# Instituto Tecnológico y de Estudios Superiores de Monterrey Campus Ciudad de México

Escuela de Graduados en Ingeniería y Arquitectura

Maestría en Ciencias de la Computación



Propuesta de Tesis



## ANÁLISIS DEL SENTIMIENTO PARA LA TOMA DE DECISIONES BURSÁTILES

AUTOR: Jorge Rodríguez Vázquez
ASESORES: Dr. Rafael Lozano Espinosa
Dr. José Martín Molina

México D.F., Diciembre del 2013

# Contenido

# CAPÍTULO 1

# Introducción

## 1.1 Antecedentes

La vertiginosa Era de la información ha desencadenado una brecha entre la capacidad de procesamiento computacional y el volumen de información disponible, dando lugar a un sin número de técnicas, espacios y modelos que permitan hacer uso de ésta para los distintos fines que constituyen el conocimiento humano.

Dentro de este conjunto de datos informativos existen opiniones que pueden influir de manera cualitativa la percepción pública de algún tema en particular debido al impacto positivo o negativo que ejerza sobre el tema tratado. Una de estas áreas donde la expresión e información juega un papel preponderante para la toma decisiones podemos encontrar la relacionada al ámbito bursátil (e.g. Bolsa Mexicana de Valores), en donde la evaluación de los activos de una empresa para la generación de capital así como para la formulación de estrategias que permitan predecir el valor de una acción en el futuro, se basan tanto en la información disponible entre los participantes así como el entorno social en el que se desarrollan.

La información financiera continuamente permite acrecentar la comprensión y el conocimiento del mercado, del tal forma que esta influye en el sentimiento del inversionista para reaccionar oportunamente a los cambios que este pueda presentar y así ajustar pertinentemente sus portafolios de inversión.

La incorporación del análisis de las noticias financieras a los modelos cuantitativos para la toma de decisiones es algo que actualmente se encuentra en etapas muy tempranas de investigación [7]. Ryan y Taffler han determinado que el 65% de los cambios en el precio y el volumen operado durante las operaciones bursátiles, se puede relacionar a la publicación y disponibilidad de las noticias afines dentro del medio bursátil.

A continuación se presenta una gráfica en donde se muestra los cambios en los precios (Figura 1.1) y el volumen (Figura 1.2) operado de la emisora denominada APPL (Apple Inc.) al momento en que fue anunciada la muerte de Steve Jobs el 5 de Octubre del 2011. Derivado de este conjunto de datos, se puede apreciar que los precios presentaron movimientos a la baja con respecto a los días anteriores, resultando un descenso en la operatividad del volumen accionario durante las horas de subasta; esto por la incertidumbre sobre la nueva dirección que habría de tener dicho emporio empresarial.

El proceso general para la toma de decisiones financieras (Figura 1.3) es originado por medio de la interacción de los datos de mercado, con los modelos de predicción para el retorno de inversión y cálculo de riesgo. Así mismo y como consecuente de la versatilidad en las soluciones tecnológicas que integran las instituciones privadas y públicas que participan en dicho proceso, es importante resaltar que aquellas enfocadas al análisis de la información del mercado, son un elemento fundamental y competitivo para contribuir en la comprensión de la estructura y dinamismo de los mercados financieros.

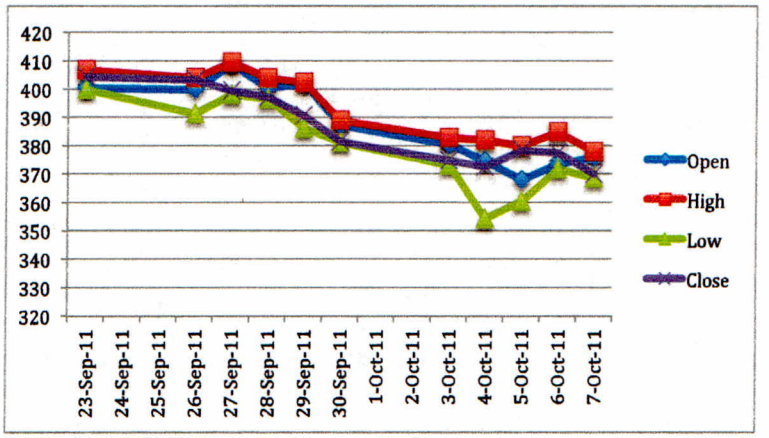



Figura 1.1 – Precios (USD) de las acciones de Apple del 23 de Septiembre al 7 de Octubre

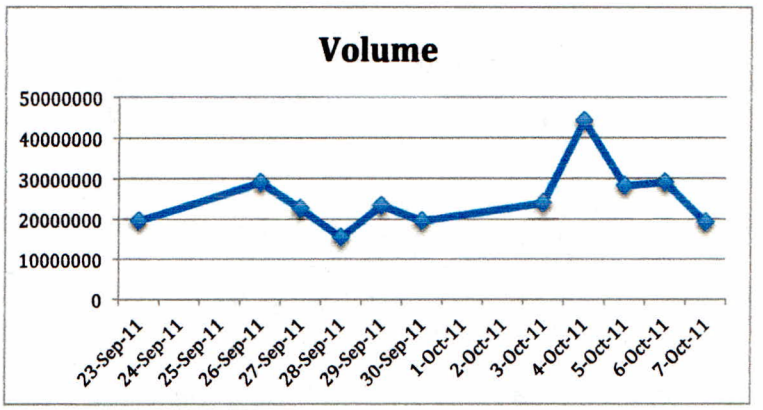



Figura 1.2 – Volumen operado de las acciones de Apple del 23 de Septiembre al 7 de Octubre

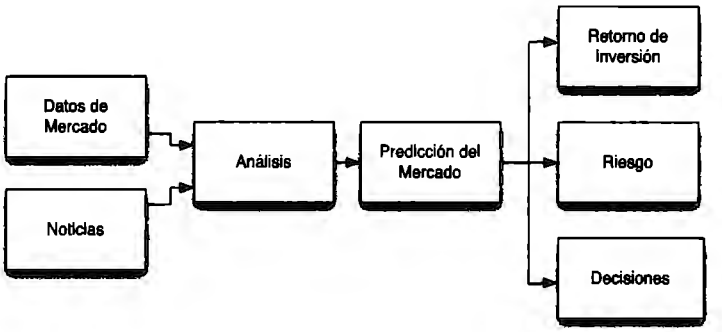



Figura 1.3 – Flujo de información para la toma de decisiones en el ámbito financiero propuesto por Mitra, G. & Mitra, L [26].

Así mismo y como consecuente de la versatilidad en las soluciones tecnológicas que integran las instituciones privadas y públicas que participan en dicho proceso, es importante resaltar que aquellas enfocadas al análisis de la información del mercado, son un elemento fundamental y competitivo para contribuir en la comprensión de la estructura y dinamismo de los mercados financieros.

De acuerdo a Siering [1] los mercados financieros son ambientes clasificados como sumamente complejos y cambiantes, debido a que los precios de las acciones se ajustan rápidamente por la disponibilidad de nueva información. Esta perspectiva está fundamentada en la teoría de un Mercado Eficiente por Eugine Fama [2], donde se argumenta que los cambios del precio dependen intrínsecamente de la información que los participantes poseen, para que éstos realicen evaluaciones en términos de su función de retorno. Dicha función originada por el autor Sharpe Lintner, describe la relación lineal positiva entre el valor esperado de una acción y su riesgo inherente (i.e beta), lo cual ha sido adscrita a lo que hoy se conoce como CAMP (*Capital Asset Pricing Model*).

Investigaciones financieras [3] han demostrado que los mercados al reaccionar a la exposición de nuevas noticias, generan movimientos bursátiles y finalmente difunden su efecto en el estado emocional de los participantes, como consecuencia, se han realizado esfuerzos intelectuales [4] orientados al estudio del contenido no estructurado de las noticias financieras para detectar los múltiples patrones que lo conforman y así interpretar el efecto que esta pueda ejercer.

Este enfoque abre una gama de posibilidades en la forma de analizar y extraer la percepción que contiene una noticia, de la cual debe considerarse su subjetividad, su estructura semántica, así como un mecanismo para cuantificar la forma en que puede ser empleada para los diversos modelos de predicción relacionados al análisis y la toma de decisiones financieras.

Por la naturaleza de la información, las técnicas basadas en aprendizaje asistido han sido empleadas por su nivel de maduración, como una alternativa para representar el efecto que estas ejercen sobre el lector una vez que estas son clasificadas e ingresadas en dichos algoritmos de inteligencia artificial. Sin embargo, éstos no han podido resolver [1] la disyuntiva entre la semántica del texto y el papel que desempeña cada elemento dentro de la oración de manera dinámica y flexible.

Como consecuencia, la presente investigación tiene como principal móvil reducir la brecha entre la estructura natural del texto y la interpretación que se deriva a partir de los elementos que lo constituyen, formulando señales positivas o negativas sobre el cambio de precio asociado al conjunto de noticias evaluadas.

Para lograr esto se planteará un análisis estadístico sobre el papel que ejerce cada unidad gramatical, así como su posible relación con los elementos adyacentes mediante la conceptualización e introducción de una polaridad entre ellos, que determinará la dirección general del sentido del texto (i.e. positivo o negativo) y será reforzado por medio de las señales asíncronos y síncronos que se encuentran en torno al desarrollo de un conjunto de noticias afines a una acción (i.e. datos de mercado).

## 1.2 Definición del problema

Las principales actividades del mercado bursátil se basan en la evaluación del riesgo y retorno de inversión que puede generar la compra y venta de acciones a largo plazo. Dicha evaluación está sustentada por el análisis del contenido informativo que se ha distribuido entre los participantes en condiciones igualitarias, es decir, asumiendo que cualquiera de ellos tiene acceso a la misma información.

La adopción de tecnologías para la distribución y captación de información ha sido una de las causante del incremento en el volumen disponible de datos para su análisis y toma decisiones, por lo cual es importante considerar que estos cambian abruptamente y su cuantificación constituye un desafío para las actuales Ciencias Computacionales, ya que la naturaleza del texto está fundamentada en una semántica no estructurada y cuya relación sus elementos puede determinar una orientación negativa o positiva según sea el caso.

Es importante determinar que la clave del desarrollo de esta rúbrica yace en la definición de interpretación que pueda generarse a partir de los elementos del texto, así como su relación con las señales o acciones que el mercado bursátil va concretando entre los participantes.

El eje principal para el desarrollo de la presente tesis yace en el concepto abstracto sobre el mecanismo general de aprendizaje expuesto por Chao [6] (Figura 1.2.1), el cual establece los pasos que preceden a la generación del conocimiento y por ende la interpretación del mismo, del cual se derivará un mecanismo que permita reconocer los elementos epímones mas representativos del texto.


Figura 1.2.1 – Proceso de aprendizaje propuesto por Chao [6]

Para dichos fines se fundamenta en el presente documento una estrategia computacional que permita analizar el papel que desempeña cada elemento lingüístico dentro de la noticia, hasta la formulación de sus relaciones entre las estructuras gramaticales que la contienen, de tal forma que a partir de un conjunto de palabras aisladas, se pueda determinar las categorías emocionales inherentes al documento mediante el reconocimiento de aquellas que permita establecer una relación de contrapeso con las de mayor predominancia, por medio de la polaridad desarrollada durante la redacción del texto.

Para abordar el presente paradigma se diseñará una arquitectura computacional (Figura 1.2.2) que describa cada una de las etapas de evaluación en las que será sometidas el texto no estructurado con el propósito de lograr una interpretación más concisa y donde se denote claramente la participación del modelo de interpretación en los mecanismos de análisis financiero.

Para abordar el presente paradigma se diseñará una arquitectura computacional (Figura 1.2.2) que describa cada una de las etapas de evaluación en las que será sometidas el texto no estructurado con el propósito de lograr una interpretación más concisa y donde se denote claramente la participación del modelo de interpretación en los mecanismos de análisis financiero.

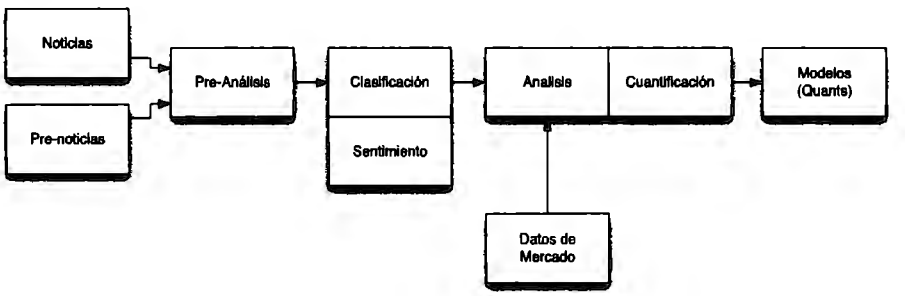



Figura 1.2.2 – Arquitectura general para el análisis de sentimiento

## 1.3 Objetivos

Durante el desarrollo del presente trabajo se proponen implementar un método computacional complementario a los ya existentes que permita determinar el o los sentimientos existentes entre los elementos del texto informativo, teniendo como característica principal la detección e interpretación de las ideas que lo componen por medio de la polaridad entre sus elementos.

Debido a que esta rama de la investigación aún se encuentra en etapas muy tempranas de desarrollo [7], el alcance la investigación será el proveer una representación del sentimiento que pueda emplearse para toma de decisiones y a su vez que sirva a estructuras computacionales más abstractas para la resolución de relaciones entre un conjunto de datos afines.

De tal forma, que la introducción del presente algoritmo constituirá una alternativa para resolver el proceso de la interpretación de un texto informativo sin la correlación de otros, dejando para investigaciones subsecuentes una forma más simple de crear relaciones entre un conjunto de datos semejantes.

## 1.4 Justificación

La información proporcionada a través de las noticias desempeña un papel crucial para la toma decisiones en los mercados financieros y la capacidad de poder analizarla de manera volumétrica, marcaría una diferencia sustancial en el nivel de competencia empresarial así como en las diversas estrategias empleadas para la predicción del retorno de inversión; ya que se reduciría la latencia en el tiempo de respuesta ante los diversos y cambiantes entornos en los que se desenvuelven los participantes.

Aunque el desarrollo de este tipo estrategias aún distan mucho de un desempeño deseable, el pináculo del problema radica principalmente en cómo se relaciona la información y qué se puede concluir a partir de ella, sin embargo esto no puede ser solucionado sin antes afinar las bases sobre el análisis de las unidades lingüísticas del texto y su semántica para detectar el sentido general de éste.

Durante el 2012 el crecimiento económico presentando en México ha sido ponderado dos veces el de Brasil [5], siendo esto una oportunidad de expansión para el mercado bursátil local en materia de captación de capital y el preámbulo para la generación de expectativas que traerán como consecuencia un mayor flujo de información sobre las variables económicas locales, así como el impacto de estos a nivel global.

Por lo tanto, el alcance de estos objetivos es sustentable en un mercado donde este concepto no existe inmerso dentro de las plataformas tecnológicas actuales y que se encuentra influenciado fuertemente por el dominio de mercados ya maduros como lo es EE.UU y la Unión Europea, por lo que esta tesis pretende ser precursora en el contexto nacional mexicano y uno de los motivadores principales para su aplicación, inversión e investigación.

Por lo tanto la necesidad de analizar la información que atañe a las decisiones mercantiles se torna indispensable para facilitar las actividades de los participantes bursátiles en el mercado local. Es de suma importancia resaltar que su aplicación puede extenderse a ciencias o áreas donde su ejecución podría retroalimentar la inclusión de objetivos muy específicos en un conjunto poblacional, tal como la introducción de nuevos productos en el mercado, reformas sociales o bienes sondeos donde se busque cuantificar la percepción pública.

## 1.5 Hipótesis

Los estudios actuales han demostrado que puede establecerse mecanismos computacionales que permitan representar la relación entre los elementos gramaticales, de tal manera que se deduzca por medio del aprendizaje supervisado su interpretación, sin embargo es primordial considerar que durante la redacción de cualquier escrito el emisor modula el tono de la redacción dependiendo de la intención del escrito.

Basado en lo anterior, el presente estudio establece que existe una relación intrínseca entre la distribución de las palabras a lo largo del texto y el uso de modificadores semánticos, de tal forma que se puede conceptualizar estos como partículas con propiedades que permitirán ejercer una fuerza de atracción, para asociar los elementos contiguos y así identificar la formación de ideas que establecerán los limites entre un sentimiento con respecto a otro.

Esta fuerza será denominada polaridad y a partir de los desplazamientos entre los posibles sentimientos por medio de su ponderación, podrá determinarse la combinación de mayor frecuencia y por lo tanto la de mayor probabilidad.

Es por esta razón que la necesidad de analizar la información que atañe a las decisiones mercantiles se torna indispensable para facilitar las actividades de los participantes bursátiles en el mercado local

# CAPÍTULO 2

# Marco Teórico

## 2.1 Generalidades

En el ámbito financiero la información constituye una de las variables más importantes para la toma de decisiones, considerado que ésta se encuentra disponible para todos y entre cada uno de los participantes [2], su influencia permite que los precios de las acciones denoten una tendencia a la alza (mantener posición larga y vender) o a la baja (mantener posición corta); lo cual se contraponte con los resultados de Muntermann [3,4] que sugerían la necesidad de aguardar un lapso de 30 a 15 minutos para que la información disponible en el mercado se encuentre completamente reflejada en los precios.

La participación de la información en el calculo del valor esperado del precio y retorno de inversión, se puede describir como una variable más dentro de la curva de distribución y denotarse como a continuación se enuncia.

$$E(p'_{j,t+1} \mid \Phi_t) = [1 + E(r'_{j,t+1} \mid \Phi_t)]p_{jt} \qquad (2.1.1)$$

Fórmula del del valor Esperado del precio descrita por Eugine Fama [2]

Donde **E** es valor esperado; $p_{jt}$ es el precio de la acción ***j*** en un tiempo ***t***; $p_{j,t+1}$ es el precio de la acción en el tiempo ***t+1***; $r_{j,t+1}$ es el porcentaje de retorno $(p_{j,t+1} - p_{jt})/p_{jt}$ en un tiempo determinado; $\Phi_t$ es la información que refleja el precio de una acción en el tiempo ***t***; las tildes indican que $p_{j,t+1}$ y $r_{j,t+1}$ son variables aleatorias en el tiempo ***t***.

Para fundamentar la eficiencia del mercado, Fama establece los siguientes supuestos en torno al precio y el retorno de inversión esperados, en donde la esperanza es igualada a cero para establecer que las variables tienen la misma probabilidad de ocurrencia y que entre los participantes existirá una participación equitativo.

$$x_{j,t+1} = p_{j,t+1} - E(p_{j,t+1} \mid \Phi_t),\ E(x'_{j,t+1} \mid \Phi_t) = 0 \qquad (2.1.2)$$

Supuesto referenciado al precio y adscrito por Eugine Fama [2]

$$x_{j,t+1} = p_{j,t+1} - E(p_{j,t+1} \mid \Phi_t),\ E(x'_{j,t+1} \mid \Phi_t) = 0 \qquad (2.1.3)$$

Supuesto referenciado al retorno y adscrito por Eugine Fama [2]

El evento $\{x_{j,t+1}\}$ representa en términos económicos el exceso del valor de la emisora ***j*** en el tiempo ***t+1*** con respecto al precio observado y el precio esperado en el tiempo ***t*** de acuerdo a la información $\Phi_t$ disponible. Deforma similar $\{z_{j,t+1}\}$ es el retorno de inversión en el tiempo ***t+1*** proyectado en el tiempo ***t***.

Si consideramos que $\alpha(\Phi_t) = [\alpha_1(\Phi_1), \alpha_2(\Phi_2), ..., \alpha_n(\Phi_t)]$ denotan la cantidad de fondos $\alpha_j(\Phi_t)$ ha invertir, el valor generado en el tiempo ***t+1*** estará determinado de la siguiente manera.

$$V_{t+1} = \sum_{j=1}^{n} \alpha(\Phi_t)[r_{j,t+1} - E(r'_{j,t+1} \mid \Phi_t)] \text{, y por lo tanto } E(V'_{t+1} \mid \Phi_t) = \sum_{j=1}^{n} \alpha(\Phi_t)[r_{j,t+1} - E(r'_{j,t+1} \mid \Phi_t)] = 0 \quad (2.1.4)$$

y

(2.1.5)

El valor generado a partir de la información disponible con los activos financieros sugerido por Eugine Fama [2]

De las ecuaciones anteriores podemos interpretar que la información es una variable fundamental para el dinamismo y el desarrollo del merado.

## 2.2 Información

La información ha sido un tema de constante evaluación por su papel estratégico en la generación de conocimiento. Autores como Zhong [10] establecen que la información representa los posibles estados de un objeto y la forma en que estos pueden ir cambiando, así como la importancia de generar conocimiento a través de la comprensión de sus estos, por medio de la generación de estrategias que permitan resolver los problemas creados a partir del conocimiento recolectado.

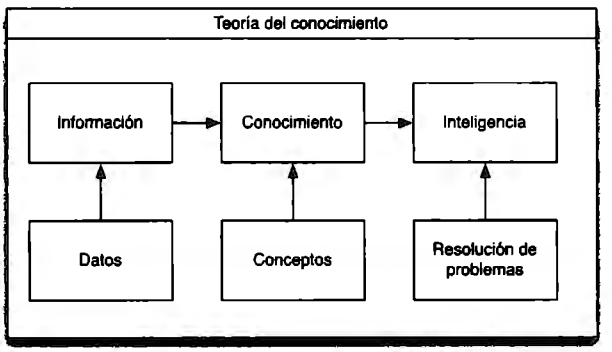



Figura 2.2.1 – Representación del conocimiento de acuerdo a los estados mencionados por Zhong[10]

## 2.3 Clasificación del Texto

Basado en la Teoría del conocimiento[10] que esquematiza el proceso general para conceptualizar mecanismos que a partir de un conjunto de datos logren interpretar el conocimiento acumulado, la clasificación del texto es una pieza fundamental para brindar las bases que permitan generar los conceptos requeridos para dicho propósito.

Diversos estudios abordan la clasificación del texto de acuerdo al efecto que éste ejerce sobre el dominio de interés, siendo la interpretación del contenido un desafío para identificar las categorías inherentes a éste [11]. Zhong establece que debe existir una relación intrínseca entre el algoritmo o herramienta que extrae los conceptos con la estrategia para la formulación del conocimiento, de tal forma que podamos establecer la siguiente relación.

$$K_C : K_F \alpha K_U \tag{2.3.1}$$

Elementos que constituyen el conocimiento de acuerdo a Zhong [10]

Donde $K_F$ es el conocimiento generado, $K_U$ es el algoritmo empleado para dicho propósito y $K_C$ establece la relación lógica entre ambas variables. De esta forma se podrá cumplir con la relación de los tres elementos esenciales para la formulación del conocimiento anteriormente descritos (i.e. información, conocimiento e inteligencia).

Para Nithya el principio en la Clasificación del Texto (TC) [12] yace en el aprendizaje supervisado por medio de las redes neuronales, donde, los datos provenientes de la fuente de información para lograr agruparlos, son etiquetados manualmente de acuerdo al dominio de la aplicación (e.g. antigüedad, tipo, nivel de sentimiento, etc.).

La mayoría de los elementos que integran el texto para este autor no son informativos, por lo que el análisis debe estar basado en la detección de conceptos comunes a lo largo del cuerpo del documento y entre aquellos que pertenecen al conjunto de evaluación, de tal forma que se asigne una frecuencia (i.e. TF–IDF Text frequency–inverse document frequency ) que permita cuantificar el número de apariciones y por lo tanto su grado de contribución.

La detección de similitud entre conceptos, se logra a partir de la relación entre los tres aspectos críticos dentro del algoritmo de Nithya; los conceptos seleccionados a lo largo del cuerpo del documento, la frecuencia del concepto es la medición empleada para determinar su contribución y el número de documentos analizados que contienen dichos conceptos.

De tal forma que por medio del empleo de la máquina de aprendizaje conocida como *K-Nereast Neighbour* se identifique la propiedad con mayor influencia por medio de los elementos contiguos (i.e. *neighbours*) y por lo tanto los elementos más cercanos serán aquellos que contribuyan más que los más distantes.

Sin embargo para lograr una interpretación del significado que aporta los elementos del texto, no es suficiente que sean analizados por estructuras que determinen patrones en su comportamiento o constitución [11]; deben ser expuestos a un nivel de detalle desde el cómo está constituida su semántica hasta la relación que sostienen éstos en el conjunto de datos como lo ha expuesto Nithya et al [12].

Emplear únicamente estrategias como las provistas por la Inteligencia artificial resultaría inexacto, puesto que existen palabras que pueden otorgar un significado ambiguo sin tener la relación completa entre la estructura semántica que se analiza.

Por ejemplo en el idioma Inglés se tiene la palabra "*rising*", la cual por si sola se podría clasificarse como algo positivo; sin embargo si su semántica tiene como sustantivo la palabra "*debt*" este implica un significado completamente negativo.

Chao, Sam [6] aborda esta perspectiva mediante un análisis sofisticado de las estructuras que componen el documento y la relación entre éstas, mediante el empleo de Lenguajes Naturales para el Procesamiento (NLP), donde se denota que el punto central en cualquier implementación debe estar enfocado a la manera introspectiva en que el cerebro humano procesa la información, relacionando las sentencias y unidades gramaticales de algún contexto en cuestión; tal y como se realiza en la adopción de una segunda lengua materna por medio de la triada que Zhong[10] expone (i.e. extracción de información, generación de conocimiento e inteligencia).

Por otra parte Thakur [13] contrapone a esta problemática un método alternativo para la extracción del contenido en documentos no estructurados; el cual consiste en el uso de gramáticas libres de contexto (GLC) determinadas por la probabilidad de aparición y similitud entre los elementos del texto.

Una gramática libre de contexto se encuentra definida por $GLC = \{N, T, P, S\}$, donde $N$ representa los elementos no terminales, $T$ el conjunto de símbolos terminales, $P$ el conjunto de producciones y $S$ el símbolo inicial. El lenguaje $L(G)$ es el conjunto de todos las cadenas terminales $w$ que tengan al menos una derivación del símbolo inicial, siendo que las reglas de producción gramatical estarán conformadas por las cadenas generadas a partir de los símbolos terminales y no terminales.

$$\{R_i : N^{ij} \rightarrow C^j\}_{i=1}^{r} N^{ji} \in N \wedge C^j \in (N \cup T) \tag{2.3.2}$$

Símbolos no terminales del árbol semántico enunciado por Thakur [13]

Thakur [13] adhiere a este concepto computacional el uso de una puntuación $S(R_i)$ basada en la función logaritmo de la probabilidad de $P(R_i)$ asociada a cada producción gramatical generada, es decir, dado que la resultante de este procesamiento es un árbol semántico cuyos nodos $R_j : N^{ij} \rightarrow C^j$ internos serán los símbolos no terminales y sus nodos externos serán los símbolos terminales, en función de sus secuencias $w$ $(N^{ij} \rightarrow w_a w_{a+1} ... w_b)$ se le asignará una probabilidad de acuerdo a lo siguiente.

$$S(R_i) = \sum_{k=1}^{F} \lambda_k(R_i) f_k((w_a, w_{a+1} ... w_b, R_i)) \tag{2.3.3}$$

Probabilidad definida por cada produción gramatical en base a la secuencia de palabras descrita por Thakur [13]

Donde $\lambda_k(R_i)$ es un vector de características distintivas del texto, el cual puede ser determinado por medio de estrategias que detecten la similitud entre los términos analizados con el fin de agruparlos y determinar cuáles son aquellos con mayor importancia.

La generación de producciones gramaticales para establecer las reglas que deben seguir los elementos del texto no estructurado para su clasificación, es referida por Devasena y Hemalatha [14] al comparar las estructuras semánticas de lenguajes como el Inglés y el Español, denotando que la segmentación de las oraciones radica principalmente en la función que aporta cada elemento dentro de la sentencia. Esta segregación permite que la redacción contribuya enormemente a la contextualización de los elementos que la componen, para esto, la forma de etiquetar los elementos deberá seguir las reglas morfológicas y gramaticales que son establecidos en cada lenguaje.

Devasena y Hemalatha [14] sugieren que el conjunto de reglas gramaticales que describen el idioma inglés pueden ser esquematizadas para conformar un árbol semántico más robusto como a continuación se describe.

| | |
|---|---|
| Sustantivos | $NP \rightarrow Det * Adj * NPP *$ |
| Posesivos | $Adj \rightarrow N's$ |
| Preposiciones | $PP \rightarrow PNP$ |
| Verbos | $VP \rightarrow Adj * V(NP \mid PP) * (PP \mid NP) * Adj *$ |
| Sustantivos + Verbos | $S \rightarrow NPVP$ |

Donde $(Det)$ incluye los artículos determinados, $(Adj)$ los adjetivos, $N$ los sustantivos, $PP$ las preposiciones.

Actualmente los sistemas de extracción de contenidos buscan representar el conocimiento por medio de objetos relacionados conocidos como entidades, los cuales dependiendo de los parámetros determinados durante el análisis del texto, definirán aquellos elementos con el mayor grado de importancia.

De acuerdo a Liu Hui y Zhang [15] la idea principal de la concepción de estos mecanismos es lograr generar un marco de trabajo que permita identificar los atributos que se desconocen dentro de las oraciones, de tal forma que por medio del entrenamiento asistido se logre clasificar correctamente la información y crear relaciones multifuncionales entre las entidades.

$$T = \{e, t.l, a, s, u, p\} \tag{2.3.4}$$

Elementos principales para la clasificación del texto por Liu Haui y Zhang [15]

En la relación $T$, $e$ representa la entidad buscada, $t$ un punto en el tiempo, $l$ es la ubicación, $a$ es una frase o palabra que es descrita por $e$, $s$ es la sentencia contenida por $T$, $u$ es la fuente de información donde $s$ fue encontrada y $p$ es la probabilidad de que $s \subset T$. De tal forma que dado un grupo de entidades $E = \{e_1, e_2, e_3..., e_n\}$ y el cuerpo del texto $C = \{s_1, s_2, s_3..., s_n\}$ el proceso de extracción obtendrá $s_i$ como una oración efectiva para $E$.

Al igual que Nithya [12], Chao [6], Lui y Zhang [15] denotan que la forma de agrupar la similitud entre las entidades resultantes y establecer las relaciones multifuncionales para facilitar la categorización, es mediante el uso del clasificador binario KNN (*k-nearest neighbor)*, sin embargo el cómo empatar las sentencias de forma que se pueda determinar un grado de semejanza entre los elementos de texto, es algo que dichos autores abordan por medio de lo que se le conoce como *Part-of-Speech Tagging* (POS).

A diferencia de lo propuesto por Devasena y Hemalatha [14], *Part-of-Speech Tagging* (POS) tiene como objetivo principal analizar las partículas gramaticales que constituye las frases, oraciones y párrafos para determinar su función semántica, ya que estas se encuentran definidas con respecto a los elementos contiguos [16]. El resultado obtenido de este procesamiento estará constituido por diferentes etiquetas que describirán las 9 partes fundamentales del habla; sustantivos, verbos, adjetivos, adverbios, preposiciones, pronombre, conjunciones, artículos e intersecciones.

El árbol de dependencia obtenido durante el análisis semántico, tiene como principal objetivo transformar el lenguaje natural en una forma de representación donde se pueda inferir el papel que desempeña cada elemento [15].

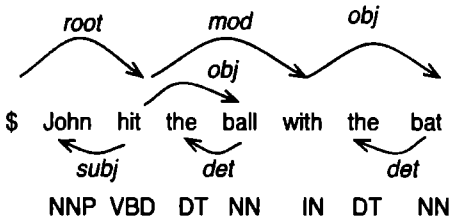



Figura 2.3.5 – Proceso general de *Part-of-Speech Tagging*

Actualmente POS constituye una alternativa a las producciones gramaticales propuestas por Devasena y Hemalatha [14], y Thakur [13], sin embargo la resultante de este procesamiento, sólo representa a un subconjunto de términos sin relación alguna, por lo que posterior a este tratamiento de información el empelo de herramientas como Penn Treebank facilitará la tarea artesanal de identificar qué clasificación le corresponde a cada uno de los elementos dentro del conjunto genera do por medio de una etiqueta.

Finalmente esta salida podrá ser sometida a la generación del árbol semántico de acuerdo a los pesos determinados y así por medio de una reducción del árbol semántico, encontrar el texto con mayor relevancia dentro del texto.

En 1965 Vladimir Levenshtein fue el primero en considerar que dicha estrategia podría ser sujeta a evaluaciones subsecuentes para determinar entre un conjunto de datos arbitrarios sus posibles similitudes.

Para tal propósito Levenshtein fundamentó un algoritmo que está relacionado directamente con el número de operaciones (i.e. inserción, eliminación y sustitución) requeridas para transformar una cadena con respecto a otra, por lo cual su desempeño se espera sea más óptimo en contenidos donde las diferencias entre la distancia de sus términos sean mínimas.

$$Sim(S_a,S_b) = 1 - \frac{2 * ed(S_a,S_b)}{|S_a| + |S_b|} \tag{2.3.6}$$

Formula para determinar la similitud entre dos conjutos evaluados y establecida por Levenshtein

Han surgido variantes del algoritmo de Levenshtein al adaptar diferentes tipos de operaciones para incrementar la precisión en su clasificación; tal es el caso del método Damerau–Levenshtein, cuya diferencia principal radica en la adición de una cuarta operación conocida como trasposición; Hamming, limita la aplicación a sólo la operación de la sustitución y por ende a palabras cuya longitud es la misma; Smith–Waterma, determina el costo asociado a la distancia Levenshtein mediante la extracción y comparación de todos los posible subsegmentos de las cadenas evaluadas.

Por otra parte la identificación de la raíz morfológica podría contribuir a la detección de semejanzas entre los términos que componen una oración, debido a que en ciertos lenguajes como el Inglés la inflexión de la palabras [12] puede ser determinada por medio de estrategias que se aboquen a la reducción de términos. Entre la vertientes más sobresalientes se encuenetra el algoritmo de Porter Stemming, cuyo objetivo es remover todos los sufijos para obtener la raíz de la palabra (e.g. agreed, agreeing, disagree, agreement, disagreement serán reducidos a la gramática agree)

Este tipo de estrategias brindan la posibilidad de identificar las palabras que no aportan información alguna a la semántica de la oración, generalmente estas son conocidas como *stop words* y pueden incluir desde preposiciones, artículos y algunos verbos (i.e. *the, is, at, which, and, on*). Finalmente esta opción de evaluación gramatical permite a las demás estrategias de clasificación, determinar la similitud entre oraciones con un mayor grado de precisión y a su vez excluir términos que introducen ruido en la redacción.

*The house was empty when the explosion occurred;* ***therefore*** *there are no reports of casualties.*

En este contexto el adverbio *therefore* no está agregando información extra a la oración debido a que su naturaleza en el idioma Inglés sólo es la de adverbio de conjunción, su omisión será completamente justificada y en beneficio tanto para la eficiencia del sistema en su precisión como en los recursos computacionales necesarios para su procesamiento.

La mayoría de las técnicas requiere de algún contexto en particular para determinar las características preponderantes en éste, sin embargo Hagenau et al. [7] sugieren que la importancia de la palabra en el documento se refleja en forma de estadístico al contemplar el número de apariciones contenidas en este y basadas en diccionarios que incluyen una amplia gama de palabras relacionadas al dominio en cuestión.

Por otra parte, el proceso de selección de palabras con mayor influencia en el texto no estructurado no solo se fundamenta para determinar la similitud entre un espacio finito de elementos morfológicos, sino para denotar la necesidad de la extracción o selección de información (i.e. selección de características comunes) para incrementar la precisión con respecto al dominio de evaluación [7].

## 2.4 Selección de Características

La producción de conocimiento ha sido abordado por áreas no solo de interés computacional sino también por aquellas donde el objeto de estudio es el hombre y su entorno social. Es en esta variante del conocimiento humano donde se destaca la participación de los Sociólogos Barney Glaser y Anselm Strauss al formular una teoría sobre la abstracción y relación entre los términos más representativos de cualquier texto.

Su teoría conocida como Fundamental, establece que los términos deben ser extraídos y agrupados en base a una serie de etiquetas derivadas del propio documento para describir con mayor grado de precisión los conceptos inmersos en el texto no estructurado, lo cual representa un esfuerzo por reducir la brecha entre el análisis de la información y la generación del conocimiento.

Basado en lo anterior, autores como Zhong [10], Hagenau et al [7] y Lui et al [15] hacen uso de este enfoque para dotar sus algoritmos de una mayor grado de confiabilidad durante el proceso de evaluación.

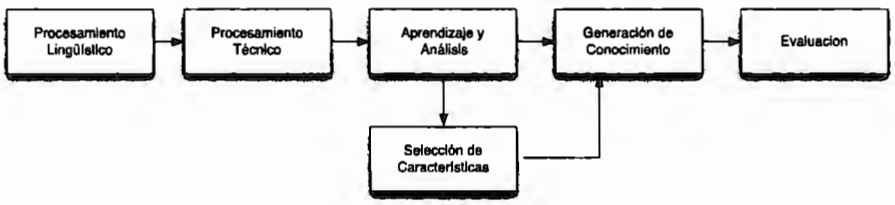



Figura 2.4.1 – Proceso de aprendizaje propuesto por Chao [6] con la selección de características propuesta de Zhong [10], Hagenau et al [7] y Lui et al [15]

Partiendo del mecanismo propuesto por Zhong [10] para la identificación de conceptos, podemos resaltar la ejemplificación en el siguiente diagrama

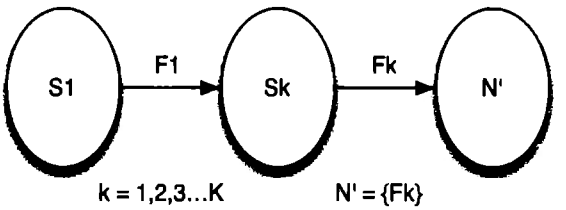



Figura 2.4.2 – Proceso general de la selección de características

Donde $S_1$ constituye el conjunto de datos aleatorios a analizar, $k$ representa las características distintivas del conjunto $S_1$ depuradas por medio de algún criterio de evaluación para preservar su significancia en las subsecuentes comparaciones $S_k$. Finalmente al conjunto de datos derivado de todas las características seleccionadas se extrapolaran para formular $N'$ y que servirá de base para la generación de los conceptos afines.

Es importante denotar que el uso de la extracción de características en un espacio finito de elementos morfológicos, han sido documentadas por Hagenau et al [7] para resaltar el incremento en la precisión de la clasificación como consecuencia.

| **Autor** | **Característica** | **Selección** | **Precisión** |
|---|---|---|---|
| Schumaker et al. 2009 | Noun Phrases | Mínimo de ocurrencias por texto | 58.2% |
| Muntermann et al. 2009 | Bag-of-words | Eliminación de stop words | 56.2% |

| Autor | Característica | Selección | Precisión |
|---|---|---|---|
| Mittermayr 2004 | Bag-of-words | TF IDF | No son directamente comparables |
| Antweiler et al. 2004 | Bag-of-words | Criterio de mínima información | No disponible |
| Tetlock et al. 2008 | Bag-of-words | Diccionario predefinido | No disponible |
| Hagenau et al [7] | Word combinations | $Chi^2$ | 65.1% |
| Groth et al. | Bag-of-words | $Chi^2$ | No son directamente comparables |
| Butler et al | N-Gram | Mínimo de ocurrencias por texto | No son directamente comparables |

Tabla 1 – Cuadro comparativo de la precisión obtenida por diversos métodos de selección de características, presentados por Hagenau et al [7]

Dentro del proceso de la Selección de Características Comunes se derivan etapas que robustecen la manera en cómo son seleccionadas las propiedades más distintivas del texto.

De acuerdo a Hagenau et al [7] la primer fase de este proceso es catalogada como la Extracción, la cual ha sido ejemplificada y abordada por las diversas estrategias previamente descritas a lo largo del texto (e.g. árboles gramaticales, número de incidencias de acuerdo a un diccionario del dominio en cuestión, función semántica que desempeña cada término, entre otros.).

La segunda fase permite depurar los resultados obtenidos por la Extracción mediante la asignación de una ponderación determinada por factores externos o internos al texto como punto de referencia conceptual (positivo, negativo, etc.). Posteriormente la etapa que nos permitirá representar computacionalmente las abstracciones obtenidas a lo largo de las etapas descritas, será la representación.

La implementación de este tipo de definiciones es la parte medular en cualquier marco de trabajo cuyo objetivo sea la interpretación del texto, ya que sin estos conceptos la categorización resultaría someramente exacta y no aportaría ningún progreso al campo de la lingüística computacional.

## 2.5 Aprendizaje

Ha sido corroborado a través de los estudios de Hagenau et al [7], [11] Johnson, Barry, Nithya [12] y Preethi et al [17] que la máquina de aprendizaje supervisado constituye el mecanismo con mayor viabilidad y fiabilidad para la clasificación de contenido a partir de la agrupación de conceptos similares. Dichos mecanismos son conocidos como máquinas vectoriales (SVM) y están diseñadas para analizar el ingreso de cada nuevo data al espacio de estudio mediante la identificación de los límites que demarca su similitud.

A diferencia de la selección de características, esto ha dado pauta a investigaciones para enfocarse en la mejora de este tipo de mecanismos para acrecentar la precisión en la clasificación de la información, tal es el caso de la especialización conocida como PSO-SVM (Optimización de la partícula de Swarm) cuyo aplicación ha sido orientada a los temas vínculos de algún foro digital Preethi et al [17].

### 2.5.1 Medición

Por otro lado la medición de éxito debe ser cuantificada por algún método que nos permita establecer la relación entre los conjuntos provistos por el mecanismo de aprendizaje supervisado y lo esperado.

Para esta consideración y evaluación Nithya et al [12] y Liu Hui y Zhang [15] emplean *F-Measure*, que consiste en una combinación de conceptos de Precisión y Recuperación. El primero estriba en obtener el porcentaje de documentos obtenidos que en realidad son relevantes (lo esperado) y lo segundo es con respecto al porcentaje de documentos que son relevantes y que realmente se obtuvieron como resultado de la clasificación. De esta forma se podrá construir un criterio de evaluación sobre la categorización y fundamento de la teoría.

## 2.6 Desempeño

Finalmente uno de los temas de mayor trascendencia en el actual entorno computacional es la capacidad de procesamiento que se tiene para resolver estas disyuntivas. Algunos autores como Xiaohui [18] presentan la posibilidad de emplear unidades graficas de procesamiento (GPU) para la solución de problemas mediante la explotación del uso de hardware en paralelo.

Sin embargo, la propuesta de Xiaohui [18] sobre un marco de trabajo colaborativo es mucho más detallado en el sentido de recursos más que de implementación, su viabilidad en este contexto es estudiado en su investigación para determinar cuales son las segmentos que realmente pueden ser sujetos al paralelismo.

En contraste tenemos la perspectiva de Sakurai et all [11], que arguye la generación de reglas y detección de patrones en un ambiente regido por el paralelismo y en tiempo real, sería innecesario para los métodos de aprendizaje e implicaría un cálculo complejo por la cantidad y tipo de datos en las secuencias de información.

Una de las ideologías predominantes en el uso de lenguajes que permiten la explotación del hardware en paralelo, es la de mantener los recursos el mayor tiempo posible ocupados.

Es por esto que el marco de trabajo colaborativo propuesto por Xiaohui [13] se enfoca en la productividad al procesar volumétricamente la información por medio del TF–IDF y la manera en como debe proveerse una arquitectura entre software y hardware para lograr la convivencia entre ambas perspectivas.

# CAPÍTULO 3

# Generación del Conocimiento

## 3.1 Elementos Fundamentales

De acuerdo a lo establecido por Zhong [10], es mandatorio definir un mecanismo que permita describir los pasos en los que será sometido el texto para la generación de conocimiento a partir de la información disponible, por lo tanto y para efectos del presente documento, dicho mecanismo se definirá como un marco de trabajo que sintetizará los elementos más importantes del contenido para lograr una categorización cualitativa y cuyos elementos principales se describirán a continuación.

La parte inicial para la creación de conocimiento cualitativo yace en el objetivo buscado y denominado como $G$, el cual podrá ser descompuesto en un múltiples tareas $G = \{G_n\}_{n=1}^{N}$ que podrán ser evaluadas de manera individual y que permitirán que cada pieza de información $X$ aporte significancia sintáctica dentro de la descripción cualitativa general $D(X)$ del documento.

Es importante denotar que la evaluación subsecuente $\{G_n\}$ de cada pieza de información que se añada a la descripción $D(X)$, imprimirá un efecto cualitativo de índole positivo $u_n^+$ o negativo $u_n^-$ que permitirá ajustar la descripción general del documento para alcanzar $G$. Por lo tanto el marco de trabajo describirá en términos de este conjunto de elementos $\{D(X),G;u\}$ los pasos relacionados que se deben llevar a cabo entre las partículas que conforman el texto para la obtener como resultado la interpretación de su contenido semántico y por lo tanto su clasificación por medio de la generación de conocimiento.

## 3.2 Procesamiento lingüístico y semántico

El primer paso para lograr la clasificación del contenido estará sustentado en la forma en que el proceso lingüístico realiza el análisis de los datos para establecer un grado de comprensión y así proveer al mecanismo de aprendizaje un conjunto de datos descriptivos y relevantes.

La comprensión de la información de acuerdo a Frank Smith [19] es el proceso estructurado de relacionar nuevo conocimiento al ya preexistente de tal forma que pueda generase un sentido positivo o negativo de toda la información agrupada, el cual constituye la base cognitiva para ostentar la construcción de abstracciones mediante la interrelación de diversas fuentes de datos no homogéneos. Cabe destacar que la diversidad de algoritmos abordados por los autores mencionados, tienen la particularidad que han sido proyectados para escenarios donde el dominio de la información es la clave para incrementar la precisión y mejorar la evaluación del conocimiento a fin.

Entre los ejemplos más representativos podemos encontrar que la agregación de otros documentos con un contexto de aplicación similar, robustecen ciertos algoritmos probabilísticos [21] que determinan la definición de la categoría que mejor se ajusta en base a la evaluación de la palabra.

Adwait Ratnaparkhi en su trabajo [21] expone la necesidad de crear un modelo capaz de usar durante la clasificación del texto las características del conjunto evaluado con el propósito de identificar la función semántica que le corresponde a cada palabra y de esta forma reducir el riesgo de una ambigüedad, maximizando el modelo de aprendizaje no asistido. Este enfoque fue desarrollado pensando en escenarios donde la evaluación del contexto incluirá palabras cuya aparición en el modelo de entrenamiento haya sido nula, de tal forma que se busque deducir a posteriori y mediante un conjunto de factores, la función semántica dentro de la oración de dichas palabras desconocidas. A este modelo formalmente se le conoce como el de Máxima Entropía y tiene como particularidad desde el punto de vista estadístico, que no fuerza ningún tipo de distribución sobre el conjunto de evaluación.

La aplicación de esta estrategia ha llevado a sintetizar el conjunto de posibles funciones semánticas que el modelo de Máxima Entropía [22] puede asignar a un texto analizado, las cuales fueron obtenidas de la aplicación de dicho modelo en los periódicos de Wall Street y del cual se llegó obtener hasta un 96.5% de precisión.

No obstante, dentro del procesamiento natural de la información existen elementos cuya semántica no contribuyen en alguna manera al sentido de la oración y que se encuentran distribuidas a lo largo de cualquier texto [20].

Autores como [7] y [20] aluden que el uso de diccionarios con este tipo de palabras, conocidos como *bag-of-words*, son insuficientes para eliminar completamente la estática que generan en la distribución del contenido, esto debido a que su aplicación es independiente de la semántica a la que está asociada en el contenido.

Debido a que nuestro modelo pretende reducir el ruido en el contexto de evaluación sin recurrir a la asociación de otro, representa una desventaja ante la mayoría de los algoritmos que actualmente buscan definir sus propias condiciones al ir adhiriendo progresivamente nuevos documentos (e.g. ontologías) y a su vez nos imposibilita el ir agregando al modelo de entrenamiento dichas incidencias, sin embargo, al ser palabras cuya aparición es escasa, su baja probabilidad de asertividad no influenciará lo suficiente para cambiar la interpretación del documento.

Por otra parte, el análisis semántico ha buscado incluir una gama sustancial de palabras que permitan la identificación del rol de cada una de ellas dentro de la estructura gramatical. Diversos autores han presentado alternativas para confrontar dicho reto semántico, tal es el caso del esfuerzo realizado en 1961 por Henry Kucera y W. Nelson Francis en la universidad de Brown en Rhode Island, en la cual recopilaron de 15 géneros literarios como Matemáticas, Política, Novelas, entre otros, conjuntos de palabras (~1,014,312) que no sólo por su frecuencia fueran relevantes, sino que incluyeran la mayor cantidad de conocimiento descriptivo y funcional [21][22].

Cabe destacar que el rango de frecuencias de una palabra está asociado al número de documentos analizados, de tal forma que dicha asociaciones está representada por medio de la distribución Zipfian.

Esta distribución alude a que la frecuencia de cualquier palabra es inversamente proporcional a su rango en la tabla de frecuencias, es decir, la palabra más frecuente ocurrirá aproximadamente el doble de veces que la segunda palabra más frecuente y aproximadamente tres veces más que la tercera palabra más frecuente y así sucesivamente. De esta forma se podría determinarse en un conjunto de documentos evaluados correlacionados con los temas vinculados, cuales son realmente las palabras que generan ruido dentro del espacio muestral e incrementar así la precisión en la evaluación lingüística[20].

La aplicación de esta estrategia semántica ha llevado a sintetizar el conjunto posible de ctiquetas quc cl modclo dc Máxima Entropía [22] pucdc asignar a un tcxto analizado. Éstas sc han derivado del análisis que Ratnaparkhi [21] aplicó a los periódicos de Wall Street y del cual se llegó obtener hasta un 96.5% de precisión. La siguiente lista muestra el conjunto de etiquetas que se emplearán como estándar, para la identificación del espacio muestral y que se derivan del estudio de [21].

| | | | | | |
|---|---|---|---|---|---|
| 1 | CC | *Coordinating conjunction* | 21 | PRP | *Pronoun personal* |
| 2 | CD | *Cardinal number* | 22 | PRP$ | *Pronoun possessive* |
| 3 | DT | *Determiner* | 23 | RB | *Adverb* |
| 4 | EX | *Existential there* | 24 | RBR | *Adverb, comparative* |
| 5 | FW | *Foreign word* | 25 | RBS | *Adverb, superlative* |
| 6 | IN | *Preposition or subordinating conjunction* | 26 | RP | *Particle* |
| 7 | JJ | *Adjective* | 27 | SYM | *Symbol* |
| 8 | JJR | *Adjective, comparative* | 28 | TO | *to* |
| 9 | JJS | *Adjective, superlative* | 29 | UH | *Interjection* |
| 10 | LS | *List item marker* | 30 | VB | *Verb, base form* |
| 11 | MD | *Modal* | 31 | VBD | *Verb, past tense* |
| 12 | NN | *Noun, singular or mass* | 32 | VBG | *Verb, gerund or present particle* |

| | | | | | |
|---|---|---|---|---|---|
| 13 | NNS | *Noun, plural* | 33 | VBN | *Verb, past participle* |
| 14 | NP | *Proper noun, singular* | 34 | VBP | *Verb, non-3rd person singular present* |
| 15 | NNP | *Proper noun, singular* | 35 | VBZ | *Verb, 3rd person singular present* |
| 16 | NPS | *Proper noun, plural* | 36 | WDT | *Wh-determiner* |
| 17 | PDT | *Predeterminer* | 37 | WP | *Wh-pronoun* |
| 18 | POS | *Possessive ending* | 38 | WP$ | *Possessive wh-pronoun* |
| 19 | PP | *Personal pronoun* | 39 | WRB | *Wh-adverb* |
| 20 | PP$ | *Possessive pronoun* | | | |

Tabla 2 – Penn Treebank [29]

## 3.3 Raíz Morfológica de la palabra

Una de las principales características que una palabra tiene en su estructura gramatical gira en torno a lo que se le conoce como raíz morfológica. Comúnmente todas aquellas lenguas que se derivan de la base estructural grecolatina provista por las lenguas romances, tienen la bondad de que sus derivaciones y funciones permiten una mayor comprensión y adopción dentro del campo del razonamiento lingüístico. Por otra parte lenguas como el Inglés cuya conformación dista de dicha herencia, se han elaborado algoritmos como el de Porter Stemming [24], donde se puede identificar por medio de la inflexión básica de la palabra, cuál es la raíz que la conforma.

Como consecuencia de esta identificación puede usarse la inflexión de cada uno de los elementos que integran el texto a evaluar para correlacionar su raíz gramatical con las diversas categorías sociales que suelen emplearse en los textos informativos y así determinar la probabilidad de aparición dentro del conjunto de oraciones. Es imperante considerar que no todas las palabras obtenidas en esta fase de análisis se encontrará contenidas en las categorías disponibles, por lo que estas no serán empleadas por los subsecuentes evaluaciones.

## 3.4 Categorías de la palabra

El diccionario Harvard IV-4 [25] contiene una diversa gama de clasificaciones de las cuales sólo tomaremos aquellas cuya orientación sea positiva o negativa para conformar nuestra tabla de relación con respecto a las rúbricas que suelen ser empleadas en los textos informativos; si alguna de esta recae en más de una categoría, su ponderación será adicionada entre aquellas que la compartan. Este diccionario incluye lo relacionado al compendio de categorías de valor dirigido por Lasswell donde se considera la división de la lengua en 8 dominios. Cada uno de estos dominios alude al poder, rectitud, respeto, afecto, salud; riqueza, bienestar, entendimiento y habilidad.

| #.Words | Category | #.Words | Category | #.Words | Category | #.Words | Category |
|---|---|---|---|---|---|---|---|
| 1915 | Positiv | 136 | Rel | 719 | Virtue | 1266 | PowTot |
| 2291 | Negativ | 20 | Yes | 685 | Vice | 35 | AffGain |
| 1045 | Pstv | 7 | No | 696 | Ovrst | 11 | AffLoss |
| 1160 | Ngtv | 217 | Negate | 76 | Need | 55 | AffPt |
| 557 | Affil | 65 | PowGain | 53 | Goal | 96 | AffOth |
| 833 | Hostile | 109 | PowLoss | 70 | Try | 196 | AffTot |
| 1902 | Strong | 30 | PowEnds | 224 | Means | 52 | WlthPt |
| 2591 | Power | 53 | PowAren | 64 | Persist | 53 | WlthTrain |
| 755 | Weak | 228 | PowCon | 81 | Complet | 271 | WlthOth |
| 1039 | Submit | 118 | PowCoop | 137 | Fail | 378 | WiltTot |
| 168 | Pleasur | 134 | PowerAuPt | 56 | Begin | 37 | WlbGain |
| 254 | Pain | 81 | PowPt | 98 | Vary | 60 | WlbLoss |
| 49 | Feel | 79 | PowerAuth | 111 | Increas | 226 | WlbPhys |
| 311 | EMOT | 332 | PoeOth | 82 | Decreas | 139 | WlbPsyc |
| 125 | Stay | 146 | EnlGain | 112 | Casual | 270 | EndsLw |
| 25 | Raise | 27 | EnlLoss | 26 | Ought | 30 | Anomie |
| 194 | Exert | 18 | EnlEnds | 192 | Percieve | 193 | NegAff |

| #.Words | Category | #.Words | Category | #.Words | Category | #.Words | Category |
|---|---|---|---|---|---|---|---|
| 79 | Fetch | 585 | EnlOth | 21 | Compare | 126 | PosAff |
| 42 | Fall | 835 | EnlTot | 205 | Eval@ | 175 | SureLw |
| 81 | Think | 129 | TrnGain | 189 | Solve | 25 | NotLw |
| 348 | Know | 113 | TrnLoss | 314 | EVAL | 132 | If |
| 46 | FREQ | | | | | | |

Tabla 3 – Diccionario Harvard IV-4 [25]

## 3.5 Selección de Características

Una de las ventajas al tener identificado la función que desempeña el elemento dentro de la oración es el de generar por medio de su representación un conjunto de reglas que describan aquellas con mayor preponderancia para la generación de conocimiento.

Es importante denotar que las correlación de diversas características semánticas tendría efectos positivos dentro del estudio de la series financieras, ya que si se aplicaran en modelos como los expuestos por Linda M. Medina [23], donde se denota la existencia de una relación entre cada de los componentes subyacentes a un determinado sector económico dentro de la BMV, se definirían portafolios de inversión con estrategias para reducir el riesgo por las relaciones intrínsecas de los activos.

Siendo esto la primicia para la selección de los elementos más representativos de un texto se identificarán aquellos cuya función sea la de modificador cualquier de los elementos circundantes a estos.

De esta forma podremos describir de manera general las combinaciones que en cualquier texto informativo se hace uso para concretar las ideas principales, así como describir sus asociaciones en torno a la idea principal.

Todo conjunto de elementos descriptivos es susceptible a ser representado mediante la formación de una Gramática Libre de Contexto que establecerá los posibles símbolos del alfabeto, los estados iniciales, terminales y finalmente las producciones gramaticales que definirán las combinaciones válidas en el lenguaje. Aunque no se aplicará un modelo probabilístico como el empleado en [13] por la falta de asociación entre modelos, se tendrá en consideración estructuras sintácticas para conformar las reglas de producción.

## 3.6 Clasificador

El léxico y la semántica conjuntamente determinan el contexto del sentimiento, sin embargo, el mayor desafío consiste en cuantificar las cualidades positivas o negativas del texto de acuerdo al entorno financiero que sustenta al contenido.

Los clasificadores representan una alternativa analítica para transformar la naturaleza del texto en un espacio de medición; entre los más relevantes para el análisis de texto no estructurado, podemos encontrar los siguientes. El clasificador *Naive* está basado en el conteo positivo o negativo de la connotación de las palabras, por lo que si una palabra se clasifica como negativa, es porque el contexto lo requiere. El clasificador basado en discriminantes remplaza el simple conteo de palabras por un peso basado en una función discriminante, la cual se construye a partir de la semejanza entre diferentes categorías.

El clasificador de frases adjetivales – adverbiales, se fundamenta en otorgar un mayor peso al conjunto de palabras que cumplan con dicha estructura semántica. El clasificador Bayesiano es una variación del teorema de Bayes ya que emplea la probabilidad condicional para determinar si una palabra pertenece o no a una categoría. Del conjunto de mensajes disponibles se puede determinar el número de apariciones que tiene cada elemento gramatical dentro del mensaje para deducir a partir de esta frecuencia, la proporción del cuerpo del documento que pertenece a una categoría.

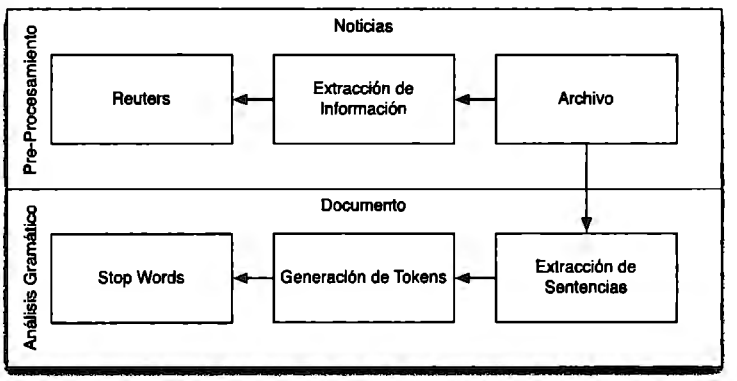
Finalmente podemos determinar que a partir de este conjunto de estrategias se puede integrar un marco de trabajo capaz de sintetizar la información y generar conocimiento de las categorías que se obtengan, a su vez, sentar las bases de los elementos que construirán el pilar fundamental.

# CAPÍTULO 4

# Algoritmo

## 4.1 Generalidades

A partir de la información recabada a lo largo del presente documento se describe a continuación el proceso implementado para analizar los textos informativos provenientes de la fuente conocida como Reuters (www.reuters.com), el cual estará acotado a las emisoras suscritas en la Bolsa Mexicana de Valores. Dicho algoritmo se encuentra dividido de a cuerdo a las siguientes etapas; Pre-procesamiento, análisis gramático, análisis semántico, categorización y representación.



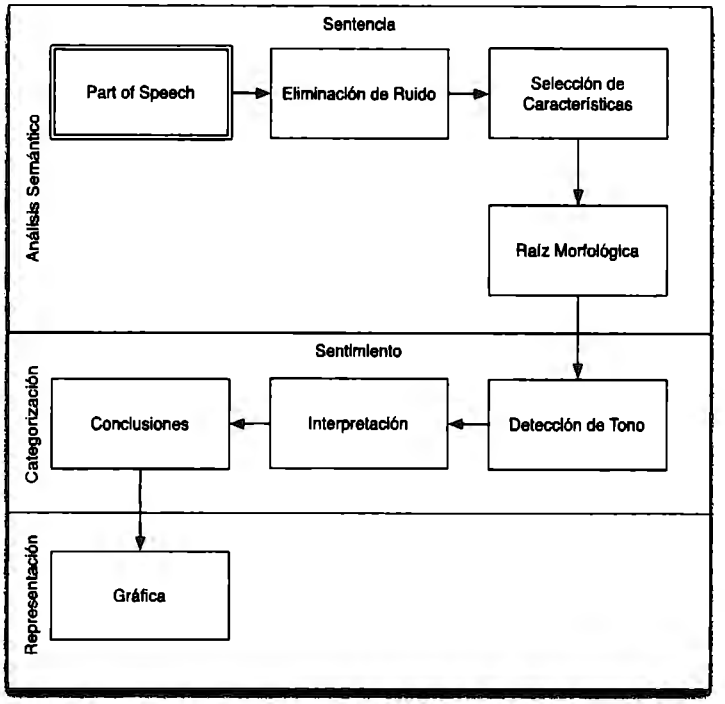



Figura 4.1.1 – Arquitectura del marco de trabajo implementado para la detección de sentimiento

El marco de trabajo representado en la Figura 4.1.1 ha sido implementado mediante el uso del lenguaje de programación conocido como JAVA en su versión 1.7, del cual se desprende un conjunto de APIS que facilitan la integración de un determinado grupo de funciones y de cuales se vale la presente propuesta para esquematizar la solución.

Se enlistan a continuación el conjunto de APIS empleadas para fundamentar los principios arquitectónicos que se emplearon (Tabla 4).

| | |
|---|---|
| Spring Framework | Su principal función será la de administrar la instancias que proveen servicio al motor principal de tal forma que se administre y segregue en un solo hilo de ejecución el procesamiento de información. |
| Apache OpenNLP | Apache diversifica su marco de trabajo mediante recursos que contienen entidades entrenadas por el aprendizaje asistido, de los cuales se puede destacar su implementación del precepto de la Máxima Entropía. |
| AspectJ | Debido a que la fuente de información puede contener caracteres especiales y pueden existir casos de excepción que no se hayan contemplado, se empleará esta perspectiva para procesar dichos escenarios que abruptamente interrumpan el procesamiento. |
| OpenCSV | El diccionario Harvard IV-4 se encuentra disponible en formato CSV por lo cual y para facilitar su lectura este marco de trabajo representa una opción viable. |
| Apache http | Los datos informativos e históricos se encuentran disponibles en sitios WEB (i.e. Reuters) por lo que realizar peticiones http es imperante para procesar el contenido. |
| Jsoup | Con la inmersión de XHTML los contenidos pueden ser extraídos y analizados con la ejecución simple de consultas sobre el XML, para lo cual este conjunto de funciones serán clave para dicho propósito. |

Tabla 4 – APIs empleadas para la implementación del marco de trabajo

## 4.2 Pre-procesamiento

Uno de los desafíos para procesar los datos informativos provenientes del mercado yace en la disponibilidad y obtención de la información, para lo cual se ha diseñado un programa capaz de extraer los datos históricos por medio de la comunicación del protocolo http del conjunto de páginas que disponen de dicho contenido (i.e. Reuters) a partir de Julio del 2011.

El compendio de datos obtenidos fue administrado por medio de archivos físicos para su posterior análisis, de tal forma que se pudiera indicar al programa principal el grupo de textos a evaluar de acuerdo a la fecha de publicación y orden de aparición, no obstante, es importante resaltar que la presente solución contempla un procesamiento posterior a la publicación física de la información y no en tiempo real de ejecución. Esto debido a que la implementación funge como un motor de análisis estacional donde solo interactúa una de las variables que integran la percepción pública de algún tema de interés del mercado financiero.

## 4.3 Análisis gramático

La mayoría de los documentos se encuentran conformados por conjuntos de sentencias que integran párrafos y que a su vez definen la estructura sintáctica del contenido, por lo que el objetivo principal de esta fase durante el procesamiento consistirá en analizar y extraer todas las sentencias que usualmente se encuentran diferenciadas por la puntuación gramática (e.g. punto y seguido, punto y a parte, exclamación, interrogación, etc.) con el propósito de proveer entidades granulares al mecanismo subsecuente y así facilitar la selección de características.

Como ejemplo a continuación se muestra un fragmento de la noticia publicada el 29 de Mayo del 2013 sobre la compra de la compañía Austriaca KPN por parte del millonario Carlos Slim.

*May 29 (Reuters) - Telekom Austria has rebutted reports it is planning further expansion in eastern Europe with major shareholder Carlos Slim after the Mexican billionaire took a 23 percent stake in Austria's leading telecoms operator last year. Analysts have been eyeing Telekom Austria and Dutch KPN for signs of major new strategic moves since Slim's America Movil Latin American telecoms group bought into the pair in a 4 billion-euro ($5 billion) foray into Europe last year [...].*

De acuerdo a lo enunciado anteriormente se determinará que dicho fragmento está integrado por dos sentencias que permitirán segregar su evaluación y que a su vez facilitaran la identificación de aquellas bien estructuradas.

Uno de los aspectos fundamentales al momento de identificar la función de la palabra dentro del contexto de la oración, es la de analizar cada uno de los elementos que la integran, de tal forma que el algoritmo de máxima entropía pueda inferir la función semántica de mayor probabilidad de ocurrencia en base a cada elemento del conjunto.

Para esto, el algoritmo transformará cada uno de los elementos sintácticos en cadenas que sean compatibles con las funciones de identificación semántica, buscando principalmente eliminar la puntuación que comúnmente se emplea para introducir pausas en el desarrollo del texto. Del ejemplo anterior a través de la implementación se puede obtener 42 y 44 elementos gramaticales respectivamente que serán empleados para evaluar si contribuyen o no semánticamente en el texto.

### 4.3.1 Stop words

Durante el desarrollo de cualquier texto informativo en el idioma Inglés se emplean recursos gramaticales que permiten realizar conexiones entre las diversas ideas que lo integran, de tal forma y de acuerdo al tema en el que se especializa su contenido, fungen como conectores o determinadores para los sustantivos y verbos.

*Shall, seem, should, besides, below, can, elsewhere, again, along, among, yet, why.*

No obstante, este conjunto de palabras puede ser enriquecida por nombres personales, días, meses, países o estados que simplemente no aportan algún valor durante la detección de sentimiento en documentos financieros.

Es importante mencionar que la definición de aportación se basa en lo que el presente algoritmo fundamenta para detectar el sentimiento y que este puede cambiar dependiendo del grado de especialización al que pueda ser dirigda la presente investigación.

Mc Donald [27] ha recopilado un bagaje de palabras del tipo *stop words* a largo del análisis de miles de documentos financieros para facilitar su identificación durante el análisis del texto y así incrementar la precisión al momento de determinar la función semántica de las palabras. A continuación se muestra el resultado de haber removido este tipo de palabras del texto informativo sobre la adquisición de KPN por Carlos Slim.

*29 ( ) - Telekom rebutted reports planning expansion eastern shareholder Slim billionaire took 23 percent stake 's leading telecoms operator. Analysts eyeing Telekom Dutch KPN signs strategic moves Slim's Movil telecoms group bought pair 4 billion-euro ($ 5) foray [...].*

## 4.4 Análisis semántico

Como su nombre lo indica, su principal área de interés es la de evaluar cada partícula con respecto a la posición en la que se encuentra dentro de la oración, de tal forma que se obtenga una clasificación sobre su aportación a la significancia del enunciado y así permita dotar de secuencias lógicas que sirvan a la selección de características determinar aquellas con un mayor valor semántico.

### 4.4.1 Part-of-Speech

Como se ha indicado en los capítulos anteriores para clasificar las funciones existentes de los elementos de la oración, se empleará el modelo probabilístico de máxima entropía cuyo aprendizaje asistido ha sido sometido al análisis de los documentos financieros de *Wall Street Journal* y *Brown corpus* con el propósito de incrementar su precisión. Cabe destacar que existen un conjunto de etiquetas (i.e. Penn Treebank) del cual este paradigma se valdrá para segregar la funciones gramaticales.

Dicho de esta forma podemos describir podemos denotar la salida resultante de aplicar POS al fragmento derivado del procesamiento de *Stop Words*.

*29_CD (_-LRB- )_-RRB- -_: Telekom_IN rebutted_VBN reports_NNS planning_VBG expansion_NN eastern_JJ shareholder_NN Slim_NNP billionaire_NN took_VBD 23_CD percent_NN stake_NN 's_POS leading_VBG telecoms_NNS operator_NN ._.*

*Analysts_NNS eyeing_VBG Telekom_NNP Dutch_NNP KPN_NNP signs_VBZ strategic_JJ moves_NNS Slim_NNP 's_POS Movil_NNP telecoms_NNS group_NN bought_VBD pair_NN 4_CD billion-euro_NN (_-LRB- $_$ 5_CD )_-RRB- foray_NN ._.*

Es de suma importancia resaltar que para fines de esta investigación se considerará que los elementos que desempeñen funciones semánticas, tales como modificadores, verbos y sus correspondientes adverbios, serán aquellos candidatos a ser considerados como eje principal para la selección de características.

### 4.4.2 Eliminación de ruido

Existen elementos dentro de la oración que no necesariamente son determinadores o conjunciones, sino caracteres o representaciones numéricas que refuerzan el sentido propio del texto, no obstante, para fines del presente algoritmo, no existe una particularidad o razón por la cual su permanencia sea mandatoria.

Este principio tiene como móvil reducir el procesamiento durante la categorización de las palabras y promover con mayor claridad la identificación de las características a seleccionar del texto. A continuación se muestra el texto resultante de remover las cadenas que representan ruido dentro del texto, es importante destacar que ciertos nombres personales persisten (e.g. Slim, Telekom, KPN) debido a que no son lo suficientemente comunes o ambiguos para colocarlos en el diccionario de *stop words*.

El siguiente fragmento muestra la salida esperada después de la eliminación del ruido asociado a la redacción del contenido.

*rebutted reports planning expansion eastern shareholder Slim billionaire took percent stake leading telecoms operator Analysts eyeing Telekom Dutch KPN signs strategic moves Slim Movil telecoms group bought pair billion-euro foray*

### 4.4.3 Selección de características

La selección de características es el medio para garantizar que las sentencias que se están evaluando cumplen con ciertos criterios de importancia o distinción entre el conjunto de oraciones disponibles, de tal forma, que una vez obtenidas éstas se podrá extraer aquellas partículas con un mayor poder explicativo. La resultante del mecanismo *Part-of-speech* al proporcionar un conjunto de funciones posibles, constituye la base primordial para identificar aquellas que posean dicha característica, de las cuales el modificador es el candidato idóneo por su capacidad de enfatizar o describir la situación actual de los verbos, sustantivos y adverbios que circundan al sujeto u objeto de la oración ha procesar.

Por lo tanto el marco de trabajó tendrá como principal objetivo identificar de las estructuras resultantes aquellas que posean las propiedades mencionadas para dotar al siguiente proceso de suficientes recursos para lograr una clasificación con mayor precisión.

De lo anterior podemos enunciar que dichos criterios estarán conformados por reglas semánticas o producciones gramaticales que reconocerán si contiene los elementos básicos para considerar que una oración representa un hecho concreto y bien estructurado.

$s \rightarrow nvs$
$nvs \rightarrow (advs)^*(ns)$
$nvs \rightarrow (advs)^*(vs)$
$advs \rightarrow <ADVERB>$
$ns \rightarrow (adjs)^* <NOUN> nvs$
$ns \rightarrow (adjs)^* <NOUN><EOF>$
$adjs \rightarrow <ADJECTIVE>$
$vs \rightarrow (<MODAL>)^* <VERB> nvs$

Las cadenas obtenidas después de procesar el ruido en cada una de las oraciones estarán conformadas por las siguientes secuencias de caracteres, las cuales serán empleadas para la generación de una matriz que identificará por medio de las probabilidades contenidas en los diccionarios aquellas que tengan una continuidad y por ende las que determinan el sentido general del texto (conceptualizado como tono).

*VBN NNS VBG NN JJ NN NNP NN VBD NN NN VBG NNS NN*

*NNS VBG NNP NNP NNP VBZ JJ NNS NNP NNP NNS NN VBD NN NN NN*

### 4.4.4 Raíz morfológica

Con el objetivo de reducir el número de comparaciones durante la detección del tono en las cadenas resultantes y por ende incrementar el desempeño del algoritmo, se ha sometido cada palabra a la reducción de su raíz morfológica, de tal forma que se obtenga la siguiente relación (Tabla 5).

| ***Word*** | ***Stem*** | ***Tag*** | ***Word*** | ***Stem*** | ***Tag*** |
|---|---|---|---|---|---|
| *rebutted* | *rebut* | *VBN* | *eyeing* | *ey* | *VBG* |
| *reports* | *report* | *NNS* | *Telekom* | *telekom* | *NNP* |
| *planning* | *plan* | *VBG* | *Dutch* | *dutch* | *NNP* |
| *expansion* | *expans* | *NN* | *KPN* | *kpn* | *NNP* |
| *eastern* | *eastern* | *JJ* | *signs* | *sign* | *VBZ* |
| *shareholder* | *sharehold* | *NN* | *strategic* | *strateg* | *JJ* |
| *Slim* | *slim* | *NNP* | *moves* | *move* | *NNS* |
| *billionaire* | *billionair* | *NN* | *Slim* | *slim* | *NNP* |
| *took* | *took* | *VBD* | *Movil* | *movil* | *NNP* |
| *percent* | *percent* | *NN* | *telecoms* | *telecom* | *NNS* |
| *stake* | *stake* | *NN* | *group* | *group* | *NN* |
| *leading* | *lead* | *VBG* | *bought* | *bought* | *VBD* |

| *Word* | *Stem* | *Tag* | *Word* | *Stem* | *Tag* |
|---|---|---|---|---|---|
| *telecoms* | *telecom* | *NNS* | *pair* | *pair* | *NN* |
| *operator* | *oper* | *NN* | *billion-euro* | *euro* | *NN* |
| *Analysts* | *ana* | *NNS* | *foray* | *forai* | *NN* |

Tabla 5 – Reducción morfológica de las palabras resultantes de la selección de características

Cabe destacar que las comparaciones subsecuentes durante la siguiente etapa de procesamiento consideran que todos los diccionarios deberán ser reducidas a su raíz morfológica para que dichas operaciones se ejecuten satisfactoriamente.

## 4.5 Implementación

Para representar la forma en que el análisis gramatical y el semántico se han implementado dentro del marco de trabajo, emplearemos entidades descriptivas del lenguaje JAVA para generalizar la forma en que estas colaboran y logar la plasmar la conceptualización del conocimiento previo a la identificación del tono inmerso en el texto.

Como se ha plasmado durante el desarrollo del presente marco de trabajo, cada nivel procesamiento depende intrínsecamente del resultado del anterior, por lo que la arquitectura deberá expresar exactamente la misma necesidad.

En el siguiente diagrama (Diagrama 4.5.1) denotaremos que el análisis gramatical será responsabilidad de la clase *SentenceAnalyzer*, y de la cual *SentimentFeatureSelection* actuará como intermediario para procesar todo lo correspondiente a la evaluación semántica.

El resultado final de ambos procesos será expuesto a la implementación responsable de procesar el tono para finalmente obtener una conclusión (i.e. positiva, negativa o positiva).

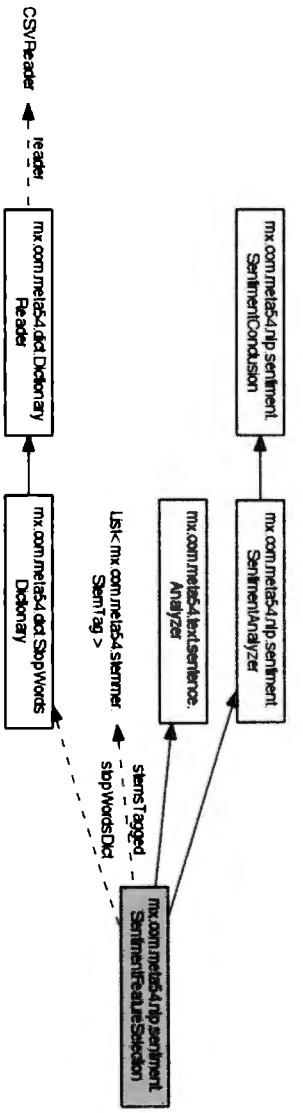



**Diagrama 4.5.1 Representación del análisis gramatical y semántico**

Derivado a que la interfaz gráfica requiere de obtener dicho resultado, se construyó una clase abstracta *SentimentConclusion* que permitiera exponer el resultado a cualquier objeto que así lo requiriera. En el siguiente diagrama 4.8.2 podemos encontrar la interacción del API enlistado que la clase *SentimentChart* requiere.

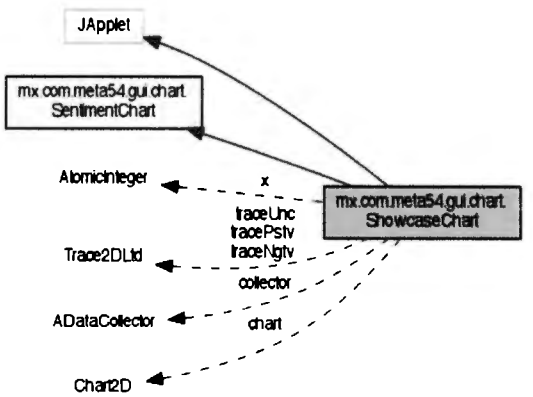



Diagrama 4.5.2 Componentes de la interfaz gráfica

## 4.6 Categorización

La categorización será la expresión en cualquiera de sus tres variantes (i.e. positiva, negativa o incertidumbre) del resultado obtenido a través de la detección del tono y en colaboración con el clasificador Bayesiano propuesto por la Universidad de Stanford. Por otra parte, la presente implementación busca introducir una perspectiva innovadora que permita reforzar los elementos que serán enviados al clasificador, por medio de la identificación probabilística de los términos más representativos con respecto a un conjunto de posibles categorías definidas en diccionarios a fines a la Psicología.

El argumento que precede a este enfoque está orientado al momento de transmitir una idea en particular se realiza bajo un razonamiento orientado a la agrupación de las ideas y en las cuales los sentimientos son expresados. A fin de proveer una solución que permita abstraer esta lógica, se ha fundamentado bajo una método de selección estadística la detección de probabilidades contiguas que dan a lución a dicha agrupación.

### 4.6.1 Detección de tono

El resultado final de la selección de características representa la información imprescindible dentro del conjunto de hechos que están explícitos en el contenido, de tal forma que el análisis a partir de esta fase buscará identificar qué oración o conjunto de oraciones poseen una influencia decisiva dentro del texto. Para este propósito, a cada elemento se le asignará una ponderación explicativa construida a partir del diccionario Harvard IV-4 [25], la cual contempla diversas categorías basadas en el ámbito psicológico, que fungirán como el umbral para determinar la relación con el desarrollo del contenido y el sentimiento que se desea plasmar.

Por cada una de las palabras que estén catalogadas como modificadores ya sea de un sustantivo o de un verbo (derivado de *Part-of-speech* y de la selección de características), se obtendrán las categorías que se encuentren registradas en el diccionario y se ordenarán conforme al desarrollo del texto (Tabla 6).

Posteriormente se comparará cada uno de los elementos que fueron resultante de la selección de características con las categorías definidas por los modificadores, con el fin de cuantificar el número de coincidencias entre estas. A su vez se contabilizará el número de coincidencias entre cada una de las categorías y se determinará la probabilidad de ocurrencia correspondiente.

Subsecuentemente el algoritmo iniciará con la detección del tono iniciando con las categorías de mayor probabilidad y se ira desplazando conforme existe una continuidad entre ellas, si esta no existe se volverá a considerar la de mayor probabilidad y así sucesivamente (Tabla 7).

| | Categoría A | Categoría B | Categoría C | Categoría D |
|---|---|---|---|---|
| Sustantivo | X | | | |
| Modificador | | X | | |
| Sustantivo | | X | X | |
| Modificador | | | | X |

Tabla 6 – Ejemplo de clasificación función de la palabra vs categoría de Harvard IV-4

*Negativ (1), Hostile (2), PowCon (3), PowTot (4), Strong (5), Solve (6), Means (7), Power (8), PowOth (9), Causal (10), Positiv (11), Virtue (12), Ovrst (13), EVAL (14), PosAff (15), Persist (16), WlbPhys (17), WlbTot (18), Perceiv (19), EnlGain (20), EnlTot (21), EndsLw (22).*

| ***Word*** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **21** | **22** |
|---|---|---|---|---|---|---|---|---|---|
| *rebutted* | 0.4961 | 0.1804 | 0.0494 | 0.2741 | 0 | 0 | 0 | 0 | 0 |
| *reports* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.8512 | 0 |
| *planning* | 0 | 0 | 0 | 0 | 0.8146 | 0.0809 | 0.1045 | 0 | 0 |
| *expansion* | 0 | 0 | 0 | 0 | 0.8757 | 0 | 0 | 0 | 0.1243 |
| *Slim* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *stake* | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| *leading* | 0 | 0.082 | 0.0225 | 0.1247 | 0.1873 | 0 | 0.024 | 0 | 0 |
| *operator* | 0 | 0 | 0 | 0 | 0.5149 | 0 | 0.0661 | 0.226 | 0 |
| *eyeing* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.7118 | 0 |
| *signs* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| *strategic* | 0 | 0 | 0 | 0 | 0.7873 | 0 | 0.101 | 0 | 0.1118 |
| *moves* | 0 | 0 | 0 | 0.3022 | 0.454 | 0 | 0 | 0 | 0 |
| *Slim* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *group* | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

Tabla 7 – Ejemplo de probabilidades asignadas por categoría y detección de tono

Para fines de presentación se han eliminado aquellas columnas que no tienen continuidad y aquellas cuya incidencia fue nula, ya que las primeras implican que no existen elementos que tengan una incidencia entre estas categorías; las últimas representan la inconsistencia entre la etiquetas generada por POS y la categoría que se tiene registrada en el diccionario Harvard IV-4 [25], por lo que en ambos casos las palabras que cumplan con alguno de estos escenarios no serán consideradas en el resto del procesamiento.

El conjunto de probabilidades calculadas así como las categorías resultantes representarán el tono del texto extraído y por ende, describirá los elementos que predominan con el conjunto de sentimientos inherentes al contenido, de tal forma que puedan ser empleados en sistemas que requieran de una representación abstracta del texto evaluado (e.g. ontologías).

No obstante, el resultado de esta interpretación puede ser introducida en trabajos futuros para corroborar la clasificación que será realizada en la siguiente etapa del procesamiento.

*+[ 0.2143*Negativ +0.0714*PowTot +0.2857*Strong +0.2143*Means +0.2143*EnlTot ]-*

### 4.6.2 Interpretación y conclusiones

A partir de la categorización podemos emplear mecanismos que nos permitan obtener por medio de la inferencia lógica, un espacio vectorial que represente la clasificación de los elementos que han sido seleccionados a lo largo del presente algoritmo, de tal forma, que obtengamos probabilísticamente una interpretación del texto evaluado en base a un conjunto de diccionarios que permitan el aprendizaje y evaluación de cualquier contenido.

Para dicho fin emplearemos un mecanismo conocido como clasificador Bayesiano, cuya implementación fue concebida en las inmediaciones de la universidad de Standford [28] y que recientemente ha sido puesto a disposición del publico en general para su uso. Como todo mecanismo de aprendizaje requiere un conjunto de evidencia que le permita razonar las secuencias de inferencia y un conjunto de datos sobre el cual validar dicho aprendizaje.

El conjunto de palabras que estadísticamente son empleadas en los contextos financieros como positivas y negativas, han sido recopiladas por Mc Donald [27] para facilitar a cualquier mecanismo la segmentación clara de estas, de tal forma que serán empleadas para los fines de esta investigación como un medio para el entrenamiento asistido.

De acuerdo a lo obtenido en la detección del tono, las siguientes palabras serán sometidas al clasificador para determinar la interpretación con mayor preponderancia.

*rebutted reports planning expansion stake leading operator eyeing signs strategic moves group*

| **Interpretación** | **Probabilidad** |
|---|---|
| Positiva | 0.500 |
| Negativa | 0.250 |
| Incertidumbre | 0.249 |

Tabla 8 – Categorización y probabilidades asociadas

De a cuerdo los resultados obtenidos se puede concluir que el texto tiene una interpretación más orientada a lo positivo. Como puede percibirse en la representación abstracta del texto, el elementos negativo tiene una probabilidad de 0.2143 en contraste con los 0.25 determinados con el clasificador, lo cual implica que la aproximación tiene un delta de 0.0357 y por lo tanto podemos confirmar que la orientación del texto es completamente positiva.

## 4.7 Implementación

En el siguiente conjunto de diagramas se esquematiza la forma en que es introducido el proceso de categorización mediante el apoyo de clases como *SentimentFeatureSelection* y *SentenceAnalyzer,* donde se efectúa el análisis gramatical y semántico del texto evaluado. La entidad que orquestará los resultados de ambos procesos (i.e. *SentimentProcess*) es inyectada a través de la tecnología de Spring dentro del espacio de trabajo de la clase *SentimentFeatureSelection,* para iniciará la detección del tono descrito con anterioridad (Diagrama 4.7.1).

El orquestador *SentimentProcess* tiene una relación intrínseca con el paradigma de la detección de tono tal y como puede observarse por la estructura jerárquica de las clases *Classification, SentimentTone y Sense* (Diagrama 4.7.2). La definición de las categorías para la matriz de detección de tono se lleva acabo durante la invocación a la implementación de *Sense,* de tal forma que la instancia de *SentimentTone* pueda extraer el fragmento que alimentará el clasificador Bayesiano y así logremos obtener las categorías buscadas (Diagrama 4.7.2).

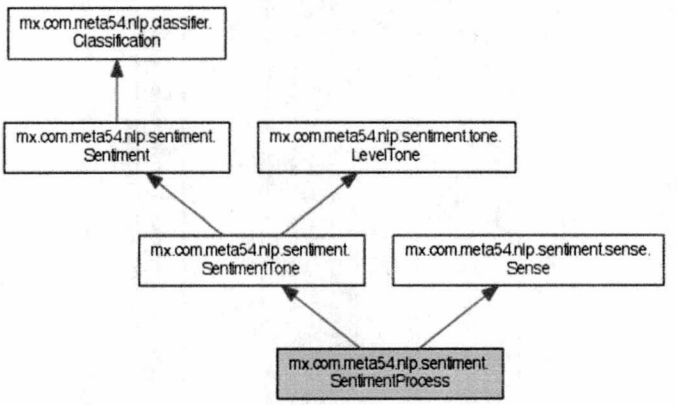



Diagrama 4.7.2 Componentes de la interfaz gráfica

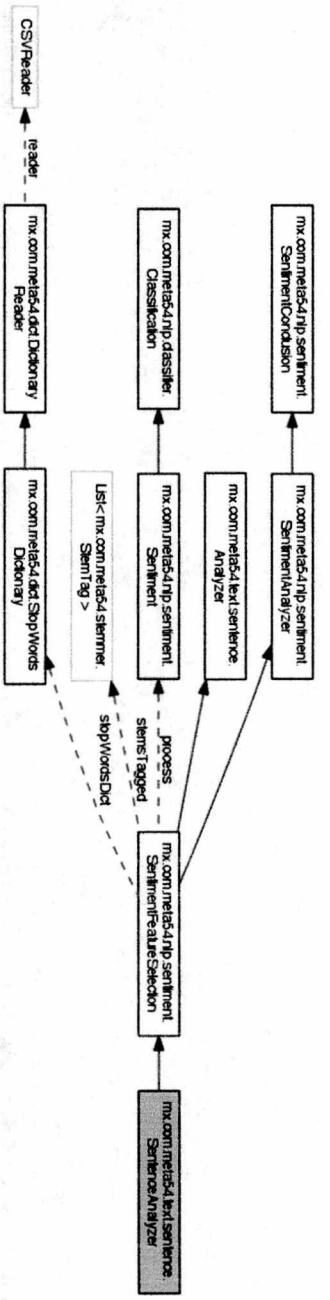



Diagrama 4.7.1 Inyección del proceso orquestador

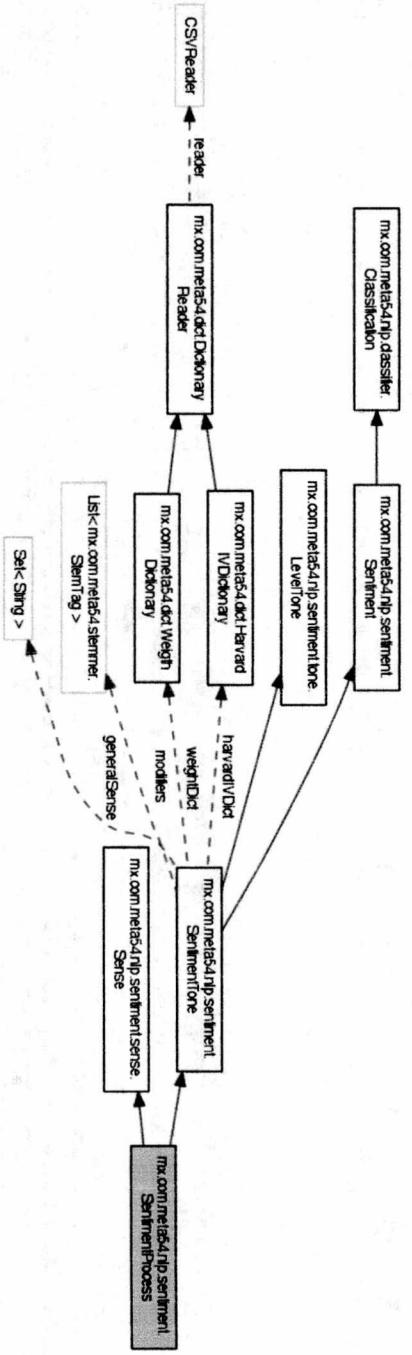



Diagrama 4.7.3 Detección del tono y clasificación

# CAPÍTULO 5

# Resultados

## 5.1 Caso práctico

A continuación se muestran los resultados obtenidos de aplicar la detección del sentimiento en el conjunto de noticias financieras comprendidas del 13 de Julio del 2011 al 12 de Septiembre del 2013, sobre dos de las emisoras con mayor influencia bursátil dentro de la Bolsa Mexicana de Valores (e.g. América Móvil SAB de CV y Wal Mart de México SAB de CV).

Con el propósito de comparar el comportamiento del precio con respecto a las noticias emitidas durante dicho ciclo, se ha consultado por medio del servicio electrónico de Reuters (www.reuters.com) los precios de apertura y cierre de ambas emisiones bursátiles. Es importante considerar que esta mecánica sigue los preceptos que Fama expuso sobre la integración del efecto de la información al estar disponible durante la compra y venta de acciones, por lo que será un excelente punto de referencia para el presente modelo.

De acuerdo al resultado obtenido la evaluación contempla tres categorizaciones de índole positivo, negativo e incertidumbre, por lo cual éstas serán los ejes que permitirán contribuir en la definición de las acciones a tomar de acuerdo a las condiciones del mercado.

En la siguiente gráfica puede denotarse el comportamiento de los precios del cierre de América Móvil durante el periodo del 19 de Julio del 2011 al 12 de Septiembre del 2013 (Figura 5.1.1), los cuales servirán de guía para comparar las categorías obtenidas por el algoritmo con respecto a la fecha de publicación de cada artículo analizado.

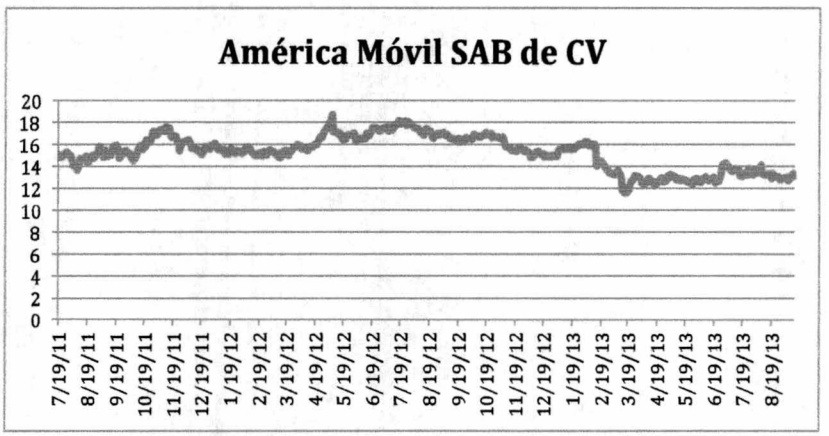



Figura 5.1.1 – Precios del 19 de Julio del 2011 al 12 de Septiembre del 2013

A continuación se presenta las categorías resultantes del análisis del sentimiento, las cuales son graficadas con respecto a la probabilidad de incidencia y en base a la fecha de publicación de cada artículo (Figura 5.1.2).

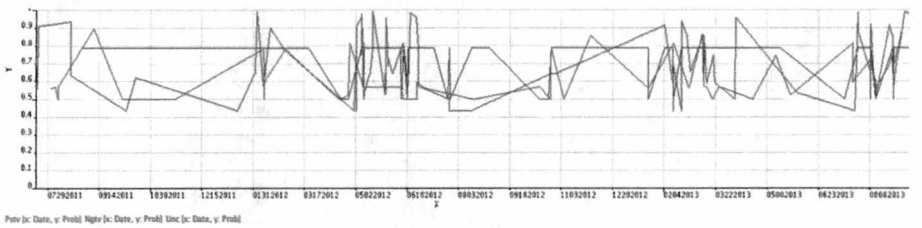


Figura 5.1.2 – Interpretación del sentimiento de la emisora de América Móvil

Contrastando ambas gráficas se puede denotar de forma agrupada la siguiente correlación entre los precios y el sentimiento obtenido (Tabla 5.1.1).

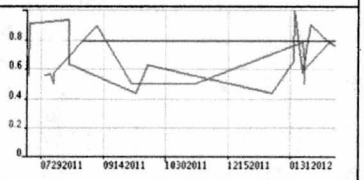
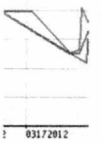
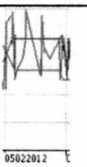
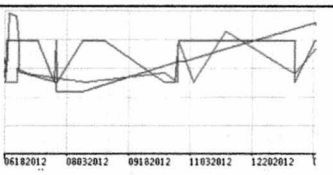
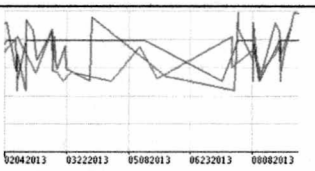
| Inicio | Fin | Meses | Precio | Sentimiento | Gráfica |
|---|---|---|---|---|---|
| 7/19/11 | 3/2/12 | 7.4 | $15.02 - $15.34 | Positivo |  |
| 3/3/12 | 5/1/12 | 2 | $15.34 - $17.39 | Negativo |  |
| 5/2/12 | 6/18/12 | 1.5 | $17.58 - $17.18 | Incertidumbre |  |
| 6/19/12 | 2/4/13 | 7.5 | $17.47 - $16.19 | Negativo |  |
| 2/5/13 | 9/12/13 | 7.2 | $15.87 – $13.11 | Incertidumbre |  |

Tabla 5.1.1 – Análisis del sentimiento para la toma de decisiones de la emisora de América Móvil

Derivado de lo anterior se puede deducir que a mayor número de noticias evaluadas sin considerar el intervalo de tiempo se tendrá una mayor precisión dando como resultado una mejora en la predicción de la tendencia del precio. Este argumento se corrobora partiendo del resultado negativo obtenido durante el periodo de 3/3/12 al 5/1/12, en donde los eventos son sumamente escasos y el mercado tiene una reacción positiva ante factores no cuantificados en la presente evaluación.

En lo que se refiere a los momentos cuya categoría se encuentra identificada como incertidumbre, se puede notar que el grupo de noticias con mayor influencia y próximas a esta, constituyen un determinante en el precio, de tal forma, que al término de los periodos del 6/18/12 y 9/12/13 la interacción con señales negativas anuncian los decrementos inminentes pese a los anuncios positivos previos a dichas eventualidades.

El número de apariciones de noticias positivas aunque representan numéricamente una mayor cantidad con respecto a las negativas, la incertidumbre genera la sensación de una especulación mayor, por lo que los inversionistas presentarán aversión a mantener la posición y por ende el precio del activo sufrirá caídas hasta que mejoren la perspectiva de las variables intrínsecas a la evaluación del instrumento.

Un aspecto determinante del análisis del sentimiento radica en la cantidad de información disponible lo cual está intrínsecamente relacionado a la liquidez que los activos representan en el mercado. La liquidez se define como la cualidad de convertir en dinero, de forma inmediata y sin pérdida significativa del valor de un conjunto de acciones; por lo tanto, las emisoras que no tengan esta propiedad no podrán ser sometidas a la evaluación por el presente algoritmo.

En la siguiente gráfica se presentan los precios de la compañía Wal Mart de México SAB de CV durante el periodo adscrito del 7 de Julio del 2011 al 10 de Septiembre del 2013.


Figura 5.1.3 – Precios del 7 de Julio del 2011 al 10 de Septiembre del 2013

El sentimiento obtenido para dicha emisora es representado en la siguiente gráfica, en la cual podemos apreciar que existe una influencia negativa predominante y reforzada por la presencia de incertidumbre entorno a ésta, lo que implica que las señales positivas son minimizadas y su aportación escaza.

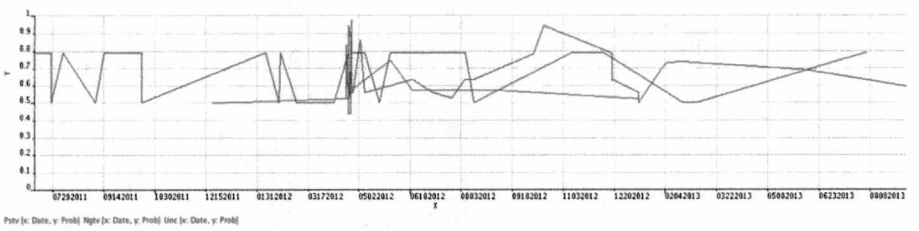


Figura 5.1.4 – Interpretación del sentimiento de la emisora de Walmart de México

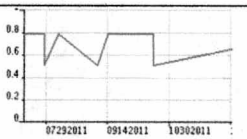
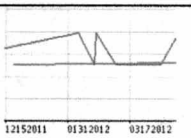
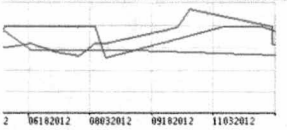
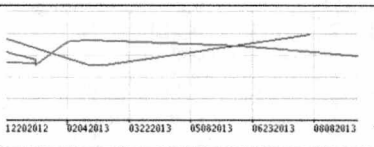
| Inicio | Fin | Meses | Precio | Sentimiento | Gráfica |
|---|---|---|---|---|---|
| 7/7/11 | 12/20/11 | 5.4 | $33.74 – $32.71 | Negativo |  |
| 12/21/11 | 4/2/12 | 3.3 | $33.12 - $43.95 | Negativo |  |
| 4/3/12 | 5/31/12 | 1.9 | $44.00 - $34.4 | Incertidumbre |  |
| 6/1/12 | 12/18/12 | 6.5 | $35.48 - $42.43 | Incertidumbre |  |
| 12/19/12 | 09/10/13 | 9.3 | $41.06 - $34.94 | Incertidumbre |  |

Tabla 5.1.2 – Análisis del sentimiento para la toma de decisiones de la emisora de Walmart de México

Como se ha mencionado anteriormente la incertidumbre rige la mayoría de las eventualidades durante las sesiones informativas para la empresa Wal Mart de México (Tabla 5.1.2), no obstante es claro que el dominio de textos negativos durante el periodo del 7/7/11 al 12/20/11 prevé una caída en las cotizaciones para dicha emisora. Aunque el precio cambia abruptamente en el siguiente trimestre y el algoritmo no predice dicho impacto posiblemente por la falta de información, puede observarse que durante la incertidumbre del periodo 4/3/12 al 5/31/12 el efecto negativo permanece y pone fin a la especulación sobre la evaluación de la emisora.

Los eventos que circundan al periodo del 6/1/12 al 12/18/12 muestran la predominancia de aspectos negativos con incertidumbre sobre los eventos positivos, aunque estos últimos no fueron estimados con exactitud para predecir el alza en el precio que se tiene registrada para finales de dicho periodo, se mantuvo la tendencia para visualizar la baja que se presentaría entre los periodos del 12/19/12 al 09/10/13.

De ambos análisis se concluye que el algoritmo al no determinar el precio sugerido para la compra y venta de acciones, logra establecer una relación que puede ser empleada como herramienta de soporte para determinar si la posición debe permanecer larga o corta. No obstante, el funcionamiento puede ser mejorado si se integran diversas fuentes de información o bien se adoptara un modelo que permitiese consolidar en una sola curva la influencia de las 3 categorías, más un modelo económico que introduzca esta relación en el cálculo de precios, ya sea mediante un portafolio o bien por medio de estrategias algorítmicas empleadas bajo los criterios o señales del mercado accionario.

Las capacidades de aplicación abordadas en la presente investigación han sido acotadas sólo en los contenidos del ámbito financiero, debido al uso de diccionarios que se han empleado para seleccionar las características más preponderantes del texto, así como la especialización de la infraestructura algorítmica y probabilística para lograr definir la categoría más adecuada para el texto analizado.

Durante la implementación se han aplicado diseños arquitectónicos que permitan la fácil sustitución de las entradas para brindar una escalabilidad viable, en las futuras investigaciones o aplicaciones prácticas.

## 5.2 Diccionarios

La siguiente distribución de componentes permite brindar al lector las entradas que requiere el sistema para su funcionamiento, de tal forma que puedan ser modificadas a conveniencia del área de aplicación o bien para cualquier otro propósito académico (Figura 5.2.1).

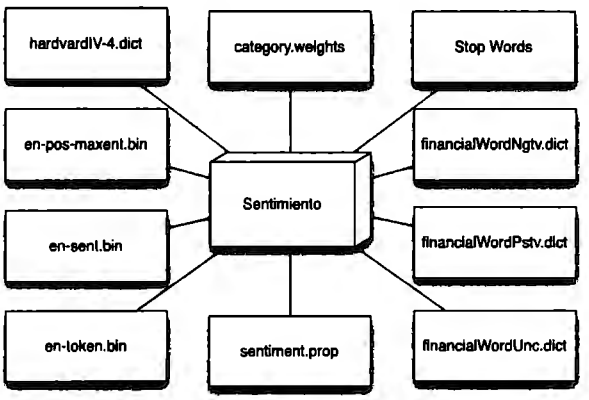



Figura 5.2.1 – Principales diccionarios para el funcionamiento del marco de trabajo

El mecanismo para el entrenamiento asistido del clasificador Bayesiano consta de los siguientes archivos de configuración.

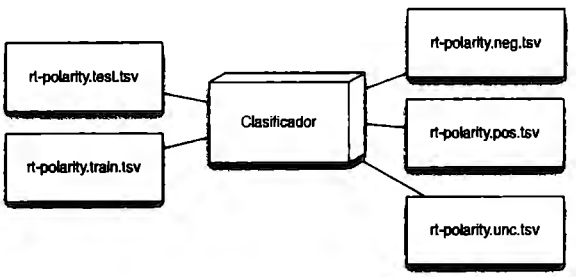



Figura 5.2.2 – Archivos empleados para el entrenamiento asisto del clasificador de Stanford

# CAPÍTULO 6

# Conclusiones y Trabajo Futuro

## 6.1 Conclusiones

Los resultados obtenidos en el presente estudio demuestran la forma en que numéricamente se ha logrado extraer y representar el conocimiento cualitativo de textos de índole financiera, los cuales permiten categorizar el sentimiento intrínseco al conjunto de conceptos evaluados y por lo tanto podrán ser empleados en modelos cuya definición permita insertar variables de esta naturaleza.

Es importante resaltar que esta incursión representa una iniciativa para impulsar tanto el área de investigación mexicana sobre el análisis sistemático del contenido, así como los fundamentos que requiere la generación de nuevos productos o servicios orientados a satisfacer aquellas áreas del conocimiento humano que requiera la evaluación de grandes volúmenes de información, así como la predicción de tendencias que permitan segregar áreas de oportunidad para cada uno de los campos de aplicación.

En materia de avance científico podemos denotar que la selección de características no solo constituye un papel preponderante dentro de la detección de emociones en el campo de la abstracción semántica, sino que éste debe ser perfeccionado de tal forma que permita lograr asociar aquellos elementos que sean las partículas distintivas del texto procesado.

En dicha fase es donde la presente investigación busca introducir un modelo innovador para profundizar y dotar de mayor precisión la forma en que se identifican aquellos elementos más representativos, aludiendo a una secuencia de tono determinada por la continuidad de que aquellas ocurrencias que describen los elementos con respecto a sus posibles definiciones.

Finalmente y sujeto a los preceptos de la teoría del conocimiento, con lo anterior se consolida el objetivo de ofrecer una herramienta que permita representar los conceptos afines al texto y de los cuales se pueden desprender un conjunto de evaluaciones tangibles que logren apoyar la toma de decisiones en el ámbito de aplicación.

## 6.2 Trabajo Futuro

Aunque los conceptos seleccionados poseen cualidades distintivas con respecto al resto del contenido del texto, estos solo describen puntualmente un único evento en línea del tiempo de un grupo de acontecimientos, es decir, si cotejamos esto en el ámbito financiero no es factible establecer que las variaciones del precio de una emisora están sujetos a efectos de un sola variable, por lo que el modelo expuesto deberá satisfacer las necesidades en tiempo real y de forma dinámica incluir la definición de nuevos términos que sean adoptados en base al área aplicación, así como un grupo de modelos que permitan asociar los conceptos y por ende consolidarlos en un único resultado.

En consecuencia, el procesamiento que se refiere a la extracción de datos puede ser robustecido mediante un grafo que le permita identificar las fuentes de información afines al grupo de evaluación, posteriormente, la inmersión de nuevas términos gramaticales podrá ser acrecentada si estos son definidos constantemente por los usuarios finales.

En materia de mejora sobre la relación puntual de cada evento, se pueden introducir algún modelo que dado un comportamiento ya histórico de un conjunto de datos, pueda nuevamente evaluar las condiciones que le han llevado a una determinada conclusión para ratificarla o bien para realizar los ajustes necesarios dentro de los clasificadores.

Cabe destacar que la abstracción de los conceptos realizados pueden emplearse en sistemas cuyo objetivo sea la representación del conocimiento asociado y del cual pueden desprenderse diversas actividades como *web semántica* o la elaboración de abstracciones más sofisticadas como lo son las ontologías. Más aún, el marco de trabajo implementado dota de mayores posibilidades de comprensión a los futuros proyectos que sean derivados de este, al representar por medio de una ecuación, los principales componentes emocionales que se encuentran inmersos en el texto, lo cual facilitará sustentar las bases en las que estructuras semánticas obtendrán su conocimiento.

Finalmente el lenguaje con el que se ha ejecutado y planeado la arquitectura expuesta se basa en el dominio de aplicación del Inglés, al ser uno de los más usados para la expresión y propagación de noticias de índole financiera, por lo que si se tuviera la necesidad de cambiar dicha naturaleza, los principales elementos que se deben considerar es el entrenar por medio de NLP un nuevo grupo de textos de tal forma que sea lo suficientemente robusto para que su precisión sea del 99% al momento de determinar las funciones semánticas, posteriormente los diccionarios del área de aplicación deben ser divididos en la misma gama que actualmente se presenta (positivo, negativo e incertidumbre) y especializados en el área que se desea evaluar, así como aquellas partículas cuya a portación gramatical sea prácticamente nula. De esta forma se podrá consolidar y delimitar de manera práctica la aplicación de esta perspectiva para lograr futuros desarrollos y por ende acrecentar el acervo científico en torno a esta rúbrica de la semántica, que actualmente se encuentra en plena evolución.

# Bibliografía

[1] Siering, M. "'Boom' or 'Ruin'–Does It Make a Difference? Using Text Mining and Sentiment Analysis to Support Intraday Investment Decisions." In 2012 45th Hawaii International Conference on System Science (HICSS), 1050 –1059, 2012.

[2] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work", The Journal of Finance (25, 2), 1970, pp. 383–417.

[3] J. Muntermann and A. Guettler, "Intraday stock price effects of ad hoc disclosures: the German case", Journal of International Financial Markets, Institutions and Money (17, 1), 2007, pp. 1–24.

[4] S. S. Groth and J. Muntermann, "Supporting Investment Management Processes with Machine Learning Techniques", Proc. of the 9th Internationale Tagung Wirtschaftsinformatik, vol. 2, Vienna, Austria, 2009, pp. 275–284.

[5] "From Darkness, Dawn." The Economist. Accessed April 10, 2013. http://www.economist.com/news/special-report/21566773-after-years-underachievement-and-rising-violence-mexico-last-beginning.

[6] Chao, Sam, Fai Wong, and C.C. Martins. "Data Mining Model in Analyzing Portuguese Studies as the Second Language Acquisition." In 2010 International Conference on Machine Learning and Cybernetics (ICMLC), 1:427 –432, 2010.

[7] Hagenau, M., M. Liebmann, M. Hedwig, and D. Neumann. "Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Specific Features." In 2012 45th Hawaii International Conference on System Science (HICSS), 1040 –1049, 2012.

[8] Goth, G. (2012). Digging Deeper into Text Mining: Academics and Agencies Look Toward Unstructured Data. IEEE Internet Computing, 16(1), 7 –9. doi:10.1109/MIC.2012.6

[9] Xinli, W. (2010). Corporate Financial Warning Model Based on PSO and SVM. In 2010 2nd International Conference on Information Engineering and Computer Science (ICIECS) (pp. 1 –5). doi:10.1109/ICIECS.2010.5677775

[10] Zhong, Y.X. "Knowledge Theory and Information-knowledge-intelligence Trinity." In Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02, 1:130 – 133 vol.1, 2002.

[11] Johnson, Barry. Algorithmic Trading & DMA: An Introduction to Direct Access Trading Strategies. First Ed. 4Myeloma Press, 2010.

[12] Nithya, K., P.C.D. Kalaivaani, and R. Thangarajan. "An Enhanced Data Mining Model for Text Classification." In 2012 International Conference on Computing, Communication and Applications (ICCCA), 1 –4, 2012.

[13] Thakur, R., N.S. Chaudhari, S. Jain, and R. Singhai. "Information Extraction from Semi-structured and Un-structured Documents Using Probabilistic Context Free Grammar Inference." In 2012 International Conference on Information Retrieval Knowledge Management (CAMP), 273 –276, 2012.

[14] Devasena, C.L., and M. Hemalatha. "Automatic Text Categorization and Summarization Using Rule Reduction." In 2012 International Conference on Advances in Engineering, Science and Management (ICAESM), 594 –598, 2012.

[15] Liu, Hui, and Chuanyan Zhang. "A Web Entity Activity Recognition Approach Based on K-nearest Neighbors Classifier." In 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), 848 –852, 2012.

[16] Sakurai, S., K. Makino, and S. Matsumoto. "A Discovery Method of Trend Rules from Complex Sequential Data." In 2012 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA), 950 –955, 2012.

[17] Preethi, T., K. Nirmala Devi, and V. Murali Bhaskaran. "A Semantic Enhanced Approach for Online Hotspot Forums Detection." In 2012 International Conference on Recent Trends In Information Technology (ICRTIT), 497 –501, 2012.

[18] Cui, X., Mueller, F. C. S. U., Zhang, Y., & Potok, T. E. (2009). GPU-Accelerated Text Mining. Retrieved from http://www.osti.gov/energycitations/servlets/purl/962625-jb99NQ/

[19] Smith, F. (1975). Comprehension and learning : a conceptual framework for teachers. New York: Holt, Rinehart and Winston.

[20] Ayral, H., and S. Yavuz. "An Automated Domain Specific Stop Word Generation Method for Natural Language Text Classification." In 2011 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 500 –503, 2011.

[21] Adwait Ratnaparkhi. "A Maximum Entropy Model for Part-Of-Speech Tagging". In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 133-142, May 17-18, University of Pennsylvania, 1996.

[22] Heyan, Huang, and Zhang Xiaofei. "Part-of-speech Tagger Based on Maximum Entropy Model." In 2nd IEEE International Conference on Computer Science and Information Technology, 2009. ICCSIT 2009, 26 -29, 2009. doi:10.1109/ICCSIT.2009.5234787.

[23] Medina Herrera, L. M., & Díaz Hernández J.B. (2011, June). "Journal of Management, Finance and Economics", Volume 5(Issue 1), 23-32.

[24] Jones, K. & Willett, P. (1997). "Readings in information retrieval". San Francisco, Calif: Morgan Kaufman.

[25] Philip Stone. (September 12, 2002). Harvard IV-4 . In Inquirer Home Page. Retrieved September 23, 2013, from http://www.wjh.harvard.edu/~inquirer/Home.html.

[26] Mitra, G. & Mitra, L. (2011). "The handbook of news analytics in finance". Hoboken, N.J. Chichester: Wiley.

[27] McDonald, B. (n.d.). Bill McDonald's Word Lists Page. University of Notre Dame. Retrieved October 5, 2013, from http://www3.nd.edu/~mcdonald/Word_Lists.html

[28] The Stanford NLP (Natural Language Processing) Group. (n.d.). The Stanford NLP (Natural Language Processing) Group. Retrieved October 7, 2013, from http://nlp.stanford.edu/software/classifier.shtml

[30] Penn Treebank Project. N.p., n.d. Web. 19 Nov. 2013. <http://www.cis.upenn.edu/~treebank/>.