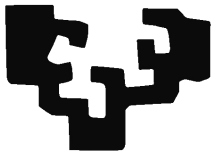

Modelando las decisiones de contribución de contenido de usuarios en una red social

Por

Pablo Andre Cleveland Ortega

Depositado en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad del País Vasco para optar al grado de
Doctor en Informática

Bajo la dirección de:

Prof. Dr. Manuel Graña Romay
Dr. Sebastián Ríos Pérez

Universidad del País Vasco
Euskal Herriko Unibertsitatea
Donostia - San Sebastián

2022

Modelando las decisiones de contribución de contenido de usuarios en una red social

por
Pablo Cleveland Ortega

Depositado en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad del Pais Vasco para optar al grado de Doctor en Informática

Resumen

El creciente uso de los servicios de internet, particularmente de las redes sociales en linea (OSN), ha generado una gran oportunidad para entender mejor el comportamiento de los usuarios como también de los flujos de información. El modelado de flujos de información no es un tema nuevo, pero la aparición de OSNs y comunidades virtuales de practica (VCoPs) proporcionan nuevas fuentes de datos que han revitalizado la investigación en esa area. La mayoría, si no todos, de los estudios revisado modelan las OSN a un nivel macroscópico, donde la agregación de eventos no permite observar comportamientos a un nivel de usuario. Nuestra hipótesis es que es posible modelar la difusión de información a nivel microscópico mediante un modelo derivado de la neurofisiología.

El objetivo principal de este trabajo es desarrollar e implementar una metodología para predecir el intercambio de información entre usuarios de una VCoP a un nivel microscópico explotando el contenido de texto mediante técnicas de mineria de texto. Una posible aplicación es apoyar el proceso de administración de una VCoP.

La metodología propuesta combina dos procesos, por un lado un proceso de

descubrimiento de conocimiento (Knowledge Discovery in Databases (KDD)) y por otro analisis de redes sociales (Social Network Analysis (SNA)). Hemos realizado la demostración sobre una VCoP real llamada Plexilandia. En la etapa de KDD se efectuó la selección, limpieza y transformación de los posts de los usuarios, para luego aplicar una técnica de análisis de tópicos usando *Latent Dirichlet Allocation* (LDA), que permite describir cada post en términos de los tópicos descubiertos de forma no supervisada. En la etapa de SNA se aplicó un modelo neurofisiológico de toma de decisiones adaptado a preferencias de texto para predecir la formación de relaciones entre hilos de conversación en la VCoP y usuarios usando la información obtenida en la etapa anterior.

Los resultados de los experimentos muestran que es posible predecir con un alto porcentaje de éxito las interacciones entre usuarios basándose en la similaridad de los textos producidos por ellos. Hemos obtenido precisión en la predicción de la contribución de un usuario a un hilo de conversación que varia entre el 65 % al 80 % cuando hay poco ruido, y del 40 % al 60 % cuando la conversación tiene elevado ruido. Esto permite vislumbrar la forma en que se difundirá un mensaje e identificar a usuarios que potencialmente estén interesados en un hilo.

Keywords: *Redes sociales en linea, Comunidad de Práctica, Análisis de redes sociales, Análisis semántico, Latent Dirichlet Analysis.*

.

Agradecimientos

El trabajo de esta tesis ha sido parcialmente financiado por ANID (CONICYT-PCHA/MagísterNacional/2015-22151700, CONICYT-PFCHA/Doctorado Nacional/2019-21190971), fondos FEDER para el proyecto MINECO TIN2017-85827-P, el proyecto KK-2018/00071 de la convocatoria Elkartek 2018 del Gobierno Vasco, y el proyecto H2020-MSCA-RISE CybSPEED de numero 777720.

Pablo Cleveland Ortega

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En este capítulo se revisa la motivación principal para estudiar las decisiones de generación de contenido por parte de los usuarios en las redes sociales en línea, seguido de la exposición de los objetivos principales y específicos de esta tesis doctoral. Después de introducir la metodología utilizada para el desarrollo de esta investigación presentamos el caso de estudio sobre el que hemos trabajado en las demostraciones prácticas. Finalmente, presentamos las contribuciones identificadas de la tesis, los resultados en forma de publicaciones que respaldan la tesis, y se da una breve descripción de los capítulos restantes.

1.1. Contexto general

El masivo uso de los servicios de Internet, como las redes sociales, permitió a las personas comunicarse e interactuar entre sí, sin preocuparse por su ubicación geográfica. Es posible que alguien encuentre personas con quienes conversar, personas con intereses comunes, ayudar a otros en ciertos proble-

mas, compartir información, participar en discusiones, etc. Estas actividades cambiaron el uso de la computadora de una actividad individual a una colectiva. uno, este a su vez se ha encargado de crear diferentes vínculos de interacción y cooperación con otras personas [16]. Todo lo anterior ha contribuido a una creciente relevancia de Internet en nuestra vida cotidiana.

La importancia de Internet ha llevado al surgimiento de nuevas instituciones sociales [9, 16]: Redes Sociales en Línea (OSN), Comunidades Virtuales (VC), junto con otro tipo de entidades sociales. Aunque, en base a las existentes, poseen características específicas [11] que deben ser consideradas al momento de realizar un estudio de las mismas. Estas diferencias se deben al uso de un medio diferente a la interacción cara a cara, lo que genera muchos rituales sociales del mundo real [10] que no existen o están limitados en el mundo virtual [14].

Para apoyar estas últimas nuevas estructuras sociales, es posible utilizar diferentes tecnologías. Por ejemplo, uno puede usar un sistema wiki o un foro, un sistema de blogs, un sistema de mensajería, listas de correo electrónico, entre muchos otros. Además, es posible utilizar una combinación de más de una de estas tecnologías para apoyar una comunidad o una red social en línea. El uso de estas tecnologías no solo enriquece la información compartida, sino que también acelera el proceso de difusión.

El uso generalizado de estas tecnologías ha generado una oportunidad para estudiar la forma en que los usuarios interactúan entre sí, la influencia que tienen sobre los demás, la forma en que cierta información se propaga en la red, etc. Aunque se sabía que estos problemas son importantes, eran muy difíciles de medir antes de OSN. Nos centramos en el problema de comprender cómo se propaga o se comparte la información dentro de una OSN. Este problema es de particular interés y encuentra aplicaciones en muchas áreas, tales como: generar con éxito una campaña de marketing viral, campañas políticas, poder detectar rumores maliciosos o inexactos e incluso prevenir

ataques terroristas, medir y rastrear eventos sociales como olas revolucionarias, enfermedades difusión, entre muchos otros.

Las OSN se pueden separar en diferentes grupos según las características que tengan, por ejemplo, Facebook, Sina-weibo y Twitter poseen características (relaciones de amigos o seguidores) que permiten reconstruir el gráfico de la red social de manera directa. También se puede decir que las relaciones sociales juegan un papel importante en estos OSN. Además, la información compartida en estas redes suele ser más ligera en contenido y profundidad y está muy influenciada por las interacciones sociales. Esto nos lleva a seleccionar otro tipo de OSN, Foros Web. Estos tienen muchas características deseables para estudiar la difusión de información, por ejemplo, debido a la falta de amistad o relaciones de seguidores, el enfoque principal de los usuarios cuando navegan en un foro web es el contenido de las conversaciones (hilos) contenidos en él, lo que hace que el contenido del foro sea -impulsado. Otra característica deseable es que brindan una plataforma verdaderamente abierta y de libre acceso para la difusión de información, como se indica en [69] porque cualquiera puede comenzar un nuevo hilo o participar en uno existente, lo que facilita que las opiniones se formen y compartan libremente. Además, la información publicada en OSN como Facebook tiene el problema de que primero es filtrada por un algoritmo de recomendación, y es bien sabido que este tipo de algoritmos, basados en el historial de navegación, restringen el contenido que presenta y navega el usuario, lo que por supuesto , aportará información sesgada a nuestro estudio. Consideramos que el contenido de la información es de suma importancia para comprender el proceso de difusión, razón por la cual esta investigación se centrará en los foros web.

Los estudios realizados en OSN se pueden separar según describan el comportamiento macroscópico, mesoscópico o microscópico de la estructura. Ejemplos de los dos primeros son los estudios de densidad de la representación de la red de la comunidad u otras características de la red global como la

distribución del grado de la ley de potencia que se puede asociar a la observación empírica de que la mayor parte del contenido es producido por un pequeño porcentaje de usuarios de la web, como mencionado en el trabajo de Baeza-Yates [74].

A medida que tratamos de profundizar en los mecanismos de difusión, tratando de comprender el papel de cada agente, el problema se vuelve cada vez más complejo debido principalmente a la cantidad exponencialmente creciente de posibles interacciones. En este trabajo nuestro interés es modelar la toma de decisiones de los agentes en el proceso de difusión, por lo que entramos en la última categoría. Podemos ver claramente el aumento de la complejidad si tomamos, por ejemplo, un problema típico que cae en la categoría de nivel microscópico y uno que cae en lo mesoscópico-macroscópico. Por ejemplo, tome el problema de obtener el conjunto de arcos formados en una red de N nodos para el nivel micro y el problema de obtener la densidad de borde para el nivel macro-meso. Si hay k arcos en la red, la probabilidad de obtener el conjunto correcto es:

$$P(\text{Acertar los k arcos}) = \frac{1}{N+1}\frac{1}{\binom{N}{k}} \tag{1.1}$$

Y la probabilidad de obtener la densidad correcta de nodos es:

$$P(\text{Acertar la densidad de nodos}) = \frac{1}{N+1} \tag{1.2}$$

Donde $P(\text{Acertar todos los k arcos}) \ll P(\text{Acertar la densidad de nodos})$ para la mayoría de los casos.

También cabe destacar que con los resultados de un modelo micro se pueden obtener posteriormente resultados a nivel meso o macro, sin embargo no es posible al revés.

Recapitulando, nuestro trabajo se centrará en modelar la toma de decisiones

de los agentes en el proceso de difusión en los foros web, con un enfoque basado en el contenido y el punto de vista de un administrador web. Creemos que conocer las conversaciones que pueden ser de interés para un usuario en particular es de gran relevancia para el administrador de un foro web. Por ejemplo, le permite recomendar discusiones recién creadas en el foro. Además, conocer los detalles de la difusión de información dentro del foro le permite tener un mejor juicio en el desempeño de sus funciones de administrador.

Las principales problemáticas a las que nos enfrentamos al tratar de modelar la difusión de información son las siguientes:

- Las redes sociales suelen ser extremadamente grandes y complejas, además, si agrega el componente de información (contenido de texto) a la ecuación, el problema se vuelve aún más difícil de modelar y procesar.
- La mayoría de las técnicas se enfocan solo en una pequeña porción de las fuentes de información disponibles. Por ejemplo, por un lado, la mayoría de los algoritmos y estadísticas de análisis de redes sociales (SNA) realizan un análisis automatizado para recopilar información valiosa sobre la estructura de la comunidad en función de las relaciones entre los miembros de la comunidad. Por otro lado, el enfoque de minería de datos, en particular la minería web (WM), que es la aplicación de algoritmos de minería de datos a los datos generados en la web, donde se pierde la estructura de las interacciones sociales pero nos permite encontrar patrones interesantes de textos en los miembros. publicaciones o patrones de navegación [31, 34, 38, 43].
- La gran mayoría de los modelos se centran en obtener resultados a nivel agregado, p. cantidad de personas con interés en un determinado tema. Sin embargo, o no es posible adaptarlos al nivel microscópico o cuando es posible los resultados ya no tienen un nivel de calidad aceptable.

- Hay una falta de estandarización en el campo, es decir, la mayoría de los trabajos sobre difusión están hechos de tal manera que no es posible hacer comparaciones. En el trabajo realizado por Guille et al [53] se intentó estandarizar los trabajos con Sondy una plataforma en línea que permite a los investigadores realizar la mayoría de los algoritmos SNA. Sin embargo, no fue adoptado por la comunidad de investigadores.
- No hay documentación sobre los muchos pasos que deben tomarse durante el proceso de desarrollo de un modelo para la difusión de información, lo que hace que el proceso en sí sea mucho más propenso a errores.

1.2. Motivación

Nuestro problema es complejo desde el punto de vista del modelado y procesamiento de datos. Pero al mismo tiempo es sumamente relevante para poder comprender la dinámica del comportamiento humano en las redes sociales en línea, que hoy ya se han consolidado como un mecanismo fundamental en muchos ámbitos de la vida cotidiana.

1.3. Hipótesis y Objetivos de la Tesis

La hipótesis de trabajo en esta Tesis es la siguiente:

H1: es posible usar el conteido semántico para predecir las decisiones de contribución de contenido de usuarios de un foro web

El objetivo general de este trabajo de tesis es desarrollar e implementar una metodología para predecir el intercambio de información entre usuarios a

nivel microscópico utilizando el contenido de texto extraído mediante técnicas de Minería de Texto, para soportar el proceso de administración de un Foro Web.

Los objetivos específicos que se desprenden del objetivo general son los siguientes:

1. Revisar la literatura de difusión de información para pintar una imagen del estado actual y crear un punto de referencia de modelos
2. Desarrollar una metodología para el modelado de difusión de información en foros web utilizando el modelo LCA.
3. Implementar un modelo que nos permita capturar el proceso de toma de decisiones de los agentes que participan en la red

1.4. Metodología Utilizada

La metodología de trabajo se basa en el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD) y SNA. Trabajaremos con los datos obtenidos de un Foro Web real. Sobre los datos provenientes de esta misma red social, se han realizado ya otros trabajos de análisis de redes sociales como la búsqueda de miembros clave en el Foro Web [34].

Posteriormente, utilizando la estrategia de minería de datos y texto antes mencionada, se calcularán las entradas del modelo LCA. A continuación, se realizará la calibración del modelo LCA mediante el uso del algoritmo genético y luego se ejecutarán las simulaciones de la red como se muestra en la Fig. 1.1.

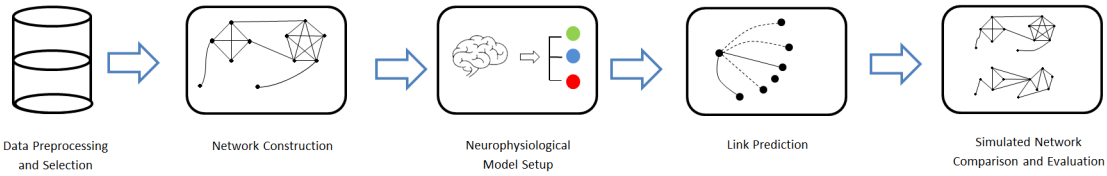

Figura 1.1: Metodología Propuesta para Modelar la Difusión de Información en Foros Web.

La metodología utilizada para el desarrollo de esta tesis se lleva a cabo en los siguientes pasos:

1. **Avances previos en el área y Difusión de la Información**
   Para el desarrollo de este trabajo de tesis, se realizará una breve introducción de las aplicaciones del SCN, describiendo sus áreas de estudio. Posteriormente se hará una revisión de los diferentes métodos que abordan el problema de la difusión de información basados en SNA y algunos de los últimos avances en el área. Además, se revisarán algunos métodos que buscan combinar la minería de texto que utilizan modelos temáticos o conceptos para mejorar la predicción de enlaces en redes.

2. **Representaciones de Grafos**
   Las representaciones de las redes sociales pueden ser diferentes dependiendo de las características del problema del SNA a abordar. Por eso es necesario utilizar estrategias que permitan la construcción de grafos que capturen toda la información contenida en una red social y permitan el posterior análisis con herramientas SNA con las que se aborde el problema de la difusión de la información.

3. **Personalización de algoritmos**
   El modelo LCA debe integrarse y personalizarse para el problema de

predicción de enlaces utilizando contenido de texto como la información disponible de la red. Además, se debe implementar un esquema de optimización para calibrar los parámetros del modelo.

4. **Simulación de la red**
   El algoritmo se probará en un foro web real. Primero, utilizando los datos del trabajo previo realizado por Ríos et al. Se realizará la modificación [31, 34, 41, 48, 71] para extraer los gráficos de red y características de los nodos necesarios para la posterior implementación del algoritmo. A continuación, se ejecutará el algoritmo personalizado y se obtendrán las redes simuladas de los foros de forma que permita la evaluación del algoritmo y, por tanto, de la metodología.

5. **Análisis de resultados y conclusiones**
   Una vez implementada la metodología propuesta, sus resultados serán evaluados mediante el uso de 4 métricas diferentes. Finalmente, se presentarán las conclusiones de cada una de las etapas descritas anteriormente.

1.5. Contribuciones de la Tesis

En el marco de la resolución del problema planteado, este trabajo de tesis contribuye con:

(a) una revisión del estado del arte, de los aspectos de analisis de redes sociales existentes y sus estrategias actuales de medición,

(b) un método para describir el funcionamiento de una estructura social tipo comunidad de práctica en la que cada ususario puede contribuir nuevos conocimientos, mediante la combinación de análisis estructural y análisis semántico, y

(c) una guía para interpretar los resultados obtenidos de la evaluación cuantitativa.

1.6. Publicaciones

La siguiente publicación es el resultado directo del trabajo reportado en esta Tesis:

- Neuro-semantic prediction of user decisions to contribute content to online social networks. Pablo Cleveland, Sebastián A. Ríos, Felipe Aguilera, Manuel Graña. Neural Computing and Applications DOI: 10.1007/s00521-022-07307-0

1.7. Estructura de la Tesis

La estructura de la Tesis consta de los siguientes capítulos:

- **Capítulo 2**: se presenta una revisión de la bibliografía más relevante en materia de difusión de la información
- **Capítulo 3**: se describe un foro web real, llamado Plexilandia
- **Capítulo 4**: la metodología propuesta para implementar el modelo de difusión de información es descrita
- **Capítulo 5**: Los detalles de los experimentos realizados y sus resultados son explicados.

- **Capítulo 6**: se presentan las principales conclusiones de esta tesis, incluyendo las principales contribuciones de este trabajo así como las próximas líneas de investigación y trabajo futuro.

Capítulo 2

Estado del arte

En este capitulo se presenta el marco teórico y los trabajos relacionados con esta tesis. En primera instancia, introducimos algunas definiciones necesarias para enteder el trabajo realizado.

2.1. Redes sociales

Hay una gran cantidad de investigaciones que se enfocan de las redes sociales. Parte de este trabajo se centra en descubrir comunidades dentro de la red [41, 43, 58], influencers y descubrimiento de miembros clave [18, 26, 34, 36, 40, 60, 61, 71], propiedades macroestructurales de redes [12].

Otros estudios se han centrado en describir la evolución de ciertas redes como [30, 31, 54]. por ejemplo, Jianwei Niu et al [54] lleva a cabo un análisis descriptivo de un OSN chino similar a Facebook llamado Renren donde se describen las propiedades evolutivas y estructurales macroscópicas de la red y parecen coincidir con los resultados de trabajos anteriores realizados en

redes similares.

Nos centraremos en trabajos que modelan la difusión de información y vinculan la predicción, ya que adoptaremos un enfoque combinado en nuestro modelo propuesto. A continuación se da una explicación más detallada de los mismos.

2.2. Difusión de Información

Comenzamos esta revisión definiendo difusión, que se refiere al proceso mediante el cual un fenómeno de interés (p. ej., información, innovación o enfermedad) se propaga entre los miembros de una red social [69]. Guille et al [52] presentan una revisión de investigaciones previas sobre difusión de información en el que dan algunas definiciones básicas y clasifican trabajos previos según su respectiva contribución y novedad. También describen los diferentes enfoques utilizados para modelar la difusión de información y definen tres preguntas que están en el centro del campo, a saber:

1. ¿Qué piezas de información o temas son populares y más difundidos ?
2. ¿Cómo, por qué y por qué vías se difunde y se difundirá la información en el futuro?
3. ¿Qué miembros de la red juegan un papel importante en el proceso de difusión?

En esta investigación tratamos de responder a la segunda pregunta. Como se señaló en [52], la difusión de información se ha abordado de muchas maneras. Por lo tanto, haremos una clasificación de los modelos propuestos en familias como se muestra en la Fig. 2.1 para comprender mejor cada uno de estos enfoques y poder presentar investigaciones previas de manera ordenada.

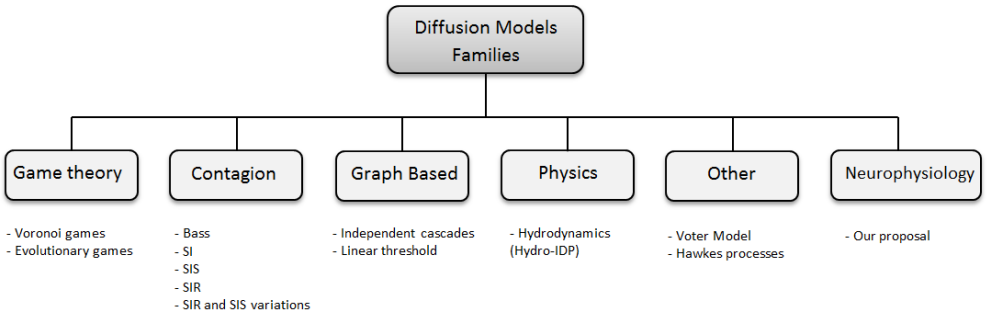

Figura 2.1: Clasificación de modelos de difusión

2.2.1. Modelos de teoría de juegos

Un enfoque que se usa con frecuencia para lidiar con la difusión de información se basa en la teoría de juegos, en particular, los juegos de Voronoi en grafos [21, 32, 49, 55, 56]. Por ejemplo, en Nora Alon et al [32] se muestra un modelo teórico de juego competitivo para la difusión que puede ser útil para comprender situaciones en las que los productos de la competencia se anuncian a través de campañas de marketing viral. También afirman una relación entre el diámetro de una red dada y la existencia de un equilibrio de Nash puro en el juego que luego fue corregido por Reiko Takehara et al [49]. Después de eso, Lucy Small y Oliver Mason [56] demostraron que la afirmación es cierta si el gráfico subyacente es un árbol y luego, en [32], ampliaron sus resultados extendiendo el modelo a un gráfico que sigue la transitividad local iterada modelo para redes sociales. Demostraron que para 2 agentes que compiten, un equilibrio de Nash independiente en el gráfico inicial sigue siendo un equilibrio de Nash para todos los tiempos posteriores. En [57] utilizan el juego evolutivo sobre redes de grado uniforme para modelar las estrategias de reenvío de información de los usuarios, es decir, reenviar la información o no. Ellos prueban este modelo sobre el conjunto de datos de hashtags de

Twitter validando su modelo propuesto.

2.2.2. Modelos de contagio

Otra familia de modelos que se utiliza con frecuencia cuando se trata de la difusión de información son los modelos de propagación de epidemias o de contagio. El primer modelo al que nos referiremos es el Modelo Bass, en [3] presenta un modelo para las ventas de un nuevo producto en función del tiempo que se origina a partir de modelos de contagio. Los modelos más comunes que se encuentran en esta categoría son: susceptible-infectado (SI), modelo susceptible-infectado-susceptible (SIS) y modelo susceptible-infectado-recuperado (SIR), entre otras variaciones.

Ye Sun y otros en [59] estudiaron el impacto del uso de bordes ponderados para representar relaciones de funciones múltiples en una red al estudiar dos métricas macroscópicas representativas, umbral de brote y prevalencia epidémica. Realizaron experimentos utilizando 2 conjuntos de distribuciones de peso, distribución Uniforme y Poisson, en una red de mundo pequeño y una red sin escala. Probaron los modelos SIS y SIR. Sus resultados muestran una buena concordancia con los resultados teóricos excepto que los resultados de la simulación muestran que el efecto de distribución del peso es muy débil. El principal resultado de este trabajo es que, en redes completamente mixtas, la distribución del peso en los bordes no afectaría los resultados de la epidemia una vez que se fija el peso promedio de toda la red. Saxena et al [62] propusieron un modelo con probabilidades jerárquicas de infección en los bordes que dependen de la posición relativa de los nodos finales en la red (núcleo o periferia). Probaron el modelo con datos de twitter obteniendo patrones de comportamiento de difusión similares.

Kubo et al. [23] aplicó el modelo SIR para capturar el comportamiento hu-

mano en una comunidad virtual, específicamente "de 2 canales"el sistema de tablón de anuncios (BBS) anónimo abierto más grande de Japón.

Jiyoung Woo et al. en [44] implemente el modelo SIR a nivel de tema para modelar la difusión de temas violentos. Probaron este modelo en el conjunto de datos Ummah del portal del foro web oscuro desarrollado por el laboratorio de inteligencia artificial de la universidad de Arizona. Posteriormente, en [50] prueban y modifican el Modelo SIR anterior para incorporar el efecto de las noticias online, proponiendo el modelo SIR impulsado por eventos. Este modelo captura el efecto de las noticias en línea sobre la tasa de infección, el crecimiento de la población y el crecimiento del grupo infectado. Probaron este modelo en Yahoo! Tablero de mensajes de Finance-Walmart y use noticias relacionadas con Walmart del Wall Street Journal. A continuación, en [69] proponen un modelo de contagio (epidemia) SIR para modelar la difusión de información en Web-Forums debido a patrones similares en la difusión de información y procesos de contagio social. Se basan en trabajos anteriores [23, 44, 50] cambiando la vista de difusión de información de nivel posterior a nivel de autor. En Xiong et al. [51] proponen otra variación de un modelo de contagio, el modelo susceptible-contactado-infectado-refractario (SCIR). Probaron el modelo mediante simulaciones numéricas. En [68] Qiu et al. incorporar un mecanismo de olvido y refutación en el modelo SIR para describir la propagación de rumores con mayor precisión. Probaron su modelo tanto en simulaciones numéricas como en Renren OSN.

Otro enfoque se presenta en [47] donde presentan un modelo que incorpora el efecto de fuentes externas de influencia en el proceso de infección, que modelan con funciones de riesgo, complementando la descripción de difusión de información. Prueban este modelo en datos sintéticos y en Twitter. En sus experimentos concluyeron que alrededor del 70 % de la difusión en Twitter se puede atribuir a un efecto de red y el resto (30 %) se debe a fuentes externas como noticias online, Facebook, etc.

2.2.3. Modelos de Grafos

Principalmente 2 modelos entran en esta categoría, a saber, cascadas independientes (IC) [13] y umbral lineal (LT) [4]. Como se describe en [52], ambos asumen la existencia de una estructura gráfica estática que subyace a la difusión y se enfocan en la estructura del proceso. IC asocia una probabilidad a cada borde que representa la posibilidad de que la información se difunda. LT define un umbral para cada usuario (nodo) y un grado de influencia para cada borde. La información se difunde, o un nodo se activa en LT si la suma de influencia de los vecinos activos de un usuario supera su umbral. No profundizaremos en estos modelos porque nuestro objetivo es modelar la difusión en foros web donde no existe un grafo explícito.

2.2.4. Otros

En esta categoría clasificamos todos los modelos propuestos que no encajan en las categorías antes mencionadas. Por ejemplo, el modelo de votante. En [21] utilizan este modelo para representar la difusión de opiniones en una red social. Sin embargo, su objetivo es resolver el problema del conjunto de maximización de la dispersión que difiere de nuestro objetivo. Otro modelo que entra en esta categoría es el que utiliza procesos multidimensionales de Hawkes, que son una clase de modelos de procesos puntuales autoexcitantes o mutuamente excitantes. Capturan la influencia subyacente del usuario con este modelo y validan el modelo probándolo en conjuntos de datos sintéticos y de Twitter.

Hu et al. probaron un enfoque diferente [70] en el que implementan un modelo hidrodinámico no paramétrico adaptado a la difusión de información (Hydro-IDP) mediante la correlación de las características de la evolución del flujo de densidad de fluidos en el espacio-tiempo físico con la de la difu-

sión de información en el espacio-tiempo cibernético. Para ello, plantean la analogía entre la energía de la fuente inicial, el radio de la fuente inicial, la velocidad del flujo inicial y la popularidad de la información, la influencia del editor, la difusividad de la plataforma social, respectivamente, definiendo simultáneamente nuevas características de interés para ser estudiadas en una red social. Prueban su modelo con datos de OSN Sina-weibo de China. Lee et al. [42] usa modelos de interacción espacial típicamente usados en el campo de la economía y la geografía económica para estudiar la relación entre la distancia y el conocimiento entre estudiantes universitarios usando datos de StudiVZ.

2.2.5. Clasificación de trabajos anteriores

Como podemos notar en la Tabla 2.1, la mayoría de las investigaciones previas revisadas se centran en OSN como Facebook y Twitter, las cuales poseen la particularidad de que un gráfico de red social explícito se puede extraer fácilmente mediante el uso de amistad o seguidor. relaciones respectivamente. A su vez, esto hace que los modelos propuestos en estos OSN dependan en gran medida de esta información y no está claro si estos modelos propuestos se pueden aplicar a los OSN en los que falta esta información.

Como se mencionó anteriormente, nuestro objetivo es estudiar los foros web, por lo que ponemos énfasis en los trabajos que se aplican a ellos, como [23, 44, 50, 69]

2.3. Predicción de arcos

El problema de predicción de enlaces consiste en poder predecir las relaciones en una red. La gran mayoría de las investigaciones realizadas con respecto a

Tabla 2.1: Clasificación de trabajos anteriores

Reference	**OSN**				**Model**			**Level**		
	Facebook or Twitter-like	Simulated network	Web-Forum-like	Theoric (no data)	Deterministic	Probabilistic	None	Macroscopic	Mesoscopic	Microscopic
C. Jiang et al (2014) [57]	✗				✗			✗		
N. Alon et al (2010) [32]				✗	✗					✗
Y. Sun et al (2014) [59]		✗				✗		✗		
A. Saxena et al (2015) [62]	✗					✗			✗	
S. Myers et al (2012) [47]	✗					✗			✗	
Y. Hu et al (2017) [70]	✗				✗				✗	
L. Small and O. Mason (2013) [55]				✗	✗					✗
F. Xiong et al (2012) [51]		✗			✗			✗		
J. Woo and H. Chen (2012) [50]			✗		✗				✗	
J. Woo and H. Chen (2016) [69]			✗		✗				✗	
J. Woo et al (2011) [44]			✗		✗				✗	
X. Qiu et al (2016) [68]	✗				✗			✗		
M. Kubo et al (2007) [23]			✗		✗			✗		
Proposed Model			✗			✗				✗

este problema utilizan la estructura de la red local para obtener la probabilidad de que dos nodos de la red formen un enlace. El problema acepta dos definiciones clásicas:

uno está relacionado con la evolución de la red en el que la pregunta que buscamos responder es si el estado actual y la topología de la red se pueden usar para predecir el estado y la topología futuros, es decir, se pueden predecir los enlaces futuros. La otra definición se refiere a una situación en la que falta

información sobre la red, es decir, algunos de los enlaces, y la pregunta que debemos responder es si es posible inferir los enlaces faltantes utilizando la información que tenemos disponible. En esta investigación trabajamos con la segunda definición.

Se realizó una encuesta extensa en [39] donde examinaron varios enfoques del problema, como modelos de características, modelos gráficos bayesianos y el enfoque algebraico lineal, comparando la complejidad del modelo, el rendimiento de la predicción, entre otros. En [73] se aborda el problema de predicción de enlaces como un problema de optimización con restricciones de cardinalidad y se proponen frameworks ITERCLIPS para obtener soluciones. En [65] abordan el problema de la evaluación para estandarizar las métricas utilizadas y hacer que los resultados sean comparables entre investigaciones. Los trabajos que pudimos encontrar que más se asemejan a nuestro enfoque son los siguientes: En [72] proponen un nuevo modelo que utiliza un conjunto de metarutas disponibles para estimar la probabilidad de enlace. Prueban este modelo en una red de bibliografía donde se pueden obtener metadatos. Los resultados obtenidos por este modelo parecen muy prometedores. En [67] utilizan la difusión de contenido de texto, como si estuvieran expuestos a una publicación enviada originalmente por un usuario no relacionado, para mejorar la predicción de enlaces. Finalmente, en [24] se propone una modificación de LDA donde se incluyen contenidos de texto, principalmente palabras clave de investigación, para mejorar la predicción de futuras colaboraciones entre autores. No hemos encontrado un trabajo que se asemeje más a nuestro enfoque del problema.

Capítulo 3

Caso de estudio y preparación de datos

En este capítulo se comienza por describir el caso de estudio utilizado para el desarrollo del trabajo experimental de esta Tesis. Posteriormente, se explicita el proceso de selección y preprocesamiento de los datos cuyo objetivo es extraer el contenido semántico de los posteos y transformarlo de modo que sea utilizable por el modelo reportado en el Capítulo 4. Estos datos procesados son los empleados para realizar los experimentos computacionales reportados en el Capítulo 5.

3.1. Foro Web Plexilandia

Plexilandia es una OSN formada por un grupo de personas que se han unido en torno a la construcción de efectos musicales, amplificadores y equipos de audio (estilo "Hazlo tú mismo"). Al principio nació como una red social para

compartir experiencias comunes en la construcción de plexies[1]. Hoy, plexilandia cuenta con más de 2500 miembros en más de 15 años de existencia. Todos estos años han estado compartiendo y discutiendo sus conocimientos sobre construcción. sus propios plexos y efectos, además de otros temas relacionados como luthier, audio profesional y compra/venta de piezas.

Si bien cuentan con una página web de información básica de redes sociales, la mayoría de las interacciones de sus miembros se producen en el foro de discusión.

Durante sus nueve años de vida, esta OSN ha experimentado un gran crecimiento sostenido en miembros y sus aportes. En Table 3.1 podemos ver el número de publicaciones para cada uno de estos 9 años incluyendo el número total de publicaciones. Debemos tener en cuenta que para este trabajo solo usamos los datos del año 2013 y 2014, y los Sub-Foros 2-6 como se puede ver más adelante en la Fig.5.1.

Tabla 3.1: Actividad en Plexilandia

Forum	***2006***	***2007***	***2008***	***2009***	***2010***	***2011***	***2012***	***2013***	***2014***	***TOTAL***
Aplifiers (2)[2]	392	2165	2884	3940	3444	3361	2398	1252	985	20822
Effects (3)	184	1432	3362	3718	4268	5995	4738	2317	1331	27345
Luthier (4)	34	388	849	1373	1340	2140	926	699	633	8382
General (5)	76	403	855	1200	2880	5472	3737	1655	1295	17573
Pro Audio (6)	—	—	—	—	—	342	624	396	219	1579
Synthesizers (7)	—	—	—	—	—	—	—	104	92	196
TOTAL	**686**	**4388**	**7950**	**10231**	**11932**	**17310**	**12423**	**6423**	**4555**	**75898**

Además, los administradores de redes sociales nos proporcionaron una lista de 64 miembros clave. Como se indica en [71], hay muchas definiciones de lo que es un miembro clave, p.e. los usuarios que más participan, los que responden a las preguntas de los demás, etc. pero está claro que juegan un

[1] "Plexi.[es] el apodo dado a los amplificadores Marshall modelo 1959 que tienen el panel de control de perspex transparente (también conocido como plexiglás). con una hoja de respaldo dorada que se ve a través de las placas de metal de los modelos posteriores.

papel primordial para mantener viva la red. Se establecieron tres grupos de importancia de la siguiente manera:

- Expertos Tipo A: que son los key-members más importantes. Hay 34 miembros para el año 2013 según el criterio de los administradores.
- Expertos Tipo B: que son los más importantes en menor medida que los miembros clave tipo A, sin embargo, también son miembros clave. Estos son 21 para el mismo período.
- Expertos Tipo C: finalmente, los key-members tipo C son aquellos que son key-members históricos, ya que han estado involucrados en la red social desde sus orígenes, pero no están participando continuamente. En esta clase hay alrededor de 11 miembros para el año 2013.
- Expertos Tipo X: esta clase son todos los miembros de la red social que no son miembros clave. No pertenecen al núcleo de la red social y normalmente hacen preguntas en lugar de publicar respuestas o tutoriales.

Es importante remarcar que esta información fue entregada durante 2013 y 2014 [71] por lo que se minimizó la probabilidad de que se olvidaran de miembros clave. Esta constituye la razón principal por la que decidimos utilizar solo los datos de estos años para nuestros experimentos. Usamos esta clasificación de usuarios como una segmentación de la población con respecto al comportamiento como se indica en la sección 4.2.2.

3.2. Preparación de datos

3.2.1. Selección y preprocesamiento de datos

El primer problema que debemos afrontar es el preprocesamiento de los datos del foro para poder utilizarlos para el modelo.

Hay varias formas de obtener datos de las redes sociales. En general, dependerá del problema que estemos abordando para determinar el método adecuado para este proceso. Culotta [20] presenta una metodología para obtener datos de una red social a partir de correos electrónicos y Ríos et al. [64] presenta los pasos básicos para la extracción de datos de un VCoP, que suele basar su funcionamiento, en sistemas de foros como VBuletin[3], MyBB[4] y phpBB[5], entre otros. Un foro en el contexto de la web, se refiere a un lugar virtual en el espacio donde los miembros interactúan, discuten ideas, comparten y generan conocimiento. Por lo general, los temas dentro de un foro se ordenan de forma jerárquica, con diferentes categorías según el interés de los miembros que lo frecuentan. En el caso de VCoP, las categorías están directamente relacionadas con el propósito de la comunidad. Cada conversación del foro, dentro de las categorías, se llama Thread o tema de discusión y en ellas los integrantes exponen sus opiniones y discuten en torno a una idea central. Cada mensaje entre miembros realizado dentro de un hilo se llama publicación, que es la unidad básica dentro de un foro. Un tema de discusión comienza con una publicación de un usuario de la comunidad, que generalmente contiene una pregunta o la presentación de una idea que desea discutir. Luego, los diferentes integrantes de un VCoP, relacionados con el usuario o el tema de discusión, realizan sus publicaciones para ser debatidas y así generar conocimiento sobre el tema central de la comunidad.

[3]https://www.vbulletin.com [Fecha de acceso 10 de abril de 2018]
[4]http://www.mybb.com [Fecha de acceso 10 de abril de 2018]
[5]http://www.phpbb.com [Fecha de acceso 10 de abril de 2018]

Cada publicación está compuesta por algunos elementos básicos como el identificador de usuario (ID), que permite saber qué miembros de la comunidad están interactuando en la discusión; el contenido del post, que dependiendo del foro puede ser texto, imágenes, enlaces a otras páginas, vídeos, etc.; y la información del sistema de foros, como la fecha de creación del post, el Thread y la categoría a la que pertenece, entre otros. De acuerdo con la metodología expuesta por Ríos et al. [64], los elementos básicos que acabamos de mencionar y el contenido se seleccionarán como datos solo en formato de texto disponible en el foro. Para la limpieza de los datos, Ríos et al. [64] indica que el primer filtro debe hacerse con respecto a las respuestas (citas) de otras publicaciones. En un foro, un usuario puede responder a otra publicación creando un nuevo mensaje con la copia de la publicación citada más el texto adicional expuesto por este nuevo usuario. Por lo tanto, es necesario eliminar la parte replicada de la nueva publicación y solo almacenar la nueva entrada de texto.

Luego deberás transformar las siglas o abreviaturas, eliminar las faltas de ortografía y todos los elementos de los posts que hagan que estos no sean comparables. Este proceso se lleva a cabo por dos métodos: El primero es un proceso denominado stemming que consiste en la transformación de cada palabra a su raíz, la modificación de las palabras del plural al singular y el cambio de todas las frases o palabras que no representan un aporte. a los datos del texto, asignándole un término representativo como ïnutilizable.º "misceláneo". El segundo método llamado filtro de palabras vacías donde se eliminan todas las palabras que no aportan información a la red, como artículos, pronombres y palabras ïnutilizables". El objetivo de los métodos mencionados es, por un lado, hacer comparables los posts para evaluar si ambos hablan de lo mismo y, por otro lado, reducir el número de palabras utilizadas para realizar la comparación. Sin embargo, la cantidad de palabras resultantes podría ser demasiado grande para relacionar las publicaciones solo en función de las palabras útiles. Para reducir el contenido del vocabulario

utilizado en la red y poder comparar adecuadamente cada publicación, se utilizarán los métodos de procesamiento de texto.

3.2.1.1. Procesamiento de texto basado en minería de texto

Luego de la selección, limpieza y transformación de los datos, es necesario reducir la gran cantidad de palabras existentes en el foro para realizar una comparación exitosa del contenido de la red. Sin embargo, no se puede realizar ninguna estrategia de reducción de contenido, debes utilizar métodos que te permitan extraer las ideas centrales de cada post. Una forma de hacerlo es definir una serie de conceptos, temas o categorías y clasificar cada mensaje según su proximidad a cada una de las categorías. Al realizar correctamente este proceso, los mensajes podrían estar descritos por un número reducido de conceptos, lo que sería muy útil para hacer comparaciones entre ellos. El problema es que, las distintas categorías que se formen dependerán de la estructura del foro, del objetivo que persiga y de los temas que se traten en él. Para solucionar lo anterior, se propone el uso de modelos temáticos o conceptuales como estrategia de reducción del contenido de la red.

Los modelos de temas y conceptos permiten describir el contenido temático de los documentos sin clasificación previa [25], sin embargo, es necesario representar los datos obtenidos en la etapa anterior de forma que permita su aplicación. A continuación, se presentará la notación y representación de los datos para implementar los modelos de tópicos, seguida de una estrategia de reducción de contenido y clasificación de publicaciones.

3.2.1.2. Representación de los datos

Cada uno de los mensajes obtenidos en la sección 3.2.1 contiene una serie de palabras que caracterizan el post. Si se toman todas las diferentes palabras

de todos los mensajes, se obtendrá el vocabulario completo utilizado por la red.

Sea $\mathcal{V}$ el vector de tamaño $|\mathcal{V}|$ en el que cada fila representa una palabra diferente utilizada en la red. Sea w_i la palabra en el lugar i del vector $\mathcal{V}$. Es posible representar post p_j como una secuencia de un conjunto de S_j palabras de $\mathcal{V}$, con $S_j = |p_j|$, donde $j \in \{1, \ldots, |\mathcal{P}|\}$ y $\mathcal{P}$ corresponde al conjunto de publicaciones en la red. Un "corpus"se define como una colección de publicaciones $\mathcal{C} = \{p_1, \ldots, p_N\}$. En términos de composición podemos definir la matriz $\mathcal{W}$ de tamaño $|\mathcal{P}| \times |\mathcal{V}|$ donde cada elemento de la matriz se define como:

$$w_{i,j} = \text{ number of times } w_i \text{ appears in } p_j \tag{3.1}$$

Entonces $\sum_{i=1}^{|\mathcal{V}|} w_{i,j} = S_j$. Asimismo, podemos definir $\sum_{j=1}^{|\mathcal{P}|} w_{i,j} = T_i$ que representa el número total de apariciones que tiene el término w_i en un corpus.

Un corpus puede ser representado por el producto entre la frecuencia de un término en el corpus y el logaritmo del recíproco de la frecuencia del documento que contiene la palabra (TF-IDF) [2]. Podemos definir la matriz tf-idf que representa el corpus como $\mathcal{M}$ de tamaño $|\mathcal{P}| \times |\mathcal{V}|$ donde cada $m_{i,j}$ se determina como

$$m_{i,j} = \frac{w_{i,j}}{T_i} \times \log\left[\frac{|\mathcal{P}|}{1 + n_i}\right] \tag{3.2}$$

Donde n_i es el número de publicaciones que pertenecen al corpus en el que aparece la palabra w_i. El término IDF presentado en 3.2 corresponde a una corrección habitual con respecto al término IDF original $\log\left[\frac{|\mathcal{P}|}{n_i}\right]$ porque, después de la limpieza y selección de datos, algunas publicaciones pueden no contener palabras, lo que hace que este término quede indefinido.

3.2.1.3. Latent Dirichlet Allocation

A continuación, se presentará el proceso de reducción de contenido mediante el uso de un modelo de tópicos, basado en el trabajo realizado por Ríos et al. [64], Álvarez [33] y L'Hullier [35]. Un modelo de tópicos puede ser considerado como un modelo probabilístico que relaciona documentos y palabras a través de variables que representan los principales tópicos inferidos del propio texto. En este contexto, un documento puede ser considerado como una mezcla de temas, representados por distribuciones de probabilidad que pueden generar las palabras en un documento dados estos temas. El proceso de inferencia de las variables latentes, o temas, es el componente clave de este modelo, cuyo objetivo principal es aprender de los datos de texto la distribución de los temas subyacentes en un corpus dado de documentos de texto.

La asignación de Dirichlet latente (LDA) [17, 28] es un modelo bayesiano donde los temas latentes en los documentos se infieren a través de la estimación de distribuciones sobre un conjunto de datos de entrenamiento. El propósito es que cada tema sea modelado como una distribución de probabilidad sobre un conjunto de palabras representadas por el vocabulario ($w \in \mathcal{V}$), y cada documento como una distribución de probabilidad sobre un conjunto de temas ($\mathcal{T}$). El muestreo de estas distribuciones se realiza con distribuciones multinomiales de Dirichlet.

El proceso se realiza de forma automatizada y solo es necesario etiquetar cada tema descubierto por el algoritmo con la ayuda de expertos de la comunidad.

Usando las definiciones de la sección 3.2.1.2 y considerando que un mensaje contenido en una publicación puede representarse como una secuencia de S palabras definidas como $\mathbf{w} = (w^1, ..., w^S)$, donde w^s representa la palabra s^{th} en la publicación, podemos describir el proceso generativo para el modelo LDA ideado por Blei [17].

Para cada publicación del corpus:

1. Elija un número S de multinomios ($S \sim Poisson(\xi)$) que representará la cantidad de palabras en la publicación.
2. Elija $\theta \sim Dir(\alpha)$.
3. Para cada $w^s \in (w)$:
   a) Elija un tema $z_s \sim Multinomial(\theta)$.
   b) Elija una palabra w^s de $p(w^s|z_s, \beta)$, que es una probabilidad condicional multinomial sobre el tema z_s.

Para LDA, dados los parámetros de suavizado β y α, y una distribución conjunta de una mezcla de temas θ, la idea es determinar la distribución de probabilidad a generar a partir de un conjunto de temas $\mathcal{T}$, un mensaje compuesto por un conjunto de S palabras $\mathbf{w}$,

$$p(\theta, z, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{s=1}^{S} p(z_s|\theta) p(w^s|z_s, \beta) \tag{3.3}$$

Donde $p(z_s|\theta)$ puede ser representado por la variable aleatoria θ_i, tal que el tema z_s se presenta en el documento i ($z_s^i = 1$). Se puede deducir una expresión final integrando (3.3) sobre la variable aleatoria θ y sumando sobre los temas $z \in \mathcal{T}$. Dado esto, la distribución marginal de un mensaje se puede definir como se muestra en (3.4):

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{s=1}^{S} \sum_{z_s \in \mathcal{T}} p(z_s|\theta) p(w^s|z_s, \beta) \right) d\theta \tag{3.4}$$

El objetivo final de LDA es estimar las distribuciones descritas anteriormente para construir un modelo generativo para un corpus de mensajes dado. Hay

una gran cantidad de métodos desarrollados para realizar la inferencia sobre esta distribución de probabilidad, como la maximización de la expectativa variacional [17], o una aproximación variacional discreta de la ecuación 3.4 empíricamente por Xing [27] ya través de un muestreo de Gibbs (modelo Monte Carlo basado en cadenas de Markov) [19] que Phang y Nguyen han implementado y aplicado de manera eficiente [29].

Según la metodología de Ríos [64], el siguiente paso para la formación de la matriz [Temas Publicaciones]. Esta matriz entregará una reducción en el contenido de las publicaciones. Usaremos el método desarrollado por Phang y Nguyen [29] que requiere una serie de parámetros de entrada, como número de iteraciones, hiperparámetros α, β, número de temas requeridos $\mathcal{T}$, cantidad de palabras n por tema que se deben guardar, publicaciones del corpus, entre otros. El método devolverá como salida las distribuciones de las palabras sobre los temas, la distribución de estos sobre el documento, k temas con las n palabras más importantes que representan el tema y sus correspondientes probabilidades de pertenecer a él, entre otros cosas. En particular, con los k temas y sus n palabras o términos representativos se pueden formar vectores de tamaño $|\mathcal{V}|$ completando con ceros en $|\mathcal{V}| - n$ palabras no representativas. Con estos vectores se formará la matriz semántica (SM) [Topics$\times$Terms], luego esta matriz se operará con la matriz transpuesta tf-idf $\mathcal{M}^t$ definida por la ecuación (**??**) que representa la participación de cada término en cada publicación, obteniendo así la matriz [Temas$\times$Publicaciones]. Esta matriz permite conocer la composición de un post en función de los diferentes temas encontrados y nos permite obtener todas las representaciones vectoriales de texto del post (ρ_p) donde p es un post del conjunto de Posts.

3.2.2. ETL para datos de entrada del modelo

En los foros, el gráfico de la red social no está definido explícitamente como en otras comunidades (Facebook, Twitter, etc.). Por lo tanto, primero debemos definir una topología de red para usar. Con esto en mente, la representación de red más habitual (y más directa) utilizada es aquella en la que cada nodo representa un usuario de la red y se agrega un enlace entre nodos para representar una relación o interacción entre los usuarios que representan. Sin embargo, hay muchas formas factibles de representar la red de esa manera. Algunas de estas formas de definir los enlaces en la red en foros web se muestran en la Fig. 3.1

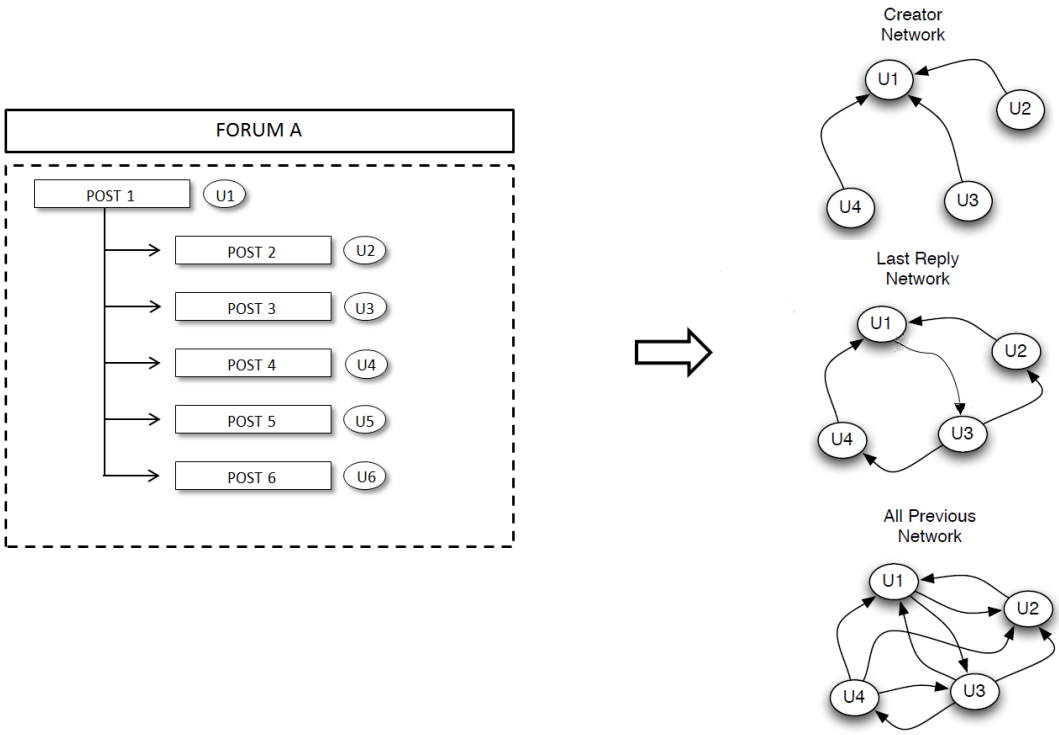

Figura 3.1: Posibles representaciones de red para foros web.

Sin embargo, de acuerdo con nuestro objetivo de hacer un modelo centrado en el contenido, proponemos una nueva topología de red que pone énfasis en el contenido y en cómo los usuarios interactúan con él. Esta topología, como se puede apreciar en la Fig. 3.2, consiste en distinguir entre cuatro tipos de nodos, a saber, nodo Foro, nodos Sub-Foro, nodos Subproceso y nodos Usuario. Estos nodos siguen una jerarquía en la que un tipo de nodo solo puede formar un enlace con un nodo de un tipo perteneciente a una capa directamente encima de ellos. Nos centraremos principalmente en los enlaces formados entre los nodos de usuario y los nodos de subproceso. Tenga en cuenta que los usuarios ahora interactúan (forman enlaces) directamente con las conversaciones (Subprocesos) que captan su interés, lo que representa con precisión lo que sucede en los foros web.

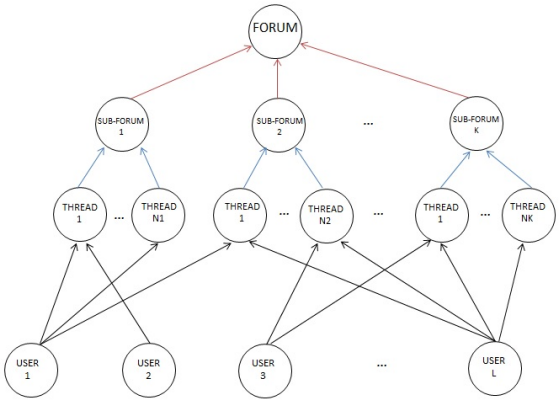

Figura 3.2: Topología propuesta para foros web.

Además, siempre podemos derivar la representación de la red donde los usuarios interactúan entre sí como si hubiéramos elegido la representación de la red orientada al creador. Esto se puede ver fácilmente en la Fig. 3.3 donde mostramos la equivalencia (en términos de enlaces formados) entre la topología habitual y la topología que propusimos. Además, si se eligiera cualquiera de las formas habituales de representar la red, si el número de usuarios activos

en la red es n, entonces el número de arcos posibles estaría dado por

$$\text{number of possible arcs in the network} = n(n-1) \tag{3.5}$$

Podemos reducir el número de posibilidades adoptando nuestra topología de red propuesta. Al representar las redes con nuestra topología propuesta, tenemos que si la cantidad de usuarios activos en la red es n y la cantidad de subprocesos activos es m, entonces la cantidad de arcos posibles estaría dada por

$$\text{number of possible arcs in the network} = nm \tag{3.6}$$

donde normalmente $m \ll n$ (si no es así, siempre es posible acortar la ventana de tiempo considerada como un período y así hacer que la desigualdad anterior sea verdadera). Esto es de particular importancia cuando se hace que el modelo elija el enlace correcto a formar.

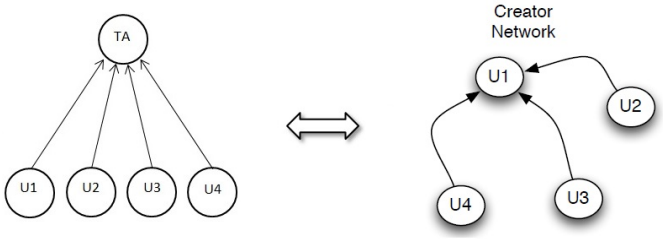

Figura 3.3: Representación de red heterogénea equivalente.

Como comentario final, el uso de la topología que proponemos ayuda a incorporar aspectos adicionales de la red, como se muestra en la Fig. 3.2, en particular con respecto a la estructura del Foro. En este trabajo, extrajimos la información directamente de la estructura del foro para las primeras tres

capas, es decir, solo intentaremos predecir enlaces entre los nodos de usuario y los nodos de hilo.

Después de establecer la representación de la red que se usará en este trabajo, tuvimos que enfrentar el problema de usar el contenido de texto generado por las publicaciones de los usuarios para extraer pistas sobre qué arcos tienen más probabilidades de existir, es decir, qué conversaciones tienen más probabilidades de ser atractivas. a qué usuarios.

Teniendo la representación vectorial de texto de cada una de las publicaciones disponibles en los datos dentro del marco de tiempo establecido como resultado de la etapa de preprocesamiento, nos enfrentamos al problema de obtener la representación vectorial de texto de cada usuario y la representación vectorial de texto de hilo. Existen múltiples alternativas para resolver este problema, pero para este trabajo decidimos utilizar el siguiente enfoque propuesto principalmente debido a que supera a otros enfoques.

Primero, subdividimos las publicaciones del Foro en diferentes grupos según el Sub-Foro al que pertenecen. A continuación, subdividimos el período de tiempo en períodos y, posteriormente, subdividimos las publicaciones dentro de cada subforo en diferentes grupos según el período de tiempo al que pertenecen. Luego, extraemos los usuarios activos y los hilos del período considerando un usuario como activo si hace una publicación durante el período y un hilo como activo si un usuario publicó en el hilo durante el período. Después de eso, calculamos la representación vectorial de texto de un hilo durante un período, ν_T^P, como la media de todas las representaciones vectoriales de texto de la publicación, ρ_p, de las publicaciones que pertenecen a ambos el hilo, T, y el punto, P, es decir

$$\nu_T^P = \frac{1}{|T \cap P|} \sum_{p \in T \cap P} \rho_p \tag{3.7}$$

donde p es una publicación. Por otro lado, para calcular la representación vectorial de texto de un usuario, primero debemos hacer una subdivisión más. Subdividimos las publicaciones realizadas por un usuario durante el período en diferentes grupos según el hilo en el que se publicaron. Es importante notar que un usuario tendrá tantas representaciones de vectores de texto para un período como subgrupos de sus publicaciones durante el período. período mencionado. Ahora podemos calcular la representación vectorial de texto de un usuario para un período y un subgrupo de publicaciones, μ_S^P, como la media de todas las representaciones vectoriales de texto de la publicación, ρ_p, de las publicaciones que pertenecen tanto al subgrupo de puestos, S, como al período, P, es decir

$$\mu_{U,S}^P = \frac{1}{|U \cap S \cap P|} \sum_{p \in U \cap S \cap P} \rho_p \tag{3.8}$$

donde p es una publicación.

Ahora que tenemos ordenadas las representaciones del vector de texto tanto para los usuarios como para los hilos, debemos aplicar el proceso que se muestra en la Fig. 3.4 para poder obtener algo útil para el modelo.

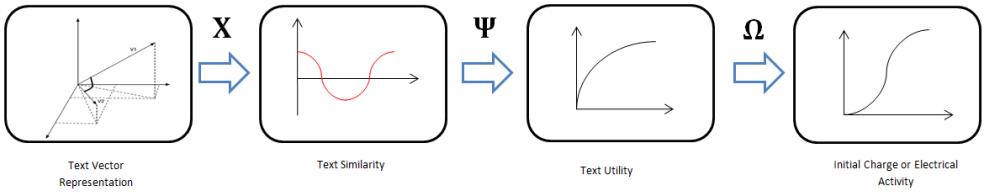

Figura 3.4: Transformaciones aplicadas para obtener la entrada del modelo.

Primero, necesitamos definir una función χ que nos dé una medida de qué tan relacionadas o distantes están dos representaciones vectoriales de texto entre sí. Para calcular esto, utilizamos la medida de similitud del coseno que nos da el coseno del ángulo formado entre dos representaciones vectoriales de texto

en un espacio de k-tema. Así, si tenemos dos representaciones vectoriales de texto μ_1 y μ_2 la similitud entre ellas viene dada por

$$\chi(\mu_1, \mu_2) = similarity(\mu_1, \mu_2) = cos(\theta) = \frac{\mu_1 \cdot \mu_2}{|\mu_1||\mu_2|} \tag{3.9}$$

donde θ es el ángulo entre μ_1 y μ_2. Ahora nos enfrentamos a un problema diferente que es definir una función Ψ para obtener la similitud de la representación del vector de texto con la utilidad del usuario. La respuesta a esta pregunta no es sencilla, principalmente debido a que no está clara la forma exacta en que un usuario reacciona ante textos similares. Este problema no ha sido considerado en investigaciones previas al menos al formular modelos de difusión. Proponemos, como una forma de salvar este obstáculo, realizar una transformación a través de una función que permita extraer de mejor manera las características o diferencias entre alternativas teniendo en cuenta que esta utilidad servirá como entrada para un modelo logit. Una pregunta interesante para responder es cómo podemos determinar la forma funcional óptima para transformar la similitud del texto en utilidad para el usuario, pero esta pregunta va más allá del alcance de esta investigación. Entonces, una vez planteado el problema, podemos conceptualizarlo como la determinación de una función Ψ_1 que toma la similitud del texto, x, y proporciona utilidad al usuario como salida. Proponemos la siguiente forma funcional

$$\Psi_1(x) = \frac{1}{1-x} \tag{3.10}$$

porque maximiza el impacto generado por las diferencias en la entrada a las probabilidades logit. Sin embargo, teniendo en cuenta que la función anterior puede tener un rango de valores muy amplio, podemos intentar controlar este rango y restringirlo al intervalo $[0, a]$ introduciendo la siguiente transformación a la utilidad por definiendo la función Ψ_2 tal que

$$\Psi_2(u) = a\frac{u}{w} \tag{3.11}$$

donde $w = \text{máx}_{j \in Threads}\, u_j$ y a es un parámetro que determina la longitud del intervalo en el que variarán los valores de la utilidad transformada. Esto hace que a sea un parámetro fundamental en términos de importancia porque dependiendo del valor elegido puede restringir la dispersión de las probabilidades logit o por el contrario permitir mucha variabilidad. Teniendo en cuenta esta información, podemos considerar a como una medida de estocasticidad permitida en las probabilidades iniciales

Para el modelo, a representa la sensibilidad de los usuarios de la red sobre cuán similar es el tema de una conversación al contenido de texto generado por el usuario en términos de satisfacción extraída de esa conversación.

Antes de pasar al siguiente paso, optamos por reducir la cantidad de alternativas que un usuario considera durante su proceso de toma de decisiones. Para lograr esto, para cada elección que hace el usuario, clasificamos las alternativas (subprocesos activos) de acuerdo con su utilidad reportada (V_j) para el usuario de forma decreciente y decidimos mantener solo las mejores k alternativas. Esta reducción de alternativas que considera un usuario se basa en investigaciones sobre la memoria de trabajo y la capacidad de atención [1].

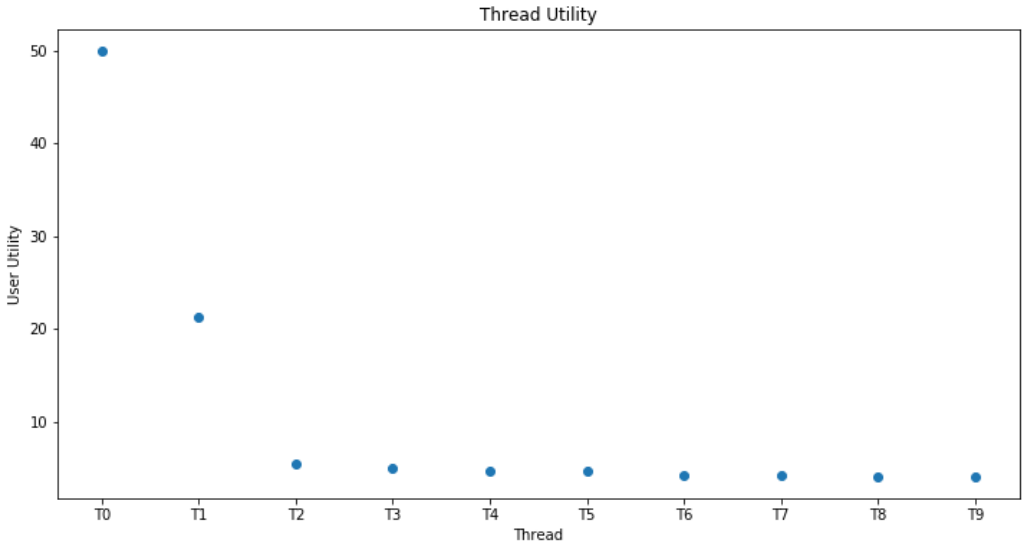

Figura 3.5: Ejemplo 1 de utilidad de un hilo

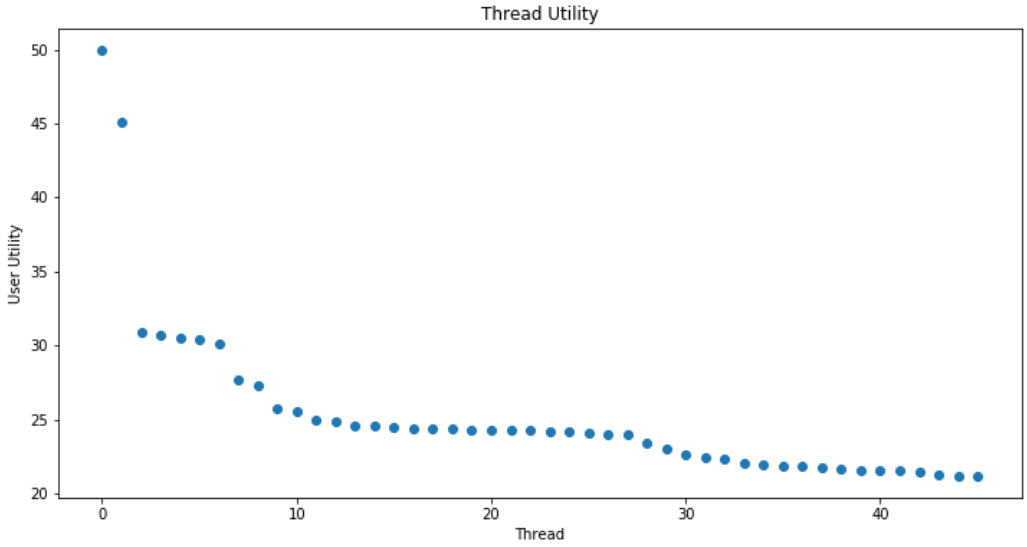

Figura 3.6: Ejemplo 2 de utilidad de un hilo

En las Fig. 3.5 y 3.6 podemos ver 2 ejemplos de la utilidad que encuentra un usuario en todos los hilos que están activos en ese momento. Podemos

notar que solo unos pocos subprocesos son de gran interés para el usuario y la mayoría de los subprocesos se apilan en la cola de la figura, lo que agrega ruido a la decisión. Por lo tanto, mantener solo las alternativas k con mayor utilidad ayuda a reducir el ruido que agregamos al modelo.

Como paso final, necesitamos definir una función Ω tal que obtengamos algo parecido a la actividad eléctrica inicial de la región neuronal asociada con una determinada decisión. Para ello decidimos hacer uso de la teoría de la utilidad aleatoria en el sentido de que I_i es proporcional a la probabilidad de elegir la alternativa i como podemos apreciar en

$$\Omega(\mathbf{V}, i) = I_i = \frac{e^{V_i}}{\sum_{j \in Threads} e^{V_j}} \tag{3.12}$$

Capítulo 4

Técnicas y metodología para la implementación del modelo ELCA

En este capítulo se presentan los métodos computacionales de los que nos hemos servido para el análisis de una Red Social en Línea (OSN, por sus siglas en inglés), incorporando la explicación de los aspectos metodológicos más relevantes. Primero, se muestra una descripción general del objetivo del modelado de la red social, para posteriormente explicar con detalle cada paso y método computacional que se considera. Introducimos los conceptos básicos del analisis semántico de documentos de texto y las dos aproximaciones que hemos probado, una basada en lógica difusa y la otra en modelos gráficos probabilísticos. A continuación presentamos las técnicas seguidas para generar la red social a partir de las comunicaciones entre los miembros, seguido del proceso de filtrado de la red. Por ultimo referimos la construcción del grafo de la red social guiado por la información semántica siguiendo tres aproximaciones distintas.

4.1. Proceso computacional

Como se mencionó anteriormente en la Sección 1.3, la pregunta principal de este trabajo es cómo modelar la toma de decisiones de los agentes en el proceso de generación de contenido en los foros web, con un enfoque basado en el contenido semantico de los posteos.

Sin embargo, para lograr eso, primero debemos abordar los procesos que se ilustran en la Fig. 4.1.

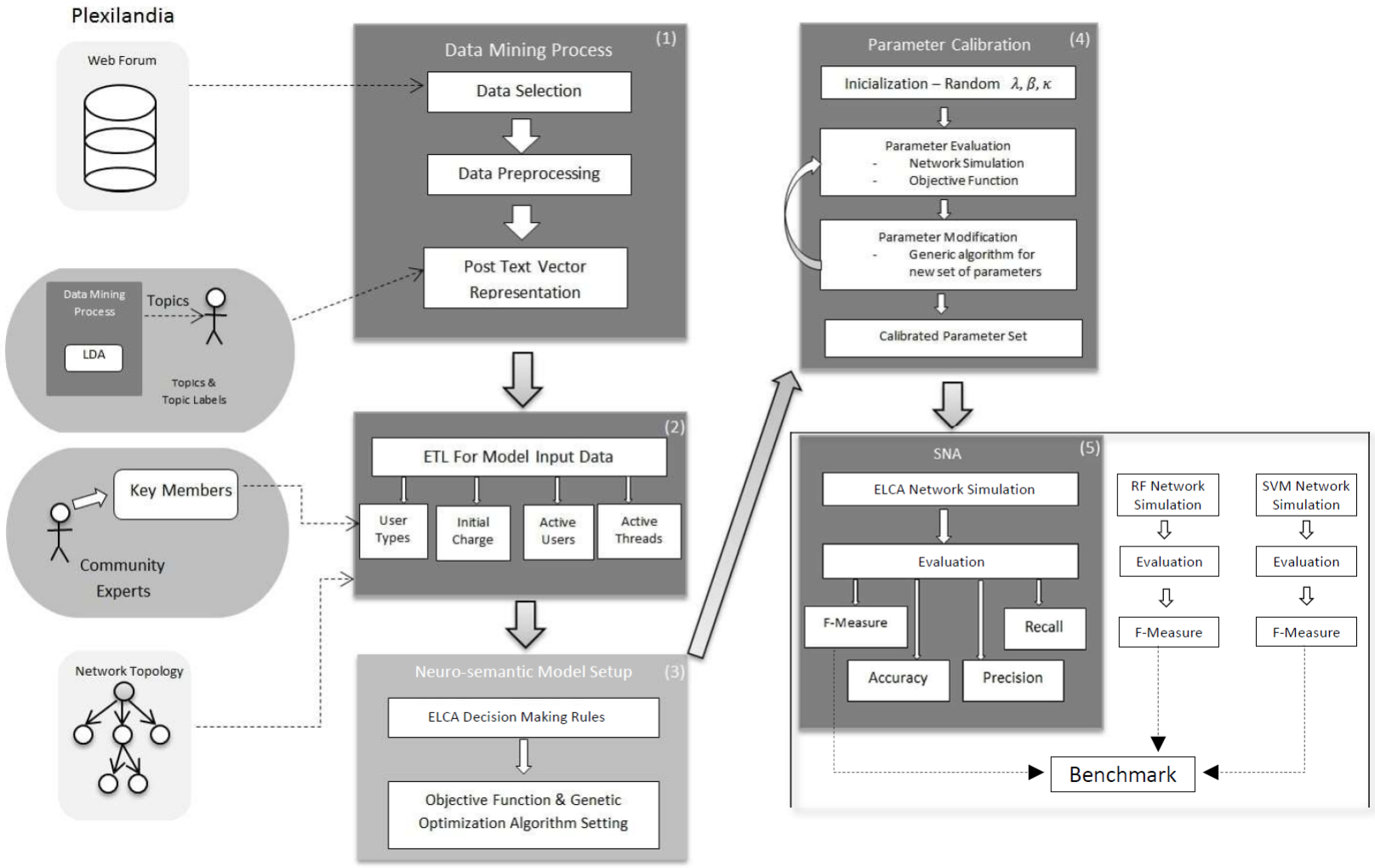

Figura 4.1: Proceso computacional del estudio

4.2. Extended Leaky Competing Accumulator (ELCA)

4.2.1. Leaky Competing Accumulator

Esta propuesta se basa en modelos Psicológicos de toma de decisiones construidos a partir del trabajo realizado en [48], y consideraremos que los usuarios de OSN corresponden a agentes cognitivos en el sentido explicado en esta sección. La psicología considera los procesos cognitivos (a partir de la operación mental) y perceptuales (a partir de los dispositivos sensoriales), así como las acciones motrices. La recuperación/almacenamiento de la memoria, la producción/interpretación del lenguaje, la atención, la resolución de problemas y la toma de decisiones son ejemplos de algunos procesos cognitivos. El modelo LCA fue presentado por Usher y McClelland [15] en 2001 como un modelo de difusión para la toma de decisiones. Su trabajo considera una unificación teórica de conceptos a partir de los procesos cognitivo-perceptuales y la neurofisiología subyacente. El modelo describe la evolución estocástica de las actividades neuronales eléctricas promedio $\{X_i\}$ de ciertas regiones del cerebro $\{i\}$. Cada i etiqueta una posible decisión que un agente cognitivo va a decidir si tomar o no. El proceso evoluciona de acuerdo a (4.1), comenzando con $X_i(t=0) \sim 0^+$, y deteniéndose en el tiempo $t = T(i^*)$. La condición de parada se activa cuando un valor de actividad neuronal alcanza por primera vez un umbral determinado $X_{i^*} = Z$, en cuyo caso la decisión que se toma es i^*.

$$dX_i = \left[I_i - \sum_j \omega_{ij} X_j \right] dt + \sigma_i dW_i, \quad i = 1, \ldots, M \tag{4.1}$$

Otros términos en (4.1) corresponden a parámetros exógenos. I_i es un valor de entrada a favor de la alternativa i que se acumula desde otros dispositivos, como la corteza visual, y sirve como entrada para el proceso LCA. Se supone

que esos valores están restringidos como $I_i \geq 0$ según el razonamiento neurofisiológico. Los valores de entrada externos se acumulan en la variable X_i a favor de la alternativa i. El parámetro ω está representado por dos valores, como se muestra en (4.2).

$$\omega_{ij} = \begin{cases} \kappa & i = j \\ \lambda & i \neq j \end{cases} \tag{4.2}$$

El parámetro κ de (4.2) tiene en cuenta el decaimiento [5].

La inhibición lateral entre las unidades del acumulador está controlada por el parámetro λ y considera el mismo efecto para todas las unidades. Los valores acumulados se consideran valores biológicos, como la actividad neuronal (tasa de picos), que luego se restringen a ser positivos ($X > 0$). De acuerdo con [22], la toma de decisiones perceptual consta de tres etapas de procesamiento cerebral. Primero, se lleva a cabo una representación de la evidencia sensorial después del recuerdo del sistema sensorial. En segundo lugar, la integración de la información sensorial disponible se realiza a lo largo del tiempo en un búfer adecuado. Finalmente, o se hace una comparación entre las pruebas a favor de las decisiones involucradas, o un umbral desencadena la elección en algunos casos.

4.2.2. ELCA

Realizamos dos modificaciones al modelo LCA expuesto en la sección 4.1. En primer lugar, hacemos una segmentación de los usuarios según su nivel de implicación en el foro. Esta clasificación en tipos de usuarios se obtiene de los expertos de la comunidad del foro.

Creemos que es razonable suponer que su proceso de toma de decisiones se

ve afectado de manera diferente por el contenido dado que se comportan de manera diferente en su participación en la comunidad.

Además, modificamos el modelo LCA incluyendo un término que da cuenta de la formación de hábitos reforzando la probabilidad de elegir una alternativa que se ha elegido antes. Esto se hace como mostramos en (4.3)

$$X_i(t=0) = 0 + (1+\gamma)^{\text{number of successes}} - 1 \tag{4.3}$$

donde $\gamma \geq 0$. Es importante entender que esta es una primera aproximación a la incorporación de la formación de hábitos en el modelo. De acuerdo con la literatura sobre formación de hábitos, el aumento de la automaticidad por repetición sigue una curva asintótica [37] pero esto es a largo plazo. Teniendo en cuenta que deseamos incorporar los efectos de la formación de hábitos durante una etapa inicial, nuestra aproximación es aceptable.

Podemos ver la estructura de pseudocódigo del modelo propuesto en el Algoritmo 1.

Algorithm 1 Predicción de contribución de posts via ELCA basado en la valoración de los hilos de conversación por parte de los usuarios, para cada periodo en el horizonte temporal, después de estimar $\hat{\beta}_c$, $\hat{\kappa}_c$, y $\hat{\lambda}_c$ mediante el algoritmo genético descrito en el Algoritmo 2.

Input: conjunto de posts agrupados por periodo de tiempo
$\mathcal{C} = \{\mathcal{C}_t; t = 1, \ldots, T\}$ donde $\mathcal{C}_t = \{p_1^t, \ldots, p_{N_t}^t; p_k \in \mathcal{P}\}$ después de la curación de la data; cada post es una tupla $p = [u, h, \{v\} \subset \mathcal{V}]$

Preprocesamiento Semántico: Aplicar LDA para obtener la representación semántica de los post como una combinación de tópicos $\{\rho_p \subset \mathcal{T}; p \in \mathcal{P}\}$.

For each $t \in \{2, \ldots, T\}$

1. Calcular la representación semántica de cada hilo de conversación en cada periodo de tiempo $\nu_h^t = \frac{1}{|\mathcal{P}(h,t)|} \sum_{p \in \mathcal{P}(h,t)} \rho_p$,

2. Calcular la representación semántica de las preferencias de cada usuario en cada periodo de tiempo $\mu_{u,s}^t = \frac{1}{|\mathcal{P}(u,s,t)|} \sum_{p \in \mathcal{P}(u,s,t)} \rho_p$,

3. Computar la utilidad de cada hilo de coversación para cada usuario $\Psi_1(\mu_{u,s}^t, \nu_h^t) = \frac{1}{1-\chi(\mu_{u,s}^t, \nu_h^t)}$, donde $\chi(\mu_{u,s}^t, \nu_h^t) = \frac{\mu_{u,s}^t \cdot \nu_h^t}{|\mu_{u,s}^t||\nu_h^t|}$ es la distancia coseno,

4. Calcular las utilidades normalizadas $V_{u,s,h}^t = \Psi_2(a, \mu_{u,s}^t, \nu_h^t) = a \dfrac{\Psi_1(\mu_{u,s}^t, \nu_h^t)}{\max\limits_{j \in \mathcal{TH}_f^t} \Psi_1(\mu_{u,s}^t, \nu_j^t)}$,

5. Computar las valoraciones de cada hilo de conversación por cada usuario $I_{u,s,h}^t = \Omega(\mathbf{V}_{u,s}^t(m), h) = \hat{\beta}_{(c(u))} e^{V_{u,s,h}^t} \left(\sum\limits_{j \in \mathcal{TH}_f^t(u,m)} e^{V_{u,s,j}^t} \right)^{-1}$.

6. Para cada usuario u e hilo h integrar usando el método de Euler las ecuaciones diferencias del modelo ELCA

$$dX_h^{(u)}(\tau) = \left[I_{u,s,h}^t - \sum_{j \in \mathcal{TH}_f^t} \hat{\omega}_{hj}^{(c(u))} X_j^{(u)}(\tau) \right] d\tau + \sigma_h^{(u)} dW_h,$$

   hasta que $X_h^{(u)}(\tau^*) > Z$, donde Z es el umbral de decisión, para cada usuario u.

7. Los arcos predichos del grafo de publicación de contribuciones están dados por $PG_t = \left\{ (u,h) \,\middle|\, X_h^{(u)}(\tau^*) > Z \right\}$. Se calculan metricas de desempeño (Sección 5.2) comparando con la vardad fundamental de las publicaciones de posts $GT_t = \{(u,h) \,|\exists\, [u,h,] \in \mathcal{C}_t\}$.

4.3. Estimación de parámetros del modelo ELCA

Finalmente, implementamos una heurística de algoritmo genético para estimar los parámetros para el modelo ELCA como se ilustra en la Fig. 4.2. La implementación está basada en el trabajo realizado en [69] donde usamos el algoritmo de clasificación lineal de Baker [63] como la función de aptitud que indica qué tan bien la población actual se ajusta a la función objetivo. Para la reproducción en el algoritmo usamos la selección de la rueda de la ruleta [45] como mecanismo de selección y el punto único [7, 8] como la rutina de cruce. Finalmente, para realizar la operación de mutación usamos la mutación de valor real [6].

Algorithm 2 Algoritmo genético para la estimación de los parámetros del ELCA $\hat{\beta}_c$, $\hat{\kappa}_c$, y $\hat{\lambda}_c$.

Input: conjunto de posts del primer periodo de tiempo
$\mathcal{C}_1 = \{p_1^1, \ldots, p_{N_1}^1; p_k \in \mathcal{P}\}$ después de la curación de la data; cada post es una tupla $p = [u, h, \{v\} \subset \mathcal{V}]$; representación semantica de los posts como una combinación de tópicos $\{\rho_p \subset \mathcal{T}; p \in \mathcal{P}\}$.

1. Construir una población inicial aleatoria
   $\mathbf{P}\,(k=0) = \{P_g\,(k)\,; g = 1, \ldots, 100\}$, donde
   $P_g\,(k) = \left\{ \left(\hat{\beta}_c\,(k)\,, \hat{\kappa}_c\,(k)\,, \hat{\lambda}_c\,(k) \right), c \in \{A, B, C, X\} \right\}$,
2. Estimar la función de aptitud inicial $f_g\,(k=0)$ para cada individuo mediante
   a) Correr la predicción en el Algoritmo 1 sobre $\mathcal{C}_1$ usando $P_g\,(k)$ como los parámetros del ELCA.

b) La aptitud $f_g\,(k=0)$ es la *accuracy* de la predicción de PG_1 contra GT_1 después de alcanzar la convergencia.

3. Para las generaciones $k = 1, \ldots, 1000$

   a) seleccionar mediante *roulette wheel* 10 individuos viejos a ser preservados para la siguiente generación $\mathbf{P}_{old,10}\,(k)$

   b) seleccionar 90 pares de cruza mediante *roulette wheel* sobre los valores de aptitud $\{f_g\,(k-1)\}$ de la generación progenitora $\mathbf{P}\,(k-1)$

   c) aplicar *single point crossover* para obtener la descendencia $\mathbf{P}_{cross,90}\,(k)$

   d) aplicar *real valued mutation* a $\mathbf{P}_{cross,90}\,(k)$ para obtener $\mathbf{P}_{mut,90}\,(k)$

   e) computar la función de aptitud $f_g\,(k)$ de cada individuo en $\mathbf{P}_{mut,90}\,(k)$ como se especifica en el paso 2.

   f) $\mathbf{P}\,(k) = \mathbf{P}_{old,10}\,(k) \cup \mathbf{P}_{cross,90}\,(k)$

4. Retornar el individuo $P_g^*\,(k)$ con mayor aptitud $f_g^*\,(k) = \underset{g,k}{\text{máx}}\,\{f_g\,(k)\}$.

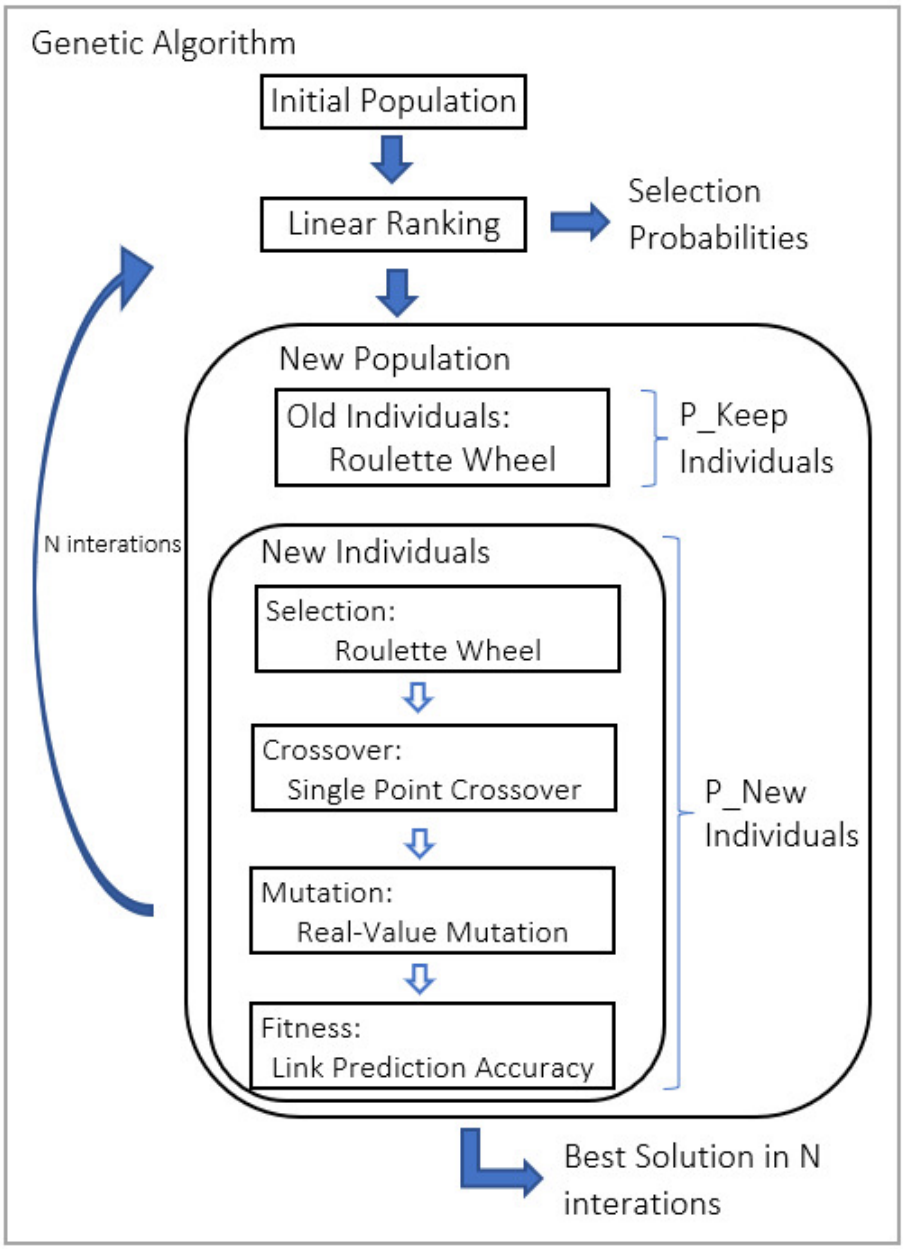

Figura 4.2: Diagrama de flujo del algoritmo genético usado para la búsqueda de parámetros óptimos del modelo ELCA

Capítulo 5

Experimentos, Resultados y Evaluación

Este capítulo presenta los resultados de los experimentos computacionales realizados sobre datos de una red social real como demostradores del modelo propuesto en la presente Tesis. Primero, introducimos la configuracion experimental utilizada en este estudio seguido por las métricas utilizadas para la evaluación comparativa y la metodología de evaluación de los resultados obtenidos. Posteriormente, se presentan los resultados obtenidos al aplicar los modelos para cada subforo del caso de estudio. Finalmente, se presenta una discusión de los resultados obtenidos

5.1. Configuración Experimental

Para dar respuesta a la pregunta de este estudio se dispuso realizar una comparación del desempeño del modelo propuesto en esta Tesis con 2 modelos clásicos de la literatura de machine learning, a saber, Random Forest y Sup-

port Vector Machines, en la tarea de predecir las decisiones de generación de contenido de los usuarios de una red social. Para la realización de los experimentos se extrajo un año del total de la data entregada por los administradores de Plexilandia descrita en el capítulo 3, en específico, para cada publicación realizada en el foro entre enero de 2013 y enero de 2014 se obtuvo: ID de usuario, ID de publicación, ID de hilo, ID de subforo, contenido de texto de publicación y hora de publicación. En particular, al contenido de texto de cada publicación se le debe aplicar el preprocesamiento de datos descrito en la sección 3.2.1 con el objetivo de obtener representaciones vectoriales del contenido de las publicaciones. Por otra parte, se crea un atributo denominado Tipo de usuario, en concordancia con la información obtenida sobre los miembros clave de la red, distinguiendo entre 4 tipos de usuario como se explica en la sección 3.1.

Posteriormente, en concordancia con la topología de red propuesta en este trabajo, se divide el conjunto de datos por subforos. Al explorar la cantidad de publicaciones durante diferentes períodos de tiempo (1 semana, 2 semanas, 1 mes, 2 meses, 4 meses) y el comportamiento de los hilos durante ese tiempo, decidimos elegir un tamaño de período de tiempo de 1 mes obteniendo un total de 13 periodos de tiempo. En la Fig. 5.1 mostramos la forma en que dividimos los datos y cómo elegimos realizar los experimentos, usando el primer mes de 2013 (enero) como datos de calibración para el modelo y el resto de los meses como datos de prueba.

Después de realizar la separación, pudimos calcular la cantidad de usuarios activos, hilos activos y publicaciones realizadas durante cada uno de estos 13 meses para cada uno de los subforos, como se muestra en las tablas 5.1, 5.2 y 5.3.

Procedemos, como se describe en la sección 4.2.2, a asignar la representación del vector de texto para cada mes a cada hilo activo usando el vector medio de todas las publicaciones que pertenecen a ese hilo durante el mes que se

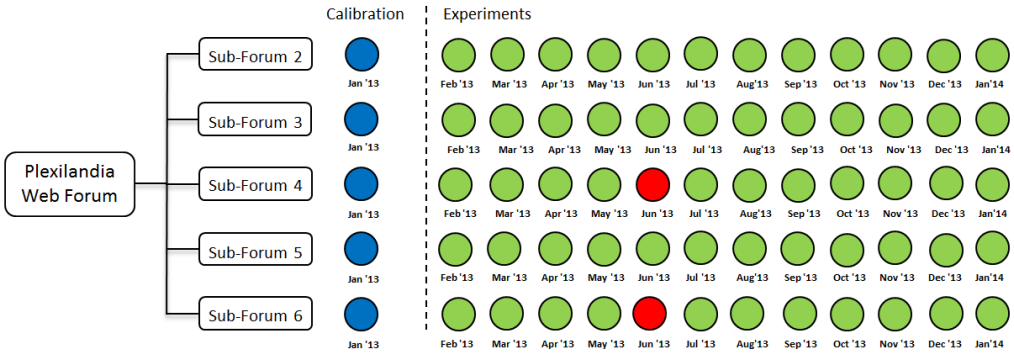

Figura 5.1: Configuración Experimental

está probando. En cuanto a los usuarios, asignamos m representaciones de vectores de texto agrupando las publicaciones según el hilo al que pertenecen y calculando el vector medio del grupo de publicaciones. De esta forma también recuperamos el número de hilos (m) con los que un usuario forma un enlace en nuestra representación de red propuesta.

5.2. Métricas y Metodología de Evaluación

Para evaluar la calidad de nuestra propuesta, decidimos utilizar el siguiente marco de evaluación.

Calcularemos 4 medidas para evaluar el rendimiento del modelo. A saber, recuperación, exactitud, precisión y medida F [65]. A continuación se proporciona una descripción de estas métricas.

La recuperación da una medida de la probabilidad de detección del modelo y se define como:

$$Recall = \frac{\text{Number of true positive links}}{\text{Number of real links}} \tag{5.1}$$

Tabla 5.1: Usuarios activos, hilos activos y publicaciones realizadas en los subforos (a) 2 y (b) 3

Month	**Users**	**Threads**	**Posts**
1	45	25	103
2	19	10	51
3	35	20	83
4	38	27	133
5	32	22	55
6	33	22	94
7	26	14	57
8	38	24	127
9	35	17	94
10	35	23	110
11	38	22	121
12	31	19	94
13	27	14	59
Total	**168**	**221**	**1181**

(a) Estadísticas del Sub-Foro 2

Month	**Users**	**Threads**	**Posts**
1	49	43	145
2	46	29	169
3	51	46	252
4	53	43	196
5	51	44	184
6	52	38	208
7	49	32	173
8	42	37	171
9	43	33	174
10	44	29	138
11	43	24	124
12	49	38	156
13	31	30	102
Total	**174**	**351**	**2192**

(b) Estadísticas del Sub-Foro 3

La precisión da una medida de la veracidad de los resultados en el sentido de que describe errores de observación sistemáticos o sesgos estadísticos en el modelo. La precisión se define como:

$$Accuracy = \frac{\text{Number of true positive links} + \text{Number of true negative links}}{\text{Number of possible links}} \tag{5.2}$$

La precisión da una medida de la variabilidad estadística o, en otras palabras, describe el error de observación aleatorio del modelo. Se define como:

$$Precision = \frac{\text{Number of true positive links}}{\text{Number of predicted links}} \tag{5.3}$$

La medida F o puntuación F_1 combina las medidas de precisión y recuerdo obteniendo una medida alternativa de la precisión del modelo y se define

Tabla 5.2: Usuarios activos, hilos activos y publicaciones realizadas en los subforos (a) 4 y (b) 5

Month	**Users**	**Threads**	**Posts**
1	32	40	115
2	25	8	81
3	20	13	60
4	22	15	50
5	12	8	23
6	5	3	7
7	19	10	46
8	21	17	57
9	19	10	52
10	20	9	30
11	22	9	72
12	12	8	33
13	28	17	104
Total	**96**	**134**	**730**

(a) Estadísticas del Sub-Foro 4

Month	**Users**	**Threads**	**Posts**
1	60	37	164
2	47	27	131
3	58	30	182
4	36	23	84
5	55	28	145
6	53	36	202
7	55	35	176
8	45	29	116
9	25	19	72
10	34	25	66
11	25	13	41
12	42	25	105
13	38	24	98
Total	**171**	**282**	**1582**

(b) Estadísticas del Sub-Foro 5

como:

$$\text{F measure} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \tag{5.4}$$

Para realizar la comparación de desempeño con los otros modelos se utilizará la medida F debido a que es una métrica más confiable cuando se trabaja con data que presenta desequilibrio de clases, que se corresponde con la situación del caso de estudio, tal como suele suceder al utilizar data de redes, donde el número de arcos negativos o inexistentes es mucho mayor que el número de arcos positivos o existentes.

Tabla 5.3: Usuarios activos, hilos activos y publicaciones realizadas en el subforo 6

Month	**Users**	**Threads**	**Posts**
1	14	11	49
2	7	5	13
3	16	6	33
4	6	5	13
5	11	9	30
6	11	5	13
7	10	7	52
8	9	3	13
9	11	7	41
10	15	5	27
11	8	5	37
12	15	6	36
13	11	6	27
Total	**50**	**47**	**384**

5.3. Resultados Experimentales

Ejecutamos el GA para obtener los siguientes parámetros optimizados para el modelo ELCA

Tabla 5.4: Valores calibrados de (a) β y (b) κ

Sub-Forum	$\beta_{\mathbf{A}}$	$\beta_{\mathbf{B}}$	$\beta_{\mathbf{C}}$	$\beta_{\mathbf{X}}$
2	0.863	0.148	0.511	0.553
3	0.584	0.906	0.389	0.029
4	0.586	0.833	0.352	0.476
5	0.628	0.184	0.000	0.429
6	0.516	0.126	0.490	0.595

(a) β valores calibrados

Sub-Forum	$\kappa_{\mathbf{A}}$	$\kappa_{\mathbf{B}}$	$\kappa_{\mathbf{C}}$	$\kappa_{\mathbf{X}}$
2	0.174	0.055	0.070	0.965
3	0.684	0.340	0.217	0.588
4	0.642	0.389	0.866	0.981
5	0.707	0.733	0.047	0.623
6	0.287	0.692	0.087	0.401

(b) κ valores calibrados

Con estos valores de parámetros, se usa el modelo para simular las redes de cada subforo y de cada mes entre febrero de 2013 y enero de 2014. Como se especifica en Algoritmo 1, el resultado de la modelo ELCA son pares

Tabla 5.5: λ valores calibrados

Sub-Forum	$\lambda_{\mathbf{A}}$	$\lambda_{\mathbf{B}}$	$\lambda_{\mathbf{C}}$	$\lambda_{\mathbf{X}}$
2	0.491	0.137	0.399	0.189
3	0.146	0.951	0.189	0.949
4	0.639	0.478	0.107	0.245
5	0.0935	0.864	0.847	0.640
6	0.956	0.869	0.044	0.315

compuestos de hilo de conversacion y usuario

$$PG_t = \left\{ (u,h) \,\middle|\, X_h^{(u)}\left(\tau^*\right) > Z \right\}$$

que deben ser interpretados como predictores de los pares reales que se pueden extraer de la verdad fundamental de las publicaciones de posts

$$GT_t = \{(u,h) \,|\exists\, [u,h,] \in \mathcal{C}_t\}$$

Se realizan predicciones independientes para cada período de tiempo y subforo. Estos pares se pueden visualizar como los arcos de gráficos bipartitos, que son los gráficos de publicación predichos y reales. Podemos definir verdaderos positivos como las aristas que están en ambos gráficos, verdaderos negativos como las aristas que están ausentes de las dos gráficas, falsos positivos son las aristas que aparecen en la predicción pero están ausentes en la verdad fundamental, y falsos negativos las aristas que están ausentes en la predicción pero aparecen en la verdad fundamental.

Luego, de estos resultados entregados por el modelo se extraen las reglas de decisión y se reconstruyen los gráficos de red simulados y reales de acuerdo con la representación de red propuesta. Posteriormente, se calculan las 4 métricas para evaluar rendimiento con respecto a los arcos previstos. Para cada Sub-Foro se presentan los resultados obtenidos para estas 4 métricas y 2 imágenes de red representativas del mejor y peor resultado en la medida F

para el marco temporal considerado, como también los resultados obtenidos para la medida F de los modelos RF y SVM.

5.3.1. Sub-Foro 2

En la Tabla 5.6 se muestran los resultados obtenidos para cada una de las métricas evaluadas para el Sub-Foro 2. Como podemos notar, el mejor resultado con respecto a la medida F se obtiene en el mes 2 y el peor en mes 4.

Tabla 5.6: Resultados del Sub-Foro 2

Month	**Recall**	**Accuracy**	**Precision**	**F-measure**
2	0.724	0.916	0.724	0.724
3	0.525	0.924	0.554	0.539
4	0.435	0.910	0.457	0.446
5	0.511	0.939	0.523	0.517
6	0.473	0.924	0.5	0.486
7	0.643	0.920	0.659	0.651
8	0.527	0.928	0.557	0.542
9	0.566	0.928	0.6	0.583
10	0.556	0.937	0.603	0.579
11	0.485	0.917	0.493	0.489
12	0.526	0.910	0.536	0.531
13	0.667	0.934	0.684	0.675
Mean	**0.553**	**0.924**	**0.574**	**0.564**
Max	**0.724**	**0.939**	**0.724**	**0.724**
Min	**0.435**	**0.910**	**0.457**	**0.446**

Por contraparte, en la Tabla 5.7 se muestran los resultados obtenidos de la medida F para el modelo RF y SVM para el Sub-Foro 2. Como se puede constatar, ambos modelos obtienen resultados similares entre ellos y mediocres en comparación al modelo ELCA a lo largo de todo el horizonte temporal evaluado.

Tabla 5.7: Resultados del Sub-Foro 2

Month	**RF**	**SVM**
2	0.20	0.21
3	0.16	0.19
4	0.10	0.11
5	0.14	0.12
6	0.13	0.11
7	0.17	0.13
8	0.10	0.11
9	0.15	0.17
10	0.14	0.15
11	0.12	0.18
12	0.13	0.11
13	0.17	0.15
Mean	**0.14**	**0.145**
Max	**0.20**	**0.21**
Min	**0.10**	**0.11**

5.3.1.1. Mejor resultado en el Subforo 2

En la Tabla 5.8 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 2 durante el mes 2

La Fig. 5.2 muestra la red del subforo 2 para el mes 2, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.8. Los nodos correspondientes a hilos de conversación se muestran en color violeta, mientras que los nodos correspondientes a usuarios en color negro. Por su parte, los arcos de color negro corresponden a los arcos que la simulación predijo correctamente, los arcos de color verde son los arcos que la simulación predijo incorrectamente y, por último, los arcos de color rojo corresponden a los arcos que la simulación no pudo predecir. Se puede observar que la mayoría de los arcos de la red son negros y que hay aproximadamente la misma cantidad de enlaces previstos que de enlaces reales.

Tabla 5.8: Reglas de Decisión de Publicación de Post para Subforo 2 Mes 2. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U1	T239,T256,T389	U62	T230	U137	T256
U2	T256	U67	T230,T283,T289	U141	T230
U17	T239	U72	T259	U178	T253
U22	T256,T273	U75	T283	U196	T262
U24	T239	U86	T273	U228	T273,T283
U43	T230	U106	T283		
U49	T256,T257,T259,T262	U107	T239,T273		

5.3.1.2. Peor resultado en el Subforo 2

En la Tabla 5.9 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 2 durante el mes 4

La Fig. 5.3 muestra la red del subforo 2 para el mes 2, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.9. Al igual que en la imagen anterior, los hilos de conversación son representados por nodos de color violeta y los usuarios por nodos de color negro. Asimismo, los arcos siguen el emparejamiento: arco color negro : arco predicho correctamente, arco color verde : arcos predicho incorrectamente, arco color rojo: arco no predicho. En esta imagen se ve un aumento de la proporcion de arcos verdes y rojos con respecto a la imagen correspondiente al mes 2. Sin embargo, aún se mantiene una alta cantidad de arcos negros. Cabe destacar que el modelo ELCA sigue dominando en cuanto a desempeño a los modelos de referencia usados.

Tabla 5.9: Reglas de Decisión de Publicación de Post para Subforo 2 Mes 4. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U1	T383,T392,T410, T441	U46	T408	U131	T392
U2	T383	U47	T449	U134	T289
U6	T415,T440	U49	T74,T257,T316, T383,T404,T418	U154	T257,T383,T410
U8	T74,T375,T433	U62	T374	U188	T289,T404
U9	T440	U64	T375	U228	T273,T283
U13	T374	U67	T374,T375,T404, T412	U190	T391
U15	T257	U72	T257	U209	T397,T401
U17	T257,T383,T392, T399,T441	U75	T374,T391,T392, T401,T404,T415, T420	U228	T74,T257,T375, T397,T426,T433
U19	T392	U76	T257,T410,T418	U229	T392
U24	T74,T415	U78	T408	U233	T283,T433
U32	T316	U110	T373	U245	T316
U34	T401,T418	U117	T397,T404,T410	U251	T375
U43	T257,T401	U128	T373,T412	U254	T316,T415

5.3.2. Sub-Foro 3

En la Tabla 5.10 se presentan los resultados obtenidos para cada una de las métricas evaluadas para el Sub-Foro 3. Es posible notar que, el mejor resultado con respecto a la medida F se obtiene en el mes 13 mientras que el peor se consigue en el mes 11. Adicionalmente, se observa un desempeño ligeramente inferior en este Sub-Foro en contraste al obtenido para el Sub-Foro 2.

Analogamente, en la Tabla 5.11 se muestran los resultados obtenidos de la medida F para el modelo RF y SVM para el Sub-Foro 3. Al igual que para

Tabla 5.10: Resultados del Sub-Foro 3

Month	**Recall**	**Accuracy**	**Precision**	**F-measure**
2	0.432	0.909	0.453	0.442
3	0.453	0.929	0.477	0.465
4	0.496	0.943	0.515	0.506
5	0.431	0.939	0.445	0.438
6	0.496	0.939	0.508	0.502
7	0.429	0.925	0.441	0.435
8	0.455	0.929	0.495	0.474
9	0.440	0.911	0.451	0.445
10	0.556	0.939	0.568	0.562
11	0.410	0.917	0.445	0.427
12	0.471	0.943	0.485	0.478
13	0.632	0.951	0.672	0.652
Mean	**0.475**	**0.931**	**0.496**	**0.486**
Max	**0.632**	**0.951**	**0.672**	**0.652**
Min	**0.410**	**0.909**	**0.441**	**0.427**

el Sub-Foro 2, ambos modelos de referencia obtuvieron resultados similares entre ellos, con SVM obteniendo una media levemente mejor a la obtenida por RF. Sin embargo, se registra una baja considerable al contrastar el desempeño de estos modelos con repecto al obtenido por los mismos para el Sub-Foro 2. Por último, el modelo ELCA sigue mostrándose significativamente superior, de manera sostenida a lo largo de todo el horizonte temporal evaluado.

5.3.2.1. Mejor resultado en el Subforo 3

En la Tabla 5.12 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 3 durante el mes 13 derivadas de los resultados del modelo.

En la Fig. 5.4 se observa la red del subforo 3 para el mes 13, reconstruida a

Tabla 5.11: Resultados del Sub-Foro 3

Month	**RF**	**SVM**
2	0.08	0.10
3	0.07	0.05
4	0.09	0.11
5	0.07	0.09
6	0.08	0.10
7	0.12	0.15
8	0.10	0.12
9	0.10	0.13
10	0.11	0.16
11	0.15	0.19
12	0.08	0.11
13	0.14	0.13
Mean	**0.10**	**0.12**
Max	**0.15**	**0.19**
Min	**0.07**	**0.05**

partir de la reglas de decisión de publicación presentadas en la Tabla 5.12. Continuando con la convención utilizada en las imágenes anteriores, los hilos de conversación se muestran como nodos de color violeta, los usuarios como nodos de color negro, los arcos verdaderos positivos son de color negro, los arcos falsos positivos son verdes y los arcos falsos negativos son presentados en color rojo. En el grafo se capta una abundante proporción de arcos verdaderos positivos (negros) con respecto a los otros tipos de arcos. Al realizar una comparación con lo ocurrido en el Sub-Foro 2 se ve que a pesar de no alcanzar un rendimiento tan elevado como en el mejor de los meses de ese Sub-Foro, el modelo ELCA logra un buen desempeño para una situación similar, en cuanto a cantidad de usuarios e hilos de conversación activos, al peor mes del Sub-Foro 2 .

Tabla 5.12: Reglas de Decisión de Publicación de Post para Subforo 3 Mes 13. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U1	T38,T58	U111	T38,T979,T986, T1004	U210	T5,T798,T963, T973,T991,T1014
U3	T990	U130	T972,T1014	U215	T1014
U8	T9,T1026	U137	T990	U228	T964,T973,T979, T1002,T1004
U9	T38	U151	T50	U229	T961,T962,T979
U13	T1009	U157	T33	U275	T1021
U19	T798,T993	U159	T5,T1004	U278	T963,T964,T990, T993,T998
U43	T1014	U161	T962,T963,T993, T1003	U290	T50
U49	T962	U165	T972,T990	U291	T1003
U99	T968	U173	T33,T76	U299	T58
U104	T133,T991,T1026	U189	T14,T1021		
U107	T961,T968	U198	T38,T964		

5.3.2.2. Peor resultado en el Subforo 3

En la Tabla 5.13 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 3 durante el mes 11

Por su parte, en la Fig. 5.5 se observa la red del subforo 3 para el mes 11, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.13 obtenidas de los resultados del modelo ELCA. Este grafo presenta el peor caso, con respecto a desempeño del modelo, del Sub-Foro 3. De la misma manera que se explicó antes los hilos de conversación se muestran como nodos de color violeta, los usuarios como nodos de color negro, los arcos verdaderos positivos son de color negro, los arcos falsos positivos son verdes y los arcos falsos negativos son presentados en color rojo. En la imagen se percibe una mayor cantidad de usuarios e hilos de conversación que en el

Tabla 5.13: Reglas de Decisión de Publicación de Post para Subforo 3 Mes 11. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U1	T877	U107	T25,T894	U228	T14,T577,T853, T891,T900
U8	T853,T896	U108	T811	U229	T49,T891
U9	T858	U111	T66,T853,T858, T900	U242	T811
U11	T172	U121	T894	U259	T857
U17	T858,T870	U134	T577	U268	T577
U18	T857	U148	T26	U275	T891
U19	T856	U151	T25,T798,T870	U293	T14
U20	T891	U161	T26,T870,T877	U298	T853
U34	T577,T896	U165	T870,T877	U302	T26,T66
U43	T49,T853,T858	U189	T798,T856	U304	T868
U46	T894	U196	T879	U305	T14
U61	T858	U202	T811,T900	U306	T876
U73	T811	U208	T857,T879,T891	U307	T66,T877
U88	T14,T870	U210	T66,T172,T811, T856		
U103	T49	U220	T858		

mes 13, y cabe mencionar que se obtiene un resultado levemente inferior al obtenido para el mes 4 del Sub-Foro 2.

5.3.3. Sub-Foro 4

En la Tabla 5.14 mostramos los resultados obtenidos para cada una de las métricas evaluadas para el Sub-Foro 4. Como podemos notar, el mejor resultado con respecto a la medida F se obtiene en el mes 5 y el peor en mes 3. Debido al bajo número de publicaciones, usuarios e hilos, se descartaron los resultados obtenidos para el mes 6. Se distingue que los resultados de medi-

da F obtenidos para este Sub-Foro superan a los obtenidos en los Sub-Foros analizados previamente.

Tabla 5.14: Resultados del Sub-Foro 4

Month	**Recall**	**Accuracy**	**Precision**	**F-measure**
2	0.600	0.825	0.698	0.645
3	0.454	0.831	0.500	0.476
4	0.632	0.915	0.632	0.632
5	0.778	0.917	0.778	0.778
6	—	—	—	—
7	0.581	0.879	0.643	0.610
8	0.634	0.927	0.703	0.667
9	0.636	0.884	0.677	0.656
10	0.692	0.917	0.720	0.706
11	0.650	0.874	0.708	0.675
12	0.700	0.885	0.737	0.718
13	0.537	0.876	0.563	0.550
Mean	**0.627**	**0.885**	**0.669**	**0.647**
Max	**0.778**	**0.927**	**0.778**	**0.778**
Min	**0.454**	**0.825**	**0.500**	**0.476**

Complementariamente, en la Tabla 5.15 se muestran los resultados obtenidos de la medida F para el modelo RF y SVM para el Sub-Foro 4. Una vez más, ambos modelos de referencia obtuvieron resultados similares entre ellos, con SVM obteniendo una media levemente mejor a la de RF. Al igual que el modelo ELCA, los modelos RF y SVM presentan mejores resultados para este Sub-Foro que para los estudiados con anterioridad. Además, se puede volver a constatar que el modelo ELCA sigue superando en desempeño a los modelos de referencia.

Tabla 5.15: Resultados del Sub-Foro 4

Month	**RF**	**SVM**
2	0.22	0.18
3	0.19	0.22
4	0.24	0.28
5	0.40	0.38
6	****	****
7	0.30	0.33
8	0.19	0.22
9	0.22	0.19
10	0.27	0.23
11	0.26	0.25
12	0.31	0.28
13	0.17	0.18
Mean	**0.23**	**0.25**
Max	**0.40**	**0.38**
Min	**0.17**	**0.18**

5.3.3.1. Mejor resultado en el Subforo 4

En la Tabla 5.16 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 4 durante el mes 5

Tabla 5.16: Reglas de Decisión de Publicación de Post para Subforo 4 Mes 5. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U6	T365,T453	U67	T453,T457,T485	U177	T485
U9	T365,T367,T488	U114	T365,T438	U198	T367,T488
U22	T438	U155	T488	U229	T438,T453,T457, T488
U34	T365	U163	T470	U233	T438

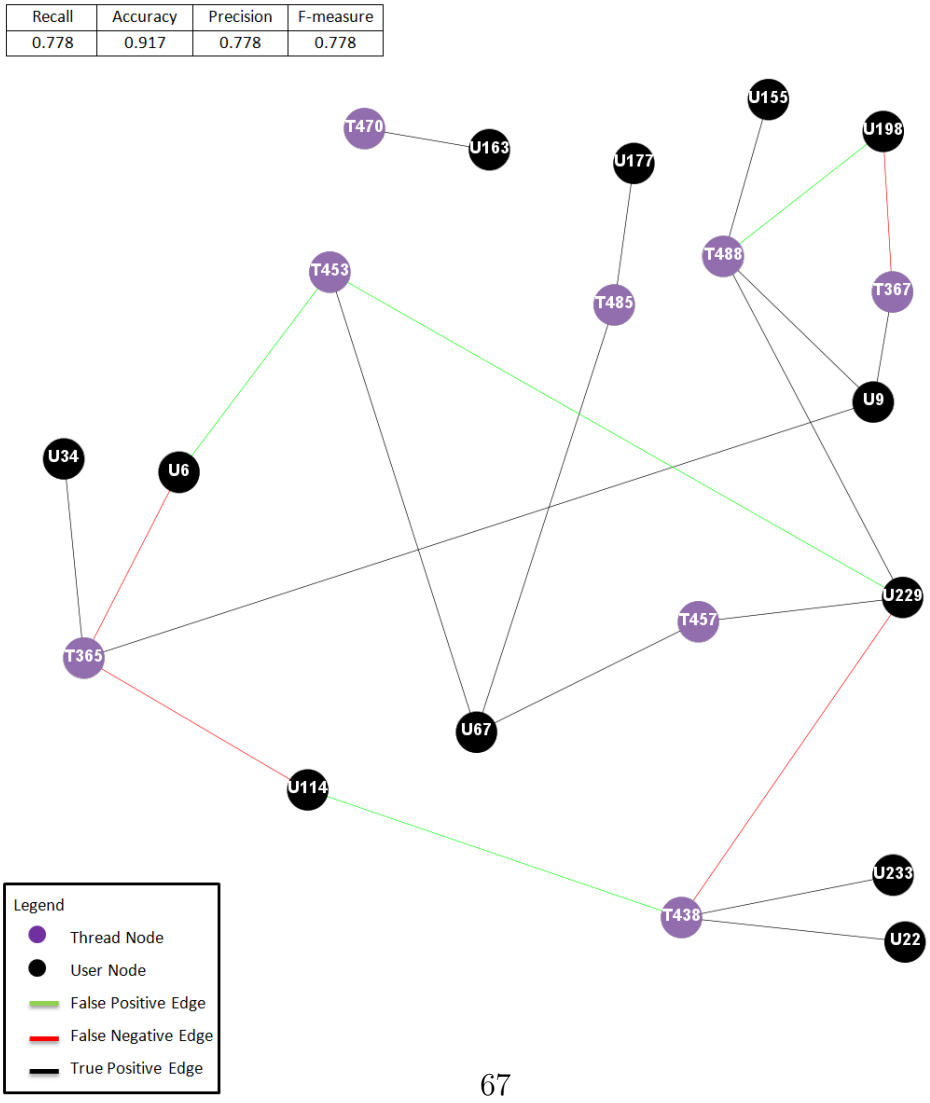

Figura 5.6: Red del Sub-Foro 4 para el Mes 5

En la Fig. 5.6 se observa la red del subforo 4 para el mes 5, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.16, correspondiente al caso de mejor desempeño para este Sub-Foro. Tal como en los otros Sub-Foros, los hilos de conversación se muestran como nodos de color violeta, los usuarios como nodos de color negro, los arcos verdaderos positivos son de color negro, los arcos falsos positivos son verdes y los arcos falsos negativos son presentados en color rojo. Se divisa en la imagen que este mes corresponde a uno más acotado en actividad que los analizados previamente. No obstante, llama la atención la poco cantidad de arcos verdes y rojos. Al comparar con los Sub-Foros anteriores se advierte un desempeño considerablemente mejor que en los mejores casos de ambos.

5.3.3.2. Peor resultado en el Subforo 4

En la Tabla 5.17 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 4 durante el mes 3 obtenidas de los resultados del modelo ELCA.

Por su parte, en la Fig. 5.7 se observa la red del subforo 4 para el mes 3, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.17. Al igual que antes, los hilos de conversación se muestran como nodos de color violeta, los usuarios como nodos de color negro, los arcos verdaderos positivos son de color negro, los arcos falsos positivos son verdes y los arcos falsos negativos son presentados en color rojo. Compratívamente, este caso tiene mejor desempeño que los peores casos de ambos de los Sub-Foros estudiados con anterioridad.

Tabla 5.17: Reglas de Decisión de Publicación de Post para Subforo 4 Mes 3. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U1	T107,T351	U67	T351	U154	T266
U3	T288	U118	T266,T295,T305	U181	T111,T320
U13	T266,T295,T305, T320	U127	T78,T236,T266, T288,T305	U201	T23,T78,T107, T260,T288,T295, T320,T351
U15	T295,T320	U129	T23,T288	U228	T23
U22	T111,T288,T351	U133	T78,T111,T288, T295,T351	U229	T16,T78,T320, T351
U43	T288,T295,T305, T320	U150	T111,T260	U233	T23,T266,T320
U56	T78,T111,T236, T260,T266,T288, T295,T320,T351	U151	T78,T266		

5.3.4. Sub-Foro 5

En la Tabla 5.18 se muestran los resultados obtenidos para cada una de las métricas evaluadas para el Sub-Foro 5. Como se puede notar, el mejor resultado con respecto a la medida F se obtiene en el mes 9 y el peor en mes 6.

De la misma forma, en la Tabla 5.19 se muestran los resultados obtenidos de la medida F para el modelo RF y SVM para el Sub-Foro 5. Manteniendo la tendencia evidenciada en los Sub-Foros el modelo SVM obtiene un rendimiento medio levemente superior al del modelo RF, ambos manteniendo desempeños similares a lo largo del horizonte temporal evaluado. Comparativamente, el desempeño de los 3 modelos en este Sub-Foro es ligeramente superior al del Sub-Foro 3 e inferior al de los Sub-Foros 2 y 4. Por último, una vez más el modelo ELCA domina en desempeño a los otros modelos.

Tabla 5.18: Resultados del Sub-Foro 5

Month	Recall	Accuracy	Precision	F-measure
2	0.474	0.939	0.500	0.487
3	0.431	0.931	0.448	0.439
4	0.557	0.936	0.567	0.562
5	0.453	0.928	0.475	0.464
6	0.377	0.919	0.402	0.389
7	0.470	0.938	0.478	0.474
8	0.457	0.926	0.483	0.470
9	0.674	0.939	0.689	0.681
10	0.615	0.955	0.640	0.627
11	0.647	0.926	0.647	0.647
12	0.507	0.933	0.521	0.514
13	0.443	0.907	0.461	0.452
Mean	**0.509**	**0.931**	**0.530**	**0.517**
Max	**0.674**	**0.955**	**0.689**	**0.681**
Min	**0.377**	**0.907**	**0.402**	**0.389**

5.3.4.1. Mejor resultado en el Subforo 5

En la Tabla 5.20 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 5 durante el mes 9 obtenidas de los resultados del modelo ELCA.

Se observa en la Fig. 5.8, la red del Sub-Foro 5 para el mes 9, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.20. Nuevamente, los hilos de conversación se muestran como nodos de color violeta, los usuarios como nodos de color negro, los arcos verdaderos positivos son de color negro, los arcos falsos positivos son verdes y los arcos falsos negativos son presentados en color rojo. Comparativamente, este caso tiene mejor desempeño solamente que el mejor caso del Sub-Foro 3.

Tabla 5.19: Resultados del Sub-Foro 5

Month	**RF**	**SVM**
2	0.13	0.11
3	0.10	0.13
4	0.11	0.15
5	0.11	0.11
6	0.07	0.11
7	0.08	0.10
8	0.09	0.11
9	0.14	0.13
10	0.14	0.17
11	0.22	0.26
12	0.10	0.12
13	0.11	0.16
Mean	**0.11**	**0.14**
Max	**0.22**	**0.22**
Min	**0.07**	**0.11**

5.3.4.2. Peor resultado en el Subforo 5

En la Tabla 5.21 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 5 durante el mes 6 obtenidas de los resultados del modelo ELCA.

Por su parte, en la Fig. 5.9 se observa la red del Sub-Foro 5 para el mes 6, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.21. Tal como en los casos previos, los hilos de conversación se muestran como nodos de color violeta, los usuarios como nodos de color negro, los arcos verdaderos positivos son de color negro, los arcos falsos positivos son verdes y los arcos falsos negativos son presentados en color rojo. Al hacer una comparación con los otros Sub-Foros se advierte que este corresponde al caso con peor desempeño de todos. No obstante, sigue siendo mejor resultado que los obtenidos por los modelos de referencia.

Tabla 5.20: Reglas de Decisión de Publicación de Post para Subforo 5 Mes 9. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U1	T747,T780	U56	T780,T783	U210	T728,T759,T780
U7	T743	U67	T540,T718,T728, T759	U228	T61,T728,T732, T741,T769,T775
U9	T540,T728,T732, T775	U81	T728	U229	T741,T771,T775
U13	T780	U88	T718	U242	T775
U23	T743	U94	T773	U245	T724,T784
U24	T741	U97	T775	U289	T679
U34	T61	U102	T61	U294	T732
U43	T769	U151	T743,T778		
U52	T732,T780	U179	T724		

5.3.5. Sub-Foro 6

En la Tabla 5.22 mostramos los resultados obtenidos para cada una de las métricas evaluadas para el Sub-Foro 6. Como podemos notar, el mejor resultado con respecto a la medida F se obtiene en el mes 10 y el peor en mes 13. Tener en cuenta que hubo problemas con los datos y no se pudo ejecutar el modelo para el mes 6.

Adicionalmente, en la Tabla 5.23 se muestran los resultados obtenidos de la medida F para el modelo RF y SVM para el Sub-Foro 6. Consolidando la tendencia, nuevamente el modelo SVM obtiene un desempeño medio ligeramente mejor al del RF, ambos manteniendo un comportamiento similar durante todos los meses. Cabe destacar que en este Sub-Foro es en el único en el cual los modelos de referencia obtienen buenos desempeños, mucho mejores que los obtenidos en cualquiera de los otros Sub-Foros. No obstante, el modelo ELCA sigue mostrándose significativamente superior, de manera sostenida a lo largo de todo el horizonte temporal evaluado.

5.3.5.1. Mejor resultado en el Subforo 6

En la Tabla 5.24 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 6 durante el mes 10 obtenidas de los resultados del modelo ELCA.

Se observa en la Fig. 5.10, la red del Sub-Foro 6 para el mes 10, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.24. De nuevo, los hilos de conversación se muestran como nodos de color violeta, los usuarios como nodos de color negro, los arcos verdaderos positivos son de color negro, los arcos falsos positivos son verdes y los arcos falsos negativos son presentados en color rojo. Este caso no solamente es el mejor caso del Sub-Foro 6, sino que, es el caso con mejor desempeño de todos, casi no presentando arcos falsos positivos ni arcos falsos negativos.

5.3.5.2. Peor resultado en el Subforo 6

En la Tabla 5.25 se muestran las reglas de decisión de publicación de posts para los usuarios del Subforo 6 durante el mes 13 obtenidas de los resultados del modelo ELCA.

Por su parte, en la Fig. 5.11 se observa la red del Sub-Foro 6 para el mes 13, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.25. Al igual que antes, los hilos de conversación se muestran como nodos de color violeta, los usuarios como nodos de color negro, los arcos verdaderos positivos son de color negro, los arcos falsos positivos son verdes y los arcos falsos negativos son presentados en color rojo. Este mes, a pesar de corresponder al peor desempeño dentro del Sub-Foro 6, sigue siendo un buen desempeño incluso superando los mejores casos de los Sub-Foros 3 y 5.

5.4. Discusión

Las imágenes de los grafos, representativos de las decisiones de contribución de contenido en los respectivos Sub-Foros, presentadas en las secciones anteriores facilitan la apreciación cualitativa de los resultados del estudio. De esas imagenes se puede observar que, en general, la mayoría de los arcos de las reded corresponden a arcos verdaderos positivos (arcos en color negro) y que hay aproximadamente la misma cantidad de aristas pronosticadas que la cantidad de aristas en la verdad fundamental de las publicaciones, que es una propiedad estructural muy importante con la que se debe cumplir. Hay pocos arcos falsos positivos en comparación con la gran cantidad de arcos inexistentes (verdaderos negativos). Esto explica los altos valores de la medida de rendimiento de precisión, en la Tablas 5.6,5.10,5.14,5.18 y 5.22, relativa a las demás medidas que sólo tienen en cuenta los verdaderos positivos. De estas tablas también se puede concluir que los conjuntos de datos de subforos empleados pueden considerarse como conjuntos de datos de dos clases muy desequilibrados para el objetivo de predicción de arcos entre usuarios e hilos de conversación. Es bién sabido, que la mayoría de los clasificadores están sesgados hacia la clase mayoritaria (En este caso: los arcos no existentes). Undersampling de la clase mayoritaria o Oversampling de la clase minoritaria se proponen como medidas a tomar para mejorar el rendimiento en la clase minoritaria, sin embargo no está claro cómo llevar a cabo estos procedimientos sobre los datos de Sub-Foros utilizados.

Al repasar los resultados obtenidos en los experimentos en términos de la medida F se advierte que el modelo neuro-semántico ELCA propuesto supera en desempeño consistentemente a lo largo del horizonte temporal y a través de los distintos Sub-Foros a los modelos de machine learning, RF y SVM usados como referencia. El modelo ELCA logra un 61 % de puntaje de medida F en promedio en todos los subforos modelando con éxito las decisiones microscópicas de generación de contenido por parte de los usuarios del

foro web con gran precisión. Por lo tanto, la hipótesis de investigación de este trabajo ha sido validada. También es importante recalcar, que los mejores resultados para la medida F son obtenidos en el Sub-Foro 6. Al parecer el menor número de publicaciones permite un análisis semántico más eficiente y facilita que el modelo encuentre los hilos de conversación en los que un usuario encuentra interés. Una observación relevante es que a medida que aumenta la cantidad de publicaciones en un Sub-Foro, los resultados predictivos empeoran. Esto se puede interpretar de manera cualitativa como que se vuelve más difícil predecir si un usuario publicará en un hilo basádonse en la descripción semántica del contenido, porque está contaminado con mensajes espurios sin filtrar. En la Fig. 5.9 se mostró el grafo de red correspondiente al mes y Sub-Foro con los peores resultados de rendimiento. En esta se advierte una gran cantidad de arcos falsos positivos. Esto condujo a profundizar la investigación, en consecuencia, en la Fig. 5.12 se muestra el diagrama de dispersión de la cantidad de publicaciones realizadas en un período de unidad de tiempo (mes) versus la puntuación de la medida F lograda por el modelo neuro-semántico ELCA en el mismo período. Todo indica que a medida que aumenta el número de publicaciones, el rendimiento de la predicción del modelo ELCA disminuye. Como antes, la causa de esta dismunición se puede interpretar como consecuencia del aumento heterogeneidad del contenido semántico en el hilo, que se convierte en muy ruidoso.

Una forma en la que se podría mejorar el modelo neuro-semántico es incorporar un comportamiento de discriminación para los usuarios, que permita filtrar las publicaciones que difieren demasiado con el vector de preferencia semántica del usuario [46]. Si se considera el comportamiento temporal de los resultados de la medida F dentro de un Sub-Foro, las puntuaciones no se desvían mucho del valor medio, por lo tanto el modelo ELCA es muy robusto en términos de decaimiento temporal. Se asocia este comportamiento con el parámetro a. En esta investigación, se fija el valor de $a = 50$ sin más búsqueda de una configuración óptima. Sin embargo, este parámetro también

podría optimizarse mediante el enfoque GA.

Recall	Accuracy	Precision	F-measure
0.724	0.916	0.724	0.724

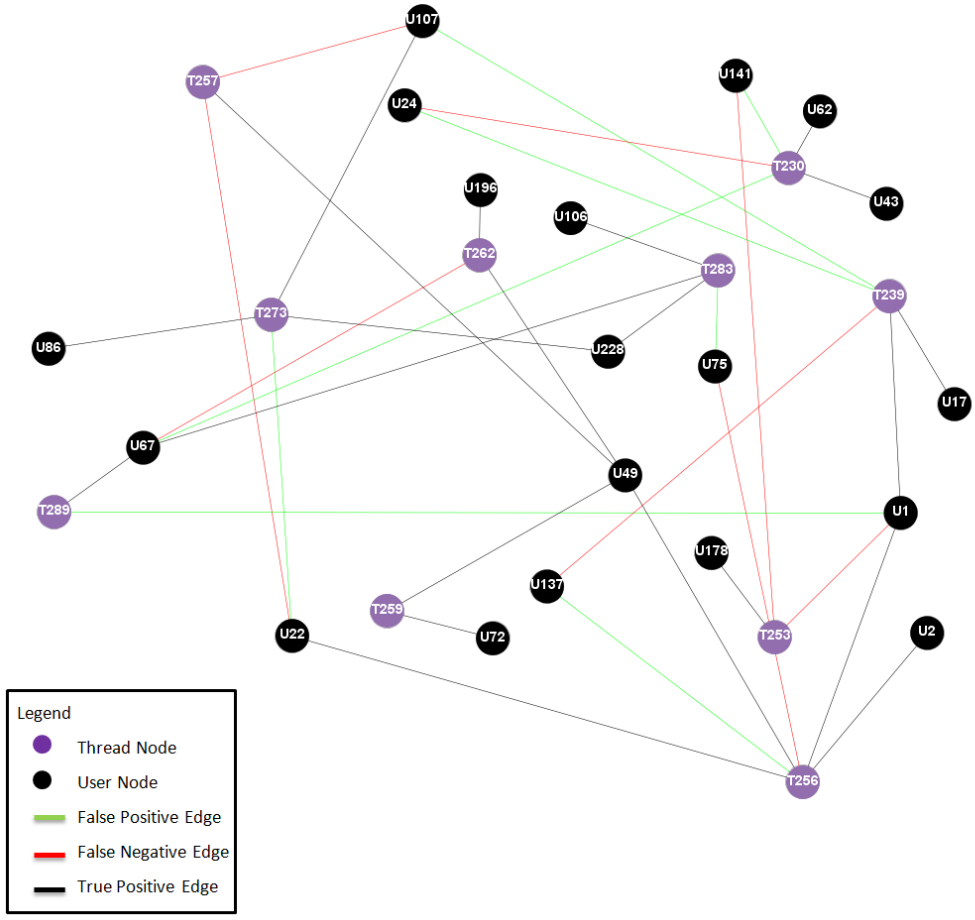

Figura 5.2: Red del Sub-Foro 2 para el Mes 2

Recall	Accuracy	Precision	F-measure
0.435	0.910	0.457	0.446

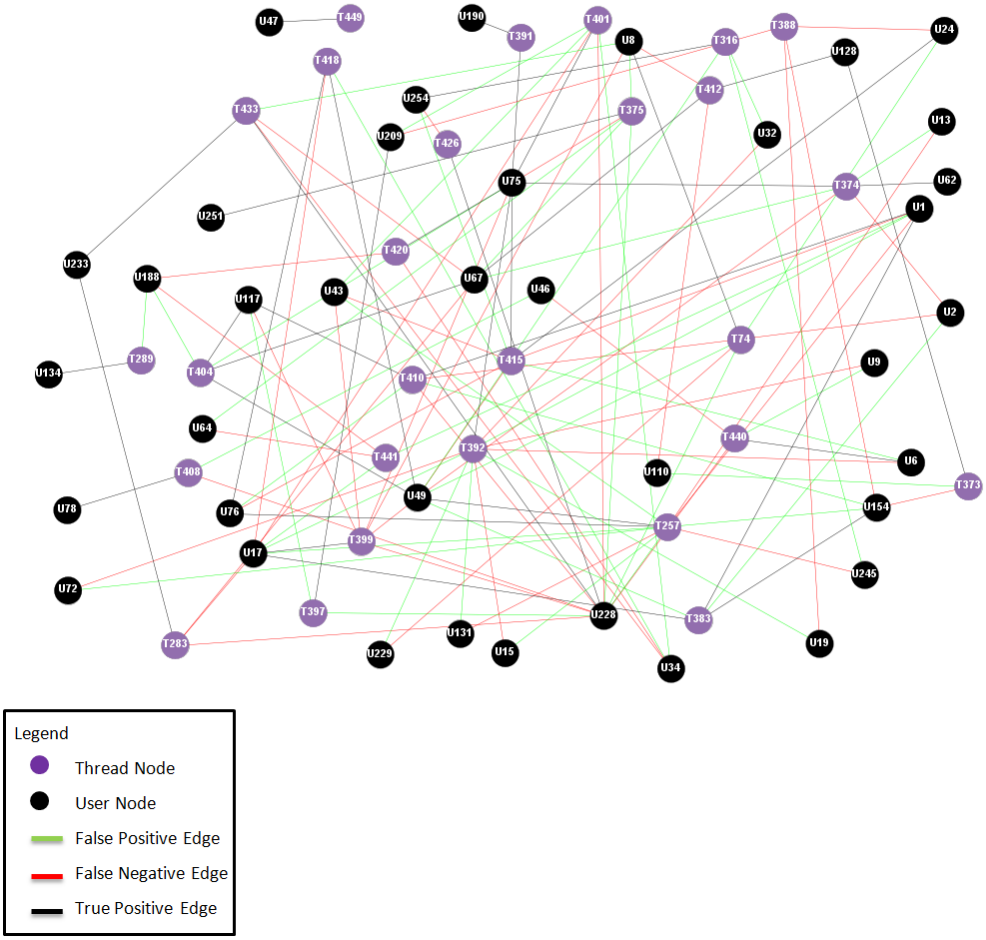

Figura 5.3: Red del Sub-Foro 2 para el Mes 4

Recall	Accuracy	Precision	F-measure
0.632	0.951	0.672	0.652

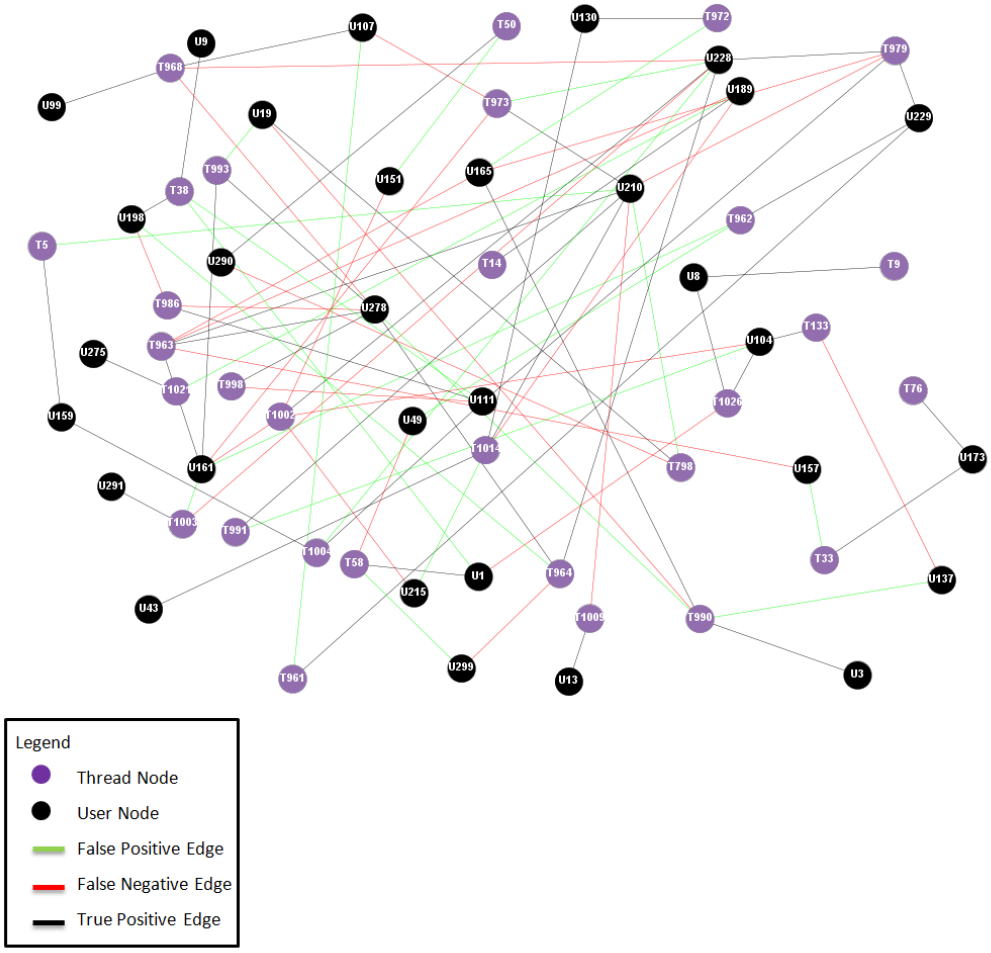

Figura 5.4: Red del Sub-Foro 3 para el Mes 13

Recall	Accuracy	Precision	F-measure
0.410	0.917	0.445	0.427

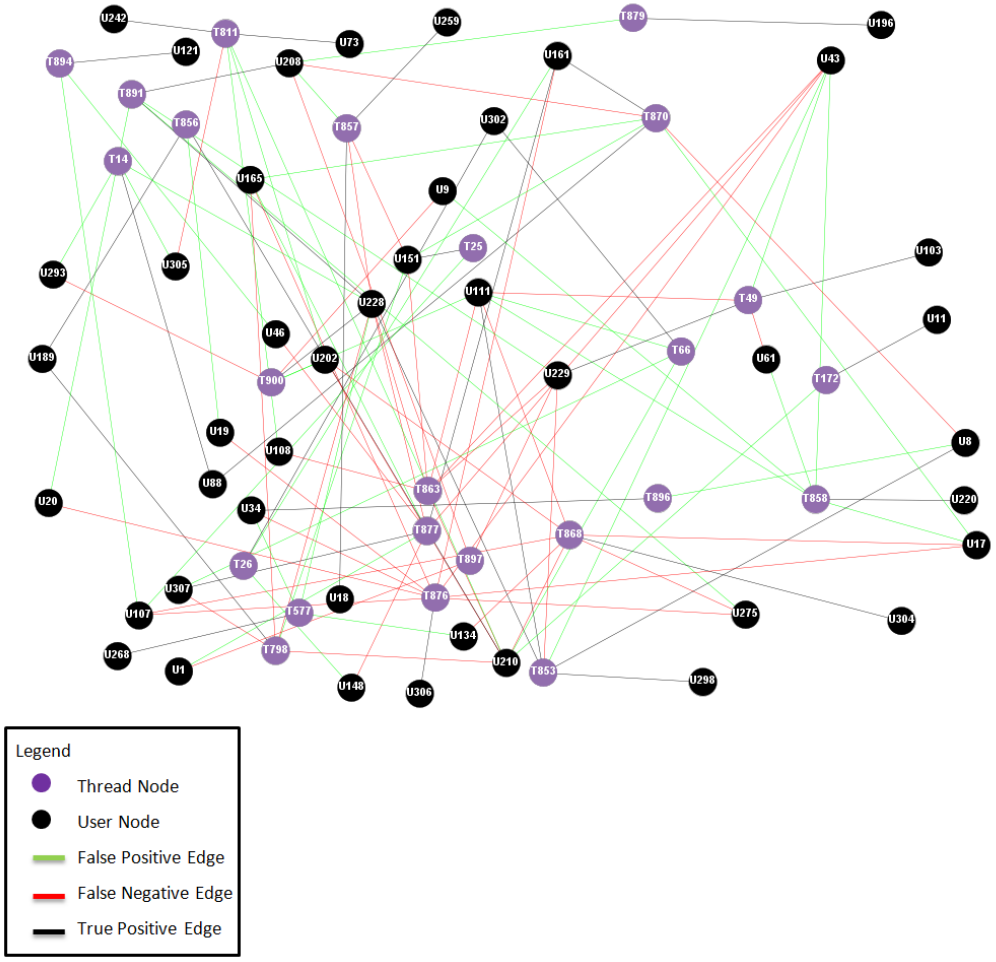

Figura 5.5: Red del Sub-Foro 3 para el Mes 11

Recall	Accuracy	Precision	F-measure
0.454	0.831	0.5	0.476

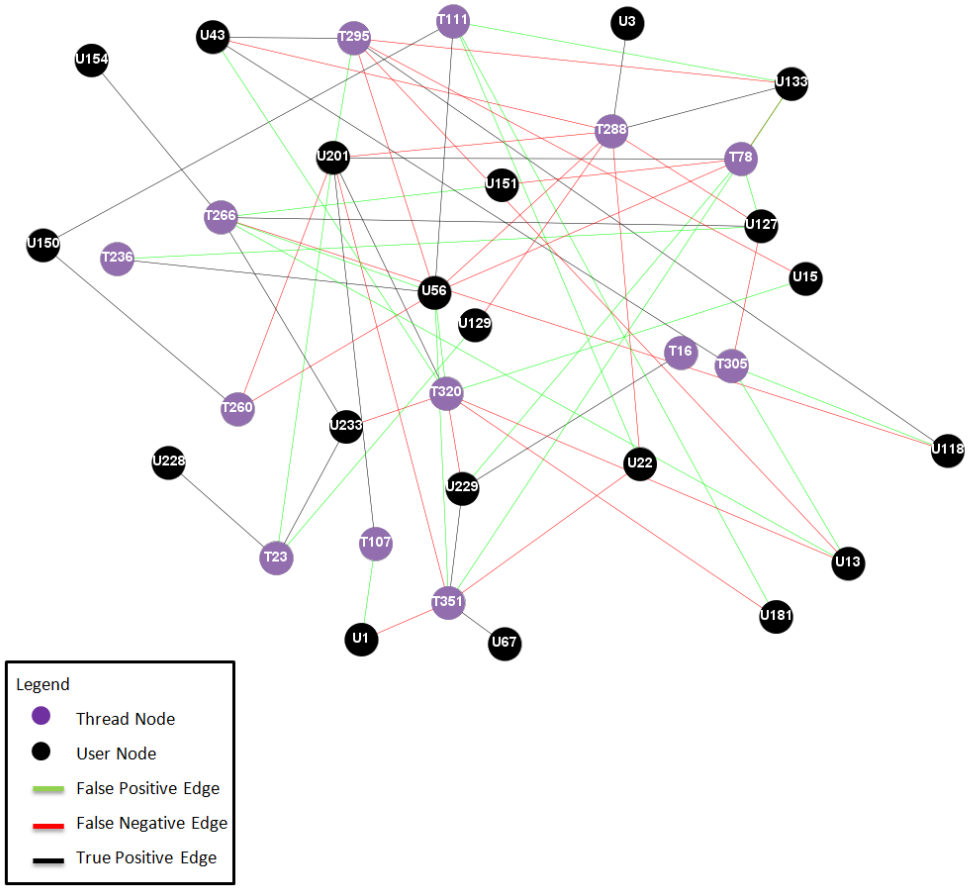

Figura 5.7: Red del Sub-Foro 4 para el Mes 3

Recall	Accuracy	Precision	F-measure
0.674	0.939	0.689	0.681

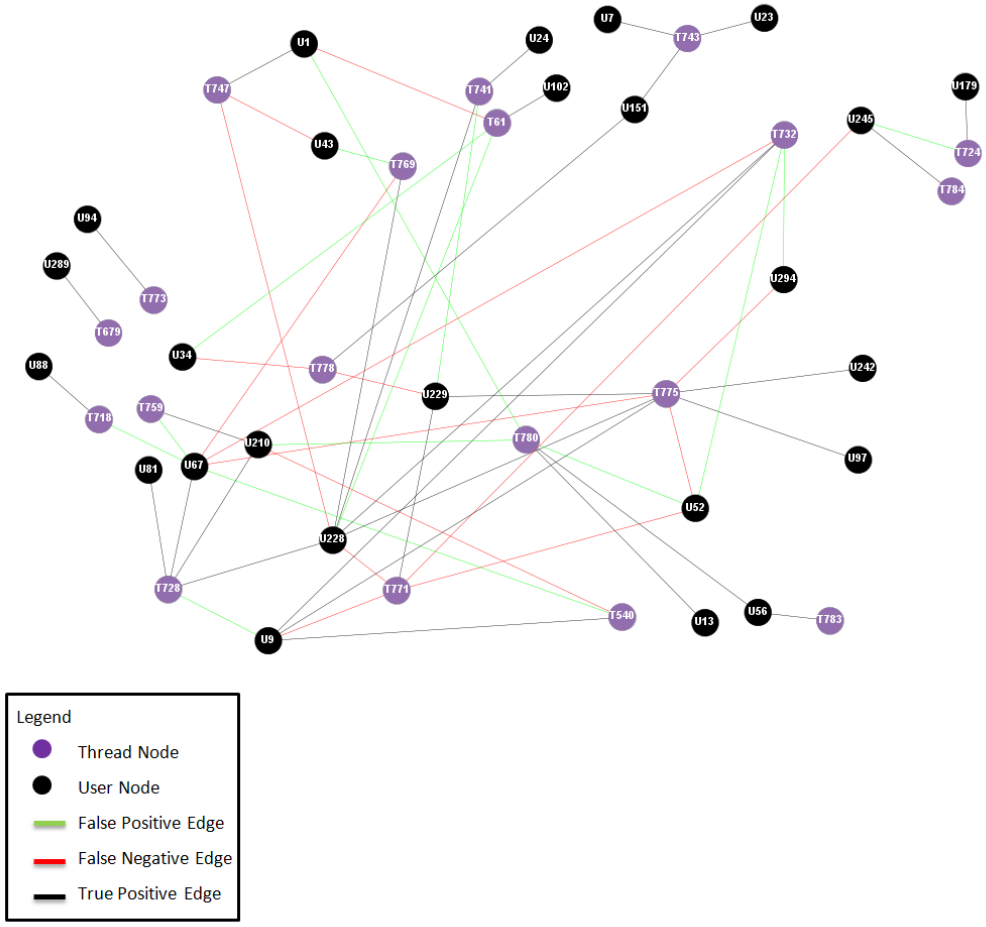

Figura 5.8: Red del Sub-Foro 5 para el Mes 9

Tabla 5.21: Reglas de Decisión de Publicación de Post para Subforo 5 Mes 6. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U1	T527,T550,T552, T565	U72	T461,T552	U158	T555
U2	T501,T548,T550, T569	U73	T550	U161	T90,T243,T501,T522, T527,T540,T541,T549, T550,T551,T555,T565, T569,T578
U3	T548,T578	U84	T131	U163	T544
U8	T36,T523,T541, T550,T569	U86	T131,T527,T535, T552	U178	T552,T555
U9	T131,T522,T523, T535,T536,T537, T540,T544,T550, T551,T561,T565	U97	T491,T540	U179	T131,T429,T491,T550, T552,T555,T561,T565,569, T578
U13	T131,T550	U99	T537,T555,T569, T578	U188	T549
U14	T520,T525	U101	T549,T564	U198	T243,T421,T522,T527,539,T561
U17	T523,T548	U109	T390,T551	U201	T429,T565
U22	T535,T565	U110	T540,T564	U209	T561,T565
U24	T243,T421,T461, T550,T551,T558	U111	T552	U210	T131,T491,T523,T525, T527,T537,T540,T541, T550,T551,T552,T561, T565,T578
U30	T520,T540	U116	T429,T520,T539, T550,T564	U228	T90,T461,T501,T522, T527,T540,T541,T550, T555,T569
U42	T523,T558,T569	U120	T537,T555	U229	T539,T540,T548,T551
U43	T243,T429,T551, T565,T569	U128	T501,T535,T555, T571,T578	U245	T540,T552,T555,T565
U45	T131,T569	U131	T36	U260	T551
U46	T565,T571	U135	T491,T569	U264	T131,T461,T522,T527, T533,T539,T540,T548, T551,T569
U49	T390,T525,T527, T533,T537,T550, T565,T578	U148	T491,T501,T551, T558	U265	T535,T536,T537
U62	T535,T550	U151	T540	U267	T390
U67	T131,T522,T527, T565	U154	T523,T527,T540, T549,T550,T558, T569		

Recall	Accuracy	Precision	F-measure
0.377	0.919	0.402	0.389

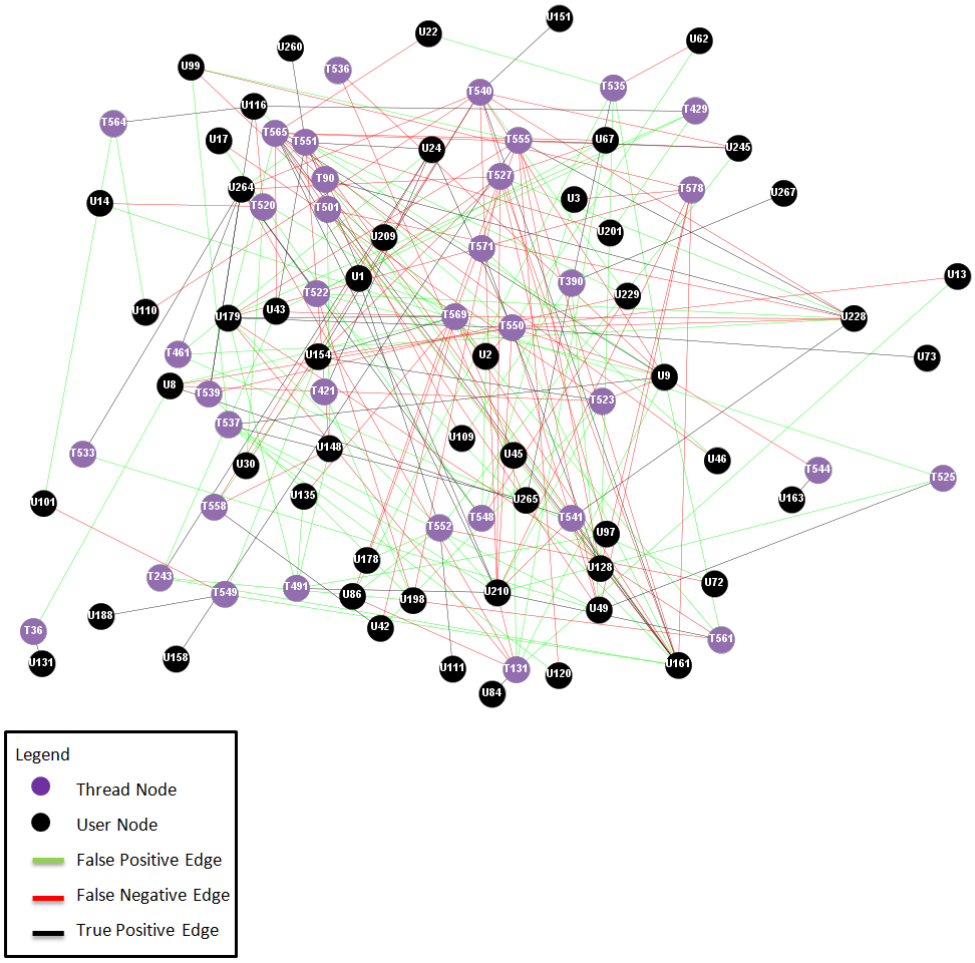

Figura 5.9: Red del Sub-Foro 5 para el Mes 6

Tabla 5.22: Resultados del Sub-Foro 6

Month	**Recall**	**Accuracy**	**Precision**	**F-measure**
2	0.818	0.886	0.818	0.818
3	0.789	0.927	0.833	0.811
4	0.857	0.933	0.857	0.857
5	0.842	0.939	0.842	0.842
6*	-	-	-	-
7	0.842	0.914	0.842	0.842
8	0.800	0.852	0.800	0.800
9	0.895	0.961	0.944	0.919
10	0.947	0.973	0.947	0.947
11	0.846	0.900	0.846	0.846
12	0.842	0.933	0.842	0.842
13	0.647	0.848	0.733	0.688
Mean	**0.830**	**0.915**	**0.846**	**0.837**
Max	**0.947**	**0.973**	**0.947**	**0.947**
Min	**0.647**	**0.848**	**0.733**	**0.688**

Tabla 5.23: Resultados del Sub-Foro 6

Month	**RF**	**SVM**
2	0.43	0.39
3	0.30	0.31
4	0.55	0.45
5	0.30	0.31
6	****	****
7	0.29	0.25
8	0.59	0.61
9	0.32	0.33
10	0.36	0.39
11	0.35	0.33
12	0.60	0.63
13	0.28	0.27
Mean	**0.38**	**0.39**
Max	**0.60**	**0.63**
Min	**0.29**	**0.25**

Tabla 5.24: Reglas de Decisión de Publicación de Post para Subforo 6 Mes 10. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U1	T46,T610,T840	U151	T788	U229	T610
U9	T840	U163	T840	U237	T46
U16	T610	U180	T703,T788	U241	T46
U32	T703	U207	T610	U257	T788
U75	T788	U228	T840	U279	T46,T610,T840

Tabla 5.25: Reglas de Decisión de Publicación de Post para Subforo 6 Mes 13. Usuario = U**, Hilos de conversación en los que el usuario ha publicado posts = T***

User	Posts in:	User	Posts in:	User	Posts in:
U1	T667,T1005	U72	T899	U208	T1005
U9	T1005	U75	T667,T967	U210	T899
U19	T967	U144	T667,T899	U229	T967,T996
U23	T899	U180	T1025		

Recall	Accuracy	Precision	F-measure
0.947	0.973	0.947	0.947

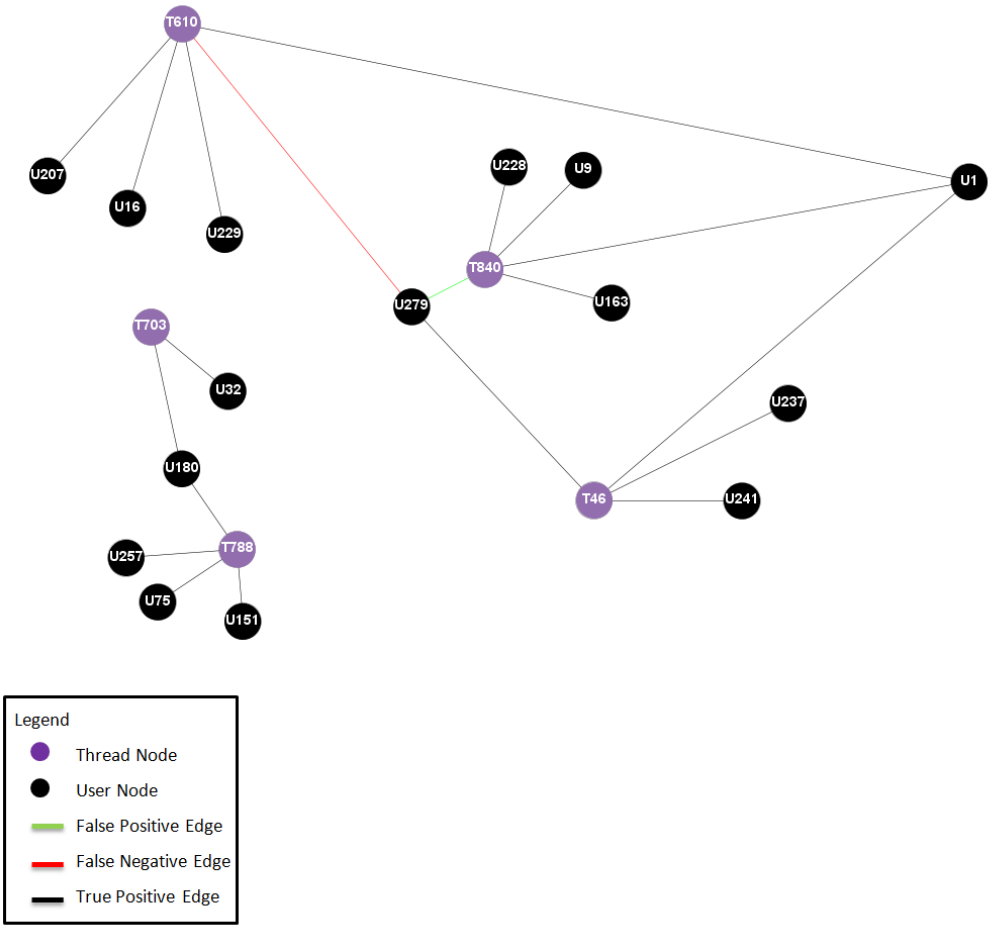

Figura 5.10: Red del Sub-Foro 6 para el Mes 10

Recall	Accuracy	Precision	F-measure
0.647	0.848	0.733	0.688

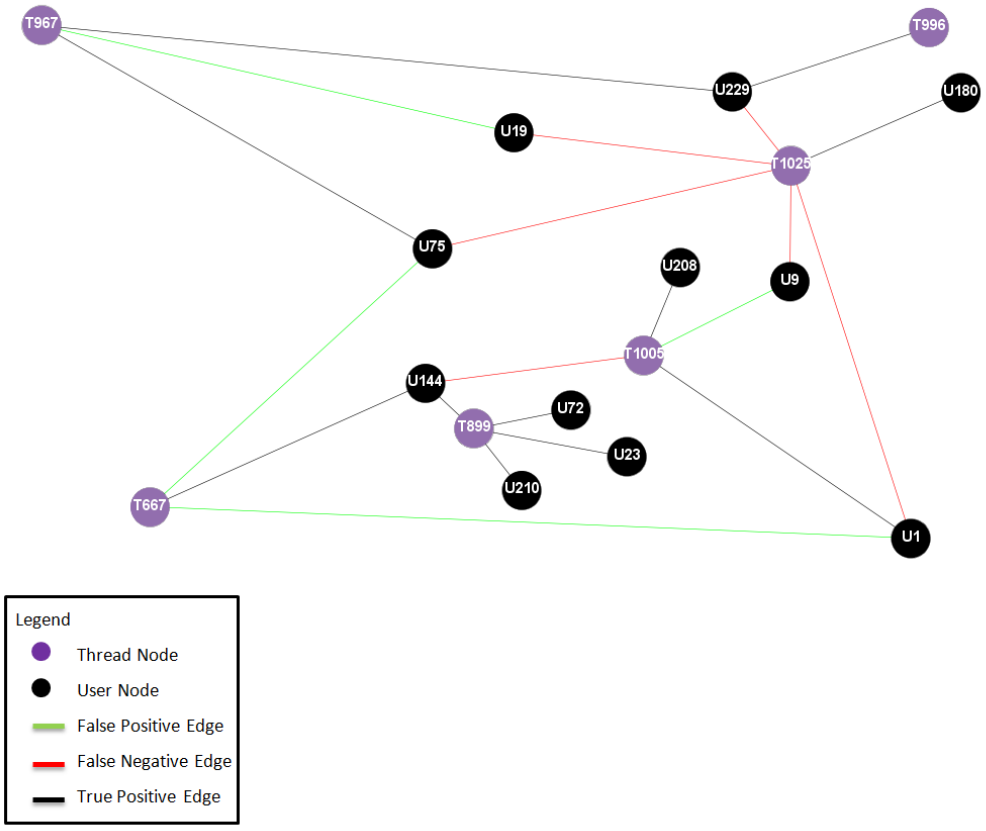

Figura 5.11: Red del Sub-Foro 6 para el Mes 13

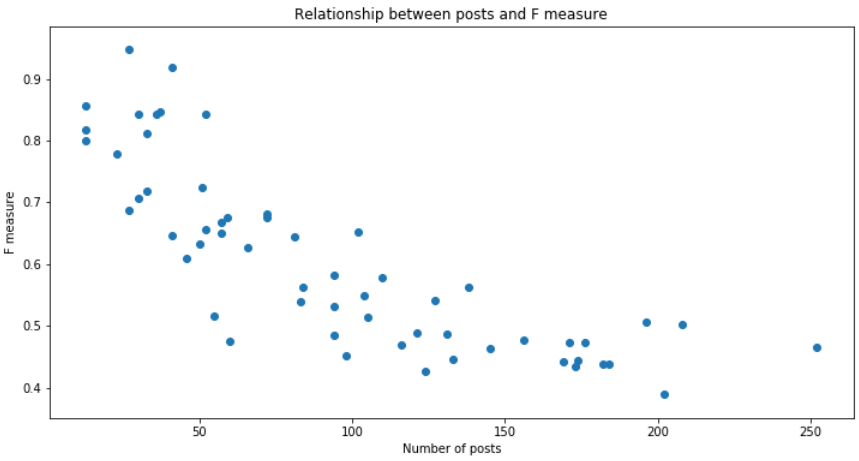

Figura 5.12: Relación entre el número de publicaciones y F-measure score

Capítulo 6

Conclusiones y Trabajo Futuro

En este capítulo recogemos las conclusiones de esta Tesis y damos algunas indicaciones de trabajo futuro.

6.1. Conclusiones

En esta Tesis se estudia el problema de modelar las decisiones de contribución de contenido de los usuarios de una red social en línea. En la literatura se encuentran distintos enfoques para enfrentar este problema, la mayoría de los cuales se realiza a nivel macroscopico o bien mesóscopico. En este trabajo se presenta un modelo neurosemántico de las decisiones de publicación de contenido de los usuarios en un foro web OSN en el nivel microscópico, es decir, el modelo predice la decisión específica de un usuario de publicar un mensaje en un hilo de conversación específico de algún Sub-Foro. Se propone el modelo neuronal extended leaky competing accumulator (ELCA) que implementa la competencia de los diversos hilos de conversación por la atención del usuario como un proceso dinámico. La estimación de los parámetros del

modelo se llevó a cabo mediante un proceso de optimización via un algoritmo genético. Una de las novedades de este trabajo consiste en que se estiman los parámetros del mdoelo LCA a partir de datos con el objetivo de lograr rendimiento predictivo óptimo para la tarea de predicción de generación de contenido de redes sociales. En este aspecto, la literatura revisada contiene enfoques con ajustes cualitativos de los parámetros con el objetvio de estudiar el comportamiento emergente de acuerdo con las teorías de la elección basada en valores. Por otro lado, no se detectaron algunos fenómenos bien conocidos propios de la elección como las inversiones de preferencia. Un análisis más detallado podría descubrir tales fenómenos en nuestro dominio del problema.

La similitud semántica que subyace al mecanismo de atención se modela mediante un análisis de tópicos no supervisado; por lo tanto, está completamente automatizado. Los resultados sobre los datos extraídos de un OSN de la vida real son bastante prometedores. Específicamente, el modelo ELCA mejora en gran medida con respecto a los enfoques estándar de aprendizaje automático, a saber, Random Forest (RF) y Support Vector Machines (SVM), que utilizan el mismo tipo de información semántica como características de entrada. La puntuación F mejor y media del modelo ELCA fue 0,95 y 0,61, respectivamente, mientras que para RF y SVM la puntuación F mejor fue 0,60 y 0,63, respectivamente, y la puntuación F media fue 0,19 y 0,21, respectivamente.

Finalmente, se valida la hipótesis de investigación planteada al mostrar que el modelo ELCA, un modelo que ocupa el contenido semántico de las publicaciones, es capaz de modelar las decisiones de contribución de contenido de los usuarios en una red social (foro web) de manera exitosa.

6.2. Trabajo futuro

El trabajo futuro se puede subdividir en 4 aristas principalmente.

Primeramente, es de fundamental importancia investigar el uso de enfoques de máxima verosimilitud para la estimación de parámetros del modelo LCA para poder contrastar con el enfoque utilizado en este trabajo.

En segundo lugar, otra área de investigación bastante interesante es la métrica del espacio temático. El trabajo futuro podría dirigirse a la definición de una distancia adecuada entre representaciones vectoriales de texto multitemáticas que permiten la extracción del contenido más valioso generado por los usuarios. Además, el enfoque desarrollado en este trabajo podría combinarse con otros métodos existentes que capturan características topológicas de la red buscando una mejora en el rendimiento de predicción por un sistema híbrido de este tipo.

En tercer puesto, se hace necesaria una exploración más profunda de los fundamentos de los algoritmos de procesamiento del lenguaje natural (PLN) para lograr una captura mas fidedigna del significado real de los textos posteados por los usuarios del foro web. Uno de los obstaculos a superar es el uso de enfoques frecuentistas para modelar la ocurrencia conjunta de palabras en un documento [66]. Dentro de las posibilidades a explorar se encuentra la utilización de word embeddings lo que permitiría capturar relaciones ocultas entre las palabras a cambio de sacrificar poder interpretativo. Por otro lado, la creación automática de ontologías para un dominio específico también sería una alternativa viable para abordar este problema.

Por último, como se mencionó en la sección 5.4 aplicar medidas para lidiar con un dataset con clases desbalanceadas resulta de interés puesto que podría conducir a la obtención de mejores resultados. El desafío en este caso radica en idear una estrategia que se pueda adaptar al tipo de datos con los que se

esta trabajando en esta Tesis, es decir, datos de publicaciones en Sub-Foros.

Bibliografía

[1] George A Miller. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." En: *Psychological review* 63.2 (1956), pág. 81.

[2] Gerard Salton, Anita Wong y Chung-Shu Yang. "A vector space model for automatic indexing". En: *Communications of the ACM* 18.11 (1975), págs. 613-620.

[3] FM Bass. "A new product growth model for consumer durables, Mathematical Models in Marketing". En: *Lecture Notes in Economics and Mathematical Systems* 132 (1976), págs. 351-253.

[4] Mark Granovetter. "Threshold models of collective behavior". En: *American journal of sociology* 83.6 (1978), págs. 1420-1443.

[5] James L McClelland. "Toward a theory of information processing in graded, random, and interactive networks." En: (1993).

[6] Heinz Mühlenbein y Dirk Schlierkamp-Voosen. "Predictive models for the breeder genetic algorithm i. continuous parameter optimization". En: *Evolutionary computation* 1.1 (1993), págs. 25-49.

[7] Colin R Reeves. "Genetic algorithms and neighbourhood search". En: *AISB Workshop on Evolutionary Computing*. Springer. 1994, págs. 115-130.

[8] Mandavilli Srinivas y Lalit M Patnaik. "Genetic algorithms: A survey". En: *computer* 27.6 (1994), págs. 17-26.

[9] Barry Wellman y col. "Computer networks as social networks: Collaborative work, telework, and virtual community". En: *Annual review of sociology* 22.1 (1996), págs. 213-238.

[10] Barry Wellman y Milena Gulia. "Virtual communities as communities". En: *Communities in cyberspace* (1999), págs. 167-194.

[11] Amy Jo Kim. *Community building on the web: Secret strategies for successful online communities*. Addison-Wesley Longman Publishing Co., Inc., 2000.

[12] Lada A Adamic y col. "Search in power-law networks". En: *Physical review E* 64.4 (2001), pág. 046135.

[13] Jacob Goldenberg, Barak Libai y Eitan Muller. "Talk of the network: A complex systems look at the underlying process of word-of-mouth". En: *Marketing letters* 12.3 (2001), págs. 211-223.

[14] Christopher M Johnson. "A survey of current research on online communities of practice". En: *The internet and higher education* 4.1 (2001), págs. 45-60.

[15] Marius Usher y James L McClelland. "The time course of perceptual choice: the leaky, competing accumulator model." En: *Psychological review* 108.3 (2001), pág. 550.

[16] Barry Wellman. "Computer networks as social networks". En: *Science* 293.5537 (2001), págs. 2031-2034.

[17] David M Blei, Andrew Y Ng y Michael I Jordan. "Latent dirichlet allocation". En: *Journal of machine Learning research* 3.Jan (2003), págs. 993-1022.

[18] David Kempe, Jon Kleinberg y Éva Tardos. "Maximizing the spread of influence through a social network". En: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2003, págs. 137-146.

[19] Thomas L Griffiths y Mark Steyvers. "Finding scientific topics". En: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), págs. 5228-5235.

[20] Aron Culotta, Ron Bekkerman y Andrew McCallum. *Extracting social networks and contact information from email and the web*. Inf. téc. MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2005.

[21] Eyal Even-Dar y Asaf Shapira. "A note on maximizing the spread of influence in social networks". En: *International Workshop on Web and Internet Economics*. Springer. 2007, págs. 281-286.

[22] Joshua I Gold y Michael N Shadlen. "The neural basis of decision making". En: *Annual review of neuroscience* 30 (2007).

[23] Masao Kubo y col. "The possibility of an epidemic meme analogy for web community population analysis". En: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2007, págs. 1073-1080.

[24] David Mimno, Hanna Wallach y Andrew McCallum. "Community-based link prediction with text". En: *Proc. of NIPS*. 2007.

[25] Sebastián A Rıos. "A study on web mining techniques for off-line enhancements of web sites". Tesis doct. 2007.

[26] Xiaodan Song y col. "Information flow modeling based on diffusion rate for prediction and ranking". En: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, págs. 191-200.

[27] Dongshan Xing y Mark Girolami. "Employing Latent Dirichlet Allocation for fraud detection in telecommunications". En: *Pattern Recognition Letters* 28.13 (2007), págs. 1727-1734.

[28] Loulwah AlSumait, Daniel Barbará y Carlotta Domeniconi. "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking". En: *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE. 2008, págs. 3-12.

[29] X. H. Phang y CT. Nguyen. "Gibbslda++". En: (2008).

[30] Haibo Hu y Xiaofan Wang. "Evolution of a large online social network". En: *Physics Letters A* 373.12-13 (2009), págs. 1105-1110.

[31] Sebastián A Ríos, Felipe Aguilera y Luis A Guerrero. "Virtual communities of practice's purpose evolution analysis using a concept-based mining approach". En: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer. 2009, págs. 480-489.

[32] Noga Alon y col. "A note on competitive diffusion through social networks". En: *Information Processing Letters* 110.6 (2010), págs. 221-225.

[33] Héctor Alvarez. "Detección de miembros clave en una comunidad virtual de prática mediante análisis de redes sociales y mineria de datos avanzada". En: *Master's thesis, University of Chile* (2010).

[34] Héctor Alvarez y col. "Enhancing social network analysis with a concept-based text mining approach to discover key members on a virtual community of practice". En: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer. 2010, págs. 591-600.

[35] GASTON ANDRÉS L'HUILLIER CHAPARRO y col. "CLASIFICACION DE PHISHING UTILIZANDO MINERÍA DE DATOS ADVERSARIAL Y JUEGOS CON INFORMACION INCOMPLETA". En: (2010).

[36] Maksim Kitsak y col. "Identification of influential spreaders in complex networks". En: *Nature physics* 6.11 (2010), pág. 888.

[37] Phillippa Lally y col. "How are habits formed: Modelling habit formation in the real world". En: *European journal of social psychology* 40.6 (2010), págs. 998-1009.

[38] Eduardo Merlo y col. "Finding inner copy communities using social network analysis". En: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer. 2010, págs. 581-590.

[39] Mohammad Al Hasan y Mohammed J Zaki. "A survey of link prediction in social networks". En: *Social network data analytics*. Springer, 2011, págs. 243-275.

[40] Phil E Brown y Junlan Feng. "Measuring user influence on twitter using modified k-shell decomposition". En: *Fifth international AAAI conference on weblogs and social media*. 2011.

[41] Lautaro Cuadra, Sebastín A Rios y Gaston L'Huillier. "Enhancing community discovery and characterization in vcop using topic models". En: *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03*. IEEE Computer Society. 2011, págs. 326-329.

[42] Conrad Lee, Thomas Scherngell y Michael J Barber. "Investigating an online social network using spatial interaction models". En: *Social Networks* 33.2 (2011), págs. 129-133.

[43] Gastón L'huillier y col. "Topic-based social network analysis for virtual communities of interests in the dark web". En: *ACM SIGKDD Explorations Newsletter* 12.2 (2011), págs. 66-73.

[44] Jiyoung Woo, Jaebong Son y Hsinchun Chen. "An SIR model for violent topic diffusion in social media". En: *Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on*. IEEE. 2011, págs. 15-19.

[45] Rakesh Kumar. "Blending roulette wheel selection & rank selection in genetic algorithms". En: *International Journal of Machine Learning and Computing* 2.4 (2012), pág. 365.

[46] Lin Li y col. "Phase transition in opinion diffusion in social networks". En: *Acoustics, speech and signal processing (ICASSP), 2012 IEEE international conference on*. IEEE. 2012, págs. 3073-3076.

[47] Seth A Myers, Chenguang Zhu y Jure Leskovec. "Information diffusion and external influence in networks". En: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2012, págs. 33-41.

[48] Pablo E Román, Miguel E Gutiérrez y Sebastián A Rios. "A model for content generation in On-line social network." En: *KES*. 2012, págs. 756-765.

[49] Reiko Takehara, Masahiro Hachimori y Maiko Shigeno. "A comment on pure-strategy Nash equilibria in competitive diffusion games". En: *Information processing letters* 112.3 (2012), págs. 59-60.

[50] Jiyoung Woo y Hsinchun Chen. "An event-driven SIR model for topic diffusion in web forums". En: *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*. IEEE. 2012, págs. 108-113.

[51] Fei Xiong y col. "An information diffusion model based on retweeting mechanism for online social media". En: *Physics Letters A* 376.30-31 (2012), págs. 2103-2108.

[52] Adrien Guille y col. "Information diffusion in online social networks: A survey". En: *ACM Sigmod Record* 42.2 (2013), págs. 17-28.

[53] Adrien Guille y col. "Sondy: An open source platform for social dynamics mining and analysis". En: *Proceedings of the 2013 ACM SIGMOD international conference on management of data*. ACM. 2013, págs. 1005-1008.

[54] Jianwei Niu y col. "An Empirical Study of a Chinese Online Social Network–Renren". En: *Computer* 46.9 (2013), págs. 78-84.

[55] Lucy Small y Oliver Mason. "Information diffusion on the iterated local transitivity model of online social networks". En: *Discrete Applied Mathematics* 161.10-11 (2013), págs. 1338-1344.

[56] Lucy Small y Oliver Mason. "Nash equilibria for competitive information diffusion on trees". En: *Information Processing Letters* 113.7 (2013), págs. 217-219.

[57] Chunxiao Jiang, Yan Chen y KJ Ray Liu. "Modeling information diffusion dynamics over social networks". En: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE. 2014, págs. 1095-1099.

[58] Sebastián A Ríos y Ricardo Muñoz. "Content patterns in topic-based overlapping communities". En: *The Scientific World Journal* 2014 (2014).

[59] Ye Sun y col. "Epidemic spreading on weighted complex networks". En: *Physics Letters A* 378.7-8 (2014), págs. 635-640.

[60] Li-Jen Kao y Yo-Ping Huang. "Mining influential users in social network". En: *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*. IEEE. 2015, págs. 1209-1214.

[61] Chuan Luo, Xiaolong Zheng y Daniel Zeng. "Inferring social influence and meme interaction with Hawkes processes". En: *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*. IEEE. 2015, págs. 135-137.

[62] Akrati Saxena, SRS Iyengar y Yayati Gupta. "Understanding spreading patterns on social networks based on network topology". En: *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE. 2015, págs. 1616-1617.

[63] Anupriya Shukla, Hari Mohan Pandey y Deepti Mehrotra. "Comparative review of selection techniques in genetic algorithm". En: *Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on*. IEEE. 2015, págs. 515-519.

[64] John Breslin Tope Omitola Ríos Sebastián. *Social Semantic Web Intelligence*. Morgan & Claypool Publishers, 2015.

[65] Yang Yang, Ryan N Lichtenwalter y Nitesh V Chawla. "Evaluating link prediction methods". En: *Knowledge and Information Systems* 45.3 (2015), págs. 751-782.

[66] Constanza Contreras-Piña y Sebastián A Ríos. "An empirical comparison of latent sematic models for applications in industry". En: *Neurocomputing* 179 (2016), págs. 176-185.

[67] Dong Li y col. "Exploiting information diffusion feature for link prediction in sina weibo". En: *Scientific reports* 6 (2016), pág. 20058.

[68] Xiaoyan Qiu y col. "Effects of time-dependent diffusion behaviors on the rumor spreading in social networks". En: *Physics Letters A* 380.24 (2016), págs. 2054-2063.

[69] Jiyoung Woo y Hsinchun Chen. "Epidemic model for information diffusion in web forums: experiments in marketing exchange and political dialog". En: *SpringerPlus* 5.1 (2016), pág. 66.

[70] Ying Hu, Rachel Jeungeun Song y Min Chen. "Modeling for information diffusion in online social networks via hydrodynamics". En: *IEEE Access* 5 (2017), págs. 128-135.

[71] Sebastián A Ríos y col. "Semantically enhanced network analysis for influencer identification in online social networks". En: *Neurocomputing* (2017).

[72] Hadi Shakibian y Nasrollah Moghadam Charkari. "Mutual information model for link prediction in heterogeneous complex networks". En: *Scientific Reports* 7 (2017), pág. 44981.

[73] Jiawei Zhang y col. "Link prediction with cardinality constraint". En: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM. 2017, págs. 121-130.

[74] Ricardo Baeza-Yates. "Bias on the web". En: *Communications of the ACM* 61.6 (2018), págs. 54-61.