UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

DESARROLLO DE UN MÉTODO PARA CALCULAR EL *RANKING* DE POPULARIDAD DE PERSONAJES POR TEMAS EN *TWITTER*

Trabajo final de investigación aplicada sometido a la consideración de la Comisión del Programa de Estudios de Posgrado en Computación e Informática para optar al grado y título de Maestría Profesional en Computación e Informática

FRANCIS ADRIÁN VARGAS BARRANTES

Ciudad Universitaria Rodrigo Facio, Costa Rica

2022

Agradecimientos

Quiero agradecer a todas las personas involucradas en mi formación académica en la Universidad de Costa Rica, a todos los profesores y a mis padres que constantemente estuvieron a mi lado. Especialmente quiero agradecer al profesor Dr. Edgar Casasola Murillo por la guía y el apoyo durante todas las etapas de este proyecto. A la Dra. Gabriela Marín Raventós y al Dr. Gustavo López Herrera, por su gran ayuda a lo largo de toda la investigación, por su tiempo en la lectura y revisión del documento. Un agradecimiento muy especial a mi novia Mariana Fallas Acosta quien me apoyó, motivó y ayudó con la lectura y corrección del documento final. También, agradezco a los expertos en análisis político costarricense, Dr.Carlos Murillo Zamora, Dr. Sergio Salazar Araya, Dr. Oscar Fernández González y M.Sc. Walther Herrera Cantillo por su valioso aporte.

¡Muchas Gracias!

Este trabajo final de investigación aplicada fue aceptado por la Comisión del Programa de Estudios de Posgrado en Computación e Informática de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Computación e Informática.

KRYSCIA RAMIREZ BENAVIDES (FIRMA)
Firmado digitalmente por KRYSCIA RAMIREZ BENAVIDES (FIRMA)
Fecha: 2022.07.29 10:00:50 -06'00'

Dra. Kryscia Ramírez Benavides

Representante de la Decana Sistema de Estudios de Posgrado

EDGAR ENRIQUE CASASOLA MURILLO (FIRMA)
PERSONA FISICA, CPF-01-0729-0430.
Fecha declarada: 29/07/2022 11:25:06 AM
Esta representación visual no es fuente de confianza. Valide siempre la firma.

Dr. Edgar Casasola Murillo

Profesor Guía

GABRIELA MARIN RAVENTOS (FIRMA)
Firmado digitalmente por GABRIELA MARIN RAVENTOS (FIRMA)
Fecha: 2022.07.28 17:16:37 -06'00'

Dra. Gabriela Marín Raventós

Lectora

GUSTAVO LOPEZ HERRERA (FIRMA)
Firmado digitalmente por GUSTAVO LOPEZ HERRERA (FIRMA)
Fecha: 2022.07.29 09:16:28 -06'00'

Dr. Gustavo López Herrera

Lector

ALLAN FRANCISCO BERROCAL ROJAS (FIRMA)
PERSONA FISICA, CPF-02-0553-0149.
Fecha declarada: 29/07/2022 11:58:10 AM
Esta representación visual no es fuente de confianza. Valide siempre la firma.

Dr. Allan Berrocal Rojas

Representante de la Directora del Programa de Posgrado en Computación e Informática

FRANCIS ADRIAN VARGAS BARRANTES (FIRMA)
Firmado digitalmente por FRANCIS ADRIAN VARGAS BARRANTES (FIRMA)
Fecha: 2022.07.26 11:04:37 -06'00'

Francis Adrián Vargas Barrantes

Sustentante

Índice general

Índice de cuadros

Índice de figuras

Índice de fórmulas

.

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Francis Adrián Vargas Barrantes, con cédula de identidad 1 1135 0140, en mi condición de autor del TFG titulado DESARROLLO DE UN MÉTODO PARA CALCULAR EL RANKING DE POPULARIDAD DE PERSONAJES POR TEMAS EN TWITTER

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [X] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

INFORMACIÓN DEL ESTUDIANTE:

Nombre Completo: Francis Adrián Vargas Barrantes.

Número de Carné: A04368 Número de cédula: 1 1135 0140.

Correo Electrónico: f.adrianvargas@gmail.com.

Fecha: 19 / 7 / 2022. Número de teléfono: 6042 6943.

Nombre del Director (a) de Tesis o Tutor (a): Dr. Edgar Casasola Murillo.

FRANCIS ADRIAN VARGAS BARRANTES (FIRMA)

Firmado digitalmente por FRANCIS ADRIAN VARGAS BARRANTES (FIRMA)
Fecha: 2022.07.29 14:58:22 -06'00'

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

Capítulo 1

Introducción

Una de las tendencias actuales dentro del área de análisis de redes sociales es determinar la popularidad de usuarios, ya que es común que la toma de decisiones se vea influenciada por lo que opinan las personas que se consideran populares. Según [Wasserman et al., 1994] el análisis de redes sociales se enfoca principalmente en las interrelaciones entre las entidades de una red social. Además, indican que el objetivo del análisis de redes sociales es la **recopilación**, **procesamiento** e **interpretación** de los datos. Con estos datos se pretende adquirir un conocimiento estructurado que conduzca a decisiones de negocio, como por ejemplo mejorar el servicio al cliente.

Por su parte, [Vise, 2007] muestra que los sitios o páginas web son un medio para la toma de decisiones, sin embargo, deja claro que mostrar las sitios más relevantes a los usuarios en función de sus consultas cada vez es más difícil. Esta problemática se debe a que algunas páginas web no son auto-descriptivas y algunos enlaces existen únicamente con fines de navegación. Por lo tanto, es muy difícil encontrar las páginas adecuadas a través de un motor de búsqueda que se basa en los contenidos web o hace uso de la información de hipervínculos. Con la idea de solucionar este problema han surgido algoritmos como *PageRank* para calcular el *ranking* de las páginas y poder ordenarlas con relación a su popularidad. [Page et al., 1999] también explica como Larry Page y Sergey Brin desarrollaron *PageRank* en la Universidad de Stanford en 1996, esto como parte de un proyecto de investigación sobre un nuevo tipo de motor de búsqueda. En ese proyecto tuvieron la idea de que la información en la web podría ordenarse en una jerarquía por **popularidad de enlaces**, es decir, **se considera una página de rango más alto en la medida que existan más enlaces a ella**.

De manera similar, se pretende adoptar la idea de Larry Page y Sergey Brin [Page et al., 1999] para decir que **"se considera un usuario de rango más alto en la medida que existan más menciones a su nombre"**. Esto se haría utilizando el algoritmo *PageRank* en conjunto con características de *Twitter* menos utilizadas (menciones y *hashtags*) en la literatura, pero no menos importantes para determinar

popularidad y calcular el *ranking*. Este trabajo se basa en proyecto personal que se inició en el curso de Recuperación de Información PF-3394 de la Maestría Profesional en Computación e Informática.

Una vez introducidos en el tema central y un poco en la problemática identificada se procede con los antecedentes. El objetivo de la siguiente sección es mostrar un poco lo que existe relacionado al tema y lo que se identificó en la revisión de la literatura.

1.1. Antecedentes

En esta sección se lleva a cabo una revisión de los antecedentes con la intención de identificar: modelos o algoritmos para calcular el *ranking* de popularidad de personajes por temas en *Twitter*. La diversidad de modelos y algoritmos es amplia, esto se debe a que el estado del arte evidencia diferentes técnicas o adaptaciones de modelos y algoritmos utilizados para calcular *ranking*.

En el trabajo de [Islam et al., 2014] se calcula el *PageRank* de los usuarios de la red social *Twitter* utilizando enlaces retuits, construyen un grafo de estos enlaces y luego este grafo se utilizan como entrada para los cálculo de *PageRank*. Por su parte, en [Chien et al., 2014] hacen una propuesta donde usan los algoritmos *HITS* y *PageRank* para el análisis de clasificación de los usuarios de *Twitter*. En sus resultados comparan los 20 usuarios de *Twitter* mejor clasificados utilizando los dos algoritmos. Años después, [Amati et al., 2019] estudia la evolución de los usuarios más influyentes o con mejor *ranking* en la plataforma de la red social *Twitter* basándose en la cantidad de retuits y menciones utilizando *PageRank*.

Por otra parte, entre las características más conocidas y utilizadas de *Twitter* para calcular *ranking* son: tuit, retuit, menciones, seguimientos. En el estudio de [Amati et al., 2019] utilizan el DRG *(Dynamic Retweet Graph)*, este modelo tiene una característica principal que se basa en el gráfico de retuit, ya que permite representar mejor las relaciones entre los usuarios y el flujo de información en *Twitter*. De manera más elaborada [Li et al., 2020] desarrollaron un método de *Ranking* de confianza de redes duales acopladas (*CoRank*) para evaluar los valores de confianza de los usuarios y tuits involucrados. Construyeron una red de usuarios y una red de tuits que se acoplan entre sí. La red de usuarios está construida con la relación de seguimientos y la red de tuits está construida con relaciones de retuits y respuestas.

Se realizó una revisión de literatura la cual muestra los siguientes resultados.

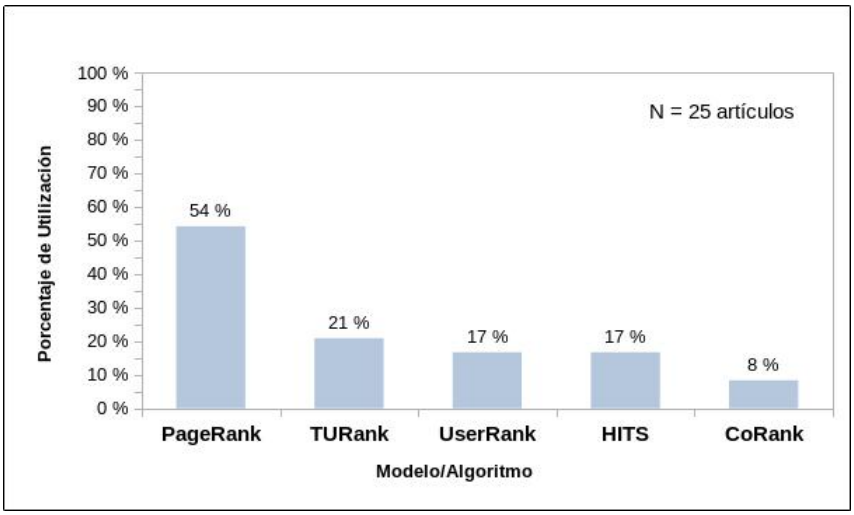


Figura 1.1: Top 5 de Modelos o Algoritmos más utilizados.

La figura 1.1 muestra que de 25 artículos revisados, *PageRank* es el algoritmo más utilizado para calcular *ranking* de usuarios o identificar usuarios influyentes en *Twitter*. En el segundo lugar se puede observar que *TURank* es el segundo modelo más utilizado.

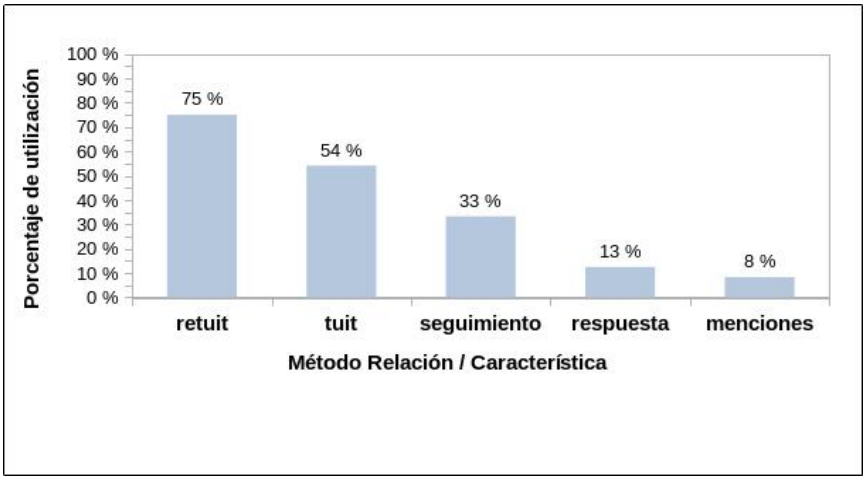


Figura 1.2: Método relación más utilizados.

La figura 1.2 nos indica que los tuit y retuit son las características de *Twitter* más utilizadas para identificar usuarios influyentes o populares. También, es importante

resaltar que las características menciones (@) y *hashtags* (#) no se muestran dentro las más utilizadas.

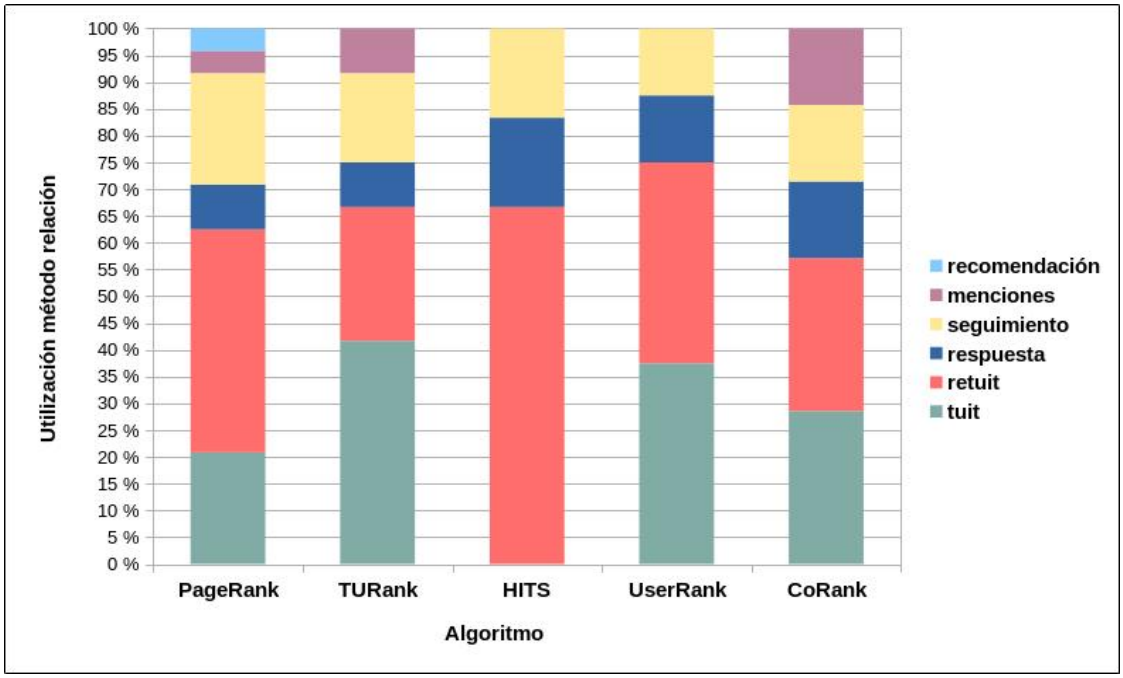


Figura 1.3: Algoritmo y característica más utilizados.

La figura 1.3 resulta de la combinación o cruce de variables presentadas en la figura 1.1 y 1.2. Los anteriores hallazgos muestran dos resultados de interés. Primero, *PageRank* es el algoritmo más utilizado para hacer *ranking* de popularidad en la red social *Twitter*. Segundo, en su mayoría los tuits y retuits son las característica más utilizadas cuando se trata de identificar personajes relevantes o calcular *ranking* de usuarios. Entonces, basados en las ideas anteriores se presenta la justificación y el planteamiento de la problemática identificada en la siguiente sección.

1.2. Justificación y planteamiento del problema

Calcular el *ranking* de personajes relevantes en *Twitter* ha sido un tema interés ya que esto le permite identificar personajes influyentes o relevantes en termino de la cantidad de menciones que estos tengan. Extraer este tipo de información es de importancia

porque personas populares en la mayoría de los casos influyen en las decisión que otros toman.

Es difícil determinar cuales usuarios son más populares en *Twitter* y más aún cuando se esta hablando de algún tema en particular. Esta es una problemática similar a la expuesta por [Page et al., 1999] en su estudio, porque cualquier usuario podría opinar con frecuencia sobre un tema y no necesariamente ser popular o reconocido como popular en la temática.

Se sabe de la revisión de literatura dos cosas: *PageRank* es uno de los algoritmos más utilizados para calcular *ranking* en *Twitter* y que los tuits y retuits son las característica o *features* más utilizadas cuando se trata de calcular *ranking*. Sin embargo, las menciones y los *hashtag* son características de *Twitter* no menos importantes que se pueden utilizar para el mismo objetivo. Entonces, en este trabajo se pretende desarrollar nueva propuesta que se enfocará en la utilización de *PageRank* y estas dos características. La primera, las menciones (@) se utiliza para identificar relaciones entre nodos (usuarios) y la segunda, los *hashtags* (#) se utilizará para identificar temas de interés en los tuits. Con esto se pretende calcular de manera general un *ranking* de usuarios en todos los comentarios y también calcular un *ranking* de usuarios por temas identificados. Esto hace que el modelo sea distinto y permita filtrar por tema en los comentarios, siendo así más específico en los resultados. También aquí, cabe aclarar que además del algoritmo PageRank que se basa en un grafo generado de relaciones, existe un segundo modelo utilizado y mencionado en la literatura que es el TURank [Yamaguchi et al., 2010], esto permite contar con un modelo de referencia para comparar los resultados.

Con el fin de proponer un nuevo modelo y para dar solución al problema mencionado se plantea la siguiente pregunta de investigación: **¿Cómo calcular un *ranking* de personajes en temas de *Twitter* utilizando el algoritmo *PageRank* y las características menciones y *hashtags*?**. Teniendo clara la justificación del problema, a continuación se presentan los objetivos de la investigación.

1.3. Objetivos

Objetivo General

Desarrollar un nuevo método para calcular *ranking* de personajes relevantes a un tema en comentarios de *Twitter*.

Objetivos Específicos

1. Adaptar el algoritmo *PageRank* para usar menciones y *hashtags* de *Twitter* para calcular el *ranking* de personajes relevantes a un tema.

2. Calcular la correlación entre el *ranking* propuesto y el *ranking* de referencia existente (TURank) sobre un conjunto de datos de prueba.

3. Comparar y validar el *ranking* obtenido con el método desarrollado y el de referencia haciendo uso de la valoración de cuatro expertos en tres temas del conjunto de datos de prueba.

Una vez definidos los objetivos de la investigación y la problemática se procede a definir los alcances y limitaciones de la propuesta.

1.4. Alcances y limitaciones

El modelo propuesto en este trabajo es un algoritmo no supervisado, por lo tanto, la investigación no incluye o explica otros tipos de algoritmos de aprendizaje existentes.

Las evaluación del modelo se realizará sobre un conjunto de datos en un dominio específico. No se pretende generalizar las conclusiones a otros dominios.

1.5. Descripción del resto del documento

En el siguiente capítulo 2 se explican los conceptos más relevantes para la comprensión de este trabajo. Se explica la idea principal del análisis de redes sociales, el algoritmo *PageRank* y funcionamiento del coeficiente de correlación de Kendall-Tau que se utilizará para hacer la correlación de los modelos.

En el capítulo 3 de metodología se explican los pasos para recolectar los datos por medio del API de *Twitter*, el preprocesamiento que se hace de los datos, el funcionamiento del programa que se encarga de calcular el *ranking* y la forma de evaluación del modelo.

La descripción del modelo TURank y BOPRank se hace en el capítulo 4. TURank que es uno de los más mencionados en la literatura y es el que se utilizará para hacer la comparación con el modelo desarrollado. BOPRank es el nuevo modelo desarrollado y acá se explica como funciona y cómo muestra los resultados.

En el capítulo 5 se presentan los resultados de la ejecución de los modelos y la correlación obtenida entre ambos. Se muestran los ranking generados para tres temas de política nacional y los resultados de la correlación de ambos usando Kendall-Tau en cada tema presentados en cuadros de ranking y tablas comparativas.

Para tener un punto de vista experto con más criterio sobre los temas analizados, se hizo una presentación de los modelos, sus resultados y una entrevista a expertos en política nacional. En el capítulo 6 se describe estos resultados. Se pretendía comparar BOPRank con TURank, sin embargo, el tema recurrente entre los expertos fue la utilidad de contar con los resultados de ambos a la vez.

En el capítulo 7 se evalúa y analiza la correlación de los modelos y la opinión de los expertos. Se hace una síntesis e interpretación de los resultados finales del proyecto.

Para finalizar las conclusiones y trabajo futuro se exponen en el capítulo 8. Adicionalmente, en el capítulo **??** se encuentra el protocolo usado para la entrevista que se le hizo a los expertos y una lista de temas que se abarcaron durante la entrevista con cada uno de ellos.

Capítulo 2

Marco Teórico

En este capítulo se presentan los conceptos teóricos necesarios para la comprensión del presente trabajo. Se define en qué consiste el análisis de redes sociales fundamentado en las teorías que le dieron origen. Se describe el algoritmo de análisis de enlace *PageRank*, su teoría y a una ejecución simple del mismo para entender su funcionamiento. También, se comenta sobre los orígenes de la red social *Twitter*, como funciona, los *features* o características que posee y que son utilizados para hacer calculo de *ranking*. Se presta especial atención a las menciones y *hashtags* que son características que se van a utilizar en el modelo a proponer.Para finalizar se da una explicación de la métrica conocidas como *Kendall-tau* que se utiliza para hacer correlación de *ranking* o resultados.

2.1. Análisis de redes sociales

Según [Khan, 2015] el análisis de redes sociales es el arte y la ciencia de la extracción de valioso conocimiento oculto en grandes cantidades de datos semi-estructurados y no estructurados de los medios sociales. Estos datos ayudan a la toma de decisiones informadas y acertadas. Es una ciencia, ya que implica una forma sistemática de identificar, extraer y analizar datos de los medios sociales tales como tuits. Es también un arte, porque la interpretación y la alineación de los conocimientos adquiridos es normalmente utilizado con objetivos de mejorar negocios. Para obtener el valor de los datos se debe dominar tanto su arte como su ciencia.

Para el análisis de redes sociales se utilizan canales de redes sociales como *Facebook*, *Twitter*, *blogs*, entre otros. Todas estas redes o medios sociales tienen o utilizan diferentes tipos de datos. Algunas son visibles o fácilmente identificables, por ejemplo, texto o acciones, y otras son invisibles como los hipervínculos y redes. Entre los tipos más conocidos tenemos:

- datos textuales (como los tuits y comentarios).
- datos de la red (tales como amistades de Facebook, seguimiento de *Twitter*).
- acciones (tales como me gusta, acciones, opiniones).
- hipervínculos (por ejemplo, enlaces incrustados dentro del texto).

Según [Ganis y Kohirkar, 2015] hay tres pasos principales en el análisis de las redes sociales: identificación de datos, análisis de datos e interpretación de información. Los analistas pueden definir una pregunta a ser respondida: ¿Quién? ¿Qué? ¿Dónde? ¿Cuándo? ¿Por qué? ¿Cómo?. Estas preguntas ayudan a determinar las fuentes de datos adecuadas para evaluar, lo que puede afectar en gran medida el tipo de análisis que se puede realizar. [Khan, 2015] sugiere que el análisis de redes sociales es un proceso iterativo de seis pasos :

1. **Identificación**: Búsqueda e identificación de la fuente de información correctas para el análisis.
2. **Extracción**: una vez que se identifica una fuente confiable de datos de redes sociales, se procede a la extracción de los datos a través de API o manualmente.
3. **Limpieza**: este paso implica la eliminación de los datos no deseados de los datos extraídos automáticamente.
4. **Análisis**: Se analizan los datos limpios para obtener información importante. Dependiendo de la capa de análisis de medios sociales que se considere y de las herramientas y algoritmos empleados.
5. **Visualización**: Dependiendo del tipo de datos, la parte de análisis conducirá a visualizaciones relevantes para una comunicación efectiva de los resultados.
6. **Interpretación y consumo**: este paso se basa en los juicios humanos para interpretar el conocimiento valioso de los datos visuales. La interpretación significativa es de particular importancia cuando se trata de análisis descriptivos que dan lugar a diferentes interpretaciones.

Ahora bien, en la siguiente sección se presenta una explicación de cómo funciona el algoritmo *PageRank*.

2.2. Algoritmo *PageRank*

El algoritmo de análisis de enlaces *(Link Analysis Algorithm)* se basa en la estructura de enlaces de los documentos. La calidad de los resultados de los motores de búsqueda es generalmente más baja de lo que el usuario espera y esta calidad puede mejorarse enormemente si las páginas se clasifican según algunos criterios basados en enlaces entre las páginas. Es decir, una página que tiene muchas referencias debe tener algo importante que decir. *PageRank* es un ejemplo de algoritmo de análisis de enlaces.

La idea detrás de *PageRank* es que las páginas buenas hacen referencia a páginas buenas. Por lo tanto, las páginas a las que hacen referencia las páginas buenas tienen un *PageRank* más alto. En [Page et al., 1999] Brin y Page desarrollaron el algoritmo *PageRank* basándose en el análisis de mención. El famoso motor de búsqueda Google utiliza el algoritmo de *PageRank*, que es el más utilizado para clasificar las distintas páginas. El funcionamiento del algoritmo de Page Rank depende de la estructura de enlaces de las páginas web. El algoritmo de *PageRank* se basa en los conceptos de que, si una página rodea enlaces importantes hacia ella, los enlaces de esta página cerca de la otra página también deben considerarse importantes. El *PageRank* se refleja en el enlace hacia atrás para decidir la puntuación del rango o *ranking*. Por lo tanto, una página adquiere un alto rango si la adición de los rangos de sus enlaces de vuelta o hacia atrás es alta. Una versión simplificada de la fórmula de *PageRank* se muestra en (2.1).

$$PR(u) = \sum_{v \epsilon B_u} \frac{PR(v))}{L(v)} \tag{2.1}$$

Donde:

- **u** representa un nodo.
- **B(u)** es el conjunto de nodos que apuntan a **u**.
- **PR(u)** y **PR(v)** son *ranking* alcanzados del nodo **u**, **v** respectivamente.
- **Lv** indica el número de enlaces salientes del nodo **v**.

Con el paso del tiempo *PageRank* se adaptó y mejoró dado que observó que no

todos los usuarios siguen los enlaces directos en las paginas. La versión modificada se muestra en la fórmula (2.2).

$$PR(u) = (1 - d) + d \sum_{v \epsilon B_u} \frac{PR(v))}{L(v)} \tag{2.2}$$

De la fórmula (2.2) podemos ver una nueva variable d. d es el factor de amortiguamiento. Este factor de amortiguamiento es importante porque los usuarios solo continuarán dando *click* en los enlaces durante un tiempo limitado antes de distraerse y comenzar a explorar algo que no esté relacionado. Con la probabilidad restante $(1 - d)$, el usuario hará *click* en uno de los enlaces en la página al azar. El factor de amortiguamiento generalmente se establece en 0,85. Por lo tanto, es fácil inferir que cada página distribuye el 85 % de su *PageRank* original de manera equitativa entre todas las páginas a las que apunta. Es decir, el voto total se amortigua al multiplicarlo por 0.85. [Del Corso et al., 2005] indican que este algoritmo es que es altamente susceptible al *spam* y no favorece las páginas importantes con solo unos pocos enlaces internos. También comentan que algunas desventajas identificadas del algoritmo de *PageRank* son las siguientes. Primero, es un algoritmo estático, porque las páginas populares tienden a seguir siendo populares en general. Segundo, a veces la popularidad de un sitio no garantiza la información específica y deseada para el buscador, por lo que también debe incluirse otro factor de relevancia. Finalmente, el algoritmo no es lo suficientemente rápido y en Internet los datos disponibles son enormes. En la siguiente sección se presenta el funcionamiento del algoritmo *PageRank*.

2.2.1. Funcionamiento de *PageRank*

Una manera simple de aplicar y entender el algoritmo de *PageRank* es por medio de un grafo. Los nodos son las entidades, es decir, paginas o personas a las que se le calcula el PR (*PageRank*). Las aristas representan las relaciones entre los nodos que pueden ser referencias o menciones.

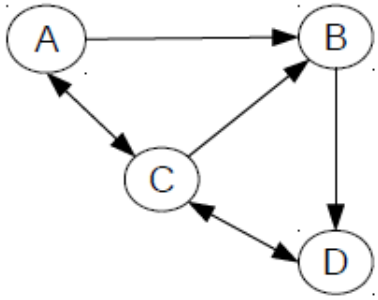


Figura 2.1: Grafo de conexiones o relaciones.

La figura 2.1 es un grafo que representa las referencias o menciones entre paginas o personas. En el caso de las páginas web serian referencias entre páginas y en el caso de los tuits serían las menciones entre usuarios. Una línea punteada significa una relación en solo un sentido, la doble línea punteada siguiente una relación en ambos sentidos. Por ejemplo, en la figura 2.1 A hace referencia o menciona a C y también hace referencia o menciona a B.

Cuadro 2.1: Matriz de Referencias o Menciones.

Nodo	A	B	C	D	...	Lv
A	0	1	1	0	=	2
B	0	0	0	1	=	1
C	1	1	0	1	=	3
D	0	0	1	0	=	1

El grafo también se puede representar en una matriz NxN donde N es la cantidad de nodos, como se muestra en el cuadro 2.1. Las referencias o menciones (aristas en el grafo) se representan por medio de los valores 0 o 1, donde 1 significa que existe referencia o mención entre los nodos y 0 donde no existe ninguna relación. De esta matriz también se puede obtener el Lv o número de referencias o menciones salientes. El Lv se calcula sumando los valores en las filas de manera horizontal. Es decir, sumar las referencias o menciones que se hacen de un nodo a los demás nodos. Debemos recordar que el Lv es muy importante en la fórmula del *PageRank*. A continuación se explica paso a paso la ejecución de una iteración del algoritmo *PageRank*.

2.2.2. Ejemplo primera iteración *PageRank*

En esta sección se ejemplifica el cálculo del *PageRank* (PR) en una iteración aplicando la fórmula para cada uno de los nodos. En este caso sería la iteración 1 dado que la iteración 0 siempre es el PR inicial que se calcula como 1/N, donde N es el total de nodos.

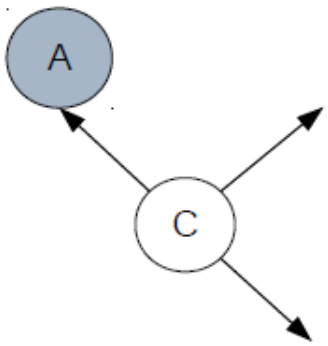


Figura 2.2: *PageRank* I-1 Nodo A.

En la figura 2.2 se observa el cálculo del PR (*PageRank*) del nodo A. Donde el único nodo que apunta a A es C y la cantidad de referencias o menciones que se hacen desde C son 3. Por lo tanto, el PR de A en la iteración 1 sería el PR anterior de A que es 1/4 entre Lv o número de referencias que salen de C, en este caso 3, como se muestra en el cuadro 2.1.

De igual manera se calcula el PR de los demás nodos y después de realizar varias iteraciones aplicando el algoritmo se puede tener un *ranking* de los nodos. Este *ranking* determina la popularidad del nodo.

Cuadro 2.2: Iteraciones *PageRank*.

Nodo/Iteración	I-0	I-1	I-2	I-3	...	Ranking
A	1/4	1/12	1,5/12	0.125	=	4
B	1/4	2,5/12	2/12	0.167	=	3
C	1/4	4,5/12	4,5/12	0.375	=	1
D	1/4	14/12	4/12	0.334	=	2

En el cuadro 2.2 muestra los resultados de aplicar la fórmula simple de *PageRank* en tres iteraciones. La primera iteración siempre es 1/N, donde N es el número de nodos, en este caso 4. En cada iteración se usa el *ranking* de la iteración anterior. En este caso solo se muestran 3 iteraciones, sin embargo, la idea es hacer tantas iteraciones como sean necesarias hasta que alguno de los valores converja a 1. La última columna del cuadro 2.2 muestra el *ranking* de los nodos después de las 3 iteraciones. Se puede observar que después de 3 iteraciones el nodo C es quien tiene el *ranking* más alto. Con esto podemos decir, que es el nodo con más popularidad.

En la red social *Twitter* es muy común que las personas hagan menciones de otras personas, empresas o marcas. Además, fácilmente podemos identificar temas por medio de los *hashtags*. Estas menciones generales o filtradas por temas o *hashtags* se podrían representar en un grafo y por ende podríamos aplicar el algoritmo de *PageRank* y determinar un ranking de popularidad entre usuarios. En las siguientes secciones se hablará sobre la red social *Twitter*, los tuits y los tipos de menciones que existen.

2.3. Red Social *Twitter*

Según [Duan et al., 2010] *Twitter* que es una de las redes sociales más populares en la actualidad, solo permite mensajes de 280 caracteres o menos. Por esta razón se le llama tuits a las publicaciones o mensajes porque al ser mensajes cortos se parece al sonido corto y dulce que puedes escuchar de un pájaro. Cualquier cosa que publiques en *Twitter*, ya sea desde la web o desde un dispositivo móvil se considera un tuit.

Existen diferentes tipos de tuits entre los que se pueden mencionar: Tuits generales, Tuits de imágenes, video tuits, Tuits de localización, menciones y retuits. En este trabajo solo se utilizan las menciones entre usuarios como la manera de determinar la

popularidad entre los mismos. Las menciones se dan en *Twitter* cuando se mantiene una conversación entre dos o más usuarios. Como parte del formato establecido al realizar una mención se debe agregar un signo (@) antes del nombre del usuario. Esto se hace para identificar a quien se hace mención en el tuit. Las menciones solo son públicas entre los usuarios que se siguen y para el usuario que se está mencionando. Siempre que se hace una mención se genera una notificación en el perfil del usuario al que se hace mención. Según [Duan et al., 2010], existen tres formas diferentes en que los usuarios pueden hacer menciones en *Twitter* :

1. **@NombreUsuario**: Cuando las personas utilizan el término "mención de *Twitter*" esto es a lo que se refieren normalmente. Se utiliza el signo @ seguido inmediatamente por un usuario de *Twitter*.

2. **Mención de Marcas**: Esto es cuando alguien menciona una empresa o el nombre de una marca en un tuit sin el @NombreUsuario. Es común que las personas twitteen un enlace en un artículo de blog desde un sitio o que los clientes comenten algo sobre una empresa.

3. **Hashtag**: Una forma común en que las personas hacen conversaciones sobre un tema particular o mencionan marcas en *Twitter* es con un hashtag (#). Como parte del formato establecido al utilizar un hashtag se debe agregar un signo (#) antes del nombre del tema o marca.

En la siguiente sección se explicará la métrica a utilizar para evaluación de resultados.

2.4. Coeficiente de correlación de Kendall

El coeficiente de correlación de rango Kendall-Tau [Kendall, 1938] es una estadística utilizada para medir la asociación ordinal entre dos cantidades. Intuitivamente, la correlación de Kendall al comparar dos *ranking* entre si, será alta si tienen un rango similar (o idéntico para una correlación de 1) o baja cuando las observaciones tienen un rango diferente (o completamente diferente para una correlación de -1), el cuadro 2.3 [Kendall, 1938] muestra valores detallados y grado de correlación. La fórmula matemática de Kendall-Tau se muestra en (2.3).

$$\tau = \frac{C_n - NC_n}{C_n + NC_n} \tag{2.3}$$

Donde:

- n número de observaciones.
- C_n número total de pares concordantes.
- NC_n número total de pares no concordantes (discordantes).

Cuadro 2.3: Rango de valores Kendall-tau

Correlación	**Valores**
fuerte	0.30 o superior
moderada	0.20 a 0.29
débil	-0.10 a 0.19
muy débil	menor que -0.10

Una vez claros los fundamentos teóricos utilizados para la creación de este trabajo, como el análisis de redes sociales, el algoritmo *PageRank*, las características de *Twitter* y coeficiente de correlación Kendall-Tau, en el siguiente capítulo de metodología (Cap.3) se explican los pasos para alcanzar los objetivos planteados.

Capítulo 3

Metodología

Basados en lo que indican [Saunders y Tosey, 2013] este trabajo se considera una **investigación exploratoria** enmarcada dentro de la filosofía del **interpretativismo**. Esto porque se utilizarán características de los tuits poco utilizadas para calcular *ranking* en conjunto con el algoritmo de análisis de enlace *PageRank* para proponer un modelo novedoso. Se utilizará un **caso de estudio** como el instrumento o método de investigación enfocado en un tema específico de **política**.

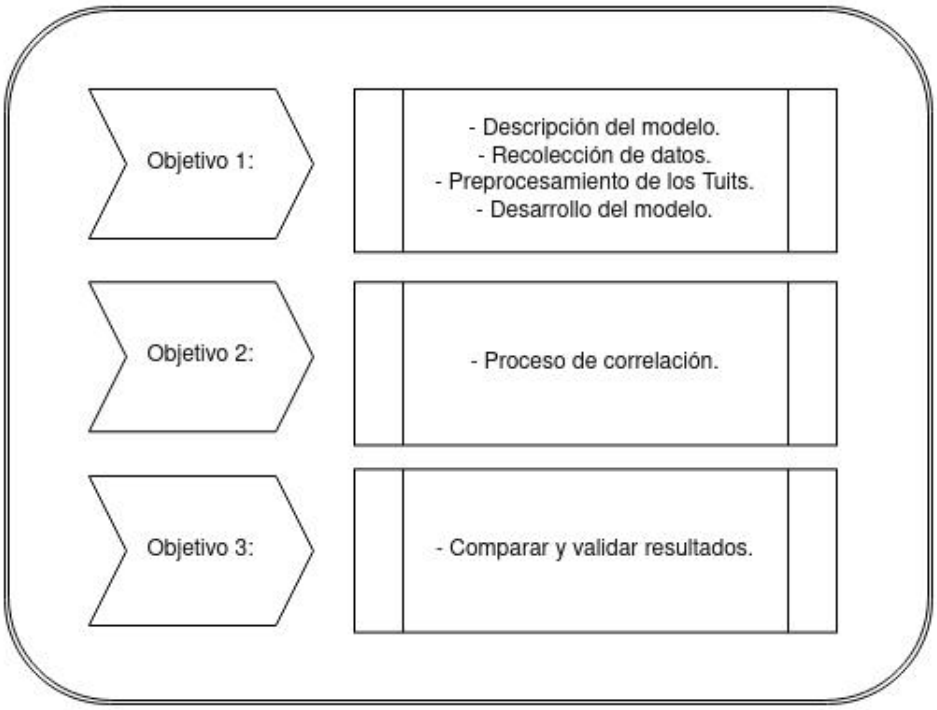


Figura 3.1: Metodología.

Ahora bien, la metodología utilizada para lograr el objetivo principal y cada uno de los objetivos específicos en este proyecto está dividida en las secciones descritas breve-

mente en la figura 3.1. Primero, se describe de manera general el modelo a implementar. Seguidamente se describen los pasos para llevar acabo la recolección de los datos por medio del API de *Twitter*. Tercero, se realiza la descripción de la estructura de los tuits a utilizar para la creación del modelo. En cuarto lugar, se describe el programa desarrollado en Java para la ejecución del algoritmo *PageRank* y obtención de los resultados. Para finalizar, se detalla el proceso de correlación, comparación y validación de los modelos utilizado en este proyecto.

3.1. Descripción del modelo

Con este trabajo se pretende proponer un método para identificar de manera no supervisada a los involucrados o usuarios más populares o relevantes alrededor de un área de conocimiento específicamente en comentarios de la red social *Twitter*.

No supervisado porque no utilizaremos *sets* de prueba ni entrenamiento previo, el modelo recibirá los datos y genera un resultado. Para calcular el *ranking* se utilizará una adaptación de un algoritmo de *ranking* a redes sociales que se llama PageRank.

La relevancia significa quien tenga más menciones en los comentarios analizados es el usuario más relevante o popular. El área de conocimiento se identificará con los *hashtags* (#) en los comentarios alrededor de algún tema como política. Entonces, los tuits se pueden agrupar por temas identificados y el *ranking* de los usuarios se puede calcular en relación a uno de los temas identificados. La recolección de los datos se detalla en la siguiente sección.

3.2. Recolección de datos

Para la extracción de los tuits se utilizará el API de Twitter. En el *query* se puede filtra la búsqueda usando parámetros que se ajusten a los datos que se requieren. Un ejemplo de cómo se extrae la información con el API de Twitter se muestra en la figura 3.2.

```
curl --request POST \
--url https://api.twitter.com/1.1/Tweets/search/fullarchive/adrTweetappdevfullarch.json \
--header 'authorization: Bearer AAAAAAAAAAAAAAAAAAAAAALvdEg8' \
--header 'content-type: application/json' \
--data '{
"query":"#coronavirus has:mentions lang:en",
"maxResults": "100",
"fromDate":"202001012315",
 "toDate":"202006012315",
"next":"eyJtYXhJZCI6MTI2NzMyNjkzODQxNjMyMDUxM30="
}' > covid.json
```

Figura 3.2: Comando para extracción de tuits usando el API de Twitter.

La figura 3.2 es un ejemplo donde se indica que el texto sea en ingles y que tenga menciones. El *maxResults* es el número máximo de tuits que se extraen por ejecución. Se define un rango de fechas con inicio *fromDate* y fin *toDate* para los datos que se van a extraer. El valor *next* se usa para hacer referencia al último request que se hizo y de esta manera en la búsqueda no se repitan los datos y se pueda extraer el siguiente grupo de tuits. El API extrae los datos o tuits en formato JSON y la estructura de los tuits se explica en la siguiente sección.

3.3. Preprocesamiento de los Tuits

Los tuits se convierten de formato JSON a CSV. El archivo CSV se preprocesa y limpia con un programa desarrollado en JAVA que extrae el identificador del tuits, el campo del texto y el nombre de usuario o *username* y el *screenname*. Con estos datos se identificarían relaciones entre los usuarios, y los temas para calcular popularidad de usuarios.

El archivo de tuits debe tener como mínimo los siguientes campos en formato CSV: *tweetid*, *text*, *username*, *screenname*. El *tweetid* se usa para identificar cada tuits, el *screenname* es el nombre del usuario que hace o escribe el tuit y el *text* que corresponde al mensaje o comentario donde están las menciones y los *hashtags* que corresponden a las menciones y temas a identificar, respectivamente.

Una vez claros con la estructura de los datos y la recolección de los tuits en la siguiente sección se procede a detallar el modelo a desarrollar. Se construye un diagrama y se hace una explicación de cómo va a ser el desarrollo.

3.4. Desarrollo del modelo

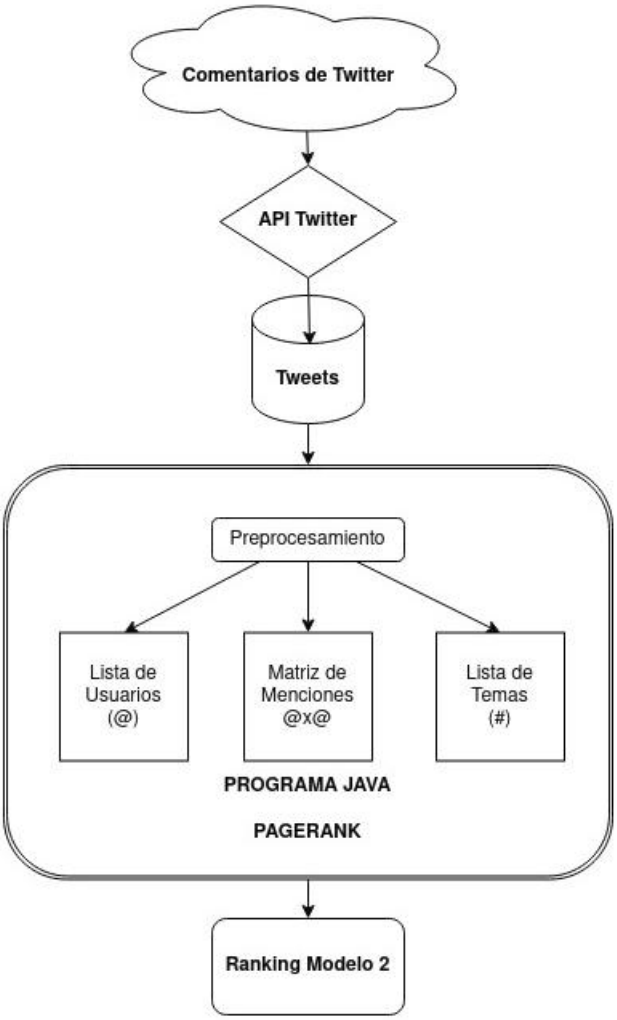


Figura 3.3: Diagrama del modelo.

La metodología de desarrollo a utilizar sería el modelo **prototipado**. Esto porque el objetivo es tener la funcionalidad básica del programa de forma temprana, sin necesidad de incluir toda la lógica o características del modelo terminado.

En resumen, en la figura 3.3 el programa JAVA carga un archivo CSV y hace una lectura de los datos para extraer los campos necesarios. En el código *nextRecord*[0] es equivalente al *id* del tuit, *nextRecord*[3] es el usuario que hace el tuit y *nextRecord*[1]

el texto del tuit que contiene los usuarios a los que se hace referencia y los *hashtags* correspondientes a los temas. Con estos datos se crea un archivo limpio que se usará para crear la matriz de de menciones o relaciones.

La lista de usuarios debe ser única para que la matriz de menciones sea **NxN**. El programa JAVA usa una lista de usuarios única y la lista de tuits donde se hacen las menciones. La primera fase del programa es hacer un barrido de los tuits y crear la matriz de relaciones que es la base para ejecutar el algoritmo *PageRank*. En este caso los usuarios son los nodos y las menciones en el tuit las relaciones entre ellos. En la matriz de menciones un **1** significa que un usuario menciona a otro y **0** en caso contrario.

La matriz de menciones se crea recorriendo el archivo de tuits limpio y usando la lista de usuarios. Por cada tuit se busca el usuario que escribe el tuit y en la matriz se coloca un 1 en la columna donde exista una mención a otro usuario. En esta prueba no se contabiliza la cantidad de veces que un usuario menciona otro en diferentes tuit, es decir, es lo mismo que un usuario mencione a otro en un tuit o en varios tuits. Después de tener la matriz de menciones se lee y se genera el *ranking* aplicando el algoritmo PageRank.

Además, en el programa se crea una lista de *hashtags* y su frecuencia en todos los tuits. Estos *hashtags* corresponden a los temas que hablan en los tuits. La frecuencia se usa para determinar cual *hashtag* o tema es más usado o más mencionado. Se elige el tema más popular, es decir, el *hashtag* con la frecuencia más alta, se extraen los tuits donde se habla sobre este tema y se ejecuta el algoritmo de PageRank para determinar la popularidad de los usuarios involucrados en el tema.

Se debe tener presente que en los tuits debe existir interacción o menciones entre los diferentes usuarios. Además, que existan *hashtags* para identificar temas. De esta manera la matriz de menciones tendrá datos suficientes que permitan ejecutar el algoritmo de manera eficiente.

La evaluación y correlación del modelo desarrollado se explica en la siguiente sección.

3.5. Proceso de correlación

La evaluación es muy importante para validar y comparar el modelo propuesto con un modelo ya existente. El coeficiente de Kendall-Tau [Kendall, 1938] nos permitirá hacer una correlaciones del modelo propuesto en este proyecto. Para el cálculo de

Kendall-Tau se utilizara la formula (3.1).

$$\tau = \frac{C_n - NC_n}{C_n + NC_n} \tag{3.1}$$

Donde:

- n número de observaciones.
- C_n número total de pares concordantes.
- NC_n número total de pares no concordantes (discordantes).

TURank es el segundo modelo más utilizado después de PageRank según la revisión de literatura que se realizó, razón por la cual utilizaremos este modelo para comparar los resultados. La evaluación se realizará de la siguiente manera. Primero, se construirá un conjunto de datos compuesto por tuits pertenecientes a un tema específico de política nacional para que el modelo desarrollado calcule el *ranking* de popularidad. Segundo, se calculará el *ranking* con los modelos, el desarrollado y el modelo TURank. Tercero, se calculará la correlación entre los dos *rankings* generados utilizando el *Rank Correlation Metric* específicamente el *Kendall-Tau coefficient* como métrica para determinar si existe correlación entre los modelos.

3.6. Comparar y validar resultados

Se elegirá un grupo de cuatro expertos en el área de política nacional para comparar los resultado de los modelos y determinar cual muestra mejores resultados en temas que se utilizaron para la construcción del conjunto de datos. Para la selección de los expertos utilizaremos un criterio de conveniencia recurriendo a expertos en Ciencias Políticas de la Universidad de Costa Rica. Sin indicar el algoritmo de procedencia se le solicitará a cada experto que valore los resultados de los modelos sobre cada tema y que indique cual resultado considera mejor.

En la sección **??** de los anexos se especifica el protocolo utilizado para las entrevistas con cada experto. En resumen, se ejecutaron los siguientes pasos:

1. Se les envió un correo solicitando su colaboración.

2. Se definió la fecha de la entrevista.

3. El día de la entrevista se completó el formulario de consentimiento informado para grabar la sesión y/o hacer referencia sus comentarios de manera publica o anónima.

4. Se procedió a la entrevista oral siguiendo la guía del protocolo.

5. Se transcribió la grabación con las respuestas literales.

Para la entrevista se preparó una presentación, que se encuentra en los anexos, sección **??** para guiar al experto con los modelos, los resultados y las preguntas definidas previamente. La presentación y las preguntas se aplicaron de forma intercalada como lo indica el protocolo. La transcripción se facilitó dado que todos los expertos autorizaron a grabar y utilizar su nombre para hacer referencia a sus opiniones, por lo tanto en los resultados se pudo citar el nombre de cada experto con sus respuestas. Con esto se completa la última fase relacionada con la recopilación de criterio experto como se muestra en la figura 3.4.

3.7. Diagrama completo de la metodología

La figura 3.4 muestra el diagrama detallado de la metodología a utilizar para la implementación de la propuesta.

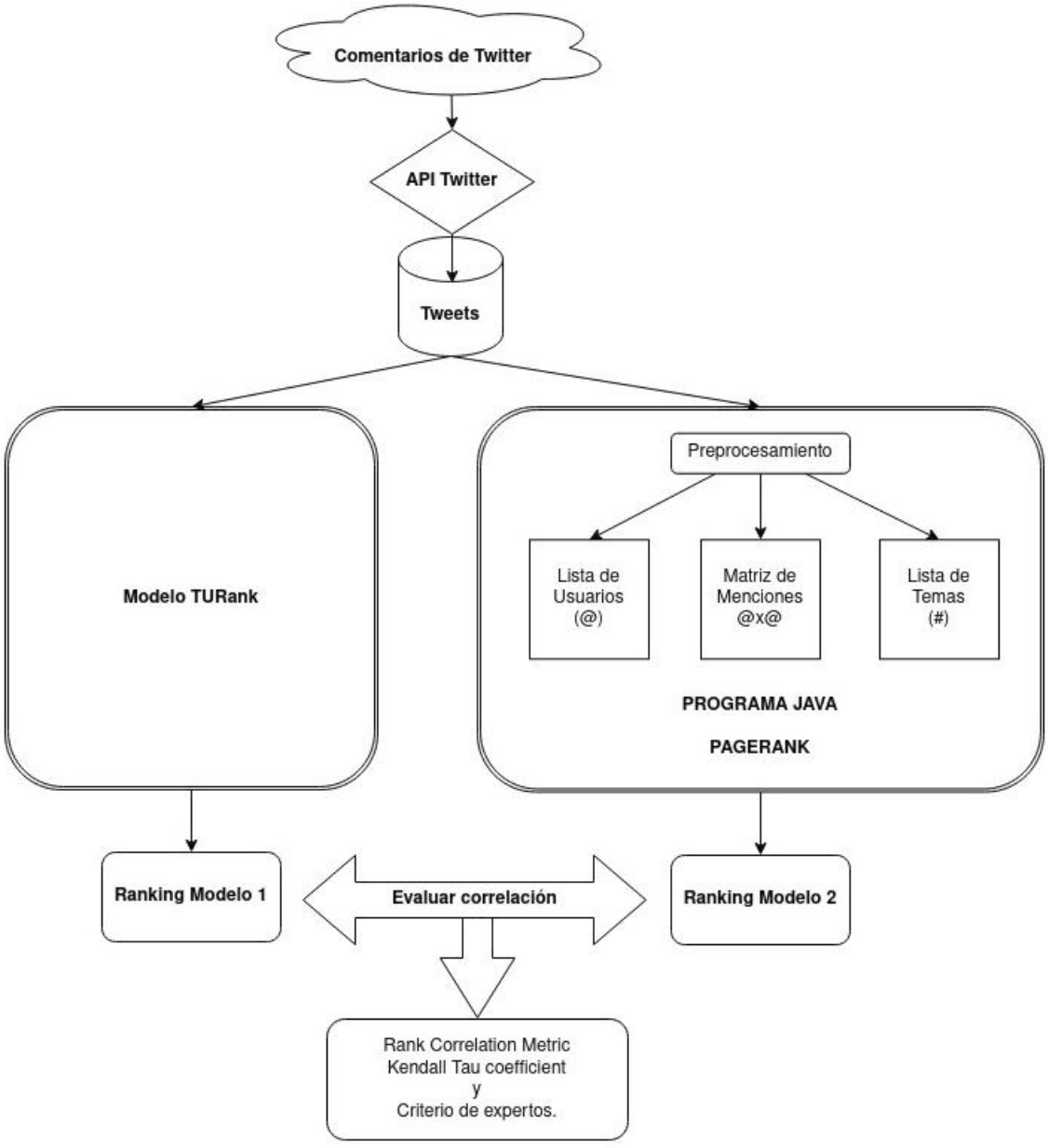


Figura 3.4: Diagrama completo de la metodología

Este capítulo describió la metodología utilizada en el proyecto. La concreción de la

misma brinda los resultados que serán descritos y analizados en los siguientes capítulos. Para comenzarse describen los modelos utilizados en el proyecto en el siguiente capítulo.

Capítulo 4

Descripción de los modelos

Parte del primer objetivo de este proyecto es proponer un nuevo modelo que se describe en este capítulo. Además, se describe el modelo TURank que es uno de los más mencionados en la literatura y es el que se utilizará para hacer la comparación con el modelo desarrollado.

4.1. Modelo TURank

TURank se basa en la idea de que un usuario autorizado que haga muchos tuits, cuyos tuits sean retuiteados por muchos usuarios y que dicho usuario tenga muchos seguidores, es un usuario popular y por ende con mejor ranking [Yamaguchi et al., 2010].

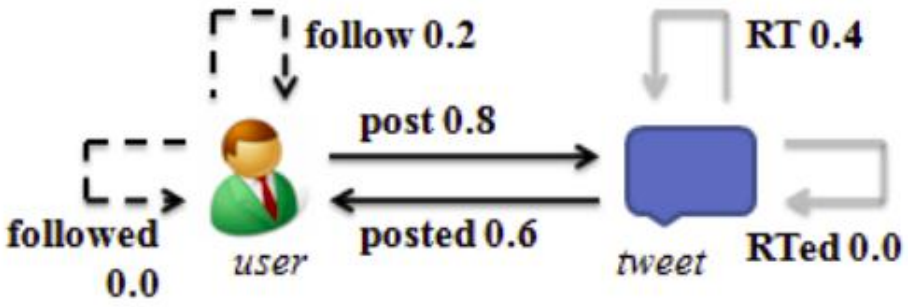


Figura 4.1: Gráfico Modelo TURank.

La figura 4.1 es el gráfico de esquema de tuit de usuario del algoritmo TURank. Este esquema define la estructura y pesos entre los nodos del gráfico. Aquí, existe el conjunto de nodos que consta de nodos de usuario y nodos de tweets, y el conjunto de bordes que consta de publicación, seguidores y ReTuit. Una **arista de publicación** es de un usuario a un tweet publicado por el usuario. Una **arista de seguimiento**

es de un usuario **u** a un usuario seguido por **u**. Un ReTuit es de un tuit **t** a un tuit retuiteado por **t**. Las aristas de publicación, seguimientos y ReTuiteado son los inversos correspondientes a los de seguimiento de ReTuit [Yamaguchi et al., 2010].

El ranking se calculan aplicando la ecuación (4.1) al gráfico de transferencia de datos construido como se muestra en la figura 4.1. La ecuación (4.1) es la misma que la empleada por BOPRank,

$$r = dAr + \frac{(1-d)}{|V|}e \tag{4.1}$$

donde **r** es el vector de puntuación, **d** es la probabilidad de salto aleatorio y **A** es la matriz de transición. Tener en cuenta que el elemento **ij** de la matriz de transición **A** es el peso del borde desde el nodo **i** al nodo **j** si existe; de lo contrario, a **ij** es 0. **V** contiene todos los tipos de objetos de destino, y el conjunto de bordes **e** contiene todos los tipos de aristas existentes entre nodos en **V** [Yamaguchi et al., 2010].

Habiendo obtenido un gráfico de tuits de usuario, intentamos calcular las puntuaciones de cada usuario para la clasificación posterior. El cálculo de la puntuación se basa en la Ecuación (4.1). Aquí, si existe el borde e del nodo **i** al nodo **j**, el elemento a **ij** de **A** es **w(e)** o peso, de lo contrario, a **ij** es 0.

TURank
  $r^0 \leftarrow [1, \cdots, 1]$
  $\alpha \leftarrow 0$
  Repeat
    $\alpha \leftarrow \alpha + 1$
    foreach $r_i^\alpha \in r^\alpha$
      $r_i^\alpha \leftarrow \sum_{e=(j,i)\in E} w(e) r_j^{\alpha-1} + (1-d)/|V|$
    end
    $r^\alpha \leftarrow r^\alpha / ||r^\alpha||_1$
  until $||r^\alpha - r^{\alpha-1}||_1 < \epsilon$
  return r^α
end

Figura 4.2: Algoritmo TURank.

La figura 4.2 muestra el algoritmo para calcular las puntuaciones de TURank en detalle. El marcador del nodo **i** en el paso α se calcula sumando las puntuaciones de todos los nodos que tienen el borde **i** en el paso α - 1 y puntúa por salto aleatorio. Este cálculo es iterado hasta que todas las puntuaciones converjan, donde el umbral de convergencia se establece en ser suficientemente pequeño. Usando este algoritmo se mide los ranking basado en la estructura de enlaces de las relaciones entre usuarios y tuits [Yamaguchi et al., 2010]. Este modelo genera un archivo de texto que tiene el nombre del usuario y el ranking asociado. Con estos datos podemos hacer la comparación con el modelo implementado que se ajustará para que tenga la misma salida. A continuación se describe el modelo implementado.

4.2. Modelo *BOPRank*

El modelo original desarrollado en esta investigación fue denominado BOPRank ya que se basa en el Page Rank, es un seudónimo de la descripción en inglés *Based On PageRank*. La idea detrás de *PageRank* es que las páginas buenas hacen referencia a páginas buenas. Por lo tanto, las páginas a las que hacen referencia las páginas buenas tienen un *PageRank* más alto [Vise, 2007]. De manera similar, se pretende adoptar la idea de Larry Page y Sergey Brin en [Page et al., 1999] para decir que **"se considera un usuario de rango más alto en la medida que existan más menciones a su nombre"**. Esto se haría utilizando el algoritmo *PageRank* en conjunto con características de *Twitter* menos utilizadas (menciones y *hashtags*) en la literatura, pero no menos importantes para determinar popularidad y calcular el *ranking*.

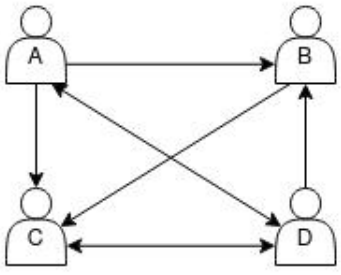


Figura 4.3: Gráfico de menciones.

La figura 4.3 es el gráfico de esquema de menciones entre usuarios en los tuits que

utiliza el modelo implementado como datos de entrada (archivo CSV). Este esquema define la estructura de usuarios como nodos del gráfico y aristas como las menciones entre los mismo. Aquí, existe el conjunto de nodos que consta de usuario y el conjunto de aristas que corresponden a las menciones que se hagan. Una arista en una dirección corresponde a un usuario haciendo mención de otro y las aristas dobles cuando ambos usuarios se mencionan mutuamente en los tuits.

El ranking se calcula usando la ecuación (4.2) que es la misma que utiliza el algoritmo PageRank, con la diferencia de que en este caso se calcula el ranking en relación a las menciones.

$$PR(u) = (1 - d) + d \sum_{v \epsilon B_u} \frac{PR(v))}{L(v)} \tag{4.2}$$

El modelo utiliza un archivo CSV de Tuits y hace una lectura de los datos para extraer los campos necesarios. El identificador del tuit, el usuario que hace el tuit y el texto del tuit que contiene los usuarios a los que se hace referencia y los *hashtags* correspondientes a los temas. Con estos datos se crea un archivo limpio que se usará para crear la matriz de de menciones o relaciones.

PR	A	B	C	D	=Lv
A	0	0	0	1	1
B	1	0	0	1	2
C	1	1	0	1	3
D	1	0	1	0	2

Figura 4.4: Matriz de menciones.

La matriz de menciones 4.4 se crea recorriendo el archivo de tuits limpio y usando la lista de usuarios. Por cada tuit se busca el usuario que escribe el tuit y en la matriz se coloca un 1 en la columna donde exista una mención a otro usuario. No se contabiliza la cantidad de veces que un usuario menciona otro en diferentes tuits, es decir, es lo mismo que un usuario mencione a otro en un tuits o en varios tuits. Después de tener la matriz de menciones se lee y se genera el ranking aplicando el algoritmo PageRank

(4.2).

Además, en el programa se crea una lista de *hashtags* y su frecuencia en todos los tuits. Estos *hashtags* corresponden a los temas que hablan en los tuits. La frecuencia se usa para determinar cual *hashtags* o tema es mas usado o más mencionado. Se elige el tema más popular, es decir, el *hashtag* con la frecuencia mas alta, se extraen los tuits donde se habla sobre este tema y se ejecuta el algoritmo de PageRank para determinar la popularidad de los usuarios involucrados en el tema.

El modelo implementado genera un archivo de texto que tiene el nombre del usuario y el ranking asociado que será utilizado para comparar y evaluar los resultados. En la siguiente sección se hace la evaluación de ambos modelos.

Una vez claros con el funcionamiento de cada uno de los modelos, el existente y el implementado, en el siguiente capítulo se muestra los resultados de la correlación entre ambos.

Capítulo 5

Correlación entre los modelos

Este capítulo describe los resultados de la ejecución de los modelos y la correlación obtenida entre ambos. Se describen los datos utilizados por los modelos, la forma en que se realiza la correlación de los rankings, se muestran el ranking generado por cada modelo y los resultados de la correlación de ambos ranking usando Kendall-Tau.

5.1. Datos utilizados

Como ya se mencionó anteriormente los rankings generados por cada modelo son el producto del análisis de la interrelación existente entre un conjunto de tuits generado a partir de un tema particular. En esta sección se describe la forma en que se procesaron los datos a partir de los tuits para la generación de cada modelo. Los tuits recolectados se almacenan en un archivo tipo CSV (*Comma Separated Values*). Este archivo se genera por medio de un programa desarrollado en R que en resumen hace lo siguiente. Recibe una palabra que sería un *hashtag* o tema a buscar, luego, por medio de un *shell script* se invoca el API de Twitter usando un comando *curl* como se muestra en la figura 3.2 para extraer 100 tuits por solicitud o ejecución. El *shell script* adicional a la palabra recibe un valor siguiente para extraer los 100 tuits, se necesita el valor siguiente para ejecutar el comando *curl* después de la primera ejecución las veces que se desea hacer las solicitudes o las veces que exista el valor siguiente en la solicitud. El valor siguiente se encuentra después de los últimos 100 tuits descargados. Cada solicitud crea un archivo .json y al finalizar todas las solicitudes se crea un solo archivo .json unificado. El archivo final se utiliza para generar un archivo .csv que contiene los valores a utilizar por los modelos.

5.2. Correlación de rankings

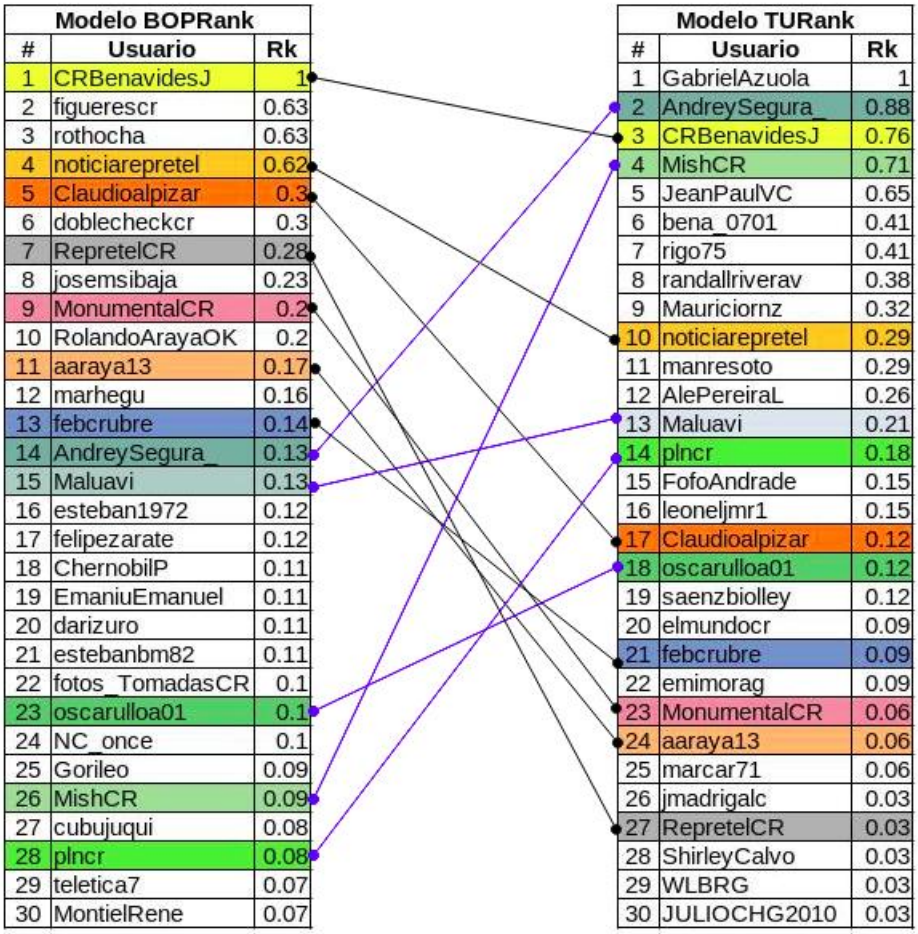
La correlación de ranking se hace con los resultados de cada uno de los modelos. El archivo que genera cada uno que contiene la lista de usuarios y su ranking asociado, ambos archivos se unifican para generar un solo archivo. Es decir, una lista de usuarios con dos ranking correspondientes a cada modelo. Luego, por medio del lenguaje de programación R se calcula la correlación de los ranking usando la formula de Kendall-Tau (3.1) para así determinar si existe similitud entre los modelos. Se hace un primer calculo con todos los usuarios, un segundo con los primeros 30 y un tercer cálculo con los usuarios en común del grupo de los primeros 30 de cada modelo.

Aspecto importante a mencionar antes de visualizar los resultados es que en los cuadros se utilizan colores para identificar los usuarios en común. En ambos modelos hay usuarios con fundo blanco, esto nos indica una de las siguientes opciones. La primera, en el caso del modelo **BOPRank**, el usuario fue mencionado pero no estuvo activo haciendo publicaciones, motivo por el cual no está en los primeros 30 del modelo **TURank**. La segunda, en el caso del modelo **TURank**, el usuario estuvo activo haciendo publicaciones pero fue poco mencionado en los tuits, motivo por el cual no esta en los primeros 30 del modelo **BOPRank**.

A continuación se presentan los resultados analizando tres grupos de temas, mostrando el ranking de cada modelo. Los temas o *hashtags* analizados son: **#DebatePLN**, **#DebateRepretel** y **#DebateTN7**. Se descargaron 2000 tuits por cada uno de los temas para realizar el estudio.

5.3. Tema 1: Debate PLN convención 2021

Para el tema o *hashtag* **#DebatePLN** los datos fueron recolectados en Diciembre del 2021. Estos corresponden a un debate del Partido Liberación Nacional (PLN), presentado en Noticias Repretel, donde participaban todos los aspirantes a candidato presidencial para las elecciones nacionales del 2022. El debate se realizó el día 6 de Junio del 2021 y los precandidatos participantes fueron: Carlos Ricardo Benavides, Roberto Thompson, José María Figueres, Claudio Alpízar y Rolando Araya.



Modelo BOPRank		
#	Usuario	Rk
1	CRBenavidesJ	1
2	figuerescr	0.63
3	rothocha	0.63
4	noticiarepretel	0.62
5	Claudioalpizar	0.3
6	doblecheckcr	0.3
7	RepretelCR	0.28
8	josemsibaja	0.23
9	MonumentalCR	0.2
10	RolandoArayaOK	0.2
11	aaraya13	0.17
12	marhegu	0.16
13	febcrubre	0.14
14	AndreySegura_	0.13
15	Maluavi	0.13
16	esteban1972	0.12
17	felipezarate	0.12
18	ChernobilP	0.11
19	EmaniuEmanuel	0.11
20	darizuro	0.11
21	estebanbm82	0.11
22	fotos_TomadasCR	0.1
23	oscarulloa01	0.1
24	NC_once	0.1
25	Gorileo	0.09
26	MishCR	0.09
27	cubujuqui	0.08
28	plncr	0.08
29	teletica7	0.07
30	MontielRene	0.07

Modelo TURank		
#	Usuario	Rk
1	GabrielAzuola	1
2	AndreySegura_	0.88
3	CRBenavidesJ	0.76
4	MishCR	0.71
5	JeanPaulVC	0.65
6	bena_0701	0.41
7	rigo75	0.41
8	randallriverav	0.38
9	Mauriciornz	0.32
10	noticiarepretel	0.29
11	manresoto	0.29
12	AlePereiraL	0.26
13	Maluavi	0.21
14	plncr	0.18
15	FofoAndrade	0.15
16	leoneljmr1	0.15
17	Claudioalpizar	0.12
18	oscarulloa01	0.12
19	saenzbiolley	0.12
20	elmundocr	0.09
21	febcrubre	0.09
22	emimorag	0.09
23	MonumentalCR	0.06
24	aaraya13	0.06
25	marcar71	0.06
26	jmadrigalc	0.03
27	RepretelCR	0.03
28	ShirleyCalvo	0.03
29	WLBRG	0.03
30	JULIOCHG2010	0.03

Figura 5.1: Ranking con primeros 30 usuarios #DebatePLN

Modelo BOPRank		
#	Usuario	Rk
1	CRBenavidesJ	1
4	noticiarepretel	0.62
5	Claudioalpizar	0.3
7	RepretelCR	0.28
9	MonumentalCR	0.2
11	aaraya13	0.17
13	febcrubre	0.14
14	AndreySegura_	0.13
15	Maluavi	0.13
23	oscarulloa01	0.1
26	MishCR	0.09
28	plncr	0.08

Modelo TURank		
#	Usuario	Rk
2	AndreySegura_	0.88
3	CRBenavidesJ	0.76
4	MishCR	0.71
10	noticiarepretel	0.29
13	Maluavi	0.21
14	plncr	0.18
17	Claudioalpizar	0.12
18	oscarulloa01	0.12
21	febcrubre	0.09
23	MonumentalCR	0.06
24	aaraya13	0.06
27	RepretelCR	0.03

Figura 5.2: Ranking en común #DebatePLN

En la figura 5.1 muestra los ranking de los primeros 30 usuarios de cada uno de los modelos ordenado de mayor a menor. Se marca por colores y se unen con una linea los usuarios en común para identificar su posición en lista. Se identifica que los usuarios en común son bastantes, sin embargo, su posición en la lista es muy distinta. Por dar un ejemplo **RepretelCR** se encuentra en la posición 7 del modelo **BOPRank** pero en el modelo **TURank** se encuentra en la posición 27, denotando que el usuario fue bastante mencionado y no estuvo ente los primeros 10 mas activos publicando.

En la figura 5.2 muestra los usuarios en común de la lista de los primeros 30. Se observa que a pesar de que los ranking no son similares debido a que los modelos se basan en características o *features* distintos de los tuits, ambos modelos identifican bastantes usuarios en común lo cual nos hace pensar que el nuevo modelo también funciona para determinar popularidad.

Con los datos recolectados y usando la formula de Kendall-Tau (3.1) se determina la correlación entre los modelos, usando la lista completa de usuarios, los primeros 30 y de estos 30 los usuarios en común.

Cuadro 5.1: Correlación de Rango de Kendall

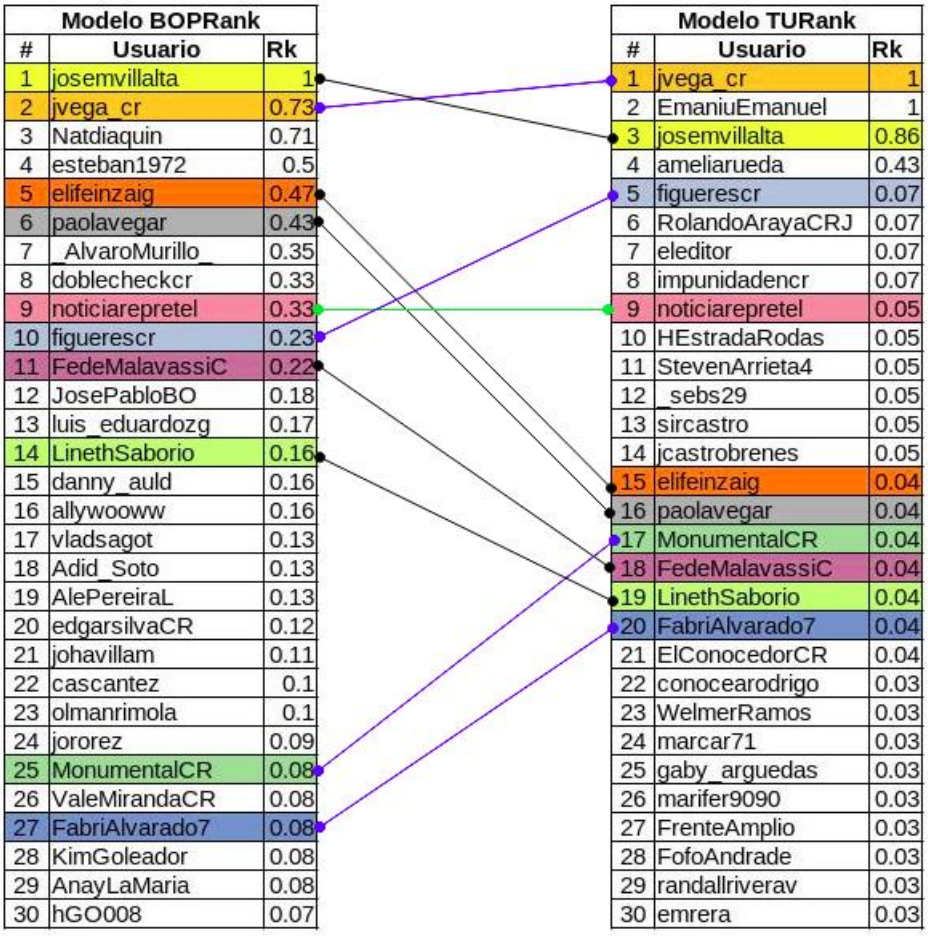
Variables	**Datos**
data rx	BOPRank
data ry	TURank
p-value	2.2e-16
Hipotesis	tau distinto de 0
TAU completo	**0.4048387**
TAU 30 usuarios	**0.2266658**
TAU usuarios en común	**-0.09231862**

El cuadro 5.1 muestra el resultado de la correlación de rango de Kendall para tres casos. En el primero, la lista completa de usuarios es de un **0.40** y basándonos en el cuadro 2.3 donde se definen los rangos de valores de Kendall-Tau para establecer si la correlación es: muy débil, débil, moderada o fuerte; podemos decir que el grado de correlación es fuerte. En la segunda correlación, los primeros 30 usuarios, el valor es de un **0.22**, acá podemos decir que la correlación es moderada. En la tercera, los usuarios en común, podemos ver un valor negativo de **-0.09**, indicándonos que no la correlación es muy débil. Por efecto de la dispersión de la cantidad de candidatos en este tema, la probabilidad de que estuvieran agrupados o que coincidieran en las primeras 30 posiciones los mismos era mas baja y por eso la correlación es negativa. Además porque los pares discordantes son mas que los pares concordantes. Sin embargo, si lo vemos desde un punto de vista de identificación de usuarios relevantes ambos modelos hacen bien su trabajo.

5.4. Tema 2: Debate Repretel primera ronda elecciones 2022

El segundo grupo de datos corresponde al tema o *hashtag* **#DebateRepretel**, fue recolectado en Febrero 2022 días antes de las elecciones presidenciales (primera ronda), el debate fue presentado por Noticias Repretel. El debate se realizó el día 3 de Febrero del 2022 y los candidatos participantes fueron: José María Figueres, del Partido Libe-

ración Nacional (PLN); Lineth Saborío, del Partido Unidad Social Cristiana (PUSC); Fabricio Alvarado, del Partido Nueva República (PNR); José María Villalta, del Frente Amplio (FA); Rodrigo Chaves, del Partido Progreso Social Democrático (PPSD) y Eliécer Feinzaig, del Partido Liberal Progresista (PLP).



Modelo BOPRank

#	Usuario	Rk
1	josemvillalta	1
2	jvega_cr	0.73
3	Natdiaquin	0.71
4	esteban1972	0.5
5	elifeinzaig	0.47
6	paolavegar	0.43
7	_AlvaroMurillo_	0.35
8	doblecheckcr	0.33
9	noticiarepretel	0.33
10	figuerescr	0.23
11	FedeMalavassiC	0.22
12	JosePabloBO	0.18
13	luis_eduardozg	0.17
14	LinethSaborio	0.16
15	danny_auld	0.16
16	allywooww	0.16
17	vladsagot	0.13
18	Adid_Soto	0.13
19	AlePereiraL	0.13
20	edgarsilvaCR	0.12
21	johavillam	0.11
22	cascantez	0.1
23	olmanrimola	0.1
24	jororez	0.09
25	MonumentalCR	0.08
26	ValeMirandaCR	0.08
27	FabriAlvarado7	0.08
28	KimGoleador	0.08
29	AnayLaMaria	0.08
30	hGO008	0.07

Modelo TURank

#	Usuario	Rk
1	jvega_cr	1
2	EmaniuEmanuel	1
3	josemvillalta	0.86
4	ameliarueda	0.43
5	figuerescr	0.07
6	RolandoArayaCRJ	0.07
7	eleditor	0.07
8	impunidadencr	0.07
9	noticiarepretel	0.05
10	HEstradaRodas	0.05
11	StevenArrieta4	0.05
12	_sebs29	0.05
13	sircastro	0.05
14	jcastrobrenes	0.05
15	elifeinzaig	0.04
16	paolavegar	0.04
17	MonumentalCR	0.04
18	FedeMalavassiC	0.04
19	LinethSaborio	0.04
20	FabriAlvarado7	0.04
21	ElConocedorCR	0.04
22	conocearodrigo	0.03
23	WelmerRamos	0.03
24	marcar71	0.03
25	gaby_arguedas	0.03
26	marifer9090	0.03
27	FrenteAmplio	0.03
28	FofoAndrade	0.03
29	randallriverav	0.03
30	emrera	0.03

Figura 5.3: Ranking con 30 usuarios #DebateRepretel

Modelo BOPRank		
#	Usuario	Rk
1	josemvillalta	1
2	jvega_cr	0.73
5	elifeinzaig	0.47
6	paolavegar	0.43
9	noticiarepretel	0.33
10	figuerescr	0.23
11	FedeMalavassiC	0.22
14	LinethSaborio	0.16
25	MonumentalCR	0.08
27	FabriAlvarado7	0.08

Modelo TURank		
#	Usuario	Rk
1	jvega_cr	1
3	josemvillalta	0.86
5	figuerescr	0.07
9	noticiarepretel	0.05
15	elifeinzaig	0.04
16	paolavegar	0.04
17	MonumentalCR	0.04
18	FedeMalavassiC	0.04
19	LinethSaborio	0.04
20	FabriAlvarado7	0.04

Figura 5.4: Ranking en común #DebateRepretel

En la figura 5.3 muestra los ranking de los primero 30 usuarios de cada uno de los modelos. Podemos observar que los dos primero usuarios del modelo **BOPRank** se encuentran en la posición 1 y 3 en el modelo **TURank**. Cuatro de los primeros usuarios en común se encuentran en las 10 primeras posiciones en ambos modelos. Tres de los usuarios en común se encuentran abajo en la lista en ambos modelos.Entonces, podríamos decir que los usuarios en común de ambos modelos se encuentran en posiciones muy similares.

El cuadro 5.4 muestra los usuarios en común de los primeros 30 de cada modelo. Acá podemos observar que los primeros 6 usuarios son los mismos en ambos modelos, con diferentes posiciones debido al ranking pero fueron identificados por ambos modelos.

Con estos datos y usando la formula de Kendall-Tau (3.1) se determina la correlación entre los modelos.

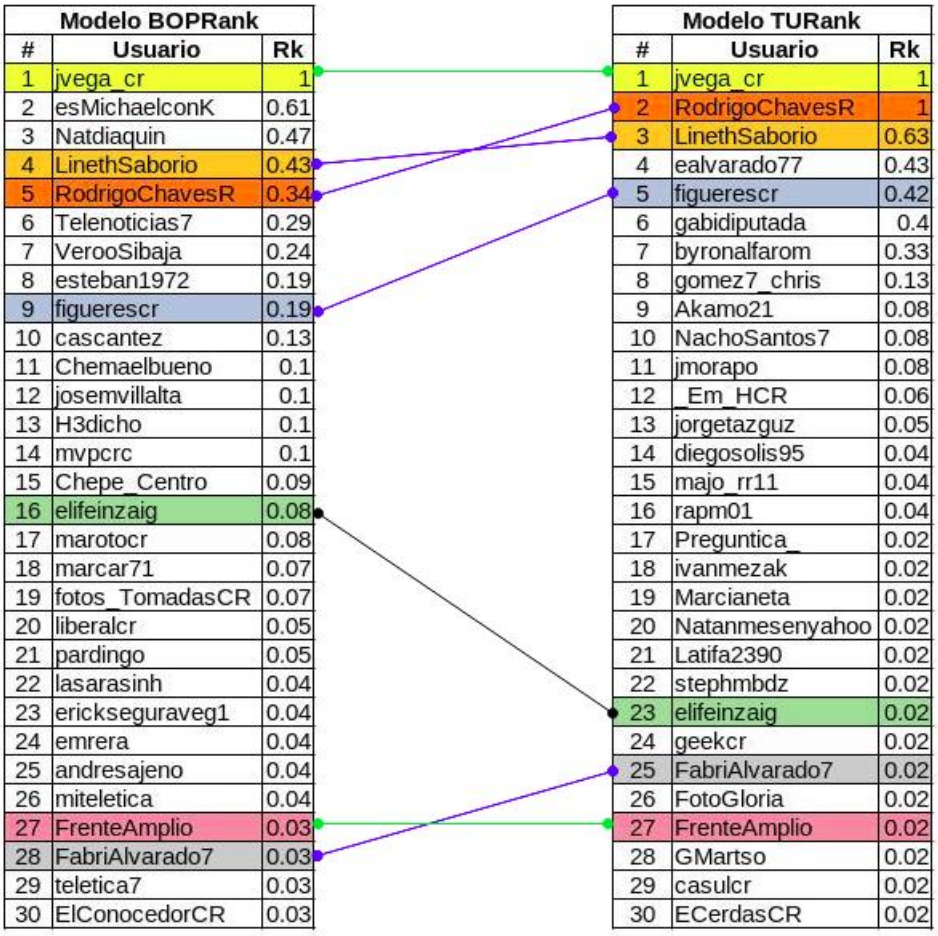
Cuadro 5.2: Correlación de Rango de Kendall

Variables	**Datos**
data rx	BOPRank
data ry	TURank
p-value	1.02e-12
Hipotesis	tau distinto de 0
TAU completo	**0.1868319**
TAU 30 usuarios	**0.1346341**
TAU usuarios en común	**0.6444444**

En el cuadro 5.2 se muestra el resultado de la correlación de rango de Kendall para tes casos. Podemos observar que en los primeros cálculos el valor de correlación es muy bajo, esto con la lista completa y con los primeros 30 usuarios. Sin embargo, la correlación de los usuarios en común muestra un valor de **0.64** lo que demuestra lo mencionado con anterioridad de que los valores en común son similares y por ende la correlación es fuerte.

5.5. Tema 3: Debate Telenoticias primera ronda elecciones 2022

El tercer grupo de datos corresponde al tema o *hashtag* **#DebateTN7**, fue recolectado en Febrero 2022 días antes de las elecciones presidenciales (primera ronda) durante el debate presentado por Telenoticias. El debate se realizó el día 4 de Febrero del 2022 y los candidatos participantes fueron: José María Figueres, del Partido Liberación Nacional (PLN); Lineth Saborío, del Partido Unidad Social Cristiana (PUSC); Fabricio Alvarado, del Partido Nueva República (PNR); José María Villalta, del Frente Amplio (FA); Rodrigo Chaves, del Partido Progreso Social Democrático (PPSD) y Eliécer Feinzaig, del Partido Liberal Progresista (PLP).



Modelo BOPRank		
#	Usuario	Rk
1	jvega_cr	1
2	esMichaelconK	0.61
3	Natdiaquin	0.47
4	LinethSaborio	0.43
5	RodrigoChavesR	0.34
6	Telenoticias7	0.29
7	VerooSibaja	0.24
8	esteban1972	0.19
9	figuerescr	0.19
10	cascantez	0.13
11	Chemaelbueno	0.1
12	josemvillalta	0.1
13	H3dicho	0.1
14	mvpcrc	0.1
15	Chepe_Centro	0.09
16	elifeinzaig	0.08
17	marotocr	0.08
18	marcar71	0.07
19	fotos_TomadasCR	0.07
20	liberalcr	0.05
21	pardingo	0.05
22	lasarasinh	0.04
23	ericksseguraveg1	0.04
24	emrera	0.04
25	andresajeno	0.04
26	miteletica	0.04
27	FrenteAmplio	0.03
28	FabriAlvarado7	0.03
29	teletica7	0.03
30	ElConocedorCR	0.03

Modelo TURank		
#	Usuario	Rk
1	jvega_cr	1
2	RodrigoChavesR	1
3	LinethSaborio	0.63
4	ealvarado77	0.43
5	figuerescr	0.42
6	gabidiputada	0.4
7	byronalfarom	0.33
8	gomez7_chris	0.13
9	Akamo21	0.08
10	NachoSantos7	0.08
11	jmorapo	0.08
12	_Em_HCR	0.06
13	jorgetazguz	0.05
14	diegosolis95	0.04
15	majo_rr11	0.04
16	rapm01	0.04
17	Preguntica_	0.02
18	ivanmezak	0.02
19	Marcianeta	0.02
20	Natanmesenyahoo	0.02
21	Latifa2390	0.02
22	stephmbdz	0.02
23	elifeinzaig	0.02
24	geekcr	0.02
25	FabriAlvarado7	0.02
26	FotoGloria	0.02
27	FrenteAmplio	0.02
28	GMartso	0.02
29	casulcr	0.02
30	ECerdasCR	0.02

Figura 5.5: Ranking con 30 usuarios #DebateTN7

Modelo BOPRank			
#	Usuario	Rk	
1	jvega_cr	1	■
4	LinethSaborio	0.43	▼
5	RodrigoChavesR	0.34	▼
9	figuerescr	0.19	▼
16	elifeinzaig	0.08	▲
27	FrenteAmplio	0.03	■
28	FabriAlvarado7	0.03	▲

Modelo TURank			
#	Usuario	Rk	
1	jvega_cr	1	■
2	RodrigoChavesR	1	▲
3	LinethSaborio	0.63	▲
5	figuerescr	0.42	▲
23	elifeinzaig	0.02	▼
25	FabriAlvarado7	0.02	▼
27	FrenteAmplio	0.02	■

Figura 5.6: Ranking en común #DebateTN7

En la figura 5.5 muestra los ranking de cada uno de los modelos. Se observa que tres usuarios en común se mantienen entre los primeros 5 en ambos modelos. Dos usuarios

se encuentran en la misma posición en ambas listas. Tres de los usuarios en común se encuentran en la parte baja de ambos modelos. Con estos datos y usando la formula de Kendall-Tau (3.1) se determina la correlación entre los modelos.

Cuadro 5.3: Correlación de Rango de Kendall

Variables	**Datos**
data rx	BOPRank
data ry	TURank
p-value	2.2e-16
Hipotesis	tau distinto de 0
TAU completo	**0.2784119**
TAU 30 usuarios	**0.1704573**
TAU usuarios en común	**0.7807201**

En el cuadro 5.3 muestra que el resultado de la correlación de rango de Kendall. Podemos observar que el primero cálculo correspondiente a la lista completa es de **0.27** indicándonos una correlación moderada. En el caso de los primeros 30 usuarios la correlación baja a **0.17**, convirtiéndose en débil. Sin embargo, la correlación de los usuarios en común muestra un valor de **0.78**, comprobando lo mencionado con anterioridad de que en los valores en común los pares concordantes son similares y por ende la correlación es fuerte.

Después de presentar los resultados de cada uno de los modelos y la correlación entre los mismos, en el siguiente capitulo se exponen los resultados del análisis que se hizo con los expertos.

Capítulo 6

Valoración por parte de expertos

En este capitulo se describe los resultados de la evaluación con los expertos, los mismos fueron presentados en cuadros de ranking y tablas comparativas para que los expertos tuvieran claro el trabajo que se hizo. Al obtener información de los expertos se pretendía comparar BOPRank con TURank, sin embargo, el tema recurrente fue la utilidad de contar con ambos a la vez.

Tal como se definió en la metodología. Se procedió a seleccionar a los 4 expertos en Ciencias Políticas de la Universidad de Costa Rica para llevar acabo la evaluación de los resultados. La selección se realizó con el fin de elegir expertos en política nacional que conocieran los temas analizados.

Los entrevistados fueron: Dr.Carlos Murillo Zamora (politólogo y profesor de la UCR), Dr.Sergio Salazar Araya (politólogo y profesor de la UCR), MSc.Walther Herrera Cantillo (figura pública en política y profesor de la UCR) y el Dr.Oscar Fernández González (profesor catedrático de la UCR), todos expertos en política nacional.

A los expertos se les hizo una entrevista donde se les explica el funcionamiento de los modelos y se les muestra los resultados de cada uno de los temas analizados. Después de dejar claro el funcionamiento y los resultados se les hizo consultas para obtener su criterio experto, recomendaciones o cualquier observación. A continuación se presentan el análisis de los resultados y un resumen de los datos mas importantes de las entrevistas.

6.1. Particularidades de los modelos

Como ya sabemos el modelo implementado usando BOPRank se basa en las menciones entre los usuarios en los tuits, mientras que el modelo TURank se base en los seguidores, cantidad de tweets y retweets de quienes publican los tuits. Esto hace que los usuarios identificados por el modelo TURank sean quienes están más activos pu-

blicando tuits, mientras que el modelo implementado adicionalmente toma en cuenta usuarios que sean mencionados en los tuits. Esto se puede observar en el cuadro 5.3 donde el usuario @**figuerescr** hizo muy pocos tuits motivo por el cual el modelo TURank no lo identificó entro los primeros 30 usuarios que se muestran en la lista, sin embargo, el modelo implementado si lo hace y lo coloca en segundo lugar de popularidad en los tuits ya que fue muy mencionado en los mismos. Por otro lado, podemos decir que en todos los temas analizados se identifican usuarios o entes no conocidos por ejemplo @**NachoSantos7** que es un usuario cuyo nombre hace referencia al señor Ignacio Santos, sin embargo, el perfil no corresponde al periodista de canal 7. Esto nos lleva a preguntarnos, qué tan confiables son los tuits cuando el filtrado de usuarios no se hace y que debemos tener claro que nos podemos encontrar entes o autoridades no confiables. Es importante dejar claro que en este proyecto solo se está trabajando con un ranking de popularidad sin importar si el ente o usuario es confiable, real o un perfil falso.

6.2. Conocimiento y consultas de los expertos

De la entrevista hecha a los expertos podemos rescatar que solo uno había trabajado con ranking de popularidad, pero en otras redes sociales como Facebook e Instagram, es decir, ninguno ha trabajando ranking de usuarios en Twitter. Sin embargo, todos están interesados en saber como funciona el ranking de usuarios en Twitter y ver los resultados obtenidos.

Entre las dudas más comunes fueron, saber si el retuit se utilizaba para hacer el calculo en algunos de los modelos, cosa que si se hace en el TURank. Consultaron si la mención debe ir estrictamente con @ para ser utilizada por el modelo BOPRank, cosa que sí funciona de esa manera, las menciones por nombre no se toman en cuenta. Les interesaba saber si se podía identificar tuits positivos y negativos, cosa que no es parte de los objetivos de este trabajo.

6.3. Comentarios sobre temática y fuentes de los datos

Con respecto a los datos utilizados les interesó saber el por qué solo 2000 tuits y si se podía descargar más datos, para lo cual se les indicó que se puede crear una cuenta Académica en Twitter y esto permite descargar 10 millones de tuits por mes. Los temas o datos utilizados les parecen interesantes y adicional a los utilizados, dos de los entrevistados indican que el análisis del debate de la segunda ronda hubiese sido muy bueno ya que es la ronda decisiva. Consideran interesante ver como José María Figures y Rodrigo Chaves se movían en el ranking y ver las tendencias que se podían dar. Uno de los expertos indicó le gustaría analizar temas como la guerra en Ucrania y los ataques de hackers en Costa Rica. En relación con los temas analizados todos los expertos indican que ambos resultados reflejan afinidad con el tema en análisis. El experto Dr. Carlos Murillo lo resume en la siguiente frase: "Ambos tienen ventajas, el tema es saber interpretar los datos y a hacia quien van dirigidos. TURank ayuda a los equipos de campaña mientras BOPRank permite ver la reacción de la gente en redes, en términos de cómo me mencionan".

6.4. Importancia de contar con varios modelos

Con respecto a los resultados de los debates los expertos dijeron que se requieren los dos modelos porque ambos dan información importante y complementaria. Uno dice cuan mencionado fue y el otro la capacidad del manejo de redes sociales. Resaltan el hecho de que los modelos muestran realidades diferentes que se pueden complementar. Sergio Salazar dice: "El cruce de los modelos permite hacer una tipología de los usuarios. Queda claro quienes son generadores de opinión pero no son tan mencionados y quienes pueden ser trolls, quienes se mueven por coyunturas específicas o grupos ". Además, Sergio indica que: "me parece interesante a partir de los resultados de ambos modelos hacer entrevistas a los entes o usuarios. Hacer preguntas en términos de los usos políticos que hacen de las redes sociales, sea o no dentro durante la campaña. Se ha estudiado muy poco y sería interesante saber cómo los actores políticos usan sus redes sociales dentro de un esquema político más amplio". Dos de los expertos les interesa el contenido del tuit y les gustaría saber si la mención es positiva o negativa, esto porque podría ser

que el usuario sea muy mencionado sólo por referencias negativas, les queda claro que esto implica mejoras en el programa. Un experto indica que seria interesante filtrar los usuarios conocidos para comparar los ranking con y sin filtros.

6.5. Ventajas y desventajas de cada modelo

Se le consultó a los entrevistados si observan ventajas entre los modelos pero todos coinciden con el hecho de que ambos miden una realidad distinta, uno los más mencionados y el otro los más activos. Resaltan el hecho de que ambos tienen ventajas, el tema es saber interpretar los datos y a hacia quien van dirigidos. En este caso particular, Carlos Murillo indica: "TURank ayuda a los equipos de campaña mientras BOPRank permite ver la reacción de la gente en redes, en términos de cómo es mencionado ". Indican que con esto datos se puede ver el movimiento del usuario Twitter, cosa que es muy importante en la política. Por su parte el Dr. Oscar Fernández indica que: "en BOPRank no se analiza el contenido de los mensajes, entonces, no se sabe si los comentarios son positivos o negativos, esto lo ve como una desventaja, también indicó que para TURank no se sabe si en realidad es la persona dueña del perfil la que está publicando, esto para él es una debilidad". Adicional a lo anterior todos los entrevistados indican que les gustaría utilizar ambos modelos ya que los resultados de la combinación brindan más información que cada uno por separado.

6.6. Comentarios generales

En cuanto a los resultados en general, al Dr. Sergio Salazar, uno los entrevistados, le surgen la idea de hacer preguntas a los usuarios identificados y le gustaría profundizar en el contenido publicado. El Dr. Sergio indica que: "en temas políticos en particular seria una buena idea ver como se mueve el actor o usuario en diferentes momentos de una campaña (arranque, primero ronda, segunda ronda), esto para ver el movimiento del usuario en los diferentes momentos, definiendo hitos como un debate o escándalo de corrupción, luego hacer una medición, ver el comportamiento del ente y después hacer preguntas al mismo sobre las estrategias del uso de las redes sociales ". Por otro lado el MSc. Walther Herrera indica que: "viéndolo desde el punto de vista de un producto, parecen muy buenos los resultados, porque se puede ver que hace un usuario o ente para

estar mejor que otro y así girar la campaña o cambiar las estrategias para mejorar", entonces, piensa que el complemento es muy bueno. Piensa que sería muy interesante ver una campaña donde salen muy mencionados y muy activos. El MSc. Walther indica que: "con TURank puedo saber cómo estoy en relación a quienes publican y BOPRank me dice que tan eficiente es mi participación en esa red social, que tan bueno es lo que publico para ser mencionado, y que ambos dicen por qué el ente se vuelve popular. También, indica que se podría ver qué se tiene que publicar y cómo para que a la gente le guste, lo mencionen o le hagan un retuit. Indica que con estos datos se le puede dar una línea de cómo hacer publicidad para promocionar un producto y estar ubicado en un buen ranking". Adicionalmente, todos los expetos resaltan el hecho de que ambos modelos son útiles para saber cómo se está manejando la opinión política, quienes participan y verificar quienes son los que publican.

Una vez presentado el ranking generado por cada uno de los modelos, la correlación entre ambos y opiniones obtenidas de la entrevista con los expertos, en el siguiente capitulo se hace un análisis general de los resultados.

Capítulo 7

Análisis de resultados

La evaluación de los resultados del modelo propuesto en comparación al modelo existente y el análisis de la opinión de los expertos es muy importante. En este capítulo se hace una síntesis de la correlación de los modelos y la evaluación de los expertos.

7.1. Síntesis de la correlación

Con respecto a la medición de correlación entre los ranking propuesta en el objetivo 2 se pudo observar lo siguiente. La evaluación de correlación con Kendall sobre los usuarios en común medido para las primeras 30 posiciones, mostraron una alta correlación en el ranking generado por los algoritmos TURank y BOPRank lo cual valida la correctitud de BOPRank como algoritmo de ranking. Se lograron correlaciones de 0,64 y 0,78 en dos de los temas.

Sin embargo, cuando en el cálculo de correlación se toman en cuenta los 30 usuarios en total (no necesariamente en común) la detección de actores que no opinaron pero son populares (novedad) la correlación es moderada. Esto demuestra que BOPRank detecta actores que el otro algoritmo no puede identificar y aún así se mantiene la correlación entre los rankings.

Algunos aspectos importantes de la correlación de rango de Kendall-Tau (cuadro 2.3) son los siguientes:

1. En el **#debatePLN** para la lista completa de usuarios la correlación es de un **0.40**, entonces, podemos decir que el grado de correlación es alto. En la segunda correlación, los primeros 30 usuarios, el valor es de un **0.22**, acá podemos decir que la correlación es moderada.

2. En el **#debateRepretel** en los primeros cálculos el valor de correlación es bajo, 0.1 en la lista completa y con los primeros 30 usuarios. Sin embargo, la correlación

de los usuarios en común muestra un valor de **0.64** y por ende la correlación es fuerte.

3. En el **#debateTN7** en el primero cálculo correspondiente a la lista completa es de **0.27** indicándonos una correlación moderada. En el caso de los primeros 30 usuarios la correlación baja a **0.17**, convirtiéndose en débil. Sin embargo, la correlación de los usuarios en común muestra un valor de **0.78**. De lo anterior comprobamos que cuando los valores en común de los pares concordantes son similares la correlación es fuerte.

4. A nivel cuantitativo basado en los resultados de Kendall-Tau se pudo identificar que ambos modelos tienen similitudes y diferencias importantes que los hacen complementarse apropiadamente para análisis de popularidad en política y muchos otros temas.

Entonces, basándonos en los resultados de los capítulos 5 y 6, podemos decir que el modelo BOPRank es original, se comporta similar al modelo TURank y existe correlación entre ambos demostrada mediante Kendall-Tau. Esto significa que BOBRank es un nuevo aporte a la literatura porque es una manera distinta de generar un ranking. Se están utilizando características de los tuits poco usadas (menciones y *hashtags*) como se mostró en la revisión de literatura (ver figura 1.2) y estamos obteniendo resultados similares. Importante recordar que BOBRank se basa en las menciones entre los usuarios y TURank se enfoca más en actividad y los tuits que haga el ente o usuario activo.

Por otra parte, después de analizar las respuestas y comentarios obtenidos de los expertos se logra determinar que el modelo nuevo esta produciendo resultados de utilidad, sea individual o en conjunto con algún otro modelo como TURank. Individualmente como lo indicaron los expertos BOPRank ayudaría a determinar la popularidad de los usuarios basados en que tan mencionados son y en el impacto que producen sus comentarios o tuits con respecto a un tema. En conjunto con otro modelo como TURank también sería de gran utilidad ya que el cruce de los resultados permitiría visualizar estrategias para mantenerse activo publicando y de igual manera que las publicaciones generen un impacto y los entes o usuarios sean más mencionados. En el ámbito político como lo mencionaron los expertos es de gran importancia en una campaña tener este tipo de información para influir en las personas que usan esta red social.

Lo importante de esta sección es que queda claro que el nuevo modelo se comporta en forma similar a uno existente pero sus resultados son distintos. Es un algoritmo de ranking valido que incorpora la detección de nuevos involucrados en la temática.

También se debe dejar claro que conforme se eliminan los entes o usuarios que no están en común la correlación aumenta demostrando que estamos ante un algoritmo de ranking valido, sin embargo, esa correlación no llega a ser idéntica lo que demuestra que estamos ante un nuevo modelo de ranking. El objetivo de este proyecto era demostrar que el modelo se comporta en forma similar a otros conocidos, pero es diferente.

7.2. Síntesis de la evaluación con expertos

En cuanto a los resultados asociados al objetivo 3 donde se evalua con expertos la calidad del algoritmo BOPRank cabe mencionar que inicialmente se buscaba comparar los modelos entre si, sin embargo, ambos modelos resultaron interesantes y útiles para los entrevistados. Esto resultó ser un aspecto importante a considerar para quienes desarrollan nuevos programas de análisis de redes sociales en el dominio político.

El nuevo modelo es diferente, se enfoca en las menciones y resultó ser interesante para los expertos. Esto nos hace pensar que el modelo podría ser utilizado de manera individual como se pensaba en un inicio o sino también en conjunto con modelos existentes como TURank para obtener resultados que se puedan cruzar y generar información valiosa para quienes lo utilicen.

Se consideró útil el modelo para identificación de entes o usuarios involucrados en alguna temática, esto para realizar posteriores análisis más profundos. Por ejemplo, se podría hacer análisis de sentimiento o análisis de la polaridad de la popularidad, es decir, determinar si la popularidad era positiva o negativa basada en los comentarios. También se podría analizar el origen de esa popularidad.

Después de la síntesis final e interpretación de los resultados, en el siguiente capitulo se presentan las conclusiones y el trabajo futuro que se puede realizar para agregar más funcionalidades al modelo implementado.

Capítulo 8

Conclusiones y trabajo futuro

Al concluir el trabajo de investigación aplicada, se exponen las siguientes conclusiones y recomendaciones como trabajo futuro.

8.1. Conclusiones

Se crea el modelo BOPRank para medición del ranking de popularidad de usuarios en Twitter, basada en el algoritmo PageRank con el uso de menciones y hashtags, cumpliendo con el objetivo general de la investigación.

Se llevó a cabo una comparación del algoritmo nuevo BOPRank contra el algoritmo existente TURank. Y mediante una correlación, se verificó que ambos modelos funcionan correctamente para la generación de ranking. Adicional a lo anterior, el modelo BOPRank tiene como ventaja sobre el modelo TURank que cuando se quiere determinar la popularidad de todos los involucrados en los tuits, este toma en cuenta aquellos usuarios que son mencionados y no sólo los que escriben los tuits como lo hace el modelo TURank. De esta manera se logró cumplir con el objetivo 2 de la investigación relacionado al análisis de la correlación entre TURank y BOPRank.

De la revisión de literatura se determinan 2 puntos importantes: primero, existen otros modelos que hacen ranking de popularidad y segundo, las características menciones (@) y *hashtags* (#) son las menos utilizadas para hacer ranking. Sin embargo, con el modelo BOPRank se pudo demostrar que también se puede generar un ranking valioso usando PageRank, en conjunto con estas características poco usadas.

De acuerdo a lo antes mencionado, se puede decir, que el nuevo modelo es novedoso y de interés ya que produjo un avance en el estado del arte, al demostrar que con las menciones y *hashtags* también se puede generar un ranking desde otra perspectiva. Además, los expertos lo consideran como una manera de ver la opinión de los usuarios y el impacto de las buenas publicaciones al ser mencionados. Esto significa, que el trabajo

tuvo un resultado positivo, al proponer un nuevo modelo que funciona y es útil tanto individualmente como en conjunto con otros métodos de análisis de redes sociales.

Parte de los resultados importantes en esta investigación, es que los expertos indican que del cruce de ambos modelos de ranking se puede obtener información más exacta y completa, que de manera individual. Adicionalmente, concluyen que sería interesante trabajar en aspectos relacionados con el filtrado de usuarios falsos y saber si los comentarios son positivos o negativos para determinar el tipo de popularidad que se está midiendo.

Se tiene claro, que los tuits podrían ser fuentes no confiables, ya que existen usuarios falsos, especialmente durante las campañas políticas, como en los temas analizados. En ellas, estos usuarios son creados y utilizados para dar popularidad a otros, o para hacer comentarios fuera de contexto y que no aportan al tema que se está tratando. Estas son funcionalidades adicionales que se podrían agregar al modelo o posibles ideas a trabajar en un futuro.

En conclusión, se identificaron temas interesantes a nivel de aplicación de los algoritmos de ranking de usuarios en redes sociales, con temas en el dominio político, adicional, se logra cumplir con los objetivos de la investigación.

8.2. Trabajo futuro

Al concluir esta investigación, se visualizan nuevas funcionalidades o ideas a implementar en un futuro. Esas funcionalidades se citan a continuación.

Primero, una funcionalidad adicional al programa es determinar si las menciones son positivas o negativas y de esta manera mostrar resultados dependiendo de lo que se quiera medir. Lo anterior, debido a que actualmente solo se está tomando en cuenta si existen o no menciones, sin lograr identificar si son positivas o negativas. De igual forma, no se puede determinar si la popularidad es buena o mala.

Segundo, ajustar el programa para mostrar los temas o *hashtags* principales y que se pueda elegir el tema para calcular la popularidad.

Tercero, hacer análisis en tiempo real de reacciones a temas específicos que se discuten en Twitter, y así determinar la popularidad de los usuarios.

Cuarto, incluir en el programa un filtro para validación de usuarios reales, que

permita detectar impostores o perfiles falsos.

Quinto, filtrar los usuarios relacionados al tema que se está analizando. De esta manera no solo se eliminan perfiles falsos, sino también aquellos que no corresponden al tema. Esto permitirá obtener resultados específicos en un grupo reducido de entes o usuarios que se quieren estudiar.

A continuación se prestan los anexos del proyecto que corresponden al protocolo de la entrevista realizada a los expertos y los temas abordados durante la entrevista que se le hizo a los mismos.

Bibliografía

[Amati et al., 2019] Amati, G., Angelini, S., Gambosi, G., Rossi, G., y Vocca, P. (2019). Influential users in twitter: detection and evolution analysis. Multimedia Tools and Applications, 78(3):3395–3407.

[Chien et al., 2014] Chien, O. K., Hoong, P. K., y Ho, C. C. (2014). A comparative study of hits vs pagerank algorithms for twitter users analysis. En 2014 International Conference on Computational Science and Technology (ICCST), pp. 1–6. IEEE.

[Del Corso et al., 2005] Del Corso, G. M., Gulli, A., y Romani, F. (2005). Fast pagerank computation via a sparse linear system. Internet Mathematics, 2(3):251–273.

[Duan et al., 2010] Duan, Y., Jiang, L., Qin, T., Zhou, M., y Shum, H. Y. (2010). An empirical study on learning to rank of tweets. En Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 295–303.

[Ganis y Kohirkar, 2015] Ganis, M. y Kohirkar, A. (2015). Social media analytics: Techniques and insights for extracting business value out of social media. IBM Press.

[Islam et al., 2014] Islam, M., Ding, C., y Chi, C.-H. (2014). Personalized recommender system on whom to follow in twitter. En 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, pp. 326–333. IEEE.

[Kendall, 1938] Kendall, M. G. (1938). A new measure of rank correlation. Biometrika, 30(1/2):81–93.

[Khan, 2015] Khan, G. F. (2015). Seven Layers of Social Media Analytics. CreateSpace.

[Li et al., 2020] Li, P., Zhao, W., Yang, J., Sheng, Q. Z., y Wu, J. (2020). Let's corank: trust of users and tweets on social networks. World Wide Web, 23(5):2877–2901.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., y Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

[Saunders y Tosey, 2013] Saunders, M. y Tosey, P. (2013). The layers of research design. Rapport, (Winter):58–59.

[Vise, 2007] Vise, D. (2007). The google story. Strategic Direction.

[Wasserman et al., 1994] Wasserman, S., Faust, K., et al. (1994). Social network analysis: Methods and applications, volume 8. Cambridge university press.

[Yamaguchi et al., 2010] Yamaguchi, Y., Takahashi, T., Amagasa, T., y Kitagawa, H. (2010). Turank: Twitter user ranking based on user-tweet graph analysis. En International Conference on Web Information Systems Engineering, pp. 240–253. Springer.

A P É N D I C E S

Apéndice A

Protocolo para entrevista

Protocolo para contacto inicial con expertos.

Se enviará al correo electrónico de los potenciales expertos la siguiente invitación a participar en la investigación.

> Estimado señor/señora [Nombre Apellido] :
>
> La Dra. Gabriela Marin, Directora del Posgrado en Computación e Informática de la UCR me sugirió su nombre para colaborar emitiendo su opinión sobre dos modelos que estoy analizando como parte de mi trabajo final de investigación aplicada en el Programa de Posgrado en Computación de la UCR.
>
> Los modelos que se están comparando se utilizan para generar ranking de usuarios en Twitter. Esto permitiría identificar quienes son personas o entes populares en las publicaciones. En el primer modelo los más populares son los más mencionados y en el segundo quienes están más activos haciendo publicaciones son los más populares.
>
> Los resultados son interesantes y queremos compartirlos con expertos en ciencias políticas como ustedes, ya que las publicaciones utilizadas en la investigación son temas de política nacional recientes. Creemos que podría ser interesante para ustedes participar ya que las herramientas podrían ser de utilidad para el análisis de involucrados en temas de la política actual. Adicional a esto, en caso de que les llame la atención algún modelo, podríamos en un futuro cercano, aplicarlo en algún tema que les interese analizar y obtener los resultados que les ayude a tomar alguna decisión o sino evidenciar algún comportamiento.
>
> Por otra parte, es muy importante para mi investigación tener la opinión de personas como ustedes ya que me permitirá mostrar los resultados con más propiedad teniendo un criterio experto en los temas analizados. El tiempo estimado de la entrevista o intercambio de opiniones es de aproximadamente 30 minutos.
>
> En esta investigación estoy trabajando con la Dra. Gabriela Marin, Directora del Posgrado en Computación e Informática de la UCR, el Dr. Gustavo Lopez y el Dr. Edgar Casasola, profesor guía.En caso de aceptar la invitación favor indicarnos si prefiere la sesión via ZOOM o presencial, y su disponibilidad para coordinar fechas y los detalles.
>
> Agradezco su colaboración.
>
> Saludos cordiales,

Protocolo sesión con expertos.

Descripción:

Este protocolo tiene como objetivo definir el proceso que se utilizará para evaluar la calidad de dos modelos de ranking de usuarios en Twitter. La evaluación se hará por medio de la opinión de expertos en los temas analizados.

El tiempo estimado para aplicar la entrevista es de aproximadamente 30 minutos.

Pasos:

1. Llenar hoja de consentimiento informado.

Consentimiento Informado

Nombre:

Profesión:

Lugar de trabajo:

Su nombre puede ser mencionado en el trabajo final haciendo referencia a sus opiniones. Si así lo desea por favor marcar con x la opción Sí. En caso contrario marcar con x la opción No y mantendremos sus comentarios de manera anónima.

- Sí ( )
- No ( )

Con el fin de transcribir sus respuestas y opiniones se le informa que esta sesión va a ser grabada. Las grabaciones serán borradas una vez transcrita la información.

2. Entregar la presentación adjunta al entrevistado para que tenga acceso a la información.(https://docs.google.com/presentation/d/1MIkKZ3rI4JDFutjSbAM_ydu9YmJi1Rskyf5w4jIU3P8/edit#slide=id.p).

3. Leer la diapositiva 2 (Definición de Ranking) para que la persona comprenda que es un ranking.
   3.1. Consultar: ¿Ha trabajando con ranking de usuarios en Twitter o en algún otro ámbito ?
   3.2. Consultar: ¿Le interesaría tener alguna herramienta que identifique quiénes son los principales involucrados en una temática ordenados de mayor a menor prioridad?

4. Leer la diapositiva 3 (Diagrama) para que la persona comprenda el trabajo que se está realizando.
   4.1. Consultar: ¿Queda claro lo que hace el programa?
   4.2. Consultar: ¿Tiene alguna duda o comentario al respecto?

5. Leer la diapositiva 4 (Trabajo a realizar) y 5 (Su opinión) para que la persona comprenda el trabajo que se está realizando.
   5.1. Consultar: ¿Antes de ver a detalle cada modelo, tiene alguna duda o comentario al respecto?

6. Leer la diapositiva 6 (Modelo **BOPRank**) para entender cómo funciona.
   6.1. Consultar: ¿Tiene alguna duda o comentario de cómo funciona este modelo?

7. Leer la diapositiva 7 (Modelo **TURank**) para entender cómo funciona.
   7.1. Consultar: ¿Tiene alguna duda o comentario de cómo funciona este modelo?

8. Leer la diapositiva 8 (Datos recolectados) para explicar los datos que se utilizaron y la cantidad de tuits.
   8.1. Consultar: ¿Tiene algún comentario?
   8.2. Consultar: ¿Qué otro tema le interesaría analizar?

9. Leer la diapositiva 9 (Resultados #DebatePLN) para mostrar los resultados correspondientes a ese tema.
    9.1. Consultar: ¿Qué modelo le parece más apegado a la realidad?
    9.2. Consultar: ¿En cuál de las dos listas considera usted que las primeras posiciones están ocupadas por entes que se relacionan más al tema?
    9.3. Consultar: ¿Por qué opina así?

10. Leer la diapositiva 10 (Resultados #DebateRepretel) para mostrar los resultados correspondientes a ese tema.
    10.1. Consultar: ¿Qué modelo le parece más apegado a la realidad?
    10.2. Consultar: ¿En cuál de las dos listas considera usted que las primeras posiciones están ocupadas por entes que se relacionan más al tema?
    10.3. Consultar: ¿Por qué opina así?

11. Leer la diapositiva 11 (Resultados #DebateTN7) para mostrar los resultados correspondientes a ese tema.
    11.1. Consultar: ¿Qué modelo le parece más apegado a la realidad?
    11.2. Consultar: ¿En cuál de las dos listas considera usted que las primeras posiciones están ocupadas por entes que se relacionan más al tema?
    11.3. Consultar: ¿Por qué opina así?

12. Preguntas finales

    1. ¿Observa alguna ventaja del BOPRank (más mencionados) sobre el TURank (más activos)?
        1.1. ¿Por qué opina de esa manera?

    2. ¿Observa alguna ventaja del TURank (más activos) sobre el BOPRank (más mencionados)?
        2.1. ¿Por qué opina de esa manera?

    3. ¿Qué modelo utilizaría y en qué casos?

4. ¿Qué opina de los resultados?
    4.1. ¿Por qué opina así ?
    4.2. ¿En qué se basó para dar su opinión ?

5. ¿Al día de hoy existe algún tema para el cual le interesaría identificar en forma automática quienes están involucrados y ordenarlos del más al menos popular?
    5.1. ¿Los resultados de cuál modelo le gustaría observar ?
        - Opción a: Modelo BOPRank, que se basa en los más mencionados.
        - Opción b: Modelo TURank, que se basa en los más activos.
        - Opción c: Ambos modelos.
    5.2. ¿Por qué ?

13. Pedir retroalimentación al entrevistado.

¿Tiene algún comentario final sobre los modelos, la presentación o el trabajo en general ?

14. Agradecimiento

Se le agradece mucho su tiempo y colaboración con las respuestas.

Apéndice B

Descripción de la presentación

En la presentación que se hizo a los expertos se abarcaron los siguientes puntos para dejar claro el trabajo realizado, los modelos, lo que se esperaba de ellos y los resultados. En resumen se habló de los siguientes aspectos.

1. Definición de *ranking*.
2. Que hace el programa implementado.
3. El trabajo a realizar.
4. Lo que nos interesa de los expertos.
5. Modelo BOPRank.
6. Modelo TURank.
7. Resultados DebatePLN.
8. Resultados DebateRepretel.
9. Resultados DebateRepretel.