e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

A STUDY ON UNSECURE CONVERSATIONS DETECTION INSOCIAL MEDIA PLATFORMS

Sanika H. Lokare[*1], Sai G. Mane[*2], Jyotirmoyee Rout[*3], Vaishnavi M. Chikhale[*4]

[*1,2,3,4]Computer Engineering Department Sinhgad College Of Engineering, Pune, India.

DOI : https://www.doi.org/10.56726/IRJMETS50357

ABSTRACT

This survey synthesizes key findings from research papers on online safety, with a focus on Instagram. Topics include risk detection in private conversations amid end-to-end encryption, identification of fake profiles, analysis of hacking behaviors, machine learning for hate speech detection, and the role of impersonators in generating fake content. Collectively, these studies offer insights for developing a project on the detection of harmful conversations, emphasizing the integration of metadata, linguistic cues, and machine learning algorithms in addressing evolving encryption challenges and emerging online risks.

Keywords: Social Media, Hate Speech Detection, Natural Language Processing (NLP), Text Analysis, Text Classification, Machine Learning, Naive Bayes Algorithm, Online Safety, SocialMedia Communities.

I. INTRODUCTION

Social media platforms has recently come under scrutiny for potentially being harmful to the safety and well-being of our younger generations. This paper focuses on the detection of unsafe youth conversations on social media, utilizing machine learning algorithms for risk detection. With the imminent switch to end-to-end encryption for private conversations on social media chatting applications, the project aims to address the challenges posed by limited data availability. The paper explores various feature sets, including metadata, linguistic cues, and image features. Overall, we found that the metadata features, such as conversation length, were the best predictors of risky conversations.[1] The study also emphasizes the importance of considering lived risk experiences of youth in developing robust solutions for risk detection in the contextof end-to-end encryption. Building on the broader landscape of social media risks, the project draws insights from related research on fake profiles and social media crime. This includes a survey of work dedicated to identifying fake profiles across popular social media platforms, emphasizing the rising concerns associated with misleading details and fraudulent activities perpetrated by such profiles.[2] To enhance the machine learning component of the project, a state-of-the-art review is conducted on hate speech detection, focusingon applications of machine learning algorithms. The review identifies Long-Short Term Memory, random forest, and convolution neural network as highly useful algorithms for detecting hate speech across various social platforms, including Twitter and Facebook.[5] This project amalgamates insights from multiple research domains to develop a comprehensive approach for the detection of unsafe youth conversations on social media, incorporating machine learning techniques for addressing challenges posed by evolving privacy measures and emerging online risks.

II. LITERATURE SURVEY

A. Getting Meta: A Multimodal Approach for Detecting Un- safe Conversations within Instagram Direct Messages of Youth

In this paper, the authors address the growing concernsabout the potential harms to the safety and well-being of younger generations on Instagram, particularly as the platform transitions to end-to-end encryption for private conversations. The study involves the collection of data from 172 youth aged 13-21, who identified private message conversations that made them feel uncomfortable or unsafe. The dataset com- prised 28,725 conversations with 4,181,970 direct messages, including text and images. The research aims to investigate which indicators are most effective in automatically detecting risks in Instagram private conversations, focusing on high- level metadata that will still be available under end-to-end encryption. The key components of an Instagram conversationare outlined, including metadata (recipient(s), conversationlength), linguistic cues (text messages), and image features (images shared). The study employs a multimodal approach, combining visual and textual content to predict the safety of a conversation. Metadata features, such as conversation

length and participant engagement, emerge as strong predictors of risky conversations. Linguistic cues and image features play a crucial role in distinguishing between different types of risks within conversations. The authors employ a two-step pipeline system for risk detection, involving binary classification to determine if a conversation is safe or unsafe, followed by multi-class classification to identify specific risk types. Meta- data features are crucial for understanding user engagement in conversations, especially in the private realm, where unsafe conversations tend to be shorter. The study employs machine learning classifiers, including Decision Trees, Random Forest,and Linear SVM, to analyze metadata features. Additionally, linguistic cues are assessed using a Convolutional Neural Network (CNN), while image features are analyzed using image classifiers and Optical Character Recognition (OCR). The paper introduces an ensemble classifier that combines the strengths of metadata, linguistic cues, and image features to provide a holistic risk detection system. Three ensemble models majority-vote, average-vote, and weighted-vote areexplored, each contributing to the overall prediction of con- versation safety. The authors extend their analysis to multi- class classification, identifying the major risk types within conversations based on participant annotations. The top five risk categories include harassment, sexual messages/ solicita- tion, nudity/porn, hate speech, and sale or promotion of illegal activities. Overall, the study contributes to the literature on adolescent online safety by proposing a robust multimodal approach for risk detection in private Instagram conversations, even in the context of end-to-end encryption. The findings offer valuable insights into the design implications for AI risk detection systems and emphasize the importance of consider- ing the lived risk experiences of youth in developing effective safety measures on social media platforms.[1]

B. Malicious Profile Detection on Social Media

The pervasive use of popular social media platforms like Facebook, Twitter, Instagram, and LinkedIn has led to anincrease in social media-related crimes, particularly through the creation of fake profiles. These fake profiles disseminate misleading information, attempt fraudulent activities, and poserisks to users' privacy. Researchers have extensively explored solutions to identify and address the issue of fake profiles across various social media applications. This paper aims to provide a comprehensive overview of research efforts dedicated to fake profile identification and offers a survey of work assessing the authenticity of user profiles on social media platforms. The introduction highlights the dual nature of social media, serving as a means for communication but also harboring malicious user accounts. The focus is on supervised and unsupervised machine learning algorithms employed for fake profile detection, with an emphasis on selecting the target profile for review and extracting relevant features such as user history, posts, and interactions. The paper references 19 research papers to illustrate the diversity of methodologies and datasets used in this domain. The subsequent sections delve into the features considered for fake profile detection, including user-based features, content-based features, time- zone-based features, and graph-based features. The importance of feature selection and dataset choice is underscored, with Twitter and Facebook datasets being prominent in the research landscape. The paper provides an insightful discussion on the significance of different feature types in identifying fake profiles. The related work section reviews existing literature on malicious account identification, covering approaches such as analytical ranking systems, sentiment analysis, graph-based learning algorithms, and methods specific to platforms like Facebook, Twitter, and LinkedIn. A diverse range of tech- niques, including support vector machines, random forests, neural networks, and clustering algorithms, are explored acrossvarious studies to enhance accuracy in fake profile detec-tion. The system overview outlines the four key stages: data collection, data pre-processing, application of classification algorithms, and evaluation of the model's performance. The supervised and unsupervised algorithms discussed include support vector machines, naive bayes, random forests, neural networks, decision trees, and k-nearest neighbors. Feature selection methods, hybrid approaches, and the use of confusion matrices for evaluation are highlighted. Results from different studies showcase successful detection of various maliciousprofiles, such as bot users, duplicate accounts, spam profiles, and false identities. The accuracy of the models is calculated using evaluation criteria like confusion matrices, precision, F1-score, and recall. A comparative analysis of research outcomesis presented in a tabular format. In conclusion, the paper emphasizes the critical role of social media in communication and networking but acknowledges the associated risks posed by fake profiles. The surveyed research contributes valuable insights into the methodologies and algorithms employed to detect malicious profiles. The study highlights the importance of feature selection, dataset diversity, and the ongoing needfor

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

innovative approaches to safeguard social media platformsfrom the proliferation of fake identities.[2]

C. Detection of Hacking Behaviors and Communication Pat- terns on Social Media

This paper addresses the detection of hacking behaviors within online communities by proposing a set of indicators thatanalyze communication patterns on social media platforms. The indicators encompass technical discussions, sentiment expression, threats, recruitment activities, and user profiling. The study evaluates these indicators using Twitter data and reveals significant variations across different types of hackers,supporting the hypothesis that detection of hacking behaviors should consider differences in intentions, motivations, andskills. The proposed indicators cover diverse aspects of hacker communication patterns, providing a comprehensive approachto identify and categorize hacking behaviors. The paper in- troduces indicators focused on technical communication, an- alyzing language used in cybercrime forums. Topics such as attack types, stages, tools, and platforms are considered, usingspecific terms to identify participants with expertise or interestin these areas. Sentiment-based indicators aim to identify motivations, ideologies, and revenge, using sentiment analysistools to detect negative and positive opinions. Social media communication indicators focus on threats, recruitment, and coordination activities, leveraging natural language processingtools to extract relevant information. User profiling indicators analyze characteristics indicative of hacking behaviors, such as reputation focus and demographic features. The analysis of technical communication indicators reveals distinctions among hacker types, with higher-skilled hackers using technical top- ics more frequently. Social media communication indicators show variations in recruitment and threat language across different motivations, with profit and ideology-driven hackers expressing recruitment more often. User profiling indicators, examining hacker names and group hashtags, prove useful for predicting attacks planned by hacktivists. Sentiment indicators highlight differences in negative sentiment expression towardstarget organizations, with insiders exhibiting comparable neg- ativity levels to other hacker types. The proposed indicators offer a comprehensive framework for detecting hacking be- haviors on social media platforms, taking into account the diverse motivations, skills, and intentions of different types of hackers. The evaluation using Twitter data demonstrates the effectiveness of these indicators in capturing variations across various hacker categories. This research contributes to enhanc- ing the understanding and identification of hacking behaviors, providing valuable insights for cybersecurity efforts.[3]

D. Cyberbullying Detection on Social Networks Using Ma- chine Learning Approaches

The rapid expansion of social media in the 21st century has positioned it as a dominant force in global connectivity. However, this surge in social interaction has been accompa- nied by detrimental phenomena such as cyberbullying, online abuse, and harassment. These issues, particularly affectingwomen and children, have led to severe mental and physical distress, emphasizing the urgent need for effective detection and mitigation strategies. This research addresses the impera- tive task of identifying online abusive and bullying messages by proposing a methodology that combines natural language processing (NLP) and machine learning (ML). Focusing on major social media platforms like Facebook and Twitter, the study utilizes two key features, Bag-of-Words (BoW) and termfrequency-inverse text frequency (TF IDF), to evaluate the accuracy of four prominent ML algorithms. The cyberbullying detection framework comprises two fundamental components: Natural Language Processing (NLP) and Machine Learning (ML). In the NLP phase, datasets containing bullying textsare systematically collected and processed for ML algorithms.This involves cleaning the data by removing irrelevant char- acters, such as stop-words, punctuation, and numbers, fol- lowed by tokenization and stemming. The processed data is then transformed into crucial features: Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Moving to the ML phase, the study employs diverse ML approaches, including Decision Tree (DT), Random Forest, Support Vector Machine (SVM), and Naive Bayes, to detect harassing or bullying messages effectively. The classifier withthe highest accuracy is identified for a specific public cyber- bullying dataset. The exploration of common ML algorithms for detecting cyberbullying delves into the Decision Tree, a versatile classifier for both classification and regression. Naive Bayes, an efficient algorithm grounded in Bayes' theorem,is discussed for its swift resolution of binary and multi-class classification problems. The Random Forest classifier, consisting of multiple decision tree classifiers, stands out for its accuracy through majority voting. Lastly, the Support Vector Machine (SVM), applicable to both classification and regression, excels in distinguishing classes in an n-dimensionalspace. In conclusion, this research contributes to the ongoing efforts to combat cyberbullying on social media by proposing an innovative

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

detection framework. The integration of NLP and ML techniques, alongside the exploration of diverse ML algorithms, aims to enhance the accuracy and efficiency of identifying online abusive and bullying messages. This holisticapproach seeks to foster a safer and healthier online environment.[4]

III. RELATED WORK

In our survey paper, we extensively reviewed existing re- search on the detection of unsafe and harmful conversations onsocial media platforms, focusing on the identification of onlinerisks encountered by youth. We draw inspiration and insights from various studies that have explored the prevalence of online risks, emphasizing the need for automated approaches to detect and mitigate such risks. Our work is particularly motivated by the gaps identified in current automated approaches for risk detection, as discussed in the literature.

We analyzed various algorithms from four different researchpapers are as follows:

A. Term Frequency/Inverse Document Frequency (TF-IDF)

One prominent algorithm extensively explored in the realm of information retrieval and text mining is the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF serves as a numerical metric that gauges the significance of a term within either a single document or a broader collection ofdocuments. This algorithm has found widespread application in diverse domains, including information retrieval, searchengines, text mining, and document similarity measurement.

The TF-IDF approach comprises two fundamental components: Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency quantifies the occurrence frequency of a particular term within a given document. Computed as the ratio of the number of times a term appears in a documentto the total number of terms in that document, TF provides a local measure of term importance.

On the other hand, Inverse Document Frequency assesses the importance of a term across an entire collection of documents. Calculated as the logarithm of the ratio of the total number of documents to the number of documents containing the term, IDF offers a global perspective on term significance. The TF-IDF score for a term in a specific document is then determined by multiplying its Term Frequency (TF) and In- verse Document Frequency (IDF), providing a comprehensiveevaluation of its relevance within the document.

In practical applications, it is customary to normalize TF- IDF scores to mitigate biases toward longer documents. One prevalent normalization technique involves dividing the TF- IDF score by the Euclidean norm of the vector of TF-IDF scores for all terms in the document. This normalization step enhances the robustness of the TF-IDF algorithm across documents of varying lengths, making it a versatile and widely adopted tool in the field of information retrieval and textanalysis.

B. Bag-of-Words (BoW)

In the realm of natural language processing, the initial step often involves transforming raw textual data into a formatsuitable for machine learning algorithms. This necessity arisesfrom the algorithms' inability to directly operate on raw text. One common method for this conversion is the Bag-of-Words (BoW) model, widely employed in applications such as text classification, information retrieval, and sentiment analysis.

Tokenization is a crucial preprocessing step within the BoW framework, breaking down text into individual words or tokens. The resulting vocabulary comprises a unique set of words across the entire corpus. The subsequent representation takes the form of a Document-Term Matrix (DTM), where each document becomes a row, and the columns correspondto words in the vocabulary, with entries indicating word frequencies or presence (binary).

To refine this representation, Term Frequency (TF) is employed, counting the occurrences of a term within a document. This count is often normalized by the document's total word count. Simultaneously, Inverse Document Frequency (IDF)measures the importance of a term across the entire document collection, assigning lower IDF values to words appearingfrequently in many documents. The combination of TF and IDF yields the TF-IDF score, highlighting a term's importancein a specific document relative to the entire corpus.

Sparse Representation is a common outcome of BoW, resulting in a matrix with numerous zero entries due to the extensive vocabulary. To capture contextual information, N- grams are introduced, considering sequences of adjacent wordsrather than individual ones.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

The preprocessing pipeline often involves the removal of Stop Words, such as common conjunctions and articles, to reduce noise in the analysis. Despite its utility, the BoW model has limitations, as it ignores word order and semantics, making it sensitive to variations in writing style. These considerations underscore the ongoing challenges in text representation for effective natural language processing applications.

C. Decision Tree

Decision trees are widely employed as a tree-like model for both classification and regression tasks, finding applications across diverse domains. The versatility of decision trees makes them an invaluable tool for data-driven decision-making in various fields.

The structure of a decision tree consists of distinct components, starting with the root node, which represents the feature that best splits the data at that point. Internal nodes follow, testing specific features or attributes, and branches connect these nodes, indicating the outcome of the corresponding decision or test. Terminal nodes, also known as leaf nodes, mark the culmination of the tree, signifying the final decision or prediction.

At each internal node, a decision is made based on a particular feature or attribute, leading to the division of the data into subsets according to the associated decision rule. The determination of decision rules is critical and is often guided by optimization criteria such as Gini impurity for classification tasks or mean squared error for regression.

To avoid overfitting, the process of constructing a decision tree continues until predefined stopping criteria are met. These criteria may include reaching a maximum depth for the tree, ensuring a minimum number of samples in a leaf node, or halting the process when further improvement in the impurity measure is not achieved.

Pruning, a technique that involves removing branches from the tree, is commonly employed to enhance the generalization performance of decision trees. By eliminating unnecessary branches, pruning helps prevent overfitting and ensures that the model generalizes well to unseen data.

Ensemble methods, such as Random Forests and Gradient Boosted Trees, have emerged as powerful extensions of deci- sion trees. These methods enhance performance and general- ization by combining multiple decision trees. Random Forests build multiple trees and aggregate their predictions, while Gradient Boosted Trees sequentially build trees to correct the errors of the previous ones, resulting in robust and accurate models.

D. Random Forest

Random Forest is an ensemble learning method that is widely applied in various domains, particularly in classification and regression tasks. One of its key strengths lies in its effective handling of large datasets. The composition of a Random Forest involves constructing a collection of decision trees, each built independently and operating in parallel.

The ensemble leverages randomization techniques during both the feature selection and data sampling processes. For feature selection, each tree in the Random Forest is constructed using a random subset of features. Additionally, bootstrap sampling is employed, where subsets of data are randomly selected with replacement, contributing to the diversity of the individual decision trees within the ensemble.

Decision aggregation is a crucial step in the Random Forest methodology. In classification tasks, the ensemble utilizes a majority voting mechanism, where the most frequent class among the trees is chosen as the final prediction. In regression tasks, the predictions from individual trees are averaged to obtain the final output.

The training process of a Random Forest involves building each decision tree on a different subset of the dataset. This sub- set is randomly sampled from the original data, contributing to the diversity of the trees within the ensemble. Furthermore, the random selection of features for each split in a tree adds an additional layer of variability, enhancing the overall robustness and generalization capabilities of the Random Forest model.

E. Linear Support Vector Machine (SVM)

The Linear Support Vector Machine (SVM) stands out as a pivotal supervised learning algorithm employed in both classification and regression tasks. Its versatile applications include text and image classification, as well as bioinformatics, among others. The primary objective of SVM is to identify the optimal hyperplane that effectively separates data into distinct classes within the feature space.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

The hyperplane, serving as the decision boundary, is strategically positioned to maximize the margin between classes, ensuring robust classification. This approach hinges on supportvectors, which are data points situated closest to the hyper- plane. These support vectors play a critical role in defining the margin and, consequently, contribute significantly to the algorithm's overall effectiveness.

To quantify the performance of SVM, a hinge loss function is commonly employed. This loss function penalizes misclassified points, reinforcing the algorithm's ability to discernbetween classes accurately. The tuning of the C parameter in SVM is crucial, as it controls the trade-off between achieving a smooth decision boundary and accurately classifying training points. This parameter plays a pivotal role in optimizing the algorithm's performance based on the specific characteristics of the dataset at hand.

F. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) stand out as powerful deep learning algorithms tailored for image and spatial data processing. Their versatility extends to applications such as image classification, object detection, and facial recognition,among others. The architecture of CNNs comprises distinct layers, each serving a specific purpose.

The Convolutional Layers within CNNs utilize convolutional filters to identify and extract patterns and features from input data. This approach enables hierarchical feature learning, allowing the network to discern increasingly complexstructures. Pooling Layers play a crucial role in down sampling feature maps, effectively reducing spatial dimensions. Common types of pooling include max pooling and average pooling.

Activation Functions, often implemented using Rectified Linear Units (ReLU), introduce non-linearity to the model, enhancing its capacity to learn intricate patterns. Fully Connected Layers, akin to traditional neural network layers, are employedfor tasks like classification or regression. These layers flattenand connect high-level features to facilitate decision-making.Parameters such as Strides in convolution determine the step size during the operation, influencing the size of the output feature maps. Padding, the addition of zeros to input data,is employed to maintain spatial dimensions post-convolution. Filters or Kernels, small matrices used for convolution, capture specific patterns such as edges, textures, or shapes. The resulting Activation Maps visualize where particular features are detected, offering insights into the network's decision-making process.

CNNs excel in processing visual and spatial information, making them invaluable for various applications in computer vision. The interplay of convolutional, pooling, and fully connected layers, along with activation functions and other architectural elements, contributes to their success in extracting meaningful representations from complex data.

G. Na¨ıve Bayes

Naive Bayes, a probabilistic algorithm grounded in Bayes' theorem, finds application in diverse domains such as text classification, sentiment analysis, and document categorization, as well as spam filtering. The algorithm leverages Bayes' Theorem as a probabilistic framework to calculate the likelihood of a hypothesis given observed evidence.

A key feature of Naive Bayes is its independence assumption, which posits that features are conditionally independent given the class. This assumption simplifies calculations and model training, making it particularly efficient for various applications. The algorithm comes in different variants, including Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes, each tailored to specific data characteristics.

Likelihood estimation is a crucial aspect of Naive Bayes, involving the estimation of probabilities of feature values given the class. Various methods are employed for different Naive Bayes variants to achieve accurate probability estimates. Prior probability, representing the likelihood of a class occurring before considering features, reflects the initial belief about the class distribution.

Posterior probability, on the other hand, signifies the probability of a class given the observed features, calculated throughBayes' theorem. To handle the challenge of zero probabilities,a technique called smoothing is employed, involving the addition of a small constant to all probabilities. Text classification emerges as a prominent application of Naive Bayes, particularly in spam filtering and document categorization. Despite its widespread use, the algorithm has limitations, notably the assumption of feature independence, which may not always hold in practical scenarios. As a probabilistic and versatile classification tool, Naive Bayes continues to be a subject of interest in the realm of machine learning and data analysis.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

IV. FUTURE SCOPE

A. Enhanced Multi modal Detection

Build upon findings that emphasize combining metadata, linguistic cues, and image features for risk detection. Explore more sophisticated techniques and additional modalities to improve the overall accuracy of identifying unsafe conversations.

B. Adaptation to Evolving Behaviours

Acknowledge the possibility of malicious users altering behaviours to bypass detection systems. Develop adaptable models that can recognize new patterns and tactics employed by predators or scammers in online conversations.

C. Proactive Risk Detection

Move towards more proactive risk detection approaches. Develop systems that can anticipate potential risks before theyescalate, providing a more preventive and real-time safeguarding mechanism for users.

D. Handling Unbalanced Datasets

Address the challenge of dealing with unbalanced datasets. Explore strategies to maintain overall accuracy while minimizing false positives for safe conversations, potentially incorporating human moderation for final decision-making.

E. Cross-Platform Applicability

Extend research beyond specific platforms to evaluate the effectiveness of detection models across various social media platforms. Ensure versatility and generalizability of developedsystems.

F. User-Centric Approach

Emphasize the importance of user experience in risk detection systems. Consider user feedback, minimize false positives, and strike a balance between strict classification and not missing critical risky conversations, particularly when the target audience is youth and young adults.

G. Integration with Social Media Companies

Collaborate with social media companies to incorporate additional metadata characteristics and enhance risk mitigation measures. This collaboration can contribute to the developmentof more robust and comprehensive risk detection systems.

H. Continuous Improvement

Recognize that the landscape of social media and online interactions is dynamic. Focus on continuous improvement of detection models, incorporating feedback loops, and adapting to the evolving nature of online communication and potential risks.

By exploring these future avenues, researchers can con- tribute to the development of more sophisticated, adaptive,and user-friendly systems for detecting unsafe conversations on social media platforms.

V. CONCLUSION

In conclusion, the surveyed papers collectively highlight the growing concerns and challenges associated with online safety, especially on popular social media platforms. The first paper emphasizes the need for robust risk detection mechanisms, particularly in the context of end-to-end encryption. It underscores the importance of metadata features in predicting risky conversations and offers valuable insights into the design implications for AI systems in the presence of such encryption.

The second and third papers address issues of fake profiles and hacking behaviors, showcasing the multifaceted natureof risks in the digital space. The fourth paper provides a comprehensive review of machine learning methods for hate speech detection, emphasizing the effectiveness of specific algorithms across different languages and platforms. Lastly, the fifth paper sheds light on the role of impersonators in gen- erating fake content on Instagram, presenting a deep learning approach to distinguish between bot-generated, fan-generated,and genuine content. Moving forward, the integration of insights from these diverse studies can pave the way for the development of a comprehensive and adaptive project focusedon the detection of harmful and unsafe conversations in the dynamic landscape of online interactions. Such a project could leverage a combination of metadata analysis, linguistic cues and machine learning algorithms to enhance the effectiveness

of risk detection systems, thereby contributing to a safer digitalenvironment for users, especially the younger demographic.

VI. REFERENCES

[1] Shiza Ali, Afsaneh Razi, Seunghyun, Ashwaq Alsoubai, Chen Ling, Munmun De Choudhury, Pamela J. Wisniewski, Gianluca, "Getting Meta: A Multimodal Approach For Detecting Unsafe Conversations Within Instagram Direct Messages Of Youth", Article In Proceedings Of The ACM On Human-Computer Interaction · April 2023.

[2] Shruti Shinde, Dr. Sunil B. Mane "Malicious Profile Detection On Social Media: A Survey Paper", 2021 9th International Conference On Reliability, Infocom Technologies And Optimization (Trends And FutureDirections)(ICRITO).

[3] Steve Hinton, Taissa Gladkova , David Maimon, Olga Babko-Malaya , Rebecca Cathey ,"Detection Of Hacking Behaviours And Communication Patters On Social Media", 2017 IEEE International Conference On Big Data.

[4] Md Manowarul Islam , Md Ashraf Uddin, Linta Islam, Arnisha Akter, Selina Sharmin, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches", 2020 IEEE Asia-Pacific Conference onComputer Science and Data Engineering (CSDE).

[5] Koosha Zarie, Reza Farahbakhsh, "Impersonation On Social Media: A Deep Neural Approach To Identify Ingenuine Content", 2020 IEEE/ACM International Conference On Advances In Social Network Analysis And Mining.

[6] Gergo Hajdu, Yaclaudes Minoso, Rafael Lopez, Miguel Acosta, Abdel- rahman Elleithy,"Use Of Artificial Neural Networks To Identify Fake Profiles", Department Of Computer Science William Paterson UniversityWayne, Nj, USA.