SOFIA

ANÁLISIS DE LA RELACIÓN ENTRE EL MERCADO DE INVERSIÓN, LAS REDES SOCIALES Y LAS NOTICIAS.

JOHN NICOLÁS ZAPATA ARIZA

UNIVERSIDAD DE LOS ANDES

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE SISTEMAS Y COMPUTACIÓN

BOGOTÁ D.C.

JULIO 2014

SOFIA

ANÁLISIS DE LA RELACIÓN ENTRE EL MERCADO DE INVERSIÓN, LAS REDES SOCIALES Y LAS NOTICIAS.

JOHN NICOLÁS ZAPATA ARIZA

Tesis de grado presentada como requisito para optar al título de Ingeniero de Sistemas y Computación

Director: Claudia Lucía Jiménez Guarín, Ph.D.

Profesora Asociada

UNIVERSIDAD DE LOS ANDES

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE SISTEMAS Y COMPUTACIÓN

BOGOTÁ D.C.

JULIO 2014

Tabla de Contenidos

Tabla de Figuras

Lista de Tablas

1 INTRODUCCIÓN

El mercado de mayor liquidez en el mundo, el de divisas o *Forex* (Foreign Exchange), movió en promedio USD 5.3 trillones diarios durante abril de 2013 [10]. Las decisiones de compra y venta en dicho mercado, se basan principalmente en dos modelos: fundamental y técnico. El primero se enfoca en monitorear variables macroeconómicas, políticas y sociales para predecir su comportamiento, mientras que el segundo hace lo propio, analizando exclusivamente el comportamiento actual y pasado de las monedas [9].

Una característica importante de los mercados Forex es la gran velocidad a la que los precios fluctúan, con una ventana muy corta para el análisis y toma de decisiones, por lo que cualquier análisis hecho debe estar disponible en un lapso muy corto de tiempo, lo más cerca posible de ser en tiempo real, para que pueda ser de utilidad a los inversores.

El análisis sintáctico de las publicaciones provenientes de agencias noticiosas y del contenido generado por las redes sociales sigue siendo en gran medida ignorado en este mercado, a pesar de que existen algunos ejemplos notables de las capacidades de estos medios en predecir comportamientos futuros en el mercado de valores; en nuestro mejor saber y entender, la aplicación de esta metodología en el mercado de divisas sigue siendo bastante reducida y limitada a soluciones que no entregan datos en tiempo real.

La falta de aprovechamiento de estas fuentes de información en línea, por parte de los operadores e inversionistas en dicho mercado, deriva en la toma de decisiones que ignoran una parte considerable de la información disponible, reduciendo la eficiencia que se podría alcanzar si se utiliza toda la información disponible. El correcto empleo de esta información podría resultar en mejores predicciones sobre su comportamiento y, en consecuencia, mayores rendimientos para los inversionistas.

Por las razones expuestas, este proyecto realiza un análisis sintáctico de redes sociales, noticias y resultados de motores de búsqueda, extrae la polaridad de cada una de estas publicaciones y, a partir de la información así recopilada, efectúa un estudio sobre el comportamiento del mercado de divisas, con el fin de examinar la posible correlación entre ambos. De esta manera se pretende enriquecer los indicadores financieros con un análisis del contexto externo que rodea al mercado de divisas y, por esta vía, otorgarle más herramientas a los inversionistas del mercado Forex para tomar decisiones mejor informados, lo que puede redundar en mejores retornos en sus operaciones.

Adicionalmente, debido a la velocidad de las transacciones y el poco margen de tiempo en el que esta información puede resultar útil para los inversionistas, se plantea desarrollar la solución utilizando herramientas de *Big Data* para llevar a cabo su procesamiento y análisis, con el propósito de obtener datos en tiempo real, que puedan ser usados inmediatamente como herramientas de decisión.

Finalmente, se lleva a cabo una evaluación de las implicaciones derivadas de las conclusiones obtenidas en este trabajo.

El proyecto tiene como objetivos entonces:

1.1 Objetivo general.

Proveer un indicador financiero (*indicador Sofia*: measurement of ***SO***cial and network news ***F***eeds for ***I***nvestment market ***A***nalysis) basado en el análisis sintáctico de fuentes diversas, que permita integrar y obtener en tiempo real información sobre el estado de ánimo del mercado, tal como se refleja en redes sociales, motores de búsqueda y fuentes profesionales de noticias. Al conjunto de tal solución se le llama SOFIA.

1.2 Objetivos específicos:

- Presentar el análisis sintáctico de una fracción del contenido generado en redes sociales, noticias y resultados de motores de búsqueda, extrayendo las polaridades -también llamadas sentimientos-, contenidas en estos datos

- Procesar en tiempo real la información sobre las diversas fuentes recopiladas, de tal manera que se puedan identificar las tendencias actuales, expresadas en las diferentes fuentes de información.

- Crear un modelo predictivo que pueda ser aplicado al comportamiento del mercado de divisas, a partir del análisis de polaridad de los datos recopilados en las diversas fuentes antes mencionadas, y que pueda ser utilizado por los inversionistas en dicho mercado.

- Crear una interfaz gráfica a través de la cual se puedan visualizar los resultados de los objetivos anteriores, en particular, los resultados generados por el modelo predictivo sobre el comportamiento del mercado Forex basado en el análisis de los datos y su comparación con la tendencia actual del mercado.

2 ANTECEDENTES Y CONTEXTO

En los últimos años ha surgido un creciente interés por el uso de fuentes de información en línea con el objetivo de predecir el comportamiento del mercado accionario, en particular el análisis se ha enfocado en el empleo de la información hallada en redes sociales como Twitter [1] [4], Facebook [2], o una combinación de las anteriores y resultados de búsqueda de Google [3]. Este último enfoque también se utiliza para el análisis de los mercados de *commodities*, accionarios y divisas, en particular, el del dólar frente al euro [5]. Sin embargo, en ninguna de estas propuestas es posible incluir el análisis de múltiples redes sociales y noticias para predecir cambios en el mercado accionario, tal como se desprende de la revisión de la bibliografía más destacada en este tema, según se explica a continuación:

2.1 Twitter mood predicts the stock market [1]

Este sistema permite hacer una predicción del comportamiento del mercado accionario, basado en un análisis de tweets. Su aplicación se reduce al Down Jones Industrial (DJI) y a Twitter, dejando de lado fuentes como noticias y otras redes sociales.

2.2 Can Facebook Predict Stock Market Activity? [2]

Este trabajo utiliza una medida muy interesante, llamada el índice nacional de felicidad bruto de Facebook (GNH por sus siglas en ingles), el cual sirve como un termómetro del estado de ánimo general de los usuarios en Facebook. La investigación muestra cómo el comportamiento del mercado accionario tiene una fuerte correlación con el GNH, moviéndose en tendencias similares. Sin embargo, al igual que el anterior proyecto, solo se enfoca en el mercado accionario y en una red social.

2.3 Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data [3]

Esta solución combina un enfoque global, condensando varios estudios hechos en el tema de análisis de sentimientos y tomando un enfoque holístico. Utiliza varias fuentes de información, Facebook, Twitter, Google Search, y encabezados de noticias comparan sus valores para predecir el índice de precios del mercado (DJI). Inclusive, hacen una comparación de los resultados obtenidos por este medio contra las tradicionales encuestas a inversionistas. A pesar de que es un punto de referencia importante, se enfoca exclusivamente en el mercado de acciones, dejando de lado el mercado de divisas.

2.4 Can social microblogging be used to forecast intraday exchange rates? [4]

Esta propuesta se enfoca en Twitter y en la relación del euro frente al dólar estadounidense. Aquí se realizan comparaciones y predicciones de mercado basados únicamente en *tuits* emitidos a una alta tasa de frecuencia diaria. Si bien emplea un modelo matemático novedoso, que permite dar una predicción con una alta tasa de efectividad, este método, al igual que los dos primeros, solo se enfoca en una fuente de información en línea, ignorando otras posibles fuentes de información.

2.5 Modeling Movements in Oil, Gold, Forex and Market Indices using Search Volume Index and Twitter Sentiments [5]

Este estudio es el más efectivo de los analizados, al alcanzar una precisión de 94% en sus resultados. Se enfoca en índices accionarios, Forex de euros contra dólares y oro. Toma como fuentes de información Twitter, y los resultados de búsqueda de Google. A pesar de que este trabajo combina varios elementos del proyecto a desarrollar, abarca superficialmente el mercado de divisas, pues toma una muestra semanal del mismo e ignora otras fuentes de información que pueden ser de gran utilidad para el cálculo del mismo..

2.6 Indicadores ISAAC: Siguiendo la actividad sectorial a partir de *Google Trends* [6]

En este trabajo se aplica una metodología de pronóstico de corto plazo de series económicas, con base en el sistema de estadísticas de búsquedas por internet de Google. Tal metodología faculta a los autores para construir indicadores sectoriales que permiten anticipar las tendencias del PIB en sendas ramas de la actividad nacional en el corto plazo. En tal sentido es un indicador de alcance reducido del comportamiento del mercado colombiano.

Tomando como base los estudios recién mencionados, la Tabla 1 ilustra su relación con la solución propuesta en el presente proyecto, a partir de los requerimientos que cada uno satisface y la comparación con los campos de estudio que abarcan:

Requerimientos vs. Soluciones existentes	[1] Twitter mood predicts the stock market	[2] Can Facebook Predict Stock Market Activity	[3] Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data	[4] Can social microblogging be used to forecast intraday exchange rates?	[5] Modeling Movements in Oil, Gold, Forex and Market Indices using Search Volume Index and Twitter Sentiments	[6] Siguiendo la actividad sectorial a partir de Google Trends	SOFIA
Recopilación y análisis de datos de redes sociales	X	X	X	X	X		X
Recopilación y análisis de datos de fuentes de noticias			X				x
Recopilación y análisis de datos de resultados de búsqueda			X		X	X	X
Procesamiento en línea de fuentes de datos							x
Generar tendencias con base en información recopilada	X	X	X	X	X	X	x
Realizar comparaciones sobre el mercado de divisas respecto a las tendencias generadas					X		x
Mostrar los resultados de estas comparaciones mediante una interfaz gráfica							x

Tabla 1. Requerimientos vs. Soluciones existentes

Con base en el examen de las propuestas existentes, contenidas en la literatura estudiada, se evidencia que a pesar de que se han realizado estudios previos sobre la aplicación de la información de fuentes diversas para proveer indicadores acerca del comportamiento de los mercados, aún existe un vasto campo de acción en el que una solución como la aquí propuesta tiene cabida, según como se explica en el siguiente capítulo.

3 ESTRATEGIA DE SOLUCIÓN

La solución SOFIA es una herramienta cuyo propósito es proveer el *indicador SOFIA*, un indicador financiero que permite dilucidar el estado de ánimo del mercado inversionista colombiano a través de múltiples fuentes de datos. Para tal fin se utilizan cuatro fuentes de datos (en inglés y en español), dos de las cuales son estructuradas; *feeds* RSS y Google Trends; y dos son semiestructuradas; Facebook y Twitter. De las cuales se obtienen tuits, *posts*, noticias y estadísticas de búsquedas.

Cada uno de los tres primeros tipos de datos recolectados; tuits, *posts* y noticias; es considerado como una "unidad de información", sobre la cual se realiza un análisis de polaridad, también llamado "análisis de sentimientos". El proceso de análisis es llevado a cabo con el uso de herramientas de análisis sintáctico de texto, acordes con el idioma de los datos, lo que revela la polaridad que puede ser asociada al texto analizado, es decir si es positiva o negativa, y con qué intensidad se genera dentro de un rango de medición entre -1 y 1. A partir de la información obtenida con este análisis, se procede a calcular el sentimiento promedio para cada fuente de datos, como el promedio del sentimiento de cada una de las unidades de información que la conforman.

En el caso de Google Trends, se recopilan las estadísticas de búsqueda para tres (3) términos definidos como relevantes para la aplicación, dada su relación con el mercado de divisas, a saber: "Dolar","Dolar hoy", "precio dólar" y "TRM". Para cada una de estas estadísticas se calcula el porcentaje de cambio del último valor disponible en el momento *t* con respecto al valor de la semana inmediatamente anterior. El valor obtenido constituye una tendencia de comportamiento del mercado, tal como se define en [6].

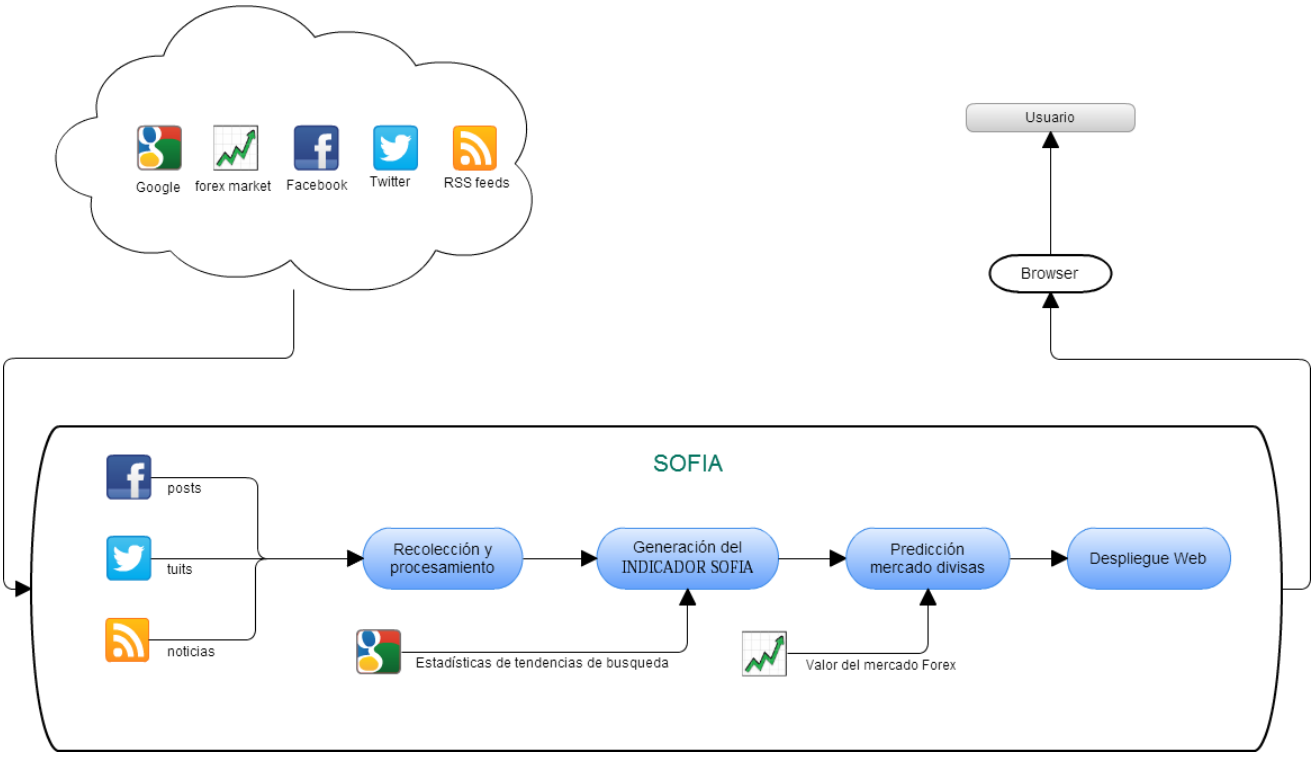

Figura 1. Diagrama de contexto

Posteriormente se calcula el *indicador Sofia*, usando los sentimientos promedio obtenidos previamente, y el valor calculado de Google Trends. Estos valores son pasados a través del modelo matemático definido en la sección 3.3 y que tiene como

resultado el valor del *indicador Sofia* y de indicadores que soportan el entendimiento del mismo, como *ValorTendencia* y *Distancia*. Esta información es mostrada luego mediante una página Web de una manera clara y descriptiva.

Es importante precisar que todo este proceso se realiza en tiempo real sobre la información recogida de las diferentes fuentes, tomando los datos de Twitter y Facebook como flujos constantes de información y los datos de Google y los *feed* RSS como consultas periódicas a las fuentes. Toda la información es procesada a medida que se recoge, utilizando para esto una instancia dedicada sobre una máquina independiente para cada fuente de información, de esta manera se reduce la posibilidad de una caída general del sistema, aumentando la fiabilidad del mismo y permitiendo que los indicadores producidos puedan ser consultados en todo momento.

Para resolver el problema planteado en la sección 1, se genera una estrategia consistente en 4 partes y descrita en la Figura 1: i) recolección de datos, ii) procesamiento de los datos, iii) Análisis de la relación entre indicadores e índices financieros y iv) visualización de la información. Tal diseño y solución se presenta en esta sección.

3.1 Recolección de datos

En esta parte del proceso se obtienen los datos de las diferentes fuentes de las que se nutre SOFIA. Todos los datos que se recolectan se almacenan desde la puesta en funcionamiento de la aplicación. Cabe recordar que las fuentes de los datos que se van a obtener son periódicos con versión en línea, Social Media, Google Trends y el mercado Forex. El flujo de datos de la solución se ilustra en la Figura 2.

Inicialmente se hace uso de *feeds* RSS por medio de los cuales se obtienen las noticias en el área económica y política de tres reconocidas publicaciones en versión electrónica en inglés (*The New York Times*[1]*, The Huffington post*[2]*, USA Today*[3]) y dos en español (Portafolio[4] y El Tiempo[5]), las cuales fueron elegidas por su amplia circulación dentro de los medios de información locales y por políticas, términos y condiciones de uso legales, los cuales permiten utilizar su información dentro del contexto de esta solución. En estos *feeds* se encuentran las noticias más recientemente publicadas por estos medios.

Dados los volúmenes de información que deben ser procesados y la gran complejidad inherente al mercado Forex, se restringe el estudio al caso del mercado del dólar estadounidense y el peso colombiano, cuyos valores son obtenidos mediante el servicio Web de Yahoo Finance[6]. Cabe precisar que se encontró que el acceso al valor de la tasa cambiaria en tiempo real suele estar restringido a servicios de pago de gran costo como *Bloomberg* y *Reuters*, por lo que después de evaluar un gran número de alternativas se optó por obtener los datos suministrados por Yahoo Finance, que los provee con una latencia cercana al tiempo real para el mercado de Estados Unidos, el cual, a pesar de tener valores diferentes a los de la tasa cambiaria, en general sigue patrones bastante similares a los del mercado colombiano .

[1] http://www.nytimes.com/
[2] www.huffingtonpost.com/
[3] http://www.usatoday.com/
[4] http://www.portafolio.co/
[5] http://www.eltiempo.com/
[6] http://finance.yahoo.com/

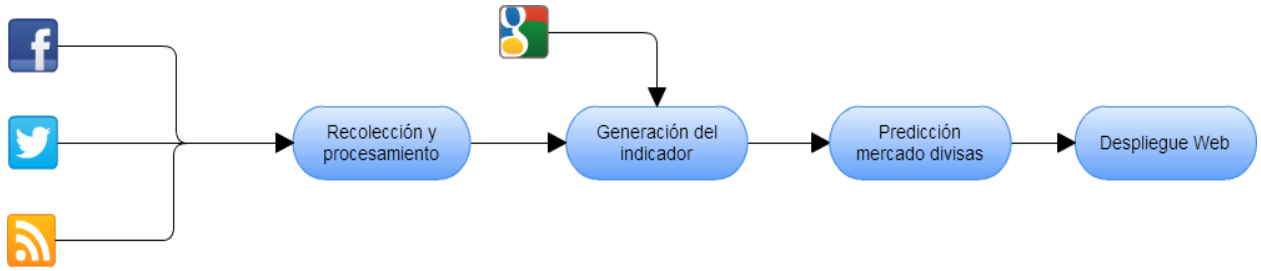

Figura 2. Diagrama de Flujo general de datos.

Paralelamente se utilizan herramientas a la medida para recopilar información de la red social más popular del mercado (*Facebook*[7]) y de la plataforma de *microblogging* con más usuarios (*Twitter*[8]). Para Facebook se utilizan las APIs públicas de la aplicación, obteniendo las publicaciones y los comentarios de estas publicaciones, de un número de páginas y grupos previamente definidos como significativos en cuanto al contenido que publican respecto al tema de este proyecto, mientras que en el caso de Twitter se hace uso del API para obtener los tuits del *firehose* (el *stream* público de tuits de Twitter), estos se filtran y se conservan solo aquellos enviados desde Colombia y Estados Unidos. Este filtrado se hace con el objetivo de obtener solamente los tuits relevantes a Estados Unidos y Colombia, los países de interés en el caso de estudio.

En este punto importa recordar que la metodología para la generación del *Indicador Sofia,* se construye sobre estudios previos (sección 2) en los cuales se concluye que el flujo total de post de Facebook y de tuits de twitter, los artículos noticiosos y las estadísticas de búsqueda de Google, tienen una fuerte correlación con el comportamiento de los mercados accionarios y Forex.

Adicionalmente se recopilan las estadísticas semanales generadas por *Google Trends*[9] sobre términos de búsqueda previamente definidos que cuentan con una alta relación con el mercado de divisas colombiano y estadounidense, las cuales son usadas debido a su utilidad como reflejo de la situación actual de la economía [6].

En general, en todo momento se tiene acceso a los datos más recientes de cada una de las fuentes: para el caso de Twitter y Facebook estos corresponden a los últimos tuits y post publicados, mientras que en el caso de los periódicos en línea se obtienen los últimos artículos publicados en sus páginas Web; en Google Trends se generan los datos de la última semana tal como son publicados por Google y en lo relacionado con el mercado Forex, se obtienen datos en tiempo real con el último valor disponible del mercado.

3.2 Procesamiento de los datos

A medida que se recopilan los datos de cada una de las fuentes, se realizan varias tareas, en primera instancia, se procesan con el objetivo de reducir el ruido en los datos estudiados[10] (Figura 3 y Figura 4), lo que se logra mediante la limpieza del texto de *hashtags*, menciones y vínculos (*links)* en el caso de los tuits, y de imágenes, vínculos y *tags* HTML en el caso de los *feed* RSS. Es importante reiterar que cada pieza de información es considerada como una unidad de información, por lo tanto, un artículo noticioso y un tuit son considerados cada uno como una unidad.

[7] https://www.facebook.com/
[8] https://twitter.com/
[9] http://www.google.es/trends/
[10] Debe entenderse "ruido" como datos irrelevantes para la tarea a realizar que pueden inducir a errores en los resultados obtenidos.

Figura 3. Ejemplo de un tuit con ruido y otro sin ruido.

Figura 4. Ejemplo de un *feed* RSS con ruido y otro sin ruido

Posterior al proceso de limpieza, los datos pasan al analizador de sentimientos de cada lenguaje, el cual obtiene la polaridad de la unidad de información y la almacena junto con el texto recogido. Tales analizadores dan un puntaje positivo o negativo a cada unidad, en función de las palabras que esta contiene, de tal manera que entre mayor número de palabras generalmente aceptadas como positivas tenga la unidad, mayor emoción positiva le será asociada por el analizador, a su vez, entre más palabras negativas tenga, mayor emoción negativa le será asociada. El valor final estará dado como una suma del promedio de ambas polaridades.

3.3 *Indicador Sofia*: Modelo Matemático

El modelo descrito a continuación permite calcular el *indicador Sofia*, para lograrlo se utilizan dos entradas:

1. A partir de las estadísticas obtenidas mediante las tendencias en resultados de búsqueda, para cada tema se realiza un cálculo del promedio de búsquedas en las últimas *n* semanas, y se calcula tambien en qué proporción varia el último dato recolectado con respecto a este promedio, el resultado obtenido es normalizado y posteriormente se promedia el resultado de los diferentes temas, obteniendo así una tendencia.

2. Se usa la polaridad calculada de las demás fuentes recolectadas (Twitter, Facebook y *feeds* RSS) para calcular la polaridad promedio de las mismas y finalmente generar un resultado combinado ponderado ($Indicador\ Sofia$).

Como resultado de este proceso se obtiene un número real en el rango [-1, 1], el cual indica el sentimiento positivo o negativo del mercado en el momento *t*. Un valor de *-0.1*, por ejemplo, indicara que en ese momento el mercado tiene un sentimiento ligeramente negativo, y por lo tanto existe una probabilidad de que el valor de mercado baje. En el otro extremo, un sentimiento de +0.8 indicara que el mercado se encuentra bastante optimista y hay una muy grande probabilidad de que el valor de mercado suba.

Se presenta a continuación una explicación de las variables usadas en este modelo:

- UI_{ij} [Unidad de Información]: Es el sentimiento calculado por el analizador de texto para la unidad de información *i* en la fuente *j.* Toma un valor en el rango [-1, 1].

- $\mathrm{CP_j}$ [Coeficiente Ponderado]: El peso relativo de la fuente de información *j* en el modelo.

- PG_{kt} [Promedio Google]: El porcentaje de aumento o decremento del valor de las búsquedas en Google con respecto a la última semana en el tiempo *t*, para un término *k* de búsqueda de relevancia para el modelo. Tiene como valores máximos [-1, 1].

- v [ventana de análisis]: Determina con que cantidad de tiempo de antelación respecto al momento actual, se toman los datos obtenidos de las diferentes fuentes para el cálculo del $indicador\ Sofia$. Este valor es expresado en minutos.

- n [semanas para el cálculo del indicador Promedio Google]: Determina cual va a ser el espacio de tiempo sobre el que se van a tener en cuenta los datos obtenidos de Google Trends con el propósito de calcular el valor promedio de los mismos.

- $IndicadorSofia_{\ t}$: Es el fin último del modelo. Un numero entre [-1, 1] que resume el sentimiento promedio ponderado de las fuentes de información sobre las que se calculan sentimientos recolectadas en los últimos *v* minutos respecto al tiempo *t*, y de la tendencia generada por el promedio de todos los datos del indicador PG_{kt} para el mismo tiempo *t*.

$$IndicadorSofia_t = \sum_{j \in p} \left[\frac{\sum_{\mathrm{i} \in z_{jt}} \mathrm{UI_{ij}}}{|z_{jt}|} * CP_j \right] + \sum_{k=0}^{n} \frac{PG_k}{|k|}$$

$z_{jt} = conjunto\ de\ las\ unidades\ de\ informacón\ en\ la\ fuente\ j\ dentro\ del\ periodo\ \ [(t-q), t]$

$p = conjunto\ de\ fuentes\ de\ información\ sobre\ las\ que\ se\ calculan\ sentimientos\ (Twitter, Facebook\ y\ feeds\ RSS)$

$k = conjunto\ de\ terminos\ de\ busqueda\ utilizados\ en\ el\ modelo$

3.4 *ValorTendencia*: modelo matemático

En esta sección se describe la construcción del indicador *ValorTendencia*, el cual representa una aplicación del $Indicador\ Sofia$ al mercado de divisas (Forex) de Colombia y Estados Unidos, específicamente a la tasa USD/COP.

- m [Mercado]: El mercado sobre el que se busca comparar la efectividad del $indicador\ Sofia$. Para este trabajo el mercado de interés es el mercado Forex del Dólar contra el peso (USD/COP).

- TVM_{mt} [Tasa de Variabilidad del Mercado]: Un porcentaje que es calculado dinámicamente como la variación típica de m en los últimos 7 días para el tiempo *t*.

- VAM_{mt} [Valor Actual de Mercado]: Se refiere al valor actual de m en el tiempo *t*.

- $ValorTendencia_{mt}$: Es la aplicación del resultado del modelo anterior al valor actual de *m* en el tiempo *t*, con el objetivo de dimensionar la precisión del $indicador\ SOFIA$. El valor resultante corresponde al posible valor de *m* en el tiempo *t+1*.

$$ValorTendencia_{m(t+1)} = VAM_m * (1 + TVM_{mt} * IndicadorSofia_t)$$

- $Distancia_{mt}$: Corresponde a la diferencia entre $ValorTendencia_{m(t)}$ y $\mathrm{VAM}_{m(t)}$. Este valor permite conocer la precisión de la predicción generada por el indicador $Valortendencia$ con respecto al valor real del mercado *m* en el tiempo *t*. Este indicador puede tomar cualquier valor real y en este trabajo se busca que su valor sea 0, dado que entre más pequeño sea este número, mayor efectividad tendría la predicción realizada.

3.5 Análisis de la relación entre indicadores e índices financieros

Esta parte de la solución se enfoca en convertir los datos recopilados en información, para lo cual se hace uso del modelo matemático definido para $el\ indicador\ SOFIA$, la parte más importante de la solución, y el indicador derivador *ValorTendencia*, una medida de la utilidad del indicador SOFIA basada en la obtención de un posible valor futuro de la tasa USD/COP usando como insumos el valor actual de la tasa, la variabilidad promedio de la misma, y el valor calculado del $indicador\ SOFIA$.

Además se desarrollan las siguientes medidas útiles para entender la relevancia del indicador: la precisión promedio de *ValorTendencia* con respecto al valor de la tasa real USD/COP, la variabilidad promedio del $indicador\ Sofia$, valor mínimo, máximo y promedio en las últimas 24 horas, y en los últimos 7 días.

3.6 Visualización de la información

Con el fin de hacer más fácil la comprensión del comportamiento de $indicador\ Sofia$, se despliega gráficamente, en tiempo real y en línea, el desempeño del indicador derivado llamado *ValorTendencia* y el del comportamiento del valor de la tasa del dólar contra el peso. Tal comparación permite dar una idea sobre la efectividad de la tasa en el tiempo. Además las medidas desarrolladas en el paso anterior también son desplegadas junto con el sentimiento promedio de cada una de las fuentes a las que se les calcula este valor. También es posible desplegar los datos que dieron origen a cada indicador.

3.7 Supuestos

En este trabajo se emplea cierto número de variables que, dado su amplio rango de valores posible, pueden afectar el resultado final de la aplicación. Con el objetivo de explicar los posibles resultados del proyecto, se tomaron decisiones respecto a cómo definir tales variables, según se explica en esta sección:

3.7.1 **Ventana de análisis *v***

Este valor determina cuál es el intervalo de tiempo que se va a usar para calcular el *indicador Sofia* para un momento de tiempo dado. Éste tiene una gran influencia sobre el comportamiento del modelo, dado que determina cuál es la frescura de los datos a usar en el cálculo de los indicadores. Para tal efecto se analizaron las implicaciones de la selección de diferentes valores de v, sin embargo su elección última se considera dados los flujos de información que se obtienen. Al contar con bastantes datos en un corto periodo de tiempo, se hace posible calcular el sentimiento del mercado para intervalos de tiempo breves, una característica altamente deseable y definida en el primer capítulo de este trabajo. Con el objetivo de no restringir la solución a un solo intervalo de tiempo, se tomaron valores de v de 10, 30, 60 y 120 minutos y para cada ventana se aplicó el modelo completo.

3.7.2 **Semanas para el cálculo del indicador Promedio Google**

Google Trends tiene un amplio poder predictivo sobre los cambios en la economía con un mínimo de rezago [5] [6]. En el trabajo realizado por el Ministerio de Hacienda y Crédito Publico Colombiano es posible observar como los cambios que se dan semana a semana en el volumen de búsquedas para términos claves para un sector, se corresponde ampliamente con el comportamiento económico del mismo. En tal sentido se afirma que el rezago de los datos proporcionados por este indicador es mínimo, por lo que las variaciones más recientes en el volumen de búsquedas respecto al valor inmediatamente anterior de las mismas proveen información valiosa sobre los cambios en la economía en este sector. En virtud de lo anterior, en este trabajo se define un horizonte de tiempo de una semana, en el que se filtran los datos recopilados y sobre estos se calcula el promedio del volumen de búsquedas para un tema en particular, con el objetivo de compararlo con el último dato disponible y verificar así el cambio porcentual del mismo.

3.7.3 **Peso de las diferentes fuentes en el modelo**

Las diferentes fuentes de datos tienen diferente influencia en el modelo de acuerdo al peso ponderado que se les asigne, para SOFIA se tomaron todas las fuentes dentro del modelo con igual peso.

3.8 Atributos de calidad

Atributo de calidad	Desempeño		
Recursos	ID	Descripción	Prioridad
Uso de una infraestructura distribuida.	D1	La solución debe entregar el valor del indicador en tiempo real, o con una demora máxima de 10 segundos respecto al tiempo del último valor Forex disponible, dado que su utilidad radica en el pequeño espacio de tiempo en el que la información puede ser usada para la toma de decisiones en el mercado de inversión.	Alta

Tabla 2. Atributo de calidad: Desempeño

Atributo de calidad	Resiliencia operacional		
Recursos	ID	Descripción	Prioridad
Manejo de fallas en código	R1	La aplicación debe poder continuar operando ante la caída o imposibilidad de acceso a una o varias fuentes de información.	Alta

Tabla 3. Atributo de calidad: Resiliencia Operacional

Atributo de calidad	Confiabilidad (calidad de la información)		
Recursos	ID	Descripción	Prioridad
Estructura de componentes modulares	C1	El indicador resultado del proceso realizado por la aplicación (*indicador Sofia*), debe tener la mayor fiabilidad posible, entendiéndose ésta como que debe estar sustentado siempre en datos fidedignos a las fuentes de donde fueron recolectados.	Alta

Tabla 4. Atributo de calidad: Confiabilidad

Atributo de calidad	Modificabilidad		
Recursos	ID	Descripción	Prioridad
Estructura de componentes modulares	M1	La aplicación debe permitir la adición de nuevas fuentes de noticias para ser monitoreadas, así como cambios en el modelo matemático usado para generar el *indicador SOFIA*.	Media

Tabla 5. Atributo de calidad: Modificabilidad.

3.9 Casos de uso

La Figura 5 muestra el principal caso de uso de la aplicación, que es la manera como el usuario final interactúa con la solución propuesta:

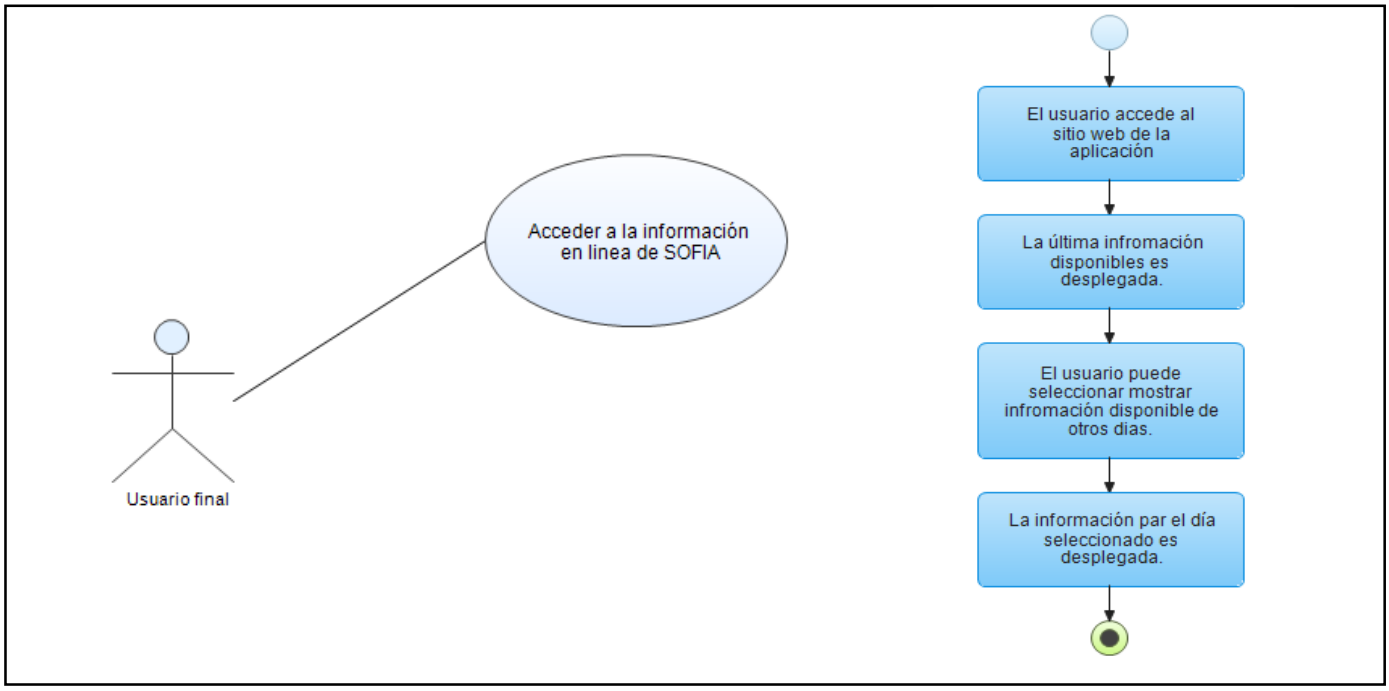

Figura 5. Diagrama de caso de uso y flujo de eventos del caso de uso.

3.10 Árbol de utilidad

En la Figura 6 se muestra el árbol de utilidad para SOFIA, considerando los atributos de calidad definidos en el capítulo 3.8.

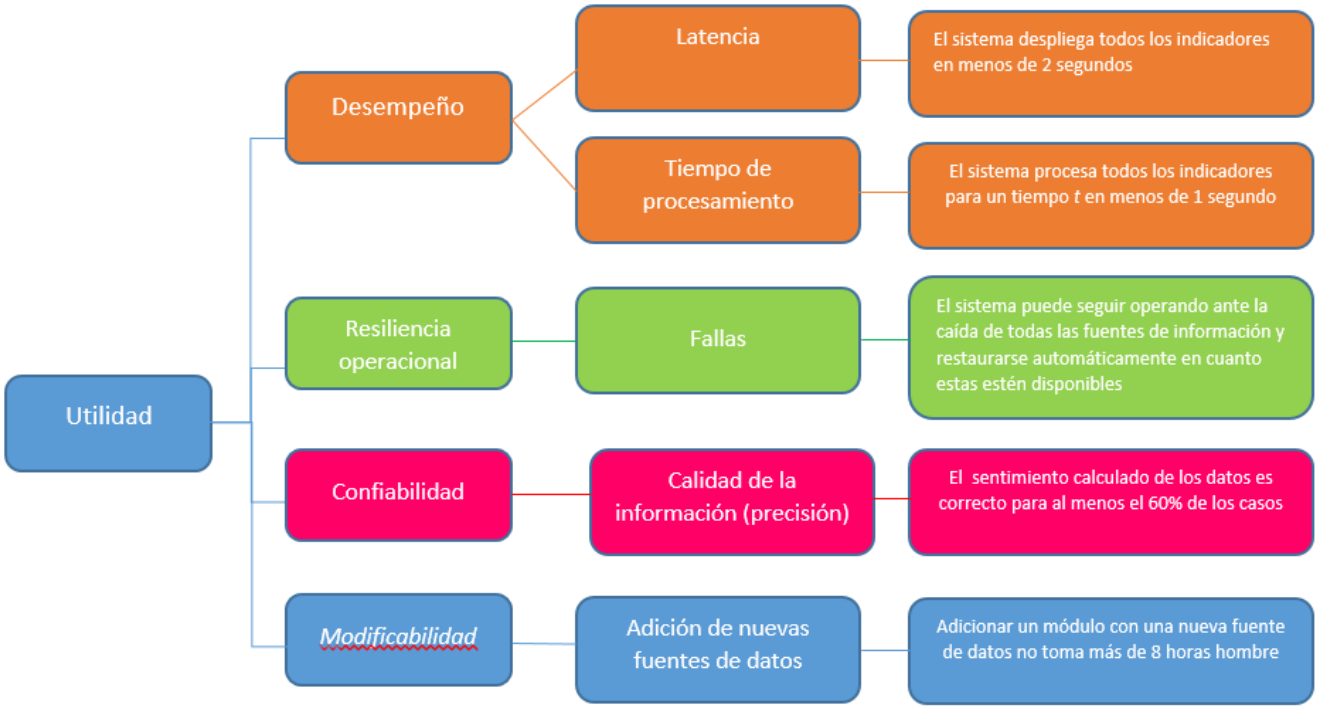

Figura 6. Árbol de utilidad.

4 DESCRIPCIÓN DE LA SOLUCIÓN

Este capítulo presenta la arquitectura de la solución propuesta, enfocándose en los diferentes componentes que forman la aplicación y detallando las relaciones entre ellos.

4.1 Diagrama de despliegue

En este diagrama es posible ver como los diferentes procesos que la aplicación realiza son llevados a cabo en máquinas independientes, cada una responsable de la ejecución de uno o dos procesos. Tal decisión se fundamenta en el requerimiento no funcional de resiliencia, ya que al funcionar en sistemas independientes si alguno de estos falla, los demás seguirán operando y recogiendo información. Contrario a lo que sucedería si todos funcionaran en una sola máquina, en cuyo caso la caída del sistema ocasionaría la caída de todos los servicios de la aplicación. Por otra parte, dado que todas las máquinas utilizan un solo repositorio central de información NoSQL, se permite a cada una utilizar sus recursos completamente para la tarea encomendada, aumentando la velocidad de procesamiento de la aplicación a costa de un mínimo de latencia causado por el uso de la red para transferir los datos entre las diferentes máquinas y servidor de datos, el cual al ser NoSQL puede manejar una mayor cantidad de consultas por segundo, en comparación con el tradicional modelo relacional.

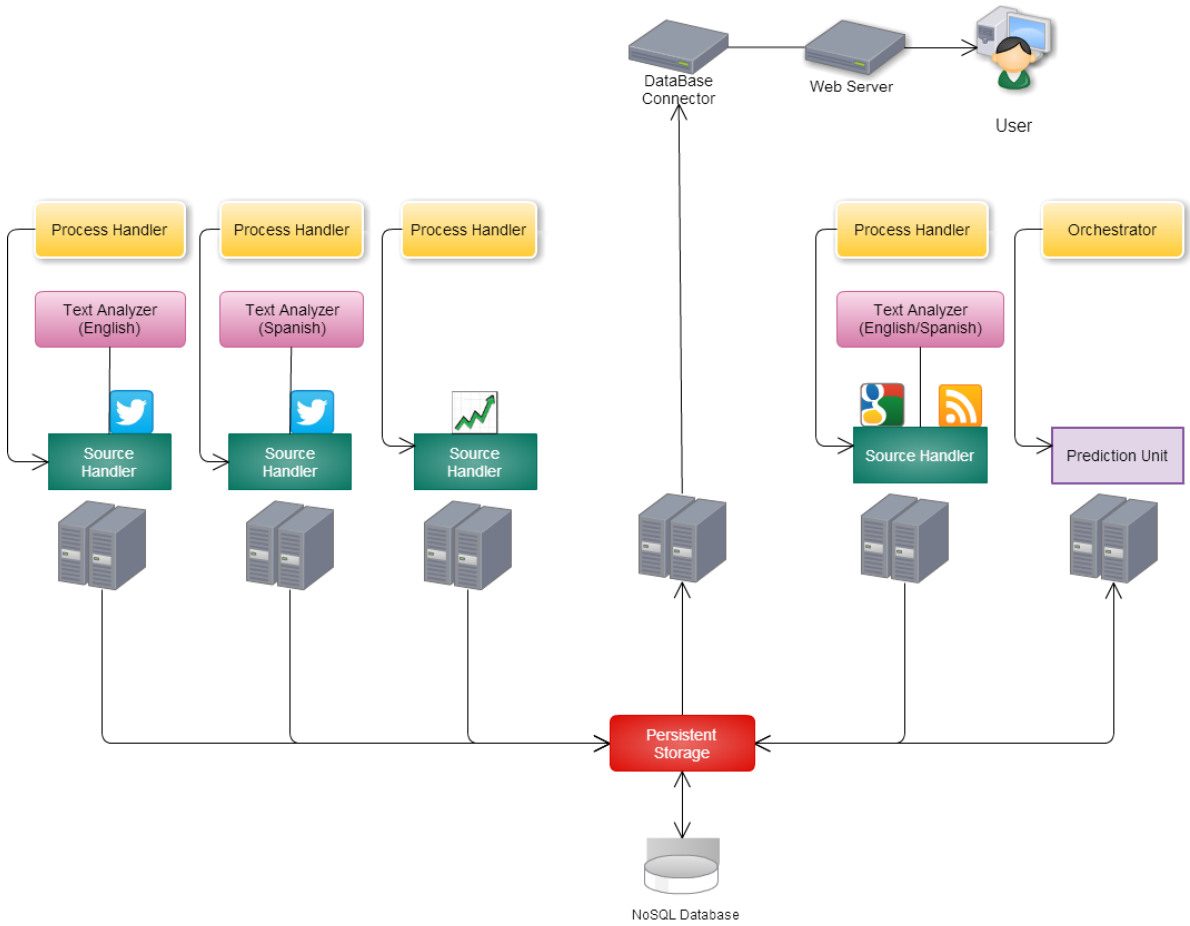

Figura 7. Diagrama de despliegue

4.2 *Arquitectura* Global

En concordancia con el atributo de calidad de desempeño, se presenta una arquitectura distribuida en varios componentes independientes que actúan sobre una misma base de datos, la cual actúa como un repositorio central desde y hacia donde la información fluye. Todo el proceso de generación del *indicador Sofía* es coordinado por un único componente, que sirve como coordinador de las tareas específicas a realizar en cada máquina sobre la cual la solución es desplegada.

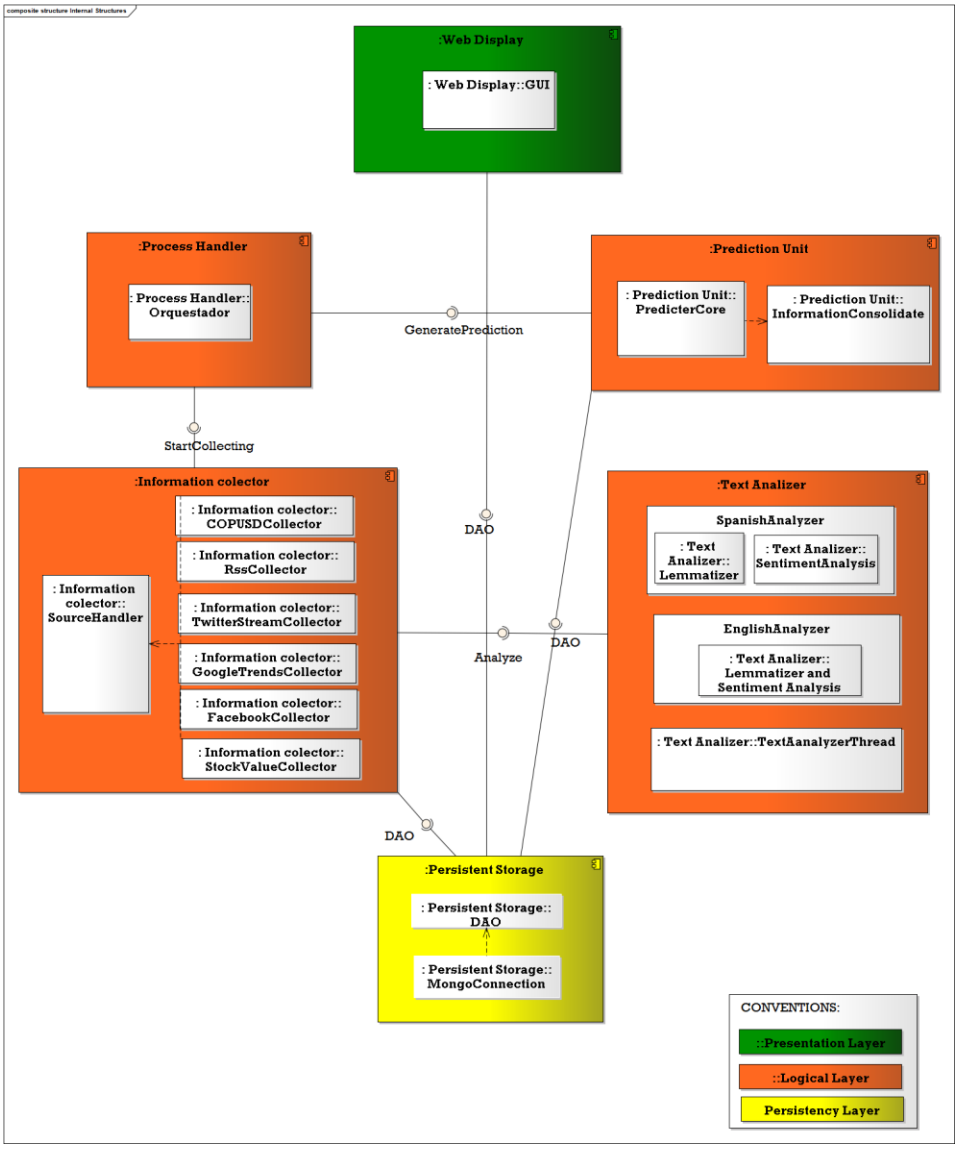

Figura 8. Modelo general de componentes

La estructura global de la aplicación creada para resolver el problema se divide en un modelo típico de 3 capas (3-*tier*), usado ampliamente por su flexibilidad y rendimiento. Tal estructura comprende una capa de persistencia, otra de lógica de negocio y finalmente una de visualización (Figura 8).

4.3 Modelo de componentes

La solución está diseñada de tal manera que depende de varios componentes que solucionan una parte del problema, cuya función se explica a continuación.

4.3.1 **Information Collector**

Este componente realiza conexiones a todas las fuentes de información que han sido definidas como importantes para la solución, dentro de éste existe un conector para cada tipo de información que va a ser descargada. Estos conectores, a su vez, hacen llamados al componente `Persistent Storage` para almacenar la información de cada fuente, clasificada según su origen e idioma. En esta solución se definen como únicos idiomas válidos para las fuentes de datos el inglés y el español. Estos procesos de recolección y procesamiento son realizados por máquinas independientes, cada una destinada a un *stream* particular con su propio analizador de sentimientos.

De otra parte, este componente también recoge los datos correspondientes al valor de la tasa de cambio entre el dólar y el peso, para lo cual se hace uso de consultas REST a Yahoo Finance, las cuales retornan el valor actual de la tasa.

4.3.2 **Persistent Storage**

Este componente establece una conexión directa con el repositorio de información, una base de datos NoSQL. Es usado a través de todos los componentes de la aplicación y es accedido a través de un objeto con una única instancia; puede ser consultado por todos los componentes tanto para almacenar como para consultar información.

4.3.3 **Text Analyzer**

La función principal de `Text Analyzer` es procesar la información recogida por el `Information Collector` y devolver un sentimiento asociado para cada unidad de información.

Dado que la aplicación maneja fuentes de información en inglés y español, un proceso de clasificación con su respectivo modelo analítico es usado para cada idioma por separado, dependiendo del origen de los datos.

4.3.4 **Prediction Unit**

El componente principal de la aplicación, es en este punto donde se genera el valor del *indicador Sofia*. Todo el contenido generado por el componente previamente descrito es agregado y procesado a través del modelo matemático descrito previamente en la sección 3.3. Como consecuencia, y basado en todas las fuentes de información consideradas, se genera el *indicador Sofia* y una aplicación del mismo, *ValorTendencia*, al valor actual del mercado de divisas, además de la correlación entre estos indicadores, *Distancia*. Estos valores son almacenados por el `Persistent Storage` para su posterior uso en la interfaz Web.

4.3.5 **Web Server**

Con el objetivo de mostrar y facilitar la interpretación de estos datos por parte del usuario final, este componente produce una representación visual de la información generada, en la forma de una gráfica del cambio del valor del dólar contra el peso y una lista de indicadores relevantes para el periodo de tiempo mostrado, el cual es accedido directamente a través de un navegador Web y es el único punto en donde el usuario tiene contacto con la aplicación.

Cabe anotar que esta parte del proceso es realizada por procesos independientes entre sí, con el fin de proveer de mayor desempeño a la aplicación. En tal sentido, un proceso independiente se encarga de leer los datos desde la base de datos y generar el *indicador Sofia* mediante el modelo matemático. Otro proceso se encarga del despliegue Web.

4.3.6 Process Handler

Finalmente, este componente se encarga de orquestar a los demás componentes de la aplicación, pues inicia y supervisa el correcto proceder de todas las tareas necesarias para producir y almacenar el valor del indicador.

4.3.7 Modelo de datos

Como se explicó previamente, se decidió utilizar un repositorio NoSQL orientado a documentos para almacenar los datos de SOFIA. En este modelo de datos se utiliza una colección para cada fuente de datos, y en el caso de los datos multilingües, se utiliza una colecciónseparada para cada idioma de cada fuente consultada. Con esta decisión se busca facilitar la creación de consultas a la base de datos, simplificando el proceso de filtrado de datos en el momento de construir los indicadores financieros.

Adicionalmente, se utilizan colecciones separadas para almacenar los valores de los indicadores financieros producidos, con el objeto de facilitar su consulta desde el componente Web. Dado que cada colección de documentos es independiente de las demás y las otras son agregadas, no hay relaciones entre las colecciones.

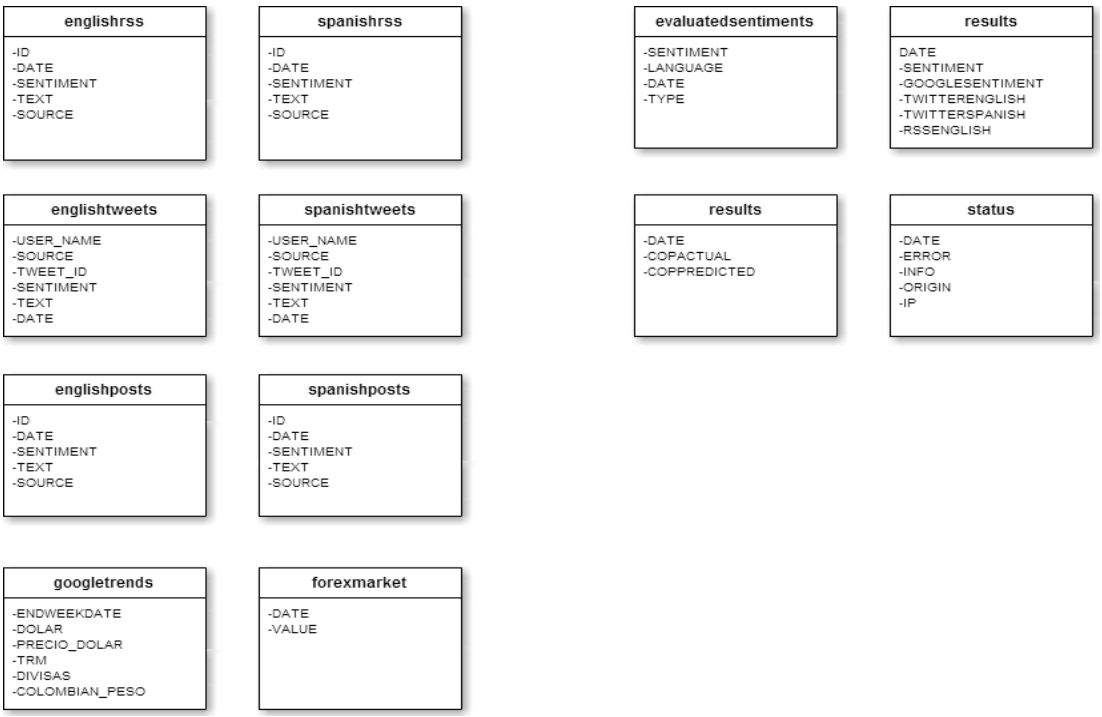

Figura 9. Modelo de datos

4.4 Diagrama de clases

A continuación se muestra el diagrama de clases. Es sabido que la densidad de la información representada en la Figura 9 hace difícil su adecuada visualización en la versión impresa de este documento, sin embargo es posible visualizarlo de manera mucho más clara en su versión digital.

Figura 10. Diagrama de clases

5 IMPLEMENTACIÓN

En este capítulo se presentan consideraciones generales respecto al desarrollo y la implementación de la solución y se aborda específicamente el trabajo realizado para entrenar el clasificador de textos en español.

5.1 Entorno de desarrollo

El desarrollo de la aplicación se hizo en su mayoría utilizando Eclipse Kepler SR1 sobre Windows 8.1 y Java 1.7u55, así mismo se utilizó Mongo 2.4.1 como la versión del repositorio de pruebas y MongoVue 1.6.9 para visualizar los resultados del mismo repositorio.

5.2 Implementación

5.2.1 Entorno de implementación

Para la implementación del proyecto se utilizaron máquinas virtuales Linux CentOS 6.5, a las cuales se conectaba a través de SSH por medio de PuTTY 0.63. Estas máquinas virtuales estaban apoyadas sobre servidores con las siguientes especificaciones:

- CPU: Intel(R) Xeon(R) CPU E7- 2850 @ 2.00GHz.
- RAM: 4 Gigabytes.
- Disco: Disco Sata de 150 Gigabytes.

5.2.2 Ubicación de los servicios:

Sobre cada máquina virtual se asignó una tarea en particular para cada instancia de la aplicación:

- 10.0.2.40: Repositorio de información y despliegue Web
- 10.0.2.41: Generar el indicador Sofía.
- 10.0.2.42: Recolectar Tweets en Español.
- 10.0.2.43: Recolectar Tweets Ingles.
- 10.0.2.44: Recolectar RSS y Google.
- 10.0.2.45: Recolectar valor actual del mercado de divisas y accionario.

5.3 Versiones de las librerías y herramientas usadas en el proyecto

Librería y herramientas	Versión	Uso en la solución
Eclipse	Kepler SR1	IDE de desarrollo
Mongo	2.4.9	Repositorio de información
Stanford Core NLP	3.3.1	Analizador de sentimientos para los textos en inglés
Lingpipe	4.1	Entrenamiento de modelo y clasificación de los textos en español
Twitter4j	4.01	Descarga de datos de Twitter
Java	1.7.55	*Framework* general de trabajo
Mongo VUE	1.6.9	Herramienta de manejo de Mongo
Node JS	0.10.26	Servidor Web
Mongoose	3.8.8	Servidor de acceso Web a Mongodb
Putty	0.63	Herramienta de comunicación con los servidores de despliegue
Centos OS	6.5 x64	Sistema operativo de despliegue

Tabla 6. versiones de las herramientas usadas en el proyecto

5.4 Flujo de información

Las fuentes de información seleccionadas para SOFIA proveen grandes flujos de datos. Recolectar, almacenar y procesar éstos volúmenes requiere considerar la eficiencia de los procesos que realizan estas tareas. Con el fin de dar una idea del volumen de información manejada se presenta la Tabla 7: Volumen de flujos de datos promedio.

	Unidades de información por minuto	Unidades de información por día	Peso típico de la unidad	Volumen acumulado al mes
Tweets de Colombia	2280	3.283.200	350 bytes	34.473,6 megabytes
Tweets de Estados Unidos	2700	3.078.000	350 bytes	29.087,1 megabytes
Noticias en español	23	33.120	0.9 kilobytes	894,24 megabytes
Noticias en Inglés	35	50.400	1.2 kilobytes	1.814,4 megabytes
Valores del mercado Forex	0,3	432	112 bytes	1,064 megabytes

Tabla 7. Volumen de los flujos de datos promedio.

5.5 Implementación de componentes

5.5.1 Implementación de Information Collector

Este componente requirió del uso de *threads* para cada uno de los subprocesos recolectores de información, iniciados de acuerdo con los parámetros pasados desde el componente `Process Handler` y funcionan independientemente entre ellos. La implementación de los mismos se realizó de tal manera que pudieran seguir funcionando aún en caso de errores graves, mediante el manejo de excepciones.

```
if(Language.contains("EN"))
{
    positiveWord = "happy";
    negativeWord = "sad";
}
else
{
    positiveWord = "feliz";
    negativeWord = "triste";
}

texto = response;

response = response.replaceAll("(:\\)|\\(:|:D|♥|❤|☺|❣|♞|😊|😃|😄|😁|❣|😍|<3|:-\\))", " " + positiveWord + " ");

response = response.replaceAll("(:s|:\\(|\\):|☹|😞|😔|😟|😕|:-\\(|:'\\()", " " + negativeWord + " ");
```

Figura 11. Porción de código en Java que reemplaza emoticones por palabras relevantes

Los procesos que recopilan texto producido por humanos llaman a la clase `Util`, la cual provee los servicios de limpieza de texto para tuits y noticias, esto quiere decir que remueve texto que no es entendible por los analizadores sintácticos, como por ejemplo las etiquetas HTML, y además realiza el reemplazo de emoticones por palabras positivas o negativas, dependiendo del emoticón encontrado y del idioma en el que se esté trabajando.

5.5.2 Persistent Storage

Para la conexión con la base de datos elegida, MongoDB, se decidió utilizar un patrón Singleton de tal manera que en todo momento solo hubiera una conexión concurrente a la base de datos. Dado que varios de los procesos de la aplicación corren en su mayoría en máquinas independientes a aquella en donde se encuentra el servidor de datos, se buscó reducir el número de conexiones abiertas a la base de datos para mejorar el desempeño y tiempo de respuesta de la misma[11].

5.5.3 Text Analyzer

La función principal de `Text Analyzer` es procesar la información recogida por el `Information Collector` y devolver un sentimiento asociado para cada unidad de información. Para realizar el análisis de texto requerido se utilizaron dos conjuntos de herramientas en función del idioma. Por una parte para el análisis de textos en inglés se utilizó la herramienta creada por la universidad de Stanford, *The Stanford NLP 3.3.1*. Esta herramienta realiza el proceso de lemmatización y análisis sintáctico de textos en inglés, que termina en la obtención de una polaridad asociada a la unidad de información estudiada.

En el caso del texto en español primero se implementó el lemmatizador de texto de Anna 3.3, posteriormente se diseñó un clasificador basado en fuentes en línea. Primero se implementaron las APIs de 5 proveedores de análisis de sentimientos en línea: AIAIOO, ApiCulture, Bipolarity, Textalytics y 140, las cuales reciben consultas por medio de peticiones REST y responden con un sentimiento asociado para cada texto. De esta manera se etiquetó un corpus de 10.000 tuits que había sido recogido previamente. Tomando como entrada este corpus y utilizando Lingpipe se entrenó un modelo basado en NaiveBayes.

El flujo de información para el caso de la recolección de tweets es bastante grande, los tweets colombianos llegan a un promedio de 38 tweets por segundo, mientras que los tweets estadounidenses llegan a un promedio de 45 tweets por segundo. Dado este flujo de información se creaban grandes cuellos de botella para el `Information Collector`, por lo que se decidió implementar dos threads que corren en simultaneo en la misma instancia en donde se recolectan los tweets, y cuyo único propósito es realizar el análisis de sentimientos de los tweets recogidos.

Finalmente dado que las herramientas de cada idioma son independientes entre sí, el proceso de análisis con su respectivo modelo analítico es realizado para cada idioma por separado, dependiendo del origen de los datos.

[11] http://docs.mongodb.org/manual/administration/monitoring/

5.5.4 **Prediction Unit**

La implementación de este componente presentó retos de eficiencia, dado que su funcionamiento básico es hacer consultas a Mongo especificando parámetros de tiempo para obtener todos los datos de un intervalo de tiempo definido *v*, luego con base en estos datos calcular el sentimiento promedio para cada fuente y con estos valores generar el *indicador Sofia* y *ValorTendencia* mediante el modelo matemático definido en la sección tres del presente documento.

La principal dificultad radicó en la necesidad de hacer consultas de agregación sobre mongo con base en intervalos de tiempo de 10, 30, 60 y 120 minutos. Éste último intervalo requiere una gran cantidad de procesamiento por parte de la base de datos, ya que una consulta típica debe procesar alrededor de medio millón de tweets y 114 noticias completas.

5.5.5 **Web Server**

En este componente se utilizó NodeJs como servidor HTTP, el cual a través de Mongoose hace consultas a la base de datos Mongo usada como repositorio principal de información, la información obtenida es entonces mostrada en una interfaz Web accesible a través de cualquier navegador moderno. Para las gráficas desplegadas se utilizaron las librerías D3.js y Amchart.

5.5.6 **Process Handler**

El componente encargado de determinar qué procesos se inician y del ciclo de vida general de la aplicación. `Process Handler` toma como entrada un archivo en el cual se especifican los parámetros de cuáles recolectores de información van a ser iniciados, la URL del repositorio Mongo de la aplicación, las direcciones de los archivos con los *feeds* URL a descargar, los datos de GoogleTrends, y las compañías a seguir en bolsa. También se especifica la ruta del archivo con los parámetros configurables del modelo, además de las credenciales de acceso a Twitter y un parámetro utilizado en las pruebas de la aplicación.

```
#path to:
#1. RSS feeds URLs 2. COmpanies to be observed within the stock market 3. Google Trends File (.csv)
#4. URL of the MongoDb to be used (if empty localhost will be used),
#5. list with feeds this instance will collect,
# separated by semicolon: GoogleTrends;RSS;TwitterEnglish;TwitterSpanish;StockTracker
#6. boolean defining wheter or not this instance will generate predictions
#7. String with the twitter credentials separated by ; to be used in the tweets collection:
# consumerkey;consumersecret;accesstoken;accesstokensecret
#8. boolean defining wheter this is a test run and the app should run regardless of the current hour
# Note that for a single instance application all boolean parameters should be set to true
#true;true;true;true;true    false;false;false;false;false
#
/root/datos/fuentes.txt
/root/datos/companies.txt
/root/datos/GoogleTrends/report.csv
10.0.2.40
false;false;false;false;false
false
S1Z89iE2hO92DIGsf529A;XwJ3djSAkt4oqruAFqBemnQKvliwsy88sYPekt0NEU;2356993034-SWBEsgckb2RXbNTPmRihczcXp
false
```

Figura 12. Ejemplo de archivo de configuración de la aplicación.

Con estos datos el componente primero procede a verificar si actualmente la bolsa de valores de Colombia esta en horario de operación y, en caso afirmativo, llama al componente `Information Collector` con los parámetros adecuados para que este ejecute los procesos específicos a esta instancia. Posteriormente, si es el caso, se invoca al componente encargado de generar el *indicador Sofia*.

5.6 Alcance logrado

En el desarrollo de la solución se presentaron obstáculos que en mayor o menor medida contribuyeron a que no se alcanzaran a cumplir todos los objetivos propuestos inicialmente. En efecto, no se realizó la integración de la aplicación con Facebook para obtener y procesar los post públicos de esta red social, lo cual obedeció, en gran medida, a dos factores: de una parte, la dificultad encontrada al trabajar con el API pública de esta red social y de otra, el tiempo de desarrollo disponible.

La estructura de la solución fue desarrollada en su totalidad, con excepción de la sección previamente mencionada. A continuación se presentan la Figura 13 y la Figura 14, las cuales muestran la interfaz web de la solución realizada

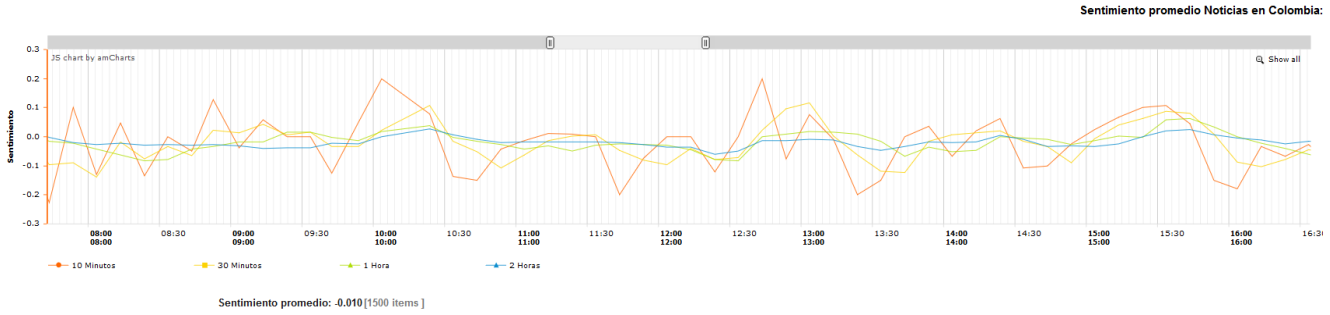

Figura 13. Sentimiento promedio de las noticias en Colombia para el día 28 de mayo

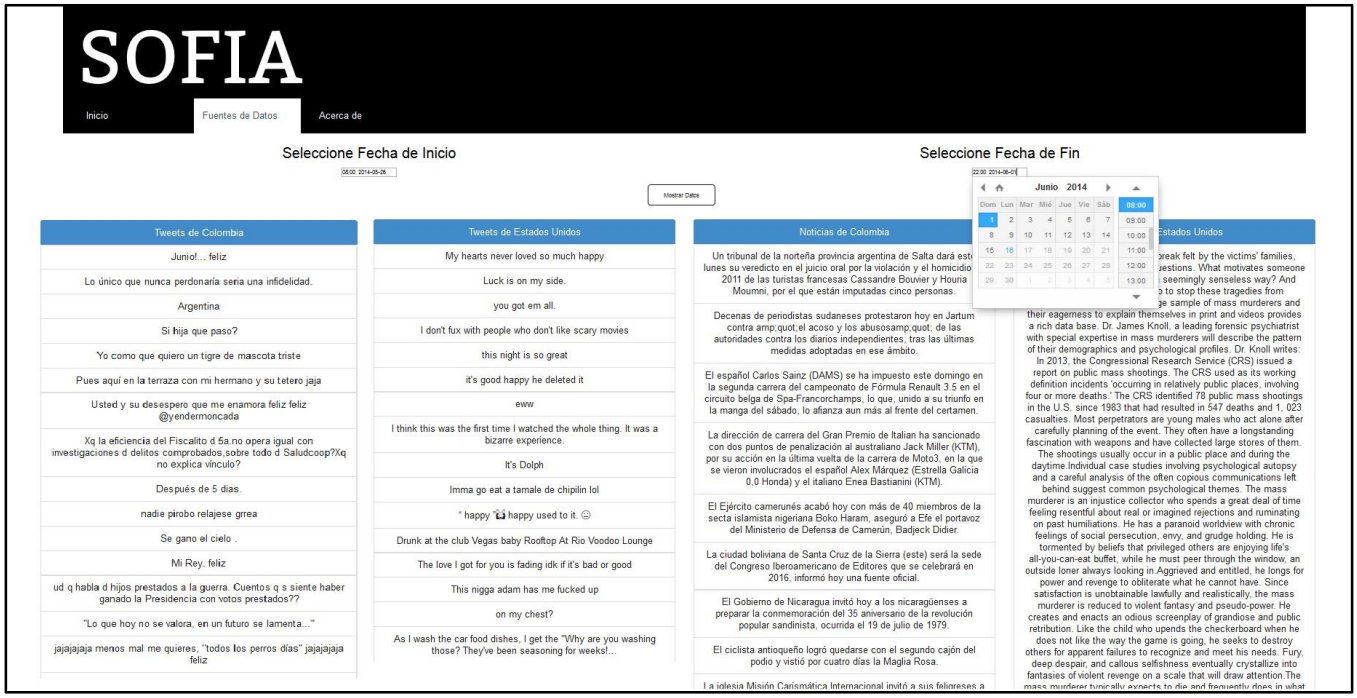

Figura 14. Vista de consulta de los datos por la solución

6 PRUEBAS Y RESULTADOS

6.1 Plan de pruebas

En este capítulo se presentan las pruebas realizadas sobre SOFIA. Estas pruebas se diseñaron con el objetivo de validar si los requerimientos más relevantes para la solución son efectivamente cumplidos.

Prueba 1: Funcionamiento de la aplicación después de la caída de todas las fuentes de datos	
Tipo	Funcional
Objetivo	Determinar si la solución, SOFIA, puede soportar la caída de las fuentes de datos y seguir operando a la espera de que estas vuelvan a estar disponibles.
Escenario	SOFIA está operando
Pasos	Cortar la conexión a internet de la máquina en donde está operando alguna de las instancias de la aplicación.
Resultados esperados	La aplicación sigue funcionando, se imprimen mensajes de error a archivo y a la base de datos.
Resultados Obtenidos	EXITOSA
Observaciones	Esta prueba se realiza sobre una máquina local de desarrollo, no sobre una infraestructura distribuida, sin embargo los resultados sobre este último ambiente deberían ser similares a los obtenidos, tomando en consideración que la red local SI debe seguir en funcionamiento.

Tabla 8. Prueba 1

Prueba 2: Tiempo de procesamiento del cálculo del *indicador Sofia* y todos los otros indicadores generados.	
Tipo	Prueba de rendimiento
Objetivo	Determinar cuánto tiempo le toma al sistema procesar los datos para generar el *indicador Sofia*
Escenario	SOFIA está operando normalmente.
Pasos	Se inicia la operación del sistema
Resultados esperados	La aplicación debería tomar menos de 10 segundos en procesar todos los datos recogidos y generar los indicadores de interés.
Resultados Obtenidos	EXITOSA
Observaciones	Esta prueba se realizó sobre una máquina local de desarrollo y sobre uno de los servidores de ejecución. En ambos casos el tiempo fue menor de 10 segundos .

Tabla 9. Prueba 2

Prueba 3: Confiabilidad de los datos obtenidos en la solución	
Tipo	Funcional
Objetivo	Verificar la efectividad de las predicciones generadas por la solución.
Escenario	SOFIA está operando
Pasos	Ingresar a la interfaz Web de la aplicación.
Resultados esperados	Los datos sobre la efectividad de la predicción son desplegados en pantalla, los mismos deben mostrar una efectividad superior al 60%.
Resultados Obtenidos	
Observaciones	.

Tabla 10. Prueba 3

6.2 Análisis de resultados

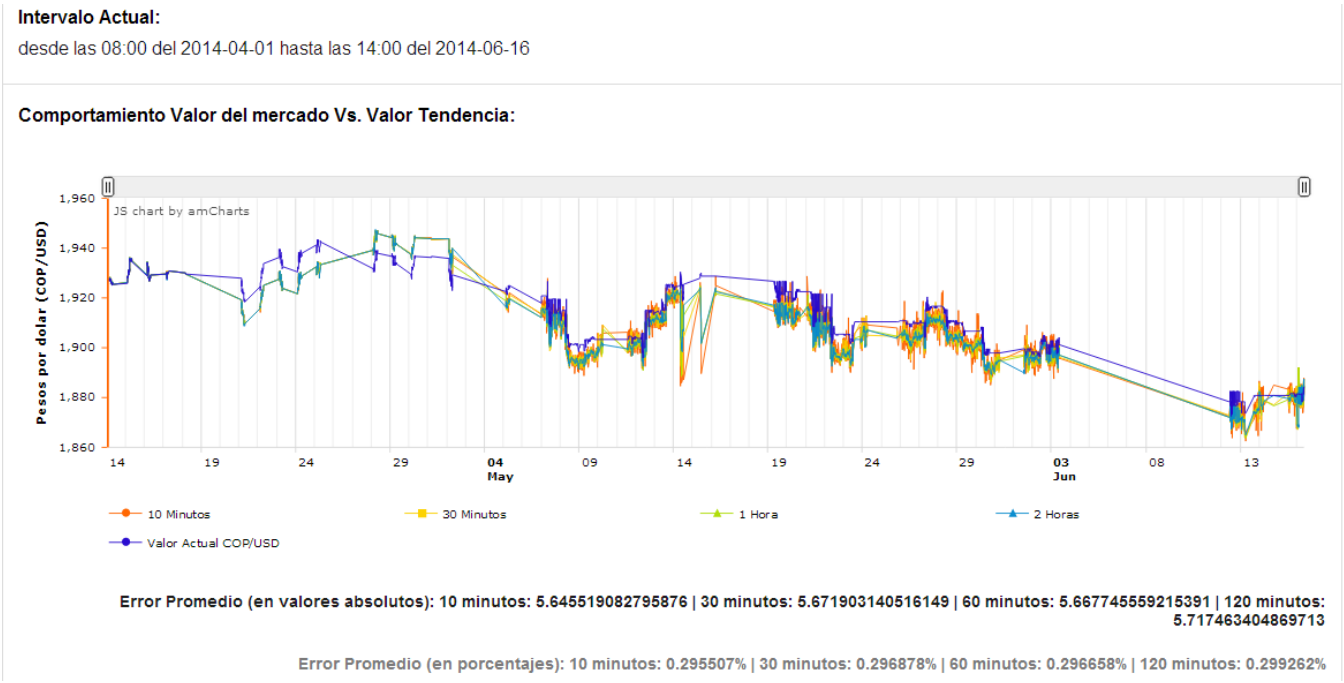

Figura 15. Ejemplo de comparación entre el *indicador Sofia* calculado con un intervalo previo de 10, 30, 60 y 120 minutos vs. el valor real de la tasa cambiaria para el periodo entre el 1 de abril y el 16 de junio

El objetivo general de la solución se alcanzó, se generó el *indicador Sofia* basado en el análisis sintáctico de fuentes diversas, obteniendo en tiempo real una medida del sentimiento del mercado inversionista basado en *social media*, periódicos en línea y resultados de búsquedas en Google.

Sin embargo los datos no son concluyentes en cuanto a la efectividad del indicador generado, en tal sentido la decisión de hacer análisis con intervalos de tiempo previos de diferentes medidas, permitió obtener un entendimiento general de cómo estos afectan el comportamiento del modelo, sin embargo existen sendas variables que afectan dramáticamente el .comportamiento del modelo, por lo que un estudio de cada una de ellas es requerida para poder tener un modelo más preciso.

El error promedio de las predicciones, definido como el promedio de la suma de las diferencias absolutas entre el valor de la predicción y el valor real de la divisa, para los últimos 3 meses fue de 0.2992% en el caso más alto, y de 0.2955% en el caso más bajo, dejando al intervalo más fiable a 10 minutos y como el menos fiable a 120 minutos. Estos resultados deben ser interpretados con suma precaución, dado que el tope máximo de error para estos resultados corresponde a la variación máxima de la tasa del dólar, la cual fue tomada como un 5%, y estos resultados no se acercan ni remotamente a éste valor. En tal sentido es evidente mediante las gráficas que los valores calculados para los diferentes sentimientos tienden a estar bastante cercanos al 0, tal es el caso del *indicador Sofia* para el mismo periodo de tiempo, el cual se situó en un promedio de -0.06, un valor lejos de los extremos posibles a los que pudo haber llegado, -1 y 1. Así mismo los valores de los sentimientos promedio de Twitter en español y en inglés se situaron en -0.046 y 0.006 respectivamente, mientras que las noticias se situaron en -0.01 para el caso de Colombia y -0.239.

Lo que estos datos revelan es que las escalas usadas en el modelo deben ser revisadas o estudiadas cuidadosamente para determinar si estos valores podrían arrojar mejores predicciones medidos bajo otras escalas.

7 CONCLUSIONES Y TRABAJO FUTURO

El trabajo desarrollado en este proyecto tiene la ventaja de que provee un punto de partida que puede ser extendido y analizado con mayor profundidad para obtener modelos matemáticos más precisos sobre los cuales calcular el valor del *indicador SOFIA.* En este aspecto se enfatiza en que a pesar de que se intentó reducir al mínimo el número de supuestos sobre los cuales trabajar, aún existen algunas suposiciones que pueden ser puestas a prueba para maximizar la utilidad de este indicador.

Las diferentes fuentes de información utilizadas en este trabajo poseen características variadas que permiten proveer indicadores sobre el estado de ánimo en el mercado, sin embargo el análisis de estás depende también de un conjunto de variables que al ser modificadas pueden alterar drásticamente el resultado obtenido, por ejemplo, en este trabajo se consideraron noticias en el área económica, política e internacional, además de algunos tópicos que tienen el potencial de influenciar el estado de ánimo de las personas, tales como los deportes. La selección de que noticias son relevantes para calcular el sentimiento general del mercado depende de estudiar en detalles cada una de las variables que afectan al mismo, lo cual escapa al alcance de este proyecto. Es por razones como ésta que el *indicador sofia* aún tiene un largo camino antes de poder ser usado en el mercado de inversión

El estado actual del modelo matemático usado en el proyecto, junto con sus supuestos, no permite la utilización del mismo en un escenario real de mercado. A pesar de que el error total mostrado por los resultados es bajo 0.3%, se debe considerar que éste está supeditado a la variabilidad del dólar en el mercado, la cual se ubica en torno al 5%, lo que supone en techo máximo para el error. Aún más, los valores calculados de los sentimientos para las diferentes fuentes de información rondan en promedios muy bajos, solo uno logra superar 0.1, lo que reduce el error absoluto posible, dado que los valores calculados tienden a alejarse menos del valor real de la tasa de cambio. i.e. un sentimiento ponderado de 0.1 multiplicado por una tasa de variabilidad de 5% daría un error máximo para el siguiente valor del dólar de 0.5%.

De cualquier manera debido al diseño realizado en la aplicación, está permite que los cambios y ajustes en el modelo sean muy fáciles de implementar, lo que da lugar a que aumentar la precisión obtenida por el modelo con el objetivo de convertirlo en un indicador más fiable, resulte una tarea de baja complejidad a nivel de programación.

A pesar de que el modelo en su estado actual no permita su uso en el mercado Forex, si presenta un planteamiento interesante sobre la posibilidad de captar el sentimiento del público como un indicador del posible comportamiento de las divisas de dos países. Aún más, es conveniente recordar que el objetivo principal de SOFIA es proveer el *indicador Sofia*, el cual es un monitor del estado de ánimo general del mercado, tal como se describe en las diferentes publicaciones sobre las que éste trabajo se apoya.

Por otra parte, la interfaz Web provee funcionalidades que pueden ser expandidas para incluir información que ayude a dar más contexto al indicador, así como incorporar nuevas fuentes de información introducidas directamente por el usuario.

Finalmente se busca que éste trabajo pueda ser usado como un punto de partida para la generación de indicadores que apoyen la toma de decisiones en los diferentes mercados de inversión, generando una herramienta que pueda ser utilizada en conjunto con otras existentes en el mercado, destinadas a proveer al inversionista con información útil para tomar mejores decisiones de inversión.

8 REFERENCIAS

[1] H. M. X.-J. Z. Johan Bollen, «Twitter mood predicts the stock market,» *Journal of Computational Science, 2(1),* pp. Pages 1-8, March 2011.

[2] Y. Karabulut, «Can Facebook Predict Stock Market Activity,» 21 octubre 2013. [En línea]. Available: https://www.ecb.europa.eu/events/pdf/conferences/140407/Karabulut_CanFacebookPredictStockMarketActivitiy.pdf?8f1a9a6dec415d891c42f5e649b7ac15. [Último acceso: 15 febrero 2014].

[3] H. Mao, S. Counts y J. Bollen, «{Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data},» *ArXiv e-prints,* #dec# 2011.

[4] P. Papaioannou, L. Russo, G. Papaioannou y C. Siettos, «Can social microblogging be used to forecast intraday exchange rates?,» *NETNOMICS: Economic Research and Electronic Networking,* vol. 14, nº 1-2, pp. 47-68, 2013.

[5] T. Rao y S. Srivastava, «{Modeling Movements in Oil, Gold, Forex and Market Indices using Search Volume Index and Twitter Sentiments},» *ArXiv e-prints,* #dec# 2012.

[6] MINISTERIO DE HACIENDA Y CRÉDITO PÚBLICO, «Notas Fiscales - Siguiendo la actividad sectorial a partir de Google Trends,» noviembre 2013. [En línea]. Available: http://www.minhacienda.gov.co/portal/page/portal/HomeMinhacienda/politicafiscal/reportesmacroeconomicos/NotasFiscales/22%20Siguiendo%20la%20actividad%20sectorial%20a%20partir%20de%20Google%20Trends.pdf. [Último acceso: 13 mayo 2014].

[7] Monetary y E. Department, «Triennial Central Bank Survey Foreign exchange turnover in April 2013: preliminary global results,» 2013.

[8] W. Essex, «Exploiting the Big Data advantage - right-sizing for FX algorithms,» *e-Forex Magazine,* vol. April, 2013.

[9] C. Altavilla y P. D. Grauwe, «Non-linearities in the relation between the exchange rate and its fundamentals,» *International Journal of Finance \& Economics,* vol. 15, nº 1, pp. 1-21, 2010.

[10] C. Restrepo-Arango, A. Henao-Chaparro y C. JimÃ©nez-GuarÃn, «Using the Web to Monitor a Customized Unified Financial Portfolio,» de *Advances in Conceptual Modeling*, vol. 7518, S. Castano, P. Vassiliadis, L. Lakshmanan y M. Lee, Edits., Springer Berlin Heidelberg, 2012, pp. 358-367.