MasonPerplexity at Multimodal Hate Speech Event Detection 2024: Hate Speech and Target Detection Using Transformer Ensembles

Amrita Ganguly*, **Al Nahian Bin Emran***, **Sadiya Sayara Chowdhury Puspo**
Md Nishat Raihan, Dhiman Goswami, Marcos Zampieri
George Mason University, USA
{agangul, abinemra}@gmu.edu

Abstract

The automatic identification of offensive language such as hate speech is important to keep discussions civil in online communities. Identifying hate speech in multimodal content is a particularly challenging task because offensiveness can be manifested in either words or images or a juxtaposition of the two. This paper presents the *MasonPerplexity* submission for the Shared Task on Multimodal Hate Speech Event Detection at CASE 2024 at EACL 2024. The task is divided into two sub-tasks: sub-task A focuses on the identification of hate speech and sub-task B focuses on the identification of targets in text-embedded images during political events. We use an XLM-roBERTa-large model for sub-task A and an ensemble approach combining XLM-roBERTa-base, BERTweet-large, and BERT-base for sub-task B. Our approach obtained 0.8347 F1-score in sub-task A and 0.6741 F1-score in sub-task B ranking 3^{rd} on both sub-tasks.

1 Introduction

In the context of polarized political discussions, when feelings and perspectives are strong, identifying offensive content is essential to moderation efforts in online communities. The challenge is increased by the use of text-embedded images in which negative emotions can be expressed both verbally and visually. Besides, in the current era of vlogging and reels, people are inclined to utilize memes and emojis or opt for text-embedded images to express their sentiments and comment on online content. As a result, the task of detecting hate speech is expanding to encompass images, posing a new challenge beyond the realm of textual content and across diverse languages.

The Shared Task on Multimodal Hate Event Detection at CASE 2024 (Thapa et al., 2024) deals with the identification of hate speech and its targets in text-embedded images during political events. The main objective is to automatically determine if an image that includes text contains hate speech (sub-task A) and, if so, to identify its targets categorized as community, individual, and organization (sub-task B). Identifying the target of offensive messages is vital to understanding their potential harm as demonstrated by annotation taxonomies such as OLID (Zampieri et al., 2019) and TBO (Zampieri et al., 2023).

In this paper, we discuss transformer-based approaches to hate speech detection in political events using the Multimodal Hate Speech Event Detection dataset (Bhandari et al., 2023). The paper sheds light on the challenges of handling multimodal content, particularly text-embedded images. For sub-task A (hate speech detection), we employ the XLM-roBERTa-large (Conneau et al., 2020) model. For sub-task B (target detection), we adopt an ensemble approach combining XLM-roBERTa-base, BERTweet-large (Ushio and Camacho-Collados, 2021), and BERT-base (Devlin et al., 2019). These models are selected to effectively address the unique challenges posed by diverse multimodal content. We report that our approach obtained a 0.8347 F1-score in sub-task A and a 0.6741 F1-score in sub-task B, ranking 3^{rd} on both sub-tasks.

2 Related Work

Offensive Content and Hate Speech Offensive content is pervasive in social media motivating the development of systems capable of recognizing it automatically. While definitions may vary, hate speech is arguably the most widely explored type of offensive content (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Several studies have proposed new datasets and models to label hateful posts on social media (Davidson et al., 2017; Zia et al., 2022). More recently, studies have focused on recognizing the specific parts of an instance that

* denotes equal contribution.
This paper contains offensive examples.

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 125–131
March 22, 2024 ©2024 Association for Computational Linguistics

may be considered offensive or hateful, as in the case of HateXplain (Mathew et al., 2021), TSD (Pavlopoulos et al., 2021), and MUDES (Ranasinghe and Zampieri, 2021). The vast majority of work on text-based hate speech detection is on English but several papers have created resources and models for languages such as Bengali (Raihan et al., 2023b), French (Chiril et al., 2019), Greek (Pitenis et al., 2020), Marathi (Gaikwad et al., 2021), and Turkish (Çöltekin, 2020).

Multimodal Hate Speech While the aforementioned studies have focused on the identification of hateful content in texts, there has been growing interest in identifying hateful content in text and images simultaneously. Hermida and Santos (2023), Ji et al. (2023), and Yang et al. (2022) highlight the significance of multimodal analysis offering a comprehensive overview of various methodologies employed to detect hate speech in images and memes. Various datasets have been introduced for multimodal hate speech detection (Grimminger and Klinger, 2021; Bhandari et al., 2023; Thapa et al., 2022) The study by Grimminger and Klinger (2021) presents a Twitter corpus with content related to the US elections of 2020. The study by Boishakhi et al. (2021) explores the combination of various modalities for hate speech detection such as text, video, and audio. While the clear majority of studies deal with English, research on different languages (Karim et al., 2022; Rajput et al., 2022; Perifanos and Goutsos, 2021).

Related Shared Tasks Thapa et al. (2023) organizes CASE 2023, a series of shared tasks identifying Multimodal Hate Speech Event Detection. There are two sub-tasks to identify hate speech and targets in the different sub-tasks. Participants present the utilization of transformer models like BERT, RoBERTa, and XLNet, as well as effective approaches such as vision transformers and CLIP which contributed to the outstanding outcomes. Similarly, different shared tasks have been organized to identify offensive language from texts i.e. (Aragón et al., 2019), (Modha et al., 2021). All of this research highlights how important it is to combine several data modalities in order to improve hate speech or offensive language detection.

3 Datasets

In sub-task A, the training dataset provided by the organizers contains 3,600 images. Additionally, a development set and a testing set were provided by the organizers each including 443 instances. Instances in the sub-task A dataset (Bhandari et al., 2023) are annotated using two labels: NO-HATE (labeled as 0) and HATE (labeled as 1). We present an example of the training data of sub-task A in Figure 1.

Figure 1: Training data example (Left: NO-HATE, Right: HATE)

The label distribution, presented in Table 1, is skewed in the dataset, with a slightly higher percentage of instances labeled as HATE in the training, testing, and evaluation sets.

	sub-task A		
Label	**Train**	**Eval**	**Test**
HATE	53.95	54.85	54.85
NO-HATE	46.05	45.15	45.15

Table 1: Distribution of labels in the training, evaluation, and test sets of the sub-task A dataset in terms of percentage.

In sub-task B, the training, evaluation, and test sets include 1,942, 244, and 242 images respectively. Instances in the sub-task B dataset (Thapa et al., 2022) are labeled into three categories: Individual (labeled as 0), Community (labeled as 1), and Organization (labeled as 2). Examples of training data for sub-task B are shown in Figure 2.

Figure 2: Training data example (Left: Organization, Top-right: Individual, Bottom-right: Community)

There is an imbalance among the three labels and the distribution is shown in Table 2. The class INDIVIDUAL is the most prevalent. The imbalance can impact the model's ability to generalize across different classes, potentially leading to biased results. Addressing this imbalance through techniques like data augmentation or re-balancing strategies may be crucial for developing robust models that perform well across all label categories.

sub-task B			
Label	**Train**	**Eval**	**Test**
INDIVIDUAL	42.38	41.80	42.15
COMMUNITY	17.25	16.40	17.35
ORGANIZATION	40.37	41.80	40.50

Table 2: label wise data percentage of sub-task B

We have used Google Vision API[1] to retrieve text from the images of all the phases of both the subtasks. Although the OCR can detect text in a variety of languages, the accuracy may change depending on the language. It's possible that some languages are more accurate and supported than others. The input image quality has an impact on OCR accuracy. In certain situations, the original formatting may not be preserved by the API.

4 Experiments

In sub-task A, we use BERTweet-large (Ushio and Camacho-Collados, 2021) (Ushio et al., 2022), BERT-base (Devlin et al., 2019), and XLM-R (Conneau et al., 2020) models. Notably, XLM-R shows the best F1 score. We also use GPT-3.5[2] zero-shot and few-shot prompting with test F1 score 0.73, 0.77. For sub-task B, we also start with BERTweet-large, BERT-base, and XLM-R using the same learning rate and epochs as in sub-task A. Later, we apply a weighted ensemble approach to these models, resulting in the 0.65 F1 score for the task. To tackle class imbalance in sub-task B, we employed back translation, converting the training data through Xosha to Twi to English and Lao to Pashto to Yoruba to English. This significantly improves overall model performance from 0.65 to 0.67.

The ensemble method with majority voting is proven helpful in this type of case where a single model may not be able to label the data correctly due to class imbalance (Goswami et al., 2023). Moreover, we follow the approach of back translation of (Raihan et al., 2023a). We follow the approach of back translation of (Raihan et al., 2023a). For this, we select languages that demonstrate limited or no cultural overlap with the original language featured in the dataset. Xosha, Twi, Lao, Pashto, and Yoruba are languages that are very diverse culturally and geographically. This diversity underscores the significance of considering a wide range of cultural and geographical influences when working with these languages. By intentionally selecting these languages without cultural overlap, we introduce a purposeful aspect of diversity, mitigating potential biases, and enhancing the dataset with a broader spectrum of linguistic expressions. Moreover, the Ensemble method with majority voting is also proven helpful in this type of case where a single model may not label the data correctly due to class imbalance (Goswami et al., 2023). For instance, when two out of three models predict a sentence as a hate event, the sentence is subsequently labeled as a hate event through the application of majority voting. We also use GPT-3.5 zero-shot and few-shot prompting with test F1 scores of 0.53, and 0.57. The prompt provided to GPT3.5 is available in Figure 3.

Role: You are a helpful AI assistant. You are given the task of <*sub-task_name*>.

Definition: <*sub-task_definition*>. You will be given a text to label either <*label1*> or <*label2*> or <*label3*>.

Task: Generate the label for this **text** in the following format: <*label*> *Your_Predicted_Label* <*label*>. Thanks.

Figure 3: Sample GPT-3.5 prompt.

We also utilize GPT-3.5 through the OpenAI API for two primary sub-tasks: Hate Speech Detection (sub-task A) and Hate Speech Target Detection (sub-task B). We fine-tune GPT-3.5 using specifically curated training and evaluation datasets, conducting the process over four epochs. It is worth noting that, no other hyper-parameter can be set other than epochs while fine-tuning GPT3.5 through the API. Notably, the OpenAI API does not provide conventional metrics such as training loss, validation loss, precision, or recall. Upon

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

[2]https://platform.openai.com/docs/models/gpt-3-5-turbo

completion of the fine-tuning, the API assigns a unique ID to our model. We use this ID to process the test dataset for both sub-tasks. For labeling and predictions, the API returns results based on the test dataset. In sub-task A, which focuses on detecting hate speech, our model achieves an F1 score of 0.82, indicating a high level of accuracy. Conversely, in sub-task B, where the objective is to identify the targets of hate speech, the model attains a lower F1 score of 0.63, reflecting the inherent challenges in this particular aspect of hate speech analysis.

Hyperparameters of all the models used excluding GPT3.5 in the experiments are available in Figure 3.

Parameter	Value
Learning Rate	$1e-5$
Train Batch Size	8
Test Batch Size	8
Epochs	5

Table 3: Training Configuration Parameters

5 Results

The detailed experimental results of the models in sub-task A and sub-task B are available in Tables 4, and 5, respectively. In sub-task A, we evaluate a BERT-base, BERTweet-large, and XLM-R model. XLM-R delivers the best performance with a 0.83 F1-score. In sub-task B, our ensemble approach was provides the best F1-score of 0.67.

Model	Eval F1	Test F1
GPT3.5 (ZERO SHOT)	–	0.73
GPT3.5 (FEW SHOT)	–	0.77
GPT3.5 (FINETUNED)	0.86	0.82
BERT-BASE	0.81	0.75
BERTWEET-LARGE	0.89	0.81
XLM-R	0.95	**0.83**

Table 4: Results of sub-task A.

6 Error Analysis

In sub-task A, our aim is to detect non-hate (labeled as 0) and hate (labeled as 1) speeches. Therefore, the task of our model is to categorize text into two categories: non-hate or hate. The confusion matrix, presented in Figure 4, illustrates both the true labels and predicted labels, indicating that our model

Model	Eval F1	Test F1
GPT3.5 (ZERO SHOT)	–	0.53
GPT3.5 (FEW SHOT)	–	0.57
GPT3.5 (FINETUNED)	0.65	0.63
BERT-BASE	0.61	0.60
XLM-R	0.63	0.61
BERTWEET-LARGE	0.68	0.64
ENSEMBLE	0.69	0.65
BERT-BASE (AUG.)	0.63	0.61
XLM-R (AUG.)	0.65	0.64
BERTWEET-LARGE (AUG.)	0.70	0.66
ENSEMBLE (AUG.)	0.71	**0.67**

Table 5: Results of sub-task B (before and after data augmentation).

excels in recognizing hate speech than the non-hate ones. The observed bias towards recognizing hate speech in the model may stem from the prevalence of HATE-labeled texts in both training and evaluation datasets. As both the training and evaluation datasets are used to train the model, the model may develop a bias, impacting its accuracy when dealing with non-hate speeches.

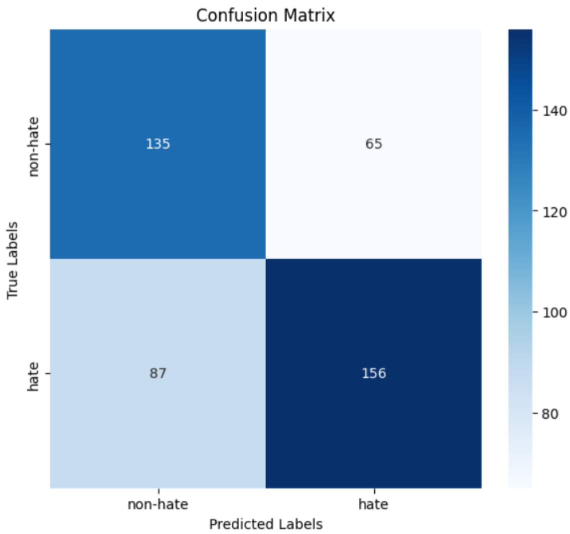

Figure 4: Confusion matrix of sub-task A evaluation set.

In sub-task B, our ensemble model is assigned the challenge of categorizing targets from text-embedded images into three labels: individual (labeled as 0), community (labeled as 1), and organization (labeled as 2). Analysis of the Confusion Matrix shown in Figure 5, indicates that our model shows difficulties in identifying community categories, compared to labeling organizations and individuals. However, the model excels in accu-

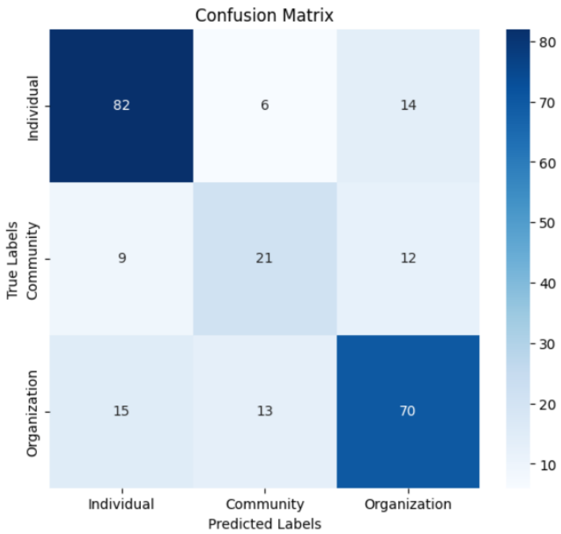

Figure 5: Confusion matrix of sub-task B evaluation set.

rately categorizing individuals. This underscores the significance of having a balanced dataset. The observed challenges in the model's performance, particularly in identifying the community category, can be attributed to an imbalance in the training and evaluation datasets.

According to our initial analysis, there are some challenges that can affect our results. Firstly, there is an imbalance in label distribution within our dataset, where certain data classes contain more Instances than others. This makes it difficult for the model to learn properties of classes that contain fewer examples. Secondly, we observed that some labels in the dataset are correctly attributed. This is the case of many offensive and hate speech datasets due to the intrinsic subjectivity of the task, as noted by Weerasooriya et al. (2023). Incorrect labels can confuse our model, making it harder for it to learn properly and leading to mistakes in the evaluation state. It may also explain why GPT3.5 underperformed, even after finetuning. Also, as this is primarily a text classification task - models like XLM-R do better than GPT3.5.

Finally, another limitation lies in the impact of external factors on the reliability of our Multimodal Hate Event Detection Model over time. The dynamic nature of online discourse and political shifts may affect its efficacy. Even though our models achieve good results, recognizing and dealing with these challenges is important when developing high-performing models that work well in the ever-changing world of online conversations and political events.

7 Conclusion and Future Work

This paper evaluated various approaches to Multimodal Hate Event Detection. We tested multiple models such as GPT, XLM-R, and BERT on subtask a and sub-task b of the competition and we addressed the difficulties associated with handling multimodal content. Our XLM-R model performed well in subtask A ranking third, achieving an F1 score of 0.83. In the same way, for subtask B, our ensemble method, which combined BERT base, BERTweet large, and XLM-R, also ranked third, achieving an F1 score of 0.67.

Despite encountering label distribution imbalances in the training and evaluation sets, our approaches successfully navigated these challenges. Future studies will focus on exploring potential biases in our models and further refining strategies for handling class imbalance as in Akhbardeh et al. (2021). Moreover, as online communication continues to increase multimodality, developing robust hate speech detection systems requires fusing information from different modalities. Future work should focus on faceted annotation schemes and semi-supervised approaches to improve generalization. Evaluating model biases, and exploring the impacts of label imbalance are also important areas needing attention. We hope our experiments provide a valuable starting point for further research towards safer online spaces.

Acknowledgment

We thank the shared task organizers for providing us with this interesting dataset. We further thank the anonymous reviewers for their insightful feedback.

Ethics Statement

This study adheres to the ACL Ethics Policy and seeks to make a contribution to the realm of online safety. The dataset is supplied to us by the organizers and has undergone anonymization to secure the privacy of the users. The technology in question possesses the potential to serve as a beneficial instrument for the moderation of online content, thereby facilitating the creation of safer digital environments. However, it is imperative to exercise caution to prevent its potential misuse for purposes such as monitoring or censorship.

References

Farhad Akhbardeh, Cecilia Ovesdotter Alm, Marcos Zampieri, and Travis Desell. 2021. Handling extreme class imbalance in technical logbook datasets. In *Proceedings of ACL*.

Mario Ezra Aragón, Miguel Angel Alvarez Carmona, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor Pineda, and Daniela Moctezuma. 2019. Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *Proceedings of IberLEF*.

Aashish Bhandari, Siddhant B 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 IEEE CVF*.

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md. Golam Rabiul Alam. 2021. Multi-modal hate speech detection using machine learning. In *Proceedings of IEEE Big Data*.

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Proceedings of TALN*.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of LREC*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings NAACL*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Saurabh Sampatrao Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. 2021. Cross-lingual offensive language identification for low resource languages: The case of Marathi. In *Proceedings of RANLP*.

Dhiman Goswami, Md Nishat Raihan, Sadiya Sayara Chowdhury Puspo, and Marcos Zampieri. 2023. nlpBDpatriots at BLP-2023 task 2: A transfer learning approach towards Bangla sentiment analysis. In *Proceedings of BLP*.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection. In *Proceedings of WASSA*.

Paulo Cezar de Q Hermida and Eulanda M dos Santos. 2023. Detecting hate speech in memes: a review. *Artificial Intelligence Review*, pages 1–19.

Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying creative harmful memes via prompt based approach. In *Proceedings of WWW*.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from bengali memes and texts. In *Proceedings of SPELL*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of AAAI*.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Proceedings of FIRE*.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of SemEval*.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7):34.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of LREC*.

Md Nishat Raihan, Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, and Marcos Zampieri. 2023a. nlpBDpatriots at BLP-2023 task 1: Two-step classification for violence inciting text detection in Bangla - leveraging back-translation and multilinguality. In *Proceedings of BLP*.

Md Nishat Raihan, Umma Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, and Marcos Zampieri. 2023b. Offensive language identification in transliterated and code-mixed bangla. In *Proceedings of BLP*.

Kshitij Rajput, Raghav Kapoor, Kaushal Rai, and Preeti Kaur. 2022. Hate me not: detecting hate inducing memes in code switched languages. In *Proceedings of AMCIS*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual Detection of Offensive Spans. In *Proceedings of NAACL*.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of SocialNLP*.

Surendrabikram Thapa, Farhan 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 CASE*.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during russia-ukraine crisis - shared task at case 2024. In *Proceedings of 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 CASE*.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of EACL*.

Asahi Ushio, Leonardo Neves, Vitor Silva, Francesco. Barbieri, and Jose Camacho-Collados. 2022. Named Entity Recognition in Twitter: A Dataset and Analysis on Short-Term Temporal Shifts. In *Proceedings of AACL*.

Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur Khudabukhsh. 2023. Vicarious offense and noise audit of offensive speech classifiers: unifying human and machine disagreement on what is offensive. In *Proceedings of EMNLP*.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of ACM MM*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023. Target-based offensive language identification. In *Proceedings of ACL*.

Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. 2022. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. In *Proceedings of ICWSM*.