Clasificación de memes misóginos

Abraham Cisneros Valladolid
Facultad de Ciencias
Universidad Autónoma de Baja California (UABC)
Ensenada, Baja California
Email: abraham.cisneros@uabc.edu.mx

***Resumen*—En este artículo científico se plantea el problema de la misoginia mediante memes en las redes sociales, proponiendo así un método usando herramientas de minera de datos para clasificar memes como misóginos o no misóginos.**

I. Introducción

En la sociedad moderna donde el internet está en todos lados, las grandes cantidades de información en nuestras manos y las redes sociales en la boca de todos, no es extraño ver este fenómeno llamado meme en todo el ciberespacio. Estos pueden ser imágenes, vídeos, audios y hasta texto con intenciones de dar risa e ironizar situaciones, todo con fin humorístico, así como podemos observar en la Figura 1. Este fenómeno ha ido creciendo con internet a la par de los años, claro, no eran como los conocemos ahora, pues los memes son un reflejo de la cultura e ideas de la sociedad en el momento, por lo que ver los memes creados por los años se puede observar ciertas tendencias, creencias, fenómenos del momento y acontecimientos históricos siendo satirizados, ironizados y puestos en contextos graciosos.

Figura 1. Una imagen con índole gracioso.

La situación con los memes es que se pueden usar para burlarse de cualquier cosa, por lo que hay comunidades enteras dedicadas a hacer este contenido sobre ciertos temas en específico. Y como se mencionó anteriormente, los memes son un reflejo de la sociedad, por lo que si se vive en una sociedad misógina, es de esperar como resultado memes misóginos. Recordando que la misoginia es definida por Johnson (2000) como:

> [...] es la parte central de los prejuicios e ideologías sexistas y, como tal, es una de las bases para la opresión de las mujeres en las sociedades dominadas por hombres. La misoginia se manifiesta de diferentes maneras, desde bromas a pornografía, violencia y el sentimiento de odio hacia su propio cuerpo al que las mujeres son instruidas a sentir.

Estos memes que ya no entraran en la definición de sátira, si no más bien estarían clasificados como machistas, violentos y sexistas, por su contenido y mensaje. Todo con el afán de hacer menos a la mujer o incitar al odio hacia ellas.

Desafortunadamente el abuso de la misoginia en la WWW (*world wide web*) es un serio problema [1], aunado que por el hecho de ser memes y estar incrustados en las redes, este tipo de contenido se masifica rápidamente [2] llegando a los ojos de todos, agrediendo con sus mensajes de odio a niñas, jóvenes y adultos que hacen uso de las redes sociales e internet.

La utilidad de una herramienta que sea capaz de determinar si una imagen con texto “graciosos“ tiene contenido de odio hacia la mujer y así facilitar el bloqueo de la publicación o simplemente que no se pueda publicar, seria de gran ayuda en las redes sociales para la sociedad.

Por los grandes beneficios que se podrían aprovechar, se piensa desarrollar un método aplicando herramientas de la minería de datos, para la detección de contenido misógino.

II. Trabajo relacionado

Debido a la cantidad de memes en la WWW y el constante crecimiento del uso de las redes sociales, se tienen grandes cantidades de información que puede ser estudiada de diversas formas, pues se puede analizar el texto, el contenido de la imagen o una combinación de ambas tal como lo hacen en [3] y [4], en donde analizan una gran base de datos de más de 11,000 memes con la herramienta BERT para usar varios métodos de observación y darle una etiqueta a cada meme sobre su contenido. Sin embargo, no solo hay datos, también existen múltiples problemáticas que se desean detectar en los memes, así como: ciberbullying [5], expresiones de odio y racismo [6], misoginia [7] y [8], entre otros más. Los resultados que presentan son bastante optimistas, con una precisión que ronda el 70 hasta el 85 por ciento al clasificar los memes, son buenos resultados, pero que aún requieren mejoras y más estudio. Muchos de estos trabajos son relativamente nuevos, ya que es un tema de estudio que está en boga gracias a los métodos desarrollados para la extracción de texto en imágenes de manera automática, a las múltiples herramientas que se tienen para clasificar esta información y

al planteamiento de las problemáticas sociales que se reflejan en la red y como afecta a los individuos [9].

III. Datos

Se planea usar al menos 100 memes en imágenes, y al menos 50 deben ser misóginos, las imágenes que se utilizaran son parte de una base de datos dada por una plataforma de desafíos orientados a programación llamada Codalab. Los memes que se tienen en la base de datos son muy variados, no solo en contenido, también en tipo de humor, dimensiones de píxeles, formato del meme, tipografía utilizada, colores y hasta contenido.

Por lo que es requerido emplear memes que tengan al menos texto legible y que describan parte de la gracia que tiene la imagen, pues existen memes sin texto o con formatos que no es posible de caracterizar.

IV. Metodología

- Bajo los requerimientos previamente mencionados se extraerá el texto de las imágenes con transcripciones manuales para finalmente generar cadenas de texto.
- Con el conjunto de cadenas de texto se realizará una bolsa de palabras *Bag of words*, este método genera una tabla en la cual se puede observar todo el vocabulario del conjunto y la frecuencia que tiene cada palabra en cada una de las cadenas.
- Dada esta información es posible aplicar el método TF-IDF *Term Frequency–Inverse Document Frequency* el cual retorna un valor estadístico que refleja la importancia de las palabras dada una colección.
- Finalmente se usará estos datos generados para compararlos mediante el uso de métodos de clasificación, tales como SVM *Support Vector Machine Algorithm*, *Naive Bayes* y *Random forest*, dichos métodos mencionados son bastante populares en el área de aprendizaje automático para la clasificación de información y reconocimiento de patrones, por lo que usar estas herramientas nos puede ayudar a lograr el objetivo esperado.

V. Experimentación y resultados

Siguiendo el esquema previamente planteado, se tiene la visualización de la bolsa de palabras separando las categorías de misógino y no misógino en nube palabras, pues este gráfico nos facilita la detección de aquellas palabras que tienen una mayor frecuencia, palabras que comparten estas categorías y palabras que son únicas en cada una de ellas. La figura 2 muestra una nube de palabras con las 100 palabras representativas de la transcripción de memes misóginos.

Figura 2. Las 100 palabras más repetidas representadas de memes misóginos en una nube de palabras. *Advertencia: Esta imagen contiene lenguaje que puede ser ofensivo*

Es importante conocer el contenido de la categoría de memes no misóginos, pues la comparación de la información contenida en cada bolsa de palabras es fundamental para lograr un muestreo completo de todo el lenguaje usado y así determinar el peso de cada palabra en todo el conjunto de cadenas de texto. La figura 3 representa la bolsa de palabras de la categoría de memes no misóginos en una nube de palabras.

Figura 3. Las 100 palabras más repetidas de memes no misóginos representada en una nube de palabras.

Al observar la figura 2 y 3, se puede detectar ciertos comportamientos y características de ambas categorías en la base de conocimientos. Lo primero que se aprecia es que su conjunto de palabras es bastante distinto, a pesar de que son más de 530 palabras solo se comparten unas cuantas palabras entre ambos tipos de memes, como: *kitchen, woman, dad, red, girls, clean....* Además, en la figura 2 se aprecian un lenguaje que tiene cierto carácter ofensivo y despectivo como lo son: *hooker, bitch, dick, rape, sex, horny*, al mismo tiempo las palabras tienen una relación, por ejemplo: *woman, girl, feminism, feminist, milf* son palabras que se refieren a las mujeres o movimientos femeninos, finalmente palabras relacionadas con estereotipos relacionados con las mujeres en la cocina: *kitchen, sandwich, beer, coke, make*. Como se puede observar estas palabras tienen cierto contexto que las une pues son usadas para hacer contenido misógino. En cambio, las palabras que aparecen en la clase de memes no misóginos

no tienen relación alguna, incluso el conjunto de palabras es mucho más amplio, pero la frecuencia es más baja.

En la tabla I se puede hacer un análisis mucho más detallado sobre la frecuencia de las palabras entre los dos tipos de memes, para lograr esto fue requerido un filtrado de palabras vacías, mejor conocidas como *stopwords*, lo cual nos permite limpiar conectores para dejarnos solo con adjetivos, sustantivos y sujetos.

Tabla I
LAS 10 PALABRAS MÁS REPETIDAS EN AMBAS CATEGORÍAS DE MEMES

Misoginos		No misoginos	
token	frecuencia	token	frecuencia
like	9	cheat	5
get	6	get	5
want	6	clean	5
women	6	house	4
make	5	kitchen	4
feminist	4	snow	4
girl	4	ban	3
hooker	4	big	3
kitchen	4	water	3
sex	4	jarvi	3

La diferencia entre ambas clases es bastante notoria, por lo que formalizar y establecer un valor para cada palabra es fundamental para llegar al objetivo, con el conocimiento sobre la frecuencia de las palabras por documento se realizó el cálculo del TF-IDF, el cual es seguido por la siguiente fórmula en la figura 4:

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Figura 4. Formula para aplicar TF-IDF

Como se mencionó previamente, TF-IDF requiere de N documentos para tomar las frecuencias de las palabras en estos, para lograr esto se suman las veces que aparece una palabra en ambos documentos y es divida por el total de palabras, posteriormente este valor es usado para calcular el logaritmo dado el número total de documentos y cuantos documentos poseen esta palabra, para así asignar un valor de peso. Se espera que las palabras representativas del documento con memes misóginos tengan un peso mayor debido a que no aparecen en documento comparado y las palabras que comparten ambos documentos deberían tener un valor menor siguiendo la lógica del método. (Teniendo en consideración que cada meme individualmente es un documento).

Como se puede observar en la tabla II, el valor que se le da a las palabras de la categoría de memes misóginos son aquellas que tenían una frecuencia alta, tenían cierta relación de tema y no aparecían en ambos documentos a excepción de *woman* y *men*, ya que estas palabras aparece muchas más veces en esta categoría a comparación de la otra. Por el contrario las palabras en memes no misóginos que tienen mayor peso son muy distinguibles que a las de los memes misóginos, tal como era esperado.

Tabla II
TABLA COMPARANDO LOS VALORES TF-IDF PARA AMBOS DOCUMENTOS

Misoginos		No misoginos	
token	tf-idf	token	tf-idf
women	4	house	4
like	4	snow	4
hooker	4	cheat	4
sex	3	water	3
men	3	jarvis	3
feminist	3	white	3
would	3	ha	3
make	2	keep	3
feminism	2	don't	2
sandwich	2	kitchen	2

Continuando con la metodología descrita, se usarán los datos de la bolsa de palabras y el TF-IDF para ingresarlos a 3 métodos de clasificación, para ello se dispondrá de 80 memes para el conjunto de entrenamiento, donde 40 son de la clase misóginos y 40 son de la clase no misóginos. Por lo que los 20 memes restantes, 10 misóginos y 10 no misóginos, serán usados como conjunto de prueba. Los datos de prueba fueron seleccionados aleatoriamente para evitar problemas de sesgo.

La tabla III muestra los resultados que se obtuvieron mediante la bolsa de palabras:

Tabla III
TABLA COMPARANDO LOS VALORES TF-IDF PARA AMBOS DOCUMENTOS

Bolsa de palabras			
Método usado	Porcentaje de precisión	Correctos	Incorrectos
Random forest	95 %	19	1
Naive Bayes	80 %	16	4
SVM	85 %	17	3

La tabla IV muestra los resultados que se obtuvieron mediante TF-IDF:

Tabla IV
TABLA COMPARANDO LOS VALORES TF-IDF PARA AMBOS DOCUMENTOS

TF-IDF			
Método usado	Porcentaje de precisión	Correctos	Incorrectos
Random forest	95 %	19	1
Naive Bayes	80 %	16	4
SVM	90	18	2

VI. TRABAJO FUTURO Y CONCLUSIÓN

Los resultados obtenidos con las dos enteradas en los distintos métodos son bastante prometedores, pues todos consiguieron un porcentaje de precisión mayor o igual al 80 %, por lo que es posible generar resultados bastante útiles al momento

de implementarlos. Haciendo un análisis sobre el método con mejores resultados: *Random forest*, vemos que se falla solo con 1 entrada de las 20, verificando estos datos se pudo extraer el meme el cual no pudo clasificar correctamente, el cual sería la figura 5.

Figura 5. Formula para aplicar TF-IDF

Este meme esta clasificado como misógino, sin embargo, el método lo etiqueto como no misógino, esto puede ser debido a que posee la palabra *kitchen*, es una de las palabras que aparece en ambos tipos de memes, al mismo tiempo poseen la misma frecuencia y en TD-IDF posee un valor mayor en la clase de memes no misóginos. Por lo que es de esperar que lo clasifique como uno no misógino aunque no lo sea. Esto resulta bastante interesante de estudiar, pues teniendo en cuenta que solo se usaron 100 memes en total, se lograron resultados muy asertivos, sin embargo, teniendo en cuenta el error que se tuvo debido a la palabra que se comparte en ambos tipos de memes, es de esperar que conforme crezca la cantidad de memes crecerá el vocabulario y por ende las palabras que comparten ambas categorías, creando la posibilidad de muchos más casos como el que ocurrió. Si se desea tener un mayor impacto en un problema real, es indispensable hacer más experimentos en un futuro con una base de conocimientos mayor y dado un análisis de esto, proponer nueva soluciones para obtener resultados igual de útiles.

VII. Referencias

[1] Citron, D. K. (2014). Hate crimes in cyberspace-introduction. Hate Crimes in Cyberspace, Harvard University Press (2014), U of Maryland Legal Studies Research Paper, (2015-11).

[2] Wiggins, B. E., & Bowers, G. B. (2015). Memes as genre: A structurational analysis of the memescape. New media & society, 17(11), 1886-1906.

[3] Rao, A. R., & Rao, A. (2022, July). ASRtrans at semeval-2022 task 5: Transformer-based models for meme classification. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 597-604).

[4] Gu, Q., Meisinger, N., & Dick, A. K. (2022, July). Qinian at semeval-2022 task 5: Multi-modal misogyny detection and classification. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 736-741).

[5] Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020, May). Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In Proceedings of the second workshop on trolling, aggression and cyberbullying (pp. 32-41).

[6] Leyva Massagué, J. (2022). Hybrid models for hateful memes classification (Master's thesis, Universitat Politècnica de Catalunya).

[7] Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P., ... & Sorensen, J. (2022, July). SemEval-2022 Task 5: Multimedia automatic misogyny identification. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 533-549).

[8] Cordon, P., Diaz, P. G., Mata, J., & Pachón, V. (2022, July). I2c at semeval-2022 task 5: Identification of misogyny in internet memes. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 689-694).

[9] Bhattacharya, P. (2019). Social degeneration through social media: A study of the adverse impact of 'memes'. 2019 Sixth HCT Information Technology Trends (ITT), 44-46.