ABSTRACT

Social media plays an integral role in our daily lives. It enables the rapid sharing of information and opinions. However, some individuals misuse these platforms to express hatred toward specific communities through their posts. Such expressions can harm these communities in various ways, creating false perceptions and subjecting them to mistreatment. Unfortunately, the online spread of such harmful information has led to devastating events like the Rohingya genocide, the Sri Lanka riots, Indian mob violence, and the Pittsburgh shooting, among others. Existing laws that prevent such situations often fail to address them effectively, resulting in limitations on freedom of speech.

In this context, developing AI systems capable of detecting and monitoring these harmful entities on social media platforms becomes essential. These systems should take preemptive actions to mitigate the risks and harm they pose to the society. Moreover, as social media content can take different forms, such as text, images, and videos, it is crucial to identify all forms of harmful content and take appropriate measures to maintain a civil platform.

While several studies have focused on automatically detecting hate speech and offensive language, most research has concentrated on the English language and textual content alone. In addition, it is essential to note that a significant portion of hateful posts originates from a small number of individuals. Therefore, the identification of these hateful users becomes another crucial task.

To address the gaps in existing research, this thesis explores various dimensions of hateful content detection. It aims to extend the focus beyond English and textual modality to encompass low-resource languages and other modalities. By developing effective detection mechanisms and understanding the behavior of hateful users, this

research aims to contribute to a safer and more inclusive online environment.

We first perform a large-scale analysis of low-resource abusive speech detection, considering eight Indic languages and utilizing 14 publicly available resources. We examine different interlingual transfer mechanisms, particularly in the resource-rich to resource-poor transfer direction. We start by analyzing each language individually. We explore various scenarios, including *zero-shot learning*, *few-shot learning*, *instance transfer*, and *synthetic transfer*, to understand how different transfer modes compensate for the lack of gold training instances. Our observations show that *model transfer yields better performance than instance transfer.* Also, the model transfer is particularly advantageous when the source and target languages belong to the same language family. The *AllBOne*[1] model demonstrates the best performance in a few-shot setting as it benefits from both fine-tuning stages. This model works as follows – in the first stage, it learns universal features; in the second stage, it learns language-specific features. For low-resource languages, we find that synthetic silver instances are useful for building classifiers for abusive language detection. However, further improvements are achieved by fine-tuning the model using gold target instances. Besides, we contribute a dataset of 10K Bengali abusive language posts from X (formerly Twitter), aiming to diversify the available resources for research in this field.

Next, we analyze the detection of abusive memes considering English and Hindi (code-mixed) languages. We explore several multi-modal models that utilize text, images, and fusion-based approaches. Like multilingual texts, we apply *interlingual transfer mechanisms* such as ELFI[2], zero-shot, and few-shot transfer. As expected, the multi-modal model outperforms using both textual and visual modalities. However, we observe that the *zero-shot cross-lingual performance of these models is not*

[1]Use the dataset of all languages except the target language for training and validation, and test on the target language.

[2]Same language for training, validation, and testing.

at par with ELFI-style training. Nevertheless, we discover significant benefits by fine-tuning existing models in another language, even with limited target-language meme data. Futher, we develop a dataset named BANGLAABUSEMEME to aid in the automatic detection of abusive memes in Bengali. This dataset comprises 4,043 memes. Subsequently, we investigate the detection of hateful videos. To accomplish this, we crawl videos from the BitChute platform and manually annotate them as hate or non-hate. Analyzing the annotated dataset, HATEMM uncovers interesting aspects of hate videos. We utilize all video modalities to determine whether a video is hateful. Our findings demonstrate that *models considering multiple modalities outperform uni-modal variants*.

Finally, we focus on detecting hateful users. We thoroughly explore the problem space and investigate various models, including purely textual, graph-based, and semi-supervised techniques using Graph Neural Networks (GNN) that leverage both textual and graph-based features. We conduct extensive experiments on two datasets: Gab, which has loose moderation, and X (formerly Twitter), which has strict moderation. Overall, the *AGNN model achieves a macro F1-score of 0.79 on the Gab dataset and 0.78 on the X (formerly Twitter) dataset, using only 5% of the labeled instances*. This performance surpasses all other models, including the fully supervised ones. We conduct a detailed error analysis on the best-performing text and graph-based models. We observe that hateful users exhibit unique network neighborhood signatures, and the AGNN model benefits from attending to these signatures. This property allows the model to generalize well across domains, even in a zero-shot setting. Lastly, we utilize the top-performing GNN model to analyze the evolution of hateful users and their targets over time on Gab. By leveraging this model, we gain insights into the patterns and changes in behavior exhibited by hateful users on the platform.

Summarizing, in this thesis, we investigate how we can enhance the detection of

low-resource multilingual abusive speech on social media. We also explore methods to detect multi-modal hateful/abusive content, specifically memes and videos. Last but not the least, we focus on detecting the users who spread such hateful posts. All the resources (datasets, models, lexicons, etc.) are available on GitHub[3] for the research community to use.

Keywords: hate speech; abusive speech; online social media; X (formerly Twitter); Gab; free speech; multilingual; multi-modal; hateful users.

[3]https://github.com/hate-alert