**2.Fundamentos de los rastros de bullying**

Para descubrir información útil sobre el acoso en las redes sociales, el primer paso es identificar el rastro del acoso entre las cantidades masivas de publicaciones en las redes sociales publicadas todos los días. Formulamos el problema como una tarea de clasificación de texto binario. Con características cuidadosamente seleccionadas y algoritmos de clasificación, somos capaces de reconocer rastros de acoso con una precisión satisfactoria. Para analizar automáticamente los rastros de intimidación, construimos clasificadores de texto para reconocer el papel de un autor, las formas y los tipos de rastros de intimidación. Incluyendo: ¿quién está publicando rastros de acoso? ¿Qué formas de acoso se mencionan o usan en Twitter? ¿Por qué la gente publica acerca del bullying en Twitter? ¿Qué temas publican las personas sobre el acoso escolar?

**2.1 Reconocimiento de los rastros del bullying.**

Dado que los rastros del acoso representan sólo una pequeña fracción de todos los tweets, es un desafío significativo encontrar suficientes rastros de intimidación sin etiquetar un enorme número de tweets. Por esta razón, nos limitamos a un "conjunto de datos enriquecido", que se obtiene mediante la recopilación de tweets a través de la API de streaming público de Twitter utilizando palabras clave. Para identificar una publicación, deben ser identificadas las palabras completas en Twitter en lugar de porciones de palabras. (Por ejemplo, después de la palabra "bull" no se identificará El término "bully")Hemos incluido las siguientes palabras clave relacionadas con el término acosador:acosado, acosar, akosado, acoso, akosador, akoso. Se incluyeron varias palabras clave mal escritas, ya que aparecen con frecuencia en las redes sociales. También incluimos otras palabras que fueron identificadas por nuestros anotadores como términos comunes en un análisis de contenido de las descripciones escritas de las experiencias de acoso de los estudiantes de secundaria. Estas palabras clave fueron: ignorado, empujado, rumores , empujón, rumor, burlarse, patear, llorar. De todos los tweets obtenidos que contienen al menos una de estas palabras clave, a continuación, filtramos los tweets de modo que sólo pusimos los que contenían una palabra que comienza con "acos-" se mantuvieron. Por ejemplo,la publicación, Acosadores empujaron al niño "sería detectado y retenido por otro lado, la publicación, "Empujó al niño" Serían recogida pero no retenida. En lugar de pre-determinar cada posible variación de la palabra acosar usada en Twitter, este proceso se utilizó para maximizar las variaciones que podrían ser utilizadas. De este conjunto de datos, eliminamos los re-tweets (la analogía de los correos electrónicos reenviados) mediante la exclusión de los tweets que contienen el acrónimo "RT".El proceso de enriquecimiento está destinado a retener muchos tipos de acoso de primera mano. También es importante tener en cuenta que este simple filtro de palabras clave y está lejos de ser perfecto: muchos tweets irrelevantes sobrevivieron y tweets relevantes se perdieron. Un rastro de acoso se definió como cualquier mención de acoso dentro del contexto de un episodio de acoso específico en el que participó el autor. Tenga en cuenta que no evaluamos la publicación para el cumplimiento de las definiciones de acoso que incluyeron nociones de poder desequilibrio y repetición (Olweus, 1993) porque esta información a menudo no es evidente en publicaciones cortas, además, no pudimos determinar si el episodio se refería a un caso aislado o un episodio continuo durante un período de tiempo. Nos apoyamos enteramente en el texto de cada publicación individual, tomándolo a su valor nominal cuando un autor participó en el o reportó "bully" ing. Algunos ejemplos marcados incluyen:

* Experiencias personales.
* Informes sobre episodios específicos.
* Publicaciones de interés periodístico.
* Mensajes claramente copiado y pegado un titular de noticias acerca de un episodio de acoso.
* Publicaciones referidas a un episodio de intimidación que puede ocurrir en el futuro.
* Las publicaciones que reflejaban sólo una opinión sobre el acoso en general.
* Publicaciones donde un comportamiento puede sonar como acoso, pero no es identificado como tal por el autor.
* Publicaciones donde un codificador reconoce los nombres mencionados en el post como ficticios.

**Datos**

Para identificar los verdaderos rastros de acoso del enriquecido conjunto de datos, lo formulamos como una tarea de categorización de texto binario. Los anotadores identificaron 1.762 tweets muestreados uniformemente del enriquecido conjunto de datos el 6 de agosto de 2011. Entre ellos, 684 (39%) fueron identificados como rastros de acoso. Este conjunto de datos y su documentación se archivan como conjunto de datos de rastreo de acoso.

**Métodos**

Después de Settles (2011), estos 1,762 tweets se doblaron los casos, pero sin ninguna eliminación de stemming o stopword. Las menciones de usuario precedidas de un "@" fueron reemplazadas por el nombre de usuario anónimo "@USERNAME". Cualquier URL que comience con "http" fue reemplazada por el token "HTTPLINK". Hashtags (palabras compuestas que siguen a "#") no se dividieron y se trataron como un solo token. Emoticones, como ":)" o ": D", también se incluyeron como fichas. Después de estos procedimientos de preprocesamiento, hemos creado tres diferentes conjuntos de representaciones de características: unigrams (1g), unigrams + bigrams (1g2g), y POS-color unigrams + bigrams (1g2gPOS). POS tagging se hizo con el Stanford CoreNLP paquete (Toutanova et al., 2003). El POS-coloring se realizó expandiendo cada token en token: POS.

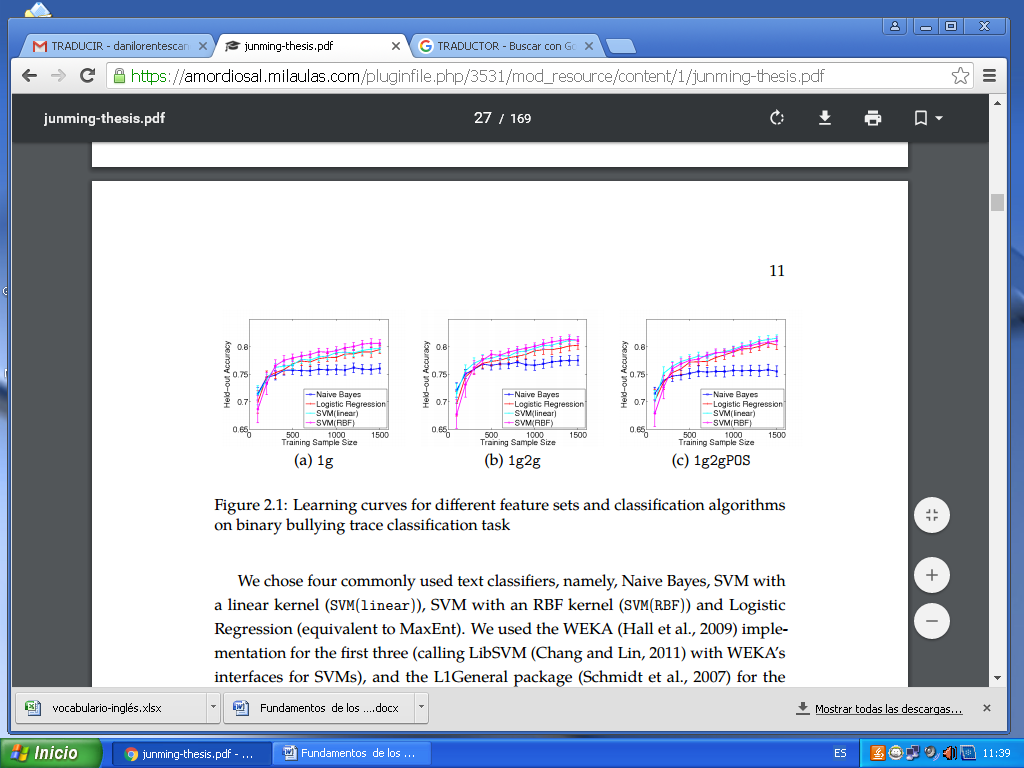


Figura 2.1: Curvas de aprendizaje para diferentes conjuntos de características y algoritmos de clasificación en la tarea de clasificación de rastreo de bullying binario.

Se eligieron cuatro clasificadores de texto de uso común, llamados, Naive Bayes, SVM con un nucleo lineal (SVM (lineal)), SVM con un núcleo RBF (SVM (RBF)) y Regresión Logística (equivalente a MaxEnt). Utilizamos la implementación de WEKA para los tres primeros (llamando a LibSVM con las interfaces de WEKA para SVMs) y el paquete L1General para el cuarto. Hemos mantenido 262 tweets para las pruebas, y sistemáticamente varió el tamaño del conjunto de entrenamiento entre los tweets restantes, de 100 a 1.500 con tamaño de paso 100. Ajustamos todos los parámetros conjuntamente por 5 veces la validación cruzada en el conjunto de entrenamiento con la cuadrícula {2 -8, 2-6,. . . , 28}. Los cuatro clasificadores de texto fueron entrenados en los conjuntos de entrenamiento y probados en el conjunto de pruebas. Todo el procedimiento se repitió 30 veces para cada representación característica.

**Resultados**

La Figura 2.1 informa la precisión del conjunto retenido a medida que aumenta el tamaño del conjunto de entrenamiento. Las barras de error son ± 1 error de desviación estándar. Con el mayor tamaño de conjunto de entrenamiento (1.500), la combinación de SVM (lineal) + 1g logró una precisión promedio de 79.7%. SVM (lineal) + 1g2g alcanzó el 81,3%, lo cual es significativamente mejor (t-test, p = 4 × 10-6).Muestra que incluir bigrams puede mejorar significativamente el rendimiento de la clasificación.   
Muestra que incluir bigrams puede mejorar significativamente el rendimiento de la clasificación. La SVM (lineal) + 1g2gPOS alcanzó el 81,6%, aunque la mejora no es estadísticamente significativa (p = 0,088), lo que indica que la coloración POS no ayuda mucho en esta tarea. SVM (RBF) da un rendimiento similar, Regresión Logística es ligeramente peor, y Naive Bayes es mucho peor, para una gran variedad de tamaños de conjunto de entrenamiento. En resumen, SVM (lineal) + 1g2g es el modelo preferido debido a su precisión y simplicidad. Lo ponemos a disposición pública en nuestro software repositorio.

|  |  |  |
| --- | --- | --- |
|  | Total | Predicted as  Yes No |
| Yes | 2102 | 1555 547 |
| No | 5219 | 503 4716 |

Tabla 2.1: Matriz de confusión de la clasificación de trazas de rastros de acoso.

También observamos que estas precisiones son mucho mejores que la línea de base de la clase mayoritaria del 61%. En el conjunto retenido, SVM (lineal) + 1g2g alcanzó la precisión P = 0,76, recuerda R = 0,79 y F-medida 0,77.

**Discusión**

En cuanto a por qué la mejor exactitud no es cercana a 1, una hipótesis es etiquetas ruidosas causadas por desacuerdo intrínseco entre los anotadores. Los tweets son cortos y algunos son ambiguos. Sin el conocimiento previo sobre los usuarios y sus otros tweets, los anotadores interpretaron los tweets en sus propias maneras. Por ejemplo, para el tweet muy corto "se siente como un matón ....." nuestros anotadores no estaban de acuerdo sobre si se trata de un rastro de acoso. Los etiquetadores pueden tener diferentes puntos de vista sobre estos tweets ambiguos, que crearon etiquetas ruidosas de rastreo de acoso.

**Numero de rastros de acoso identificados entre 2011-2013**

La Figura 2.1 muestra que las curvas de aprendizaje siguen aumentando, lo que sugiere que se puede obtener una mayor exactitud si anotamos más datos de entrenamiento. Para lograr un mejor rendimiento, nuestros anotadores identificaron 7,321 tweets (incluyendo los 1,762 tweets en el experimento anterior), seleccionados al azar entre el 6 de agosto de 2011 al 31 de agosto de 2011

|  |  |  |  |
| --- | --- | --- | --- |
|  | 9/1/11 - 8/31/12 Count % | 9/1/12 - 8/31/13 Count % | total Count % |
| Keyword-filtered tweets Bullying traces | 12,421,237 3,955,458 31.8% | 20,056,321 5,809,125 29.0% | 32,477,558 9,764,583 30.1% |

El acuerdo entre evaluadores para identificar los rastros de acoso de los tweets de palabras clave de intimidación se calculó sobre la base de dos anotadores que codifican 1.000 de los 7.332 mensajes. Se determinó que κ = 0,83. De los 7,321 puestos, 2,102, o el 28,7%, fueron etiquetados como rastros de acoso. Este conjunto de datos y su documentación se archivan como conjunto de datos de traza Bullying. Con este conjunto de entrenamiento más grande, la precisión de nuestro clasificador de texto SVM (lineal) + 1g2g mejoró a 86%. Este nivel de precisión es similar al nivel de acuerdo alcanzado por dos anotadores humanos diferentes. Por lo tanto, hemos determinado que es posible utilizar el aprendizaje automático para reconocer automáticamente y con precisión los rastros de intimidación en las redes sociales. Aplicamos el clasificador entrenado al conjunto de datos enriquecido recolectado desde el 1 de septiembre de 2011 hasta el 31 de agosto de 2013. El resultado se reporta en la Tabla 2.2. Hemos recogido 32.477.558 tweets en nuestro conjunto de datos enriquecido a través de filtrado de palabras clave durante dos años académicos consecutivos. Entre ellos, encontramos que el 30,1% (9,764,583) fueron reconocidos como rastros de acoso por nuestro clasificador de texto. Las proporciones de las rastros intimidantes y las huellas no-intimidantes identificadas fueron similares a los datos anotados en humanos en la Tabla 2.1. La principal diferencia entre los dos años es que el número total de rastros de acoso aumentó del año 1 al año 2, una tendencia que probablemente refleje la creciente popularidad de Twitter entre los dos años. Hubo 3,955,458 rastros de acoso en 2011-2012 y 5,809,125 en 2012-2013.