Ometeotl@Multimodal Hate Speech Event Detection 2023: Hate Speech and Text-Image Correlation Detection in Real Life Memes Using Pre-Trained BERT Models over Text

Jesús Armenta-Segura and **César-Jesús Núñez-Prado** and **Grigori Sidorov** and **Alexander Gelbukh** and **Rodrigo Román-Godínez**
Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico City, Mexico
{jarmentas2022, sidorov, gelbukh, rromang2019}@cic.ipn.mx,
cnunezp@ipn.mx

Abstract

Hate speech detection during times of war has become crucial in recent years, as evident with the recent Russo-Ukrainian war. In this paper, we present our submissions for both subtasks from the Multimodal Hate Speech Event Detection contest at CASE 2023, RANLP 2023. We used pre-trained BERT models in both submission, achieving a F1 score of 0.809 in subtask A, and F1 score of 0.567 in subtask B. In the first subtask, our result was not far from the first place, which led us to realize the lower impact of images in real-life memes about feelings, when compared with the impact of text. However, we observed a higher importance of images when targeting hateful feelings towards a specific entity. The source code to reproduce our results can be found at the github repository https://github.com/JesusASmx/OmeteotlAtCASE2023

1 Introduction

In recent decades, online platforms have gained increasing relevance in the worldwide sociopolitical scenario, to the extent that they have become significant representations of the so-called *soft power* (Mavrodieva et al., 2019). This growing importance has also led to an alarming spread of offensive, discriminatory, and harmful content, particularly during periods of significant political changes such as elections (Ezeibe, 2021) or wars (Aslan, 2017; Thapa et al., 2022).

Detecting hate speech, both in text and images, is crucial in order to mitigate its negative impact on digital platforms and safeguarding individuals from its harmful effects (Parihar et al., 2021). As an example of this need, in 2022, social networks witnessed a surge in activity following the outbreak of the Russo-Ukrainian war; numerous content, full of hate speech from both sides, went viral, and the need for a specific focus to that particular conflict became evident.

For this reason, the Multimodal Hate Speech Event Detection contest was proposed during the CASE 2023 workshop (Thapa et al., 2023) to tackle this problem with a dataset of manually annotated text-image memes (Bhandari et al., 2023). This shared task was divided into two subtasks A and B. In subtask A, participants were required to determine whether a meme related to the Russo-Ukrainian war constituted hate speech or not. In subtask B, participants were tasked with identifying the target of a hate speech meme, classifying it as directed against an individual (such as Volodymyr Zelensky or Vladimir Putin), an organization (such as the Ukrainian army), or a community (such as the Russian speakers in the Donbass region).

In this paper, we present out participation in both subtasks, under the name of *Team Ometeotl*. Our proposal consists on a fine-tunning of the pre-trained BERT model (Devlin et al., 2018), trained solely on the text extracted from the memes, without incorporating any image feature. Surprisingly, those experiments outperformed models that considered image features, such as ResNet152, and even multimodal ensemble learning approaches, such as ResNet152+BERT. These approaches achieved the sixth position in Subtask A, with an $F1$ score of 0.809, and the seventh position in Subtask B, with an $F1$ score of 0.567.

The structure of the paper is as follows: in Section 2, we describe the updated research on automatic hate speech detection. In Section 3, we describe the database. In Section 4, we detail the methodology used. In Section 5, we show the results of our experiments. In Section 5, we discuss the results. Finally, in Section 6 we present the conclusions.

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 53–59, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_007

2 Related Work

Hate speech detection in social media are one of the most prominent classification tasks in recent years (Schmidt and Wiegand, 2017). One of the earliest known approaches is the General Inquirer (Stone and Hunt, 1963), an IBM system developed in 1961 that enabled content analysis for behavioral sciences. It focused on pattern detection in text to categorize words based on their semantics, particularly positive or negative sentiments. In 1997, a more targeted approach was proposed with the system Smokey (Spertus, 1997), designed to detect abusive messages. Smokey utilized a rule-based approach to identify offensive language and contexts.

From there, several new approaches were proposed to address the task and its variations. In (Warner and Hirschberg, 2012), the authors proposed a lexicon-based approach for hate speech detection, starting from the hypothesis that the task can be related with word sense disambiguation. However, such approach was vulnerable in front of incomplete datasets, as they discovered when every method learnt *jew* as a inherent word for antisemitism speech. In order to deal to this sort of datasets, several methods and further methodologies has been developed: one of the most recent machine learning techniques who had brought promising results are the transformers (Vaswani et al., 2017), including BERT models (Devlin et al., 2018). In a nutshell, BERT models are is a family of language models composed of Transformer encoder layers. Such architectures has been successfully used in transphobic-homophobic speech detection, as can be seen in the LT-EDI-ACL2022 homophobia/transphobia speech detection contest in English, Tamil and Tamil-English (Chakravarthi et al., 2022). Team Sammaan (Upadhyay et al., 2022) employed ensemble transformers and obtained the second place in English; team Nozza (Nozza, 2022) obtained the third position in English and used ensemble learning over fine-tunned models of BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and HateBERT (Caselli et al., 2021).

Another hate speech detection contest in which transformers were used was in the IberLEF2023 shared task of HOMO-MEX: Hate Speech Detection towards the Mexican Spanish-Speaking LGBT+ (Bel-Enguix et al., 2023). Contrary to the previous contest, in which only the first places used transformers (the last place used TF-IDF with traditional classifiers such as Support Vector Machine (Swaminathan et al., 2022)), here team LIDOMA, the last place of the competition, employed a BERT model (Shahiki-Tash et al., 2023). However, in their paper, the authors explained how the lack of a preprocessing highly affected the efficiency of the attention mechanisms. To dive further, in this work we find a counterexample to their hypothesis in the shared task A, where preprocessing actually brought worse results.

All the related works discussed so far have focused solely on text-based hate speech detection. This is because text has historically been the most prevalent format for hate speech across the internet, especially during the early days of the worldwide web. However, it is crucial to recognize that there exists a wealth of historical data on hate speech in images, such as visual propaganda (Margolin, 1979), extensive datasets from the Second World War (Kallis, 2005; Basilio, 2014) and the Cold War (Snyder, 1995). Nevertheless, it is worth noting that these works were handcrafted by artists, hence impossible to get mass-produced during the early stages of the worldwide web, unlike the solely text-based propaganda. This landscape has since changed with the advent of text-image memes, which are pre-designed images that can have accompanying text, making possible the mass-production of visual propaganda and hence attracting the attention of researchers all across the world. For instance, Meta AI initiated the paid contest titled "Hateful Memes Challenge and Dataset for Research on Harmful Multimodal Contest" (Kiela et al., 2020), in which they provided a dataset of memes and the task to detect hate speech on them. One of the most interesting aspects of the challenge was that the dataset considered the significant phenomena of text-image interaction through phrase-sense disambiguation. For instance, a meme featuring the text *I love the way you smell today* could be classified as hate speech if accompanied by an image of a skunk (Mephitidae), but as non-hate speech if accompanied by a picture of a rose.

Another relatable example of multimodal hate speech text-image detection is (Perifanos and Goutsos, 2021). In this work, the authors combined natural language processing techniques with computer vision models to analyze both text and images in greek social media. For text processing they fine-tuned a pre-trained BERT model (Devlin

et al., 2018). For image processing, the authors fine-tuned a pre-trained ResNet118 (He et al., 2015) in the ImageNet dataset (Deng et al., 2009). Their best result was an F1 score of 0.947.

In (Yang et al., 2022), the authors proposed a multimodal hate speech detection approach that uses cross-domain knowledge transfer to improve hate speech detection accuracy. To address the semantic inconsistency between hate speech and sarcasm, the authors combined the contrastive attention mechanism with representational dissociation to design a semantic adaptive module. In addition, they applied curricular learning to accelerate the training process. Experimental results showed that the proposed approach outperformed existing multimodal hate speech detection methods in terms of accuracy and F1-score on two public datasets: the Facebook Hateful Memes dataset from the Meta AI's contest, mentioned before, and the Twitter sarcasm detection dataset (Cai et al., 2019).

3 Dataset

The dataset for both subtasks consists in $6,913$ text-image memes concerning the Ukraine-Russia conflict. These samples were collected from social media platforms such as Twitter, Reddit and Facebook with keywords for specialized searches. The labeling was done manually, and they used Cohen's Kappa statistical measure (Matthijs, 2015) to assess the agreement between two or more annotators which ranges from -1 to 1, where the value 1 indicates perfect agreement, 0 indicates casual agreement and -1 total disagreement (Bhandari et al., 2023).

3.1 Sub-task A

The main goal is to identify whether or not a text-image meme contains hate speech or not. The training set for this sub-task contains $3,600$ images in jpg format, where $1,942$ are hate speech and $1,658$ are no hate speech. There is also a evaluation set with samples, where 243 are hate speech and 200 are no hate speech. Finally, the test set consists in 443 images. Table 1 shows the statistics for this subtask.

3.1.1 Sub-task B

In this task, the goal is to identify to whom the hate speech of a given meme is directed. Possible targets to be identified are community, individual and organization. For this task, the training dataset consists of $1,942$ images in jpg format, where 335 are hate speech against a community, 823 are directed towards an individual and 728 are aimed to an organization. There is also an evaluation set with 244, where 102 are community, 40 are individual and 101 are organization. Finally, the test set has 242 images. Table 2 shows the statistics for this subtask.

Label	Amount	Data
Hate Speech	1,942	Training
No Hate Speech	1,658	Training
Hate speech	243	Evaluation
No Hate speech	200	Evaluation
–	443	Test

Table 1: Subtask A Dataset Statistics.

Label	Amount	Data
Community	335	Training
Individual	823	Training
Organization	784	Training
Community	102	Evaluation
Individual	40	Evaluation
Organization	101	Evaluation
–	242	Test

Table 2: Subtask B Dataset Statistics.

In addition to the text-image memes, the organizers also provided the texts, extracted with the Google vision API[1]. Table 3 shows examples of these extracted texts.

Label	Example
Hate	Death of Russian
No Hate	Putin recognises Ukraine rebel region
Community	Russian troop pronuons are were was
Individual	Zelenskyys massiv balls Putins balls
Organization	Love is sitting together and watching Russian tanks burn

Table 3: Example of texts extracted from the memes.

4 Methodology

The first step was to encode each labeling into a numerical value. In the case of subtask A, 0 was

[1]https://cloud.google.com/vision/

used to represent no hate speech and 1 to indicate hate speech. In subtask B, we utilized 0 for hateful messages towards a community, 1 for individual, and 2 for organization. All these labelings were chosen following the indications of the organizers.

The next step involved an optional preprocessing outside the BERT processing of the text. It consisted of a function that removed special characters, converted to lowercase, and removed the stopwords using the spacy python library[2]. The main idea behind this function was to enhance the efficiency of the attention mechanisms as mentioned in (Shahiki-Tash et al., 2023). However, as anticipated in Section 2, it only worked for subtask B.

Regarding the model specifications, we utilized the *BertForSequenceClassification* model with the *bert-base-uncased* architecture, which was pretrained on the English corpus. The employed parameters for the preparation of the data were:

- add_special_tokens = True,
- max_length = 256,
- padding = max_length,
- return_attention_mask = True,
- Truncation = True,
- return_tensors = pt.

The input tokens and attention masks were concatenated into separate tensors using the *torch.cat* and *torch.tensor* libraries.

The parameter for training the *bert-base-uncased* model were:

- number of labels = 2 (for Sub-task B number of labels = 3),
- optimizer = AdamW, with a learning rate of 2e-5,
- batch size = 16,
- with training inputs, training masks and training labels is created a *TensorDataset*,
- epochs = 4.

The system infraestructure consisted in a CPU with a AMD Rysen 2 5600x processor with six kernels, along with 46gb of RAM. With this system, the run for the subtask A spent around ten hours while the run for the subtask B spent around eight.

[2]https://spacy.io/

5 Results and Discussion

In subtask A, we did not utilized preprocessing and achieved an F1 score of 0.809. The first-place score was 0.856, which is only 0.047 points higher than ours. This difference is relatively low, especially when considering that we did not employ image features in our predictions. See Table 4 to check the full leaderboard of subtask A, with F1 score and Accuracy.

Team	F1	Accuracy
arc-nlp	0.856	0.858
bayesiano98	0.853	0.853
karanpreet_singh	0.846	0.846
DeepBlueAI	0.834	0.835
csecudgs	0.825	0.826
Ometeotl	**0.810**	**0.810**
Avanthika	0.788	0.790
Sarika11	0.782	0.759
rabindra.nath	0.780	0.783
md_kashif_20	0.729	0.736
Sathvika.V.S	0.429	0.578
lueluelue	0.522	0.526
pakapro	0.494	0.497

Table 4: Sub-task A Results. Numbers were rounded up from 6, starting on the fourth digit. (Team Ometeotl achieved a $F1$ score of 0.8099)

In subtask B, we employed preprocessing and achieved an F1 score of 0.567. This time, the difference with the first-place score was more substantial (of 0.195 points), leading us to hypothesize that visual features may have a stronger correlation when determining the target of a hateful meme. The leaderboard of this subtask can be found in Table 5.

5.1 Image features in subtask A

To incorporate visual features and improve the results, we experimented with ResNet152 on the image data alone. Initially, without data augmentation, the best F1 score achieved in subtask A was 0.55, but the model exhibited significant overfitting. To address this issue, we augmented the data ten times by performing rotations, expansion, and narrowing, which resulted in an enhanced F1 score of 0.71. However, this performance was still far below that of BERT. We attempted Voting Ensemble, but it only led to a marginal improvement, reaching an F1 score of 0.76, so we discarded it for the last submission.

Team	F1	Accuracy
arc-nlp	0.763	0.793
bayesiano98	0.741	0.773
karanpreet singh	0.697	0.723
Sarika22	0.680	0.715
csecudgs	0.653	0.690
DeepBlueAI	0.652	0.698
Ometeotl	**0.568**	**0.640**
Avanthika	0.526	0.640
Sathvika.V.S	0.433	0.529
pakapro	0.334	0.351

Table 5: Sub-task B Results. Numbers were rounded up from 6, starting on the fourth digit.

We hypothesize that the reason for the low correlation between visual features and hate speech in subtask A is that images in memes are primarily used as conceptual support for the message rather than pragmatic support. For instance, consider Figure 1. In this figure, sample $11,381$ is labeled as hate speech due to its text, but its visual features consist entirely of the well-known *The-What* meme[3], which solely portrays a woman with a funny smile, and no further information about whether the messages is hateful or not. On the other hand, sample $10,465$ consists in a frame from the movie *Star Wars I: The Phantom Menace*[4], in which an old man (Governor Sio Bibble) is sitting in a wide chamber while speaking, once again, withouth further visual information about the emotion of the message.

Figure 1: On top, sample $10,465$ labelled as no hate speech. On bottom, sample $11,381$ labelled as hate speech.

6 Conclusions

In this paper, we presented our approach to address both subtasks from the Multimodal Hate Speech Event Detection at CASE 2023, which consists in A) Detect hate speech in text-image memes spread during the Russo-Ukranian war, and B) given a hateful meme about that conflict, determine if the target is a community, an individual or an organization. We utilized text-based transformers, specifically fine-tunned pre-trained BERT models, and achieved high results in subtask A using only text features.

Our methodology involved the numerical encoding of the labels, and a preprocessing step for subtask B consisting in lowercase conversion and the removal of stopwords and special characters. Afterward, we conducted a four-epoch training of the fine-tunned pre-trained BERT model *bert-base-uncased*.

We discovered that visual features played a more significant role in determining the target of hate speech rather than determining whether the meme itself was hateful or not, at least in this particular database. As a result, further research and analysis are needed to explore this phenomenon comprehensively. Exploring other datasets could provide valuable insights into the dynamics between visual features and hate speech, offering a more comprehensive understanding of the varying impact these elements have across different contexts and social settings. Such investigations can shed light on the broader implications of visual cues and how they interact with textual content in influencing the perception and spread of hateful memes.

Acknowledgments

This work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20232138, 20232080, 20231567 of the Secretaría de Inves-

[3]https://knowyourmeme.com/memes/the-what-rug-doctor-woman-ad

[4]https://knowyourmeme.com/photos/1810076-prequel-memes

tigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Alev Aslan. 2017. Online hate discourse: A study on hatred speech directed against syrian refugees on youtube.

Miriam Basilio. 2014. *Visual Propaganda, Exhibitions, and the Spanish Civil War*. Ashgate Publishing, Ltd.

Gemma Bel-Enguix, Helena Gómez-Adorno, Gerardo Sierra, Juan Vásquez, Scott-Thomas Andersen, and Sergio Ojeda-Trueba. 2023. Overview of HOMO-MEX at Iberlef 2023: HOMO-MEX: Hate Speech Detection in Online Messages Directed Towards the MEXican Spanish Speaking LGBTQ+ Population. *Procesamiento del lenguaje natural*, 71.

Aashish Bhandari, Siddhant Bikram Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from Russia-Ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*.

Chidiebere Ezeibe. 2021. Hate Speech and Election Violence in Nigeria. *Journal of Asian and African Studies*, 56(4):919–935.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Aristotle Kallis. 2005. *Nazi propaganda and the second world war*. Springer.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *ArXiv*, abs/2005.04790.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Victor Margolin. 1979. The visual rhetoric of propaganda. *Information design journal*, 1(2):107–122.

J Warrens Matthijs. 2015. Five ways to look at Cohen's Kappa. *Journal of Psychology Psychotherapy*, 5(4).

Aleksandrina V Mavrodieva, Okky K Rachman, Vito B Harahap, and Rajib Shaw. 2019. Role of social media as a soft power tool in raising public awareness and engagement in addressing climate change. *Climate*, 7(10):122.

Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 258–264, Dublin, Ireland. Association for Computational Linguistics.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*.

Moein Shahiki-Tash, Jesús Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Lidoma at homomex2023@iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In *CEUR Workshop Proceedings*.

Alvin A Snyder. 1995. *Warriors of disinformation: American propaganda, Soviet lies, and the Winning of the Cold War: an insider's account*. Arcade Publishing.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: Studies using the general inquirer system. *Proceedings of the May 21-23, 1963, spring joint computer conference*.

Krithika Swaminathan, Bharathi B, Gayathri G L, and Hrishik Sampath. 2022. SSNCSE_NLP@LT-EDI-ACL2022: Homophobia/transphobia detection in multiple languages using SVM classifiers and BERT-based transformers. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 239–244, Dublin, Ireland. Association for Computational Linguistics.

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection - shared task 4, case 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*.

Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A multimodal dataset for hate speech detection on social media: Case-study of Russia-Ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 1–6, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ishan Sanjeev Upadhyay, Kv Aditya Srivatsa, and Radhika Mamidi. 2022. Sammaan@LT-EDI-ACL2022: Ensembled transformers against homophobia and transphobia. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 270–275, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *MM'22: Proceedings of the 30th ACM International Conference on Multimedia*.