Can The Potential for Offline Harm Events in Social Media Texts Be Identified Without knowing the Context ?

Anonymous submission

Abstract

This paper explores the capability to assess the potential for offline harm events posed by online social media texts without contextual knowledge. Engaging in the TRAC-2024 shared task : Offline Harm Potential Identification (HarmPot-ID), we investigate a vital classification challenge: determining a post's likelihood to incite offline harm, such as protests, clashes or riots and identifying the target groups(s). Our analysis encompasses two main tasks: the prediction of offline harm potential across four categories, ranging from no risk to certainty of causing harm, and identifying the probable target(s) of such harm among five broad groups including gender, religion, caste, descent and political ideology. Leveraging a dataset comprising of 4 languages - English and code-mixed Hindi, Bangla, Manipuri texts from platforms like YouTube, Twitter, Telegram, etc.., we apply and evaluate various computational approaches, including transformer models, Large language models and a keyword-based filtering strategy. Our findings offer insights into the effectiveness of these methods in identifying harmful content without context, suggesting avenues for practical applications in monitoring and mitigating online threats. The paper also discusses the strengths and weaknesses of each of the approaches and methods.

Keywords: Social Media Analytics, Hate Speech Recognition, Text Classification

1. Introduction

With an overwhelming amount of content being shared daily on various social media platforms by numerous users, identifying posts that have the potential to incite offline harm, such as riots, protests, or clashes, becomes a significant challenge. The complexity of this task is compounded by the diversity of the content, ranging from text in multiple languages to coded and code-mixed messages. This paper delves into two specific tasks within this context. The first task is classifying the text based on the likelihood of inciting offline harm events among 4 classes on a scale of 0 to 3. The second task involves binary classification among 5 categories about what group is likely being targeted by the text : gender, religion, descent, caste, political ideology. We have tested various approaches and methods including keyword based filtering, transformers, LLMs towards improving classification using micro f1 score as primary metric for both the subtasks along with recall.

2. Dataset

The dataset (Kumar et al., 2024) used in this study is detailed in three tables, providing an extensive breakdown of the collected social media posts. The texts range from single word to multiple sentences , sometimes just emojis with no text consisting on 4 languages : English(en) , code-mixed and direct versions of Bengali(bn), Meitei(mni), Hindi(hi). Table 1 shows the distribution of labels for the potential offline harm (subtask 1A) across the training, development, and test datasets. The distribution and labels for the test set were not released as of when the paper is being written. The Labels for subtask 1A imply the likelihood of the text inciting offline harm events. 0 indicating the text will never lead to offline harm, in any context. 1 indicating it could lead to an offline harm event given specific conditions or context. 2 indicating it is most likely to initiate an offline harm event in specific contexts. 3 meaning it is certainly going to incite or initiate an offline harm event in any context.

Count split ↓	0	1	2	3	total
Train	16135	21554	12211	888	50788
Dev	2017	2695	1526	111	6349
Test	?	?	?	?	6349

Table 1: Label distribution for subtask 1A

Table 2 presents the counts for the binary labels of each of the 5 classes of who is being targeted (subtask 1B) in each dataset segment. The dataset is very highly imbalanced for class 3 in subtask 1A and all the classes for subtask 1B.

Count→ column ↓	train 0/1	dev 0/1	test 0/1
Gender	41189/9599	5169/1180	?/?
Religion	45912/4876	5704/645	?/?
Descent	49332/1456	6169/180	?/?
Caste	50227/561	6291/58	?/?
Ideology	50381/407	6301/48	?/?
Total	50788	6349	6349

Table 2: Label distribution for subtask 1B

Finally, Table 3 outlines the language distribution across the same dataset divisions. The source of text, context, and the test set labels are unknown.

Split	bn	en	hi	mni	total
Train	12507	12664	14491	11026	50788
Dev	1538	1833	1526	1124	6349
Test	1522	1743	1889	1854	6349

Table 3: Language distribution in the dataset

3. Related Works

Some of the existing works in a similar direction are Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) (Masud et al., 2023) which deals with English tweets by classifying as hate and offensive or not at a span level. The previous editions were (Ranasinghe et al., 2022) and (Mandl et al., 2021) on English, Hindi and Marathi tweets. along with (Mandla et al., 2021) and (Mandl et al., 2019) were on English, Hindi and German. Another similar work is The Offensive Language Identification Dataset (OLID) (Uglow et al., 2019) which consist of just English tweets similar to the current work but classifies then as targeted or not and whether if it targeted towards an individual or a group. Lastly the ComMa (Kumar et al., 2022) which uses data of the same language as the current dataset : English, Hindi, Bengali, Meitei categorizing the aggression level, intensity, discursive role and bias based on gender, religion, etc. , while the current work classifies the text based on who is being targeted by the text. (Kumar et al., 2018) is another similar corpus mostly compromising Hindi and English code-mixed tweets and posts from Facebook.

4. Transformers Approach

For training, some samples which had no text other than URLs and emojis were excluded. We have used base and large variants (except mdeberta) of mDeBERTa (He et al., 2023), electra (Clark et al., 2020) and XLM-R (Conneau et al., 2020) to finetune over the training data with various sets of hyperparameters with and without preprocessing. Some of the results on the dev and test sets can be seen in Table 4 and Table 5. The pre-processing included removal of URLs, emojis, lowercasing, spell checking in various permutations. Since the codalab interface only provided metrics upto two decimal places for the test set, the metrics were rounded to 2 points for the dev set to match the format of the test set metrics.

Due to limitation on total number of submissions, only a few were tested on the test set. For the monolingual versions used to make predictions on the development set, few rows weren't being translated properly and had to be manually translated to test them by translating to English. BART (Lewis et al., 2019) and RoBERTa (Liu et al., 2019) were used on the translated versions but the results weren't satisfactory. Predictions on the test set were made with no further training.

5. LLM Approach

We have used GPT-4 (OpenAI, 2024), LLaMa-2 7B (Touvron and ..., 2023), Mistral 8x7B (Jiang et al., 2023) to make predictions in various ways including zero-shot , few-shot, ranged predictions.

5.1. Zero-shot Predictions

The zero shot predictions were made on both the dev and test sets using the 3 models. This required multiple iterations with slightly varying prompts instructing the model to respond with just a 0 or 1 incase of subtask-1b and a integer only in the range of 0 to 3 for subtask-1a. Though the outputs are supposed to be non-deterministic, while making integer predictions, the outputs were found to be deterministic. In the case of LLaMa and Mistral, chat and mixture of experts versions were used. These required tweaking to re-generate outputs when the outputs generated at first aren't in the required format. The performance of Zero-shot approach can be seen in Table 6.

5.2. Few-shot Predictions

For few-shot predictions 3-shot was chosen, after testing with 2 to 5 shot responses on the dev set. The samples which were misclassified in all or most of the models' predictions from transformer approach were chosen along with the prompt with atleast one of them positive and negative, a justification for each of these classifications was also added to the prompt, Then predictions were made separately for each target label. The results can be seen in Table 7.

5.3. Ranged Zero-shot Predictions

In this approach, the prompt was to predict a float value in the range of 0.00 to 1.00 based on how likely the text is directed towards a particular group incase of subtask-1b. The same was done for subtask-1a at the same range of values of how likely is the text to incite offline harm. The approach used later was similar to (Patkar et al., 2023) where the values are then assigned a threshold based on dev set predictions to classify into each class on the test set. The results in this approach with LLaMa and Mistral were found to be not satisfactory. Due to the non-deterministic nature of the outputs, the float predictions had a standard deviation of 1.6×10^{-3} on testing over 5 runs over the entire development set.

Base Model	pre-processing	1a	1b				
			Gender	Religion	Descent	Caste	Ideology
xlm-roberta-large	with	0.71	0.88	0.93	0.97	**0.99**	0.99
	without	0.70	0.86	0.93	0.97	**0.99**	0.99
electra-large-disc..	with	0.71	0.88	**0.94**	0.97	**0.99**	0.99
	without	0.71	0.87	0.93	0.97	**0.99**	0.99
mdeberta-v3-base	with	**0.72**	**0.89**	**0.94**	**0.98**	**0.99**	**1.00**
	without	0.71	**0.89**	**0.94**	**0.98**	**0.99**	0.99

Table 4: F1 scores using transformers approach: Dev set

Base Model	pre-processing	1a	1b				
			Gender	Religion	Descent	Caste	Ideology
electra-large-disc..	with	0.70	0.72	**0.83**	0.97	**0.98**	**0.99**
mdeberta-v3-base	with	**0.71**	**0.73**	**0.83**	**0.98**	**0.98**	**0.99**

Table 5: F1 scores using transformers approach: Test set

Base Model	1a	1b				
		Gender	Religion	Descent	Caste	Ideology
Dev set						
GPT-4	**0.44**	**0.81**	**0.93**	0.96	**0.98**	**0.97**
LLaMa-2 7B chat	0.37	0.79	0.88	0.94	0.97	0.91
Mistral 8x7B	0.43	**0.81**	0.90	**0.97**	0.97	0.94
Test set						
GPT-4	**0.47**	**0.75**	**0.82**	**0.97**	**0.98**	**0.99**

Table 6: F1 scores for predictions with LLMs using Zero-shot approach : Dev and Test sets

Base Model	1a	1b				
		Gender	Religion	Descent	Caste	Ideology
Dev set						
GPT-4	**0.57**	**0.84**	**0.93**	0.96	**0.98**	**0.98**
LLaMa-2 7B chat	0.41	0.78	0.89	0.94	0.96	0.92
Mistral 8x7B	0.52	0.80	0.90	**0.97**	**0.98**	0.95
Test set						
GPT-4	**0.48**	**0.76**	**0.81**	**0.98**	**0.98**	**0.99**

Table 7: F1 scores for predictions with LLMs using Few-shot approach : Dev and Test sets

Base Model	1a	1b				
		Gender	Religion	Descent	Caste	Ideology
Dev set						
GPT-4	**0.54**	**0.76**	**0.87**	**0.93**	**0.99**	**0.99**
Test set						
GPT-4	**0.56**	**0.76**	**0.83**	**0.96**	**0.98**	**0.99**

Table 8: F1 scores for predictions with LLMs using Ranged predictions approach : Dev and Test sets

6. Keywords based Approach

Another approach tried was to use a list of keywords to filter for classification in subtask-1b where if one of these words is detected, the text is classified as positive. Such lists were created separately for each category in the subtask which consist or derogatory terms and certain terms or phrases used to refer to the target groups in a negative way , while this had a recall which is almost perfect, but the F1 was lower compared to other approaches. These words were added to the lists after observing the positive labelled texts of the train set and used to make predictions on the development set.

Base Model	**1a**	**1b**				
		Gender	**Religion**	**Descent**	**Caste**	**Ideology**
electra-large-disc..	0.70	0.72	**0.83**	0.97	**0.98**	**0.99**
mdeberta-v3-base	**0.71**	0.73	**0.83**	**0.98**	**0.98**	**0.99**
GPT-4 Zero-shot	0.47	0.75	0.82	0.97	**0.98**	**0.99**
GPT-4 3-shot	0.48	**0.76**	0.81	**0.98**	**0.98**	**0.99**
GPT-4 Ranged	0.56	**0.76**	**0.83**	0.96	**0.98**	**0.99**

Table 9: F1 scores for predictions submitted on Test set

7. Error Analysis

Most of the errors originated from very short texts which consist of links and emojis and very short text. Few Texts which were one word like 'Hi' and 'Nice' were labelled as targeted towards Religion or Gender in some instances without the links and tags. Such cases are obviously prone to being misclassified. A possible work around might be replacing URLs with the description/summary of what the link is pointing to so that classification can be improved. The high error rate when translating to English and working with monolingual models was due to data loss during translation and the low-resource availability for Meitei texts which led to errors in translation. However using GPT through API for the purpose of translation did work in those case. However due to the non-deterministic nature and data-loss across translation did cause an uptick in errors when using monolingual models. Another issue is not having information on who is being responded to with the text. Here is an example from the dataset which was labelled as being targeted toward a gender.

> *"Sick and crap mentality!! I don't understand in which kind of world we are living. If this kind of people exist, they are a threat to entire humanity. Uncivilized morons."*

It is tough for any LLM or a human reviewer to understand who the targeted groups are through this text without other information. Another possible extension would be using img-to-text models to append the text with what else was attached in the post/tweet/.. to build a better classifier.

8. Conclusion

Due to the 5 submissions limit , only a few models have been tested on the test set as in Table 9. The mdeberta version was used as the official submission. While the fine-tuned transformer models had the best performance, LLMs had their own advantages, The LLMs had better beformance on texts which were very long probably due to the ability to process longer sequences of text compared to fine-tuned transformer models, likely due to their architecture's capacity to handle larger inputs. Same can be seen in Figure 1 as an example where the accuracy dropped significantly after the lenght of texts crossed the max token limit. An ensemble with other models incase of very long texts might improve the performance .In both cases, the texts which were shorter i.e 1 or 2 sentences long had a high error rate likely due to lack of enough information to classify.

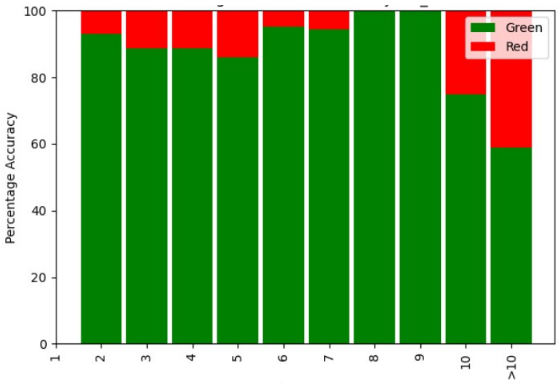

Figure 1: Accuracy vs Sentence Count : subtask-1b3 using mdeberta

A different method of evaluation where class 3 is weighed more then others , followed by 2, 1 and 0 might be a better approach as accurately detecting high risk texts should have more importance than whether or not low risk text were detected. Also due to time and cost limitations all approaches haven't been tested which include using an ensemble of above mentioned approaches where keyword-based filtering resulted in near perfect recall scores. This along with one of the transformers or LLM approaches might yield better results. Despite having no context or information regarding the texts, the results appeared quite good. But, it is very likely that with context and other details even better results can be obtained.

Due to the page limitations, some of the omitted information, more plots, prompts used, hyperparameter space explored, and other information along with the code is added in the Appendix.

9. Bibliographical References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Ritesh Kumar, Ojaswee Bhalla, Madhu Vanthi, Shehlat Maknoon Wani, and Siddharth Singh. 2024. Harmpot: An annotation framework for evaluating offline harm potential of social media text.

Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, Akanksha Bansal, and Atul Kr. Ojha. 2022. The ComMA dataset v0.2: Annotating aggression and bias in multilingual social media discourse. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4149–4161, Marseille, France. European Language Resources Association.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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 pre-training approach.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. New York, NY, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schaefer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages.

Thomas Mandla, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2021. Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages.

Sarah Masud, Mohammad Aflah Khan, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. Overview of the hasoc subtrack at fire 2023: Identification of tokens contributing to explicit hate in english by span detection.

OpenAI. 2024. Gpt-4 technical report.

Aditya Patkar, Suraj Chandrashekhar, and Ram Mohan Rao Kadiyala. 2023. AdityaPatkar at WASSA 2023 empathy, emotion, and personality shared task: RoBERTa-based emotion classification of essays, improving performance on imbalanced data. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada. Association for Computational Linguistics.

Tharindu Ranasinghe, Kai North, Damith Premasiri, and Marcos Zampieri. 2022. Overview of the hasoc subtrack at fire 2022: Offensive language identification in marathi.

Hugo Touvron and ... 2023. Llama 2: Open foundation and fine-tuned chat models.

Harrison Uglow, Martin Zlocha, and Szymon Zmyślony. 2019. An exploration of state-of-the-art methods for offensive language detection.