanced\_binary\_trend | ^GSPC | \*1 |
| most\_balanced\_binary\_finance | ^GDAXI | \*2 |
| median\_precision\_weighted |  | 0,611139 |
| median\_recall\_weighted |  | 0,611139 |

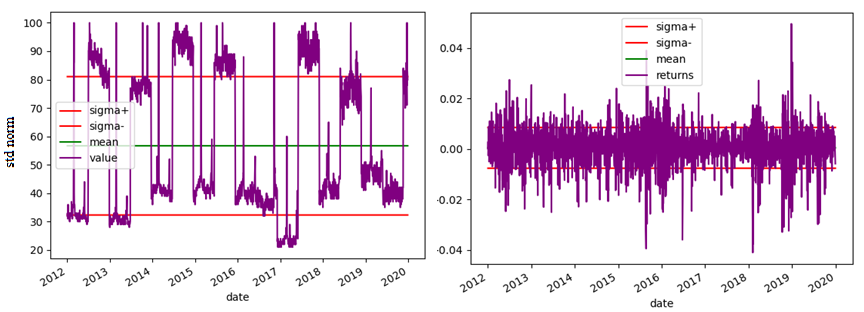
Respecto de la precisión encontramos que la mayor precisión en la predicción se encuentra en el S&P 500 con un 61.54%, y la mayor ratio de verdaderos positivos (recall) en el DAX30 con un 55,97%. La precisión y recall media se encuentran en un 61,11%.

Relativo a los índices cuyas series binarias están más balanceadas, encontramos que, para el S&P 500 la precisión es de un 61.54% y una ratio de verdaderos positivos de 52,42%, con la siguiente matriz de confusión:

Tabla 9

|  |  |  |
| --- | --- | --- |
| **S&P500** | **predicción positiva** | **predicción negativa** |
| clase positiva | 173 | 304 |
| clase negativa | 647 | 875 |

Imagen 9



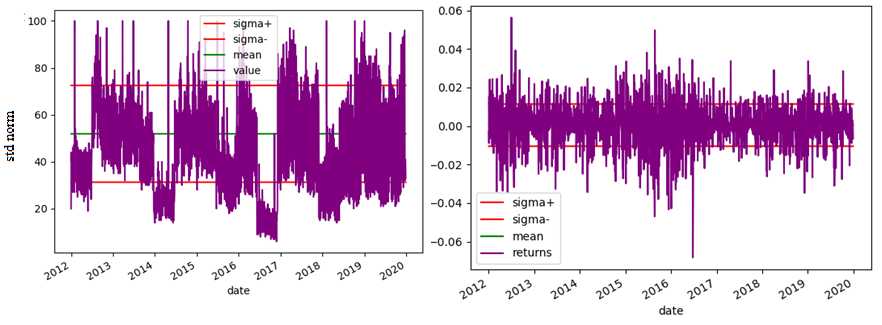
S&P500 Izquierda: Movimiento de la serie de tendencias. Derecha: Movimiento de la serie de volatilidad.

En referencia al índice DAX30 cuya serie finance es la mejor balanceada, podemos ver que, la precisión es del 61.49% y una ratio de verdaderos positivos de 55.97% muy similar a la predicción en el S&P 500. Tenemos la siguiente matriz de confusión:

Tabla 10

|  |  |  |
| --- | --- | --- |
| **DAX30** | **predicción positiva** | **predicción negativa** |
| clase positiva | 185 | 329 |
| clase negativa | 555 | 939 |

Imagen 10



DAX30 Izquierda: Movimiento de la serie de tendencias. Derecha: Movimiento de la serie de volatilidad.

Para el índice **IBEX35** el stock con mayor cantidad de data extraída de Tendencias y volatilidad entre 2012-01-01 y 2019-12-31 es el ArcelorMittal (MTS.MC) con un tamaño en cada serie de 2033. Las series más balanceadas en cuanto a misma cantidad de ceros y unos son la Ferrovial S.A. para las tendencias con 119 0’s y 91 1’s, es decir, una relación de 1.3:1, y el stock más balanceado para la volatilidad con 151 0’s y 59 1’s, es decir, una relación de 2.5:1. La precisión medida en el stock Ferrovial S.A es del 58,16%. La precisión medida en el stock ArcelorMittal es del 62,10% la cual está relativamente cerca a la obtenida en el stock Ferrovial S.A por unos 400pb. Relativo unos de los stocks más importantes está el BBVA en el cual se obtuvo una precisión del 63,26% y una relación de 0’s a 1’s en la serie de tendencias de 2.1:1 y la de finance sobre 0.4:1.

Pasando al sector energía, otra empresa representativa del índice el stock de Iberdrola en el cual se obtuvo una precisión del 65,58%, muy cercano a le media por una diferencia 4pb y una relación de 0’s a 1’s en la serie de tendencias de 2.2:1 y la de finance sobre 0.4:1, en los cuales se puede notar una diferencia muy ligera en sus resultados. Adicional cabe notar que el stock de Ferrovial S.A. teniendo la serie de tendencias y finance más balanceada entre todo el grupo de empresas, tiene la peor precisión en el modelo con un 58,16%, muy por debajo de la media de precisión en 65,54%. Cabe resaltar que siendo ArcelorMittal el stock con mayor cantidad de datos cuenta con una precisión en el modelo de 62,1%, es decir, 300pb por debajo de la media lo cual hace notar que aporta una conclusión relevante al modelo predictivo en el grupo del índice bursátil IBEX35, por otro lado, su (ArcelorMittal) ratio de 1’s esta muy por debajo de la media con una diferencia de 765pb lo cual nos devuelve un paso en cuanto a la calidad de la predicción. Con respecto a Iberdrola encontramos uno de los stocks mas aportantes a la calidad del modelo debido a su riqueza en datos igualando a ArcelorMittal, con una precisión muy cercana a la media y recall solo 500pb por debajo de la media. Hasta el momento este stock (Iberdrola) es el más aportante, ya que al ver su siguiente matriz de confusión podemos ver una buena predicción y baja frecuencia de errores.

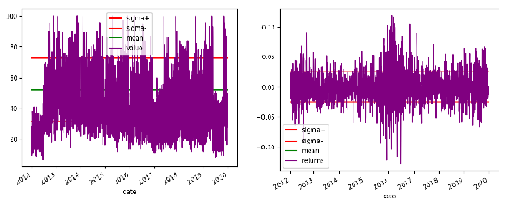
Tabla 11

|  |  |  |
| --- | --- | --- |
| **Iberdrola** | **predicción positiva** | **predicción negativa** |
| clase positiva | 141 | 309 |
| clase negativa | 493 | 1090 |

Tabla 12

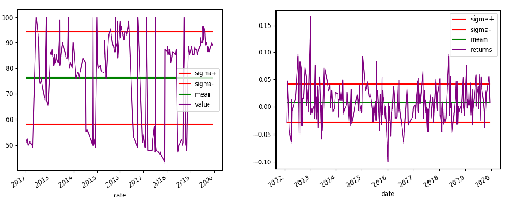
|  |  |  |
| --- | --- | --- |
|  | **stock** | **value** |
| richest\_data\_stock | MTS.MC | 4066 |
| best\_precision\_weighted | SGRE.MC | 0,805046 |
| best\_recall\_weighted | COL.MC | 0,759216 |
| most\_balanced\_binary\_trend | FER.MC |  |
| most\_balanced\_binary\_finance | FER.MC |  |
| median\_precision\_weighted |  | 0,655544 |
| median\_recall\_weighted |  | 0,655544 |

Imagen 11



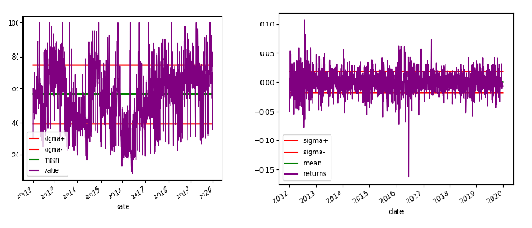
ArcelorMittal (MTS.MC) Izquierda: Movimiento de la serie de tendencias. Derecha: Movimiento de la serie de volatilidad.

Imagen 12



Ferrovial S.A (FER.MC) Izquierda: Movimiento de la serie de tendencias. Derecha: Movimiento de la serie de volatilidad.

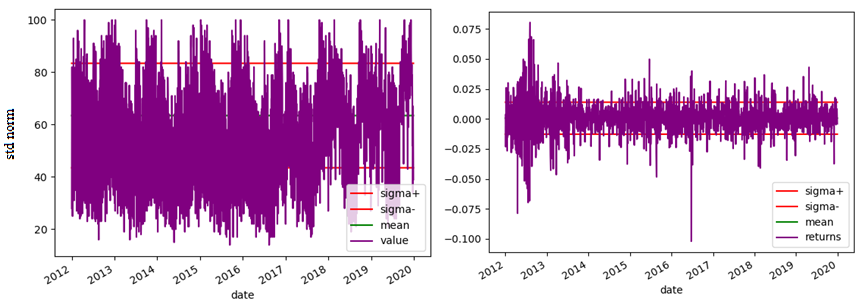
Imagen 13



BBVA (BBVA.MC) Izquierda: Movimiento de la serie de tendencias. Derecha: Movimiento de la serie de volatilidad.

Pasando al sector energía, otra empresa representativa del índice es Iberdrola tenemos una ratio de 1’s de 22,23% la cual esta casi 400pb por debajo del stock mejor balanceado del índice (Ferrovial S.A) pero con una mayor precisión que el mismo con un 65,58% con lo cual podríamos intuir que la precisión subió más por la cantidad de 0’s existentes en las dos series, por ejemplo, podemos ver que la proporción de 0’s en la binaria de tendencias es de 2,2:1 y finance aún más alto con 3,5:1. Por consiguiente la sensibilidad del modelo ejecutado frente al stock y el cual representa la calidad de predicción de la clase positiva, es de 31,3% y la especificidad representa la calidad de predicción de la clase negativa con un 8,9%, por tanto, para el stock Iberdrola podemos concluir que el resultado de la predicción es bastante aleatorizada debido a su bajo rendimiento en sensibilidad y especificidad.

Imagen 14



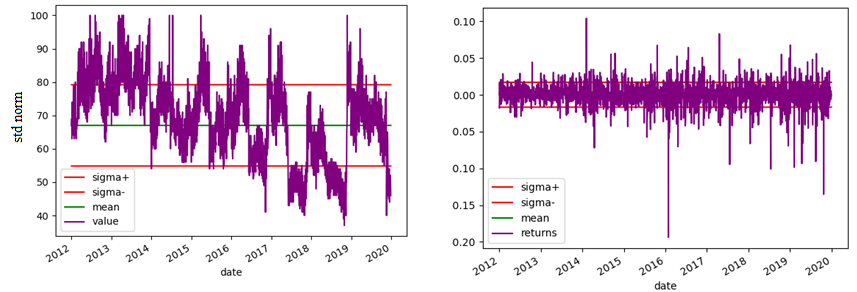
Iberdrola (IBE.MC) Izquierda: Movimiento de la serie de tendencias. Derecha: Movimiento de la serie de volatilidad.

El siguiente índice **DAX30** contamos con los stocks mas ricos en datos en las series finance y tendencias como Adidas del sector textil, Allianz de seguros, Deutsche post de correos, Deutsche Telekom de telecomunicaciones, SAP de servicios de tecnología, entre otras, por lo cual tenemos sectores bastante variados que nos ayudaran a potenciar nuestro análisis sobre el modelo predictivo. En el stock Adidas tenemos una precisión del 65,17% con una diferencia por debajo de la media de precisión del índice de mas de 1500pb el cual es de 81,19%, y aunque es una precisión importante podemos ver que las dos series están bastante desbalanceadas con la de tendencia con una proporción de 1’s de 0,46:1 y la de finance con 0,26:1 y sumado a esto cuenta con pvalue de Fisher bastante bajo con 1,28% lo que indica una alta dependencia entre las predicciones de la clase negativa y positiva, adicionalmente se le atribuye la mayor predicción a la clase positiva con una proporción del 26,74% sobre la negativa con 7% como puede verse en la siguiente matriz de confusión:

Tabla 13

|  |  |  |
| --- | --- | --- |
| **Adidas** | **predicción positiva** | **predicción negativa** |
| clase positiva | 111 | 304 |
| clase negativa | 526 | 1058 |

Imagen 15



Adidas (ADS) Izquierda: Movimiento de la serie de tendencias. Derecha: Movimiento de la serie de volatilidad.

Por otro lado, con respecto al stock Allianz si tenemos un balance ligeramente mas equidistante con una precisión del 63,35% y proporción de la serie de tendencias y finance de 0,57:1 y 0,30:1 respectivamente, pero con una relación de independencia entre la predicción positiva y negativa mas marcada con un pvalue en el test de Fisher de 44,37% e igualmente la predicción de la clase positiva se le atribuye la mejor precisión con un 35,10 como se puede ver en la siguiente matriz de confusión:

Tabla 14

|  |  |  |
| --- | --- | --- |
| **Allianz** | **predicción positiva** | **predicción negativa** |
| clase positiva | 165 | 305 |
| clase negativa | 568 | 961 |

Respecto de Deutsche post, tiene una precisión bastante alta y con un delta de únicamente 0,15pb por debajo de la media con un 81,17%, también es visible la desproporción de 1’s con respecto de 0’s en las dos series con una razón de 1:2 y 0,12:1 de 1’s para tendencias y finance respectivamente, adicionalmente su pvalue está en 22,54% lo que indica que la alta precisión se debe a la cantidad desbalanceada de 0’s en las dos binarias:

Tabla 15

|  |  |  |
| --- | --- | --- |
| **Deutsche Post** | **predicción positiva** | **predicción negativa** |
| clase positiva | 82 | 137 |
| clase negativa | 591 | 1189 |

Mirando Deutsche Telekom, el cual tiene la precisión más baja del DAX30 de 58,95%, con un balance de series respecto de DPW (Deutsche post) y muy similar a ALV (Allianz) con 0,51:1 y 0,37:1 para las binarias de tendencias y finance respectivamente, DTE (Deutsche Telekom) tiene el conjunto de datos de finance mejor balanceado entre los demás stocks del DAX30, además del mejor grado de sensibilidad con un 11,45%, pero con pvalue mucho más bajo con un 8,88% con el que podríamos intuir una alta dependencia entre la precisión en la predicción y la predicción de la clase negativa.

En concordancia con Sap, que, aunque no cuenta con tan buena precisión con 62,84%, es un caso particular ya que su pvalue indica una independencia entra la predicción negativa y positiva con 69,62% pero su desbalance es notable en las dos series con 0,25:1 y 0,32:1 para tendencias y finance respectivamente.

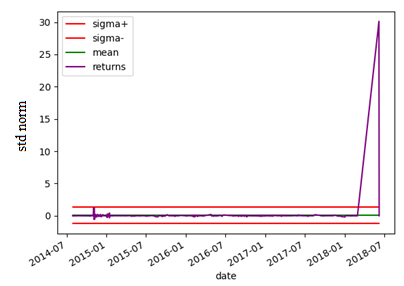
Dentro de los stocks mejor balanceados respecto de la serie finance se encuentra el stock MRK (Merck) con una razón de 1’s en su serie de 0,3:1, pero cuenta con la peor ratio de verdaderos positivos, es decir, un grado muy bajo de los que debería predecir la clase positiva que debería predecir con 54,77% como se puede ver en la siguiente matriz de confusión:

Tabla 16

|  |  |  |
| --- | --- | --- |
| **Merck** | **predicción positiva** | **predicción negativa** |
| clase positiva | 175 | 289 |
| clase negativa | 615 | 920 |

La mejor precisión en el conjunto de stocks del DAX30 esta en el RWE (RWE) con 99,82%, y se debe a su desbalance en la serie binaria de finance la cual solo cuenta con un 1 dentro de 583 ticks como se puede ver a continuación:

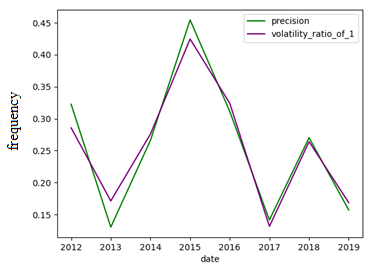
Imagen 16



RWE (RWE) Movimiento de la serie de volatilidad

Juzgando únicamente por la precisión el modelo aplicado sobre los stocks del DAX30 tienen un buen desempeño debido a que la media esta mas cerca de su pico superior que de su pico inferior (superior: 99,82%, inferior: 58,95%, media: 81,19%), y basado en su bajo balanceo en las series binarias y su pvalue con una media de 41,27% se puede concluir que la dependencia entre las predicciones de la clase negativa y la positiva son dependientes entre sí, y ya que la mayor calidad de predicción reside en su clase positiva con un 30,43% considerado como un valor bastante bajo y media de especificidad de 4,39%, podemos decir que el modelo aplicado al índice tuvo un bajo desempeño.

Imagen 17



DAX30 Frecuencia de movimiento en la precisión vs ratio de la binaria en función de la volatilidad por año

En la anterior imagen (Imagen 17) se evidencia una baja capacidad de predicción del modelo teniendo en cuenta que en la mayoría de los ticks de precisión se encuentran por debajo de la ratio de 1’s en la serie de volatilidad, y esto se debe en su mayoría al desbalanceo de clases resultantes a la aplicación del modelo, que, aplicando técnicas de machine learning como el de aplicar Regresión logística podrían afrontarse este tipo de problemas de desbalanceo para mejorar el resultado obtenido.

En el índice **S&P500** entre los stocks mas ricos en datos tenemos compañías del sector industrial como 3M, tecnología como Google, consumo discrecional como Darden Restaurants, energía como Apache Corporation, finca raíz como Boston Properties, químico como PPG Industries, entre otras, las cuales cuentan con cada serie binaria con 1999 ticks.

Tabla 17

|  |  |  |
| --- | --- | --- |
|  | **stock** | **value** |
| richest\_data\_stock | MMM,DRI,GOOG, APA,BXP,CI,EW,GRMN,INTC, KR,MRK,NRG,SBAC,TIF,VNO,PPG | 3998 |
| best\_precision\_weighted | LHX | 0,870553 |
| best\_recall\_weighted | ISRG | 0,694347 |
| most\_balanced\_binary\_trend | LW |  |
| most\_balanced\_binary\_finance | VIAC |  |
| worst\_precision\_weighted | PNW | 0,56997 |
| worst\_recall\_weighted | SWK | 0,471236 |
| median\_precision\_weighted |  | 0,645772 |
| median\_recall\_weighted |  | 0,645772 |
| best\_sensitivity | LHX | 0,583333 |
| best\_specificity | GL | 0,159596 |
| best\_odds\_ratio | LHX | 3,188889 |
| best\_pvalue |  | 1 |
| worst\_sensitivity | TMO | 0,109312 |
| worst\_specificity | WMB | 0,028555 |
| worst\_pvalue | ['TMO'] | 1,37E-05 |
| median\_sensitivity |  | 0,30067 |
| median\_specificity |  | 0,090567 |
| median\_pvalue |  | 0,434177 |

Para enriquecer el análisis contamos con la mejor precisión por L3Harris Technologies (LHX) con un 87,05%, su peor precisión por Pinnacle West Capital (PNW) con 56,99% y su media de precisión con 64,57%, la cual nos da un primer indicio de que la muestra esta mas inclinada hacia el pico bajo, es decir, la distribución posee una asimetría positiva.

Dentro de los stocks más ricos en datos tenemos a 3M (MMM) con 1999 ticks en cada serie binaria, y con una precisión del 65,82%, es decir, mas de 100pb por encima de la media, tiene un balance de binarias bastante bajo con una proporción de 1’s de 0,3:1 y 0,27:1 para tendencias y finance respectivamente, pero un pvalue considerable de 94,98% que nos indica una independencia casi perfecta entre la predicción de la clase positiva y la clase negativa, respecto de la calidad de predicción el panorama es negativo, ya que, su sensibilidad y especificidad se encuentran por debajo de la media con una diferencia de 5,75% y 2,27% respectivamente.

Respecto al stock de Google (GOOG) tenemos un panorama positivo ya que tiene una precisión con poco menos de 200pb por encima de la media con 66,31%, y, aunque el balance de las binarias no es lo esperado con una razón de 1’s de 0,38:1 y 0,27:1 para tendencias y finance respectivamente (ligeramente por encima de la razón en el stock 3M), su pvalue nos reporta una independencia favorable respecto de la predicción de clase positiva y la negativa con 80,74%. Bajo la calidad de predicción de las clases podemos decir que esta por debajo del promedio para la clase positiva y negativamente con 27,29% y 9,05% respectivamente.

Para el stock de Darden restaurants (DRI) del sector de consumo discrecional, tenemos una precisión mas baja que el stock de Google pero aun estando por encima de la media con 65,84%, además contamos con un mejor balance en las binarias con una razón de 1’s de 0,46:1 y 0,29:1 en tendencias y finance respectivamente y sensibilidad y especificidad por encima de la media con 34,64% y 10,32% frente a 30,06% y 9,05% respectivamente, pero un pvalue bastante bajo 13,76% que nos indica una baja calidad de predicción debido a la dependencia entre las predicciones de la clase negativa y positiva.

Pasando al sector energía en el stock de Apache corporation vemos que su precisión radica ligeramente por encima de la media por 8pb con 65,43%, aunque debido a su bajo balance de binarias en tendencias con razón de 1’s de 0,33:1 y finance con 0,30:1 se puede apreciar su alta dependencia entre las predicciones de las dos clases positiva y negativa:

Tabla 18

|  |  |  |
| --- | --- | --- |
| **Apache corp** | **predicción positiva** | **predicción negativa** |
| clase positiva | 131 | 337 |
| clase negativa | 372 | 1159 |

Debido al bajo balance que hemos visto en las binarias de los stocks previos pasamos a una de las empresas con una razón mas adecuada respecto de su binaria de tendencias, el stock de Lamb Weston Holdings (LW) cuya razón de tendencias esta en 1,04:1 y finance con 0,31:1 pero con una precisión por debajo de la media con 61,62% lo que nos podría dar un primer indicio de que para el índice los stocks con peor balance cuentan con mayor precisión y viceversa. Adicionalmente vemos que cuenta con una buena sensibilidad con 47,34% pero debido a su pvalue indica una alta dependencia entre ambas predicciones. Por otro lado, el stock con mejor balance en la binaria de finance es ViacomCBS (VIAC) cuenta con una alta precisión debido a que la muestra tiene muy pocos ticks (17) y el desbalance en su binaria de tendencias tira a tener mas 0’s con lo cual se genera una baja independencia entre las predicciones de las clases.

Analizando uno de los stocks con independencia perfecta entre predicción de las clases esta Abiomed Inc. (ABMD) la cual tiene una precisión del 70,87% pero una sensibilidad y especificidad bastante bajas con 22,03% y 4,7% respectivamente.

Por consiguiente, podemos concluir que la hipótesis aplicada a los stocks del S&P500 es de muy baja calidad teniendo en cuenta que la mayoría de las precisiones calculadas independientemente para cada stock se encuentran agrupadas bajo su media y que su ratio de 1’s en función de su cambio brusco en volatilidad está por encima de su precisión en función de cada año:

Imagen 18



S&P500 Frecuencia de movimiento en la precisión vs ratio de la binaria en función de la volatilidad por año



# Análisis de Inferencia Estadística:

En este apartado realizamos un análisis de inferencia estadística por medio de un ajuste de regresiones logísticas entre las series binarias extremas de Google Trends y retornos diarios para todos los activos del SP500 y del IBEX.

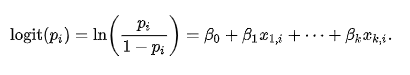


Imagen 19

Imagen 20

Tabla 19

|  |  |  |
| --- | --- | --- |
| **Precisiones Medias Estadísticas** | | |
|  | **IBEX** | **SP500** |
| **Media** | **78%** | **77%** |
| **Des. Típica** | **4%** | **3%** |
| **Max** | **87%** | **91%** |
| **Min** | **72%** | **65%** |

En los gráficos anteriores mostramos la precisión media en muestra para todos los constituyentes del IBEX y SP500 de las regresiones logísticas. En la tabla mostramos unos estadísticos de todas las precisiones medias para ambos índices. Anteriormente habíamos intentado probar la hipótesis de la relación directa entre la serie calculada para Google trends y retornos diarios al día siguiente. Ahora introducimos una ecuación para describir esa relación y pasa de ser uno a uno a una fórmula paramétrica determinada por la regresión logística, que se usa para series binaria y que tiene en cuenta la distribución de la muestra de 1’s y 0’s. Podemos ver como la precisión media es muy superior en media a el caso de las secciones anteriores de este documento cuando no utilizábamos una ecuación que explicara la relación y la relación era uno a uno. Mostramos en la tabla, la media, desviación típica y máximo y mínimo valor de las precisiones medias de los gráficos para el IBEX y SP500.

Ahora analizamos más en detalle los ganadores y perdedores en precisión con la regresión logística.

En el caso del IBEX tenemos los mejores resultados en Inmobiliaria Colonial (COL) 87%, Endesa (ELE) 85%, Inditex (ITX)84%, Red Eléctrica (REE) 87% y Siemens Gamesa (SGRE) 86%. La peor Ferrovial (FER) 72%.

En el caso del SP500 tenemos los mejores resultados en L3Harris Technologies (LHX) 91% (Empresa de Tecnología y Telecomunicaciones), Danaher Corp (DHR) 88% (Conglomerado Industrial enfocado en la tecnología y medio ambiente en todos los sectores), Eaton Corp (ETN) 88% ( Empresa del sector eléctrico, industria aeronáutica y energías), Constellation Brands (STZ) 87% ( Empresa de cerveza, vino y bebidas), Amcor (AMCR) 86% ( Empresa de empaquetados), Vertex (VRTX) 86% (Empresa farmacéutica), Williams (WMB) 86% (Empresa de Energías), DXC Technology (DXC) 85% (Empresa de servicios Tecnológicos), Under Armour (UA) 85% (Empresa de Ropa), Expedia (EXPE) 84% (Agencia de Viajes Online).

Las peores del SP500 son CBS Corporation 65% (Conglomerado de Medios de Comunicación), NortonLifeLock 70% (Empresa de Cyberseguridad), Dow Chemical 70% (Empresa Química), Pinnacle West Corp 71% (Empresa de Construcción).

Interesantemente, la compañía de peor rendimiento para predecir su retorno diario extremo con Google trends es una compañía de Noticias (CBS Corp). En cuanto a las otras compañías, vemos que por lo general las compañías que mejor se predicen son las de Electricidad, Tecnología, Energía y Ropa en ambos índices.

# Conclusiones

//opiniones

//Trabajo futuro

#TODO

- modelo de predicción con la beta para ajustar los parametros del modelo (la beta es la correlación entre las series)

- análisis de sentimiento en google trends (negativo - positivo) para prededcir la sigma atípica

- mismo analisis pero en tiempos de alta volatilidad o tiempos de crisis

# Bibliografía

[1] Stock market, \Stock market| Wikipedia, the free encyclopedia," 2004, [Online; accessed 28-July-2020]. [Online]. Available: https://en.wikipedia.org/wiki/Stock market

[2] New York Stock Exchange, \ New York Stock Exchange, | Wikipedia, the free encyclopedia," 2004, [Online; accessed 28-July-2020]. [Online]. Available: https://en.wikipedia.org/wiki/ New\_York\_Stock\_Exchange

[3] IBEX 35, \ IBEX 35, | Wikipedia, the free encyclopedia," 2011, [Online; accessed 28-July-2020]. [Online]. Available: https://en.wikipedia.org/wiki/IBEX\_35

[4] H. Markowitz, \PORTFOLIO SELECTION," The Journal of Finance, vol. 7, no. 1,

pp. 77{91, mar 1952.

[5] Nausheen S, Anil Kumar M, and Amrutha K K. (2017). “SURVEY ON SENTIMENT ANALYSIS OF STOCK MARKET.” International Journal of Research - Granthaalayah, 5(4) RACSIT, 69-75. https://doi.org/10.5281/zenodo.572298.

[6]Audrino, F.; Sigrist, F. & Ballinari, D.The impact of sentiment and attention measures on stock market volatility International Journal of Forecasting, Elsevier, **2020***, 36*, 334-357

[7] Huang, M.Y., Rojas, R.R. & Convery, P.D. Forecasting stock market movements using Google Trend searches. *Empir Econ* (2019). <https://doi.org/10.1007/s00181-019-01725-1>

[8] Chan WS. Stock Price reaction to news and no-news: drift and reversal after headlines Journal of Financial Economics, Elsevier, **2003***, 70*, 223-260

[9] Ingle V., Deshmukh S. (2017) Live News Streams Extraction for Visualization of Stock Market Trends. In: Lobiyal D., Mohapatra D., Nagar A., Sahoo M. (eds) *Proceedings of the International Conference on Signal, Networks, Computing, and Systems.* Lecture Notes in Electrical Engineering, vol 395. Springer, New Delhi. <https://doi.org/10.1007/978-81-322-3592-7_30>

[10] Vega, Clara and Albuquerque, Rui A., Economic News and International Stock Market Co-Movement (January 19, 2012). Review of Finance Vol. 13, 2009, EFA 2006 Zurich Meetings, Available at SSRN: <https://ssrn.com/abstract=1139927>

[11] Jiang, C., Liang, K., Chen, H. et al. Analyzing market performance via social media: a case study of a banking industry crisis. Sci. China Inf. Sci. 57, 1–18 (2014). <https://doi.org/10.1007/s11432-013-4860-3>

[12] S. A. Phand and J. A. Phand, "Twitter sentiment classification using stanford NLP," *2017* *1st International Conference on Intelligent Systems and Information Management (ICISIM)*, Aurangabad, 2017, pp. 1-5, doi: 10.1109/ICISIM.2017.8122138.

[12] Moat, H., Curme, C., Avakian, A. et al. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. Sci Rep 3, 1801 (2013). <https://doi.org/10.1038/srep01801>

[13] Dang, N.C.; Moreno-García, M.N.; De la Prieta, F. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics* 2020, *9*, 483.

[14]  [Xin Du](https://dblp.org/pid/18/4833.html), [Kumiko Tanaka-Ishii](https://dblp.org/pid/42/2790.html): Stock Embeddings Acquired from News Articles and Price History, and an Application to Portfolio Optimization. [ACL 2020](https://dblp.org/db/conf/acl/acl2020.html#DuT20): 3353-3363

[13] Da Z, Engelberg J, Gao P (2011) In search of attention. J Finance 66:1461–1499

[14] Vlastiakis N, Markellos RN (2010) Information demand and stock market volatility. SSRN, eLibrary

[15] Bank M, Larch M, Peter G (2011) Google search volume and its influence on liquidity and returns of German stocks. Financ Mark Portf Manag 25:239

[16] Curme, Chester and Preis, Tobias and Stanley, H. Eugene and Moat, Helen Susannah, Quantifying the Semantics of Search Behavior Before Stock Market Moves (August 12, 2014). Proceedings of the National Academy of Sciences 111, 11600-11605; DOI:10.1073/pnas.1324054111 (2014) , Available at SSRN: <https://ssrn.com/abstract=2480274>

[17] Preis, Tobias and Moat, Helen Susannah and Stanley, H. Eugene, Quantifying Trading Behavior in Financial Markets Using Google Trends (April 25, 2013). Scientific Reports, Vol. 3, pp. 1684; DOI:10.1038/srep01684 (2013), Available at SSRN: <https://ssrn.com/abstract=2260189>

[18] Borup, Daniel and Rapach, David and Schütte, Erik Christian Montes, Now- and Backcasting Initial Claims with High-Dimensional Daily Internet Search-Volume Data (September 11, 2020). Available at SSRN: <https://ssrn.com/abstract=3690832>

[18] Mahalakshmi, G.; Sridevi, S. & Rajaram, S. A survey on forecasting of time series data 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), **2016**, 1-8