Name: Ovais Bashir Gashroo **Notification No: 578/2025**

Name of the Supervisor: Prof. Monica Mehrotra **Notification Date: 08-05-2025**

Topic: Modelling Abusive Communities in Online Social Networks (OSN's)

Department / Faculty: Computer Science, Faculty of Sciences

FINDINGS

This research provides a comprehensive investigation into the detection and classification of abusive content on online social networks (OSNs); addressing one of the most pressing challenges in digital communication platforms today. With the rapid proliferation of internet technologies and the increasing engagement of users from diverse backgrounds, ensuring a safer and more inclusive online environment has become essential. The findings of this study reveal several critical insights into the nature of online abuse, the effectiveness of various detection methods, and the role of advanced machine learning and data balancing techniques.

A key finding of this thesis is the limitation of traditional binary classification techniques in effectively capturing the complexity of abusive content. To address this, the research proposed and implemented a multi-classification framework, enabling the detection and categorization of nuanced forms of abuse rather than simply labeling content as abusive or non-abusive. This approach allows for the classification of different abuse types, such as those targeting specific races, religions, nationalities, sexual orientations, and personal identities. The results demonstrate that multi-class classification is significantly more effective in identifying context-specific and subtle abusive content, which often gets overlooked by binary classifiers.

Another significant contribution is the development of a novel framework which integrates a fine-tuned transformer model to detect multiple forms of abuse in textual content. The framework leverages the DistilBERT model, a lightweight and computationally efficient version of BERT, fine-tuned specifically for the task of abusive content classification. The findings indicate that DistilBERT, due to its reduced complexity and faster performance, is highly suitable for real-time abuse detection scenarios while maintaining strong classification performance.

In the course of experimentation, the study explored the impact of feature extraction techniques and various machine learning and deep learning models, including traditional classifiers and neural networks. The performance of these models was evaluated using benchmark datasets, highlighting the importance of high-quality feature representations in boosting classification accuracy. The research found that pre-trained word embeddings—such as

contextual embeddings derived from transformer-based models—greatly enhance the models' ability to understand semantic context, thus improving detection precision.

An important aspect of this work is its detailed analysis of data imbalance, which is a major challenge in abusive content detection due to the disproportionate distribution of abusive versus non-abusive samples. The study evaluated the impact of data balancing techniques, such as Synthetic Minority Oversampling Technique (SMOTE) and contextual data augmentation using transformer-based embeddings. The findings suggest that data augmentation not only mitigates the effects of class imbalance but also improves generalization across different forms of abuse.

The research also extends its findings to low-resource languages, proposing a novel method that considers the linguistic and cultural nuances of such languages. This highlights the adaptability of the proposed framework across multilingual platforms and its applicability in regions with less-studied languages. The study emphasizes the need for language-specific approaches in abusive content detection, especially in global social networks that cater to diverse linguistic communities.

Finally, the study underscores the importance of data quality in abusive content detection. Through extensive experimentation, it was shown that high-quality, well-annotated, and diverse datasets are crucial in training robust and reliable models. The research demonstrates that datasets need to capture the subtle, contextual, and evolving nature of online abuse to ensure effective classification and minimize bias or false positives.

In summary, this thesis delivers a multifaceted contribution to the field of online safety by developing efficient, adaptable, and context-aware frameworks for abusive content detection. The findings validate the effectiveness of multi-class classification, the utility of transformer-based models, the necessity of balanced and high-quality data, and the importance of accommodating linguistic diversity. These insights can significantly aid service providers in building safer communication environments, ultimately fostering trust and harmony across user communities.