Beyond Words: A Preliminary Study for Multimodal Hate Speech Detection

Sofía Barceló and Magalí Boulanger
Facultad de Ciencias Exactas, UNICEN, Tandil, Argentina
{sbarcelo, mboulanger}@alumnos.exa.unicen.edu.ar

Antonela Tommasel and Juan Manuel Rodriguez
ISISTAN, CONICET-UNICEN, Tandil, Argentina
{antonela.tommasel, juanmanuel.rodriguez}@isistan.unicen.edu.ar

***Abstract*—Hate speech involves harmful expressions that either directly attack or endorse hatred toward a group or individual, often based on identity characteristics such as ethnicity, religion, or sexual orientation. The lack of regulation regarding hate speech raises concerns about the vulnerability of targeted groups and the potential escalation of violence. Detecting and eliminating hate speech from social media is crucial to prevent harm, negative psychological effects, and further attacks. The complexity of identifying hate speech lies in the indirect or implicit nature of its expressions, which can be influenced by context and the use of sarcasm or irony. Detection becomes even more challenging in multimodal scenarios, such as memes, where text and images are combined to convey meaning. To address this problem, this study analyzes the design of a multimodal hate speech detection technique that integrates textual and visual dimensions. This study explores how to define the textual and visual representations of *memes* and explores fusion strategies. The preliminary experiments conducted over publicly available datasets achieved encouraging results when compared to more complex approaches in the literature, while also highlighting the challenges of the task.**

***Index Terms*—Hate speech, Meme, Multi-modality, Text classification, Image classification**

I. Introducción

La libertad de expresión es un derecho fundamental que permite expresar ideas sin perjuicio alguno. Con el uso de los medios sociales, el alcance de dichas ideas creció exponencialmente y, así como esta hiper-comunicación puede ser beneficiosa, también puede ser peligrosa. En los últimos años, las medios sociales debieron hacer frente al fenómeno de *hate speech*[1], el cual puede ser definido como cualquier comunicación que desprestigie a una persona o grupo basándose en alguna de sus características, pudiendo ocurrir en formas sutiles o por medio del humor [1]. No solo los medios sociales y el ámbito académico se han pronunciado respecto a este fenómeno, sino también instituciones gubernamentales, por ejemplo Reino Unido[2] propuso planes de acción para tratar de garantizar un acceso seguro a los medios sociales, y en Argentina[3] se definieron recomendaciones para su control.

El *hate speech* no solo afecta a la persona o grupos sobre las cuales se ejerce, sino que aumenta el riesgo de otros ataques más graves. En este sentido, se han encontrado relaciones entre el *hate speech* y la disminución de rendimiento escolar, abandono y comportamientos violentos, en combinación con efectos psicológicos como depresión, baja autoestima e incluso suicidios [2]. Asimismo, el *hate speech* también se relaciona con otros fenómenos, como la desinformación, los cuales se potencian contribuyendo uno a la propagación del otro [3].

Diversos trabajos en la literatura han abordado la detección de *hate speech* en texto. Sin embargo, esta tarea no resulta sencilla dado que las palabras o expresiones no siempre se manifiestan de forma directa o explícita, sino que en muchas ocasiones, "adquieren" su agresividad de acuerdo a su contexto o mediante sarcasmo/ironía [4]. Por otra parte, recurrentes en los medios sociales, los *memes*[4] dificultan el análisis, dado que no basta con analizar la imagen y/o el texto por separado, sino que hay que analizarlos como un conjunto. Sin embargo, no todos los estudios coinciden respecto a los beneficios del análisis multimodal [5], lo que plantea la necesidad de continuar estudiando la problemática y sus posibles soluciones.

El objetivo de este trabajo es presentar una evaluación experimental preliminar de una técnica de detección de *hate speech* multimodal que considera el texto y el contexto visual que lo acompaña. Diferente a la mayoría de los trabajos en la literatura que se enfocan en tareas directas de clasificación, esta propuesta se enfoca desde la perspectiva de *recuperación de información*, en la que la selección del clasificador de *memes* depende de una etapa previa de análisis de sus características. Para dar soporte a esta evaluación se consideraron dos conjuntos de datos disponibles en la literatura. Los resultados mostraron las dificultades de la tarea, en especial, relacionados con los datos y cómo afectan a la flexibilidad y generalizabilidad de las técnica analizada.

II. Trabajos relacionados

Previo a las técnicas multimodales, la detección ha sido principalmente unimodal, enfocándose en el texto, siendo menos frecuente el análisis de lo visual. Por ejemplo, Gaydhani et al. [6] usaron *N-grams*, TF-IDF y clasificadores tradicionales como *Logistic Regression* y *Support Vector Machines*. Rajput et al. [7] y Roy et al. [8] cuantificaron el impacto de embeddings (*FastText*, *GloVe* y *BERT*) y redes neuronales. Niam et al. [9] crearon representaciones basadas en *Latent Semantic Analysis* para clasificarlas de acuerdo a su distancia.

S. Barceló y M. Boulanger contribuyeron de forma equitativa a este trabajo

[1]Por ejemplo, *Facebook*: https://transparency.fb.com/es-la/policies/community-standards/hate-speech/

[2]https://www.gov.uk/government/consultations/online-harms-white-paper

[3]https://www.argentina.gob.ar/noticias/discursos-de-odio

[4]Para la RAE, un *meme* es una imagen/video/texto con fines caricaturescos.

979-8-3315-4097-5/24/$31.00 ©2024 IEEE

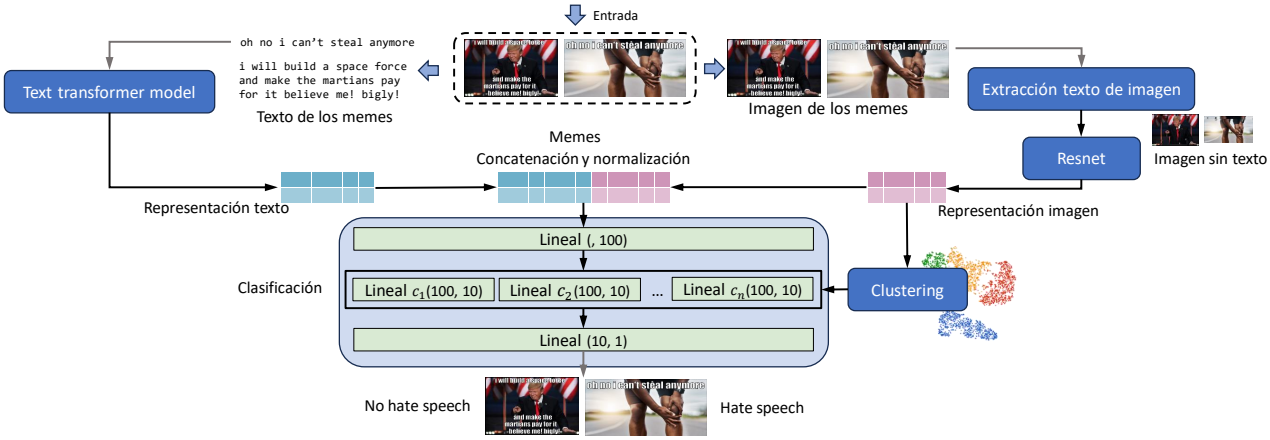

Figura 1: Representación esquemática de la propuesta para detectar *hate speech* en memes

En lo que respecta al análisis del contenido visual, Suryawanshi et al. [10] se basaron en *ImageNet* para clasificar *memes* de origen Hindú. Sabat et al. [11] fueron de los primeros en abordar la detección multimodal. La representación textual se basó en BERT, mientras que la visual en VGG-16, las cuales fueron utilizadas como entrada para un perceptrón multicapa. De acuerdo con los autores, la representación solo con imágenes alcanzó mejores resultados que la de texto. Fersini et al. [12] compararon diferentes técnicas para la identificación de *memes* sexistas. Para el texto evaluaron una representación simple de *bag-of-words* y *Support Vector Machines*, mientras que para las imágenes usaron características de bajo y alto nivel (contraste, rugosidad, varianza de croma, entropía, complejidad de la imagen y grado de enfoque, entre otros) y un clasificador de *1-Nearest Neighbours*. Los resultados mostraron que el texto aportaba más información que el contenido visual y que al combinarlos se obtenían mejores resultados. Gomez et al. [5] analizaron la presencia de *hate speech* en *memes* de *Twitter* incluyendo también el texto de los *tweets*. Para ello, utilizaron una LSTM para extraer las características de ambos textos, e *ImageNet* para las características visuales. Estas tres representaciones se combinaron con modelos CNN+RNN. En este caso, el análisis multimodal no superó al unimodal.

En 2020, *Facebook AI* lanzó una competencia de detección de *hate speech* en memes[5] para promover las soluciones multimodales. Como resultado, surgieron diversos trabajos (e.g., [13, 14]) que explotaron *image captioning*, *transformers* visuales y texuales (mayoritariamente BERT) e incluso análisis de sentimientos sobre el texto e imagen. El ganador de la competencia [15] incluyó también información de entidades y etiquetas de raza y género, buscando dar contexto al modelo.

Es importante destacar la disparidad de los resultados reportados, por ejemplo, precisión entre 0,6 [9] y 0,9 [6]. Esto podría estar asociado, a sesgos en la recolección de datos, los cuales podrían haber influenciado la "dificultad" de la tarea.

III. Propuesta

A diferencia de los enfoques en la literatura que, en su gran mayoría, se enfocan en combinar las representaciones textuales y visuales en el contexto de una tarea simple de *clasificación*, en este trabajo se analiza una propuesta (esquematizada en la Figura 1) basada en la *recuperación de información*. Esta propuesta se basa en la utilización de redes neuronales para la extración de características textuales y visuales de los *memes* y técnicas de clústering de imágenes y clasificación de texto para determinar la presencia de *hate speech*.

III-A. Procesamiento de las características

Limpieza de imágenes: Generalmente, los *memes* se conforman de una imagen con texto superpuesto. Un problema frecuente es la interferencia del texto en el análisis de la imagen quitando el foco de su contenido. Es por esto que se eliminó el texto en las imágenes combinando Keras *Optical Character Recognition* (para la detección de las zonas donde se encontraba el texto) con Navier-stokes[6], un algoritmo de *inpainting* (para repintar las zonas donde estaba el texto).

Extracción de características de las imágenes: Una vez obtenida una imagen sin texto, se obtuvieron sus características. Haciendo uso de *transfer learning*, se utilizó el modelo *ResNet152*[7] para la obtención de las características. Este modelo se encuentra pre-entrenado sobre *ImageNet* que consta de 1000 categorías y 1,2 millones de imágenes[8]. La representación de las imágenes se obtuvo seleccionado la anteúltima capa del modelo, con dimensión 2048.

Extracción de características del texto: Para las características textuales se realizaron pruebas con dos variantes de modelos pre-entrenados comúnmente utilizados en la literatura, BERT[9] y *SentenceTransformer*[10].

Concatenación y normalización: Las representaciones del texto y la imagen fueron concatenadas para ser utilizadas como entrada del modelo multimodal. Una vez concatenadas, para facilitar el aprendizaje de la red, se aplicó estandarización, reduciendo así el rango de valores que pueden adoptar.

[5]https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/

[6]https://opencv24-python-tutorials.readthedocs.io/en/latest/py_tutorials/py_photo/py_inpainting/py_inpainting.html

[7]https://pytorch.org/hub/pytorch_vision_resnet/

[8]De acuerdo a la literatura, este modelo supera a AlexNet, GoogleNet y VGG, además de evitar el problema del *vanishing gradient*.

[9]https://huggingface.co/bert-base-uncased

[10]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

III-B. Clústering de imágenes

Las representaciones de las imágenes se utilizaron como entrada para el algoritmo de clústering. El objetivo era limitar el dominio sobre el cual clasificar los memes, haciendo que la clasificación de un nuevo *meme* considere el clúster de elementos más similares a él. Se evaluaron dos técnicas para esta tarea, *K-Means* y *Bisecting K-Means*. La cantidad óptima de clústers se obtuvo utilizando el método de la silueta.

III-C. Clasificación de hate speech

El clasificador se definió como un modelo de 3 niveles lineales con las características concatenadas como entrada, donde dos niveles son comunes a todos los elementos y uno depende de la estructura de clústering. El primer nivel cuenta con una capa de $(2432, 100)$ (*SentenceTransformer*) o $(2816, 100)$ (BERT) dimensiones de entrada y salida, y una función de activación *LeakyReLu*. El segundo nivel tiene tantas capas en paralelo como clústers encontrados, cada una con $(100, 10)$ dimensiones y también una función de activación *LeakyReLu*. La capa utilizada en este nivel durante el entrenamiento o evaluación depende del clúster al que pertenece el *meme* usado para entrenar o evaluar, esto no solo integra el contexto de la imagen al análisis, sino que, en un futuro, podría servir para explicar las predicciones. Luego, el último nivel cuenta con una capa con una función de activación sigmoide que determina la predicción para el meme. Esto genera dos tipos de aprendizaje, el obtenido del entrenamiento con todos los elementos en el primer y último nivel, y el obtenido como resultado del entrenamiento sobre la capa correspondiente al clúster seleccionado en el segundo nivel.

La predicción obtenida indica la probabilidad de que el *meme* evaluado sea *hate speech*. Luego, debe definirse un *threshold* que indique cuándo se considera *hate speech*. Para evitar disparidad de resultados a partir de un *threshold random*, se realizó un análisis basado en el área debajo de la curva, eligiendo un *threshold* que maximice el Índice de Youden [16]. Este índice, con valores entre 0 y 1, informa el rendimiento de una decisión, donde 1 refiere al clasificador perfecto.

IV. Evaluación experimental

Colección de datos: Se utilizó la colección de *Facebook AI* para el *Hateful Memes Challenge*, la cual cuenta con 10 mil *memes* generados artificialmente etiquetados de forma binaria. Para dificultar la clasificación unimodal, por cada *meme hate speech* crearon "*confounders*", nuevos elementos que cambiaban el texto o la imagen original para alterar su significado, y asociarlos con la clase opuesta. El dataset se encuentra pre-dividido en: entrenamiento (85 %, 35 % *hate speech*), validación (5 %) y test (10 %, 50 % *hate speech*).

Métricas de evaluación: Los resultados se evaluaron utilizando métricas de clasificación binaria: *accuracy* (para poder comparar los resultados con los de la competencia), y precisión y *recall* para la clase *hate speech*.

Selección de baselines: Se consideraron diversos baselines. Para `Unimodal-texto`, se refinó BERT, agregando una capa densa y *Dropout*, utilizando la función de pérdida *Binary Cross-Entropy* y el optimizador *Adam* (el mismo utilizado originalmente por BERT). Para `Unimodal-imagen`, se definió un modelo con cuatro bloques de capas formados por una capa convolucional con activación lineal, una *LeakyRelu*, una *MaxPooling* y una *Dropout* seguidos por una capa *Flatten*, una capa densa con función de activación lineal, una *Dropout* y una última capa densa con función de activación *Softmax*. Se utilizó la función de pérdida *Categorical Cross-Entropy*. Se seleccionaron dos trabajos de la literatura [13, 14] basados en *image captioning* y Visual BERT. No se consideró al ganador de la competencia debido a que incorporaba información externa, resultando computacionalmente costoso y limitando la aplicabilidad del enfoque a determinados contextos. Debido al desbalanceo en el conjunto de entrenamiento, se agregó el cómputo de los pesos a la secuencia, además de la randomización de las muestras para entrenamiento.

Se consideran también variantes del modelo analizado. Primero, una en la que no se entrena un modelo de clasificación (`MSemejanza`), sino que una vez encontrado el clúster más similar, se calcula la semejanza del coseno entre el texto del *meme* y los textos en dicho clúster, utilizando la etiqueta del texto más cercano para determinar si el nuevo *meme* es *hate speech*. Segundo, una en la que no se realiza clústering (`MNo-clústering`), resultando en un único modelo común de clasificación. Finalmente, se consideró invertir el orden de análisis, es decir, clusterizar sobre el texto y clasificar sobre las imágenes. Sin embargo, se descartó debido a que no fue posible verificar la existencia de clústers distinguibles.

Detalles de implementación: Se utilizó Python[11]. El modelo de clasificación, su entrenamiento y la extracción de las características de las imágenes se implementó con PyTorch.

V. Resultados de la evaluación

La Tabla I[12] presenta los resultados obtenidos. Para el modelo presentado (`MClústering`) se reportan las diferentes combinaciones de variantes descriptas en la Sección III. Los resultados obtenidos para los enfoques unimodales son inferiores que los obtenidos por los multimodales, lo que refuerza la importancia de estudiar estas propuestas. En el caso de `MClústering`, la cantidad óptima de clústers fue 5, salvo para `MClústering`-*BiKmeans + BERT* que fue 8. Si bien una mayor cantidad de clústers permitiría hacer una mejor distinción entre los aspectos visuales de los memes, también implica que habrá menos *memes* por clúster, y por lo tanto menos texto para entrenar los clasificadores, lo cual puede afectar su capacidad predictiva. En lo que respecta a los modelos de texto, los mejores resultados fueron obtenidos por BERT, con diferencias más amplias en términos de *recall*.

Las variantes de `MNo-clústering` obtuvieron resultados inferiores que `MClústering`, lo que indica la importancia del contexto de los memes, y la potencial relación de se-

[11] Para más detalles de la implementación, ver el repositorio: https://github.com/magaliboulanger/HateSpeechDetection

[12] Para los enfoques de la literatura solo se incluye *accuracy* ya que fue la única métrica reportada en la competencia, y no se tuvo acceso a las implementaciones para replicar la evaluación. Considerando que los datos se encuentra balanceado, no se incluyen modelos como *random* o más popular.

Modelo	*Accuracy*	Precisión	*Recall*
Unimodal-Texto	0,5	0,5	0,5
Unimodal-Imagen	0,55	0,5	0,65
Das et al. [13]	0,68	-	-
Aggarwal et al. [14]	0,65	-	-
MSemejanza	0,574	0,424	0,605
MNo-clústering + *BERT*	0,566	0,566	0,707
MNo-clústering + *SentenceTransformer*	0,589	0,571	0,690
MClústering-*Kmeans* + *BERT*	0,603	0,600	0,617
MClústering-*BiKmeans* + *BERT*	0,582	0,557	0,812
MClústering-*Kmeans* + *SentenceTransformer*	0,622	0,631	0,594
MClústering-*BiKmeans* + *SentenceTransformer*	0,607	0,590	0,708

TABLA I: Resultados obtenidos en la evaluación realizada

mejanza entre las imágenes e, implícitamente, de los objetos en ellas. A diferencia de MClústering, no se observaron grandes diferencias en el modelo textual utilizado. Finalmente, la diferencia con el *accuracy* de los modelos de la literatura [13, 14] es muy pequeña como para determinar si la misma se debe a los modelos o al azar. Al no haber podido calcularse las otras métricas, no es posible determinar qué clase identificaban mejor dichos modelos, siendo que en este caso, resulta más valioso identificar las instancias de la clase *hate speech*. Por otra parte, es importante destacar la poca diferencia en los resultados siendo que dichos modelos incorporan procesamientos adicionales, incrementando su costo computacional.

Evaluación de la generalización del modelo: Para conocer la capacidad de generalización de MClústering, se realizó una evaluación tomando el mejor modelo entrenado y evaluándolo con el conjunto de datos presentado por Gomez et al. [5]. Esto se conoce como *zero-shot-learning* y permite evaluar la dependencia del modelo de los datos de entrenamiento. Esta colección contiene 150 mil *memes* extraídos de *Twitter* con un 61 % de *memes* de *hate speech* Considerando que para cada *meme* se tenían 3 etiquetas, se evaluaron diversas formas para elegir sus etiquetas, se optó por considerar que un *meme* era *hate speech*, si todas las etiquetas coincidian en eso.

Unimodal-Texto obtuvo mejores resultados que para la otra colección, alcanzando una precisión y *recall* de 0,98 y 0,55, mientras que Unimodal-Imagen tuvo una precisión de 0,61. Esto puede deberse al impacto de los factores de confusión en la colección de entrenamiento, ya que un mismo texto/imagen puede pertenecer a ambas clases a la vez. Los resultados obtenidos con MClústering no evidencian capacidad de adaptación a nuevos datos. Por el contrario, reflejan un desempeño similar al que se obtendría en una clasificación aleatoria. Esto podría deberse a la naturaleza artificial de la colección de datos (que difiere de lo que se pueda encontrar en un ambiente real) y su tamaño (puede resultar insuficiente para el entrenamiento). Es así que la colección de *Twitter* no solo puede presentar una mayor variedad de memes, sino también algunos representantes de los casos explícitamente excluídos por *Facebook AI*, como por ejemplo, violencia explícita.

VI. Conclusiones

El *hate speech* es una problemática que requiere atención debido a su constante crecimiento en los medios sociales. En este trabajo se analizó una alternativa a las propuestas en la literatura explorando la combinación de técnicas de clústering y clasificación intentando conceptualizar el contexto de los memes. Si bien los resultados obtenidos no igualaron los reportados en la literatura, las diferencias observadas fueron pequeñas como para determinar la efectiva superioridad de los modelos de la literatura. En este sentido, este trabajo marca el primer paso en un camino alternativo a los ya propuestos, con potencial de obtener mejoras sustanciales mediante ajustes simples y sin demandar cantidades prohibitivas de recursos. Futuras propuestas podrían realizar una evaluación cualitativa de los resultados y abordar alternativas para la representación base del contenido textual y visual (e.g., mediante técnicas de detección de objetos o tópicos), explorar otras técnicas de clústering incorporando no solo caraterísticas visuales, sino también de entidades detectadas; y finalmente, incrementar la cantidad de datos de entrenamiento con *memes* "reales" para favorecer la generalización del modelo.

Consideraciones éticas

El trabajo se basó mayoritariamente en una colección de datos artificial creada por terceros, la cual no contiene información personal. La colección de datos de *Twitter/X*, también fue creada y etiquetada por terceros, y ninguna identidad ha sido revelada.

Referencias

[1] J. T. Nockleby, "Hate speech in context: The case of verbal threats," *Buff. L. Rev.*, vol. 42, p. 653, 1994.

[2] H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," in *SocInfo*. Springer, 2015, pp. 49–66.

[3] J. Y. Kim and A. Kesari, "Misinformation and hate speech: The case of anti-asian hate speech during the covid-19 pandemic," *Journal of Online Trust and Safety*, vol. 1, no. 1, 2021.

[4] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggression and violent behavior*, vol. 40, 2018.

[5] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in *WACV – IEEE/CVF*, 2020.

[6] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach," *arXiv preprint arXiv:1809.08651*, 2018.

[7] G. Rajput, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Hate speech detection using static bert embeddings," in *Big Data Analytics: 9th International Conference, BDA 2021, Virtual Event, December 15-18, 2021, Proceedings 9*. Springer, 2021, pp. 67–77.

[8] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, "A framework for hate speech detection using deep convolutional neural network," *IEEE Access*, vol. 8, pp. 204 951–204 962, 2020.

[9] I. M. A. Niam, B. Irawan, C. Setianingsih, and B. P. Putra, "Hate speech detection using latent semantic analysis (lsa) method based on image," in *2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)*. IEEE, 2018, pp. 166–171.

[10] S. Suryawanshi, B. R. Chakravarthi, P. Verma, M. Arcan, J. P. McCrae, and P. Buitelaar, "A dataset for troll classification of tamilmemes," in *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, 2020, pp. 7–13.

[11] B. O. Sabat, C. C. Ferrer, and X. Giro-i Nieto, "Hate speech in pixels: Detection of offensive memes towards automatic moderation," *arXiv preprint arXiv:1910.02334*, 2019.

[12] E. Fersini, F. Gasparini, and S. Corchs, "Detecting sexist meme on the web: A study on textual and visual cues," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2019, pp. 226–231.

[13] A. Das, J. S. Wahi, and S. Li, "Detecting hate speech in multi-modal memes," *arXiv preprint arXiv:2012.14891*, 2020.

[14] A. Aggarwal, V. Sharma, A. Trivedi, M. Yadav, C. Agrawal, D. Singh, V. Mishra, and H. Gritli, "Two-way feature extraction using sequential and multimodal approach for hateful meme classification," *Complexity*, vol. 2021, pp. 1–7, 2021.

[15] R. Zhu, "Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution," *arXiv preprint arXiv:2012.08290*, 2020.

[16] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.