Identificación automática de usuarios conflictivos en grupos de Facebook

Francisco Serrano[1], Franco D. Berdun[2], Marcelo G. Armentano[2]

[1]Fac. Cs. Exactas, UNICEN, Campus Universitario,
Paraje Arroyo Seco, Tandil, Argentina
fserrano@exa.unicen.edu.ar

[2] ISISTAN Research Institute (CONICET / UNICEN), Campus Universitario,
Paraje Arroyo Seco, Tandil, Argentina
{franco.berdun, marcelo.armentano}@isistan.unicen.edu.ar

Resumen. Actualmente, las redes sociales cumplen un papel primordial en el modo en cómo interaccionan los grupos sociales. La creación de grupos específicos permite que los usuarios entren en contacto con otros usuarios con sus mismos intereses. Sin embargo, las redes sociales no están exentas de los conflictos que pueden surgir de las dinámicas grupales. Por lo cual, cada vez es más frecuente la necesidad de moderadores para identificar aquellos usuarios con conductas no deseadas y aplicar advertencias o sanciones. En este artículo, proponemos un enfoque para identificar a los usuarios conflictivos en una red social basado en la observación del comportamiento en grupos de interés de Facebook. Obtuvimos resultados preliminares prometedores que se pueden usar para asistir a administradores o moderadores con el análisis y mantenimiento de estos grupos.

1 Introducción

Las redes sociales tienen un profundo efecto en la forma en que las personas se relacionan y pasan su tiempo. Por lo tanto, constituyen un fenómeno significativo de especial interés para las organizaciones, los negocios y la sociedad. Si bien las redes sociales han traído varios beneficios en cuanto a la capacidad y flexibilidad en la comunicación entre individuos, existe una mayor conciencia de las controversias y consecuencias adversas que rodean al uso de estas redes sociales [1, 2].

A medida que las plataformas de medios sociales continúan proliferando, con beneficios aparentemente incalculables, es fácil pasar por alto las desventajas que conllevan [3, 4], que son de gran importancia. Entre dichas desventajas se pueden encontrar diversos aspectos como la presión del tiempo, la tergiversación, la adicción [5] y otros estresores psicológicos, que pueden afectar de sobremanera a los individuos, al punto de poder provocarles serias repercusiones negativas en la vida cotidiana [6]. Si bien las redes sociales han traído muchas novedades que mejoran la forma en que nos podemos

relacionar, colateralmente generan escenarios que pueden propiciar la aparición de situaciones de conductas excesivas. Estos excesos, en caso de poder ser detectados de forma automática, se podrían ver reducidos o prevenidos debido a que dicha detección, habilitaría la posibilidad de tomar decisiones correctivas o preventivas más rápidamente.

Entre las posibilidades de interacción que ofrecen las redes sociales, se destaca la colaboración de usuarios a través de grupos de interés [7]. Cada individuo puede establecer, mediante una publicación en el muro de un grupo específico, una comunicación con un círculo de individuos con intereses comunes. Esta dinámica puede ser estudiada mediante el análisis de las interacciones entre dichos individuos del grupo, para así extraer información útil sobre el comportamiento de cada uno. Un marco teórico para el estudio de este tipo de interacciones es Symlog [8, 9], que describe varios patrones de comportamiento, entre los cuales se encuentra la actitud dominante, que se compatibiliza con situaciones de excesos en la interacción dentro de los grupos de las redes sociales. Estas situaciones, correctamente modeladas, pueden ser clasificadas mediante métodos de aprendizaje supervisado, para así proveer la posibilidad de emitir alertas automáticas en caso de que se produzcan situaciones de interacción fuera de lo normal, avisando a una hipotética autoridad competente.

Además de la obtención de los indicadores Symlog, se trabaja con datos adicionales como por ejemplo el nivel de actividad que un usuario tiene en relación con el resto del grupo. Este indicador es de especial interés, ya que indirectamente repercute también en los indicadores de dominancia y sumisión asociados a un determinado individuo. Se buscaron diferentes vías de obtención de información a fin de perfeccionar el modelado del comportamiento de cada usuario, lo cual se traduce en mejores predicciones a la hora de la experimentación.

Una vez que se obtiene dicha información, se propone un enfoque materializado a través de un modelo predictivo, que pueda indicar si el comportamiento de una determinada persona es compatible o no con algún tipo de exceso, para así posibilitar la toma de decisiones correspondientes. Para la creación de dicho modelo se utiliza la teoría Symlog para el modelado del comportamiento de los individuos, además del uso de diversos indicadores que contribuyen a la detección de excesos en la interacción. Dicho modelado será automatizado a través del modelo propuesto.

La noción de modelado de usuario para la detección de diversas conductas, ya fue aplicada de forma exitosa por Berdun et al. [10] en otros contextos como por ejemplo en los juegos serios. El enfoque que allí se propuso permite mejorar el rendimiento de las tareas colaborativas a través de un correcto armado de los grupos en función a los perfiles individuales de cada individuo.

Por el momento se obtuvieron resultados experimentales, que servirán en el futuro para el desarrollo de una herramienta inteligente que pueda alertar de forma automática,

cualquier tipo de situación merecedora de atención, para que así puedan tomarse decisiones inmediatas en caso de tratarse de algún tipo de conducta. Si bien la herramienta inteligente inicialmente se desenvolverá en el contexto de grupos de Facebook, su dominio puede ser extendido a otros ámbitos como por ejemplo la comunicación verbal, en donde a través del modelado de secuencias, se podrán aplicar técnicas de aprendizaje profundo, a fin de obtener mejores resultados, sin modificar por ello la finalidad inicial, que es la detección automatizada de excesos en la comunicación entre individuos.

2 Fundamento teórico y herramientas empleadas

2.1 Trabajos Realizados

El estudio realizado por Kaplan y Haenlein [11] enumera todas las plataformas y servicios que normalmente se consideran redes sociales, así como también las tecnologías digitales emergentes, como las aplicaciones de realidad virtual/aumentada, y las tecnologías portátiles que se relacionan con las redes sociales. Entre ellas, Facebook (grupos de interés, como grupo controlado) es de nuestro interés.

Dentro de las redes sociales, otro aspecto importante son las conversaciones, en donde la interacción no se da a través de publicaciones sino también a través de mensajes entre dos o más individuos. Existen diversos estudios sobre el tema, en donde se identifica a la conflictividad en la comunicación, como por ejemplo en el realizado por Miah et al. [12], en donde se identifica a los usuarios conflictivos a través de la clasificación del texto de los mensajes, o el realizado por Michalopoulos et al. [13] en donde se proponen técnicas bayesianas y markovianas para la prevención de los ataques de grooming. Sin embargo, ambos estudios trabajan solamente sobre el texto, aplicando métodos de preprocesamiento tradicionales, además de trabajar en el contexto de conversaciones por mensajes.

2.2 Espacio compartido

En el enfoque presentado en este artículo, se utilizan los grupos de Facebook para modelar el concepto de grupos de interés. Un grupo en esta red social es generalmente creado con una determinada finalidad. Los integrantes que componen cada grupo, poseen algún tipo de intereses en común, beneficiándose de la finalidad del grupo de compartir estos intereses. Esto quiere decir que cada grupo de Facebook constituye un espacio de colaboración, por lo que dicha dinámica de colaboración que se da en cada grupo, puede ser estudiada a fin de detectar patrones de comportamiento dentro de dicho espacio. Cada publicación, junto con sus comentarios, se asemeja a la idea de una conversación por mensajes, en donde el que inicialmente publica, es el que envía el primer mensaje, siendo los comentarios de dicha publicación, los mensajes que le siguen a la publicación inicial.

Existen estudios que se han dedicado al análisis del contenido a partir del cual se podrían detectar situaciones de abusos, como por ejemplo el realizado por Rybnicek et

al. [14], en donde se detectan las fuentes de información de dichas situaciones. En la propuesta que se presenta, algunas de las fuentes detectadas están presentes, no sólo para ser analizadas, sino también para ser aplicadas para el desarrollo de un modelo predictivo que automatice la detección de situaciones de interacción excesiva.

2.3 Modelo de observación de grupos

La teoría Symlog distingue tres dimensiones estructurales en interacciones grupales: estado, atracción y orientación de objetivos. La primera dimensión analiza la actitud dominante (U) o sumisa (D) de quienes interactúan; la segunda estudia la tendencia positiva (P) o negativa (N); y finalmente, la tercera dimensión analiza la cuestión de si las personas están involucradas con la tarea (F) o con comportamientos socio-emocionales (B). Cada dimensión contempla además una posible conducta neutral, logrando de esta forma 27 combinaciones. A pesar de que Symlog fue desarrollado como una extensión del modelo IPA [15], estas dos teorías se complementan. Este último método permite codificar las conductas verbales acorde con dos categorías principales: la socio-emocional y la de tarea, para sub-clasificarlas luego en doce tipos diferentes: seis socio-emocionales y seis hacia la tarea. Adicionalmente, IPA provee una enumeración (del 1 al 12) de posibles conductas surgidas durante la actividad colaborativa y las clasifica según el tipo de reacción que significan (R1: positiva, R2: respuestas, R3: preguntas, R4: negativa).

Los atributos que define IPA para la orientación de los mensajes, son abordados a través de la tercera dimensión de Symlog (Orientación de Objetivo). Las dimensiones restantes, constituyen la extensión que Symlog realiza sobre IPA. Sin embargo, la dimensión de Atracción puede interpretarse como la consecuencia de cómo las reacciones Positivas y Negativas definen al usuario como Amistoso o No Amistoso; mientras que la dimensión de Estado, que abarca los atributos de Sumisión y Dominancia pueden interpretarse como el grado de y tipo de participación (Tabla 1). El atributo Dominancia, es de especial interés para el enfoque propuesto, ya que se compatibiliza con los excesos que se dan en las interacciones de un grupo de interés. Además, la determinación del atributo Dominancia se puede enriquecer a través de interacciones no textuales, es decir, a través de emoticones y/o reacciones (aplicado a Facebook, esto sería: me gusta, me enoja, me encanta, etc.)

Tabla 1. Formulación de los atributos Symlog a partir de IPA.

Symlog	IPA
# Dominante	# Negativo + # Requiere
# Sumiso	# Positivo + # Responde
# Amistoso	# Positivo
# No Amistoso	# Negativo
# Tarea	# Requiere + # Responde
# Socioemocional	# Positivo + # Negativo

Estos seis atributos de Symlog son los que se utilizan como modelo base para la observación de los grupos, ya que la información que proveen los atributos, resulta ser muy útil para la elaboración de diversos indicadores (desarrollados en la subsección siguiente) cuya finalidad es su posterior utilización en los Datasets del modelo predictivo que se desarrollará.

Durante el procesamiento de las contribuciones, es decir, de los posteos que cada usuario realiza en un grupo, se calculan dos tipos de indicadores: los indicadores de interacciones intragrupo y los indicadores de contribuciones individuales (ICI). Para este cálculo, se tienen en cuenta los atributos establecidos en Symlog, que a su vez son calculados a partir de las categorías IPA (ver sección 3). Los indicadores de contribuciones individuales e interacciones intragrupo, se crean con el objeto de indicar alteraciones en la normalidad de la conducta de un usuario. Para calcular el ICI, se computa la cantidad de intervenciones y el porcentaje relativo que cada uno de los usuarios manifestó en relación con cada una de las categorías. De esta forma es posible evaluar el rendimiento individual de cada uno de los miembros del grupo.

Procesar una contribución, implica realizar la clasificación de cada interacción como muestra de una determinada conducta de grupo. Una vez finalizado el procesamiento de una base de logs, se reconoce la existencia de conflictos o perturbaciones en la dinámica del grupo de usuarios. De esta manera se logra llevar a cabo acciones de alerta personalizadas para cada usuario.

Respecto al mapeo de interacciones a los modelos propuestos por Bales (IPA y Symlog) en particular, Berdun et al., [16] propusieron el registro de las interacciones del grupo basado en la clasificación de las mismas a partir de texto libre a los atributos IPA. En este trabajo, la interpretación de la conducta de los usuarios se limita en un contexto de pequeños grupos. Nuestro trabajo propone superar esta limitación, permitiendo la libre interacción entre los miembros.

Partiendo, entonces, de la caracterización hecha del modelo Symlog, de una red social y de un espacio de intereses compartidos para la interacción entre los participantes, en la siguiente sección describiremos el proceso experimental que se llevó a cabo para la clasificación automática de los usuarios.

2.4 Detección automática de perfiles conflictivos. Indicadores empleados.

Para cada individuo, inicialmente se clasificarán todas sus interacciones según la teoría IPA, utilizando para ello una herramienta destinada al preprocesamiento de los mensajes mediante el análisis del lenguaje natural desarrollada por Berdun et al. (2017). Cabe aclarar que, al referirse a interacciones, se tienen en cuenta no solo los posteos que un usuario realiza en el grupo, sino también los comentarios que se realizan en los diversos posteos. Una vez que se tienen los valores IPA asociados a cada interacción, se procede al armado de los indicadores Symlog, utilizando las fórmulas asociadas a cada uno de los atributos, expresadas previamente en la Sección 2.3.

Partiendo de estos seis atributos de Symlog, como se introdujo anteriormente, el de dominancia es de especial interés, ya que se trata de uno de los indicadores que, dentro del enfoque propuesto, contribuye mayoritariamente a la detección automatizada de conductas excesivas.

El problema que este concepto genera en un principio, es la confiabilidad del indicador, ya que una mayor cantidad de dominancia no necesariamente tiene que aumentar la probabilidad de que la conducta de un determinado individuo, sea catalogada como excesiva. Esto se debe a que, ante una eventual comparación entre grupos con cantidades muy dispares de posteos y comentarios, inevitablemente en el de mayor cantidad, la dominancia será mayor para sus integrantes, sin que efectivamente se trate de una conducta que se encuentre por fuera de lo normal.

Dicha cuestión se soluciona normalizando el indicador de dominancia, es decir, dividiendo al indicador por la cantidad de interacciones que el integrante tuvo en el grupo de interés, para así poder trabajar con valores porcentuales que no son sensibles a las cantidades de interacciones de los grupos. Para el atributo de sumisión, se procedió de forma análoga a modo de agregar un indicador más para el enfoque.

Otro indicador que se tuvo en cuenta para incorporar, fue el porcentaje de participación de un individuo dentro de un grupo, calculado a partir de la división entre la cantidad de interacciones de dicho individuo, y la del grupo. Si bien se trata de un indicador que a priori parece sencillo, su significado es de gran relevancia, ya que las conductas dominantes van emparentadas con la acaparación de la atención del grupo.

Luego se propone utilizar, el porcentaje de participación de un individuo, como penalización de la dominancia, es decir, que una participación acaparadora de un individuo, repercutirá en un aumento de su indicador de dominancia. Esta penalización se verá cada vez más agravada, conforme se aleje el porcentaje de participación del porcentaje promedio para todos los integrantes. Todo este concepto de penalización se basa en la idea de que la participación de un individuo dentro de un grupo, indica cierta intención de mostrarse activo ante el resto de los integrantes, lo cual explica que tenga una leve penalización, que en este caso sería deseable.

Presentada la idea detrás de este indicador, la fórmula para el Indicador de Dominancia Ponderado (IDP), asociado a cada integrante del grupo, es la siguiente:

$$IDP = (\#\ Dominante\ /\ Cantidad\ Mensajes\ Grupo) + Porcentaje\ de\ Participación \quad (1)$$

A fin de ejemplificar el uso de dicha penalización, se toma como persona de interés a uno de los integrantes de uno de los grupos que conforman el dataset (a detallar posteriormente en la sección 3.1). El usuario posee la cantidad de 1004 interacciones (entre posteos y comentarios), de las cuales 37 fueron clasificadas con el atributo de domi-

nancia. Esto arroja un índice normalizado de dominancia de 3.68, pero como el individuo posee un porcentaje de participación de 58.61, dicho porcentaje constituye la penalización, por lo que el IDP termina dando 62.29. El concepto que se quiere volcar, es la idea de que un alto porcentaje de participación produce una abultada penalización, lo cual se traduce en un IDP muy elevado.

Por último, se tuvieron en cuenta otros indicadores asociados a cada individuo, a fin de contribuir con el posterior armado del modelo. Uno de ellos es la división de los integrantes del grupo en siete intervalos, según el porcentaje de participación. Dentro de estos intervalos, el primero se corresponderá con una participación altamente acaparadora por parte del individuo, mientras que el último intervalo indicará participación mínima o nula. La definición de los límites inferiores y superiores de cada intervalo está parametrizada debido a la necesidad de trabajar con diferentes grupos de interés, lo cual hace variar las características de los integrantes de los mismos, así como los propósitos de cada grupo.

La forma en que se definen los intervalos a partir de los parámetros comienza con el promedio de participación de todos los integrantes de un grupo. A partir de dicho promedio, se agrega el Umbral de Participación Promedio, para constituir el límite superior del intervalo. Para el límite inferior, se parte del mismo promedio, pero restando el mismo Umbral.

Luego para el intervalo superior, por ejemplo, su límite inferior lógicamente es el mismo que el límite superior del intervalo anterior. El límite superior de este intervalo se calcula sumando, al límite inferior, el Umbral de Participación Levemente Elevada. Para el resto de los intervalos se procede de manera análoga. En caso de calcular un intervalo inferior al de participación promedio, se utiliza para ello el Umbral de Participación Levemente Inferior, y así sucesivamente hasta que haya intervalos que cubran toda la distribución de participaciones.

Otro indicador que se tomó en cuenta para el momento de realizar el entrenamiento, es el filtro de intención, que consiste básicamente en una alerta binaria en caso de que un usuario presentase vocabulario compatible con excesos verbales. Su funcionamiento es sencillo, ya que se basa en un diccionario de elementos en donde dicha alerta se emite para los casos en que un usuario postee o comente contenido textual que esté presente en el diccionario.

La última propuesta dentro del enfoque en cuanto a indicadores, fueron las clases que indican el grado de excesividad que un individuo está llevando a cabo dentro de un grupo. Estas clases son tres, a saber: positivo, negativo, y mid, siendo esta última una clase intermedia, en donde se considera a una conducta como sospechosa de ser peligrosa, para así no ser completamente descartada.

A modo de resumen, se utilizaron los siguientes indicadores asociados a cada individuo:

- Seis atributos Symlog
- Índices (ICI, IDP)
- Intervalos de Participación
- Filtro de Intención
- Clases de Conducta (positivo de abusividad, negativo y mid)

2.5 Detección Automática de Conductas Excesivas

Asociando los indicadores previamente expuestos, para cada individuo dentro de un grupo de interés, se establece que las clases de conducta serán las etiquetas de cada conducta (positivo, negativo, mid), para luego poder llevar a cabo el correspondiente aprendizaje automático. Para la obtención de resultados experimentales, se utilizaron tres algoritmos de clasificación, para tres datasets, a detallar en la sección 3. De esta manera se cierra el flujo de información para el modelo de detección de conductas excesivas, proporcionando en la Figura 1, una descripción gráfica del proceso. En dicha figura, se muestran dos clases en la salida del clasificador (Si / No), ya que se trata de la primera etapa de la experimentación. En la segunda etapa, se decidió incorporar la tercera clase (mid) para intentar mitigar los efectos negativos que supone no reconocer como positiva a una conducta que realmente lo es.

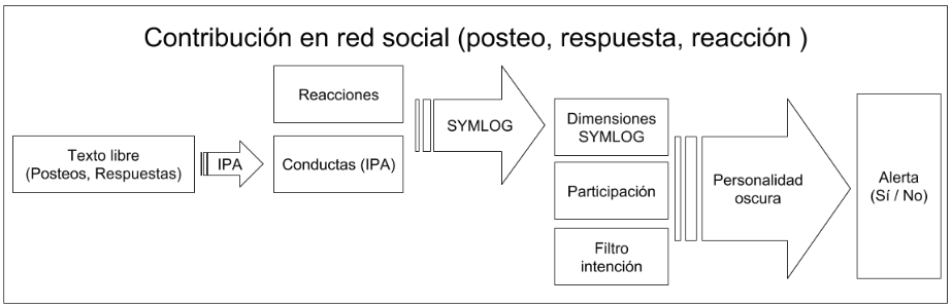

Fig. 1. Pipeline del proceso predictivo.

3 Experimentación

Esta sección se encuentra organizada de la siguiente forma. En la Sección 3.1, se explican los detalles y las características de los datasets con los que se trabajó, junto con el preprocesado de los mismos para luego, en la Sección 3.2, detallar el proceso que se llevó a cabo para efectuar el experimento. Finalmente, en la última subsección, se muestran los resultados obtenidos, y se efectúa un análisis de los resultados y sus implicancias.

3.1 Conjunto de Datos

Para realizar las experimentaciones, se recolectaron datos de tres grupos de Facebook. Se cuenta con tres datasets, donde cada uno se corresponde a grupos de trabajo distintos, a fin de garantizar la heterogeneidad en los comportamientos de los integrantes. Cada integrante, en cada grupo, tiene una determinada cantidad de interacciones, los cuales fueron sometidos al preprocesamiento detallado a continuación.

Cada mensaje de la conversación, es sometido a la herramienta de clasificación, que utiliza filtros de stemming, y elimina stopwords. El resultado es que, para cada mensaje, se obtiene su respectiva clasificación IPA (Requiere, Responde, Positiva, Negativa). A partir de dichas clasificaciones, se arman los indicadores Symlog, conforme a las fórmulas expuestas en la Sección 2.3.

Luego se arman los indicadores para cada instancia: índice de dominancia ponderada, porcentaje de participación de cada usuario, intervalo donde cae dicho porcentaje de participación, y el etiquetado manual de cada instancia, que consiste en dividir a las instancias en intervalos según el índice de dominancia ponderada. Los umbrales para determinar estos últimos intervalos, están parametrizados a fin de adaptar la métrica a la dinámica de cada grupo.

El resultado de dicho proceso es un conjunto de indicadores asociado a cada integrante, es decir, que a un conjunto de indicadores se lo va a denominar comportamiento. Como los datasets están constituidos por ejemplos que entrenan al modelo predictivo, se dice que cada dataset, está compuesto por un conjunto de comportamientos. En otras palabras, cada ejemplo es un individuo, representado a través de un conjunto de indicadores (comportamiento).

El primer grupo analizado, está compuesto por 39 individuos, es decir, que el primer dataset está compuesto por 39 comportamientos. El segundo dataset, está compuesto por 30 comportamientos, mientras que el tercer dataset, un tanto más pequeño, está compuesto por 19 comportamientos

Cabe destacar que, dentro de los comportamientos analizados en un principio, en algunos se pudieron observar niveles de participación marginales, los cuales no fueron tenidos en cuenta debido a que no eran representativos, lo cual se traduce en ruido para el dataset (outliers).

3.2 Proceso Utilizado

Partiendo de la cuestión planteada sobre los efectos colaterales que pueden producir el uso de las redes sociales, más precisamente los excesos en las interacciones, es de gran importancia la necesidad de trabajar en la detección de dichos excesos. El objetivo de esta experimentación es ofrecer un enfoque manifestado en un modelo predictivo, que permita alertar a un hipotético supervisor de un grupo de interés, cuando se detecte

comportamiento considerado compatible con conductas excesivas. De esta manera, la posibilidad de emitir una alerta instantánea permitirá reducir considerablemente el tiempo de respuesta (correctiva o no) por parte de un supervisor, ante este tipo de eventos que lamentablemente se suelen dar en las redes sociales.

La forma de obtención del modelo consistió en la ejecución de una iteración sobre cada dataset, utilizando para ello tres algoritmos a describir en la subsección siguiente, pertenecientes a la librería de aprendizaje de máquina WEKA. Para cada algoritmo, se obtuvo un promedio en la precisión obtenida en cada uno de los tres datasets, el cual se expone en la Tabla 2. En cuanto a la división del dataset para entrenar y testear, se utilizó validación cruzada con diez iteraciones, a fin de garantizar la independencia de las particiones.

3.3 Resultados

A continuación, se proveen los valores de precisión finales (junto con los clasificadores utilizados) de toda la experimentación, la cual fue dividida en etapas. El primer algoritmo utilizado fue un clásico clasificador bayesiano en el que, en promedio para los tres datasets, se obtuvo una precisión de 88.63%. Luego se experimentó con Support Vector Machines con la optimización SMO [17], en donde la precisión promedio bajó a 82.09%, mientras que por último se recurrió al aprendizaje *"lazy"*, utilizando el método KNN en donde si bien la precisión mejoró con respecto a SVM, no se alcanzaron los resultados del clasificador bayesiano, obteniendo una precisión de 85.76%.

Primera Etapa.

La experimentación se dividió en dos etapas, las cuales varían en cuanto a la salida del clasificador. En primer lugar, la idea fue polarizar totalmente las conductas, es decir, que puedan ser clasificadas como positiva o negativa de exceso, en el tipo de conducta que el individuo está llevando a cabo.

A partir de la información disponible, se entrenaron diversos clasificadores, en general con buenos resultados (ver tabla 2), siendo en el clásico clasificador bayesiano en donde se obtuvieron los mejores resultados, para todos los datasets. También se incluyen las correspondientes matrices de confusión, siendo todas construidas para el clasificador Naive Bayes (ver tablas 3, 4 y 5).

Analizando los resultados obtenidos, se llegó a dos conclusiones. La primera radica en la fiabilidad de la medida de precisión, ya que para el caso de Naive Bayes, la elevada precisión obtenida puede llegar a inferir que el clasificador fue sobreentrenado, más aún cuando se tiene en cuenta el tamaño del dataset, que por el momento no cuenta con una cantidad elevada de ejemplos .

Por otro lado, la segunda conclusión a la que se llegó, fue en relación a la semántica de la salida del clasificador, ya que el hecho de que existan solo dos clases para la

detección, puede provocar que existan conductas excesivas, que queden por fuera de la detección automatizada, o bien puede ocurrir que existan conductas que, si bien no son compatibles con los patrones de lo que es un exceso, sí sean dignas de alertar a un hipotético supervisor.

Tabla 2. Precisión de diferentes clasificadores sobre los datasets (Etapa 1).

Valores de Precisión	Dataset 1	Dataset 2	Dataset 3	Promedio
Naive Bayes	94.8718	93.3333	94.7368	94.3139
SVM (SMO)	84.6154	83.3333	89.4737	85.8075
KNN (IBk)	92.3077	86.6667	89.4737	89.4827

Tabla 3. Matriz de confusión para el primer dataset.

Valor Predicho →	Yes	No
Yes	6	1
No	1	31

Tabla 4. Matriz de confusión para el segundo dataset.

Valor Predicho →	Yes	No
Yes	6	0
No	2	22

Tabla 5. Matriz de confusión para el tercer dataset.

Valor Predicho →	Yes	No
Yes	4	0
No	1	14

Teniendo en cuenta las dos conclusiones resultantes de esta primera etapa, se tomó la decisión de agregar una clase intermedia entre el positivo y negativo, a modo de medida de mitigación de los problemas surgidos. De esta manera, la segunda etapa consiste en realizar una nueva iteración sobre los mismos datasets, pero modificando la salida del clasificador, pasando a contar con tres alternativas, en lugar de dos.

Segunda Etapa.

En principio, el algoritmo que arroja los mejores resultados de precisión es el Naive Bayes, por delante de SVM y KNN (ver tabla 6). Esto no quiere decir que sea el algoritmo que prevalezca conforme vaya avanzando la madurez de la experimentación, es decir, puede que las variaciones que vaya sufriendo el dataset, en cuanto a cantidad de información y representación de la misma, provoquen cambios en la performance de cada algoritmo.

Tabla 6. Precisión de diferentes clasificadores sobre los datasets (Etapa 2).

Valores de Precisión	Dataset 1	Dataset 2	Dataset 3	Promedio
Naive Bayes	89.7436	86.6667	89.4737	88.6280
SVM (SMO)	82.0513	80.0000	84.2105	82.0873
KNN (IBk)	89.7436	83.3333	84.2105	85.7625

Como se puede observar, si bien las medidas de precisión con respecto a la primera etapa se vieron disminuidas, el hecho de haber agregado una clase intermedia provoca que la clasificación de cada instancia se viera más distribuida, con casos en donde comportamientos que anteriormente no eran detectados, ahora pasan a ser tenidos en cuenta, no para ser clasificados como una conducta efectivamente excesiva, sino como un tipo de conducta a tener en cuenta, que no debiera ser descartada.

A continuación, se presenta, en las Tablas 7, 8 y 9, las matrices de confusión asociadas a cada dataset para el clasificador bayesiano, que es con el cual se obtuvieron los mejores resultados en la segunda etapa. Como se podrá observar en dichas matrices, si bien las conductas que decididamente son excesivas no son clasificadas como tal igualmente caen dentro de la categoría Mid, o sea que igualmente provocarán que se emita una alerta para el supervisor del grupo de interés. De todas maneras, este desempeño irá mejorando a medida que se vayan llevando a cabo las iteraciones sobre el dataset, a modo de ir afinando los mecanismos de predicción, ya sea a través de efectuar un mayor preprocesamiento del dataset, o bien a través de buscar los hiperparámetros que mejor se adapten para cada situación.

Tabla 7. Matriz de confusión para el primer dataset.

Valor Predicho →	Yes	Mid	No
Yes	0	1	0
Mid	0	6	0
No	0	3	29

Tabla 8. Matriz de confusión para el segundo dataset.

Valor Predicho →	Yes	Mid	No
Yes	0	1	0
Mid	0	5	0
No	0	3	21

Tabla 9. Matriz de confusión para el tercer dataset.

Valor Predicho →	Yes	Mid	No
Yes	0	1	0
Mid	0	3	0
No	0	1	14

4 Conclusiones

En este trabajo se presentaron resultados preliminares de clasificación de conductas de integrantes de diferentes grupos de interés, para así detectar individuos con personalidad peligrosa, que se estén desenvolviendo en dichos grupos de manera excesiva. Para el dominio de la detección automatizada de dichos excesos, está abierta la posibilidad de elaborar una mayor cantidad de indicadores que también puedan nutrir al modelo predictivo, a la vez de que también se busca ampliar el tamaño del dataset utilizado.

A su vez, también hay otras líneas de mejora en el proceso. Al obtener mayores volúmenes de datos con los cuales trabajar, el modelo predictivo puede ser perfeccionado dejando de lado las técnicas tradicionales de aprendizaje automático, para reemplazarlas por métodos de aprendizaje profundo, en donde, en caso de incluir en el dataset el contenido que cada usuario postea en algún grupo de interés (por ejemplo, el texto), se daría un contexto interesante para experimentar con redes neuronales recurrentes, utilizando para ello unidades básicas como las LSTM o GRU para construir dichas redes.

Partiendo de la base de que dentro de la literatura actual no hay mucho trabajo relacionado a la detección de conductas excesivas, creemos que este trabajo efectúa una contribución interesante al área de análisis de interacciones, sobre todo en el idioma español. Como trabajo futuro, se buscará extender el dominio de acción a cualquier tipo de interacción entre individuos. En este caso se aplicó en grupos de interés, pero podría aplicarse a cualquier tipo de conversación, en donde dicho dominio podría ir mucho más allá de las redes sociales, como por ejemplo las conversaciones entre individuos. Esto es, no sólo a través de mensajes escritos sino a través de comunicación verbal, en donde a través de la digitalización del audio y del modelado de secuencias, se podrían elaborar predicciones que indiquen excesos en la comunicación, entre muchas otras cosas.

La idea es que conforme se vayan realizando los incrementos sobre el trabajo original, el dataset se irá enriqueciendo y diversificando al contener datos de distintos contextos. Como las características de cada contexto propician una aplicación específica de técnicas de aprendizaje automático, dicha heterogeneidad deja abierta la posibilidad de desarrollar un framework integral donde se especialice y facilite el trabajo para la detección automatizada de todo tipo de excesos en las conductas, sirviéndose para ello de la base de conocimiento generada a lo largo de las iteraciones.

5 Referencias

1. Fox, J., & Moreland, J. J. (2015). The dark side of social networking sites: An exploration of the relational and psychological stressors associated with Facebook use and affordances. Computers in Human Behavior, 45, 168-176.

2. Mäntymäki, M., & Islam, A. N. (2016). The Janus face of Facebook: Positive and negative sides of social networking site use. Computers in Human Behavior, 61, 14-26.
3. Krasnova, H., Widjaja, T., Buxmann, P., Wenninger, H., & Benbasat, I. (2015). Research note—why following friends can hurt you: an exploratory investigation of the effects of envy on social networking sites among college-age users. Information systems research, 26(3), 585-605.
4. Yang, S., Liu, Y., and Wei, J. 2016, "Social capital on mobile SNS addiction: A perspective from online and offline channel integration," Internet Research (26, 4) pp. 982-1000.
5. Garcia, D., & Sikström, S. (2014). The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. Personality and Individual Differences, 67, 92-96.
6. Lai, C., Altavilla, D., Ronconi, A., & Aceto, P. (2016). Fear of missing out (FOMO) is associated with activation of the right middle temporal gyrus during inclusion social cue. Computers in Human Behavior, 61, 516-521.
7. Choi, A. (2013). Use of Facebook group feature to promote student collaboration. In American Society for Engineering Education. ASEE Southeast Section Conference.
8. Bales, R. F., Cohen, S. P., & Williamson, S. A. (1979). SYMLOG: A system for the multiple level observation of groups. Free Pr.
9. Bales, R. (2017). Social interaction systems: Theory and measurement. Routledge.
10. Berdun, F. D., & Armentano, M. G. (2018). Modeling Users Collaborative Behavior with a Serious Game. IEEE Transactions on Games.
11. Kaplan, A. M., and M. Haenlein. 2010, "Users of the World, Unite! the Challenges and Opportunities of Social Media," Business Horizons (53:1), 2, pp. 59-68.
12. Miah, M. W. R., Yearwood, J., & Kulkarni, S. (2011). Detection of child exploiting chats from a mixed chat dataset as a text classification task. In Proceedings of the Australasian Language Technology Association Workshop 2011 (pp. 157-165).
13. Michalopoulos, D., & Mavridis, I. (2010, July). Towards risk-based prevention of grooming attacks. In Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on (pp. 1-4). IEEE.
14. Rybnicek, M., Poisel, R., & Tjoa, S. (2013, October). Facebook watchdog: a research agenda for detecting online grooming and bullying activities. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on (pp. 2854-2859). IEEE.
15. Bales, R. F. Interaction process analysis; a method for the study of small groups (1950).
16. Berdun, F. D., Armentano, M. G., Berdun, L., & Mineo, M. (2017). Classification of collaborative behavior from free text interactions. Computers & Electrical Engineering.
17. Platt, John. "Sequential minimal optimization: A fast algorithm for training support vector machines." (1998).