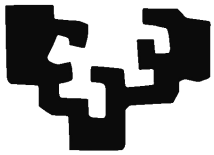

Modelando las decisiones de contribución de contenido de usuarios en una red social

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

Comprender a nivel microscópico el proceso de generación de contenidos en una red social en línea (OSN) es muy deseable para una mejor gestión de las OSN y la prevención de fenómenos indeseables, como el acoso en línea. La generación de contenido, es decir, la decisión de publicar un contenido aportado en la OSN, puede ser modelado por enfoques neurofisiológicos sobre la base de análisis semántico insesgado de los contenidos ya publicados en la OSN. Esta Tesis propone un modelo neuro-semántico compuesto por (1) un modelo que hemos denominado *Extended Leaky Competing Accumulator* (ELCA) como la arquitectura neuronal que implementa el proceso de decisión concurrente del usuario para generar contenido en un hilo de conversación de una comunidad virtual de práctica, y (2) un modelado semántico basado en el análisis de tópicos realizado por un *Latent Dirichlet Allocation* (LDA) de usuarios e hilos de conversación. Se utiliza la similitud entre las representaciones semánticas del usuario y del hilo para construir el modelo del interés del usuario en los contenidos del hilo como estímulo para aportar contenido nuevo en el hilo. Los intereses semánticos de los usuarios en los hilos de discusión son las entradas externas para el ELCA, es decir, el valor externo asignado a cada elección. Demostramos este proceso de modelado

y sus resultados sobre un conjunto de datos extraído de un foro web de la vida real dedicado a los fanáticos de los retoques con instrumentos musicales y dispositivos relacionados. El modelo neurosemántico logra un alto rendimiento al predecir las decisiones de publicación de contenido (puntuación F promedio de 0.61) mejorando en gran medida con respecto al desempeño de enfoques conocidos de *machine learning*, a saber, *random forest* y *support vector machines* que sólo alcanzan puntuaciones F promedio de 0,19 y 0,21, respectivamente.

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

Para Myriam.

Agradecimientos

Mis agradecimiento a todas las personas que de alguna u otra forma contribuyeron a este proceso, en particular:

A mi novia Karla por ser mi compañera y fuente de motivación y fortaleza.

A mi familia por su apoyo y aliento.

A mi madrina Ernestina y a su esposo Cristián por sus consejos y por impulsarme a mejorar constantemente.

A mis tutores Manuel Graña y Sebastián Ríos por todo el compromiso, paciencia y enseñanzas durante todo este proceso.

Por último, agradecer a las entidades que contribuyeron al financiamiento de este trabajo de esta tesis: 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 aplicada para el desarrollo de estos trabajos de investigación, se presenta el caso de estudio sobre el que se han realizado las demostraciones prácticas. Finalmente, se presentan 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

Hay una gran cantidad de literatura de investigación que trata sobre diversos aspectos del análisis de redes sociales en línea (OSN). Los esfuerzos dentro de la investigación clásica se dedican a identificar comunidades dentro de la red [48, 50, 67], encontrar personas influyentes o miembros clave de la comu-

nidad virtual [41, 47, 71, 22, 43, 72, 89, 31], o describir la evolución de redes específicas [37, 62, 38]. Sin embargo, existe muy poco o ningún trabajo sobre el proceso de decisión que lleva a un usuario a publicar algún contenido en la OSN, por ejemplo, publicar un mensaje en un foro de una comunidad virtual de práctica (VCoP). Una VCoP implementado como un foro web es un lugar virtual donde los miembros interactúan, discuten ideas, comparten y generan conocimiento sobre temas específicos organizados en subforos e hilos de discusión. La generación de contenido es un proceso radicalmente diferente al proceso de propagación dentro de una OSN que siguen a la publicación de algún contenido nuevo. Por ejemplo, publicar un tuit es radicalmente diferente a retuitear, compartir, gustar o cualquier otro proceso de propagación que difunda la influencia del contenido original del tweet. La generación de contenido sintético, como los modelos Markovianos de n-gramas que permiten generar tweets falsos que son difíciles de distinguir por los humanos [97], están fuera del alcance de este trabajo.

La decisión de contribuir con una publicación a un hilo de discusión dentro de una VCoP es un fenómeno afectado por múltiples factores como el conocimiento del usuario del tema, sus preferencias, otros usuarios participantes de la discusión, e incluso, la calidad de la información presentada, entre otros factores. Este proceso de decisión se puede modelar por la competención entre varios hilos de discusión en curso simultáneamente por ganar la atención del usuario, es decir, el usuario selecciona el hilo ganador para publicar una contribución. Esta competención puede ser modelada por un modelo de elección neurofisiológico. Específicamente, es relevante el modelo *Leaky Competing Accumulator* (LCA) [25, 19, 57], donde la actividad de las neuronas computacionales es impulsada por un conjunto de ecuaciones diferenciales que acumulan las contribuciones inhibidoras de otras neuronas, las contribuciones excitatorias de las unidades de entrada y las fluctuaciones de una fuente independiente de ruido blanco. Se ha demostrado que el LCA da cuenta con éxito de la distribución del tiempo de reacción observada

empíricamente en experimentos psicofísicos. Específicamente, para algunas combinaciones de parámetros de inhibición y valores de decaimiento, se ha demostrado que el LCA reproduce las violaciones del valor esperado observadas empíricamente y las reversiones de preferencia informadas en muchos experimentos sobre elección preferencial basada en valores. Estos estudios se centran en la distribución del tiempo de decisión para una tasa de error fija después de muchas repeticiones de la ejecución de LCA, tratando de imitar las distribuciones encontradas empíricamente. Los parámetros LCA se ajustan a mano (o se exploran con un *grid-search*) para encontrar los valores que reproducen el comportamiento en el tiempo de respuesta deseado y la tasa de error de elección esperada entendida como la opción con el menor valor.

En esta Tesis, el enfoque es más parecido a los enfoques de *machine learning* para modelar el proceso de decisión, es decir, usamos LCA como modelo de toma de decisiones cuyo rendimiento se mide por la precisión en la predicción de la decisiones tomadas por los usuarios de publicar una contribución de contenido a un hilo de conversación específico donde el valor semántico asignado al hilo de conversación se trata como una constante de entrada.

Para nuestro trabajo específico, proponemos un modelo LCA extendido (ELCA) en varios aspectos.

- En primer lugar, el modelo incluye muchas elecciones simultáneos de muchos usuarios, mientras que el LCA clásico considera un solo agente y un pequeño número de opciones.
- En segundo lugar, utilizamos el modelado semántico de usuarios e hilos para calcular el valor semántico excitatorio de entrada para cada cada opción, por lo tanto se vincula la valoración abstracta de las opciones con las evidencias del dominio concreto relacionado.
- En tercer lugar, implementamos una búsqueda mediante algoritmo

genético para la calibración óptima de los parámetros del modelo ELCA (también conocido como entrenamiento) utilizando datos de las decisiones de contribución de contenido en un VCoP de la vida real.

La calibración de los parámetros del LCA, realizada como la inducción de los parámetros del modelo a partir de la simulación de trayectorias del acumulador, ha sido reconocido como un problem abierto difícil [88], que ha sido abordado mediante la explotación de simetrías de Lie para una formulación modificada de las ecuaciones del LCA [111] para el caso más sencillo, y que es inaplicable en nuestro caso.

Al contrario de estos enfoques, buscamos los parámetros ELCA óptimos que reproducen las decisiones reales de los usuarios después de la convergencia de la simulación. Sin embargo, nuestro trabajo no intenta estudiar o reproducir fenómenos de elección humanos, como la inversión de preferencias, que son los dominio de estudio originales del modelo LCA [25, 19, 57].

Otro de los aspectos importantes de las contribuciones realizadas en esta Tesis es el uso del análisis semántico de los contenidos para incluirlo en el modelado cuantitativo de las decisiones. El análisis semántico del contenido publicado en OSN es una área de investigación, que actualmente está de moda, que permite detectar y prevenir usos no deseados de la OSN. Por ejemplo, el análisis semántico a nivel de palabras ha sido reportado como capaz de detectar el ciberacoso [94], ayuda a la hora de detectar tweets borrachos [105], y la edad de los usuarios [110]. Además, el análisis de contenido de los posts de las redes sociales permite predecir los niveles de depresión [82]. Específicamente, usamos análisis de tópicos no supervisado realizado mediante *Latent Dirichlet Allocation* (LDA) [21] para el modelado semántico del contenido publicado en la OSN, que permite construir representaciones semánticas vectoriales cuantitativas tanto de usuarios como de hilos de conversación, no muy diferente al modelo semántico neurobiológico social basado

en el conocimiento conceptual [102]. LDA es una poderosa herramienta que se ha utilizado para resumir y construir modelos de redes de contenidos, como gráfos semánticos que relacionan publicaciones sobre COVID-19 [101].

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 modelado 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 las decisiones de contribución de contenido de parte de los usuarios a nivel microscópico utilizando el contenido de texto extraído mediante técnicas de minería de texto y modelado semántico

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

1. Revisar la literatura en busca de enfoques útiles para abordar el problema y crear un punto de referencia de modelos.

2. Desarrollar un modelo que capture la generación de las decsiones de contribución de contenido en foros web.

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 análisis de redes sociales (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 [41].

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

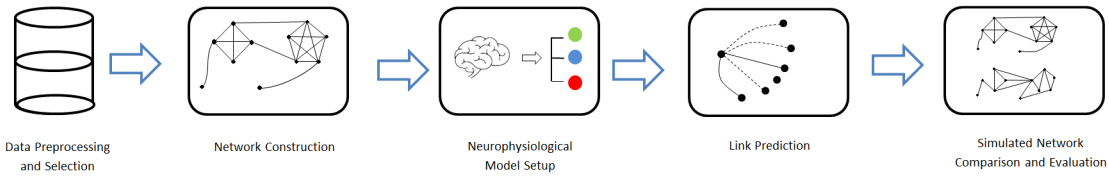

Figura 1.1: Metodología Propuesta para Modelar la Contribución de Contenido en Foros Web.

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

1. **Trabajos Relacionados**
   Para el desarrollo de este trabajo de tesis, se revisará la literatura asociada al análisis de redes sociales (SNA), poniendo énfasis en los trabajos relacionados a enfrentar el problema de la difusión de información, predicción de formación de arcos y modelamiento semántico. Esto con el objetivo de conocer lo que está hecho y poder encontrar enfoques de los cuales poder extraer inspiración para nuestro trabajo.

2. **Representaciones de Grafos**
   Las representaciones de las redes sociales pueden ser diferentes dependiendo de las características del problema de 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 modelar las contribuciones de contenido.

3. **Personalización de algoritmos**
   Se propondrá un modelo basado en el modelo LCA adaptado para recibir información del contenido semántico y adecuado para abordar el problema de modelar las decisiones de contribución de contenido de los usuarios de un foro web. Además, se debe implementar un esquema de optimización para calibrar los parámetros del modelo.

4. **Simulación de la red**
   El algoritmo se probará sobre datos provenientes de un foro web real. Primero, utilizando los datos del trabajo previo realizado por Ríos et al [38, 41, 48, 55, 89]. Se realizará la modificación 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 modelo propuesto

5. **Análisis de resultados y conclusiones**
   Una vez implementada la metodología propuesta, sus resultados serán evaluados mediante el uso de 4 métricas diferentes. También se evaluarán los resultados de 2 modelos de machine learning en la métrica más representativa del rendimiento para realizar un analisis comparativo entre los modelos. 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 a nivel microscópico 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**: donde se presenta una revisión de los trabajos relacionados y algunos elementos relevantes del estado del arte.
- **Capítulo 3**: donde se describe el foro web real, llamado Plexilandia, del que se han extraido los datos para el estudio y el procesamiento de los datos.
- **Capítulo 4**: donde se presenta el corazón de la contribución de esta Tesis, que es la formulación del modelo ELCA y el proceso computacional para implementar el modelo utilizando los resultados del modelado semántico de los datos extraidos de la red social.
- **Capítulo 5**: donde se dan los detalles de los experimentos realizados y sus resultados son explicados.
- **Capítulo 6**: donde 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. Nos centraremos en trabajos que modelan la difusión de información y la predicción de los arcos del grafo que modela la estructura de la red social, ya que para la formulación del modelo propuesto se rescató características de trabajos de estas áreas.

2.1. Machine learning

Machine learning es una disciplina, dentro de la *Inteligencia artificial*, dedicada al estudio y construcción de en un conjunto de técnicas y herramientas que permiten aprender de los datos para la toma de decisiones, ya sea mediante clasificaciones o predicciones cuantitativas, sin la necesidad de ser programadas de manera explicita para lograrlo.

Dentro de las técnicas que contempla *Machine learning*, para este estudio nos son de interés dos de ellas, a saber, *Random Forest* (RF) y *Support Vector*

Machine (SVM) de las que se dará una breve descripción. En esta Tesis Doctoral se usan estos dos modelos para compararar su rendimiento con el del modelo propuesto.

- *Random Forests* [14] son familias de modelos que se caracterizan por consistir de un conjunto de árboles de decisión los cuales son entrenados y luego mediante una combinación de los resultados de los árboles individuales se llega al resultado entregado por el RF. Esta familia de modelos pertenece a los modelos de aprendizaje supervisado y pueden ser usados para realizar regresiones y clasificaciones.
- *Support Vector Machines* [15] son una familia de modelos de aprendizaje supervisado que permiten realizar tanto clasificaciones como regresiones. Están basados fundamentalmente en la idea de separar clases de datos mediante el establecimiento de un borde de decisión que logre la mayor separación o margen entre dichas clases.

2.2. Análisis Redes sociales

Gran parte de la literatura sobre análisis dinámico de las redes sociales en linea (OSN) se ha centrado en la propagación de información a través de la red, la detección de comunidades [48, 50, 67], de usuarios clave y de *influencers* [22, 31, 41, 43, 47, 71, 72, 89]. Otra aproximación frecuentemente tomada por los investigadores [37, 38, 62] ha sido la de describir la evolución de ciertas redes. Por ejemplo, Jianwei Niu et al [62] lleva a cabo un análisis descriptivo de la OSN Renren, la cual es una OSN China parecida a Facebook. En su trabajo se describen las propiedades evolutivas y estructurales macroscópicas de la red. Las conclusiones obtenidas por su investigación guardan coherencia con otros trabajos realizado en redes similares. Para nuestros fines, resulta de interes indagar en la parte de la literatura que trata sobre la propagación

de información, puesto que está en muchos aspectos se asemeja al problema que se enfrenta en este trabajo.

2.3. Difusión de Información

La Tabla 2.1 ofrece un resumen no exhaustivo de los trabajos encontrados en la literatura desde 2007. 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 [81]. Guille et al [60] 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.

Existen dos líneas principales de investigación sobre modelos de difusión de información en redes [87], a saber, la explicativa y la de los modelos predictivos. La primera incluye el modelado inspirado en epidemias, mientras que la segunda incluye modelos de propagación como la cascada [16] o los modelos de umbral lineal [4]. Se ofrecen más detalles del enfoque inspirado en pandemias en la Sección 2.3.2 y de los modelos en cascada y modelos de umbral lineal en la Sección 2.3.3. Estas líneas de investigación son de suma importancia para áreas como marketing, publicidad, epidemiología y análisis de redes sociales [81].

Algunas aproximaciones al modelado de la difusión de la información en redes se basa únicamente en los resultados de la teoría de grafos [39, 64] asumiendo un conocimiento completo de la red, pero no reportan validación empírica sobre datos reales (ver Sección 2.3.1), algunos son puramente especulativos[66, 28, 54, 80, 73, 68, 59]. También se han propuesto predicciones agregadas de comportamiento macroscópico o mesoscópico de difusión de información [85, 58, 81, 51, 69]. Por ejemplo, modelar la difusión de información como pro-

pagación epidémica predice el número de usuarios que pertenecen a la clase infectada [58, 81, 51] en lugar de tratar de predecir la infección individual, esto es, es incapaz de modelar los procesos microscópicos.

Otros trabajos modelan la función de densidad de la distribución de usuarios influenciados [85], la influencia del nodo derivada de las propiedades topológicas de la red [69], o la difusión de información macroscópica como la propagación de una señal a través de la red donde la interferencia entre los eventos se modelan mediante convolución de señales [96]. En el nivel microscópico, obteniendo de los datos la recompensa de las decisiones de los agentes sociales permite una predicción precisa de la difusión de la información [86].

Predictores basados en *machine learning* de la actividad de Twitter han sido desarrollados [109], sin embargo, los datos no están siempre disponibles para la confirmación de los resultados. El papel de los tópicos en la adopción de Twitter se ha considerado a través de modelos predictivos de *machine learning* [70] donde los temas corresponden a *hashtags* seleccionados, descubriendo que los tópicos juegan un papel importante a nivel microscópico en la propagación de la información. Los temas de los *hashtags* también se utilizan en la construcción de la medida de similitud subyacente a un modelo de transferencia de radiación para la predicción de influencia [83], pero su papel no es aislado.

2.3.1. 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 [26, 39, 56, 64, 65]. Por ejemplo, en Nora Alon et al [39] 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 [56]. Después de eso, Lucy Small y Oliver Mason [65] demostraron que la afirmación es cierta si el gráfico subyacente es un árbol y luego, en [39], 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 [66] 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.

2.3.2. Modelos de contagio

Como se mencionó en la sección anterior una familia de modelos utilizada frecuentemente es la de los modelos inspirados en la propagación de epidemias o 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. En [68] se analiza los efectos generados sobre dos métricas macroscópicas, umbral de brote y prevalencia epidémica, el uso de arcos ponderados. Realizaron experimentos utilizando 2 conjuntos de distribuciones de peso, distribución Uniforme y Poisson. 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 afec-

taría los resultados de la epidemia una vez que se fija el peso promedio de toda la red. Por su parte, en [73] se propone un modelo con probabilidades jerárquicas de infección en los arcos, que dependen de la posición relativa de los nodos finales en la red (núcleo o periferia). Se testea este enfoque en un dataset de twitter obteniendo patrones de comportamiento de difusión similares. En [28] se aplica el modelo SIR para capturar el comportamiento humano en una comunidad virtual, específicamente "*2 channel* canales"el sistema de tablón de anuncios (BBS) anónimo abierto más grande de Japón. Otro enfoque se presenta en [54] 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. Otra aplicación del modelo SIR se ve en [51] donde se aplica a nivel de tópico 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 [58] 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! Finance-Walmart message board* y usan noticias relacionadas con *Walmart* del *Wall Street Journal*. A continuación, en [81] proponen un modelo de contagio (epidemia) SIR para modelar la difusión de información en Foros web debido a patrones similares en la difusión de información y procesos de contagio social. Se basan en trabajos anteriores [28, 51, 58] cambiando la vista de difusión de información desde una orientación sobre la publicación a un enfoque sobre el autor. En Xiong et al. [59] proponen otra variación de un modelo

de contagio, el modelo susceptible-contactado-infectado-refractario (SCIR). Probaron el modelo mediante simulaciones numéricas. En [80] se incorpora un mecanismo de olvido y refutación en el modelo SIR para describir la propagación de rumores con mayor precisión validado tanto sobre simulaciones numéricas como en la OSN Renren.

2.3.3. Modelos de Grafos

Principalmente 2 modelos entran en esta categoría, a saber, los modelos de cascadas independientes (IC) [17] y de umbral lineal (LT) [5]. Como se describe en [60], ambos asumen la existencia de una estructura de grafo estática que subyace a la difusión y se enfocan en la estructura del proceso. El modelo IC asocia una probabilidad a cada borde que representa la posibilidad de que la información se difunda. El modelo LT define un umbral para cada usuario (nodo) y un grado de influencia para cada arco. 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. El problema principal para aplicar estos enfoques a nuestro problema es que para los foros web no existe un grafo explícito.

2.3.4. Otros

Por último, hay modelos relevantes para este estudio que no encajan en la categorías definidas previamente. En [26] se utiliza el modelo del votante 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 de modelar las decisiones de contribución de contenido. Lee et al. [49] propone modelos de interacción espacial típicamente usados en el campo de la economía y la geografía económica para estudiar la relación entre la distancia y la familiaridad entre estudiantes uni-

versitarios usando datos de la OSN StudiVZ. Por otro lado, en Hu et al. [85] se prueba 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 difusión 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 la OSN Sina-weibo de China.

2.4. Modelado semántico

Por otra parte, el modelado semántico del contenido de información publicado en la OSN está ganando atención. Por ejemplo, el análisis semántico de las redes sociales weibo y twitter basado en tópicos de una sola palabra se ha aplicado para estudiar la percepción pública sobre las vacunas contra el COVID-19 [112]. Se ha demostrado que el modelado semántico de contenidos de usuario permite una mejor detección de comunidades[92, 106]. El impacto de eventos específicos en los medios sociales puede ser evaluado utilizando modelos semánticos. Por ejemplo, un modelo aproximado [103] se muestra capaz de detectar eventos en los medios sociales, mientras que el resumen de eventos sobre la base de tweets se puede lograr mediante una arquitectura de *deep learning* [104]. En concreto, el análisis de tópicos por LDA se ha utilizado para descubrir el significado de los eventos en redes sociales [79] y la evolución de los contenidos en las redes sociales [84]. En particular, el análisis de sentimiento se ha propuesto para predecir los resultados de concursos de canciones [98]. Para sistemas de recomendación, se ha propuesto

[107] un sistema de recomendación de tópico híbrido basado en LDA, y el análisis semántico para recomendaciones también se ha utilizado en entornos educativos [99]. Además, el modelado semántico de las interacciones del usuario con un chatbot permite interacciones personalizadas [95]. El análisis semántico puede extenderse en el dominio del tiempo, lo que permite medir cambios en contenidos de forma dinámica. Se aplicó la dinámica de tópicos para rastrear la aparición de tweets influyentes sobre el desastre de Fukushima [100] durante un largo período de tiempo. La consideración tanto del tiempo como del contenido permitió monitorear los cambios en una VCoP donde el usuario intercambia información sobre cosméticos [63].

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

La primera está relacionado con la evolución de la red en el que la pregunta que buscamos responder es sí 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, falta algunos de los enlaces, y la pregunta que debemos responder es si es posible inferir los enlaces que faltan utilizando la información que tenemos disponible.

Se realizó una revisión extensa en [46] donde examinaron varios enfoques del problema, como modelos de atributos, modelos gráficos bayesianos y el enfoque algebraico lineal, comparando la complejidad del modelo, el rendimiento

de la predicción, entre otros. En [91] se aborda el problema de predicción de enlaces como un problema de optimización con restricciones de cardinalidad. En [76] abordan el problema de la evaluación para estandarizar las métricas utilizadas y hacer que los resultados sean comparables entre investigaciones. Por otro lado, en [90] 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. Finalmente, en [29] 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.

2.5.1. Clasificación de trabajos anteriores

Tabla 2.1: Enfoques de modelado de los procesos difusión de información en redes sociales encontrados en la literatura. N/A = no disponible

Ref./año	Descripción del Modelo	Resultados	Data Set
[28]/2007	Modelo SIR para estimar número de accesos a un sitio.	N/A	foro web 2 channel. DATA: Número de posters por 15 min 9pm Ene 10 2007-6am Ene 11 2007.
[36]/2009	Propiedades topológicas del grafo de la OSN	N/A	data tipo Flickr `http://socialnetworks.mpi-sws.org/datasets.html`

Tabla 2.1: Enfoques de modelado de los procesos difusión de información en redes sociales encontrados en la literatura. N/A = no disponible

Ref./año	Descripción del Modelo	Resultados	Data Set
[39]/2010	Modelo de difusion de tecnologías de teoría de juegos que permite la competencia entre agentes.	N/A	No aplicable a redes implícitas.
[51]/2011	Modelo SIR basado en Tópicos. Aplicado a la difusión de tópicos violentos.	R-cuadrado: 0.57-0.8	Ummah data set Dark Web Forum Portal por AI lab de U. of Arizona. 1,263,724 posts, 76,242 hilos, 15,345 autores.
[54]/2012	Modelo generativo probabilístico de emergencia de información en redes, que captura exposiciones internas y externas. Difusión de URLs.	N/A	Probado en datos sintéticos y data set completo de Twitter de Enero 2011 . 3 billones de tweets, 18,186 URLs.
[59]/2012	Modelo SCIR.	N/A	Probado en datos sintéticos.

Tabla 2.1: Enfoques de modelado de los procesos difusión de información en redes sociales encontrados en la literatura. N/A = no disponible

Ref./año	Descripción del Modelo	Resultados	Data Set
[58]/2012	Modelo SIR basado en eventos	R-cuadrado: 0.66-0.89	Yahoo! Finance Walmart message board
[64]/2013	Modelo determinista de difusión de información competitiva sobre la Transitividad local iterada.	N/A	No aplicable a redes implícitas.
[66]/2014	Modelo de teoría de juegos evolutivos para dinámicas de difusión.	N/A	Dataset de Twitter hashtags. 1000 Twitter hashtags, númerp de menciones por hora y series de tiempo.
[68]/2014	Modelos SIS y SIR con pesos de arcos.	N/A	Data sintética.
[73]/2015	Modelo de propagación de memes basado en topología de red..	N/A	Probado en Higgs Twitter Network.

Tabla 2.1: Enfoques de modelado de los procesos difusión de información en redes sociales encontrados en la literatura. N/A = no disponible

Ref./año	Descripción del Modelo	Resultados	Data Set
[70]/2015	Probabilidad de adopción. Predicción de *machine learning*	F1=0.93	Twitter hashtags y URLs 2009
[81]/2016	Modelo SIR a nivel de tópicos.	R^2 0.52-0.75 y 0.44-0.79	Yahoo! Finance Walmart message board (139,062 hilos, 441,954 mensajes, 25,500 autores) y US Politics Online Breaking News in Politics (2192 hilos, 130,850 mensajes, 1,124 autores).
[80]/2016	Modelo SIR con mecanismos de asfixia y olvido.	N/A	Data sintética y data de OSN Renren (9,590 nodos, 89,873 arcos).
[85]/2017	Modelo hidrodinámico de predicción de difusión de información.	$\overline{ACC}$: 76,2−88	6500 video tweets de Sina-weibo.
[83]/2017	Transferencia de radiación física	N/A	Twitter dataset de cerca de 9000 usuarios

Tabla 2.1: Enfoques de modelado de los procesos difusión de información en redes sociales encontrados en la literatura. N/A = no disponible

Ref./año	Descripción del Modelo	Resultados	Data Set
[86]/2017	Modelado de pago de decisiones	Prom. Precision : 0.7	Sina Weibo y Flickr datasets
[96]	Maximización de Esperanza. Simulación Monte Carlo	R^2: 0.98	predicción de volumen de difusión de SINA microblogging
[109]/2020	Regresión logistica Bayesiana y predictores random forest	F1: 0.89-0.91	Twitter data extraída sobre tópicos informativos y de tendencias. N/A
[108]/2020	incendio forestal modificado	Núm. esparcidores	Twitter datasets

Como podemos notar en la Tabla 2.1, la mayoría de las investigaciones previas revisadas se centran en OSNs 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 las relaciones de *amistad* o *seguidor*, 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 [28, 51, 58, 81] dado que tratan con las particularidades de este tipo de OSN. Sin embargo, estos trabajos realizan un modelado macroscópico de la red en general, mientras que nuestro objetivo es el modelado a nivel microscópico

de las decisiones individuales de contribución a la OSN.

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 las publicaciones en los hilos de conversación (*posts*) 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

Los trabajos de experimentos computacionales reportados en esta Tesis doctoral se llevan a cabo sobre los datos extraídos de un foro web llamado *Plexilandia*, el cual fue implementado como una red social en línea (OSN) con más de 2500 miembros. *Plexilandia* es el hogar de una Comunidad Virtual

de Práctica (VCoP), [41, 48, 38, 89, 55] específicamente dedicada a la construcción, manipulación y retoque de aparatos musicales (tales como amplificadores, equipos de música, efectos musicales, entre otros), que lleva en funcionamiento más de 15 años. Los datos que se abordan en este trabajo correponden a los de la época de mayor actividad del foro, que abarca alrededor de 9 años. La Tabla 3.1 contiene el número de publicaciones de contenido (*posts*) por subforo a lo largo de estos 9 años, incluido el número total de publicaciones.

Tabla 3.1: Actividad en Plexilandia medida en número de publicaciones de contenido por Sub-Foro relevante por año.

Forum	***2006***	***2007***	***2008***	***2009***	***2010***	***2011***	***2012***	***2013***	***2014***	***TOTAL***
Aplifiers (2)[1]	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**

Los temas tratados dentro del foro de *Plexilandia* están organizados en Sub-Foros según el interés de los miembros de la VCoP que lo frecuentan. La Tabla 3.1 identifica los siguientes subforos: Amplificadores, Efectos, Luthiers, General, Audio para profesionales y Sintetizadores. Los contenidos publicados en dichos subforos debiesen estar estrictamente relacionados con el propósito de la comunidad, aunque de vez en cuando pueden surgir temas espurios producto de la interacción irrestricta entre usuarios. La estructura jerárquica del foro de subdivisión en Sub-Foros se ilustra en la Fig. 3.2.

Las contribuciones de contenido por parte de los usuarios son realizadas dentro de hilos de conversación que podemos llamar de forma abreviada como *hilos*. Un *hilo* discutiendo un tema comienza con un mensaje publicado por un usuario, que contiene una pregunta o la presentación de una idea para abrir una discusión. Luego, los diferentes miembros de la comunidad publican sus contribuciones aumentando así el conocimiento compartido sobre el

tema central de la conversación. Cada publicación en el *hilo* se compone de elementos como el identificador de usuario (ID); la aportación de contenido, que dependiendo del foro puede ser texto, imágenes, enlaces a otras páginas, videos; y la información de gestión del sistema de foros, como la fecha de creación de la publicación, el hilo, y el tema al que pertenece. Todos estos elementos pueden ser tomados en consideración, pero en este documento solo se explota el contenido de texto de las publicaciones para construir y analizar la red social.

3.1.1. Categorías de usuarios

Los administradores de la OSN proporcionaron una estratificación de los miembros para el año 2013 en cuatro categorías de usuario [89] según el papel que juegan en mantener vivo el foro:

- Expertos Tipo A: son los miembros clave más importantes que crean y mantienen *hilos* significativos en subforos relevantes. Hay 34 de miembros en esta categoria de acuerdo con el criterio de los administradores.
- Expertos Tipo B: que también son muy importantes pero en menor medida que los miembros clave de tipo A. Contribuyen de manera constante pero tienen un papel menos fundamental. Hay 21 miembros de esta categoría.
- Expertos Tipo C: Este tipo de usuario corresponde a aquellos que son miembros clave históricos. Han estado involucrados en la red social desde sus orígenes, pero no están participando continuamente. En esta categoría hay unos 11 miembros.
- No expertos o Tipo X: esta clase contiene a 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.

Para los experimentos se usan solo los datos de los años 2013 y 2014 porque solo se posee información información sobre los miembros clave para estos años [89]. Además, se usan solo los datos de los Sub-Foros 2 a 6. Se descartan los subforos 1 y 7 porque no tienen suficientes publicaciones para contribuir al análisis.

3.2. Preparación de datos

3.2.1. Selección y preprocesamiento de datos

El primer paso a realizar consiste en aplicar la curación y preprocesamiento de datos [75] a los datos extraídos de *Plexilandia*. Primero, filtramos las repeticiones de contenido anterior dentro de las contribuciones publicadas en el hilo. Un usuario puede responder a una publicación creando una nueva contribución de contenido que incluye una copia de la publicación citada más el texto adicional del nuevo aporte. Por lo tanto, es necesario eliminar el texto replicado de la nueva publicación para conservar solo el contenido original de la nueva publicación. A continuación, se transforman las siglas o abreviaturas, se eliminan las faltas de ortografía, y todos los elementos de las publicaciones que puedan volverlos no comparables entre ellos. Este proceso se lleva a cabo mediante dos técnicas de procesamiento del lenguaje natural: *stemming* y *stop words removal*. *Stemming* corresponde a un proceso que consiste en la reducción 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 "inutilizable" o "misceláneo". *Stop words removal* consiste en la

eliminación de todas las palabras que no aportan información a la red, como artículos, pronombres y palabras "inutilizables". Esto sirve para que las publicaciones sean comparables y para reducir el número de palabras utilizadas para calcular la comparación de publicaciones. Luego, nosotros aplicamos modelamiento de tópicos no supervisado LDA que se decribe posteriormente en la Sección 3.2.1.1 para realizar el modelado semantico del contenido de los documentos [30].

3.2.1.1. Latent Dirichlet Allocation

En esta sección damos una breve descripción de la utilización de análisis de tópicos *Latent Dirichlet Allocation* (LDA) para el modelado semántico. Sea $\mathcal{V}$ un vector de tamaño $|\mathcal{V}|$ en el que cada fila representa una palabra diferente utilizada en la red, es decir, el vocabulario. Sea v_i la palabra en lugar i del vector $\mathcal{V}$. Es posible representar la publicación p_j como una secuencia de S_j palabras de $\mathcal{V}$, con $S_j = |p_j|$, donde $j \in \{1, \ldots, |\mathcal{P}|\}$ y $\mathcal{P}$ corresponde a el número de publicaciones que se han publicado en el foro de la VCoP. Un *corpus* se define como una colección de publicaciones $\mathcal{C} = \{p_1, \ldots, p_N\}$. Podemos definir la matriz $\mathcal{W}$ de tamaño $|\mathcal{V}| \times |\mathcal{P}|$ donde cada elemento $w_{i,j}$ de esta matriz se define como el número de veces que aparece la palabra v_i en la publicación p_j. Entonces $\sum_{i=1}^{|\mathcal{V}|} w_{i,j} = S_j$. Del mismo modo, podemos definir $\sum_{j=1}^{|\mathcal{P}|} w_{i,j} = T_i$ que representa el número total de apariciones del término w_i en el *corpus*.

Un corpus puede ser representado por la matriz *term frequency and the inverse document frequency* (TF-IDF) $\mathcal{M}$ de tamaño $|\mathcal{V}| \times |\mathcal{P}|$ [2], que se define de la siguiente manera: cada elemento $m_{i,j}$ en la matriz se determina como

$$m_{i,j} = \frac{w_{i,j}}{T_i} \times \log \left[\frac{|\mathcal{P}|}{1 + n_i} \right], \tag{3.1}$$

donde n_i es el número de publicaciones que incluyen la palabra w_i, T_i es el número máximo de apariciones de la palabra w_i en cualquier publicación. El término IDF presentado en la ecuación (3.1) contiene una corrección con respecto al término IDF original, que era $\log\left[\frac{|\mathcal{P}|}{n_i}\right]$, para evitar resultados indefinidos cuando una publicación no contiene palabras después de la curación de datos. Para la reducción de dimensiones empleamos la técnica no supervisada de descubrimiento de tópicos LDA [33, 21] utilizando la implementación basada en el muestreo de Gibbs [34]. Esta implementación no determina los valores óptimos de la hiperparámetros α, β y número de temas requeridos $|\mathcal{T}| = k$, por lo que se requiere de una exploración empírica para encontrarlos. LDA nos proporciona

(1) la distribución de cada palabra sobre los temas descubiertos,

(2) la distribución de temas en las publicaciones, y

(3) las n palabras más importantes que representan cada tema junto a sus probabilidades de pertenencia.

Para tener vectores de probabilidad de tamaño fijo que representen cada tema $|\mathcal{V}|$, los rellenamos con ceros. Estos vectores son las columnas de la matriz semántica (SM) de dimensiones $[Términos \times Topicos]$. Para obtener la descripción semántica de los posts en una matriz de dimensiones $[Publicaciones \times Tópicos]$, multiplicamos la matriz SM por $\mathcal{M}^t$, la transpuesta de la matriz TF-IDF definida por ecuación (3.1). La matriz resultante de dimensiones $[Publicaciones \times Topicos]$ contiene la explicación semántica de cada publicación como un combinación lineal de los tópicos descubiertos mediante sus representaciones semánticas vectoriales dadas por las filas de la matriz, denotadas $\{\rho_p; p \in \mathcal{P}\}$.

3.2.2. ETL para datos de entrada del modelo

En los foros, el grafo de la red social no está definido explícitamente como en otras comunidades (Facebook, Twitter, etc.). Por lo tanto, primero debemos definir la topología de la red. 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

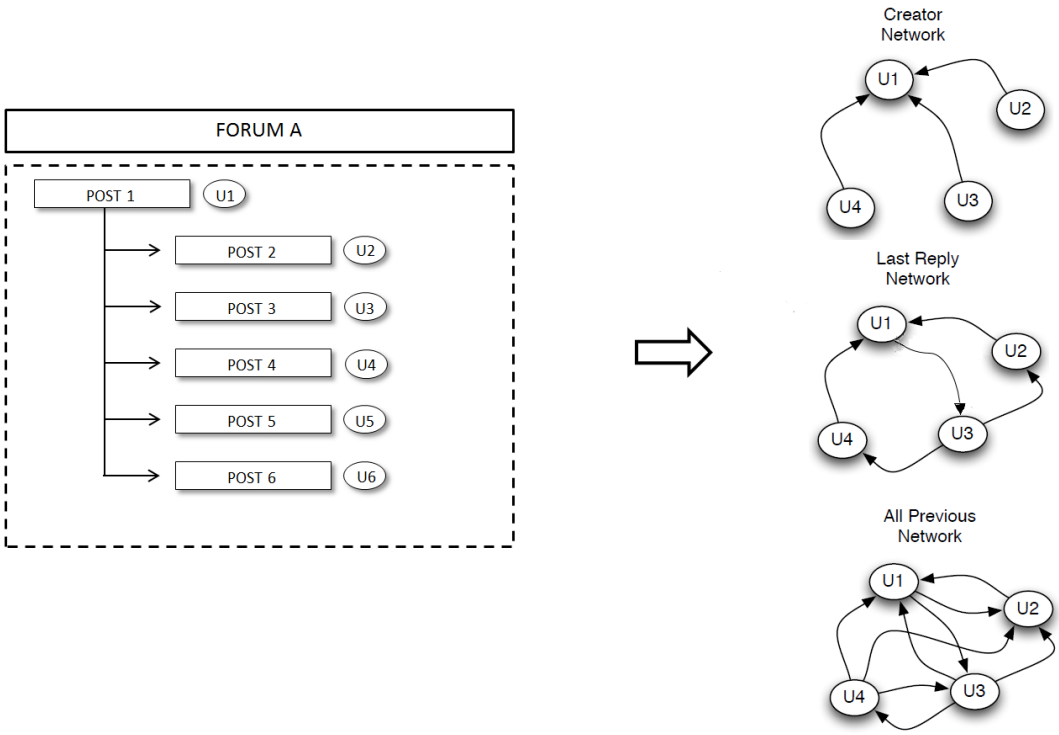

Figura 3.1: Posibles representaciones de red para foros web.

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 Hilo 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 hilo. Tengase 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.

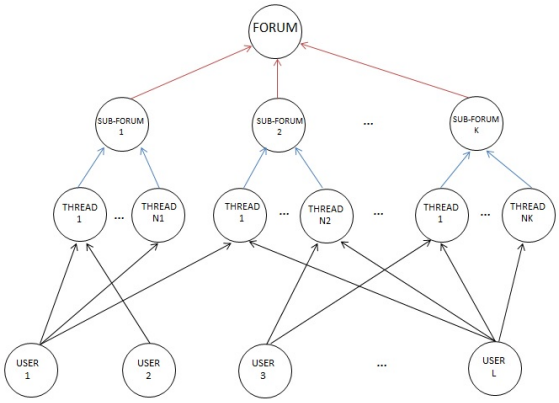

Figura 3.2: Topología propuesta para foros web.

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 propuesta. 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

$$\text{número de arcos posibles en la red} = n(n-1) \tag{3.2}$$

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 hilos activos es m, entonces la cantidad de arcos posibles estaría dada por

$$\text{número de arcos posibles en la red} = nm \tag{3.3}$$

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.

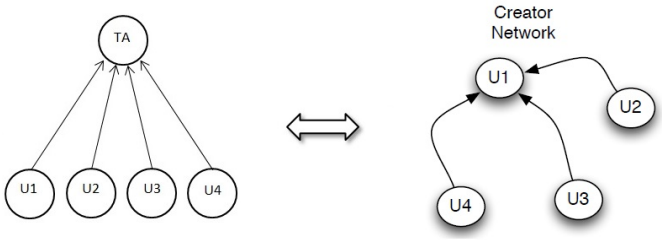

Figura 3.3: Representaciones de la red con cuatro usuarios y un hilo (TA). A la derecha la representación habitual, donde los enlaces relacionan usuarios que participan en el mismo hilo. A la izquierda la representación propuesta.

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 en forma de grado de la red social 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.

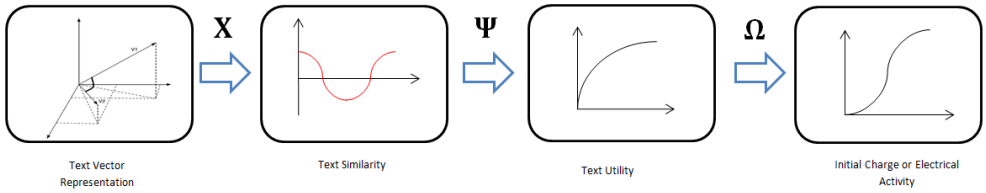

Figura 3.4: Secuencia de transformaciones aplicadas a la representación semántica de los hilos y usuarios para obtener la entrada del modelo ELCA.

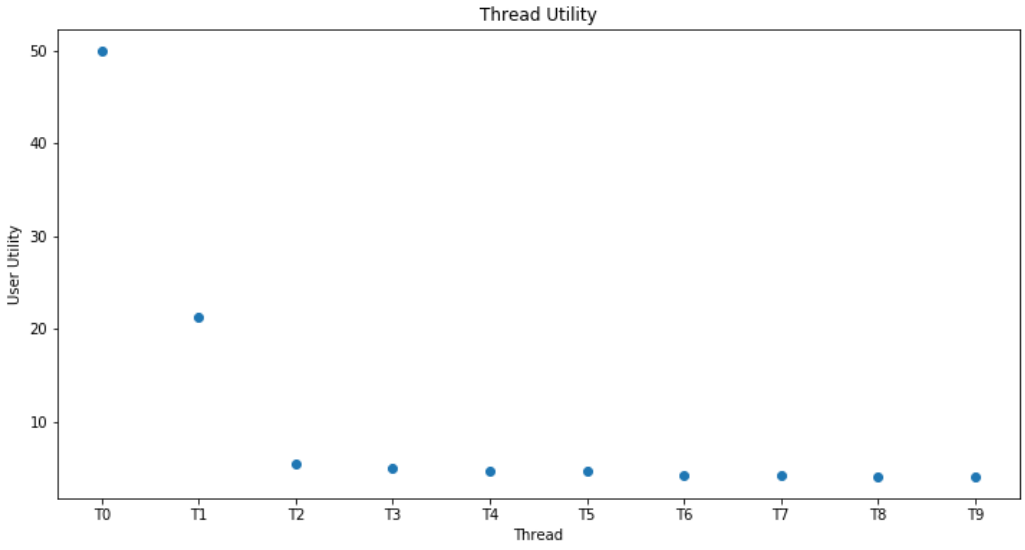

Figura 3.5: Ejemplo 1 de utilidad de los hilos

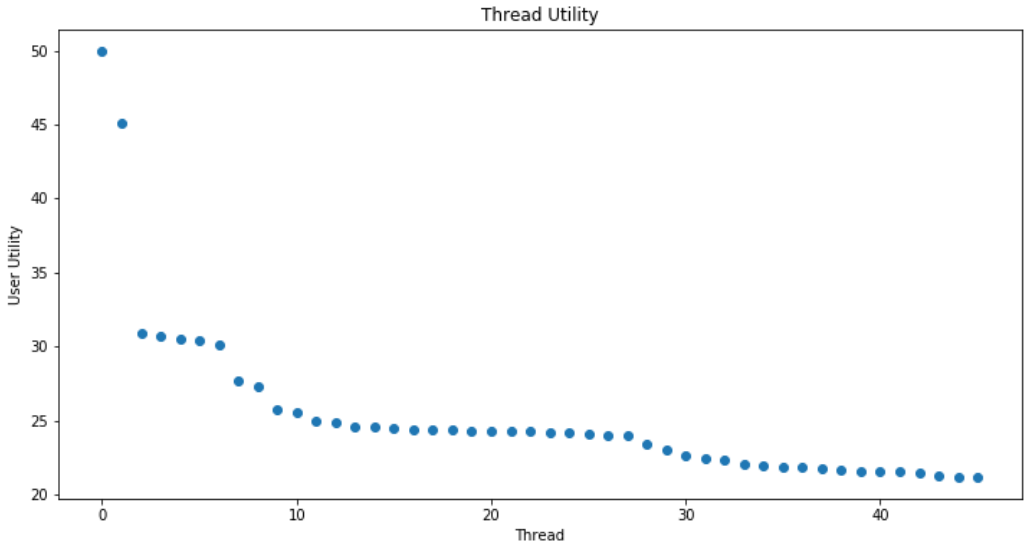

Figura 3.6: Ejemplo 2 de utilidad de los hilos

Denotemos $\mathcal{U}$, $\mathcal{TH}$ y $\mathcal{SF}$ el conjunto de usuarios, el conjunto de hilos y el conjunto de subforos de la comunidad virtual, respectivamente. Los resultados del análisis semántico LDA, a saber, los vectores ρ_p, permiten inducir para cada usuario ($u \in \mathcal{U}$) una representación vectorial de sus preferencias de multi-tópicos, y para cada hilo ($h \in \mathcal{TH}$) una representación vectorial de su contenido semántico. El proceso para calcular estos representaciones semánticas vectoriales es el siguiente:

1. Se agrupan las contribuciones de contenido de los usuarios según el subforo $f \in \mathcal{SF}$ donde fueron publicadas.

2. Discretizamos el eje de tiempo en periodos de tamaño Δt, creando de esta forma un conjunto de periodos de tiempo T. Posteriormente, agrupamos los aportes de contenido de cada subforo según el período de tiempo ($t \in T$) al que pertenecen.

3. Extraemos los usuarios ($\mathcal{U}_f^t$) y los hilos ($\mathcal{TH}_f^t$) que están activos durante cada período de tiempo. Un usuario u se considera que está activo en el

Sub-Foro f y período t si hace una contribución de contenido durante este período. Un hilo h en el subforo f está activo en el período t si hay al menos un usuario que hace una contribución de contenido al hilo durante el período t.

4. La representación vectorial del contenido semántico del hilo para un período, denotado ν_h^t, es la media de las representaciones vectoriales semánticas ρ_p de las contribuciones de contenido que pertenecen tanto al hilo h y al periodo t, formalmente:

$$\nu_h^t = \frac{1}{|\mathcal{P}(h,t)|} \sum_{p \in \mathcal{P}(h,t)} \rho_p, \tag{3.4}$$

   Donde $\mathcal{P}(h,t) = \{p \in \mathcal{P} : \ p \text{ se publica en el hilo } h \text{ durante el período } t\}$.

5. Para calcular la representación semántica del usuario, clasificamos en subgrupos, denotados por s, las contribuciones de contenido realizadas por un usuario durante un período de acuerdo con el hilo en el que se publicaron. Un usuario tendrá tantas representaciones vectoriales semánticas para un período como la cantidad de hilos en los que haya contribuido durante ese período. Denotamos la colección de estas representaciones vectoriales como S_u^t.

6. La representación semántica vectorial para un usuario en un período t y subgrupo de contribuciones de contenido s, denotado $\mu_{u,s}^t$, es la media de las representaciones vectoriales semánticas ρ_p para el contenido aportes realizados por el usuario u en este periodo de tiempo, formalmente:

$$\mu_{u,s}^t = \frac{1}{|\mathcal{P}(u,s,t)|} \sum_{p \in \mathcal{P}(u,s,t)} \rho_p, \tag{3.5}$$

donde

$$\mathcal{P}(u,s,t) = \{p \in \mathcal{P} :\ p \text{ es posteado por el usuario } u \text{ en el periodo } t \text{ y pertenece al subgrupo } s\}. \tag{3.6}$$

Ahora que tenemos la representación vectorial semántica multitópica de los usuarios y la representación semántica de los hilos, aplicamos el proceso computacional que se muestra en la Fig. 3.4 para obtener el *input* para el ELCA que implementa el modelo de decisión de contribución de contenido.

1. Primero, seleccionamos una medida de la similitud χ de dos representaciones vectoriales semánticas en el espacio de tópicos. Usamos la similitud coseno, dada por el coseno del ángulo formado entre dos vectores semánticos. Por lo tanto, para una representación vectorial de las preferencias multi-temáticas de un usuario $\mu^t_{u,s}$ y una representación vectorial del contenido semántico de un *hilo* ν^t_h, la similitud entre ellos viene dada por

$$\chi(\mu^t_{u,s}, \nu^t_h) = cos(\theta) = \frac{\mu^t_{u,s} \cdot \nu^t_h}{|\mu^t_{u,s}||\nu^t_h|}, \tag{3.7}$$

   donde θ es el ángulo entre $\mu^t_{u,s}$ y ν^t_h.

2. Luego, definimos una función Ψ_1 para mapear la similitud semántica en la utilidad del usuario. La utilidad que un usuario extrae de un *hilo* es el número esperado de veces que él elige el hilo sobre otros hilos para hacer una contribución de contenido. Considere que $\pi = 1 - \chi(\mu^t_{u,s}, \nu^t_h)$ es el parámetro de probabilidad de éxito de una distribución geométrica. La utilidad Ψ_1 de la similitud entre las preferencias de un usuario y la representación semánticas de un hilo se define de la siguiente manera [35]:

$$\Psi_1(\mu^t_{u,s}, \nu^t_h) = \frac{1}{1 - \chi(\mu^t_{u,s}, \nu^t_h)}. \tag{3.8}$$

Además, la preferencia de un usuario por un hilo, es decir, la utilidad normalizada de un usuario dada por un *hilo* h, denotada por $V^t_{u,s,h}$, toma en cuenta todos los hilos del subforo, calculado por una función Ψ_2 definida de la siguiente manera:

$$V^t_{u,s,h} = \Psi_2(a, \mu^t_{u,s}, \nu^t_h) = a\frac{\Psi_1(\mu^t_{u,s}, \nu^t_h)}{\max\limits_{j \in \mathcal{TH}^t_f} \Psi_1(\mu^t_{u,s}, \nu^t_j)}, \tag{3.9}$$

donde el parámetro a modula la preferencia de los usuarios a hilos cuyos temas son similares a los temas cubiertos por las contribuciones de contenido del usuario. Cuanto mayor sea la preferencia, mayor la satisfacción extraída de la conversación. Las Fig. 3.5 y 3.6 trazan ejemplos de los valores de utilidad que un usuario atribuye a los *hilos* que están activos en algún período de tiempo. Note que solo unos pocos hilos son de gran interés para el usuario. La mayor parte de los *hilos* activos se apilan en la cola del gráfico, lo que significa que en su mayoría contribuyen ruido al proceso de decisión. Por lo tanto, reducimos el número de *hilos* alternativos que un usuario tiene en cuenta durante su proceso de toma de decisiones para generar contenido, conservando solo los m *hilos* con valores de utilidad superiores. Esta reducción de alternativas se basa en resultados de investigaciones clásicas sobre la memoria de trabajo y capacidad de atención [1].

3. Como paso final, definimos una función Ω que mapea la utilidad normalizada percibida por el usuario de cada *hilo* en la entrada del ELCA asociada con la decisión de hacer una contribución de contenido al *hilo*, denotada $I^t_{u,s,h}$. Para este propósito, hacemos uso de la teoría de la utilidad aleatoria [35]: $I^t_{u,s,h}$ es proporcional a la probabilidad de elegir

entre *hilos* alternativos. Formalmente:

$$I^t_{u,s,h} = \Omega(\mathbf{V}^t_{u,s}(m), h) = \beta_{(c(u))} \frac{e^{V^t_{u,s,h}}}{\sum_{j \in \mathcal{TH}^t_f(u,m)} e^{V^t_{u,s,j}}} \tag{3.10}$$

Donde $\beta_{(c(u))}$ es un parámetro de proporcionalidad del modelo, que es específico para la categoría $c\,(u)$ del usuario (definido como A, B, C o X en la sección 3.1.1), y $\mathcal{TH}^t_f(u,m) = \{h \in \mathcal{TH}^t_f :$
La utilidad de h es una de los mejores m para el usuario $u\}$.

Capítulo 4

Implementación del modelo ELCA

En este capítulo se presentan el proceso computacional diseñado para la implementación del modelo y los detalles de la formulación y calibración del modelo propuesto de generación de contenidos por parte de los usuarios en la red social. Este modelo es una extensión del modelo *Leaky Competing Accumulator* (LCA) propuesto en los trabajos relativos al modelado de los procesos sicológicos de toma de decisiones, tratando de explicar fenómenos como la inversión de las preferencias. En nuestro caso, la decisión de

4.1. Proceso computacional

Como se mencionó anteriormente en la Sección 1.3, la pregunta principal que trata de responder este trabajo es cómo modelar la toma de decisiones de los agentes en el proceso de contribución de contenido en los foros web, con un enfoque basado en el contenido semantico de las publicaciones (*posts*)

realizadas por los usuarios.

Para conseguir este objetivo se diseñó el proceso computacional que se ilustra en la Fig. 4.1. El proceso computacional de este trabajo se compone de 5

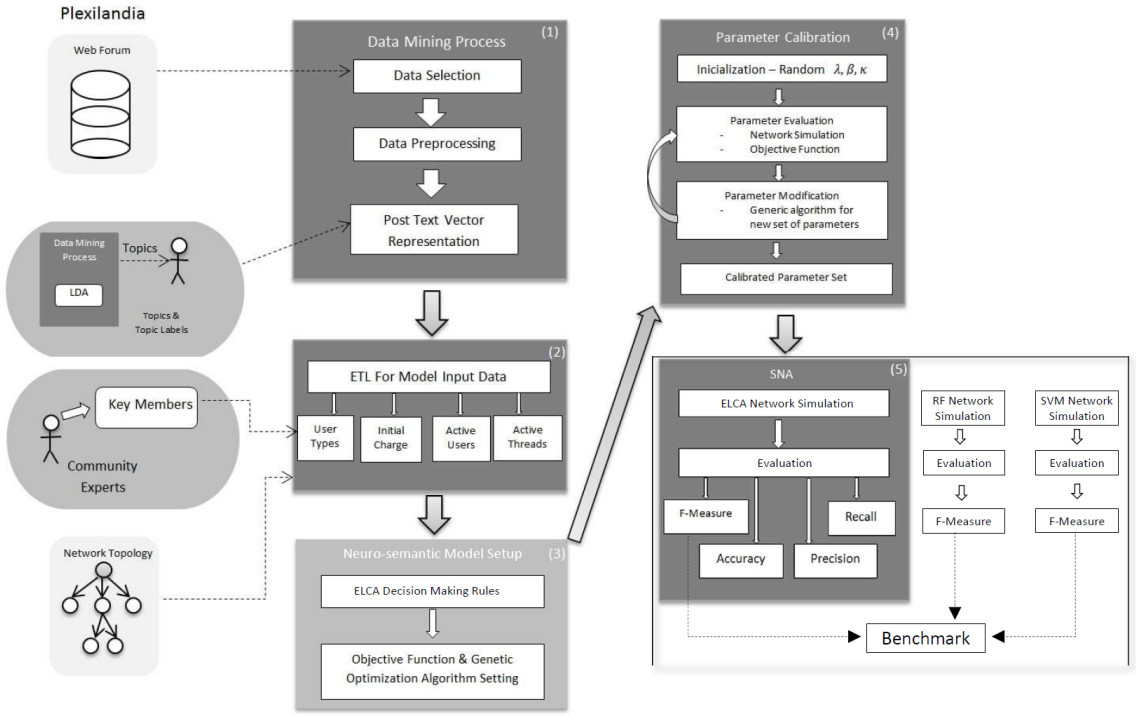

Figura 4.1: Proceso computacional del estudio

etapas correspondientes a las casillas numeradas en la figura partiendo de izquierda a derecha:

(1) Proceso de minería de datos: en esta fase se lleva a cabo la curación y el preprocesado de los datos crudos de la OSN descritos en la Sección 3. Sección 3.2.1 describe la curación y el preprocesamiento de datos. Es más, se construye una caracterización de cada contribución del foro mediante un análisis no supervisado de tópicos semánticos via LDA. La Sección 3.2.1.1 ofrece una breve descripción general de LDA.

(2) Etiquetado de datos de entrenamiento por expertos (ETL): en esta fase se prepara la categorización de usuarios utilizando información de expertos (es decir, los administradores de red) como se describe en la Sección 3.1. Esta categorización es usada para modular algunos de los parámetros del modelo ELCA como se discute a continuación.

(3) Configuración del modelo neuro-semántico: en esta fase formulamos el modelo neuronal LCA que simula el proceso de toma de decisiones de publicación de contenido en algún hilo de un subforo. Se describe el LCA extendido (ELCA) en la Sección 4.2. A partir del modelo semántico de tópicos obtenido por LDA se construye el vector de representación para cada hilo de conversación y para cada usuario relevante que servirá como entrada para el proceso de predicción de las decisiones de contribución generadas por ELCA. Esta construcción se describe en la Sección 3.2.2.

(4) Calibración de parámetros: un algoritmo genético busca la configuración óptima de los parámetros ELCA usando los datos reservados para entrenamiento. La función objetivo se define como el rendimiento predictivo sobre un subconjunto del conjunto de datos seleccionado para la calibración del modelo. El algoritmo genético se describe en la Sección 4.3.

(5) Análisis de redes sociales (SNA): se aplica el modelo ELCA para simular las decisiones de contribución de contenido tomadas por los usuarios. Los resultados de la simulación del modelo ELCA calibrado se utilizan como predicción del comportamiento real del usuario. En paralelo, a modo comparativo también se aplican modelos clásicos de aprendizaje automático, especificamente *random forest* (RF) y *support vector machines* (SVM) para simular las decisiones de contribución de contenido tomadas por los usuarios. La calidad de la predicción se evalúa frente la verdad del terreno dada por las contribuciones reales

registradas por los periodos de tiempo diseñados para la validación. El rendimiento predictivo se mide por la puntuación F1 (también se calculan otras métricas: *precision*, *accuracy* y *recall*). Se finaliza contrastando los rendimientos obtenidos por los 3 modelos. Los resultados experimentales se presentan en Sección 5.4.

4.2. Extended Leaky Competing Accumulator (ELCA)

En esta sección presentamos primero el modelo de decisión clásico, para elaborar después nuestro modelo extendido y el proceso de calibración o estimación de parámetros basado en un algoritmo genético.

4.2.1. Leaky Competing Accumulator(LCA)

Usher y McClelland [19] durante el año 2001 presentaron el modelo *Leaky Competing Accumulator (LCA)*, unificando conceptos teóricos a partir de los procesos cognitivo-perceptuales y los procesos neurofisiologicos subyacentes. En su trabajo muestran el modelo LCA como un modelo de difusión para la toma de decisiones. El LCA trata de explicar los procesos de decisión desde la perspectiva de la neurofisionlogía. Se preocupa de describir la evolución de la actividad eléctrica cerebral en determinadas regiones del cerebro, modelando la carga acumulada en la región i-esima como una variable aleatoria $\{X_i\}$. Se asume que cada decisión se puede asociar a una región distinta del cerebro, por lo que, cada etiqueta i modela una posible decisión del agente cognitivo. El proceso de simulación dinámica evoluciona de acuerdo a la ecuación (4.1), comenzando con valores iniciales $X_i(t = 0) \sim 0^+$, y deteniéndose en el instante $t = T(i^*)$. 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^*.

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

Donde I_i, es una constante que reune el efecto de la actividad sensorial que genera una propension a escoger la alternativa i, y sirve como entrada para el proceso LCA. Se impone la restricción sobre estas constantes de ser no negativas, es decir $I_i \geq 0$, siguiendo un razonamiento neurofisiológico. Por su parte, el parámetro ω está representado por dos valores, como se muestra en (4.2).

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

Donde el parámetro κ de (4.2) tiene en cuenta el decaimiento [6] de la actividad acumulada para una alternativa, y el parámetro λ da cuenta del fenómeno de inhibición lateral entre las unidades del acumulador. Se considera que el efecto es el mismo 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 ($X > 0$).

El modelo *LCA* se puede aplicar a datos provenientes de una OSN, como en el trabajo realizado en [55] en el que se considera que los usuarios de la OSN corresponden a agentes cognitivos.

4.2.2. ELCA

El proceso de decisión que conduce a la contribución de las publicaciones en los hilos de conversación está modelado por un *Extended Leaky Compe-*

ting Accumulator (ELCA). El modelo LCA original [19, 55, 25, 57], descrito en la Sección 4.1, sólo considera una decisión llevada a cabo por un único agente mientras que ELCA simula los procesos de decisión de muchos usuarios simultáneamente, es decir, ECLA extiende LCA sobre una comunidad de usuarios que están tomando decisiones simultáneamente. Se consideran procesos de simulación independientes para cada sub-foro f y cada período de tiempo t. Definimos $X_h^{(u)}$ como la activación (neuronal) asociada con la decisión del usuario $u \in \mathcal{U}_f^t$ de publicar una publicación en hilo $h \in \mathcal{TH}_f^t$. El proceso de decisión se implementa como un proceso de simulación dinámico donde las unidades de activación evolucionan hasta que alguna de ellas alcanza un umbral dado, que desencadena la decisión correspondiente. La evolución de las unidades de activación para un usuario se ilustra en la Fig. 4.2.

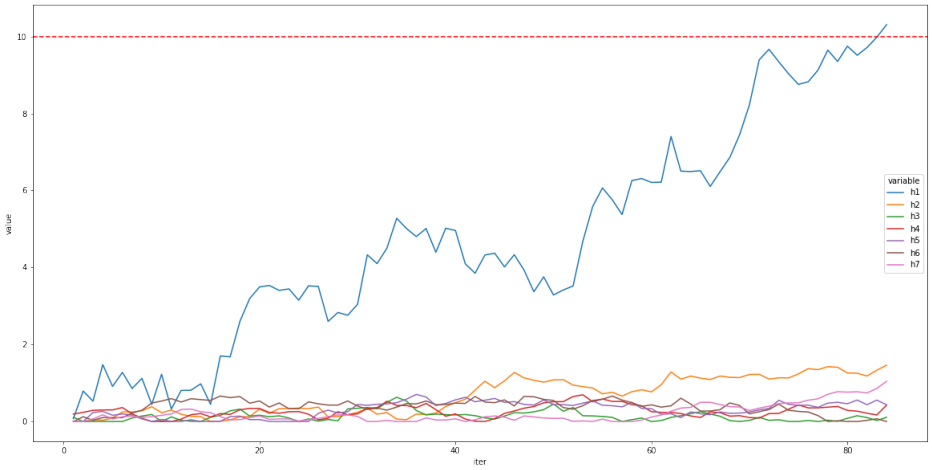

Figura 4.2: Una instancia de evolución de los acumuladores correspondientes a la decisión de publicar por parte de un usuario específico.

Además, en el modelo ELCA el término $I_{u,s,h}^t$ tiene valores semánticamente fundamentados asociados a cada elección, definidos en la ecuación (3.10), mientras que los modelos *LCA* clásicos tienen valores arbitrarios ajustados por la intuición del investigador. Finalmente, se proporciona un procedi-

miento para estimar los parámetros óptimos del *ELCA* para reproducir las decisiones reales tomadas por los usuarios, de forma similar al entrenamiento realizado por los enfoques convencionales de *machine learning*.

El modelo *ELCA* describe la evolución del proceso de decisión de todos los usuarios como la simulación del siguiente conjunto de ecuaciones estocásticas dinámicas:

$$dX_h^{(u)}(\tau) = \left[I_{u,s,h}^t - \sum_{j \in \mathcal{TH}_f^t} \omega_{hj}^{(c(u))} X_j^{(u)}(\tau) \right] d\tau + \sigma_h^{(u)} dW_h, \quad h \in \mathcal{TH}_f^t, u \in \mathcal{U}_f^t, \tag{4.3}$$

que se integran aplicando el método de Euler. Para cada sub-foro f se tienen tantas ecuaciones dinámicas implementando procesos de decisión concurrentes como la cantidad de usuarios e hilos que están activos durante el período de tiempo t. Las ecuaciones dinámicas para cada usuario u en la ec.(4.3) se integran hasta el momento τ^* cuando el usuario u toma la decisión de publicar un mensaje en el hilo h^*, es decir, cuando la unidad correspondiente supera un umbral de decisión $X_{h^*}^{(u)}(\tau^*) \geq Z$, como se ilustra en la Fig. 4.2. Se estableció empíricamente el valor del umbral como $Z = 10$. Los parámetros $\omega_{hj}^{(c(u))}$ modulan la inhibición lateral por los otros procesos de decisión en curso del usuario u, donde $c(u) \in \{A, B, C, X\}$ denota la categoría del usuario definida en la Sección 3. El término $I_{u,s,h}^t$ en la ec. (4.3) es un valor de entrada externo constante a favor de publicar una contribución en el hilo (alternativa) h basado en el análisis semántico desarrollado anteriormente. Esos valores de entrada son positivos, es decir, $I_{u,s,h}^t \geq 0$. Los valores de entrada externa se acumulan linealmente en la variable de activación $X_h^{(u)}$. Los pesos $\omega_{ij}^{(c)}$ toman diferentes valores dependiendo de la relación modelada y la categoría del usuario, como se muestra en ec. (4.4).

$$\omega_{ij}^{(c)} = \begin{cases} \kappa_c & i = j \\ \lambda_c & i \neq j \end{cases}, c \in \{A, B, C, X\}, \tag{4.4}$$

donde el parámetro κ_c modela el decaimiento de activación de cada unidad [6]. La inhibición lateral entre unidades acumuladoras está modelado por el parámetro λ_c. La Ec. (4.4) considera el mismo efecto para todas las unidades estratificadas por las diferentes categorías de usuarios definidas por los administradores de OSN. Siguiendo la inspiración biológica, las variables de activación están restringidas a valores positivos ($X_h^{(u)} > 0$). Este límite duro tiene algunas propiedades computacionales interesantes [25]. Este modelo está de acuerdo con la toma de decisiones perceptiva [27]. Las condiciones iniciales $X_h^{(u)}(\tau = 0)$ están especificados por la ec. (4.5):

$$X_h^{(u)}(\tau = 0) = (1 + \gamma)^l - 1 \tag{4.5}$$

El parámetro l en la ec. (4.5) denota el número de veces la alternativa de subproceso h ha sido elegida previamente, y el parámetro $\gamma \geq 0$ modela el efecto de elecciones repetidas de la misma alternativa acercándose a la curva asintótica definida en [44]. Trabajos recientes han mostrado convergencia a una decisión para una gran cantidad de opciones en un modelo LCA modificado [111], pero su modelo se limita a un solo agente. Ellos muestran que es posible recuperar los parámetros del modelo mediante el enfoque de máxima verosimilitud, sin embargo, se refieren a la reproducción de trazas de simulación mientras que el enfoque de este trabajo, que se mmuestra en la siguiente sección, propone la estimación de parámetros para aproximar el comportamiento de decisión del usuario extraído de los datos reales de la OSN.

Se presenta 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 $\{\rho_p \subset \mathcal{T}; p \in \mathcal{P}\}$.

Para cada $t \in \{2, \ldots, T\}$

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. Las predicciones de arcos 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 rendimiento (Sección 5.2) comparando con la verdad del terreno dada por:

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

.

4.3. Estimación de parámetros del modelo ELCA

Finalmente, se implementa una heurística de algoritmo genético (GA) [9] ilustrado en la Fig. 4.3 para estimar los parámetros del modelo ELCA. La configuración de GA utilizada es la siguiente:

- Cada individuo $P_g \in \mathbf{P}$ en la población de GA está compuesto de 12 genes de valor real, que son estimaciones de los parámetros de el modelo

ELCA para cada tipo de usuario en el subforo, es decir,

$$P_g = \left\{ \left(\hat{\beta}_c, \hat{\kappa}_c, \hat{\lambda}_c \right), c \in \{A, B, C, X\} \right\}.$$

El tamaño escogido de la población fue de 100 individuos. Los valores inciales de los genes de los individuos se generaron siguiendo una distribución uniforme en el intervalo [0,1].

- La función de aptitud de los individuos es la precisión de la predicción de contribución de contenido por el modelo ELCA utilizando la configuración de parámetros contenida en los genes de los individuos sobre el primer mes del conjunto de datos. En otras palabras, para calcular la aptitud de cada individuo en la población se ejecuta una instancia de la simulación ELCA comparando su rendimiento en la predicción de decisiones publicación de posts decisión con los datos del primer mes.

- La selección de individuos para el cruce se lleva a cabo mediante el algoritmo *linear-ranking* de Baker [74] y *roulette wheel selection* [52].

- El cruce reproductivo se implementó mediante el algoritmo *single point crossover* [8].

- El operador de mutación es *real value mutation*[7].

Se realizaron búsquedas GA independientes para la estimación de los parámetros óptimos del modelo ELCA de cada subforo. Los detalles de la implementación, como el tamaño de la población, el número de generaciones calculadas, y la implementación de políticas de selección elitista se especifican en el Algoritmo 2.

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

$$\mathbf{P}\left(k=0\right) = \left\{P_g\left(k\right); g = 1, \ldots, 100\right\},$$

donde

$$P_g\left(k\right) = \left\{\left(\hat{\beta}_c\left(k\right), \hat{\kappa}_c\left(k\right), \hat{\lambda}_c\left(k\right)\right), c \in \{A, B, C, X\}\right\},$$

se extraen como muestras de una variable aleatoria con distribución uniforme en el intervalo [0,1].

2. Calcular la función de aptitud inicial $f_g\left(k=0\right)$ para cada individuo mediante los siguientes pasos:

   a) Realizar la predicción de contribuciones descrita en el Algoritmo 1 sobre $\mathcal{C}_1$ usando $P_g\left(k\right)$ como los parámetros del ELCA.

   b) La aptitud $f_g\left(k=0\right)$ 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 $\mathbf{P}_{old,10}\left(k\right)$

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

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

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) crear la siguiente población

$$\mathbf{P}(k+1) = \mathbf{P}_{old,10}(k) \cup \mathbf{P}_{cross,90}(k)$$

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

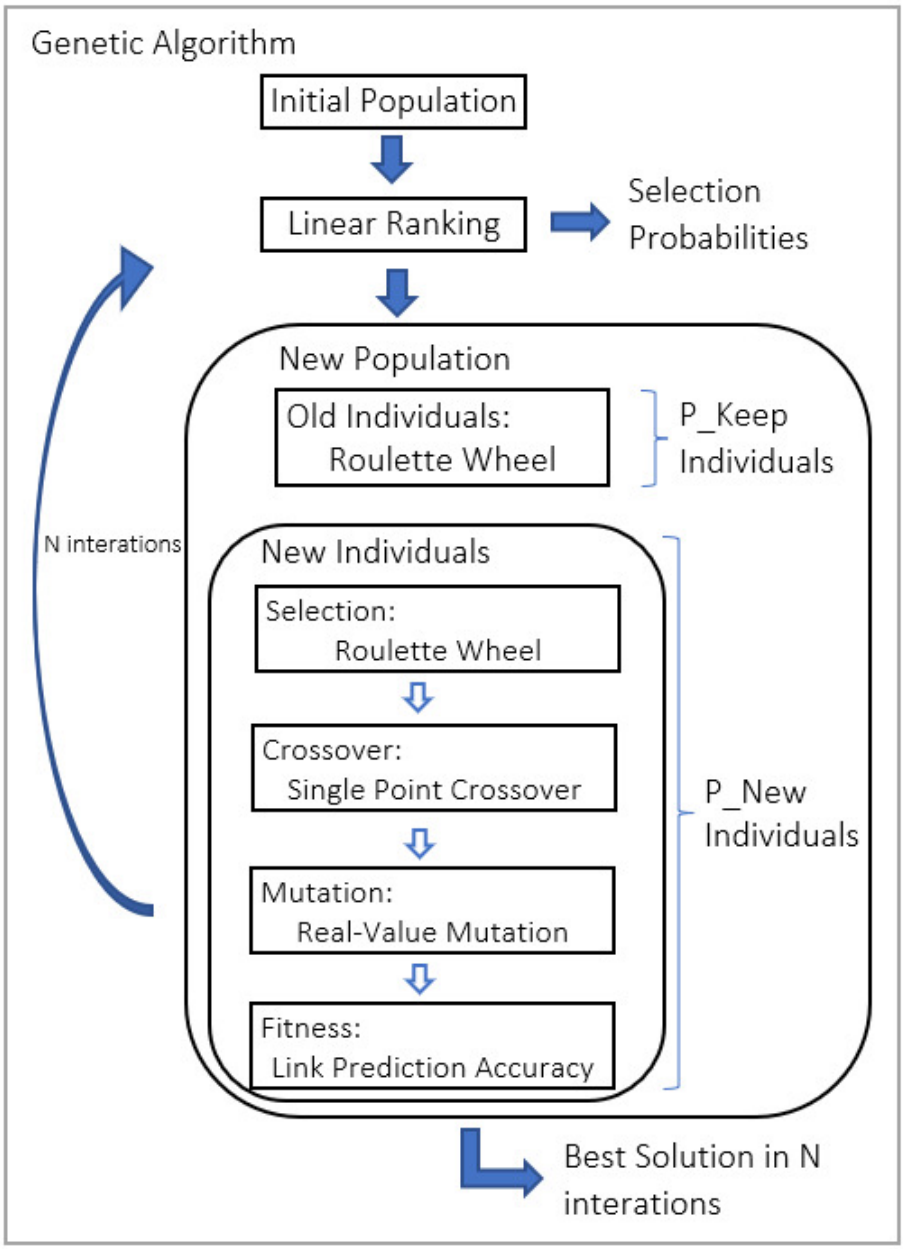

Figura 4.3: Diagrama de flujo del algoritmo genético usado para la búsqueda de parámetros óptimos del modelo ELCA

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 rendimiento del modelo propuesto en esta Tesis contra dos modelos clásicos de la literatura de *machine learning*, a saber, *Random Fo-*

rest y *Support Vector Machines*, en la tarea de predecir las decisiones de generación de contenido de los usuarios de una red social de tipo comunidad de práctica. Para la realización de los experimentos se extrajo un año del total de los datos proporcionados por los administradores de Plexilandia (que ha sido descrita en el Capítulo 3). Específicamente, 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 de conversación,
- ID de subforo,
- Contenido de texto de la publicación y
- Hora de publicación

En particular, al contenido de texto de cada publicación se le aplicó 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 proporcionada por los administradores de la red sobre los miembros clave de la red, distinguiendo entre los 4 tipos de usuario detallados en la Sección 3.1.1.

Posteriormente, de acuerdo con la topología de la red neuronal 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. Damos una aproximación de

la razón de desequilibrio (IBR) de cada sub-foro computado como el número de posibles contribuciones de contenido, es decir, número de usuarios activos multiplicado por el número de hilos activos, dividido por el número de publicaciones reales. 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 (entrenamiento) de los parámetros para construir el modelo. El resto de los meses se usa como datos de prueba para el testeo del modelo. En otras palabras, el 8 % de los datos es utilizado para la estimación de los parámetros ELCA óptimos mediante un algoritmo genético, y 92 % para testeo. Por lo tanto, la validación del modelo se establece en el marco de la escasez de datos de entrenamiento, que es más realista que abundancia de datos de entrenamiento (como cuando se usa 70 % para entrenamiento, 30 % para testeo) al intentar predecir la evolución en línea de una OSN.

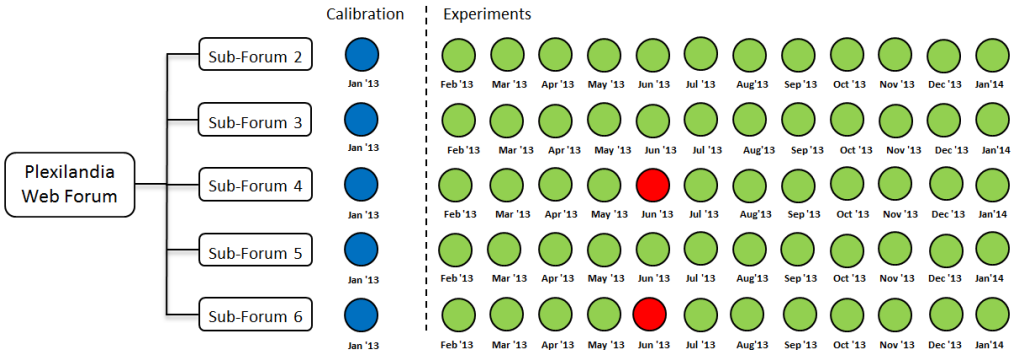

Figura 5.1: Configuración Experimental. En azul los meses cuyos datos se usan para calibrar los parámetros de la red. En verde los meses cuyos datos se usan para validar el modelo. En rojo, los meses que no tienen datos suficientes.

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.

Tabla 5.1: Usuarios activos (users), hilos activos (threads) y publicaciones realizadas (posts) 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

Procedemos, como se describe en la Sección 4.2.2, a asignar la representación del vector de texto para cada mes a cada hilo de conversación activo durante ese mes usando el vector promedio de todas las publicaciones que pertenecen a ese hilo durante ese mes. En cuanto a los usuarios, asignamos m representaciones vectoriales de los textos agrupando las publicaciones según el hilo al que pertenecen y calculando el vector medio del grupo de publicaciones generado por el usuario. De esta forma también recuperamos el número de hilos (m) con los que un usuario forma un enlace en nuestra representación de red propuesta.

Tabla 5.2: Usuarios activos (users), hilos activos (threads) y publicaciones realizadas (posts) 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

5.2. Métricas y Metodología de Evaluación

Calcularemos 4 medidas de rendimiento de los modelos. A saber, recuperación (*recall*), exactitud (*accuracy*), precisión (*precision*) y medida F (F *score*) [76]. A continuación se proporciona una descripción de estas métricas, donde PC significa Positivos Ciertos (*True Positive*) y NC son los Negativos Ciertos (*True Negative*).

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

$$Recuperación = \frac{\text{N enlaces TC}}{\text{N enlaces reales}} \tag{5.1}$$

Tabla 5.3: Usuarios activos(users), hilos activos (threads) y publicaciones realizadas (posts) 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**

Tabla 5.4: Razón de desequilibrio (IBR)

	SF2	**SF3**	**SF4**	**SF5**	**SF6**
IBR	31.43	27.86	17.6	30.48	61.32

- La exactitud 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:

$$Exactitud = \frac{\text{N enlaces PC} + \text{N enlaces NC}}{\text{N enlaces reales}} \tag{5.2}$$

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

$$Precision = \frac{\text{N enlaces PC}}{\text{N enlaces predichos}} \tag{5.3}$$

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

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

Para realizar la comparación del rendimiento se utilizará preferentemente la medida F debido a que es una métrica más confiable cuando se trabaja con conjuntos de datos (*datasets*) que presentan desequilibrio de clases, lo que se produce en el caso de estudio, donde el número de enlaces inexistentes (lo que denominamos negativos en el cálculo de las medidas de rendimiento) es mucho mayor que el número de enlaces reales (positivos). El diseño experimental es el siguiente:

1. Para cada subforo se calculan las predicciones de las contribuciones para cada mes, en las siguientes fases

   a) utilizando el primer mes de datos del subforo (enero 2013) se aplica el algoritmo genético para estimar los parámetros óptimos de la simulación de LCA para ese subforo

   b) para cada mes se realiza la simulación de LCA para predecir las contribuciones de los usuario durante ese mes

2. se calcula para cada mes y subforo las medidas de rendimiento

5.3. Resultados de Calibración

Ejecutamos el algoritmo genético obteniendo los parámetros optimos para el modelo ELCA para cada uno de los subforos usando los datos del mes

de calibración correspondiente (Enero 2013). Los valores se muestran en las Tablas 5.5 y 5.6.

Tabla 5.5: 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

Tabla 5.6: Valores calibrados de λ

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

Con estos valores de parámetros, se usa el modelo ELCA para simular el comportamiento de los usuarios en las redes de cada subforo y de cada mes entre febrero de 2013 y enero de 2014.

5.4. Resultados Experimentales

Como se especifica en Algoritmo 1, el resultado de la modelo ELCA son pares compuestos por las identificaciones del hilo de conversacion y el usuario

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

que deben ser interpretados como predictores de los pares reales que se pueden extraer de la verdad del terreno proporcionada por los administradores de

la red y que viene dada por las publicaciones de posts actualmente realizadas por los usuarios, a lo que hemos denominado enlaces reales en la definición de las medidas de rendimiento.

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

Se realizan predicciones independientes para cada mes y subforo. Estos conjuntos de pares se pueden visualizar como los arcos de grafos bipartitos, que son los grafos de publicación predichos y reales. Definimos los verdaderos positivos como las aristas que están en ambos grafos, verdaderos negativos como las aristas que están ausentes de los dos grafos, falsos positivos son las aristas que aparecen en el grafo de predicción pero están ausentes en el grafo de la verdad del terreno, y falsos negativos las aristas que están ausentes en el grafo de predicción pero aparecen en el grafo de la verdad del terreno.

A partir de las predicciones generadas por el modelo ELCA 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 de rendimiento definidas arriba. 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, así como también los resultados obtenidos para la medida F de los modelos RF y SVM entrenados sobre los mismos datos que ELCA.

5.4.1. Sub-Foro 2

En la Tabla 5.7 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 5.7: 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 5.8 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.

5.4.1.1. Mejor 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 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 5.9. 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

Tabla 5.8: 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**

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.

5.4.1.2. Peor resultado en el Subforo 2

En la Tabla 5.10 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 5.10. Al igual que en la imagen anterior, los hilos de conversación son representados por nodos de color violeta y los

Tabla 5.9: 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		

usuarios por nodos de color negro. Asimismo, los arcos siguen el codigo de color: negro = arco predicho correctamente, verde = arcos predicho incorrectamente, 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.

5.4.2. Sub-Foro 3

En la Tabla 5.11 se presentan los resultados obtenidos para cada una de las métricas en 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 rendimiento ligeramente inferior en este Sub-Foro en contraste al obtenido para el Sub-Foro 2.

Analogamente, en la Tabla 5.12 se muestran los resultados obtenidos de la medida F para los modelos RF y SVM para el Sub-Foro 3. Al igual que para el Sub-Foro 2, ambos modelos de referencia obtuvieron resultados similares

Tabla 5.10: 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

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.

5.4.2.1. Mejor 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 13 derivadas de los resultados

Tabla 5.11: 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**

del modelo. En la Fig. 5.4 se observa la red del subforo 3 para el mes 13, reconstruida a partir de la reglas de decisión de publicación presentadas en la Tabla 5.13. 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: 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**

5.4.2.2. Peor resultado en el Subforo 3

En la Tabla 5.14 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 5.14 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 mes 13, y cabe mencionar que se obtiene un resultado levemente inferior al obtenido para el mes 4 del Sub-Foro

Tabla 5.13: 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		

2.

5.4.3. Sub-Foro 4

En la Tabla 5.15 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 medida F obtenidos para este Sub-Foro superan a los obtenidos en los Sub-Foros analizados previamente.

Complementariamente, en la Tabla 5.16 se muestran los resultados obtenidos

Tabla 5.14: 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		

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	**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**

5.4.3.1. Mejor 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 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 5.17, 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

Tabla 5.16: 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**

considerablemente mejor que en los mejores casos de ambos.

5.4.3.2. Peor resultado en el Subforo 4

En la Tabla 5.18 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 5.18. 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,

Tabla 5.17: 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

este caso tiene mejor desempeño que los peores casos de ambos de los Sub-Foros estudiados con anterioridad.

5.4.4. Sub-Foro 5

En la Tabla 5.19 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 5.20 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: 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		

5.4.4.1. Mejor 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 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 5.21. 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	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**

5.4.4.2. Peor resultado en el Subforo 5

En la Tabla 5.22 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 5.22. 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 rendimiento de todos. No obstante, sigue siendo mejor resultado que los obtenidos por los modelos de referencia.

Tabla 5.20: 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**

5.4.5. Sub-Foro 6

En la Tabla 5.23 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 5.24 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,

Tabla 5.21: 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		

el modelo ELCA sigue mostrándose significativamente superior, de manera sostenida a lo largo de todo el horizonte temporal evaluado.

5.4.5.1. Mejor 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 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 5.25. 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.

5.4.5.2. Peor resultado en el Subforo 6

En la Tabla 5.26 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 5.26. 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.

5.5. 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. Se puede observar que, en general, la mayoría de los arcos de las redes 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 del terreno 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 precisión, en la

Tablas 5.7,5.11,5.15,5.19 y 5.23, frente 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 rendimiento 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 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, creemos que 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 el análisis, 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 de la 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 [53]. 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

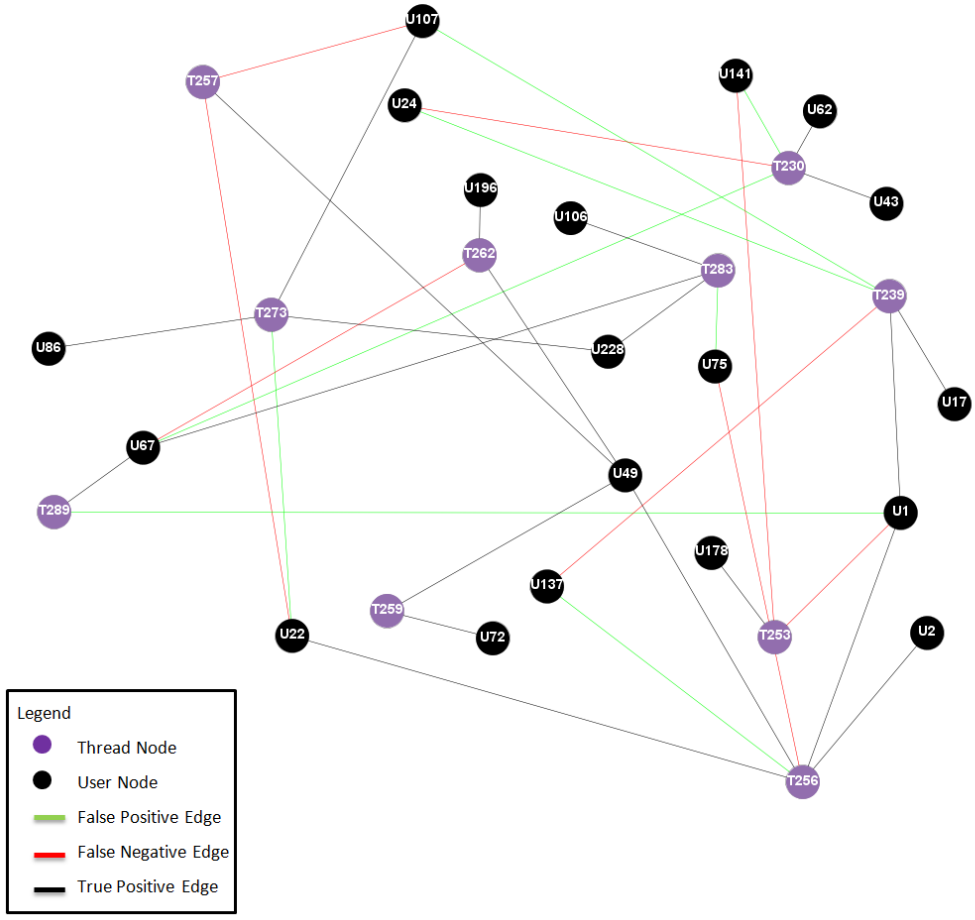

Figura 5.2: Red del Sub-Foro 2 para el Mes 2

Recall	Accuracy	Precision	F-measure
0.435	0.910	0.457	0.446

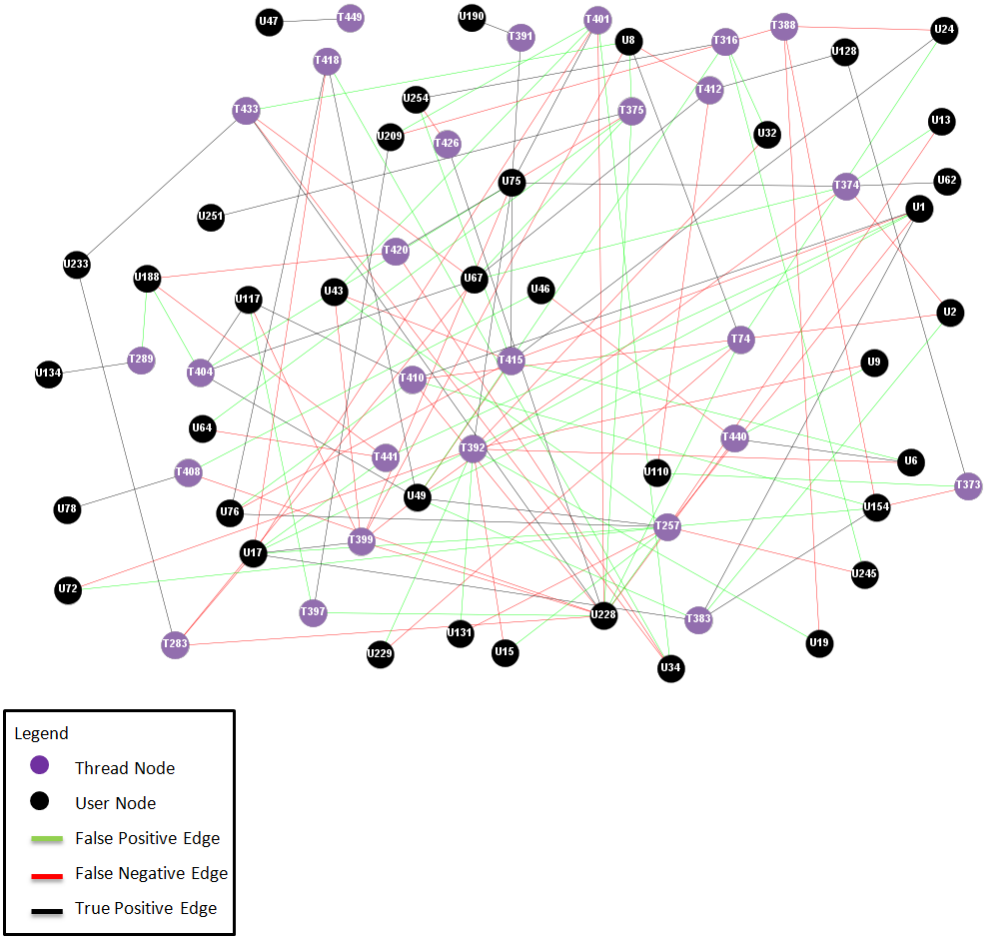

Figura 5.3: Red del Sub-Foro 2 para el Mes 4

Recall	Accuracy	Precision	F-measure
0.632	0.951	0.672	0.652

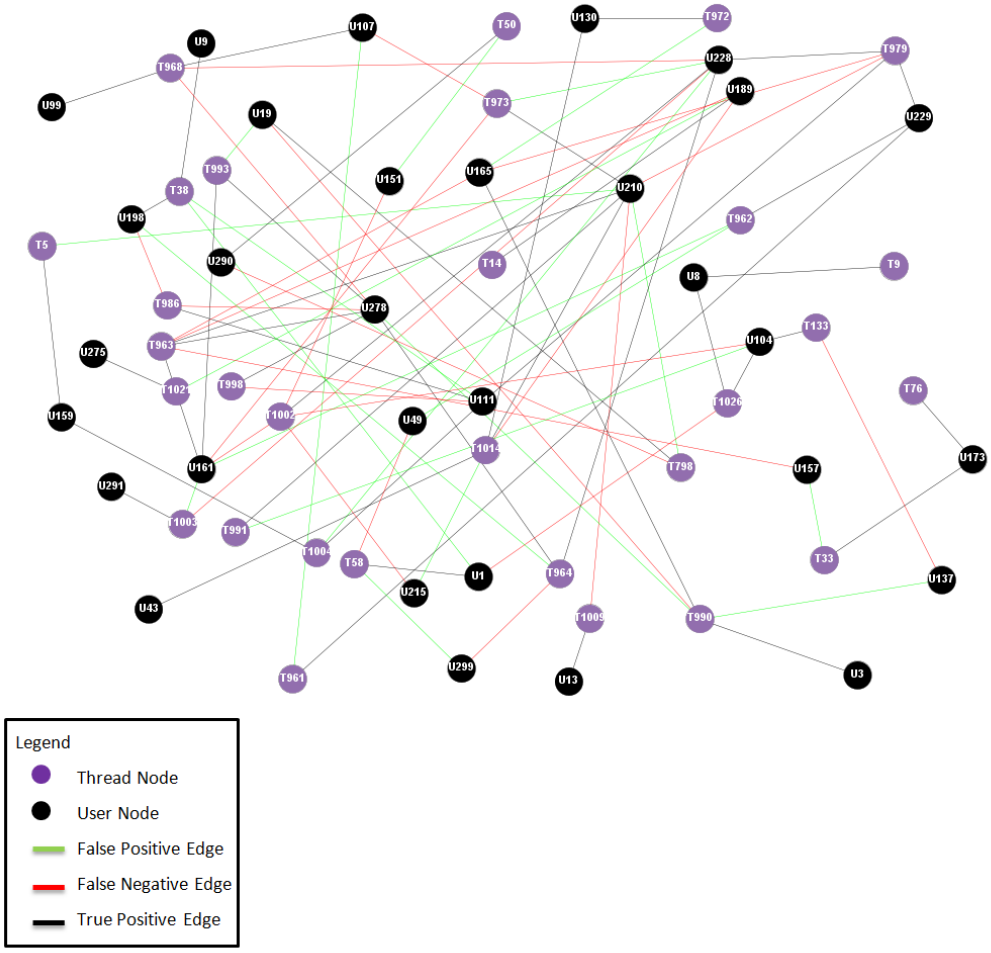

Figura 5.4: Red del Sub-Foro 3 para el Mes 13

Recall	Accuracy	Precision	F-measure
0.410	0.917	0.445	0.427

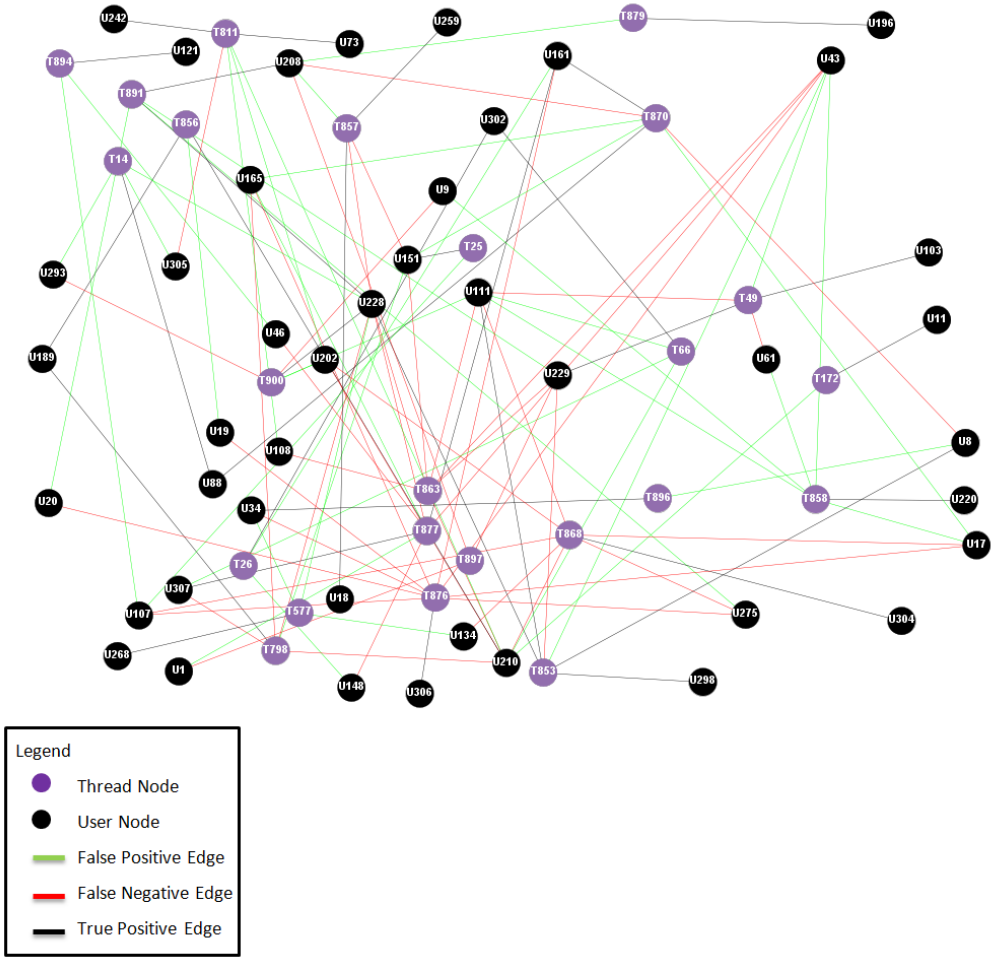

Figura 5.5: Red del Sub-Foro 3 para el Mes 11

Recall	Accuracy	Precision	F-measure
0.778	0.917	0.778	0.778

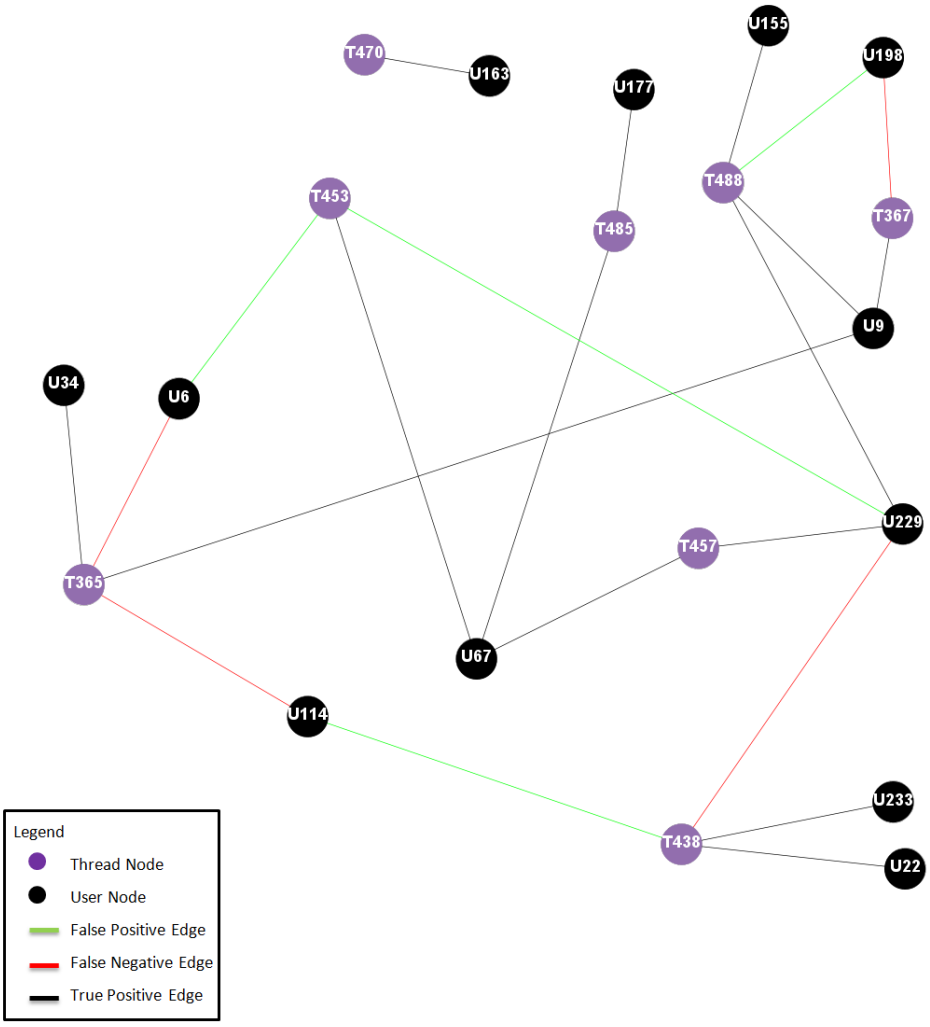

Figura 5.6: Red del Sub-Foro 4 para el Mes 5

Recall	Accuracy	Precision	F-measure
0.454	0.831	0.5	0.476

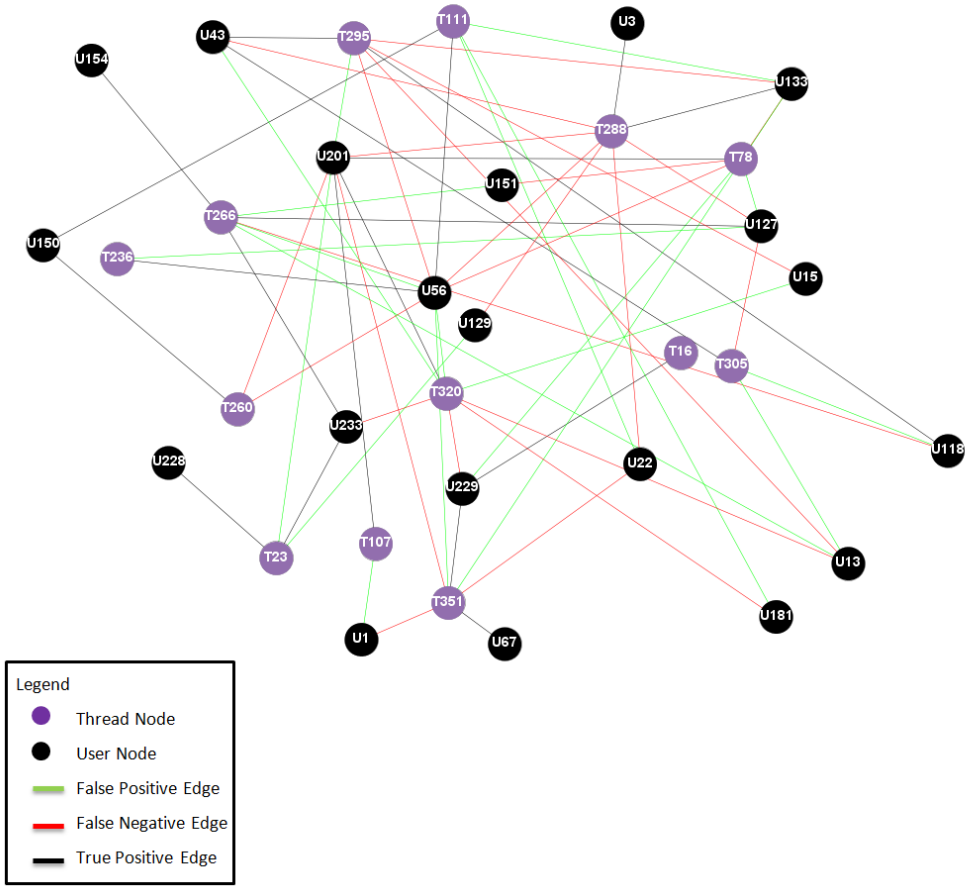

Figura 5.7: Red del Sub-Foro 4 para el Mes 3

Recall	Accuracy	Precision	F-measure
0.674	0.939	0.689	0.681

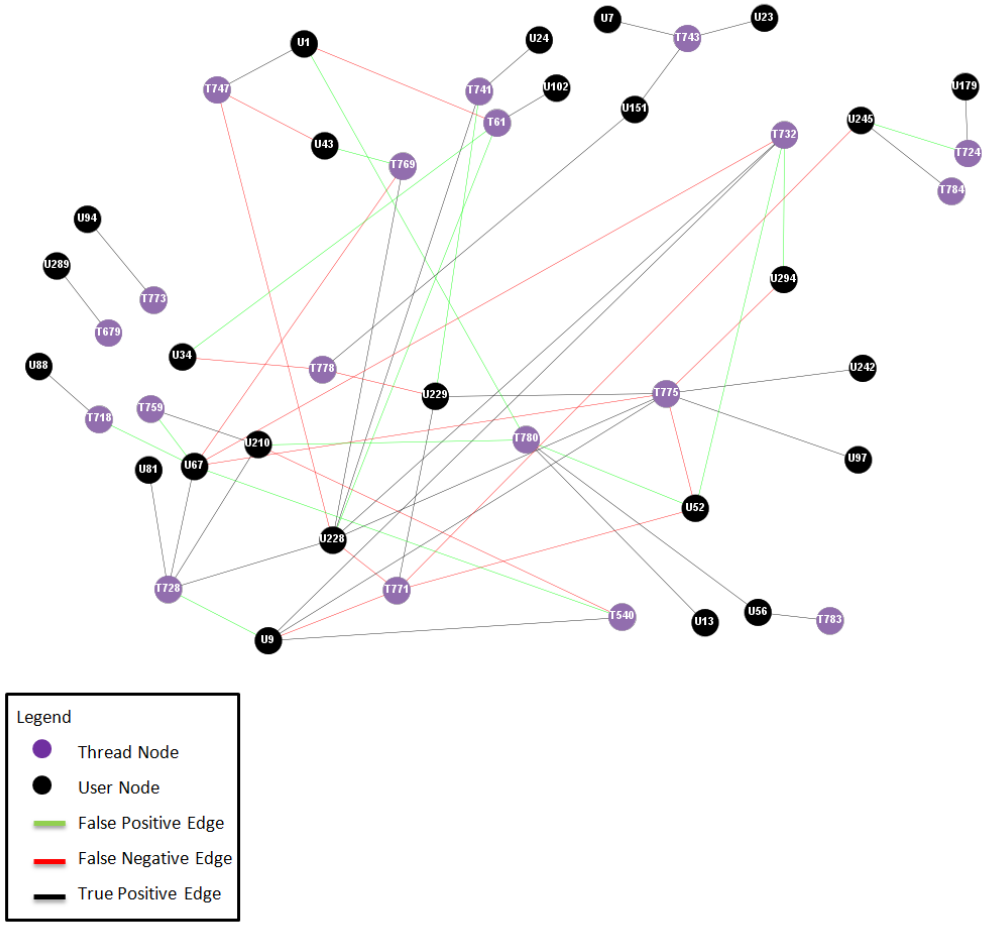

Figura 5.8: Red del Sub-Foro 5 para el Mes 9

Tabla 5.22: 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

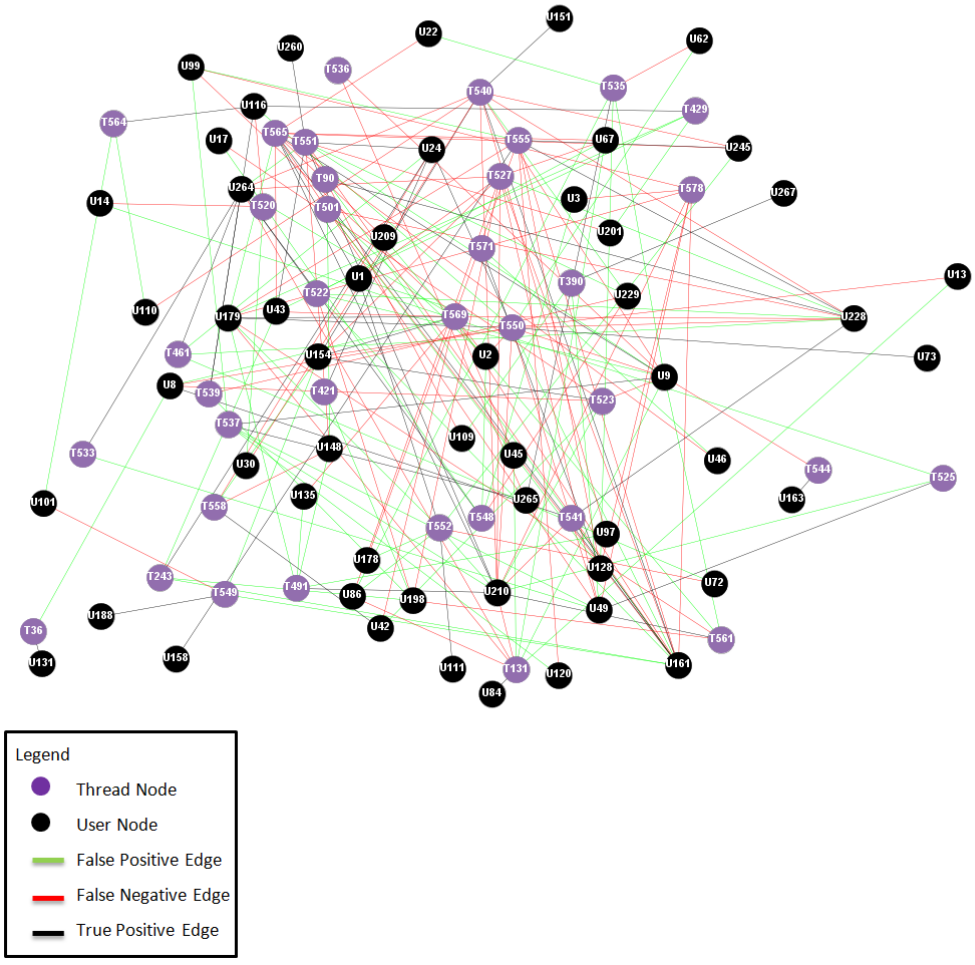

Figura 5.9: Red del Sub-Foro 5 para el Mes 6

Tabla 5.23: 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 5.24: 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 5.25: 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 5.26: 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

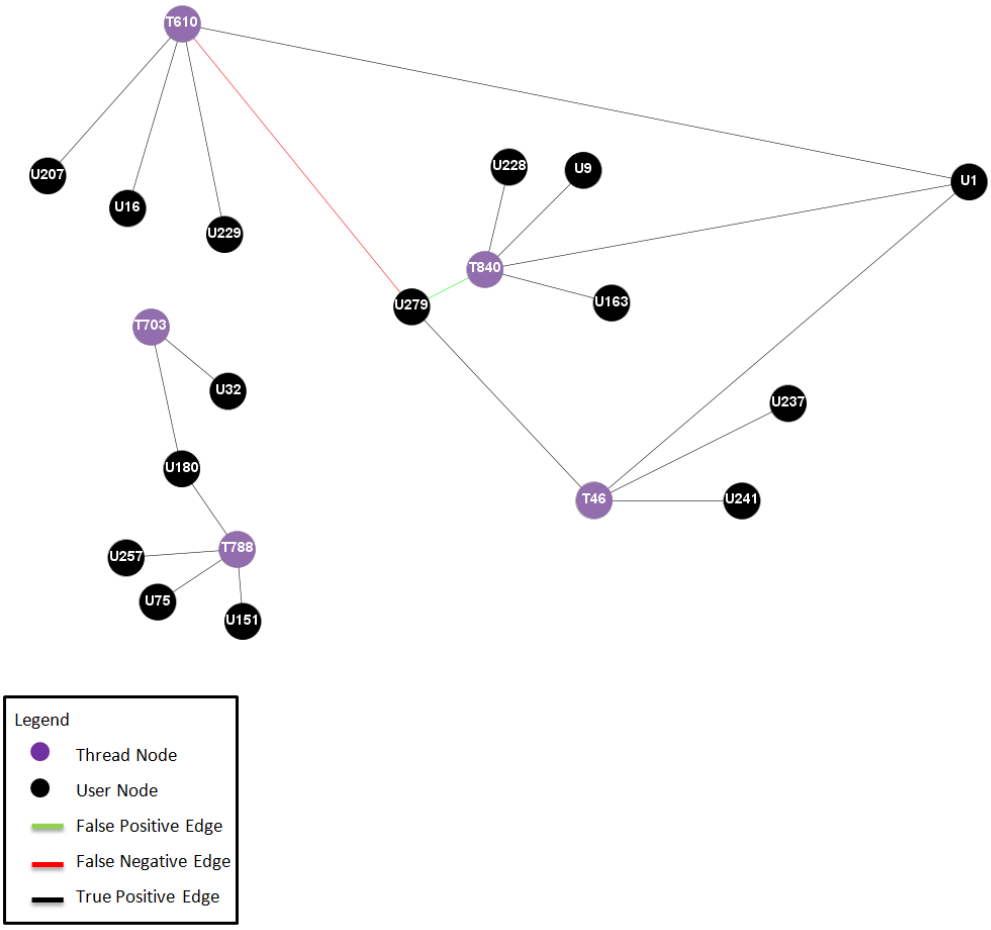

Figura 5.10: Red del Sub-Foro 6 para el Mes 10

Recall	Accuracy	Precision	F-measure
0.647	0.848	0.733	0.688

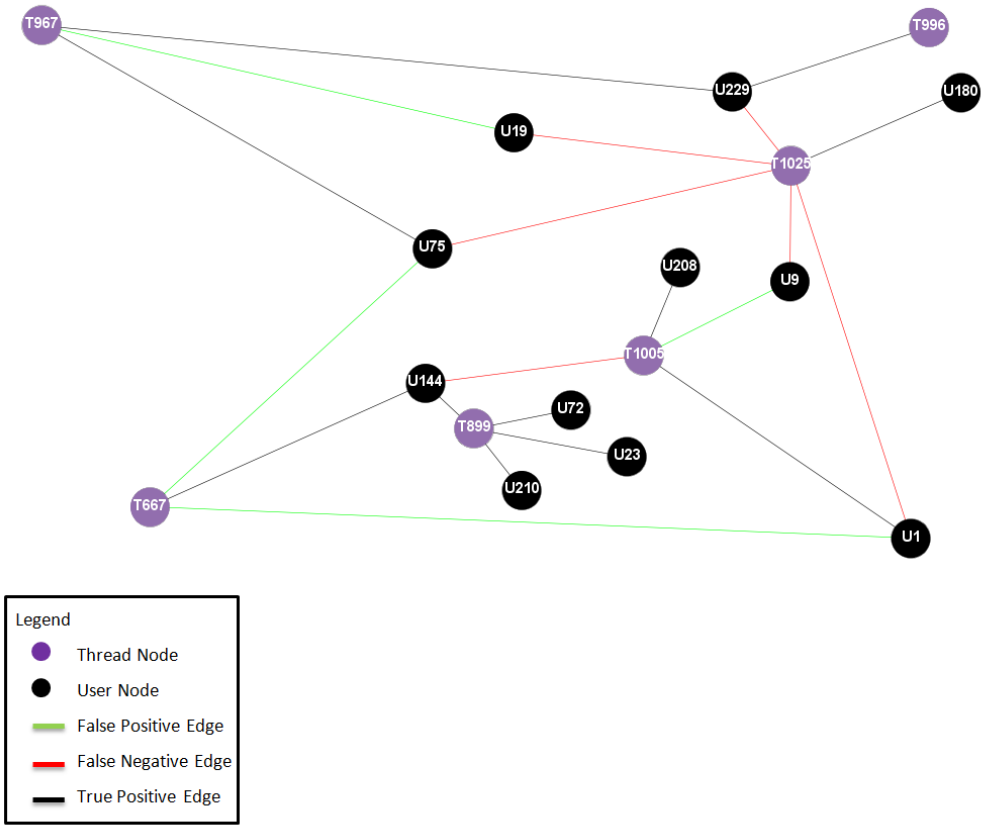

Figura 5.11: Red del Sub-Foro 6 para el Mes 13

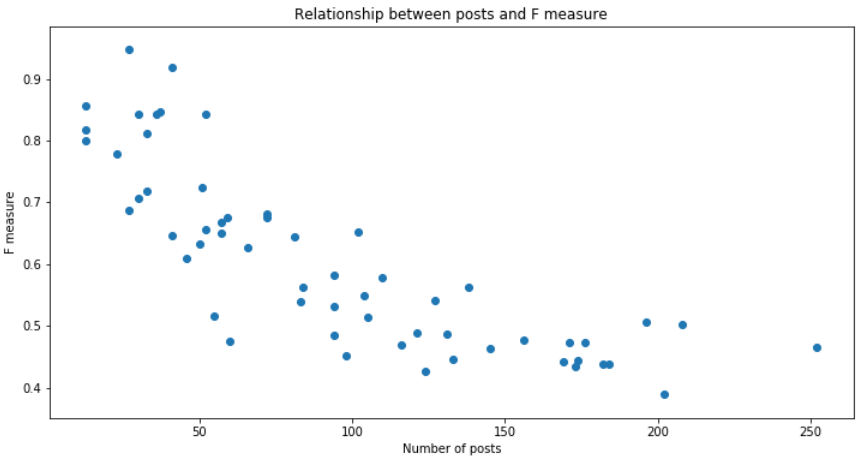

Figura 5.12: Relación entre el número de publicaciones y F-measure score

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 mesóscopico, pero no a nivel microscópico, a nivel de invividuo. 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 implementado via un algoritmo genético. Una de las novedades de este trabajo consiste en que se estiman los parámetros del modelo 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 del modelo LCA con el objetivo 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 mejor puntuación F del modelo ELCA y su promedio 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, este trabajo valida la hipótesis de investigación planteada al mostrar que el modelo ELCA, un modelo basado en el contenido semántico de las publicaciones, es capaz de modelar las decisiones de contribución de contenido de los usuarios en una red social conformada como un foro web de manera exitosa.

6.2. Trabajo futuro

El trabajo futuro se puede subdividir en 4 aspectos principales.

Primeramente, consideramos 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 interesante es la métrica definida sobre 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 lugar, 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 [77]. Dentro de las posibilidades a explorar se encuentra la utilización de *word embeddings* permitirían 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 5.5 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. Cómo generar series de contribuciones verosímiles es un problema que no se puede tratar fácilmente con aproximaciones interpolativas como SMOTE. Quizás las aplicaciones generativas como las *Generative Adversarial Networks* puedan aportar soluciones en un futuro próximo.

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] M. Granovetter. "Threshold Models of Collective Behavior". En: *The American Journal of Sociology* 83.6 (1978), págs. 1420-1443.

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

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

[7] 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.

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

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

[10] 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.

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

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

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

[14] Leo Breiman. "Random forests". En: *Machine learning* 45.1 (2001), págs. 5-32.

[15] Theodoros Evgeniou y Massimiliano Pontil. "Support Vector Machines: Theory and Applications". En: vol. 2049. Ene. de 2001, págs. 249-257. DOI: 10.1007/3-540-44673-7_12.

[16] J. Goldenberg. "Talk of the Network : A Complex Systems Look at the Underlying Process of Word-of-Mouth". En: *Marketing Letters* (2001), págs. 211-223.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.

[25] Rafal Bogacz y col. "Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice." eng. En: *Philos Trans R Soc Lond B Biol Sci* 362.1485 (2007), págs. 1655-1670. ISSN: 0962-8436 (Print); 1471-2970 (Electronic); 0962-8436 (Linking). DOI: 10.1098/rstb.2007.2059.

[26] 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.

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

[28] 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.

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

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

[31] 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.

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

[33] 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.

[34] X. H. Phang y CT. Nguyen. “Gibbslda++”. En: (2008).

[35] Ennio Cascetta. “Random Utility Theory”. En: *Transportation Systems Analysis: Models and Applications*. Boston, MA: Springer US, 2009, págs. 89-167. ISBN: 978-0-387-75857-2. DOI: 10.1007/978-0-387-75857-2_3. URL: https://doi.org/10.1007/978-0-387-75857-2_3.

[36] Meeyoung Cha, Alan Mislove y Krishna-P. Gummadi. “A Measurement-driven Analysis of Information Propagation in the Flickr Social Network”. En: *WWW 2009, April 20–24, 2009, Madrid, Spain.* Ed. por ACM. 2009, págs. 721-730.

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

[38] 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.

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

[40] 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).

[41] 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.

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

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] Konstantinos Tsetsos y col. “Using Time-Varying Evidence to Test Models of Decision Dynamics: Bounded Diffusion vs. the Leaky Competing Accumulator Model”. En: *Frontiers in Neuroscience* 6 (2012), pág. 79. ISSN: 1662-453X. DOI: 10.3389/fnins.2012.00079. URL: https://www.frontiersin.org/article/10.3389/fnins.2012.00079.

[58] 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.

[59] 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.

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

[61] 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.

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

[63] Kazumi Saito y col. "Detecting changes in content and posting time distributions in social media". En: *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. 2013, págs. 572-578. DOI: 10.1145/2492517.2492618.

[64] 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.

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

[66] 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.

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

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

[69] Y. Feng, B. Bai y W. Chen. "Information diffusion efficiency in online social networks". En: *2015 IEEE International Conference on Digital Signal Processing (DSP)*. 2015, págs. 1138-1142. DOI: 10.1109/ICDSP.2015.7252057.

[70] Przemyslaw Grabowicz, Niloy Ganguly y Krishna Gummadi. "Microscopic Description and Prediction of Information Diffusion in Social Media: Quantifying the Impact of Topical Interests". En: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: Association for Computing Machinery, 2015, págs. 621-622. ISBN: 9781450334730. DOI: 10.1145/2740908.2744106. URL: https://doi.org/10.1145/2740908.2744106.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

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

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

[77] 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.

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

[79] Weidong Liu y col. "Discovering the core semantics of event from social media". En: *Future Generation Computer Systems* 64 (2016), págs. 175-185. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2015.11.023. URL: https://www.sciencedirect.com/science/article/pii/S0167739X15003805.

[80] 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.

[81] 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.

[82] Maryam Mohammed Aldarwish y Hafiz Farooq Ahmad. "Predicting Depression Levels Using Social Media Posts". En: *2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*. 2017, págs. 277-280. DOI: 10.1109/ISADS.2017.41.

[83] Lulwah AlSuwaidan y Mourad Ykhlef. "A Novel Information Diffusion Model for Online Social Networks". En: *Proceedings of the 19th International Conference on Information Integration and Web-Based Applications & Services*. iiWAS '17. Salzburg, Austria: Association for Computing Machinery, 2017, págs. 116-120. ISBN: 9781450352994.

DOI: 10.1145/3151759.3151812. URL: https://doi.org/10.1145/3151759.3151812.

[84] Carmen De Maio y col. "Unfolding social content evolution along time and semantics". En: *Future Generation Computer Systems* 66 (2017), págs. 146-159. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2016.05.039. URL: https://www.sciencedirect.com/science/article/pii/S0167739X16301649.

[85] 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.

[86] D. Li y col. "Modeling Information Diffusion over Social Networks for Temporal Dynamic Prediction". En: *IEEE Transactions on Knowledge and Data Engineering* 29.9 (2017), págs. 1985-1997. DOI: 10.1109/TKDE.2017.2702162.

[87] M. Li y col. "A Survey on Information Diffusion in Online Social Networks: Models and Methods." En: *Information* 8 (2017), pág. 118.

[88] Steven Miletić y col. "Parameter recovery for the Leaky Competing Accumulator model". En: *Journal of Mathematical Psychology* 76 (2017), págs. 25-50. ISSN: 0022-2496. DOI: https://doi.org/10.1016/j.jmp.2016.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0022249616301663.

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

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

[91] 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.

[92] Shuai Zhao, Le Yu y Bo Cheng. "Probabilistic Community Using Link and Content for Social Networks". En: *IEEE Access* 5 (2017), págs. 27189-27202. DOI: 10.1109/ACCESS.2017.2774798.

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

[94] Pearl Keitemoge. "Technology Threats:Impacts of Cyberbullying to Today's Generation". En: *2018 15th International Conference on Service Systems and Service Management (ICSSSM)*. 2018, págs. 1-6. DOI: 10.1109/ICSSSM.2018.8464953.

[95] Bingquan Liu y col. "Content-Oriented User Modeling for Personalized Response Ranking in Chatbots". En: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.1 (2018), págs. 122-133. DOI: 10.1109/TASLP.2017.2763243.

[96] Jinshan Qi y col. "Discrete time information diffusion in online social networks: micro and macro perspectives". En: *Scientific Reports* 8.1 (2018), pág. 11872. DOI: 10.1038/s41598-018-29733-8. URL: https://doi.org/10.1038/s41598-018-29733-8.

[97] Yalin E. Sagduyu, Alexander Grushin y Yi Shi. "Synthetic Social Media Data Generation". En: *IEEE Transactions on Computational Social Systems* 5.3 (2018), págs. 605-620. DOI: 10.1109/TCSS.2018.2854668.

[98] Dimitri Demergis. "Predicting Eurovision Song Contest Results by Interpreting the Tweets of Eurovision Fans". En: *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. 2019, págs. 521-528. DOI: 10.1109/SNAMS.2019.8931875.

[99] Abdelaziz Khaled, Samir Ouchani y Chemseddine Chohra. "Recommendations-based on semantic analysis of social networks in learning environments". En: *Computers in Human Behavior* 101 (2019), págs. 435-449. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2018.08.051. URL: https://www.sciencedirect.com/science/article/pii/S0747563218304266.

[100] Hiroshi Nagaya, Kazuko Uno e Hiroyuki A. Torii. "Tracking Topics of Influential Tweets on Fukushima Disaster Over Long Periods of Time". En: *2019 International Conference on Data Mining Workshops (ICDMW)*. 2019, págs. 13-16. DOI: 10.1109/ICDMW.2019.00010.

[101] Feras Al-Obeidat y col. "Cone-KG: A Semantic Knowledge Graph with News Content and Social Context for Studying Covid-19 News Articles on Social Media". En: *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. 2020, págs. 1-7. DOI: 10.1109/SNAMS52053.2020.9336541.

[102] Richard J. Binney y Richard Ramsey. "Social Semantics: The role of conceptual knowledge and cognitive control in a neurobiological model of the social brain". En: *Neuroscience & Biobehavioral Reviews* 112 (2020), págs. 28-38. ISSN: 0149-7634. DOI: https://doi.org/10.1016/j.neubiorev.2020.01.030. URL: https://www.sciencedirect.com/science/article/pii/S0149763419301915.

[103] Aarzoo Dhiman y Durga Toshniwal. "An Approximate Model for Event Detection From Twitter Data". En: *IEEE Access* 8 (2020), págs. 122168-122184. DOI: 10.1109/ACCESS.2020.3007004.

[104] Poonam Goyal y col. "Multilevel Event Detection, Storyline Generation, and Summarization for Tweet Streams". En: *IEEE Transactions on Computational Social Systems* 7.1 (2020), págs. 8-23. DOI: 10.1109/TCSS.2019.2954116.

[105] Marcos Grzeça, Karin Becker y Renata Galante. "Drink2Vec: Improving the classification of alcohol-related tweets using distributional semantics and external contextual enrichment". En: *Information Processing & Management* 57.6 (2020), pág. 102369. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2020.102369. URL: https://www.sciencedirect.com/science/article/pii/S0306457320308645.

[106] Hui Jiang y col. "Community Detection Based on Individual Topics and Network Topology in Social Networks". En: *IEEE Access* 8 (2020), págs. 124414-124423. DOI: 10.1109/ACCESS.2020.3005935.

[107] Zafran Khan y col. "DST-HRS: A topic driven hybrid recommender system based on deep semantics". En: *Computer Communications* 156 (2020), págs. 183-191. ISSN: 0140-3664. DOI: https://doi.org/10.1016/j.comcom.2020.02.068. URL: https://www.sciencedirect.com/science/article/pii/S0140366419304062.

[108] Sanjay Kumar y col. "Modeling information diffusion in online social networks using a modified forest-fire model". En: *Journal of Intelligent Information Systems* (2020). DOI: 10.1007/s10844-020-00623-8. URL: https://doi.org/10.1007/s10844-020-00623-8.

[109] Abiola Osho, Colin Goodman y George Amariucai. *MIDMod-OSN: A Microscopic-level Information Diffusion Model for Online Social Networks*. arXiv 2002.10522. 2020. arXiv: 2002.10522 [cs.SI].

[110] Abhinay Pandya y col. "On the use of distributed semantics of tweet metadata for user age prediction". En: *Future Generation Computer Systems* 102 (2020), págs. 437-452. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2019.08.018. URL: https://www.sciencedirect.com/science/article/pii/S0167739X19304509.

[111] Chi-Fai Lo y Ho-Yan Ip. "Modified leaky competing accumulator model of decision making with multiple alternatives: the Lie-algebraic approach". En: *Scientific Reports* 11.1 (2021), pág. 10923. DOI: 10.1038/

s41598-021-90356-7. URL: https://doi.org/10.1038/s41598-021-90356-7.

[112] Chen Luo y col. "Exploring public perceptions of the COVID-19 vaccine online from a cultural perspective: Semantic network analysis of two social media platforms in the United States and China". En: *Telematics and Informatics* 65 (2021), pág. 101712. ISSN: 0736-5853. DOI: https://doi.org/10.1016/j.tele.2021.101712. URL: https://www.sciencedirect.com/science/article/pii/S0736585321001519.