PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

MODELO PROBABILÍSTICO PARA ESTIMAR LA CAPACIDAD DE INFLUENCIA DE USUARIOS EN TWITTER

REYNALDO ANDRÉS HERRERA GONZÁLEZ

INFORME FINAL DE PROYECTO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO CIVIL EN INFORMÁTICA

JULIO, 2012

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

MODELO PROBABILÍSTICO PARA ESTIMAR LA CAPACIDAD DE INFLUENCIA DE USUARIOS EN TWITTER

REYNALDO ANDRÉS HERRERA GONZÁLEZ

Profesor Guía: **Rodrigo Alfaro Arancibia**

Carrera: **Ingeniería Civil en Informática**

JULIO, 2012

A mis padres, mis mayores influyentes.
Y a todos los que de alguna forma u otra
contribuyeron con un granito de arena
para que yo pudiese llegar a estas instancias.

Resumen

El estudio y entendimiento de la influencia en Twitter puede entregar un apoyo al desarrollo de campañas virales más efectivas para esta red social. Saber cómo seleccionar a los usuarios más influyentes puede ser de gran ayuda para esparcir un mensaje importante. Este estudio se ha centrado en investigar sobre las métricas y métodos que han establecido otros autores para determinar la influencia entre usuarios. Además se han propuesto soluciones a algunos problemas que no se han abordado en las publicaciones vistas. Más específicamente, se aplicaron algunas métricas que dictan el comportamiento de los usuarios en un modelo probabilístico de influencia, el cual es un modelo para redes sociales en general, y que ha sido adaptado para Twitter. Este modelo adaptado ha sido complementado con algunas propiedades de la teoría de grafos y otras disciplinas para obtener conclusiones valiosas en cuanto al estudio de la influencia en Twitter. El modelo se ha aplicado en un contexto determinado y se ha analizado el rendimiento comparándolo con Klout, una popular métrica de influencia online. Bajo ciertas condiciones el modelo propuesto logra imitar a Klout con una precisión considerable. Sin embargo, la principal ventaja de este modelo es que este es altamente configurable y se puede aplicar en diversos contextos.

Palabras Clave: Twitter, influencia, campañas virales, estadística y probabilidad, teoría de grafos, clasificación automática de textos, minería de datos.

Abstract

Understanding the role of influence on Twitter may help to contribute to the development of more effective viral marketing campaigns. Selecting influential users may help considerably in the propagation of an important message. This research has focused on the understanding of metrics and methods proposed by other authors to measure user influence on social networks. On the other hand, problems that haven't been addressed on the works of these authors have inspired some solution proposals on this research. To be more precise, metrics related to the behavior of users have been applied to an influence probability model, which is a model adapted from social networks in general, to Twitter. This model has been complemented with some properties of graph-based theory and other disciplines in order to obtain valuable conclusions regarding the study of influence on Twitter. The model has been applied within a given context and its performance has been compared with Klout, a popular metric of online influence. Under certain conditions the proposed model achieves to imitate Klout with a fairly accurate precission. However, the main advantage of this model is that it is highly adjustable and it can be applied in a wide variety of topics.

Keywords: Twitter, influence, viral marketing, statistics and probabilities, graph-based theory, automatic text classification, data mining.

Glosario de Términos

Blog: Derivado de la palabra ingles *weblog*, es un sitio web periódicamente actualizado que recopila cronológicamente textos o artículos de uno o varios autores, apareciendo primero el más reciente, donde el autor conserva siempre la libertad de dejar publicado lo que crea pertinente.
Campaña viral: es un término empleado para referirse a las técnicas de marketing que intentan explotar redes sociales y otros medios electrónicos para producir incrementos exponenciales en reconocimiento de marcas y campañas publicitarias o políticas, mediante procesos de autorreplicación viral análogos a la expansión de un virus informático. Se suele basar en el boca a boca mediante medios como Internet para llegar a una gran cantidad de personas rápidamente. En ingles este término se conoce como *viral marketing*.
Dataset: es una colección de datos normalmente tabulada. Por cada elemento (o individuo) se indican varias características.
Microblogging: es un servicio que permite a sus usuarios enviar y publicar mensajes breves (alrededor de 140 caracteres), generalmente sólo de texto. Las opciones para el envío de los mensajes varían desde sitios web, a través de SMS, mensajería instantánea o aplicaciones ad hoc.Estas actualizaciones se muestran en la página de perfil del usuario, y son también enviadas de forma inmediata a otros usuarios que han elegido la opción de recibirlas. El usuario origen puede restringir el envío de estos mensajes sólo a miembros de su círculo de amigos, o permitir su acceso a todos los usuarios, que es la opción por defecto.
Post: un mensaje publicado en un blog o en un foro de internet.
Redes sociales: Son estructuras sociales compuestas de grupos de personas, las cuales están conectadas por uno o varios tipos de relaciones, tales como amistad, parentesco, intereses comunes o que comparten conocimientos.

Lista de Abreviaturas

API: del inglés *Application Programming Interface*, es el conjunto de funciones y procedimientos o métodos, en la programación orientada a objetos que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción. Son usadas generalmente en las bibliotecas (también denominadas comúnmente librerías).
JSON: del inglés *JavaScript Object Notation*, es un estandar para representar estructuras de datos simples y arreglos asociativos llamados objetos. Dichas estructuras y sus datos asociados pueden ser recuperados y manipulados por varios lenguajes de programación.
NLTK: del inglés *Natural Language Toolkit*, es una librería de funciones para el lenguaje de programación Python, destinadas al procesamiento del lenguaje natural.
SQL: del inglés *Structured Query Language*, es un lenguaje para realizar consultas sobre bases de datos relacionales.

Índice

Lista de Figuras

Lista de Tablas

1. Introducción

En términos generales la influencia ha sido estudiada largamente en diversos campos (sociología, marketing, política y comunicaciones). El comportamiento del mercado y la sociedad está altamente relacionado con este concepto, por lo tanto el estudio de este puede colaborar con determinar por qué algunas tendencias e innovaciones son adoptadas más rápido que otras, y cómo es que la influencia entrega un apoyo al desarrollo de campañas publicitarias más efectivas, entre otras cosas.

En este trabajo de título se pretenderá investigar sobre el rol que juega este concepto en la red social Twitter, con el fin de determinar el grado de influencia que tienen la acciones de determinados usuarios por sobre las de otros así como la probabilidad de que un usuario sea influenciado por las acciones de los demás. Se hará una descripción de las herramientas, métricas y algoritmos que se han propuesto al respecto y se pondrán a prueba. Más adelante se propondrán alternativas que permitan mejorar algunos de los modelos existentes y así contribuir con el desarrollo de esta área de estudio.

Una de las disciplinas que ayudará en el desarrollo de esta investigación será la teoría de grafos, la cual cuenta con un alto grado de madurez de estudio. Se espera que con el apoyo visual que entrega esta disciplina, así como con algunos de sus algoritmos, se puedan tener otras perspectivas que puedan aportar a la investigación.

En el capitulo 2 se presentará el marco teórico, el cual aborda el fundamento científico que respaldará tanto a las publicaciones investigadas, como a las soluciones al problema que se planteará en esta investigación.

En el capitulo 3 se verán los avances que se han hecho en cuanto al estudio de la influencia en Twitter y en las redes sociales en general. Se entrará en detalle sobre tres publicaciones al respecto en los capítulos 4.1, 4.2 y 4.3 respectivamente.

En el capítulo 4.5 se verá un estudio relacionado con la intención de los mensajes en Twitter, el cual tendrá especial importancia en los últimos capítulos.

En el capítulo 4 se identificarán algunos problemas que no se han abordado en las publicaciones investigadas.

En el capitulo 5 se definirá el objetivo de esta investigación, y parte del plan para lograrlo se verá en el capítulo 6.

La solución a los problemas identificado será abordada en el capítulo 7.

En los capítulos 8 y 9 se describirán los procedimientos realizados para implementar el modelo de solución, y en el capítulo 10 se comentarán los resultados de está investigación.

Finalmente, en el capítulo 11 se presentarán conclusiones acerca de las publicaciones investigadas, y de la aplicación del modelo propuesto para solucionar los problemas identificados.

2. Marco Teórico

2.1. Twitter

Twitter es una red social y servicio de *microblogging* de carácter gratuito, que permite a sus usuarios leer y responder *posts* conformados por texto, con un límite de 140 caracteres. Estos *posts* se conocen informalmente como *tweets*, los cuales son *tweeteados* por los usuarios registrados en el sistema. Los usuarios pueden *tweetear* desde un navegador con internet a través de un computador personal o desde una gran variedad de dispositivos. Esto gracias a que Twitter ofrece una *API* flexible para los desarrolladores de software.

Lanzado en julio de 2006 y creado por el estadounidense Jack Dorsey, el servicio ha ganado popularidad rápidamente en todo el mundo. Son más de 200 millones de usuarios los que se han suscrito al servicio desde sus inicios hasta marzo de 2011 [1]. En agosto de 2011, se estimó que se publican alrededor de 200 millones de *tweets* por día y la cifra sigue aumentando [2].

La variedad de usuarios del servicio y sus interacciones han hecho que Twitter sirva para diversos fines. Algunos usuarios lo usan para conversar y opinar sobre diversos temas, mientras que otros, lo usan como una plataforma de noticias. También están los que usan este servicio con fines publicitarios y políticos. Este último punto es de principal interés en el desarrollo de está investigación.

En la práctica, el hecho de que existan diversos fines para utilizar el servicio no implica que un usuario no pueda asumir distintos roles dentro de esta red social. Esto quiere decir que por ejemplo, un usuario que cumple con la labor de informar no necesariamente se limita a esa labor, y perfectamente puede opinar sobre ciertos temas o entablar una conversación con otros usuarios.

A causa de sus atributos y su creciente popularidad, Twitter se ha vuelto una red social atractiva para reconocidos líderes de opinión y medios informativos, los cuales han creado cuentas en el sistema y participan activamente en este servicio.

2.1.1. Seguidores y Amigos

Cuando un usuario se interesa en los contenidos de los *tweets* de otro usuario tiene la opción de seguirlo, de tal forma que un usuario recibe notificaciones de todos los usuarios que está siguiendo a medida que estos van generando nuevos *tweets*. Desde el punto de vista del usuario, se entenderá por *followers* a los usuarios que lo siguen, y por *friends* a los usuarios que él está siguiendo. En los perfiles de cada usuario de Twitter aparece la cantidad de *followers* y *friends* que este tiene, y a su vez, la cantidad de usuarios que él está siguiendo. En general, los líderes de opinión, celebridades y los perfiles de noticias tienen una gran cantidad de *followers*.

2.1.2. *Retweets*

Cuando un usuario se interesa en lo que ha *tweeteado* otro usuario, tiene la opción de volver a publicar el *tweet* original, haciendo referencia al usuario que lo creó. Esto se conoce como *retweet*. Los *followers* del usuario que *retweetea*, pueden ver este *retweet* en sus notificaciones. No es necesario que un usuario siga a otro para *retweetear* o ser *retweeteado*. Sin embargo, los usuarios con perfil privado no pueden ser *retweeteados*.

2.1.3. Menciones y Respuestas

Cuando un usuario quiere interactuar de forma más directa con otro de forma pública, puede hacer una mención, lo que vendría a ser como un intercambio de mensajes entre un usuario y otro. Una mención se considera pública por el hecho de que los *followers* de quién hace la mención pueden ver este intercambio en sus notificaciones, y si lo desean, pueden participar de la conversación mediante un *reply*. Las menciones llevan dentro del mensaje el nombre del usuario a quién se desea mencionar antecedido de una arroba (e.g., @usuario). No es necesario que un usuario siga a otro para hacer menciones o que lo puedan mencionar.

2.2. Correlación de rangos

En estadística, los coeficientes de correlación sirven para medir la intensidad con la que dos variables están relacionadas linealmente. Para casos en que existan relaciones no lineales entre dos variables es conveniente usar coeficientes de correlación de rangos. Este tipo de coeficientes permite determinar el tipo de asociación que existe entre los rangos de dos variables y no por su valores propiamente tales. Como se dijo anteriormente, los coeficientes de correlación de rangos no requieren que las relaciones entre dos variables sean lineales pero si se espera que sean relaciones monótonas. Otra razón por la que algunos autores prefieren utilizar coeficientes de correlación de rangos por sobre los de correlación general, es por el hecho de que estos tienen una mayor facilidad de cálculo [3].

2.2.1. Coeficiente de correlación de Spearman (ρ)

El coeficiente de correlación de Spearman es un coeficiente de correlación de rangos que se formula a partir del coeficiente de correlación *momento-producto* de Pearson [3].

Considérese una muestra de n individuos cuyos rangos relativos a ciertas características A y B, se encuentran descritos por las variables $X_1, X_2, X_3, ...X_n$ y $Y_1, Y_2, Y_3, ...Y_n$ respectivamente. Además, la diferencia de rangos entre cada individuo se describe por la relación $d_i = X_i - Y_i$, donde i representa al *i-ésimo* individuo de la muestra.

A partir del coeficiente de correlación *momento-producto* de Pearson, se tiene [3]:

$$\rho = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \tag{1}$$

Por otro lado, la varianza de la variable x se describe como [3]:

$$\sigma_x^2 = \frac{\sum x^2}{n} \tag{2}$$

Alternativamente, la varianza de x puede definirse como [3]:

$$\sigma_x^2 = \frac{(n^2 - 1)}{12} \tag{3}$$

De 2 y 3 puede deducirse que:

$$\sum x^2 = \sum y^2 = \frac{n(n^2 - 1)}{12} \tag{4}$$

Por otra parte:

$$\sum d^2 = \sum (X_i - Y_i)^2 = \sum (x - y)^2 = \sum x^2 - 2\sum xy + \sum y^2 \tag{5}$$

Despejando 5 se tiene:

$$\sum xy = \frac{1}{2}(\frac{n(n^2-1)}{6} - \sum d^2) \tag{6}$$

Finalmente, reemplazando 4 y 6 en 1 se tiene [3]:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{7}$$

La ecuación 7 se conoce como el coeficiente de correlación de rangos de Spearman, el cual se utiliza para relaciones monotónicas, en muestras como la descrita anteriormente y no está restringido por el tipo de distribución.

En ciertos casos, es imposible distinguir cual es el rango que le corresponde a dos o más individuos que comparten un mismo valor bajo una determinada característica. En tales circunstancias es mejor promediar dichos rangos y asignar ese promedio a cada uno de los individuos afectados [3]. Por ejemplo, en una ordenación de 10 se le asigna a un individuo el rango 1, pero no es posible decidir cuál de los dos individuos próximos debe ser el segundo o cuál el tercero. Entonces, dichos individuos aparecen empatados, y a cada uno se le da el rango $\frac{2+3}{2} = 2{,}5$. El próximo individuo tiene entonces el rango 4, y así sucesivamente. Si aparecen empatados los tres individuos siguientes, se le asignara a cada uno de ellos el rango $\frac{4+5+6}{3} = 5$, siguiendo la misma lógica del primer caso.

2.2.2. Coeficiente de correlación de Kendall (τ)

Otro coeficiente de correlación por rangos de mucha utilidad es el de Kendall. Este difiere del coeficiente de correlación de Spearman en su forma de cálculo y en los resultados que arroja, sin embargo ambos se utilizan para fines similares.

Considérese una muestra de n individuos cuyos rangos relativos a ciertas características A y B, se encuentran descritos por las variables $X_1, X_2, X_3, ...X_n$ y $Y_1, Y_2, Y_3, ...Y_n$ respectivamente. En caso de haber empates, se procederá asignar los rangos que causan conflicto con el mismo criterio que se explicó anteriormente para ρ.

El coeficiente de correlación de Kendall para dicha muestra sería [3]:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{8}$$

Donde n_c sería la totalidad de pares concordantes en la muestra, y n_d sería la suma de los pares discordantes de la misma. Sean (x_i, y_i) e (x_j, y_j) dos pares de individuos dentro de la muestra se sabrá si ambos pares son concordantes o discordantes al verificar lo siguiente [4]:

1. Si $x_i > x_j$ y $y_i > y_j$, o bien, $x_i < x_j$ y $y_i < y_j$, el par será considerado concordante.

2. Si $x_i > x_j$ y $y_i < y_j$, o bien, $x_i < x_j$ y $y_i > y_j$, el par será considerado discordante.

3. Si no ocurre ninguna de las dos anteriores, es decir, si $x_i = x_j$ o $y_i = y_j$, entonces el par no será ni concordante ni discordante.

Para poder considerar los empates entre alguna de las características de dos individuos, es necesario hacer una modificación a 8, de tal forma que los pares que no son ni concordantes ni discordantes puedan ser considerados en τ, y de esa forma tener una medida de correlación más precisa.

Se define el nuevo τ como sigue:

$$\tau = \frac{n_c - n_d}{\sqrt{\frac{1}{2}n(n-1) - U_x}\sqrt{\frac{1}{2}n(n-1) - U_y}} \tag{9}$$

Donde U_x y U_y corresponden a la suma de los empates de X e Y respectivamente y se calculan de la siguiente manera:

$$U_x = \frac{1}{2}\sum_{i=1}^{T} t_i(t_i - 1) \tag{10}$$

Donde t_i correspondería a la frecuencia con la que se repite un *i-ésimo* empate dentro de un total de grupos de empates T. Por ejemplo, si en X existe el rango 2 repetido dos veces y el rango 6 repetido 3 veces, habrían dos grupos de empates, uno para cada rango, por lo tanto $T = 2$. Como el primer rango se repite 2 veces, $t_i = 2$. Para el segundo rango repetido, $t_i = 3$. De esta forma, $U_x = \frac{1}{2}(2 + 3 * 2) = 4$. Para U_y la lógica es la misma, se calcula la frecuencia de empates que hay para cada grupo de empates en Y.

2.2.3. Interpretación y relación entre ρ y τ

Tanto para ρ como para τ, los coeficientes de correlación por rangos se mueven entre -1 y 1 [4].

1. Si tanto ρ como τ son iguales a 1, significaría que ambos rangos comparados estarían perfectamente correlacionados (son iguales).

2. Si tanto ρ como τ son iguales a -1, significaría que ambos rangos comparados estarían inversamente correlacionados (son opuestos).

3. Si tanto ρ como τ son iguales a 0, significaría no existe correlación entre ambos rangos (no hay dependencia)

Por lo tanto la cercanía que tengan los coeficientes de correlación con alguno de estos tres casos, determinará la naturaleza que hay entre ambos rangos.

Al comparar los valores de ρ y τ para una misma muestra n, se ha comprobado que aproximadamente $\rho = \frac{3\tau}{2}$. Esta relación se cumple para valores grandes de n, y siempre que ρ y τ no sean muy cercanos a 1 o -1 [3].

2.2.4. Contraste entre ρ y τ

El cálculo de τ es más tedioso que el de ρ. Sin embargo τ es más conveniente que ρ en las siguientes circunstancias [3]:

1. Existen métodos conocidos para realizar pruebas de significación sobre τ, sin embargo, salvo excepciones, es poco lo que se puede hacer respecto a ρ
2. τ puede extenderse a las correlaciones parciales por rangos.
3. Si se añade un nuevo miembro a la muestra es más fácil calcular τ que ρ, ya que en el caso de ρ habría que reordenar los rangos de la muestra y habría que recalcular las diferencias de rango entre cada individuo. Para τ sólo bastaría con agregar los nuevos pares a la correlación calculada previamente.

2.3. Teoría de grafos

Un grafo es una estructura de datos consistente en un conjunto de vértices conectados por un conjunto de aristas que se pueden utilizar para modelar las relaciones que existen entre los objetos dentro de una colección [5]. Los grafos son típicamente estudiados en el área de teoría de grafos, la cual se apoya en estudios matemáticos.

2.3.1. Definición y propiedades de grafos

Formalmente, un grafo se define como un par ordenado $G = (V, E)$, donde V es una colección de vértices $V = V_i, ..., V_n$ y E es una colección de aristas sobre V, tal que $E_{ij} = (V_i, V_j), V_i \in V, V_j \in V$. Alternativamente los grafos se conocen como redes, los vertices como nodos, y las aristas como vínculos. En la figura 1(a) se muestra un grafo, donde A, B, C, D y E son vértices y los trazos que unen estos vértices son sus artistas.

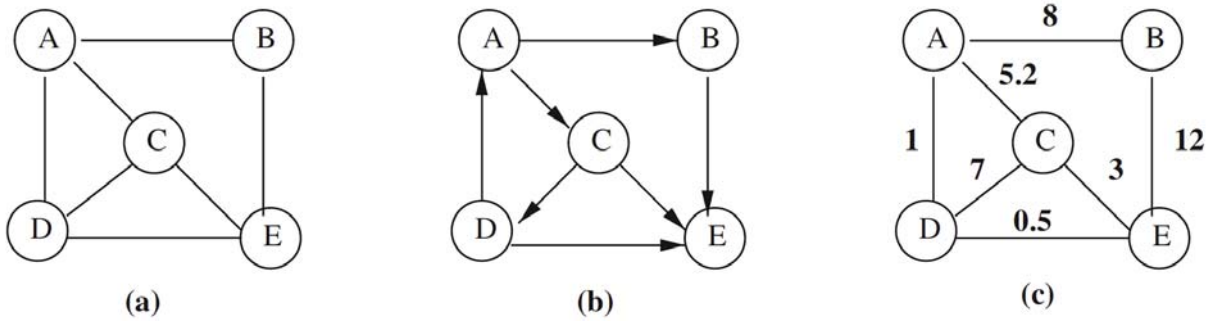

Figura 1: Grafos de ejemplo: (a) no dirigido, (b) dirigido y (c) con pesos y no dirigido

Los grafos pueden ser dirigidos o no dirigidos [5]. En un grafo dirigido una arista E_{ij} se puede recorrer desde V_i a V_j, pero no en dirección contraria. Tomando en cuenta la dirección anterior, el vértice V_i sería la cola de la arista y el vértice V_j la cabeza. También existen grafos en que sus aristas tienen valores numéricos que representan algún tipo de relación entre sus vértices. Estos valores se conocen como *pesos* [5].

Dos vértices, V_i y V_j, son considerados *adyacentes* cuando se encuentran conectados por una misma arista [5]. La arista que conecta ambos vértices se conoce como incidente [5]. Un grafo en el cual todo par de vértices es adyacente se llama grafo completo [5].

Una propiedad importante es el grado que posee un vértice de un grafo. Se define como el total de aristas que inciden sobre un vértice [5]. En un grafo dirigido el grado se puede clasificar como *in-degree* (las artistas que entran al vértice) o *out-degree* (las aristas que parten desde el vértice) [5].

El grado promedio de un grafo se puede calcular de la siguiente manera [5]:

$$a = \frac{1}{N} \sum_{i=1}^{n} Grado(V_i) \tag{11}$$

Donde N representa al total de vértices que hay en el grafo.

Una secuencia de vértices, en la que cada arista es incidente con dicha secuencia, será llamada un camino [5]. En caso de que en el camino no se repitan vértices, se entenderá que este es un camino simple [5]. Un camino cerrado en el que el primer vértice coincide con el ultimo es conocido como ciclo [5]. Un grafo en el cual cada vértice puede conectarse con otro a través de un mismo camino, es llamado grafo conexo [5].

Un subgrafo es un subconjunto dentro de un grafo, en el cual sus vértices y aristas forman un nuevo grafo $S = (V_s, E_s)$, siendo V_s y E_s subconjuntos de los vértices y artistas del grafo original [5].

2.3.2. Representación matricial

Los grafos y las matrices suelen usarse intercambiablemente para representar relaciones de datos [5]. La figura 2 muestra una representación matricial del grafo G mostrado en la figura 1(b), el cual tiene cinco nodos (A, B, C, D y E) y siete aristas (AB, AC ,BE, CD, CE, DA y DE). En la figura 2, el grafo G de la figura 1(b) se encuentra representado como una matriz en la cual las filas y columnas corresponden, de izquierda a derecha y de arriba hacia abajo respectivamente, a los vértices A hasta E. Un valor 1 indica la presencia de una arista dirigida entre dos vértices y 0 en caso de no haber conexión alguna [5]. Por ejemplo, un valor 1 para la fila 2, columna 5 ($G_{2,5}$), implica que hay una arista que va desde B a E. Por otro lado, el valor 0 en $G_{2,1}$ indica que no hay ninguna arista dirigida de B hacia A. En un grafo dirigido como el de la figura 1(b), no todas las relaciones son simétricas, por lo tanto no necesariamente $G_{i,j} = G_{j,i}$.

$$G = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Figura 2: Representación matricial del grafo dirigido de la figura 1(b)

En un grafo no dirigido, las relaciones sí son simétricas, por lo tanto una conexión de vértices es representada redundantemente en utilizando dos celdas de la matriz [5]. Por ejemplo, ambas aristas AC ($G_{1,3}$) y CA ($G_{3,1}$) llevarían un valor 1 en la matriz de la figura 3, la cual representa al grafo de la figura 1(a). Lo mismo para el resto de cada par de vértices unidos.

$$G = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Figura 3: Representación matricial del grafo no dirigido de la figura 1(a)

Los grafos con pesos utilizan una estructura similar pero los valores de la matriz llevan el valor de los pesos correspondientes en lugar del valor 1[5]. La figura 4 es una representación del grafo con pesos de la figura 1(c).

$$G = \begin{pmatrix} 0 & 8 & 5{,}2 & 1 & 0 \\ 8 & 0 & 0 & 0 & 12 \\ 5{,}2 & 0 & 0 & 7 & 3 \\ 1 & 0 & 7 & 0 & 0{,}5 \\ 0 & 12 & 3 & 0{,}5 & 0 \end{pmatrix}$$

Figura 4: Representación matricial del grafo con pesos de la figura 1(c)

En el caso de un grafo no dirigido con pesos, se seguiría la misma lógica de la matriz de la figura 3, pero lo que redunda esta vez serían los pesos de cada par de vértices.

2.3.3. *Betweenness Centrality*

El *betweenness* es una medida de centralidad que indica qué tan importante es un nodo contando la cantidad de caminos mas cortos que pasan por él [5]. *El Betweenness Centrality* de un nodo x es el cuociente de todos los caminos más cortos que pasan por x:

$$BET(x) = \sum_{i,j \in V - \{x\}} * \frac{número\ de\ caminos\ más\ cortos\ entre\ i\ y\ j\ que\ pasan\ por\ x}{número\ de\ caminos\ más\ cortos\ entre\ i\ y\ j} \quad (12)$$

2.3.4. *PageRank*

PageRank fue concebido como un método para darle rangos de importancia a las páginas web que forman parte del motor de búsqueda Google. *PageRank* se basa en un sistema de votaciones

y recomendaciones. Cuando una página web se enlaza con otra, le está dando su "voto". Mientras más votos tiene una página, mayor es su importancia. Un voto también cobra importancia si es emitido por una página importante.

Formalmente considérese un grafo dirigido $G = (V, E)$, para un vértice V_i, $In(V_i)$ y $Out(V_i)$ su *in-degree* y *out-degree* respectivamente. Con esto se define el puntaje PR para un nodo V_i como:

$$PR(V_i) = \frac{(1-d)}{|V|} + d * \sum_{j \in In(V_i)} \frac{1}{Out(V_j)} PR(V_j) \tag{13}$$

Donde d vendría siendo un factor de amortiguación que puede estar entre 0 y 1 y se entiende como la probabilidad de saltar de un nodo a otro de forma aleatoria en el grafo. Por defecto este valor es igual a $0{,}85$.

2.4. Clasificación Automática de Textos

El problema de la clasificación de textos surge de la necesidad de categorizar un conjunto de documentos dado un conjunto de clases. En otras palabras, determinar a que clase o categoría pertenece cierto documento [6]. La clasificación de textos en general tiene muchas aplicaciones, entre ellas: indexación automática de textos, análisis de sentido, filtrado de documentos, desambiguación de palabras, detección de correo *spam*, detección de contenido para adultos y en general todo lo relacionado con la organización de documentos.

En términos formales la clasificación de textos se compone de un conjunto de documentos $\mathbb{X} = \{d_1, d_2, \ldots, d_i\}$ que pueden ser categorizados en un número finito de clases, categorías o etiquetas, en un conjunto $\mathbb{C} = \{c_1, c_2, \ldots, c_j\}$. Las clases van a depender del ámbito de aplicación, o el problema que se desee abarcar con la clasificación de los documentos. A partir de lo anterior, se define un conjunto de entrenamiento $\mathbb{D}$ o *training set* [6] como pares de documentos etiquetados $(d, c) \in \mathbb{X} \times \mathbb{C}$. Por ejemplo:

$$(d, c) = (Santiago\ es\ una\ de\ las\ capitales\ con\ más\ polución\ en\ el\ mundo, Chile)$$

sería un par compuesto por el documento con la frase *Santiago es una de las ciudades con más polución en el mundo* y su clase (en este caso) *Chile*. Usando un método de aprendizaje o algoritmo de aprendizaje [6], el objetivo es implementar una función γ que sea capaz de asociar documentos con clases, de manera formal:

$$\gamma : \mathbb{X} \to \mathbb{C}$$

Este tipo de aprendizaje es llamado aprendizaje supervisado, ya que el supervisor (un humano), es el que define las clases y documentos de entrenamiento y es el que dirige el proceso de aprendizaje. Un conjunto de entrenamiento se encuentra provisto de una variedad de ejemplos típicos para cada clase los cuales sirven para entrenar a la función de clasificación γ. Una vez que γ ha sido entrenada se puede aplicar a un conjunto de datos de prueba o *test set*, el cual es un conjunto de documentos cuyas clases son desconocidas. La función γ debe ser capaz de asignar estos documentos no clasificados a alguna clase. Siguiendo el ejemplo anterior, si d es un documento nuevo, $\gamma(d)$ podría ser igual a *Chile* o alguna otra clase de ese conjunto, según

lo estime la función de clasificación. El objetivo en la clasificación de textos es lograr una alta precisión en la clasificación de datos de prueba. Más adelante se explicará en detalle el concepto de precisión.

2.4.1. *Bag of Words*

El modelo *bag of words* (bolsa de palabras) [6] es un modelo de espacio vectorial en el que cada documento es representado por un vector. Cada componente dentro del vector representaría la ocurrencia de una palabra del diccionario de términos en un documento. Entiéndase el diccionario de términos como el conjunto de todas las palabras distintas que aparecen en el conjunto de todos los documentos. Lo que se busca con este modelo es almacenar la cantidad ocurrencias de cada término, por lo tanto se pierde el orden de las palabras dentro del documento y su significado gramatical. Sin embargo, hay una alta probabilidad de que dos documentos con un mismo *bag of words* sean similares en contenido.

En términos formales, si $\mathbb{D}$ es un vector de documentos, D_i cada uno de los documentos que pertenecen al vector, y d_{in} la ocurrencia del n-ésimo término del diccionario dentro del documento i, la representación de un documento en general sería $D_i = (d_{i1}, d_{i2}, ..., d_{in})$.

El valor de cada componente del vector anterior dependerá del método utilizado para representar la existencia de un n-ésimo término en un documento. Por ejemplo, en una representación binaria, d_{in} puede valer 1 si aparece en un documento o 0 en caso contrario.

2.4.2. *Stopwords*

En ocasiones hay palabras demasiado comunes y frecuentes en un documento, las cuales no aportan ningún valor significativo en la clasificación de estos. Estas palabras son conocidas como *stopwords* [6] y es común que sean eliminadas en la etapa de procesamiento de un documento. Ejemplos de *stopwords* son determinantes, conjunciones y preposiciones, pero también pueden ser palabras que son demasiado comunes en el contexto de cierto documento. Por lo tanto, el idioma del diccionario de *stopwords* a utilizar va a depender del idioma del documento que va a ser clasificado y de la categoría a la que pertenece el documento mismo. A modo de ejemplo, si se están analizando documentos médicos en español, debería usarse un diccionario de *stopwords* en español y con términos que se consideran demasiado comunes en el lenguaje médico.

2.4.3. Naive Bayes

Naive Bayes es un método de aprendizaje supervisado basado en el teorema de Bayes, el cual asume la independencia de variables. Por esta razón es que se considera un clasificador ingenuo (naive). Este es un método probabilístico en el cual la probabilidad de que un documento d se encuentre en una clase c es:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \tag{14}$$

Donde $P(t_k|c)$ es la probabilidad condicional de la ocurrencia de un k-ésimo término t_k en c. Se puede interpretar como una medida de cuánta evidencia aporta t_k para que c sea la clase correcta. $P(c)$ vendría siendo la probabilidad anticipada de que un documento se encuentre en

la clase c. Por último n_d es la cantidad de términos distintos que hay en un documento d, una vez removidas las *stopwords*.

El objetivo en la clásificación de un documento es encontrar la mejor clase para este. En Naive Bayes la clase más probable es c_{map} (clase máxima a posteriori). Su valor se encuentra dado por la ecuación 15:

$$c_{map} = arg\,\max_{c\in\mathbb{C}}\ \hat{P}(c|d) = arg\,\max_{c\in\mathbb{C}}\ \hat{P}(c)\prod_{1\leq k\leq n_d}\hat{P}(t_k|c) \tag{15}$$

Donde las probabilidades P son estimativas y por lo tanto se describen como $\hat{P}$.

2.4.4. Métodos de Evaluación

Para medir la calidad de un clasificador, existen distintas métricas que se usan de manera estándar en el área de clasificación de textos. Para probar un clasificador e implementar estas medidas, se necesita un documento de prueba, el cual será clasificado automáticamente por el algoritmo clasificador sometido a evaluación. Por otro lado, el mismo documento será clasificado de manera manual. Finalmente se comparan ambas clasificaciones con medidas como las que se verán en esta sección.

Se necesita de una tabla de contingencia para poder interpretar dichas medidas. Esta tabla sirve para ver si la predicción de una clase para un documento se hizo correctamente o no [6].

Predicción	Relevante	No Relevante
Clasificado	Verdadero Positivo (VP)	Falso Positivo (FP)
No Clasificado	Falso Negativo (FN)	Verdadero Negativo (VN)

Tabla 1: Tabla de contingencia

Para entender mejor la tabla 1 hay que considerar 4 casos distintos:

1. Se predijo correctamente que el documento pertenecía a una clase (Verdadero Positivo).
2. Se predijo incorrectamente que el documento pertenecía a una clase (Falso Positivo).
3. Se predijo incorrectamente que el documento no pertenecía a una clase (Falso Negativo).
4. Se predijo correctamente que el documento no pertenecía a una clase (Verdadero Negativo).

Las métricas que se verán en esta sección serán precisión, *recall*, *F-measure* y *accuracy*. Todas utilizan valores de la tabla de contingencia en sus fórmulas [6].

La precisión (π) se entiende como la proporción de documentos correctamente clasificados entre todos los que fueron predichos para una determinada clase:

$$\pi = \frac{VP}{VP + FP} \tag{16}$$

El concepto de *recall* o cobertura (ρ) representa la proporción de documentos correctamente clasificados respecto de todos los que debiesen haber sido correctamente clasificados:

$$\rho = \frac{VP}{VP + FN} \tag{17}$$

El concepto de *F-measure* es un híbrido entre las medidas de precisión y *recall*:

$$F_\beta = \frac{(1 + \beta^2)\pi\rho}{\beta^2\pi + \rho} \tag{18}$$

Donde β controla la importancia relativa entre ρ y π. Usualmente se utiliza $\beta = 1$ para darle igual importancia a ambas medidas.

Finalmente, la exactitud o *accuracy*, es una medida que considera todos los casos de clasificación exitosa:

$$accuracy = \frac{VP + VN}{VP + FP + VN + FN} \tag{19}$$

Usualmente no se recomienda esta medida ya que es muy fácil tener un *accuracy* alto y esto no necesariamente representa un conjunto de documentos bien clasificados [6].

3. Estado del Arte

En una de sus acepciones, el diccionario de la lengua española define influencia como el "Poder, valimiento, autoridad de alguien para con otra u otras personas"[7]. Y aunque dependiendo del contexto la influencia se puede definir de muchas maneras, esta es una de las definiciones que más se acerca al tipo de influencia que se puede ejercer en una red social, y más específicamente, en Twitter.

En general, los autores que han contribuido en el estudio de la influencia en las redes sociales abarcan este concepto como la intensidad de persuasión que tienen algunos usuarios por sobre otros al momento de efectuar alguna determinada acción. En Twitter dicha acción podría ser un *tweet*, una mención o un *retweet*, es decir, básicamente cualquier evento que refleje la intención del usuario de esparcir sus mensajes para que cada uno de sus *followers* los vean en sus notificaciones.

La influencia como tal es un concepto difícil de definir y más aún, de medir. A raíz de esto, se ha convertido en un interesante desafío para quienes se han dedicado a estudiarla. En especial en las redes sociales, donde su estudio puede servir para entregar apoyo al desarrollo de *campañas virales* más efectivas, al seleccionar a los usuarios más influyentes dentro de una red con el fin de poder esparcir un determinado mensaje que se espera que una gran cantidad de usuarios adopten. Si bien no hay una unidad física para medir la influencia en Twitter, se ha estudiado el comportamiento y los atributos de los usuarios de esta red social, así como la forma en que se propagan los mensajes dentro de esta, lo que ha permitido poder definir ciertas métricas y *frameworks* para estudiar este concepto.

Esto ha dado varios indicios sobre el comportamiento de los usuarios de Twitter y la susceptibilidad de algunos para adoptar ciertas tendencias. Analizar la forma en que se propagan de los mensajes en Twitter no es algo fácil, sin embargo no es imposible. Twitter ofrece una *API* para sus desarrolladores la cual entre otras cosas, ofrece facilidades para la extracción de *tweets* en su red.

Ya que Twitter es una red abierta (no hay que estar registrado para revisar los *tweets* de otros usuarios), también está la posibilidad de usar *Web crawlers* para extraer información similar. Sin embargo, tanto esta *API*, como los *Web crawlers*, tienen que lidiar con el inconveniente de que algunos usuarios tienen sus perfiles cerrados (solo los usuarios que ellos están siguiendo pueden ver sus *tweets*) por lo tanto este tipo de usuarios no aportan indicio alguno sobre la forma de propagación de los mensajes, y muchos autores que están estudiando el tópico de la influencia en Twitter han dejado simplemente de lado estos perfiles.

Entre las contribuciones al estudio de la influencia en las redes sociales, se puede encontrar un *framework* propuesto por [8], el cual utiliza modelos probabilísticos para medir la influencia entre usuarios y se apoya en la teoría de grafos para representar las interacciones entre ellos, así como la forma en que se propagan sus acciones. Este modelo se propuso para redes sociales en general, pero puede ser adaptado para su funcionamiento en Twitter.

Hay otros estudios que afirman que se pueden distinguir distintos tipos de influencia entre los usuarios de Twitter. Tal es el caso de [9] y [4], los cuales proponen los tipos de influencia que puede tener un usuario y realizan varias comparaciones entre estos para establecer cuales son los más relevantes. Ambos estudios proponen métricas para cada tipo de influencia, pero no una presentan una métrica para medir la influencia en términos globales. Sin embargo, identifican ciertos atributos que pueden ser indicios de lo que se podría considerar un prototipo de usuario

influyente en Twitter.

En un estudio propuesto por [10], se definen *ratios* entre distintos tipos de influencia [9] [4] para calcular el potencial de red que tiene un usuario. Es decir, la habilidad que tienen ciertos usuarios de Twitter para influenciar a sus *followers*.

Otro estudio interesante es el de [11]. Los autores de esta publicación distinguen entre dos tipos de usuarios, los usuarios pasivos y los activos. Los usuarios pasivos son los que sólo leen los *tweets* sin compartir el contenido u opinar de este. Por el contrario, los usuarios activos son los que participan y difunden el contenido. Los autores proponen un algoritmo para clasificar a los usuarios bajo estas dos categorías en base a sus datos de perfil y el contenido que difunden. Estos llegan a la conclusión los usuarios influyentes son los que logran logran transformar usuarios pasivos en activos.

En [12], los autores proponen un algoritmo para medir influencia que se basa en *PageRank*, llamado TwitterRank. Este toma en cuenta la estructura de la red formada por los usuarios en un determinado contexto. TwitterRank mide la influencia de los usuarios según un tema en particular.

Otro aspecto importante a considerar, es que ya existen empresas que se dedican al calculo de la influencia en Twitter. Una de ellas es Klout [13], que ofrece un servicio *online* gratuito que se basa en un sistema de puntuación llamado *Klout Score*. Lo interesante de este sistema de puntuación es que los usuarios pueden agregar sus perfiles no sólo de Twitter, si no que de otras redes sociales relevantes como Facebook y LinkedIn, entre otras. Esto permite al usuario saber cual sería su influencia global en las redes sociales. Klout permite a sus usuarios ver como varia su influencia a través del tiempo y también permite comparar influencias entre usuarios. Sin embargo, tanto la forma de medir el *Klout Score* como los criterios en que se basa Klout para medir la influencia de los usuarios son de carácter confidencial y a la fecha no han sido publicados por sus creadores. Otras medidas populares y similares a Klout son Kred [14] y PeerIndex [15].

En los siguientes capítulos se profundizará sobre el *framework* propuesto por [8] y se detallarán los tipos de influencia [9] [4] que se han identificado en Twitter, así como algunos *ratios* [10] que permiten tener una visión más acabada sobre el concepto de influencia en esta red social.

4. Definición del Problema

4.1. Modelo Probabilístico de Influenciabilidad en las Redes Sociales

Entre las contribuciones más importantes del *framework* propuesto por [8] se encuentra un modelo probabilístico, el cual determina la probabilidad de que un usuario pueda influenciar a algún otro usuario en particular en una red social, así como también la probabilidad de que un usuario sea influenciado por su vecindad.

4.1.1. Supuestos del modelo

Considérese un grafo social, el cual es un grafo no dirigido $G = (V, E, \tau)$, donde los vértices V son usuarios, E es el conjunto de aristas entre usuarios y $\tau : E \to \mathbb{N}$ es una función que define el *timestamp* (tiempo) para cada arista. Una arista $(u, v) \in E$ representa un vinculo entre los usuarios u y v. El *timestamp* de aquel vinculo que etiqueta a esa arista representa el momento en que u y v realizaron un vínculo social (Por ejemplo, se hicieron amigos en Facebook, o uno sigue al otro en Twitter).

Se tiene también un registro de acciones (una tabla), *Acciones(Usuario, Acción, Tiempo)*, el cual contiene tuplas de la forma (u, a, t_u). Esta tupla indica que un usuario u realizó la acción a en el instante t_u. Los usuarios de este registro son los usuarios del grafo social, es decir, cada tupla representa una acción de uno de los usuarios definidos en el grafo social.

Se define A_u como el numero de acciones realizadas por un usuario u. Por otro lado se define A_{v2u} como la suma de las acciones que se propagan desde un usuario v a u. Se dice que una acción $a \in A$ se propaga de un usuario vi_i a v_j siempre que:

1. $(v_i, v_j) \in E$. Es decir, existe un vínculo social entre v_i y v_j.
2. $\exists(v_i, a, t_i)(v_j, a, t_j) \in$ *Acciones* con $t_i < t_j$. Esto quiere decir que v_i tiene que haber realizado la acción a antes que v_j.
3. $\tau(v_i, v_j) \leq t_i$. Lo que quiere decir que v_i tiene que haber realizado la acción después de haber formado lazos con v_j

Cuando todo lo anterior se cumple, se define la relación $prop(a, v_i, v_j, \Delta t)$. Donde $\Delta t = t_j - t_i$. Con esta relación se puede definir un grafo de propagación PG para cada acción a, como $PG(a) = (V(a), E(a))$. Este grafo se compone de lo siguiente:

1. $V(a) = \{v | \exists t : (v, a, t) \in Acciones\}$, es decir, para formar el grafo de propagación tiene que haber ocurrido una acción, la cual estaría registrada en el registro de acciones.
2. Existe una arista dirigida $v_i \xrightarrow{\Delta t} v_j$ en $E(a)$ cuando $prop(a, v_i, v_j, \Delta t)$.

Cuando un usuario realiza una acción, se dice que se ha activado en relación a esa acción, por lo tanto se ha vuelto un usuario contagioso y tiene la facultad de activar a su vecindad inactiva. Un usuario sin embargo, no tiene la facultad de desactivarse. El poder de influenciar a dicha vecindad es lo que los autores [8] quieren modelar en su propuesta. El problema abordado consiste en cómo calcular las probabilidades de influencia entre vecinos con un registro previo que permita deducir las propagaciones entre ellos.

4.1.2. *Framework* de solución

Considérese un usuario inactivo u y un conjunto S con sus vecinos activados justo después de que u se ha hecho vecino de un usuario $v \in S$. Para predecir si u se activará, es necesario determinar $p_{u\leftarrow(s)}$, es decir, la probabilidad conjunta de que u sea influenciado por el conjunto S. Esta probabilidad de influenciabilidad conjunta se define como sigue [8]:

$$p_{u(S)} = 1 - \prod_{v \in S}(1 - p_{v,u}) \tag{20}$$

Donde $p_{v,u}$ es la probabilidad de que un usuario $v \in S$ influencie a u, por lo tanto, la ecuación 20 es el conjunto de todas las influencias que impone el conjunto S sobre u. Por asuntos de facilidad de cálculo, se asume que la probabilidad $p_{v,u}$ es independiente entre todos los usuarios v que influencian a u [8].

Los autores de este *framework* concluyen en su investigación, que el modelo probabilístico más confiable es el que combina la distribución de Bernoulli con un modelo estático llamado modelo de créditos parciales. Este último modelo se basa en la suposición de que si u es influenciado por S, cada usuario $v \in S$ comparte una porción de crédito por haber influenciado a u a realizar una acción a. Por lo tanto, si $|S| = d$, se define el crédito parcial de v sobre u al influenciarlo a realizar una acción a como [8]:

$$crédito_{v,u}(a) = \frac{1}{d} \tag{21}$$

Por otra parte, la distribución de Bernoulli, se basa en un cuociente entre intentos exitosos sobre intentos totales. Para efectos de este problema, se considerará como un éxito cada acción de v que haya influenciado a u. Por lo tanto se define $p_{v,u}$ con una distribución de Bernoulli como [8]

$$p_{v,u} = \frac{A_{v2u}}{A_v} \tag{22}$$

Por lo tanto, para calcular una probabilidad $p_{v,u}$ más confiable, es necesario tener la suma de créditos de v sobre u considerando cada una de las acciones que se han propagado de v hacia u. Combinando el modelo de créditos parciales de la ecuación 21 con la distribución de Bernoulli en la ecuación 22, se tiene [8]:

$$p_{v,u} = \frac{\sum_{a \in A} crédito_{v,u}(a)}{A_v} \tag{23}$$

La probabilidad conjunta definida en la ecuación 20 asume que las probabilidades de influencia permanecen invariables con el paso del tiempo, sin embargo en la práctica, esto no es así. Es de suponer que si un usuario u se da cuenta de la acción de v inmediatamente después de ser realizada, este no reaccionará de la misma forma que si se diera cuenta un día o un mes después. Es probable que u se motive más al realizar una acción influenciada por v mientras menos tiempo haya pasado desde su realización inicial. Este supuesto es comprobado por [8], por lo tanto se define un modelo de tiempo continuo, el cual es un ajuste del modelo de probabilidad conjunta y el modelo de probabilidad individual $p_{v,u}$.

En primer lugar $p^t_{v,u}$ sería la probabilidad de influencia $p_{v,u}$ dependiente del tiempo, el cual decrece exponencialmente a medida que transcurre el tiempo. Se define de la siguiente manera [8]:

$$p^t_{v,u} = p^0_{v,u} e^{-(t-t_v)/\tau_{v,u}} \tag{24}$$

Donde $p^0_{v,u}$ es el momento de máxima influencia, es decir, $p_{v,u}$ en el modelo estático anterior. Por otro lado t_v es el momento en que v realiza la acción, por lo tanto, $p^t_{v,u}$ es máximo cuando $t = t_v$. El parámetro $\tau_{v,u}$ se denomina tiempo de vida promedio y es el tiempo mínimo esperado que debería pasar entre que v realiza una acción y u realiza la misma.

En segundo lugar y de forma análoga, se define $p^t_u(s)$ como una probabilidad conjunta de influencia $p_u(s)$ dependiente del tiempo. Esta se define como sigue [8]:

$$p_{u^t(S)} = 1 - \prod_{v \in S} (1 - p^t_{v,u}) \tag{25}$$

4.2. Tipos de Influencia en Twitter

En las publicaciones de [9] y [4], se establece que un usuario puede ser influyente de distintas maneras en Twitter. En la primera publicación se identifican tres maneras:

1. **Influencia por *followers***: El numero de *followers* de un usuario indica directamente el tamaño de la audiencia que tiene este.

2. **Influencia por *retweets***: La cantidad de *retweets* que contienen el nombre del usuario indica la habilidad que tiene este para hacer llegar el contenido valioso a los demás.

3. **Influencia por *menciones*:** Las menciones realizadas a un usuario indican la habilidad que tiene este para entablar conversaciones.

Tanto para las menciones como para los *retweets*, [4] desglosa ambos tipos de influencia en dos. Para el caso de las menciones, el autor considera las menciones totales que se le realizan a un usuario, pero también toma en cuenta la cantidad de usuarios que *retweetean* al mismo, de manera de ser más imparcial ante los usuarios que hacen demasiadas menciones. La misma consideración se toma en cuenta para los *retweets*.

Desde ahora en adelante la letra F indicará la influencia a causa de *followers*. En el caso de las menciones, M será la influencia por menciones totales, y M_u la influencia medida por cantidad de usuarios que realizan estas menciones. Finalmente RT se entenderá como influencia por *retweets*, y RT_u, la influencia en cuanto a la cantidad de usuario que realizan dichos *retweets*.

4.2.1. Comparación entre tipos de influencia

Para comprobar cuales de los tipos de influencia mencionados son los más relevantes, [9] construyó un *dataset* a partir de aproximádamente dos mil millones de *tweets*. Estos *tweets* fueron intercambiados entre mas de 54 millones de usuarios. Después de eliminar de la muestra a los usuarios que tenían sus perfiles cerrados, y los usuarios que prácticamente no usaban su cuenta (menos de diez *tweets* en sus registros), se redujo la lista de usuarios relevantes a aproximádamente 6 millones. Esos 6 millones de usuarios formaron parte del análisis, sin embargo se estudiaron sus interacciones con toda la muestra (los 54 millones).

Para analizar los datos, los 6 millones usuarios fueron ordenados por rangos bajo los tres criterios de influencia determinados por [9]. Para medir la relevancia entre cada uno de los criterios ordenados, se utilizó el coeficiente de correlación de Spearman (ρ) definido en la ecuación 7. La tabla 2 muestra la correlación que existe entre cada uno de los tipos de influencia. Ya que los usuarios con pocos *tweets* suelen tener pocas o nulas menciones y *retweets* a la vez, dan información poco valiosa. Esto por la simple razón de que, por ejemplo, si un usuario tiene cero *retweets*, y cero menciones, la correlación entre ambos parámetros será perfecta, siendo que el usuario no sería para nada influyente bajo esos criterios. Para contrarrestar esto, no sólo se probó con la población completa, si no que con una selección de los usuarios más populares según cantidad de *followers*. Se analizaron dos muestras bajo este criterio, una con el $1\,\%$ de la población, y otra con el $10\,\%$.

La tabla anterior confirma dos cosas. Lo primero es lo que se dijo anteriormente respecto de los usuarios que no aportaban nada al estudio, ya que con la población completa, tanto F con RT, como F con M, están altamente correlacionados ($\rho > 0{,}5$ en ambos casos). Sin embargo la

Correlación	Todos	10 %	1 %
F vs RT	0.549	0.122	0.109
F vs M	0.638	0.286	0.309
RT vs M	0.580	0.638	0.605

Tabla 2: Correlación por rangos ρ entre tipos de influencia

correlación es muy débil cuando se compara con los usuarios más populares. Lo segundo, una de las razones por las que [9] tenía en mente hacer el estudio, era para demostrar que tener muchos *followers* no necesariamente significa ser influyente. Esto se demuestra con la baja correlación entre F, tanto con M, como con RT, y también con el hecho de que RT y M tengan una alta correlación en las tres divisiones de la población estudiada.

Otro aspecto que se analizó, es como varia la influencia entre distintos temas. Por ejemplo, si es que un usuario que es influyente al hablar de temas de espectáculo, también lo sería al hablar de política o algún otro tema radicalmente distinto. Los autores recopilaron *tweets*de tres eventos importantes que ocurrieron durante 2009, y que son de tres áreas distintas. En primer lugar, la elección del presidente de Iran, un hecho político. En segundo lugar, la muerte de Michael Jackson, un hecho relacionado con el mundo del espectáculo. Por último se recopilaron *tweets* relacionados con el esparcimiento del virus de la *influenza H1N1*, un hecho noticioso relacionado con el área de la salud. En la tabla 3 se muestra la correlación entre todos los temas. En este análisis se consideraron sólo usuarios que hablaron de los tres temas (aproximadamente 13 mil), y se ordenaron por rangos según la cantidad de *retweets* y la cantidad de menciones para cada tema discutido. Además se consideraron separadamente sólo los usuarios populares que pertenecían al 10 % y al 1 % de los 13 mil usuarios extraídos, y no la muestra completa.

Temas	10 %	1 %	10 %	1 %
Iran vs Influenza	0.54	0.62	0.59	0.68
Iran vs M. Jackson	0.48	0.54	0.59	0.63
Influenza vs M.Jackson	0.55	0.50	0.80	0.68

Tabla 3: Correlación por rangos ρ entre distintos temas

La tabla anterior muestra que los usuarios populares tienen la habilidad de ser influyentes en diversos temas. Sólo una de las correlaciones dio un valor menor a $0{,}5$, y aún así se acercó bastante a lo que podría considerarse una correlación alta.

En el estudio realizado por [4], los autores se centraron principalmente en analizar la estabilidad de la influencia a través del tiempo. Para eso formaron un *dataset* de aproximadamente 680 mil usuarios activos en Twitter, y sus *tweets*. Dicho *dataset* se ordenó por las fechas en que se realizaron los *tweets* de la muestra, y se dividió en partes iguales formando dos nuevos *datasets*. Para comparar, se utilizaron los coeficientes de correlación de rangos de Spearman y Kendall (ρ y τ) sobre cuatro métricas de influencia, las cuales se describieron anteriormente como RT, RT_u, M yM_u al inicio de este capítulo. El criterio de comparación se realizó sobre cada tipo de influencia ordenado por rangos en ambos *datasets*, y con ambos coeficientes de correlación. La idea es ver como se comportan las métricas descritas en dos intervalos de tiempo distintos. En

la tabla 4 se comparan las cuatro métricas mencionadas, y se muestra ρ y τ para cada una de ellas. Además se muestra un porcentaje de superposición para cada métrica, el cual representa la proporción de usuarios que tienen el mismo rango en ambos *datasets*. Se trabajó con los mil usuarios más populares del *dataset* para tener resultados más representativos.

Tipos	ρ	τ	Superposición
M_u	0.722	0.526	73.9 %
M	0.614	0.433	57.8 %
RT_u	0.802	0.997	54.1 %
RT	0.634	0.451	65.6 %

Tabla 4: Correlación por rangos ρ y τ para dos intervalos de tiempo.

De esta tabla se puede destacar que los resultados son bastante estables, es decir, la influencia ha perdurado con el pasar del tiempo. Es importante notar que RT_u tiene una correlación muy alta, especialmente en el caso de τ. Esto indica que la acción de *retweetear* es más constante en el tiempo que el resto de las otras acciones. Por otro lado M_u tiene un alto porcentaje de superposición, lo que implica que es un tipo de influencia estable a través del tiempo. Además, en cuanto a ρ y τ, M_u sólo esta detrás de RT_u lo que confirma el hecho de que es un tipo de influencia confiable. Con esto último se podría deducir que los usuarios populares suelen conservar una gran cantidad de público interesado en responder sus *tweets*.

Otro aspecto importante investigado por [4] es la correlación que existe tanto entre M con M_u, como entre RT y RT_u. Este hecho que no fue considerado en [9] porque trabajaron con *retweets* y menciones basándose sólo en cantidades de mensajes y no en la cantidad de usuarios que emiten aquellos mensajes. Para estas correlaciones, se volvió a ocupar el *dataset* original (no particionado).

En la tabla 5 se comparan RT y RT_u tanto en términos de ρ y τ como en el caso de la súper posición. Además se realizan comparaciones para 3 rangos de usuarios, los mil, cinco mil, y diez mil más populares de la muestra. De esta tabla se puede concluir que los usuarios con alta cantidad de *retweets*, son *retweeteados* por una gran variedad de usuarios, y viceversa. Por otro lado deberían haber pocos usuarios que son *retweeteados* masivamente por un numero reducido de *followers*.

Criterio	Top 1000	Top 5000	Top 10000
ρ	0.833	0.817	0.795
τ	0.655	0.628	0.604
Superp.	82.5 %	81.6 %	82.1 %

Tabla 5: Correlación y superposición entre RT y RT_u

En la tabla 6, la cual realiza un análisis similar pero con M y M_u, sucede lo contrario. La correlaciones y la superposición son menores, especialmente para los rangos de usuarios mayores o iguales a cinco mil usuarios. Esto indica que hay que usuarios realizan más menciones que otros lo que podría significar que hay usuarios que conversan prolongadamente con las

mismas personas y no necesariamente involucran a gente nueva, lo que disminuye la rotación de personas que interactúan con el usuario.

Criterio	Top 1000	Top 5000	Top 10000
ρ	0.749	0.574	0.533
τ	0.571	0.409	0369
Superp.	61.1 %	55.5 %	57.2 %

Tabla 6: Correlación y superposición entre M y M_u

4.3. *Ratios* de Influencia en Twitter

En el estudio realizado por [10], se presentan *ratios* que permiten sacar nuevas conclusiones acerca de la influencia y comportamiento de los usuarios en Twitter. A partir de un usuario u, se formulan los siguientes *ratios*:

1. ***Ratio followers/friends*** **(**$r_f(u)$**):**

$$r_f(u) = \frac{Followers\ de\ u}{Friends\ de\ u} \tag{26}$$

   Mientras más alto es este *ratio*, mas gente está interesada en seguir a u, independiente del interés que tenga u en seguir a los demás. Si $r_f(u) < 1$, es probable que el usuario pueda considerarse un seguidor de masas que sólo sigue a otros usuarios en búsqueda de influencia, sin embargo, esto es relativo y depende del contexto. Además, su buena interpretación requiere de otros *ratios* para no precipitarse a sacar conclusiones erróneas.

2. ***Ratio*** **de** ***retweets*** **& menciones (**$r_{RT}(u)$**):**

$$r_{RT}(u) = \frac{Retweets,\ Menciones(o\ Replies)\ emitidos\ por\ u}{Total\ de\ Tweets\ emitidos\ por\ u} \tag{27}$$

   Este *ratio* permite saber la proporción *tweets* de u que surgen de la interacción con su audiencia.

3. ***Ratio*** **de interacción (**$r_i(u)$**):**

$$r_i(u) = \frac{Seguidores\ de\ u\ que\ retweetean,\ mencionan\ (o\ responden) a\ u}{Seguidores\ de\ u} \tag{28}$$

   El numerador de este ratio es similar a las métricas RT_u y M_u descritas en el capítulo 4.2, las cuales se comprobó que son confiables para medir influencia). Este *ratio* permite saber la proporción de usuarios que interactúan con u.

4.4. Intención de los mensajes en Twitter

En el estudio realizado por [16] se propone una taxonomía para clasificar los mensajes de Twitter dependiendo de su intención. Para dicho fin, el autor utiliza clasificación automática de textos, lo cual es un desafío en si mismo ya que por su naturaleza, los *tweets* son difíciles de clasificar por ser mensajes demasiado cortos. Al definir una taxonomía se pueden separar los mensajes de una manera más fácil y rápida para poder mejorar la toma de decisiones.

Las categorías definidas por el autor para clasificar un mensaje según su intención son las siguientes:

1. **Reporte de Noticia (RN):** Corresponde a una noticia emitida de manera objetiva y acompañada por un hipervínculo.

2. **Opinión de Noticia (ON):** Corresponde a una opinión sobre un reporte de noticia citando a la fuente y emitiendo un comentario en el mismo *tweet*.

3. **Publicidad (PU):** Similar en estructura a un reporte de noticia, pero con palabras que indican claramente que es una oferta o propaganda.

4. **Opinión General (OG):** A diferencia de una opinión de noticia, este tipo de opinión es sobre algún tema en particular. El autor no hace referencia a ninguna noticia.

5. **Compartir Ubicación / Evento (CU):** Estos mensajes suelen estar compuestos por el deseo del autor de dar a conocer su ubicación, acompañado de algún servicio de geolocalización que indica la ubicación exacta de la persona.

6. **Chat (CH):** Es una conversación entre uno o más usuarios, representado por una mención al principio del tweet.

7. **Pregunta (PR):** Puede ser una pregunta directa a algún usuario en particular, o una pregunta sin destinatario para ser respondida por cualquier seguidor del usuario.

8. **Mensaje Personal (MP):** Cualquier mensaje que no pertenezca a ninguna de las categorías anteriores. En general son *tweets* del tipo "Qué estoy haciendo" o "Qué estoy pensando".

El autor también propone dos clasificaciones jerárquicas, las cuales son agrupaciones de las categorías anteriores. Una de las jerarquías propuestas es la siguiente:

1. **Noticia (NO):** Es la agrupación de RN y PU.

2. **Opinión (OP):** Es la agrupación de ON y OG.

3. **Diario (DI):** Es la agrupación de MP y CU

4. **Social (SO):** Es la agrupación de CH y PR

El autor denominó la jerarquía anterior con el nombre de "Jerarquía I"[16]. El criterio de agrupación es por similitud de contenidos y estructuras entre estas categorías.

4.5. Problemas identificados en esta investigación

En general los estudios sobre influencia en Twitter reconocen algunas métricas que permiten identificar los atributos de un usuario influyente (Por ejemplo, es *retweeteado* y mencionado constantemente). Sin embargo, no se ha propuesto una métrica más completa para establecer qué usuario es más influyente que otro. Completa en el sentido de que no se base sólo en un criterio de medida, si no que pueda mezclar varios elementos. Por un lado que pueda basarse en más de un tipo de influencia, y por otra parte, que considere el comportamiento y atributos del usuario. Esto último podría responder preguntas como:

1. ¿Cómo darle valor a las interacciones que provienen de un usuario dependiendo de su estatus?

2. ¿Cómo restarle valor a los usuarios que no aportan un valor significativo a su entorno?

La primera pregunta se refiere a que no debería ser lo mismo si un usuario popular interactúa con uno regular y viceversa, ya que los distintos niveles de influencia que ambos tienen deberían considerarse. La segunda pregunta va dirigida a los usuarios que solo sólo se dedican a *retweetear* o mencionar a otros usuarios, pero que no producen *tweets* por iniciativa propia. Se pueden sacar conclusiones acerca del comportamiento e identificar algunos atributos de los usuarios utilizando los ratios descritos en el capítulo 4.3, y así ayudar a formar perfiles para cada usuario.

Por otro lado, se pueden modelar las relaciones entre usuarios a través de una red. Es más fácil ver la importancia de los usuarios que forman dicha red, la cual se puede corroborar de forma visual y objetivamente basándose en algunas propiedades topológicas de grafos.

Además se pueden utilizar herramientas de clasificación de texto para analizar el sentido de los mensajes de manera de poder decir si las influencias provocadas por un usuario son positivas o negativas. A lo último se le puede sumar la clasificación de la intención de los *tweets*, como se vio en el capítulo 4.5, para saber que tipo de *tweets* son los que emiten los usuarios influyentes.

Otro aspecto importante es establecer un modelo probabilístico para medir la influencia particularmente para Twitter. El modelo descrito en el capítulo 4.1 es para redes sociales en general, por lo tanto, se concentra en aspectos comunes, sin considerar los atributos diferenciadores de Twitter. Klout y similares utilizan modelos que supuestamente consideran algunas de las problemáticas expuestas en este capítulo. Sin embargo, el funcionamiento y los criterios considerados para medir la influencia, son desconocidos para el público.

5. Objetivo de la Investigación

5.1. Objetivo general

Desarrollar un modelo probabilístico basado en técnicas de grafos y apoyado en clasificación de textos para estimar rangos de influencia de distintos mensajes y usuarios en Twitter.

5.2. Objetivos específicos

1. Investigar sobre las métricas y modelos que se han establecido en relación al estudio de la influencia en Twitter.

2. Comprobar el rendimiento de tipos de influencia y *ratios* que ayudan a determinar el comportamiento y atributos de los usuarios.

3. Adaptar el modelo probabilístico propuesto por [8] para adecuarse las particularidades de Twitter.

4. Mejorar el modelo anterior considerando perfiles de usuario.

5. Construir las relaciones de influencia en un modelo basado en grafos, lo cual permitirá agregar una dimensión topológica al modelo.

6. Además del valor escalar de la influencia, incluir el sentido e intención de los mensajes. Esto implica analizar el contenido de los *tweets* y se considera como una extensión al modelo propuesto.

7. Comparar resultados con Klout y buscar ajustes en el modelo que se asimilen al Klout Score. Si bien Klout se basa en algoritmos y criterios desconocidos, puede aportar con dar una visión más amplia a los resultados de esta investigación.

6. Plan de Trabajo

Para cumplir con los objetivos propuestos, se destinaron diversas tareas. Las cuales se desarrollaron de manera secuencial entre mediados de octubre de 2011 y fines de agosto de 2012.

1. Recopilar información respecto a las métricas y modelos de influencia que existen en Twitter.
2. Construir un marco teórico que permita respaldar y comprender de manera más amplia la información recopilada.
3. Desarrollar una propuesta de solución a la problemática abordada en esta investigación.
4. Construir un *dataset* apropiado para poder aplicar el modelo propuesto.
5. Construir una red de relaciones entre usuarios del *dataset*.
6. Probar el modelo.
7. Clasificar los *tweets* emitidos por los usuarios del *dataset* para ver si influyen de manera positiva o negativa a la red de usuarios.
8. Clasificar los *tweets* emitidos por los usuarios del *dataset* por intención según una taxonomía.
9. Analizar los datos obtenidos, comparar resultados de influencia histórica con Klout y aplicar el modelo probabilístico de tiempo contínuo.
10. Buscar un ajuste similar a Klout para cada uno de los usuarios investigados.

7. Formulación del Modelo

El modelo probabilístico está dividido en tres partes o diemnsiones. Estas apuntan a representar la capacidad de influencia en forma númerica (así como lo hacen Klout y similares). Una cuarta parte, considerada una extensión, vendría a ser la que dicta si la influencia calculada tiene una tendencia positiva o negativa.

Las tres dimensiones principales del modelo son:

1. **Actividad:** La cual se refiere a la aplicación del modelo probabilístico de influencia conjunta y se basa exclusivamente en las acciones de los usuarios del *dataset*.

2. **Perfiles de usuario:** La cual pretende darle heterogeneidad a los usuarios del *dataset*.

3. **Topología de red:** La cual, a través de una red formada por los usuarios y sus relaciones en el *dataset* ayudará conocer la importancia de los usuarios, visto desde otra perspectiva.

7.1. Actividad en base al modelo probabilístico

A continuación algunos supuestos que servirán para diseñar la solución propuesta. De forma más especifica, el objetivo de esta sección es definir las consideraciones que se tomaron para transformar el modelo probabilístico de influenciabilidad conjunta en uno de influencia conjunta. Además se indican los ajustes necesarios para adaptar el modelo a las particularidades de twitter

7.1.1. Supuestos Iniciales

Considérese un grafo social, el cual es un grafo no dirigido $G = (V, E)$, donde los vértices V son usuarios, E es el conjunto de aristas entre usuarios. Una arista $(u, v) \in E$ representa un vinculo entre los usuarios u y v. El *timestamp* (τ) descrito en el capitulo 4.1 no se considerará por razones que se explicarán en breve.

Para el registro de acciones, *Acciones(Usuario, Acción, Tiempo)*, se seguirá considerando que los usuarios de este registro son los usuarios del grafo social, es decir, cada tupla representa una acción de uno de los usuarios definidos en el grafo social. Las acciones en Twitter pueden ser cuatro: Un *tweet*, una mención, un *reply* y un *retweet*.

Se define A_u como el numero de acciones realizadas por un usuario u. Por otro lado se define A_{v2u} como la suma de las acciones que se propagan desde un usuario v a u. En Twitter esto se podría traducir en un una mención o un *retweet* de u a v. Se dice que una acción $a \in A$ se propaga de un usuario vi_i a v_j siempre que:

1. $(v_i, v_j) \in E$. Es decir, v_j sigue a v_i.

2. $\exists(v_i, a, t_i)(v_j, a, t_j) \in$ *Acciones* con $t_i < t_j$. Esto quiere decir que v_i tiene que haber realizado la acción a antes que v_j.

Nótese que la tercera condición del modelo original consideraba el *timestamp* como prerrequisito para que una acción se pudiera propagar. Se optó por eliminar esta condición ya que en Twitter no es necesario seguir a alguien (o ser seguido por ese alguien) para poder *retweetearlo*

o mencionarlo. Por lo tanto, el tiempo en que ambos formaron un lazo social es irrelevante y no alterará a la formación del grafo de propagación

Con todo lo anterior, se define la relación $prop(a, v_i, v_j, \Delta t)$. Donde $\Delta t = t_j - t_i$. Con esta relación se puede definir un grafo de propagación PG para cada acción a, como $PG(a) = (V(a), E(a))$. Este grafo se compone de lo siguiente:

1. $V(a) = \{v | \exists t : (v, a, t) \in Acciones\}$, es decir, para formar el grafo de propagación tiene que haber ocurrido una acción, cual estaría registrada en el registro de acciones.

2. Existe una arista dirigida $v_i \xrightarrow{\Delta t} v_j$ en $E(a)$ cuando $prop(a, v_i, v_j, \Delta t)$.

7.1.2. Selección de un modelo probabilístico adecuado

Según [8], los modelos estáticos de créditos parciales y el de Bernoulli combinados son la opción más efectiva para el modelo de probabilidad conjunta. Sin embargo, en el caso de Twitter, el modelo de créditos parciales no es necesario por la sencilla razón de que Twitter es una red social donde las acciones de un usuario afectan a uno o a muchos usuarios, pero no se cumple lo contrario. Es decir, no se da el caso de que un grupo de usuarios pueda realizar una acción en conjunto y afectar a otro usuario, por lo tanto, las influencias se ejercen (para cada acción) en una relación de uno a uno y el usuario que influenció a otro siempre se llevaría el $100\,\%$ del crédito. Por lo tanto, si los créditos son siempre iguales, es conveniente volver al modelo de Bernoulli original de la ecuación 22, y de paso no gastar tiempo de cómputo en calcular créditos para cada acción.

7.1.3. Probabilidad de influencia conjunta

Para calcular la probabilidad de influencia conjunta se propone una sencilla modificación del modelo de probabilidad de influenciabilidad conjunta, algo así como poner el modelo anterior frente a un espejo. El modelo para influencia se define como sigue:

$$p_{u(S)} = 1 - \prod_{v \in S} (1 - p_{u,v}) \tag{29}$$

Donde $p_{u,v}$ también se basará en un modelo estático de Bernoulli, sin embargo, desde el punto de vista de la propagación de las acciones de u por sobre las de v. Se define esta probabilidad como sigue:

$$p_{u,v} = \frac{A_{u2v}}{A_u} \tag{30}$$

7.1.4. Propuestas alternativas a Bernoulli

El modelo estático de Bernoulli, tanto para influencia como para influenciabilidad, ayuda a definir la proporción intentos exitosos visto de las acciones totales que realiza el usuario que comienza el contagio. Sin embargo, sería igual de interesante ver cuál es la proporción de intentos exitosos sobre las acciones que realiza el usuario contagiado. Por lo tanto se define las variación de la probabilidad de la probabilidad de influencia alternativa a $p_{u,v}$ como:

$$p_{u,v} = \frac{A_{u2v}}{A_v} \tag{31}$$

7.2. Perfiles de usuario

Se determinó que dadas las características del problema a resolver y su contexto de aplicación, sólo uno de los ratios vistos anteriormente sería de utilidad para incorporar al modelo. El *ratio* followers/friends $r_f(u)$ descrito en el capítulo 4.3, ecuación 26, ayuda a responder la primera pregunta planteada en el capítulo 4. Esta métrica indica a grandes rasgos cuáles son los usuarios más populares en Twitter. Por lo tanto, en cuanto a las interacciones entre usuarios, esta métrica debería tomarse en cuenta ya que por ejemplo, no sería lo mismo ser *retweeteado* por un usuario popular que por un amigo (considerando que el amigo no es un usuario influyente en Twitter). No se va a usar el ratio tal cual en el modelo. Se van a hacer un par de modificaciones, de tal forma que este ratio quede normalizado según los usuarios pertenecientes al *dataset* a analizar. Por otro lado, los valores normalizados irán multiplicados en vez de divididos, pudiendo así castigar aún más a los usuarios con pocos *followers* y pocos *friends* a la vez. El producto de lo mencionado anteriormente se describe en la ecuación 32:

$$\frac{Followers}{Followers_{Max}} * \frac{Friends}{Friends_{Max}} \tag{32}$$

En el producto anterior, se multiplican los *friends* y *followers* de un usuario u, pero también estos valores van normalizados por los máximos valores de cada variable, los cuales se buscan entre los usuarios pertenecientes al *dataset*.

7.3. Topología de redes de usuario

Más allá de usar el *dataset* como punto de partida para hacer los cálculos del modelo, también se modelaron las relaciones entre usuarios con un modelo basado en grafos. En esta red, los usuarios representan los nodos y los enlaces son sus relaciones más representativas. Estos enlaces representan tres tipos de acciones en Twitter: *Retweets*, Menciones y *Replies*. Este grafo es un grafo dirigido, con las aristas representado la dirección en que se propagan las acciones, por ejemplo, si v le responde a u dicha acción será representada por una arista que va desde u a v. Los nodos con mayor influencia (determinada tras aplicar el modelo en cada usuario perteneciente a la red) serán de mayor tamaño, por otro lado se agregarán tres propiedades topológicas al modelo, las cuales son: el tamaño de la componente a la que pertenece un nodo, el *PageRank* del nodo y su *betweenness centrality*. Al igual que en el caso de los perfiles, estas propiedades topológicas serán normalizadas por los valores máximos que hay en la red de usuarios analizada. Un ejemplo de red de usuarios se verá tras aplicar el modelo en un próximo capítulo.

7.4. Modelo final

Con todas las consideraciones descritas anteriormente en este capítulo, se describe a continuación el modelo final que considera la magnitud de la capacidad de influencia. El modelo de forma simplificada sería el siguiente:

$$(\alpha * Actividad + \beta * Topologia + \gamma * Perfil) * 100 \tag{33}$$

Donde α, β y γ son valores arbitrarios que se mueven entre 0 y 1 y entre los tres no suman mas que 1 ya que su fin es ponderar la actividad de los usuarios, la topología de la red y el perfil de cada usuario. Todo está multiplicado por 100 para que la influencia quede representada en un puntaje 0 a 100. El modelo detallado queda de la siguiente forma:

$$\alpha*(p_{u(S)}*\frac{K}{K_{Max}})+\beta(\frac{Csize}{Csize_{Max}}*(\frac{\frac{BET}{BET_{Max}}+\frac{PR}{PR_{Max}}}{2}))+\gamma(\frac{Followers}{Followers_{Max}}*\frac{Friends}{Friends_{Max}})*100 \tag{34}$$

Donde K representa la cantidad de enlaces (*degree*) de u. $Csize$ es el tamaño de la componente a la que pertenece u en la red, BET y PR serían el *betweenness centrality* y *PageRank* de u (con un factor de amortiguación de $0{,}85$). Con excepción de $p_{u(S)}$ todos los valores están normalizados por sus máximos correspondientes en la red.

En el caso del modelo tiempo continuo basta con reemplazar $p_{u(S)}$ con $p_{u^t(S)}$ en el modelo anterior.

7.5. Sentido e intención de los *tweets*

Los *tweets* emitidos por los usuarios a analizar serán clasificados por sentido (positivo, negativo y neutro) y por intención, según la jerarquía vista en el capítulo 4.5 (Noticia, Opinión, Social y Diario). La idea es que además del valor numérico de la influencia de un usuario, se incluya un detalle con el porcentaje de *tweets* positivos, negativos y neutros, así como el porcentaje de *tweets* por categoría. En estricto rigor, y como se dijo anteriormente, esto último no forma parte del modelo en sí, pero sirve para tener una visión más amplia de la influencia de los usuarios y no estrictamente numérica.

8. Obtención del *Dataset* de Prueba

Para probar el modelo se optó por aplicarlo en base a algún contexto en particular, en este caso, la política. Por lo tanto se construyó un *dataset* en torno a cuatro políticos pertenecientes al senado de Chile, los cuales son: Guido Girardi, Ena von Baer, Jovino Novoa y Soledad Alvear. Estos cuatro senadores pertenecen a las circunscripciones VII y VIII del senado, y representan a la Región Metropolitana. Los cuatro senadores tienen cuentas en Twitter.

8.1. Características del *dataset*

Para armar el *dataset*, se usó como criterio de búsqueda *tweets* que mencionaran a cada uno de los senadores, tanto sus nombres de usuario como sus nombres y apellidos. Se extrajeron 11036 *tweets* emitidos por 5681 usuarios entre el 7 de abril de 2012 y 8 de junio de 2012. Además se registraron los Klout Score diarios de los cuatro senadores entre el 10 de mayo de 2012 y 8 de junio de 2012.

Alrededor de 350 *tweets* emitidos por los cuatro políticos y algunos usuarios que los mencionaron entre el 7 de abril de 2012 y 9 de mayo de 2012 fueron destinados para armar el *training set* de sentido. Estos fueron clasificados manualmente por el criterio mencionado.

Para clasificar los *tweets* por intención se uso el mismo *training set* del autor de [16]. El cual es un conjunto de 2200 *tweets* clasificados manualmente. En promedio son 300 *tweets* por cada una de las ocho categorías descritas en el capitulo 4.5.

8.2. Método de obtención y limitaciones en el proceso

Para obtener el primer *dataset* y el *training set* de sentido se utilizó la *API* de Twitter, la cual se integró con el lenguaje de programación Python. Se hizo una adaptación de algunos scripts propuestos en [17], para realizar dicha tarea. La *API* presenta algunas limitaciones, las cuales se presentan a continuación:

1. Las consultas de búsqueda tienen un limite de 1500 *tweets* y hasta una semana de antigüedad. Cuando se cumpla cualquiera de estas dos condiciones la extracción se detendrá.

2. Solo se puede extraer un máximo de 100 *followers* y 100 *friends* por usuario.

3. La red de Twitter es inestable, por lo tanto se pueden presentar caídas en plena ejecución, obligando a empezar de nuevo el proceso de extracción.

Para lidiar con esas limitaciones se optó por extraer los *tweets* de forma periódica y automatizada, sin abusar de los limites impuestos por la API, los resultados se almacenaron en una base de datos MySQL con tres tablas, las cuales se describen a continuación.

La tabla de la figura 5 muestra los *tweets* extraídos, sus autores y la fecha en que emitieron los mensajes. Estos *tweets* son almacenados continuamente a lo largo del día. La tabla de la figura 6 es una versión procesada de la tabla anterior y se utiliza para formar la red de usuarios. Esta tabla cuenta con un par de tuplas (autor de origen, autor de destino), el cual sirve para visualizar la propagación del mensaje, así como también el tipo de mensaje. El tipo de mensaje se clasifica en *Retweet*, Mención, *Reply* y Mensaje Propio (mensaje no dirigido a ningún usuario). La tabla

de la figura 7 es el resultado tras aplicar el modelo diariamente a partir de los datos de la tabla de red de usuarios. Con esto se puede tener un registro histórico de la influencia de cada usuario día a día. Los datos que se guardan aquí son variables que se ocupan en el modelo, los máximos de cada variable pueden calcularse a través de consultas SQL. Los datos fueron almacenados en los servidores de AnaliTIC, una empresa que se especializa en la extracción y análisis de datos.

fecha_publicacion	nombre_autor	cuerpo
2012-03-08 00:09:36	alejandroriosw	?@Ale_K12: @alejandroriosw @guidogirardi y Fulvio ...
2012-03-08 00:02:08	soledadalvear	@kerosut Lamentablemente se requiere para solucion...
2012-03-08 00:00:35	soledadalvear	@SOLFEOPROD Lamentablemente es necesario
2012-03-08 00:08:48	borisrifo	RT @GladysVeral: Ojalá que no veamos nunca a @guid...
2012-03-08 00:07:42	CBecerra01	Señor @guidogirardi, Lavanderos siempre será un de...

Figura 5: Tabla de *tweets*

user_1	user_2	id_conversa	id_etiqueta	id_articulo	Tipo	fecha	id_ref
frann_cisk	*NULL*	39239	73	431189	MP	2012-03-21	29803
Renolander	AbrilDelarge	39240	73	431190	RT	2012-03-21	29804
Megariu5	*NULL*	39238	73	431188	MP	2012-03-21	29802
serny8688	*NULL*	39237	73	431187	MP	2012-03-21	29801
Pablo_Donoso	Eriolo	39236	73	431186	RT	2012-03-21	29800

Figura 6: Tabla de red de usuarios

usuario	actividad	k	ccsize	pagerank	betweenness	fecha	friends_count	followers_count
RodrigoLeonSBDO	0	2	17	0.0470588	0	2012-02-15	52	15
chahuan	0	1	17	0.0515033	0	2012-02-15	926	23527
senadornavarro	0	1	17	0.0515033	0	2012-02-15	149	41237
aleguillier	0	1	17	0.0603922	0	2012-02-15	0	0
MarcelusMaximu	0.666667	9	17	0.0470588	0	2012-02-15	295	337

Figura 7: Tabla de influencia histórica

9. Aplicación del Modelo

Hay dos funciones de la *API* de twitter que sirvieron para construir el dataset. Estas son [18]:

1. **GET users/lookup:** Retorna un *JSON* con los datos de un usuario a partir de su ID único de Twitter. El campo de interés en este caso es el nombre de usuario.

2. **GET search:** Retorna un *JSON* con los datos de *tweets* que han sido recuperados de una búsqueda. Si se antecede el término *from:* a la búsqueda, se pueden filtrar los *tweets* que provienen de un determinado usuario. Ejemplos de búsqueda pueden ser términos como: *"Soledad Alvear"*, *"from:guidogirardi"* y cualquier otra palabra clave.

El algoritmo que se describe a continuación recibe como entrada un id asociado a los *tweets* que mencionan a cada senador, una fecha de inicio, y una fecha de término. Con esto se forma una selección de usuarios S, y finalmente se retorna la influencia de cada usuario perteneciente a dicha selección. El algoritmo se muestra en la figura 8.

$S \leftarrow GETvecindad$
for all $u \in S$ **do**
  for all $tweet \in u$ **do**
    $A_u \leftarrow A_u + 1$
  end for
  for all $v \in S_{(u)}$ **do**
    for all $tweet \in v$ **do**
      $A_v \leftarrow A_v + 1$
      if $destino(tweet) = screen\_name(u)$ **then**
        $A_{u2v} \leftarrow A_{u2v} + 1$
      end if
    end for
    for all $tweet \in u$ **do**
      if $destino(tweet) = screen\_name(v)$ **then**
        $A_{v2u} \leftarrow A_{v2u} + 1$
      end if
    end for
  end for
  print $(\alpha * Actividad + \beta * Topologia + \gamma * Perfil) * 100$
end for

Figura 8: Algoritmo de aplicación del modelo

Donde $destino(tweet)$ devuelve el nombre del usuario a quien se le ha respondido, mencionado o *retweeteado* un *tweet*. $screen\_name(u)$ devuelve el nombre de usuario y $GETvecindad$ forma la vecindad S a partir de un id asociado al senador y un intervalo de fechas, como se dijo anteriormente. Para reducir el ruido en los cálculos, se pensó que sería adecuado que la vecindad para cada usuario u estuviera compuesta sólo por los usuarios v más próximos, es decir los que pertenecen a su componente. Por lo tanto, se define $S_{(u)}$ como un subconjunto de S.

10. Resultados e Interpretación

10.1. Redes de usuario e influencia histórica

Con el *dataset* anterior se realizaron pruebas con los cuatro políticos entre el 7 de abril de 2012 y 8 de junio de 2012. Se armaron redes de usuario, gráficos de influencia histórica y gráficos de decaimiento de influencia utilizando el modelo de tiempo continuo. Para crear las redes, se uso la librería D3. Dicha librería está construida sobre jQuery, por lo tanto, se pueden visualizar los resultados por medio de un navegador web. Para crear las relaciones de los nodos, así como el el *PageRank* y el *Betweenness* de cada nodo, se uso una librería de Python llamada NetworkX, la cual devuelve un archivo JSON con todas las relaciones de los nodos y sus parámetros. Para todas las redes a continuación se usaron valores de α,β y γ iguales a 0,33 y el modelo estático alternativo a Bernoulli. En la figura 9 se ve una red de usuarios de Guido Girardi para los días entre el 7 y 9 de abril de 2012. Se puede ver que hay una componente bastante grande alrededor del usuario *guidogirardi* (que en efecto, es el senador Girardi), y por otro lado una componente mucho menor alrededor *Guido_Girardi*, el cual no es el senador, si no que un usuario con alcance de nombres. Como se puede ver, *guidogirardi* tiene una influencia tremenda, lo que es evidente al ver el tamaño de su nodo. Al hacer click sobre un nodo, se pueden ver sus detalles, entre ellos su influencia para esa fecha (que está sobre su imagen de perfil). El gráfico que aparece es su influencia histórica para las redes diarias que fueron generadas en esa fecha, es decir, influencias para el 7,8, y 9 de abril respectivamente. Estos detalles se muestran en la figura 10.

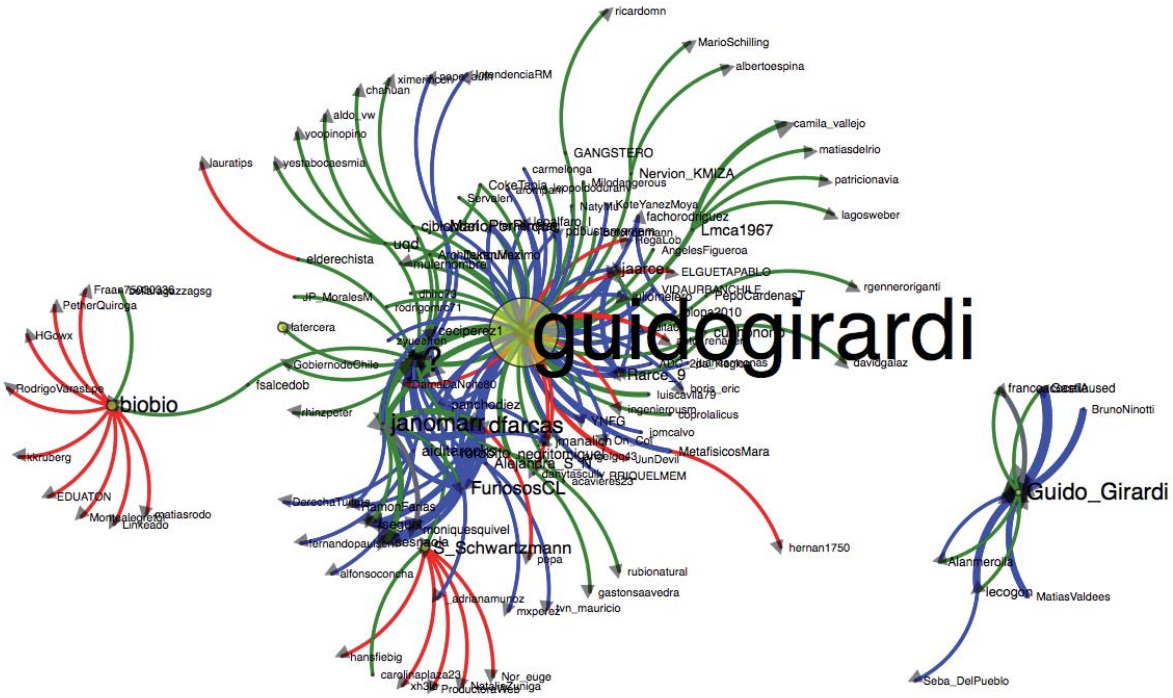

Figura 9: Ejemplo de red de usuario. Red de usuarios del senador Guido Girardi para una determinada fecha

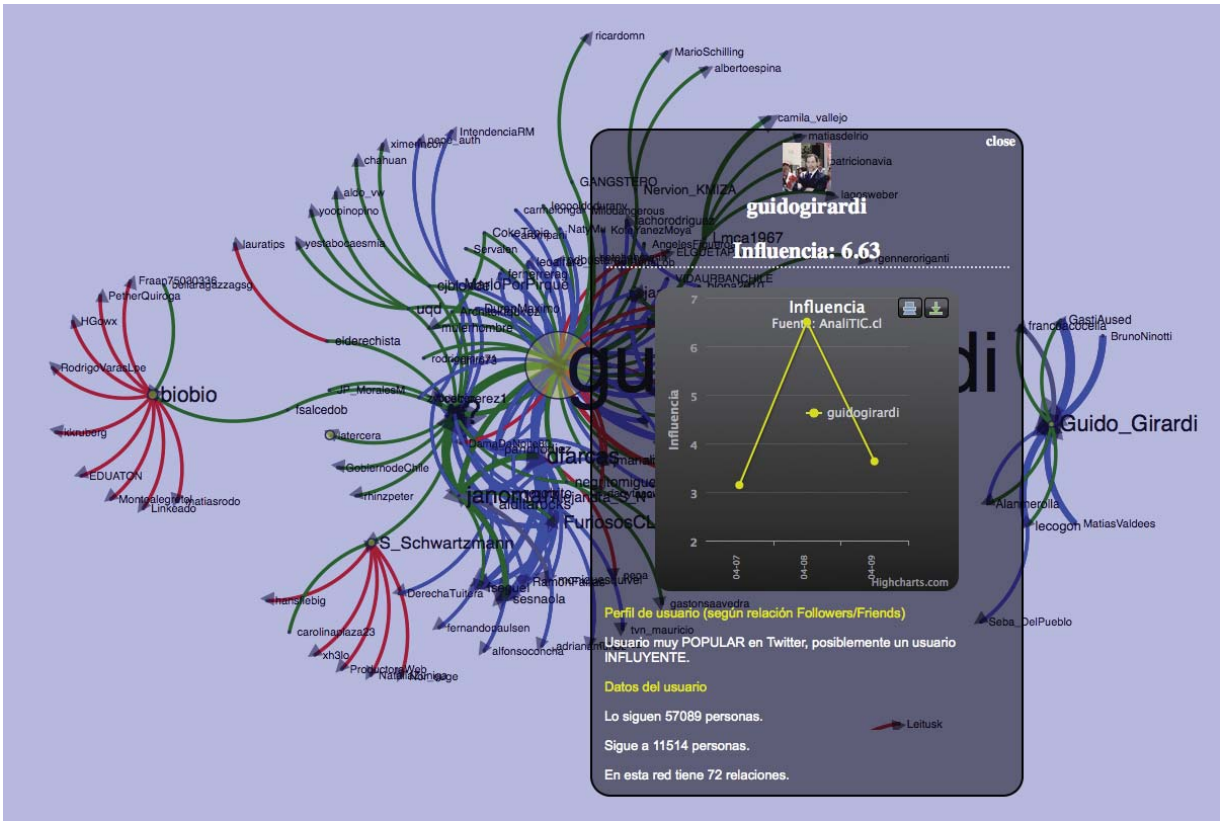

Figura 10: Ejemplo de red de usuario con detalle. Al hacer click sobre el nodo del senador Girardi aparecen sus detalles, entre ellos, su influencia.

10.2. Resultados preliminares

Antes de explicar los gráficos que vienen a continuación, cabe destacar que Soledad Alvear es quien más *tweetea* con la sorprendente suma de 789 *tweets* en aproximadamente 1 mes, muy por detrás es seguida por Guido Girardi con 64 *tweets*, Ena Von Baer, con 59 *tweets* y finalmente Jovino Novoa con sólo un *tweet*. En promedio de todos los días observados Soledad Alvear tiene 38677 seguidores y 792 amigos aproximadamente. Guido Girardi tiene 57084 seguidores y 11514 amigos. Ena von Baer tiene 162679 seguidores y 685 amigos. Finalmente, Jovino Novoa tiene 3547 seguidores y 43 amigos. Estas cifras serán importantes en la determinación de qué tan importante son la actividad y los datos de perfil en el cálculo de la influencia.

Para todos los gráficos en lo que queda de esta sección se usó el modelo estático alternativo a Bernoulli. En el siguiente apartado se usaron valores de α,β y γ iguales a 0,33.

10.2.1. Influencia histórica

Las figuras 11, 12, 13 y 14 muestran la influencia histórica diaria de los cuatro senadores en el intervalo de fechas del *dataset*. Además estos cuatro gráficos están contrastados con el Klout Score de cada senador. Nótese que en todos los casos Klout es mucho más constante que

el modelo probabilístico, el cual tiene alzas y bajas mucho más bruscas, lo que hace pensar que Klout puede estar usando algún método para suavizar sus curvas de influencia histórica.

Las figuras 15, 16, 17 y 18 muestran la curva del modelo un poco más suavizada. Para esto se ponderaron las influencias históricas con una media móvil simple de 3 períodos. Fue necesario recolectar la influencia 3 días antes que en el caso de Klout, es decir, desde el 7 de mayo. Hay dos aspectos interesantes que destacar. Lo primero es que la curva suavizada se acerca un poco más a la curva de Klout en todos los casos y los valores en el caso del modelo dejan de ser tan extremos. Lo segundo es que se puede tener la influencia de un día posterior a la fecha de finalización del *dataset*, ya que puede ser predicha por los datos de los últimos tres períodos.

Las figuras 19, 20 y 21 permiten ver la influencia en términos relativos, es decir, que tan influyente es un senador respecto de los otros cuatro. Es evidente por los gráficos que Novoa es el senador menos influyente de los cuatro, muy alejado de los otros tres, que están muy cercanos entre sí. Es importante destacar que las tendencias se mantienen en los tres gráficos, y sólo en el gráfico de influencia relativa hay más cercanía y valores extremos que en los otros dos.

Finalmente, en la figura 22 se muestra la aplicación del modelo continuo de cálculo de influencia para demostrar como disminuye esta al pasar el tiempo. En este caso se seleccionó al senador Girardi y el día 26 de marzo. Día en el que, según la curva no suavizada, su influencia bordeó los 70 puntos. Se estimó que sería prudente un tiempo promedio de respuesta de 60 segundos para los usuarios que reciben los *tweets* de Girardi en sus *timelines* y para este caso se vio como varía la influencia en 5 minutos con intervalos de 30 segundos. Los resultados indican que la influencia decae de forma suave de 70 hasta un poco menos de 40 puntos lo que parece indicar que si bien el modelo de tiempo continuo tiene un decaimiento exponencial, las partes de topología y perfiles en el modelo hacen que el decaimiento sea más suave. Sin embargo esta es una instancia particular, y los resultados podrían ser muy distintos cambiando α,β y γ, o cambiando el contexto de aplicación del modelo.

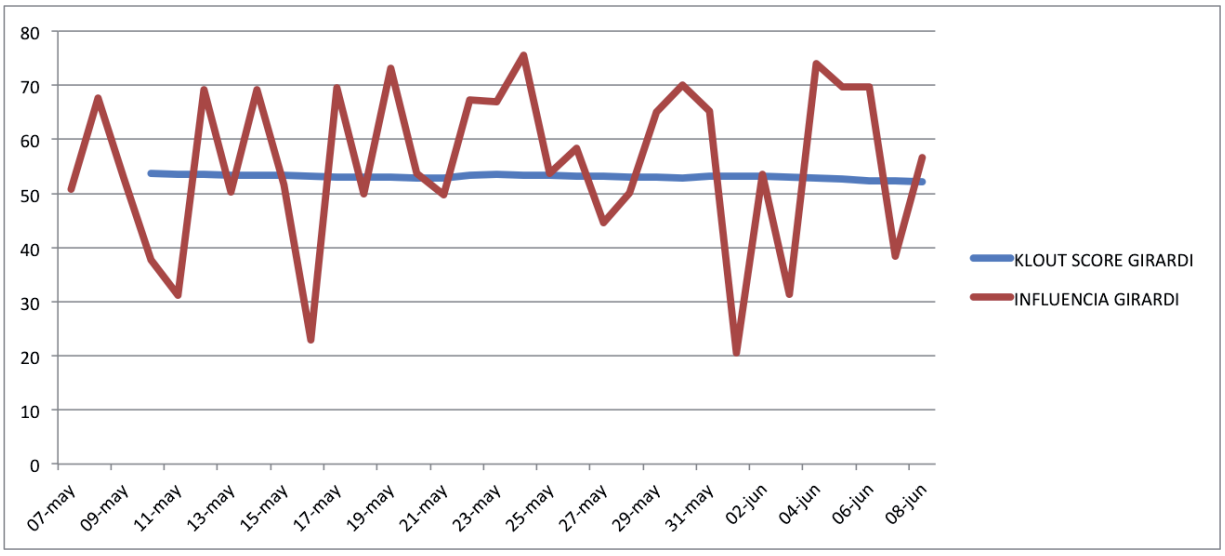

Figura 11: Influencia contrastada con Klout Score de Guido Girardi.

Figura 12: Influencia contrastada con Klout Score de Ena Von Baer.

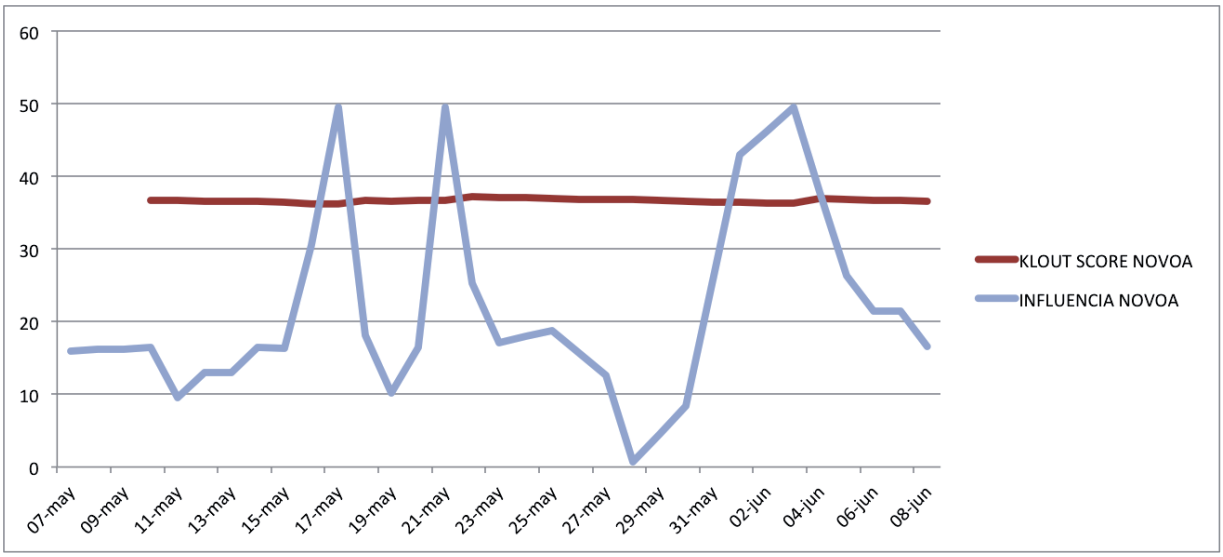

Figura 13: Influencia contrastada con Klout Score de Jovino Novoa.

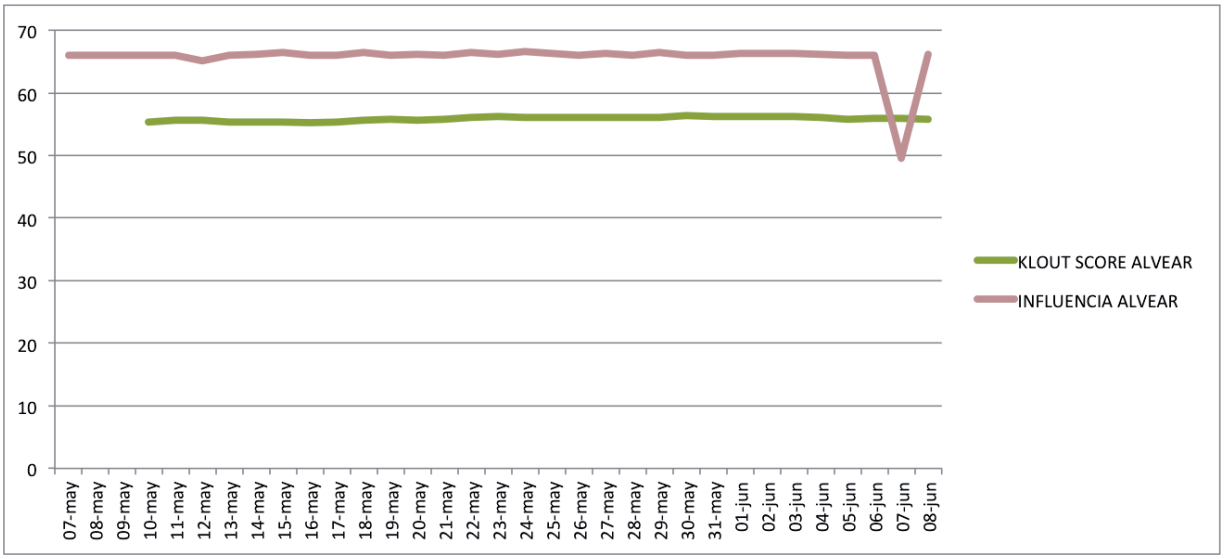

Figura 14: Influencia contrastada con Klout Score de Soledad Alvear.

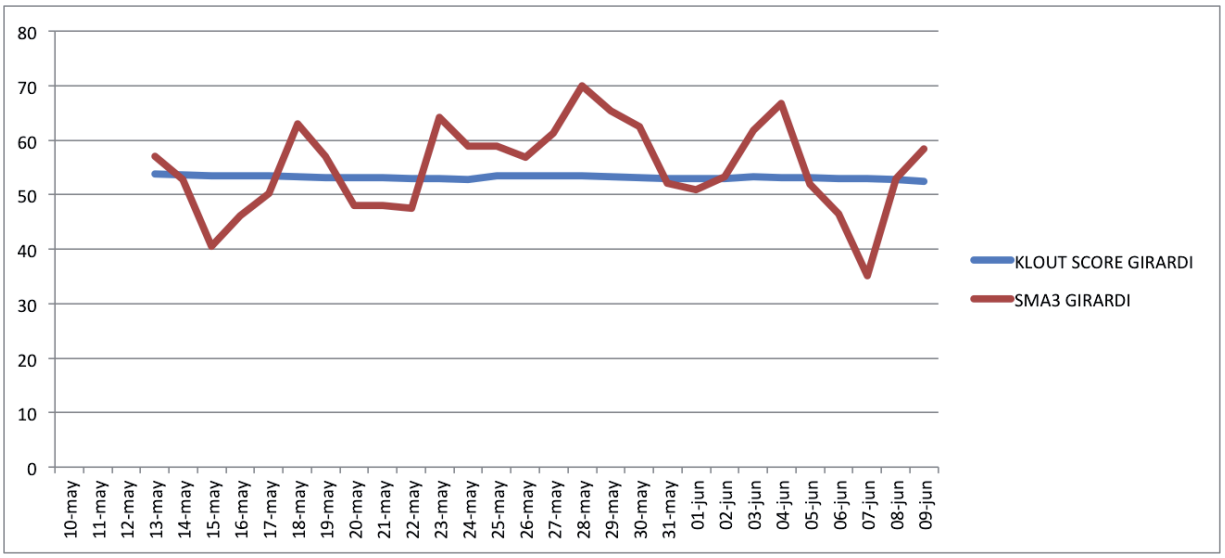

Figura 15: Influencia de Guido Girardi con media móvil simple de 3 períodos contrastada con Klout Score.

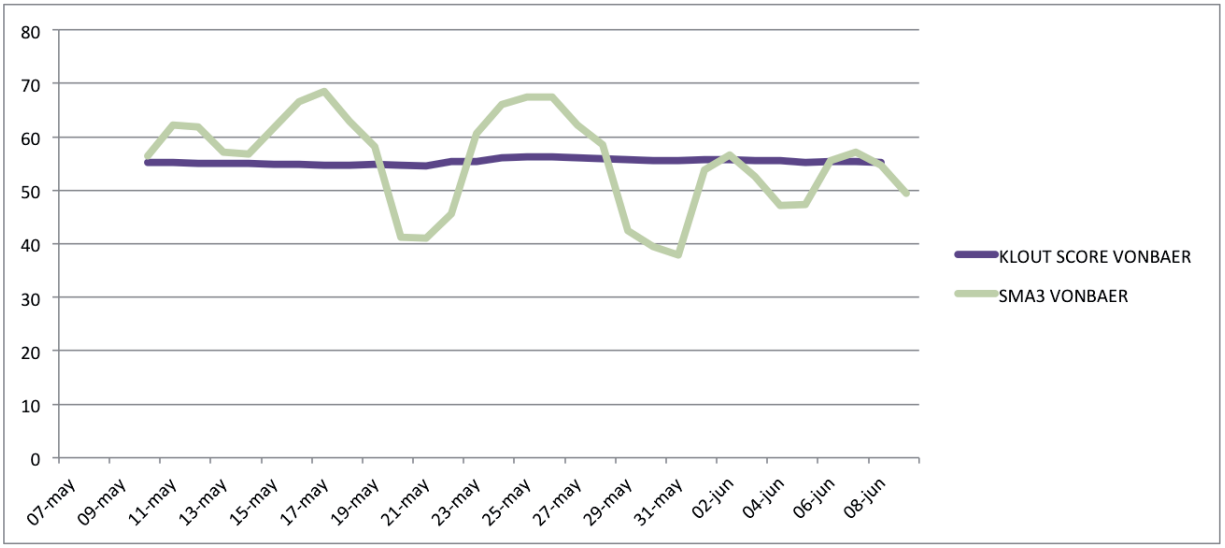

Figura 16: Influencia de Ena Von Baer con media móvil simple de 3 períodos contrastada con Klout Score.

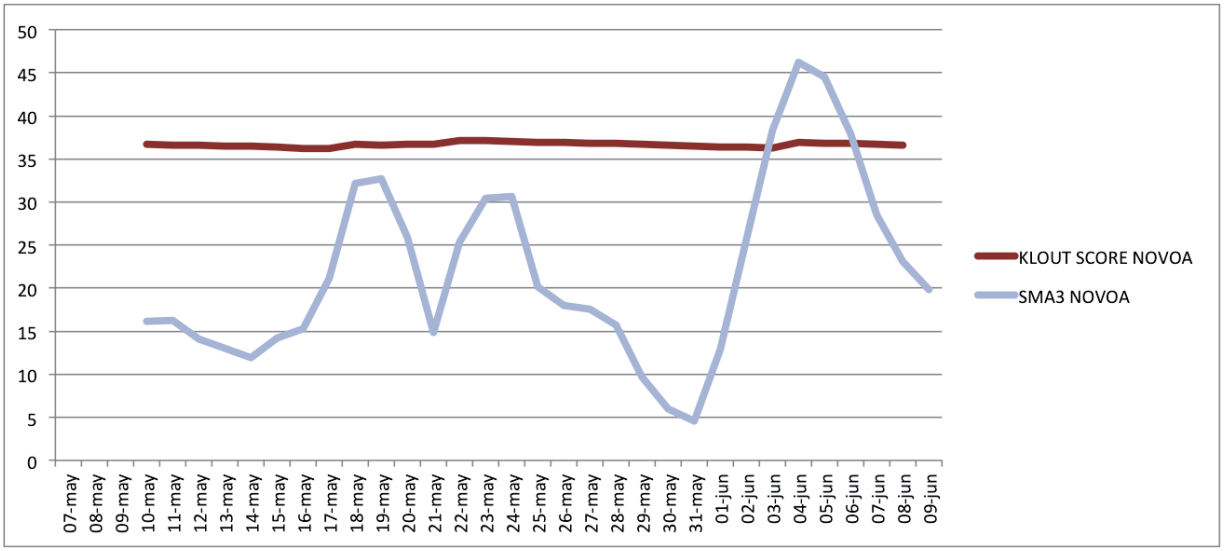

Figura 17: Influencia de Jovino Novoa con media móvil simple de 3 períodos contrastada con Klout Score.

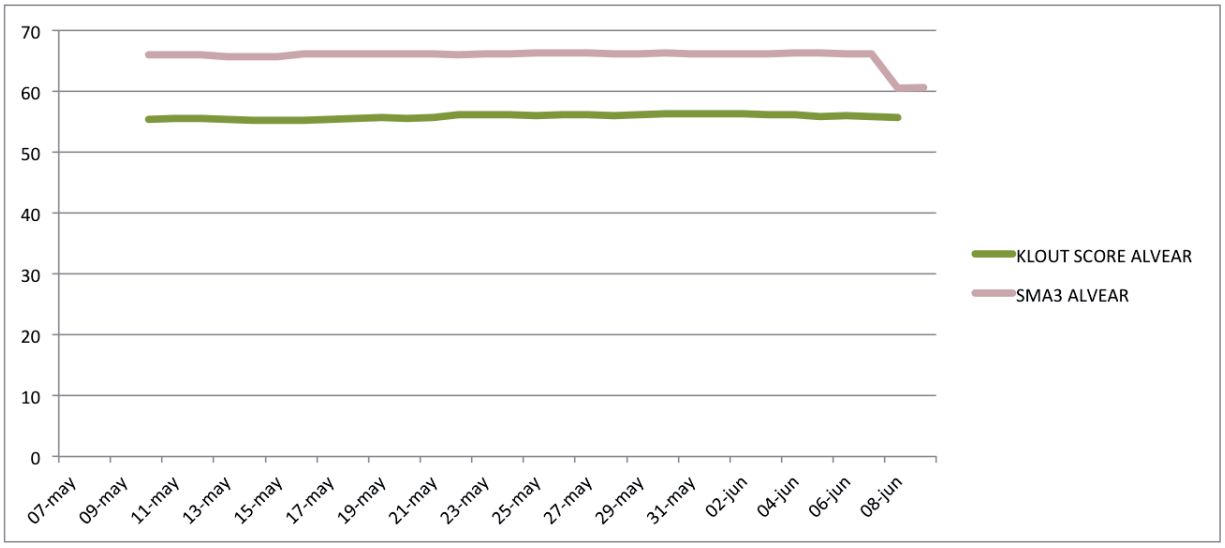

Figura 18: Influencia de Soledad Alvear con media móvil simple de 3 períodos contrastada con Klout Score.

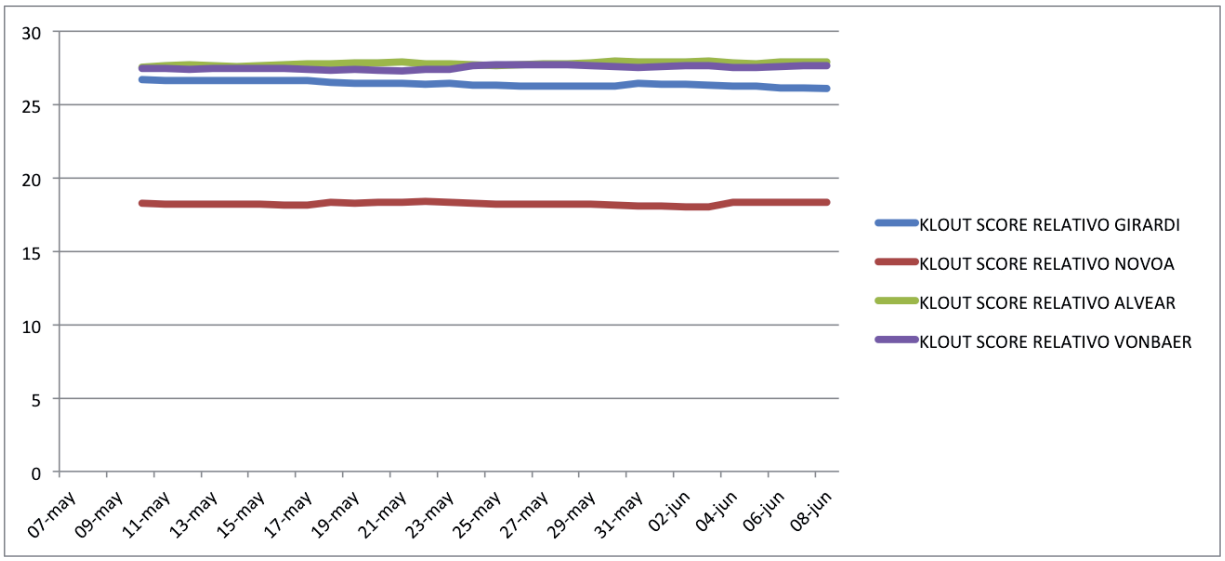

Figura 19: Klout Score relativo.

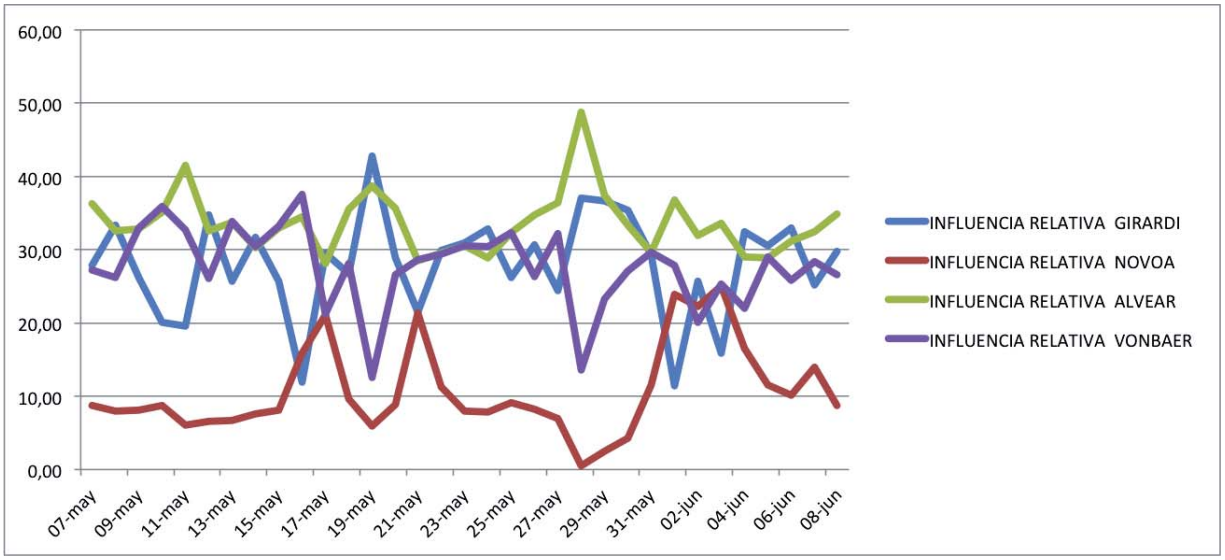

Figura 20: Influencia Relativa.

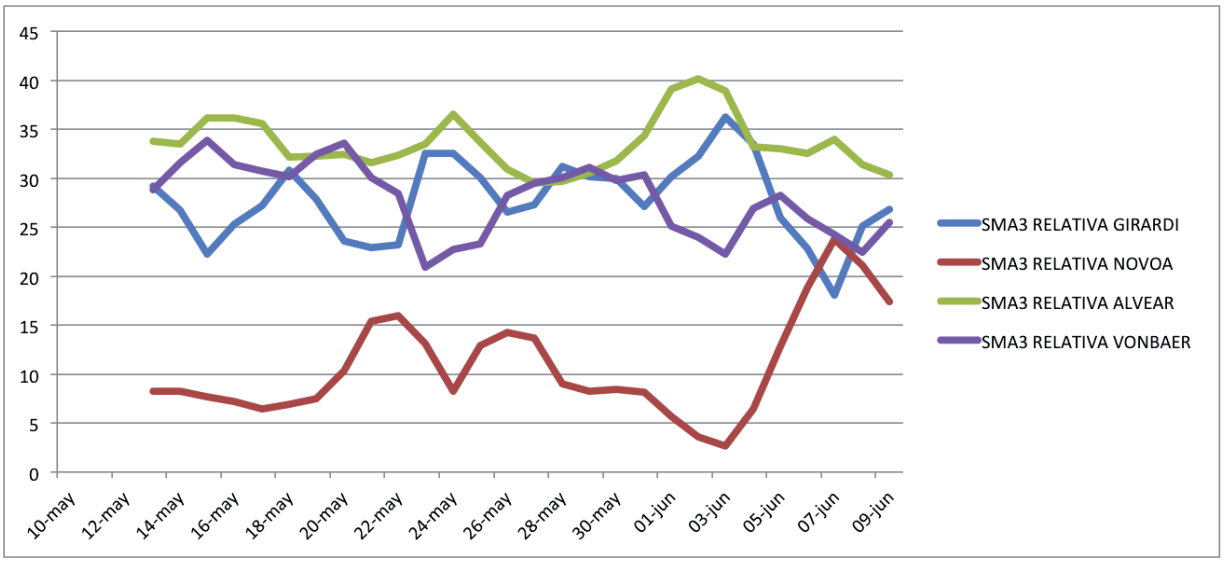

Figura 21: SMA3 relativa.

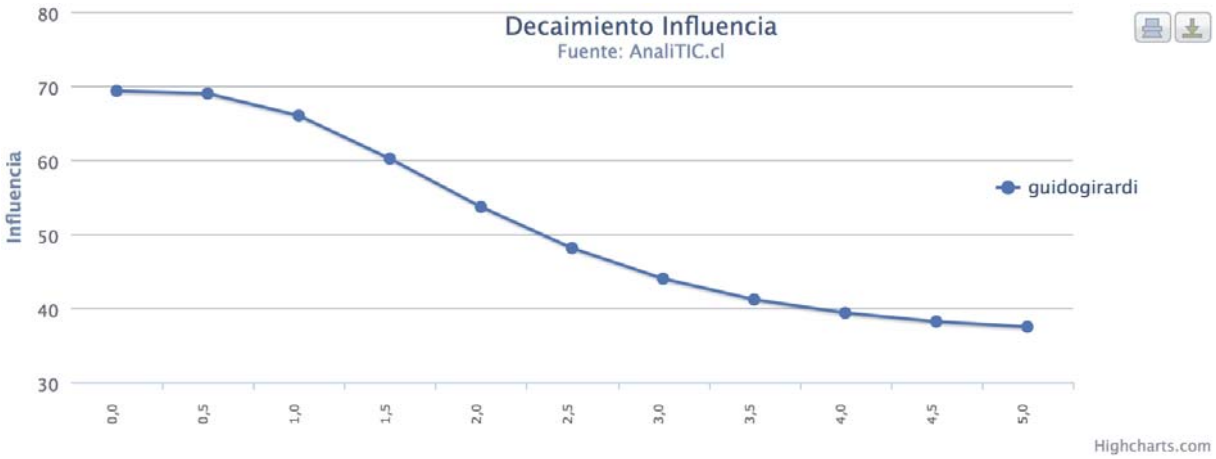

Figura 22: Decaimiento de la influencia de Guido Girardi. Se muestra como decae la influencia de Guido Girardi el día 26 de marzo con el modelo de tiempo continuo entre 0 y 5 minutos, con intervalos de 30 segundos y un tiempo promedio de respuesta esperado de 60 segundos.

10.2.2. Modificación de parámetros y análisis de correlación

Para ver como era la sensibilidad de α, β y γ en el modelo, se hicieron pruebas sobre la media y análisis de correlación de Pearson para cuatro casos:

1. Con α, β y γ iguales a $0{,}33$. Es decir, totalmente balanceado.
2. Con $\alpha = 0{,}5$, $\beta = 0{,}25$ y $\gamma = 0{,}25$. Es decir, levemente cargado hacia la actividad.
3. Con $\alpha = 0{,}25$, $\beta = 0{,}5$ y $\gamma = 0{,}25$. Es decir, levemente cargado hacia la topología de red.
4. Con $\alpha = 0{,}25$, $\beta = 0{,}25$ y $\gamma = 0{,}5$. Es decir, levemente cargado hacia los perfiles de usuario.

Para las siguientes tablas considérese la variable $K_1, K_2, K_3, ...K_{30}$ como la variable que representa el Klout Score de cada uno de los 30 días observados (desde el 10 de mayo hasta el 8 de junio). La variable $I_1, I_2, I_3, ...I_{30}$ representa la influencia para esos días, y la variable $S3_1, S3_2, S3_3, ...S3_{30}$ sería la influencia anterior con media móvil simple de 3 períodos. $\overline{k}$, $\overline{i}$, $\overline{s3}$ serían las medias de las tres variables propuestas. La correlación de Pearson entre K e I está determinada por r_{KI} y la correlación entre K y $S3$ se encuentra determinada por r_{KS3}. Las abreviaciones 0.33, Act. 0.50, Top. 0.50 y Per. 0.50 representan respectivamente a los ítems 1,2,3 y 4 de la lista anterior.

Las tablas 7, 8, 9 y 10 presentan varios hechos interesantes. En primer lugar el análisis de correlación de Pearson da bastante bajo en casi todos los casos (menor a $0{,}5$). Solo en Novoa da resultados levemente más cercanos a cierto tipo de correlación, pero en el resto se podría decir que casi no hay correlación entre la influencia histórica del modelo con la de Klout. Sin embargo para el caso de las medias hay resultados curiosos, en especial en el caso en que el modelo está balanceado. Girardi y Alvear difieren muy poco en comparando las medias con las de Klout, practicamente en 2 puntos. En el caso de Alvear la diferencia es un poco mayor pero aún así se acerca bastante a la estabilidad de Klout. En el caso de Novoa, su influencia promedio en términos del modelo es bastante menor que en términos de Klout. Esto podría deberse a la casi nula actividad que ha tenido Novoa en la muestra. Otro punto importante a destacar es como algunos senadores varian sus promedios con distintos parametros. Todos bajan su influencia promedio cuando se carga el modelo hacia γ, pero no hay un consenso entre α y β. Sin embargo en todos los casos, un α o β mayor significa un mayor promedio de influencia que con los tres parámetros balanceados.

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	r_{KI}	r_{KS3}
0.33	53.1	55.3	55.61	-0.09	-0.11
Act. 0.50	53.1	61.76	62.04	-0.1	-0.02
Top. 0.50	53.1	59.12	59.42	-0.05	-0.18
Per. 0.50	53.1	46.68	47.05	-0.09	-0.14

Tabla 7: Modificación de parámetros para Guido Girardi

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	r_{KI}	r_{KS3}
0.33	55.35	54.87	55.57	-0.108	-0.108
Act. 0.50	55.35	64.22	65.12	-0.092	-0.097
Top. 0.50	55.35	59.11	59.98	-0.13	-0.11
Per. 0.50	55.35	42.96	43.29	-0.08	-0.09

Tabla 8: Modificación de parámetros para Ena Von Baer

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	r_{KI}	r_{KS3}
0.33	36.66	22.28	22.09	-0.39	0.3
Act. 0.50	36.66	22.32	22.26	-0.47	0.23
Top. 0.50	36.66	26.96	26.93	-0.30	0.37
Per. 0.50	36.66	18.24	17.76	-0.29	0.33

Tabla 9: Modificación de parámetros para Jovino Novoa

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	r_{KI}	r_{KS3}
0.33	55.86	65.59	65.95	-0.008	0.14
Act. 0.50	55.86	74.66	74.94	0.01	0.19
Top. 0.50	55.86	74.27	74.82	-0.001	0.09
Per. 0.50	55.86	49.82	50.09	0.02	0.17

Tabla 10: Modificación de parámetros para Soledad Alvear

10.3. Resultados posteriores

Si bien los resultados anteriores entregaron información valiosa, se consideró que se podían hacer más pruebas para poder tener una visión aún más acabada del comportamiento de los usuarios observados. Además, en la mayoría de los casos, la curva suavizada con una media móvil de 3 períodos sigue siendo distante en comparación a los resultados más estables de Klout. En la figura 23 se probó con medias móviles simples de 9 y 30 períodos. Se puede ver como a medida que se le aumentan los períodos a la media móvil simple de Guido Girardi, se va asemejando más a Klout y la estabilidad del puntaje va a aumentando.

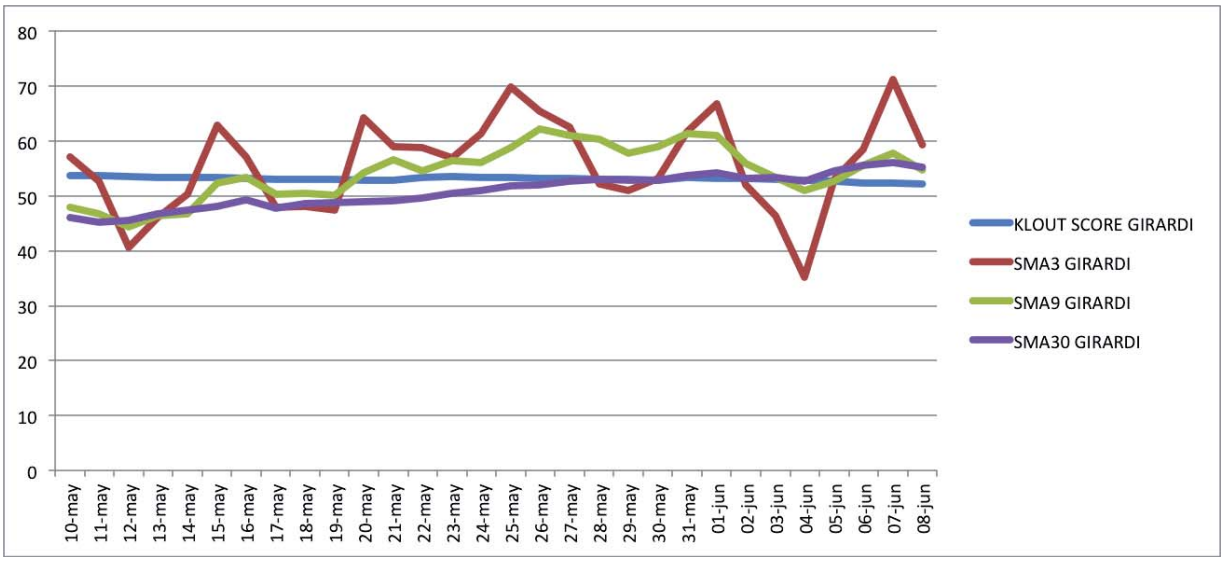

Figura 23: Influencia de Guido Girardi con media móvil simple de 3,9 y 30 períodos contrastado con Klout Score.

10.3.1. Influencia histórica con nueva media móvil simple

Las siguientes pruebas se hicieron con una media móvil simple de 30 períodos, ya que se tiene la sospecha de que Klout usa esta media considerablemente alta para mantener sus resultados tan invariables. Se calcularon las medias y las correlaciones de estos datos. En la búsqueda de información valiosa y de resultados aún más parecidos con Klout, se probó con nuevas ponderaciones. A los casos vistos previamente se sumaron los siguientes:

1. Con $\alpha = 1$, $\beta = 0$ y $\gamma = 0$. Es decir, totalmente cargado hacia la actividad.
2. Con $\alpha = 0$, $\beta = 1$ y $\gamma = 0$. Es decir, totalmente cargado hacia la topología de red.
3. Con $\alpha = 0$, $\beta = 0$ y $\gamma = 1$. Es decir, totalmente cargado hacia los perfiles de usuario.
4. Con $\alpha = 0{,}5$, $\beta = 0{,}5$ y $\gamma = 0$. Es decir, cargado equitativamente hacia la actividad y la topología de red.

5. Con $\alpha = 0{,}5$, $\beta = 0$ y $\gamma = 0{,}5$. Es decir, cargado equitativamente hacia la actividad y los perfiles de usuario.

6. Con $\alpha = 0$, $\beta = 0{,}5$ y $\gamma = 0{,}5$. Es decir, cargado equitativamente hacia la topología de red y los perfiles de usuario.

Los siguientes gráficos de este apartado muestran la influencia con la nueva media móvil simple para cada uno de los senadores. Son cuatro gráficos por senador y cuatro más para mostrar influencias relativas. Los gráficos que se mostrarán son los del modelo con ponderación de 33 % para todas las dimensiones, además de los casos 1, 2 y 3 del listado anterior.

Las figuras 24, 25, 26 y 27 muestran los resultados de Guido Girardi. La primera figura muestra cómo se comporta el modelo con una ponderación de 33 % para todas las dimensiones. Aquí se puede ver que efectivamente la influencia de Girardi es mucho más estable y comparable con Klout. La influencia en este caso parte desde el 10 de mayo un poco más abajo que Klout y a medida que pasan los días, sube con una leve inclinación, sobrepasando a Klout desde el 5 de junio. Nótese que entre el 27 de mayo y 4 de junio ambas curvas prácticamente se superponen. La figura 25 muestra la influencia de Girardi exclusivamente en términos de actividad. Tiene una inclinación similar al caso anterior, sin embargo en todo momento su influencia está considerablemente más arriba que su Klout Score. En términos topológicos como se puede ver en la figura 26, la influencia de Girardi tiene un comportamiento similar con la diferencia de que en los primeros días permanece superpuesta con su Klout Score. Si se consideran sólo los datos de perfil, como es el caso de la figura 27, la influencia de Girardi cae notablemente y muy por debajo de su Klout Score. De esto se concluye que la influencia de Guido Girardi se explica por su alta actividad y su importancia en niveles topológicos, sin embargo, son sus características de perfil de usuario las que le juegan en contra y bajan su puntaje.

Las figuras 28, 29, 30 y 31 muestran los gráficos de Ena von Baer. En el caso de iguales ponderaciones para cada dimensión, Von Baer tiene un comportamiento similar a Girardi en los primeros días, con la diferencia que desde el 28 de mayo la influencia comienza a converger con su Klout Score y prácticamente ambas curvas se superponen hasta el último día. En relación a la actividad (figura 29) como en perfiles (figura 31), su influencia sigue una tendencia similar. Ambas curvas empiezan a ascender suavemente desde el 20 de mayo y ambas están por sobre la curva de Klout durante todo el período observado. Al igual que con Girardi, los datos de perfil bajan el puntaje promedio de influencia. Como se puede ver en la figura 31, este está muy por debajo de la curva de Klout de Von Baer.

En el caso de Jovino Novoa los resultados difieren bastante con el resto. En casi todos los casos, Novoa tiene una influencia muy baja y muy por debajo de Klout. Esto se ve en las figuras 32, 33 y 35. Es evidente que dada la baja actividad del senador, corresponda que bajo este criterio la influencia de este se encuentre por debajo de Klout. Sin embargo es importante destacar también que en términos de topología (figura 34) es donde Novoa logra subir su nivel de influencia, y es donde se diferencia bastante de Klout.

Soledad Alvear es otro caso particular. Con todas las dimensiones ponderadas de igual forma, la senadora tiene una influencia ligeramente por sobre la Klout e igual de pareja en términos de estabilidad (figura 28). Lo excepcional de Alvear es que ponderados al 100 % la actividad y los perfiles prácticamente llegan a los 100 puntos durante todo el período (Figuras 29 y 30). Nuevamente, como se puede ver en la figura 31 es el perfil el responsable de bajar la influencia imponente de la senadora.

Respecto a las influencias relativas, y tomando como punto de referencia la figura 19, las tendencias en cuanto a cuál de los cuatro senadores es más influyentes se mantiene. Alvear es la mas influyente según los criterios de las figuras 40, 41 y 42. Girardi y Von Baer son los que se pelean el segundo lugar, y Novoa siempre es el menos influyente de los cuatro. Sin embargo es importante destacar, que en términos de topología es donde Novoa se acerca más a los demás senadores. El caso excepcional es en la ponderación hacia los datos de perfil. Tal y como se puede ver en la figura 43, el panorama es totalmente distinto. En este caso Girardi es el más influyente, seguido por Von Baer que desciende abruptamente hacia el final del período observado. Alvear y Novoa son los que en este caso se disputan el tercer lugar.

En síntesis, se ha visto que en todos los casos los perfiles de los usuarios analizados han afectado sus niveles de influencia, lo que puede significar que hay personas con un nivel mucho mayor de amigos y seguidores en las redes en las que participan los cuatro políticos. En el caso de Alvear, su alto nivel de participación es lo que claramente le da supremacía por sobre el resto. Por otro lado el evidente desapego de Novoa por su cuenta de Twitter, al menos durante el período observado, es lo que significó una penalización grande en su puntaje. Aún así es curioso que con una actividad casi nula, el senador obtuvo un puntaje relativamente decente.

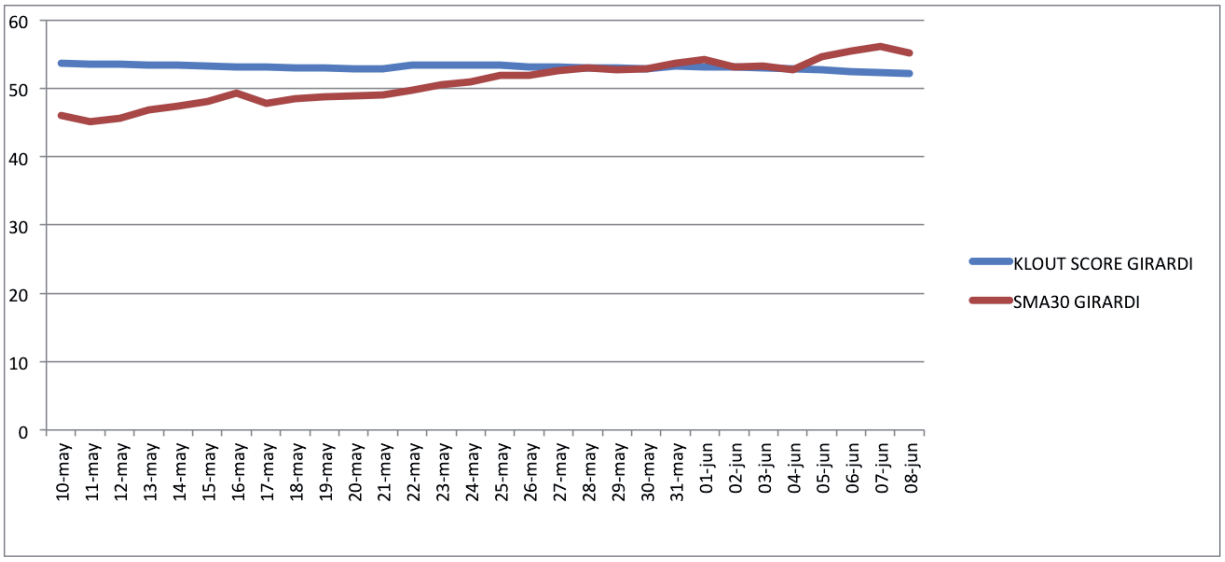

Figura 24: Influencia de Guido Girardi con media móvil simple de 30 períodos contrastada con Klout Score.

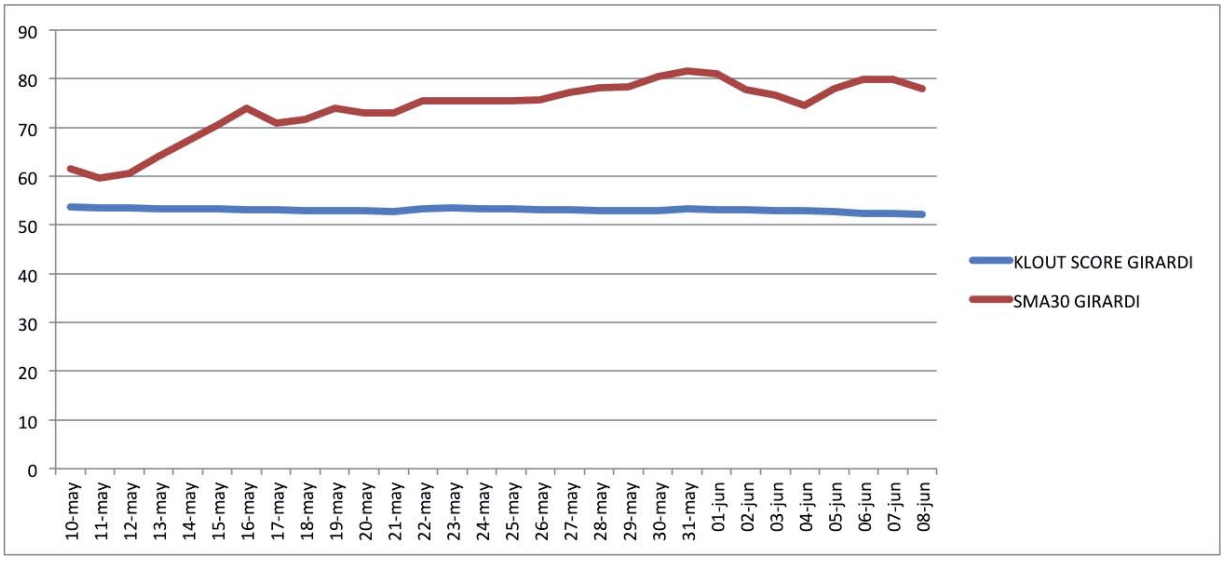

Figura 25: Influencia de Guido Girardi con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la actividad.

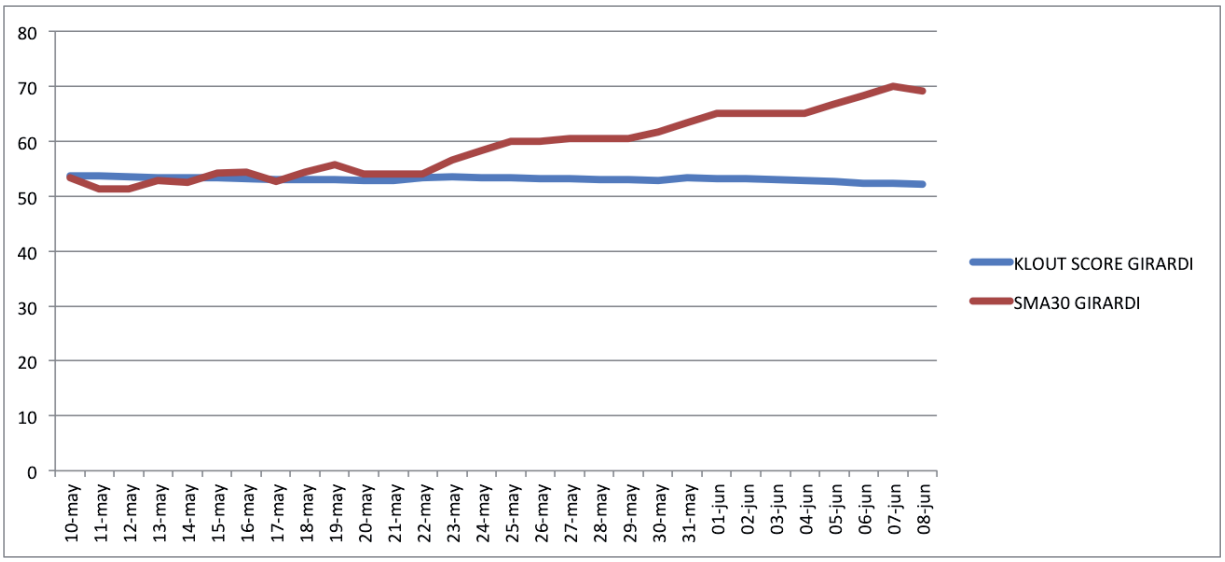

Figura 26: Influencia de Guido Girardi con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la topología de red.

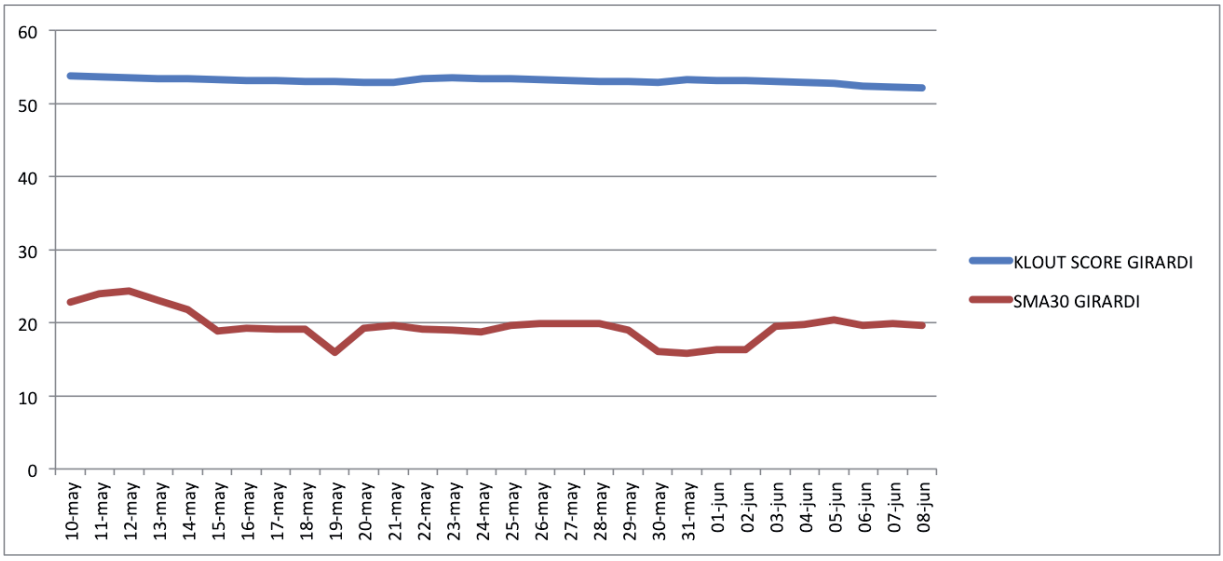

Figura 27: Influencia de Guido Girardi con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia los datos de perfiles.

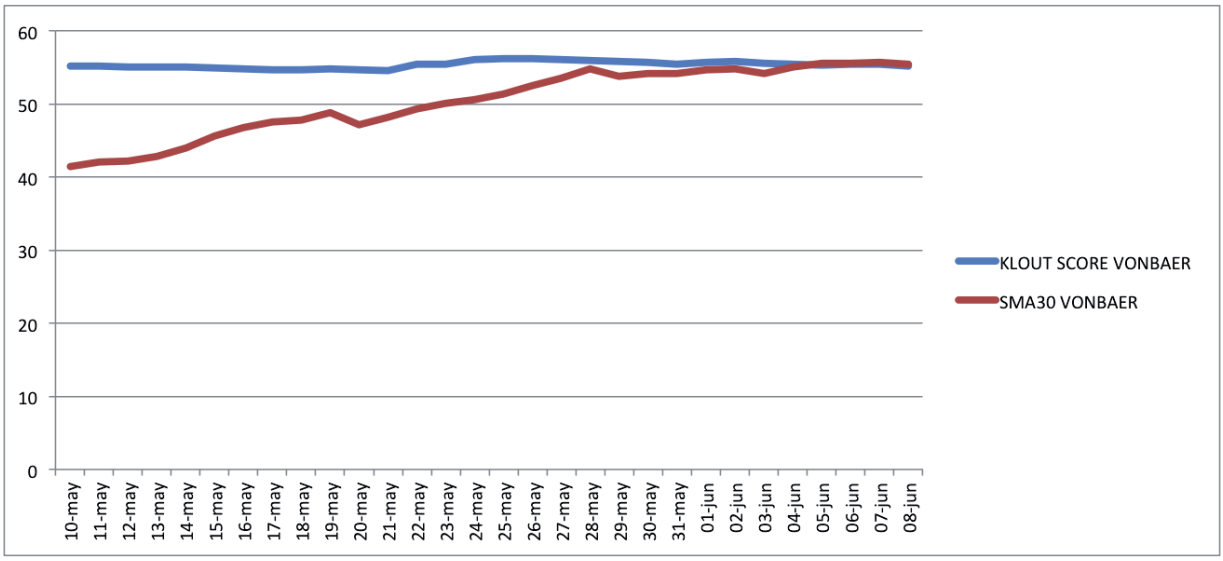

Figura 28: Influencia de Ena von Baer con media móvil simple de 30 períodos contrastada con Klout Score.

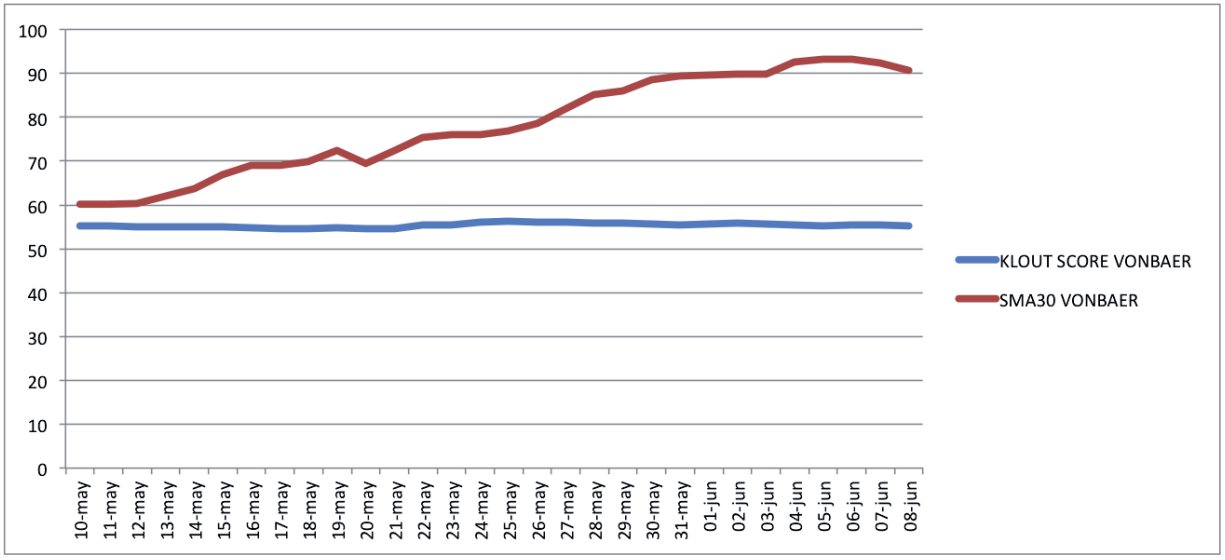

Figura 29: Influencia de Ena von Baer con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la actividad.

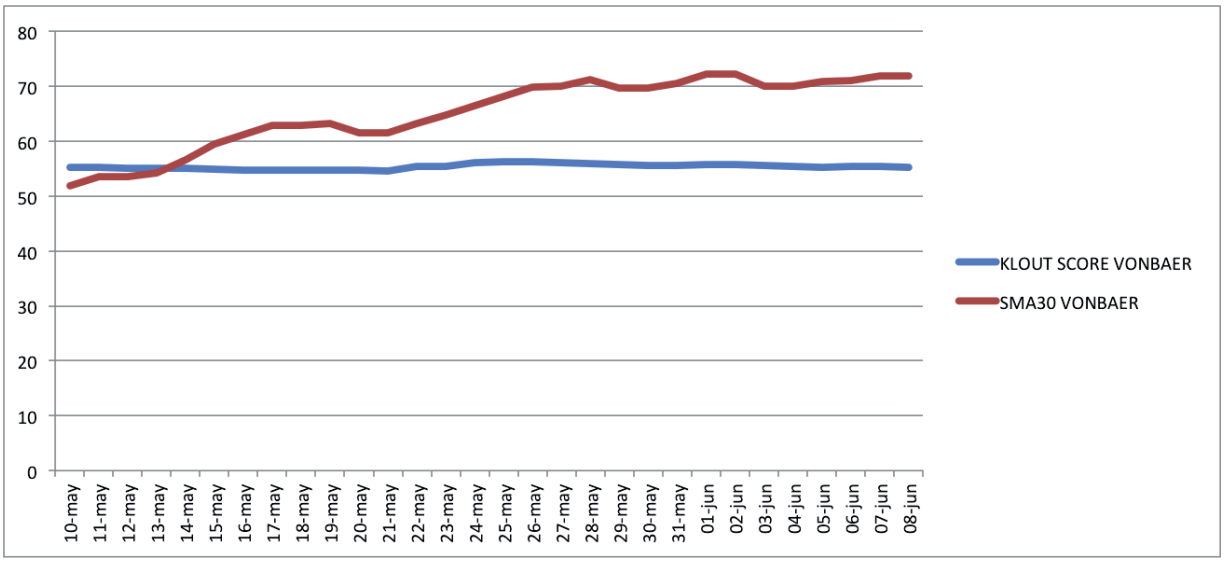

Figura 30: Influencia de Ena von Baer con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la topología de red.

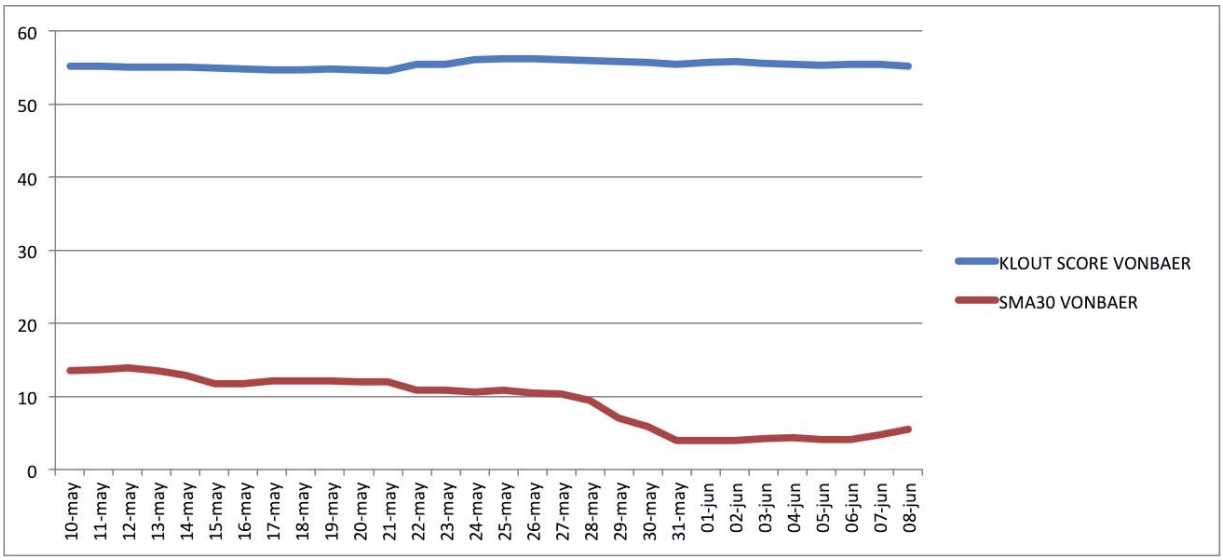

Figura 31: Influencia de Ena von Baer con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia los datos de perfiles.

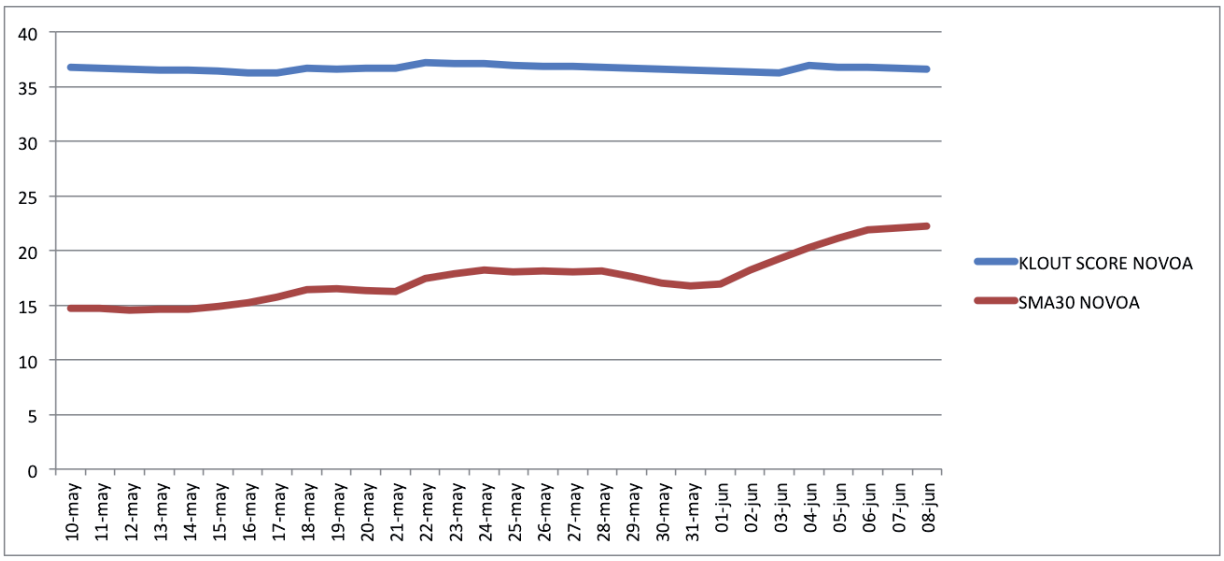

Figura 32: Influencia de Jovino Novoa con media móvil simple de 30 períodos contrastada con Klout Score.

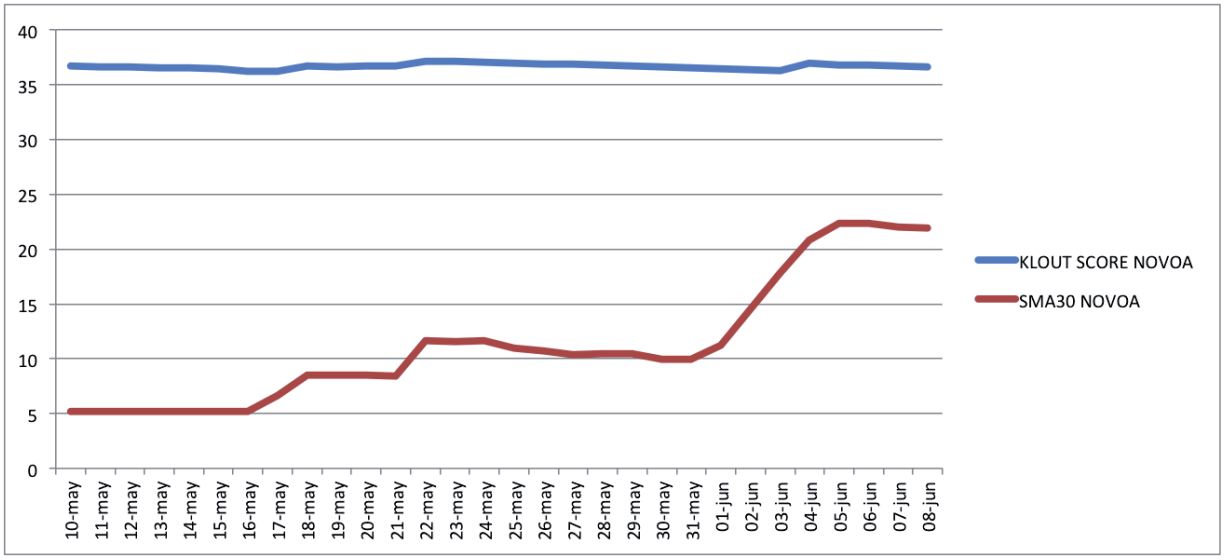

Figura 33: Influencia de Jovino Novoa con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la actividad.

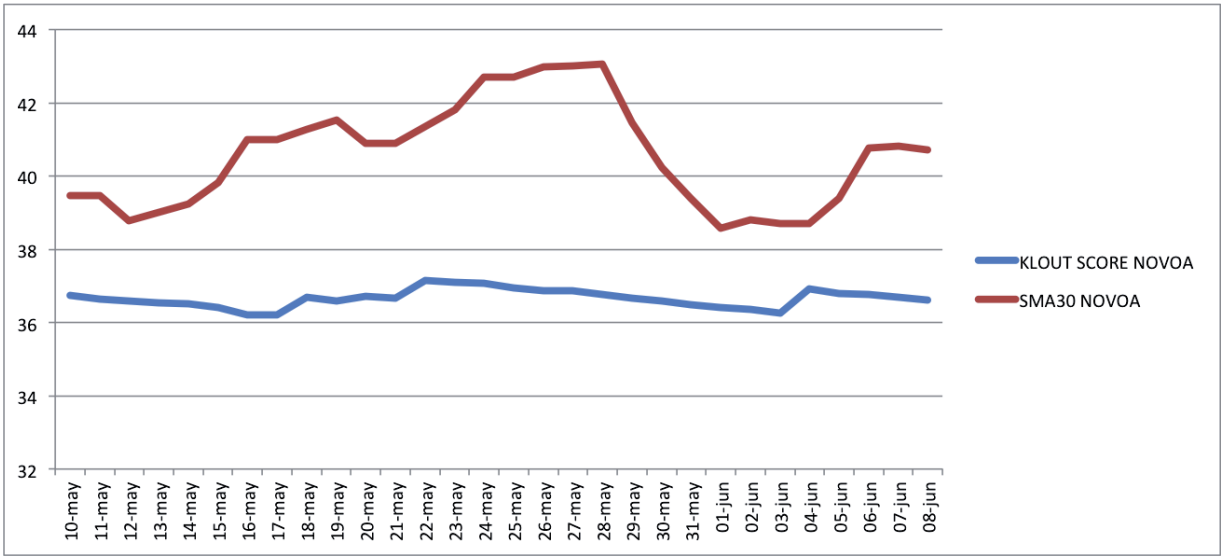

Figura 34: Influencia de Jovino Novoa con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la topología de red.

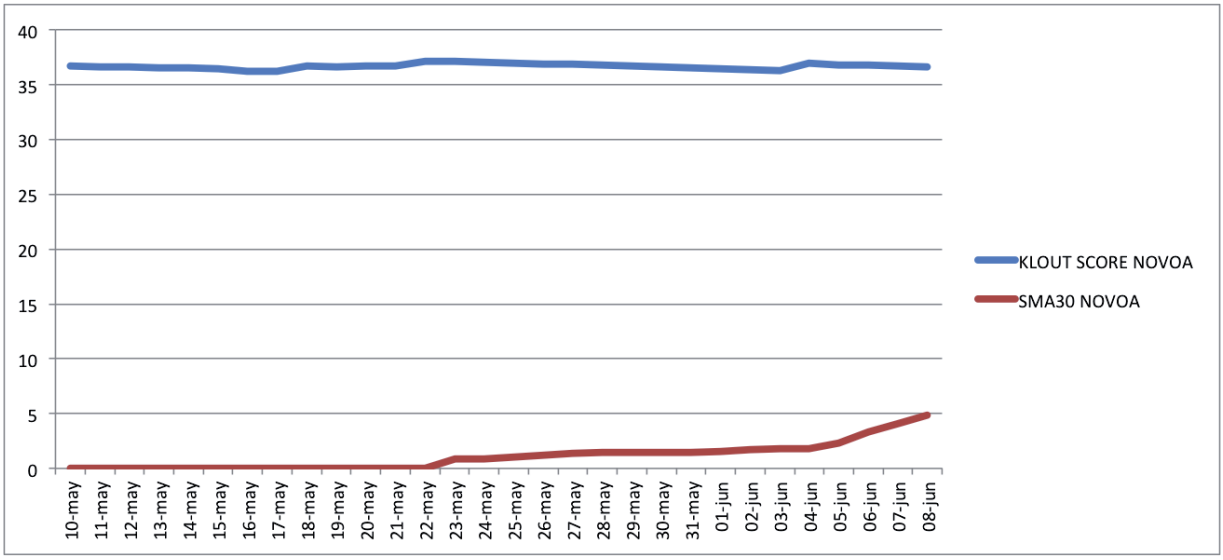

Figura 35: Influencia de Jovino Novoa con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia los datos de perfiles.

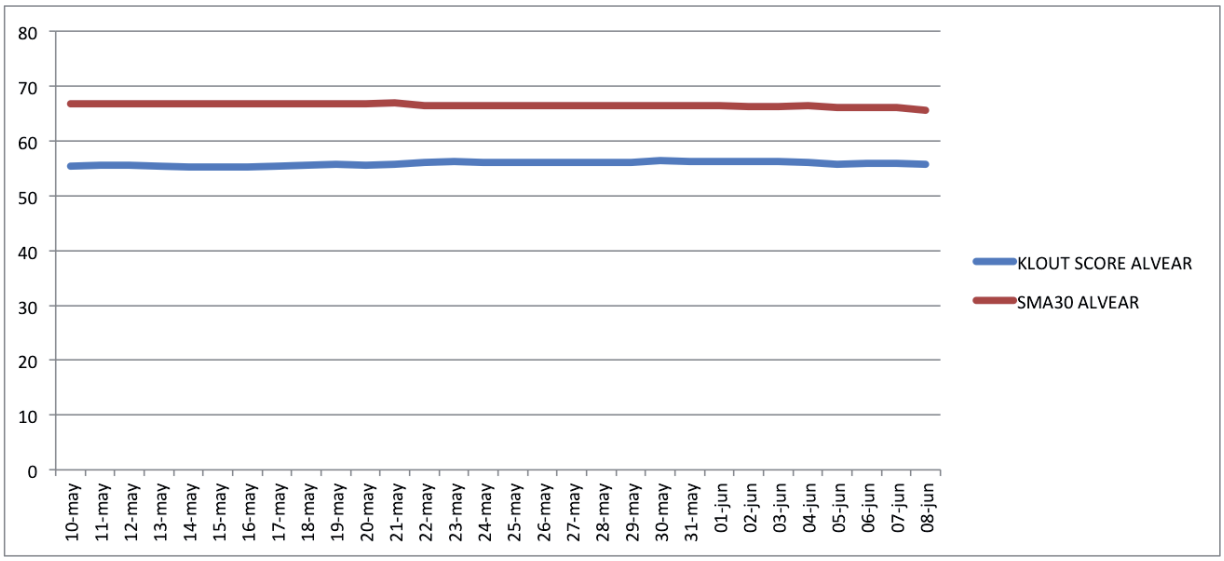

Figura 36: Influencia de Soledad Alvear con media móvil simple de 30 períodos contrastada con Klout Score.

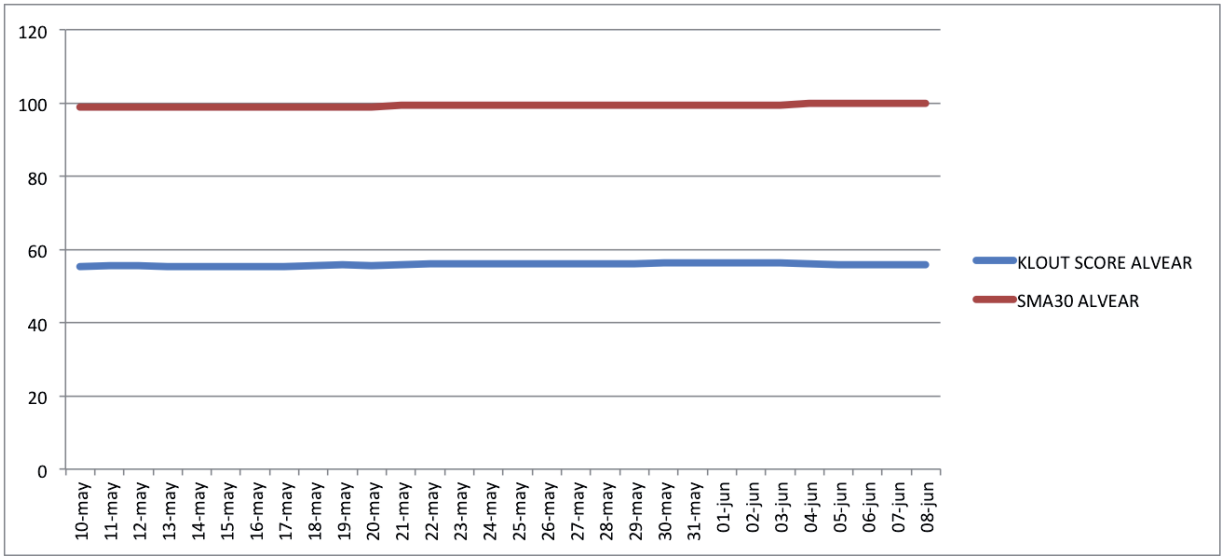

Figura 37: Influencia de Soledad Alvear con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la actividad.

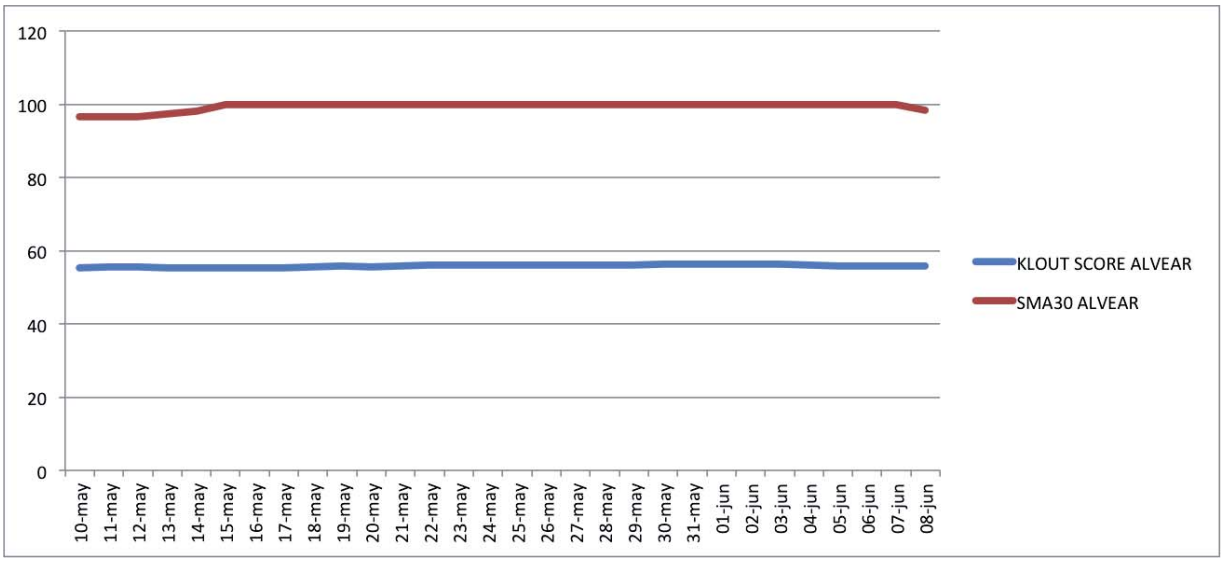

Figura 38: Influencia de Soledad Alvear con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la topología de red.

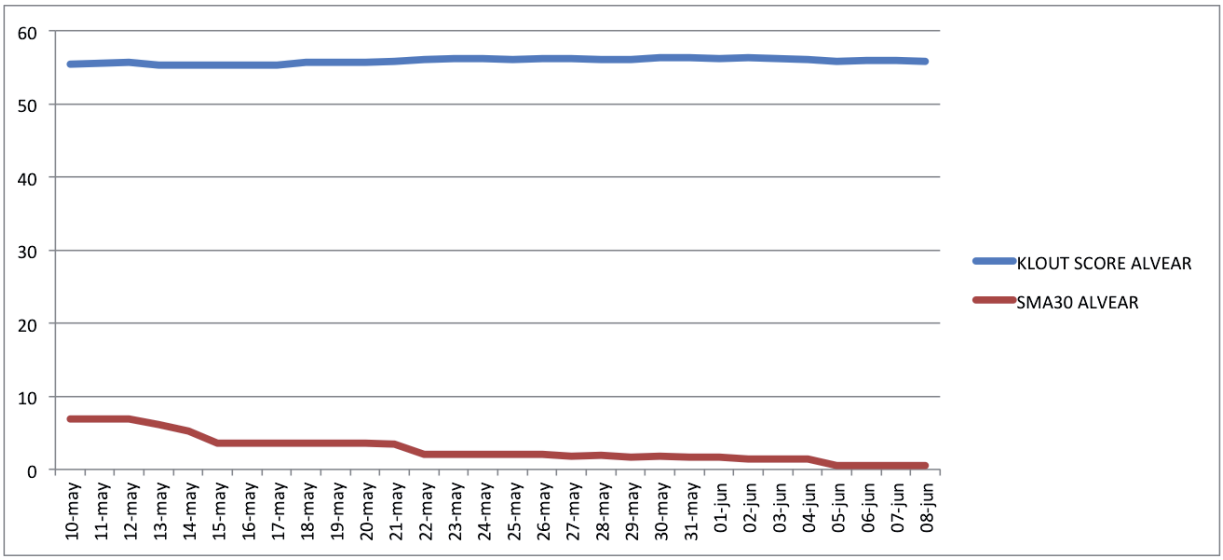

Figura 39: Influencia de Soledad Alvear con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia los datos de perfiles.

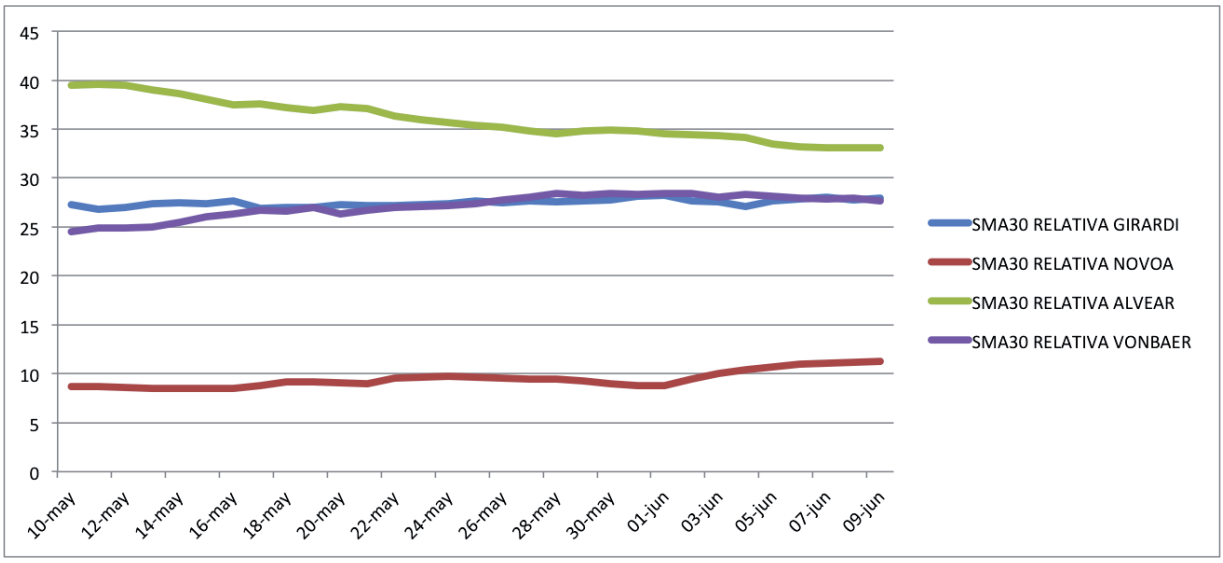

Figura 40: Influencia relativa con media móvil simple de 30 períodos contrastada con Klout Score.

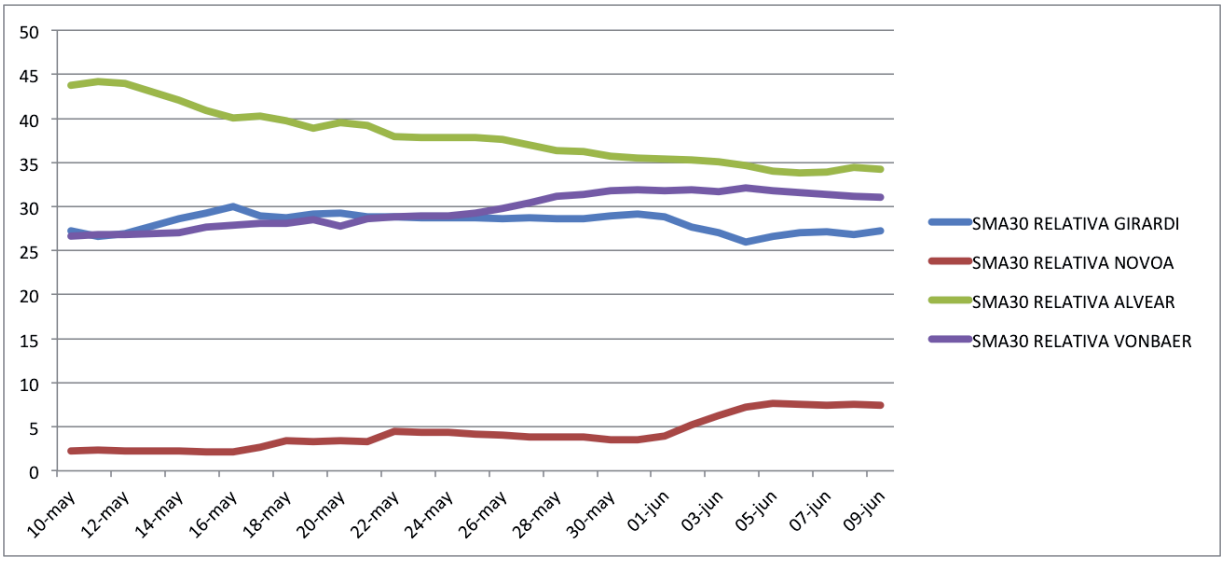

Figura 41: Influencia relativa con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la actividad.

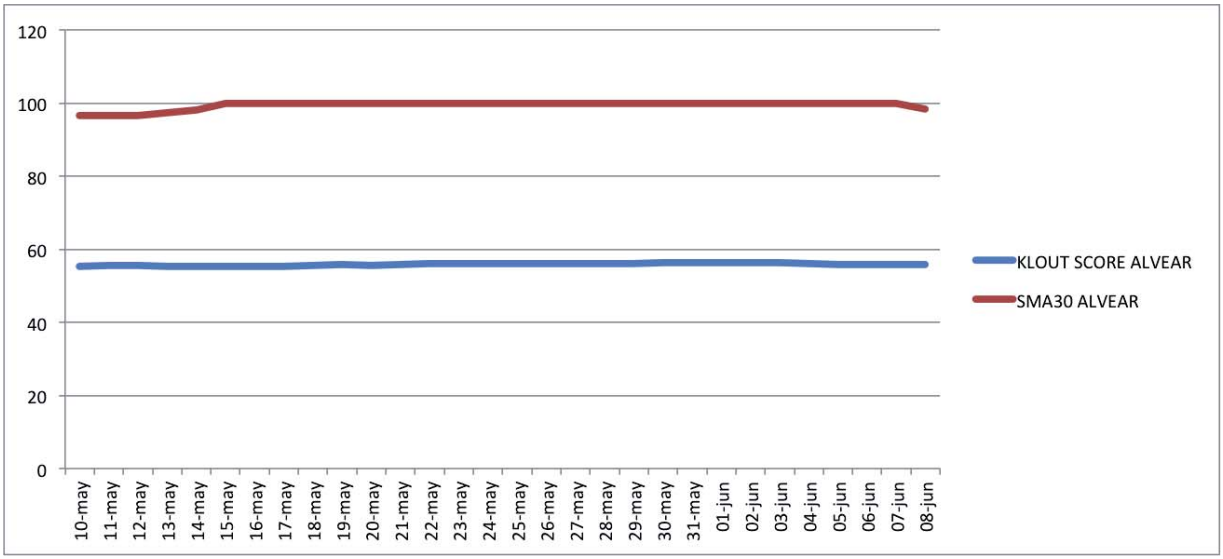

Figura 42: Influencia relativa con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia la topología de red.

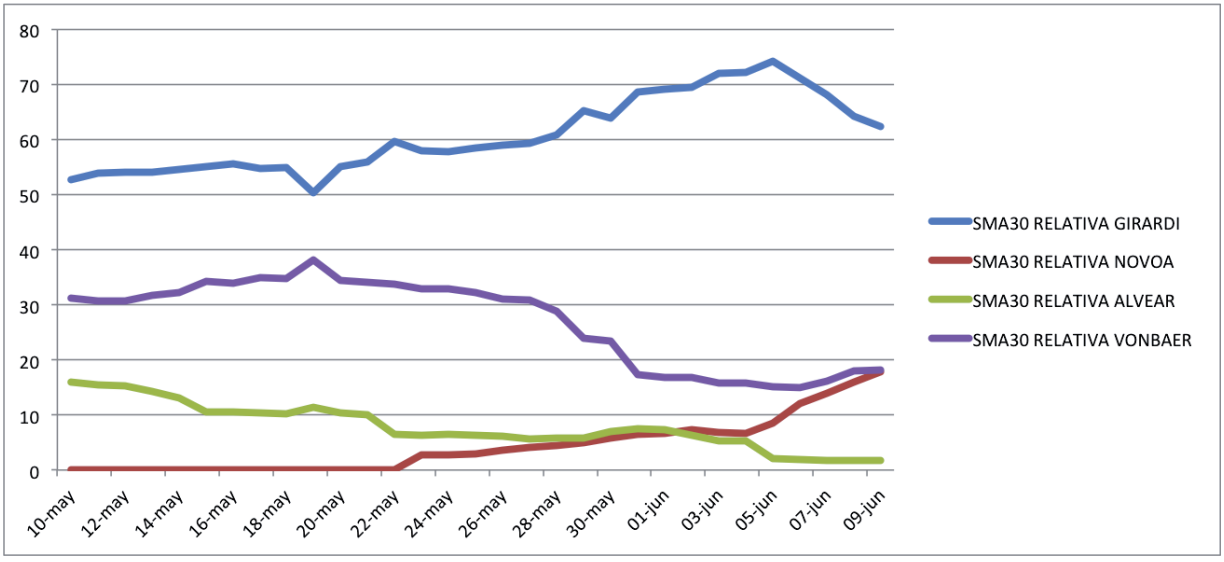

Figura 43: Influencia relativa con media móvil simple de 30 períodos contrastada con Klout Score, cargada 100 % hacia los datos de perfiles.

10.3.2. Modificación de parametros y correlacion con nuevas ponderaciones

Para las nuevas tablas de promedio y correlación se agregó $S30_1, S30_2, S30_3, ...S30_{30}$, variable la cual representa la influencia con media móvil simple de 30 períodos. Donde $\overline{s30}$ es la media de esta variable y la correlación de Pearson entre K y $S30$ se encuentra determinada por r_{KS30}.

A las ponderaciones de las tablas anteriores se le suman las 6 nuevas ponderaciones descritas anteriormente en este capítulo. Act. 1, Top. 1 y Per. 1 representan la influencia cargada 100 % hacía actividad, topología y perfiles respectivamente. Las últimas tres ponderaciones corresponden a las que consideran sólo dos dimensiones a la vez. Por ejemplo Act. - Top. es el caso en que $\alpha = 0{,}5$, $\beta = 0{,}5$ y $\gamma = 0$.

En la tabla 11 se puede ver el rendimiento del modelo para el senador Girardi. En general las correlaciones entre el modelo propuesto con Klout no son buenas (cercanas a cero). Especialmente con la media móvil simple de 30 períodos donde las correlaciones observadas son sorprendentemente negativas (aproximadamente $-0{,}7$ en la mayoría de los casos). Sin embargo, en varios casos los promedios se acercan bastante al promedio de Klout. En el caso de Top. 0.50 por ejemplo, el promedio de $55{,}3$ es muy cercano a los $53{,}1$ de Klout. Independientemente de que la correlación sea muy negativa, hay que revisar los gráficos y hacer un análisis de desviación estándar para ver que tanto difieren los puntajes observados de los de Klout. Un caso excepcional es en Per. 1 con una media móvil simple de 30 períodos donde se llegó a la correlación más alta observada ($0{,}3$) para Girardi.

Los resultados de Ena Von Baer se pueden ver en la tabla 12. La particularidad en este caso es que salvo una excepción las correlaciones con media móvil simple de 30 períodos son claramente positivas (mayores o iguales a $0{,}5$). Esto no se da en el resto de los casos. Los promedios más cercanos a Klout se encuentran concentrados en las tres primeras ponderaciones (0.33, Act 0.50 y Top. 0.50). También se puede notar que claramente el perfil es el que baja notablemente el puntaje de influencia de Von Baer y que basándose sólo en actividad se logran puntales altísimos.

Jovino Novoa tiene en general promedios bajos. Tal y como se puede apreciar en la tabla 13, se aleja bastante de los $36{,}66$ puntos promedio de Klout. Sin embargo en Act. -Top. y en Top. 1 hay una gran cercanía. Como se vio anteriormente en este capítulo, la topología de red era la dimensión que más puntaje le daba a Novoa, lo que explica este fenómeno. En cuanto a correlaciones, hay mucha variación tanto por el lado positivo como negativo. La correlación más alta alcanzada es $0{,}52$ en Top. 1 con la media móvil simple de 30 períodos.

Por último, en el caso de Soledad Alvear (tabla 14) las correlaciones son en general bajas, tendiendo en la mayoría de los casos a cero. Sin embargo hay dos casos en que las correlaciones son claramente positivas, en Act. 1 y en Act. - Per., pero esto ocurre con una media móvil simple de 3 períodos. Como se vio en los gráficos en la sección anterior de este capítulo, los puntales de Alvear con las ponderaciones cargadas totalmente hacia las dimensiones de actividad o topología llegan prácticamente a 100 puntos a lo largo de todo el período. Por lo tanto no es de extrañar que los promedios que involucran a dichas dimensiones son altísimos. La dimensión de perfil de usuario es la que baja abruptamente el promedio de Alvear.

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	$\overline{s30}$	r_{KI}	r_{KS3}	r_{KS30}
0.33	53.1	55.3	55.61	50.88	-0.09	-0.11	-0.69
Act. 0.50	53.1	61.76	62.04	57.45	-0.11	-0.02	-0.65
Top. 0.50	53.1	59.12	59.42	55.3	-0.06	-0.18	-0.7
Per. 0.50	53.1	46.68	47.05	43.43	-0.09	-0.14	-0.69
Act. 1	53.1	79.48	79.26	74	-0.12	0.06	-0.59
Top. 1	53.1	68.93	69.16	59.01	0.02	-0.27	-0.71
Per. 1	53.1	19.16	19.71	19.53	-0.04	-0.06	0.3
Act. - Top.	53.1	74.21	74.4	67.32	-0.07	-0.07	-0.67
Act. - Per.	53.1	49.32	49.67	47.58	-0.14	0.06	-0.58
Top. - Per.	53.1	44.04	44.43	39.27	-0.01	-0.32	-0.66

Tabla 11: Modificación de parámetros para Guido Girardi con SMA30 y nuevas ponderaciones

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	$\overline{s30}$	r_{KI}	r_{KS3}	r_{KS30}
0.33	55.35	54.87	55.57	50.3	-0.11	0.11	0.56
Act. 0.50	55.35	64.22	65.12	57.61	-0.09	0.1	0.54
Top. 0.50	55.35	59.11	59.98	54.41	-0.13	0.12	0.57
Per. 0.50	55.35	42.96	43.29	40.42	-0.08	0.09	0.59
Act. 1	55.35	90.59	92.08	78.02	-0.05	0.07	0.5
Top. 1	55.35	70.14	71.53	65.19	-0.16	0.12	0.58
Per. 1	55.35	5.56	4.78	9.22	0.15	-0.1	-0.4
Act. - Top.	55.35	80.36	81.8	71.61	-0.13	0.12	0.54
Act. - Per.	55.35	48.07	48.43	43.62	0.01	0.04	0.54
Top. - Per.	55.35	37.85	38.15	37.21	-0.12	0.11	0.59

Tabla 12: Modificación de parámetros para Ena von Baer con SMA30 y nuevas ponderaciones

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	$\overline{s30}$	r_{KI}	r_{KS3}	r_{KS30}
0.33	36.66	22.28	22.09	17.47	-0.39	0.31	0.32
Act. 0.50	36.66	22.32	22.26	16.01	-0.48	0.24	0.28
Top. 0.50	36.66	26.96	26.93	23.38	-0.31	0.37	0.4
Per. 0.50	36.66	18.24	17.76	13.51	-0.29	0.33	0.28
Act. 1	36.66	21.76	22.1	11.25	-0.55	0.15	0.24
Top. 1	36.66	40.31	40.76	40.59	0.11	0.4	0.52
Per. 1	36.66	5.46	4.09	1.09	0.39	0.26	0.08
Act. - Top.	36.66	31.03	31.43	25.92	-0.48	0.27	0.36
Act. - Per.	36.66	13.61	13.09	6.17	-0.46	0.2	0.21
Top. - Per.	36.66	22.88	22.43	20.84	0.29	0.42	0.44

Tabla 13: Modificación de parámetros para Jovino Novoa con SMA30 y nuevas ponderaciones

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	$\overline{s30}$	r_{KI}	r_{KS3}	r_{KS30}
0.33	55.86	65.59	65.95	66.49	0.01	0.14	0.22
Act. 0.50	55.86	74.66	74.94	75.19	0.02	0.02	0.13
Top. 0.50	55.86	74.27	74.82	75.21	-0.002	0.1	0.08
Per. 0.50	55.86	49.82	50.09	51.08	0.02	0.17	0.2
Act. 1	55.86	99.89	99.89	99.29	0.13	0.54	0.18
Top. 1	55.86	98.33	99.44	99.36	-0.01	0.05	-0.1
Per. 1	55.86	0.52	0.52	2.83	0.23	0.41	-0.16
Act. - Top.	55.86	99.11	99.67	99.33	0.004	0.11	-0.14
Act. - Per.	55.86	50.21	50.2	51.06	0.25	0.61	-0.45
Top. - Per.	55.86	59.52	59.96	60.74	0.005	0.11	0.25

Tabla 14: Modificación de parámetros para Soledad Alvear con SMA30 y nuevas ponderaciones

10.3.3. Ponderaciones aproximadas a Klout

Después de haber probado con distintas ponderaciones, se decidió buscar valores que se acercaran aún más a Klout para cada político. Si bien el objetivo principal de esta investigación no es crear un modelo similar a Klout, este apartado nace de la mera curiosidad de ver si se puede emular el Klout Score con el modelo propuesto.

En el caso de Guido Girardi, se probaron distintas combinaciones de α, β y γ. Las correlaciones nunca fueron claramente positivas así que se opto por buscar el mejor promedio y desviación estándar. La ponderación de 33 % para cada dimensión con una media móvil simple de 30 períodos se consideró satisfactoria en este caso ya que presentó algunos de los promedios mas cercanos a Klout. La desviación estándar del Klout Score de Girardi es de $0{,}37$ y la del modelo en este caso es de $3,13$ puntos, lo que se podría considerar relativamente cercano.

Para Ena von Baer se propone la combinación que pondera 50 % a la topología y el resto equitativamente para actividad y perfiles y con una media móvil simple de 30 períodos. Está ponderación, que ya se analizó en el apartado anterior, tiene $54{,}41$ puntos de promedio, uno de los más cercanos al Klout Score de Von Baer. La correlación también es una de las más altas observadas ($0{,}57$). Y finalmente la desviación estándar es aproximadamente $5{,}16$ mientras que la de su Klout Score es de $0{,}48$. Se considero que esta configuración tenía un buen equilibrio entre media, correlación y desviación estándar.

Para Jovino Novoa se consideró que había que darle mayor importancia a la topología que a las demás dimensiones. Luego de varias pruebas, la configuración más apropiada encontrada fue $\alpha = 0{,}1$, $\beta = 0{,}8$ y $\gamma = 0{,}1$ con una media móvil simple de 30 períodos. Con esta configuración se encontró la correlación más alta para Jovino Novoa ($0{,}56$) y un promedio muy cercano a su Klout Score promedio. Estos datos aparecen en la tabla 15. La figura 44 muestra evidencia de lo buena que es esta aproximación. Esta es una de las mejores aproximaciones a Klout que se han alcanzado en toda esta investigación.

Finalmente para Soledad Alvear se proponen dos alternativas. Se probó con distintas ponderaciones y dos parecieron satisfactorias. La primera alternativa se acerca más en términos de promedio y desviación estándar. Esta aproximación tiene los valores $\alpha = 0{,}35$, $\beta = 0{,}2$ y $\gamma = 0{,}45$ con una media móvil de 30 períodos. Como se puede ver en la tabla 16, el promedio de $55{,}9$ que da está configuración es tremendamente similar a los $55{,}86$ puntos promedio de Klout. La desviación estándar de esta configuración es de $0{,}57$ puntos la cual también es muy parecida a la de Klout ($0{,}35$ puntos). En la figura 45 se puede la cercanía que se logra con esta configuración.

Sin embargo la correlación en este caso no es tan buena, tal y como se puede ver en la tabla anterior. Por lo tanto, se pensó en proponer una segunda alternativa la cual difiere en promedio pero tiene la correlación mas alta encontrada para Soledad Alvear. De hecho es la correlación positiva más alta encontrada en esta investigación y es de $0{,}61$ puntos. En la tabla del apartado anterior (tabla 14) se puede ver este caso, donde la ponderación es $\alpha = 0{,}5$, $\beta = 0$ y $\gamma = 0{,}5$ (Act. - Per. en dicha tabla) y la media móvil simple es de 3 períodos. En este caso no se considera para nada la topología de red de Alvear. El gráfico para esta configuración se puede ver en la figura 46.

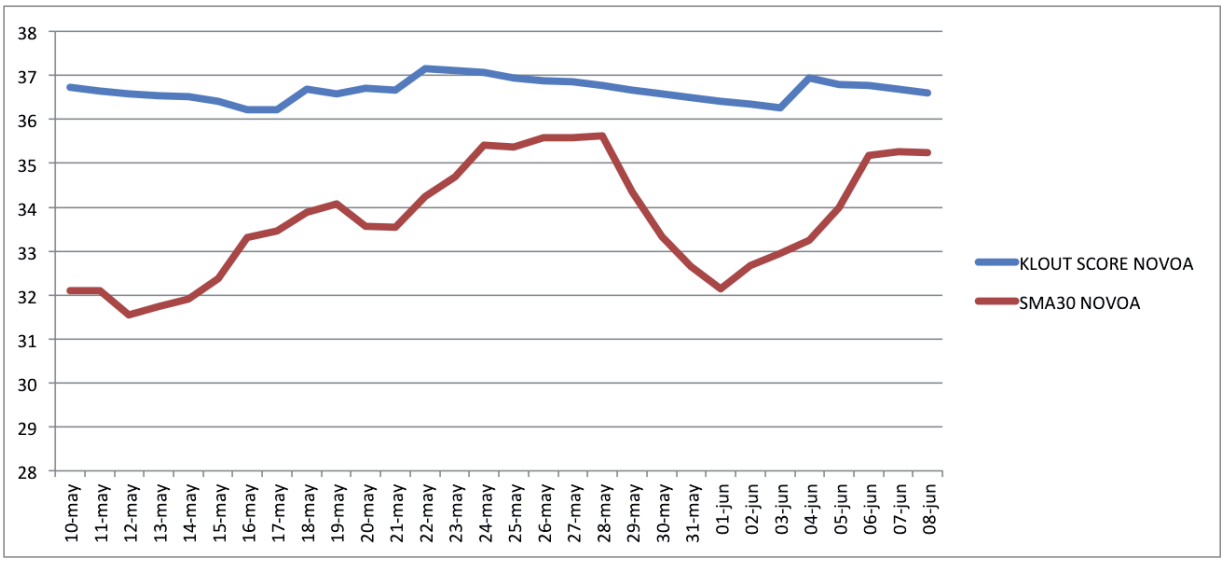

Figura 44: Influencia de Jovino Novoa con media móvil simple de 30 períodos aproximada a su Klout Score

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	$\overline{s30}$	r_{KI}	r_{KS3}	r_{KS30}
Aprox.	36.66	34.97	35.23	33.7	-0.06	0.43	0.56

Tabla 15: Aproximación a Klout Score de Jovino Novoa

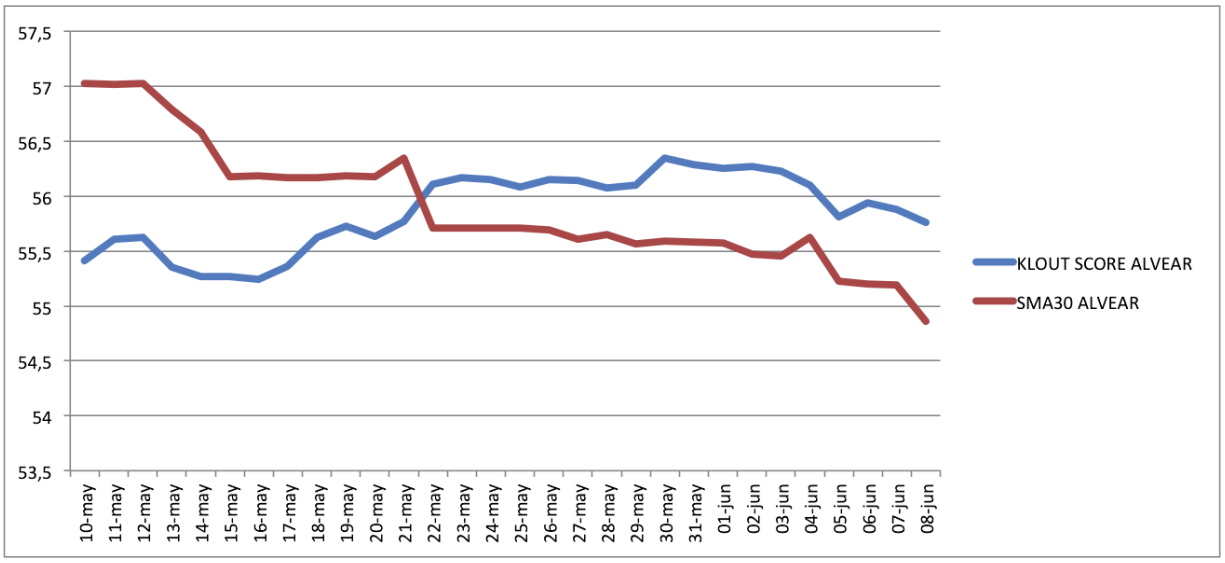

Figura 45: Influencia de Soledad Alvear con media móvil simple de 30 períodos aproximada a su Klout Score

Ponderación	$\overline{k}$	$\overline{i}$	$\overline{s3}$	$\overline{s30}$	r_{KI}	r_{KS3}	r_{KS30}
Aprox.	55.86	54.86	55.08	55.9	0.03	0.22	0.17

Tabla 16: Aproximación a Klout Score de Soledad Alvear para la primera alternativa propuesta

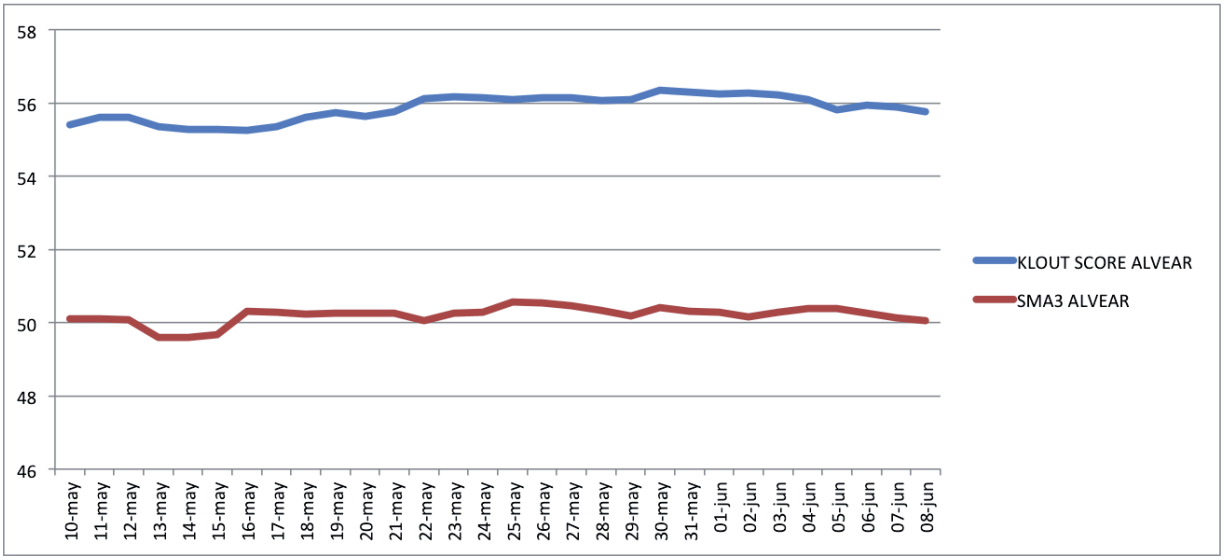

Figura 46: Influencia de Soledad Alvear con media móvil simple de 3 períodos aproximada a su Klout Score

10.4. Análisis de sentido e intención

Para clasificar los *tweets* por sentido y por intención, se utilizó la librería NLTK (*Natural Language Toolkit*) de Python la cual cuenta entre otras cosas, con un diccionario de *stopwords* en español y una función para procesar los documentos en un *bag of words*. Además, NLTK tiene varios clasificadores. Por razones de rendimiento se prefirió clasificar con Naive Bayes. Se decidió dejar a Novoa afuera de la clasificación dada su baja actividad.

Para el caso de la clasificación de intención los *tweets* del *training set* y los del *test set* fueron pre procesados según la especificación de [16].

10.4.1. Tendencias

Las tablas 17 y 18 muestran las tendencias que tienen los *tweets* de los tres políticos.

En términos de sentido, Girardi tweetea claramente contenido de carácter neutro. De hecho, viendo el *dataset*, Girardi *tweetea* muchos links acerca de entradas de sus blogs y no suele opinar mucho. Ena Von Baer tiene una mayor tendencia hacia los *tweets* positivos y un poco más bajo en neutros, y Soledad Alvear tiene una tendencia un poco mayor hacia los *tweets* positivos. Ninguno de los tres políticos se destaca por *tweetear* negativamente.

En el caso de la intención Girardi es quién tweetea más noticias, Alvear está más dedicada a opinar y a compartir con sus seguidores, y Ena Von Baer es un poco más dedicada a comentar su vida personal.

Porcentaje (%)	Positivos	Negativos	Neutros
Girardi	15.625	3.125	81.25
Von Baer	57.62	11.86	30.5
Alvear	60.96	17.99	21.03

Tabla 17: Porcentaje sentido *tweets* de Guido Girardi, Ena Von Baer y Soledad Alvear

Porcentaje (%)	NO	OP	SO	DI
Girardi	42.18	39.06	7.81	10.93
Von Baer	23.72	10.16	28.81	37.28
Alvear	9.125	35.23	43.72	11.91

Tabla 18: Porcentaje intención *tweets* de Guido Girardi, Ena Von Baer y Soledad Alvear

10.4.2. Resultados de evaluación del clasificador

En términos de *F-measure* el clasificador tiene resultados dispares, tanto entre los tres senadores, como entre las clases. Es particularmente pobre en la clasificación de sentido para Girardi y para Von Baer. En el caso de Alvear el rendimiento es levemente mejor, pero solo en el caso de los tweets positivos. En el caso de la taxonomía el clasificador anduvo particularmente mejor para clasificar tweets sociales y de noticias pero anduvo bastante mal en el caso de las

opiniones y levemente mejor en diario. Esto puede deberse a que tal vez la clasificación manual no estuvo tan bien o a que simplemente el clasificador no sirvió para este tipo de problemas de clasificación.

	Positivos	Negativos	Neutros
Accuracy	0.84	0.96	0.81
Precision π	0	0	1
Recall ρ	N/D	N/D	0.81
F-measure	N/D	N/D	0.89

Tabla 19: Evaluación sentido *tweets* de Guido Girardi

	Positivos	Negativos	Neutros
Accuracy	0.35	0.86	0.25
Precision π	0.125	0	0.66
Recall ρ	0.44	N/D	0.19
F-measure	0.19	N/D	0.26

Tabla 20: Evaluación sentido *tweets* de Ena Von Baer

	Positivos	Negativos	Neutros
Accuracy	0.67	0.80	0.56
Precision π	0.55	0.09	0.78
Recall ρ	0.85	0.36	0.29
F-measure	0.67	0.14	0.43

Tabla 21: Evaluación sentido *tweets* de Soledad Alvear

	SO	OP	DI	NO
Accuracy	0.81	0.56	0.15	0.46
Precision π	0.6	0	0.42	0.59
Recall ρ	0.23	0	0.05	0.51
F-measure	0.33	N/D	0.1	0.55

Tabla 22: Evaluación intención *tweets* de Guido Girardi

	SO	OP	DI	NO
Accuracy	0.83	0.89	0.52	0.26
Precision π	0.82	0.16	0.68	0.21
Recall ρ	0.66	0.5	0.41	0.65
F-measure	0.73	0.25	0.51	0.1

Tabla 23: Evaluación intención *tweets* de Ena Von Baer

	SO	OP	DI	NO
Accuracy	0.64	0.63	0.15	0.09
Precision π	0.96	0.02	0.52	0.08
Recall ρ	0.55	0.34	0.073	0.008
F-measure	0.70	0.05	0.12	0.01

Tabla 24: Evaluación intención *tweets* de Soledad Alvear

11. Conclusión

En este informe se estableció lo que se entiende como influencia en Twitter y su importancia para en el desarrollo de campañas virales en internet. Se revisaron las publicaciones más relevantes en cuanto al estudio de este concepto en Twitter y el marco teórico detrás de estos estudios. Algunos estudios se basan generalmente en el área de estadística y otros incluyen la teoría de grafos para modelar las relaciones entre los usuarios de redes sociales en general. Dos de las publicaciones revisadas se preocupan de buscar métricas para el cálculo de la influencia y luego compararlas para ver cuales son las más confiables.

Se definieron problemas que no se habían abordado en estudios anteriores, los cuales no consideraban el comportamiento y heterogeneidad de los usuarios. Dicho comportamiento puede afectar importantemente a los cálculos de influencia ya que no todos los usuarios se comportan de la misma forma, por lo tanto, no son igualmente influyentes. Se tomó en cuenta también la calidad de los enlaces entre los usuarios. Adicionalmente, se tomó en cuenta el contenido de los *tweets* en términos de sentido e intención.

Se adaptó el modeló probabilístico basado en redes sociales en general, a las particularidades de Twitter. Se probó la totalidad del modelo (actividad, topología, perfiles) y se analizó el contenido de los *tweets* (por sentido e intención) en un caso particular. Se pudieron resolver muchas de las limitantes impuestas para extraer grandes cantidades de datos. Tras aplicar el modelo se obtuvieron resultados satisfactorios, ya que coincidieron con muchas de las expectativas que se tenían al comparar los resultados con Klout. Se sacaron conclusiones muy relevantes respecto de la importancia que tiene la actividad, la topología y los perfiles. En todos los casos analizados, la dimensión de perfil fue la que más afectó al puntaje de los usuarios analizados, lo que quiere decir que en todos los casos hubo usuarios mucho más populares que los senadores participando activamente en las redes de usuario. También se descubrió que a pesar de los resultados extremos que presenta el modelo probabilístico, el puntaje promedio hace que se acerque bastante al Klout Score promedio en algunas instancias. El caso más curioso es el de Jovino Novoa, quien con sólo un *tweet* tiene un Klout relativamente alto y un puntaje un poco más bajo con el modelo propuesto. Esto indica que probablemente hay factores externos que no pueden ser representados por este modelo. La baja actividad de Novoa es indudablemente castigada, reduciendo importantemente su puntaje en el modelo, mientras que Soledad Alvear es "premiada"por su alto nivel de interacción.

Respecto al análisis de sentido y de intención, los resultados fueron medianamente satisfactorios. La clasificación manual de *tweets* resultó ser compleja ya que en algunos casos es difícil categorizar y desambiguar un *tweet*.

Si bien el objetivo de la investigación no era emular a Klout, se intentó buscar ponderaciones en las dimensiones que lograran imitar el comportamiento del Klout Score. Es evidente sin embargo, que no hay un ajuste único para todos los usuarios y que de usuario en usuario los ajustes pueden variar enormemente. Ya que el modelo propuesto es contextual, en lugar de medir la influencia en términos globales, los ajustes pueden variar también de tema en tema. Lo importante del modelo propuesto es que es altamente configurable y que la persona que lo implemente puede estimar qué dimensiones son las que hay que tomar más en cuenta dependiendo del usuario, tema y fecha en que se aplique.

11.1. Trabajo futuro

Hay muchas mejoras que se le pueden hacer al modelo y muchos casos de prueba que se pueden aplicar para ver como se comporta bajo determinadas condiciones. Dentro del modelo mismo sería interesante probar con otras métricas para la dimensión de perfil. Por ejemplo, un ratio de actividad diaria promedio considerando los *tweets* totales del usuario en una ventana de tiempo, o incluso desde la fecha de creación de su perfil (Estos datos pueden obtenerse con facilidad desde la API de Twitter). En cuanto a casos de prueba sería interesante ver como varía la capacidad de influencia y la topología de las redes de usuarios en otros temas. En este trabajo se vio la política, pero podría probarse con otros temas muy diferentes. Todo depende de las palabras claves y las fechas que se elijan para armar la red de usuarios.

Computacionalmente hablando, el modelo propuesto puede ser mejorado. La dimensión de actividad en particular requiere mucho tiempo de procesamiento. Por lo tanto cualquier mejora al rendimiento del modelo completo no deja de ser relevante.

En cuanto a la clasificación automática de textos para sentido e intención, se podría probar con nuevos clasificadores y definir pautas para desambiguar más objetivamente al momento de entrenar los documentos.

Referencias

[1] M. Shiels, "Twitter co-founder jack dorsey rejoins company," March 2011 (Revisada en Noviembre 23, 2011). BBC News, http://www.bbc.co.uk/news/business-12889048.

[2] S/A, "Your world, more connected," August 2011 (Revisada en Noviembre 23, 2011). Twitter Blog, http://blog.twitter.com/2011/08/your-world-more-connected.html.

[3] G. U. Yule and M. G. Kendall, *Introducción a la estadistica matemática*. Spain: Aguilar S. A., 4th ed., 1964.

[4] S. Ye and S. F. Wu, "Measuring message propagation and social influence on twitter.com," in *Proceedings of the Second international conference on Social informatics*, SocInfo'10, (Berlin, Heidelberg), pp. 216–231, Springer-Verlag, 2010.

[5] R. F. Mihalcea and D. R. Radev, *Graph-based Natural Language Processing and Information Retrieval*. New York, NY, USA: Cambridge University Press, 1st ed., 2011.

[6] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[7] R. A. E., *Diccionario de la lengua española. Serie Asociación de Academias de la Lengua Española.* Madrid, España: Espasa Calpe, 22 ed., 2001.

[8] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, (New York, NY, USA), pp. 241–250, ACM, 2010.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, (Washington DC, USA), May 2010.

[10] I. Anger and C. Kittl, "Measuring influence on twitter," in *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '11, (New York, NY, USA), pp. 31:1–31:4, ACM, 2011.

[11] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Proceedings of the 20th international conference companion on World wide web*, WWW '11, (New York, NY, USA), pp. 113–114, ACM, 2011.

[12] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, (New York, NY, USA), pp. 261–270, ACM, 2010.

[13] S/A, "The standard for online and internet influence," 2011 (Revisada en Diciembre 2, 2011). Klout, http://www.klout.com/corp/about.

[14] S/A, “We all have kred somewhere,” 2012 (Revisada en Agosto 28, 2012). Kred, http://kred.com/rules.

[15] S/A, “About us,” 2012 (Revisada en Agosto 28, 2012). PeerIndex, http://www.peerindex.com/help/about.

[16] M. Martis, “Detección automática de intención en microblogs,” Master’s thesis, Pontificia Universidad Católica de Valparaíso, December 2011.

[17] M. A. Russell, *21 recipes for mining Twitter*. Sebastopol, Calif: OReilly Media, 1st ed., January 2011.

[18] S/A, “Rest api resources,” 2012 (Revisada en Enero 12, 2012). Twitter Developers, https://dev.twitter.com/docs/api.