**[1] V. U. Gongane, M. V. Munot, and A. D. Anuse, “Detection and moderation of detrimental content on social media platforms: current status and future directions,” Social Network Analysis and Mining, vol. 12, no. 1, p. 129, 2022.**

This research paper explores the current status and future directions in the detection and moderation of detrimental content on social media platforms. Social media has become an integral part of society, providing a platform for virtual communication and expression of views. However, the dark side of social media includes detrimental content such as fake news, hate speech, cyberbullying, and rumors, impacting individuals' mental health and causing irreversible losses.

Major social media platforms have invested in manual and semi-automated moderation methods, but the increasing volume of detrimental content necessitates fully automated solutions. Artificial Intelligence (AI), particularly Natural Language Processing (NLP) with Machine Learning (ML) and Deep Neural Networks, is crucial for effective detection and moderation. The paper emphasizes the importance of addressing both detection and moderation, highlighting the limitations of current methods and proposing AI-based solutions to bridge the gaps. The three-part study covers methods for detecting detrimental components using NLP, moderating such content, and provides insights into research gaps and future directions. The research underscores the significance of AI in mitigating the negative impact of detrimental content on social media platforms.

**[2] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian and A. Alothaim, "Online Extremism Detection in Textual Content: A Systematic Literature Review," in IEEE Access, vol. 9, pp. 42384-42396, 2021, doi: 10.1109/ACCESS.2021.3064178.**

The research paper addresses the growing concern of extremist groups using social media platforms for dissemination and recruitment. Authored by Saja Aldera, Ahmad Emam, Muhammad Al-Qurishi, Majed Alrubaiyan, and Abdulrahman AlOthaim, the study focuses on understanding the landscape of online extremism through a systematic literature review of 45 studies published between 2015 and 2020.

The authors identify key research contributions, including a systematic literature review, taxonomy development, identification of challenges and opportunities, and a guideline for selecting techniques in online extremism detection. The work underscores the importance of understanding the dynamics of online extremism and offers a critical analysis of the existing research landscape. The systematic review aims to guide researchers, counter-terrorism agencies, and technology companies in advancing methods to detect and counter online extremism effectively.

**[3] A. Geiger, D. Liu, S. Alnegheimish, A. Cuesta-Infante and K. Veeramachaneni, "TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks," 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 33-43, doi: 10.1109/BigData50022.2020.9378139.**

The focus was on differentiating time series anomaly detection from traditional machine learning approaches. It was noted that time-series analysis is particularly useful for real-time systems, such as real-time phishing tweet detection. TadGAN, a method combining GANs and LSTM Recurrent Neural Networks, was introduced to tackle anomaly detection in time series data by training GANs with cycle consistency loss, which computes anomaly scores based on reconstruction errors and critic outputs.

The paper compared deep learning-based methods like LSTM AutoEncoder and MAD-GAN to traditional statistical models like ARIMA and Holt-Winters, emphasizing that traditional models require strong assumptions, while deep learning methods capture complex temporal correlations more effectively. TadGAN demonstrated superior performance in most cases, offering high accuracy and generalizability.

The research contributions include the introduction of TadGAN, evaluation on 11 datasets, and insights into using GANs for anomaly detection. Although the relevance of TadGAN for real-time Twitter data was highlighted, the paper lacked specific details about the Twitter dataset used. It employed the "Tweets" dataset from the Numenta Anomaly Benchmark (NAB) for its evaluation. Time series analysis was deemed more suitable for real-time Twitter data to capture temporal trends and identify abrupt shifts in Twitter activity. The analysis suggested that traditional machine learning approaches may be less effective in handling temporal data, potentially overlooking valuable insights in Twitter data. In some cases, combining machine learning and time series analysis with sentiment analysis could provide a more comprehensive approach for detecting sentiment trends in real-time data.

**[4] L. Safarnejad, Q. Xu, Y. Ge and S. Chen, "A Multiple Feature Category Data Mining and Machine Learning Approach to Characterize and Detect Health Misinformation on Social Media," in IEEE Internet Computing, vol. 25, no. 5, pp. 43-51, 1 Sept.-Oct. 2021, doi: 10.1109/MIC.2021.3063257.**

This article examines the spread of false information about health on social media, with a particular emphasis on the 2016 Twitter debate on Zika. After identifying popular tweets that contain both false and accurate information, the study creates a data mining strategy to extract attributes for the purpose of detecting disinformation. It takes into account signal-based features, content and language features, user features, and information dissemination features. These features are used to train classifiers such as random forest (RF) and support vector machine (SVM). When all feature categories were put together, the RF classifier detected disinformation with over 83% accuracy and over 90% AUC. Responding to false information during health emergencies such as the COVID-19 pandemic and the Zika epidemic can benefit greatly from this strategy.

**[5] Anna L. Buczak, Erhan Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," in IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153 - 1176, 26 October 2015, doi: 10.1109/COMST.2015.2494502.**

This survey paper reviews the application of machine learning (ML) and data mining (DM) methods in cyber security intrusion detection. It categorizes intrusion detection into misuse-based, anomaly-based, and hybrid techniques. The paper discusses the importance of data in ML/DM, the challenges in cyber security, and provides recommendations based on specific problem characteristics. It highlights the need for labeled data and suggests creating new datasets to advance research in this field. Additionally, the paper identifies the importance of fast incremental learning for regular model updates in cyber security applications, making it a valuable resource for those entering this domain.

This survey paper provides a comprehensive overview of machine learning (ML) and data mining (DM) methods as applied to cyber analytics, particularly in the context of intrusion detection. The paper highlights the significance of cyber datasets used in ML/DM and addresses the complexity of ML/DM algorithmsThe paper emphasizes the need for ML and DM methods in the cyber domain, both for misuse and anomaly detection, but notes the absence of a universally effective method. It underscores the importance of quality training and testing datasets, including network and kernel-level data.The paper also highlights unique challenges in the cyber realm, particularly the frequent need for model retraining. It suggests that researching fast incremental learning methods for daily model updates in misuse and anomaly detection represents a promising avenue for further investigation

**[6] N. Cao, C. Shi, S. Lin, J. Lu, Y. Lin, C. Lin, “TargetVue: Visual Analysis of Anomalous User Behaviors in Online Communication Systems”, in IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 280-289, 11 August 2015, doi: 10.1109/TVCG.2015.2467196**

This research paper introduces a novel visual analysis system called "TargetVue" that aims to detect anomalous users in online communication systems, such as email and social media platforms. The paper highlights the challenges in automated anomaly detection, including the need for proper ground truth for model training and evaluation, often requiring human judgment to fine-tune the detection. TargetVue employs an unsupervised learning model and innovative visualization techniques to present the behaviors of suspicious users in a behavior-rich context. This is achieved through the use of ego-centric glyphs, which efficiently summarize a user's communication activities, features, and social interactions. The glyphs are arranged on a triangle grid to enable easy comparisons between users, capturing similarities among them.The paper discusses the application of TargetVue in a social bot detection challenge using Twitter data and a case study based on email records. The evaluation shows that TargetVue is beneficial for the detection of users with anomalous communication behaviors.The paper concludes by suggesting future work in tuning the anomaly detection model based on user feedback and the integration of advanced methods based on active learning techniques. The authors also plan to conduct a formal user study to further evaluate the usability of the TargetVue system.

**[7] Ravneet Kaur, Sarbjeet Singh. "A survey of data mining and social network analysis-based anomaly detection techniques." Egyptian Informatics Journal, Volume 17, Issue 2, 2016, Pages 199-216. ISSN 1110-8665. doi: 10.1016/j.eij.2015.11.004.**

The paper by Ravneet Kaur and Sarbjeet Singh provides a comprehensive survey of anomaly detection techniques employing data mining and social network analysis. It discusses numerous techniques for spotting anomalies or outliers in data, with an emphasis on how they might be employed in social networks. The authors highlight the importance of these strategies in identifying anomalies as they examine their usefulness in various fields. For scholars and practitioners looking for information on anomaly detection techniques in relation to data mining and social network analysis, the survey is a great resource.

**[8] T.R Soumya, S.Solai Manohar, N.Bala Sundara Ganapathy, Leema Nelson, A. Mohan, M.Thurai Pandian. "Profile Similarity Recognition in Online Social Network using Machine Learning Approach" in International Conference on Inventive Research in Computing Applications (ICIRCA), 29 December 2022, Pages 805-809. ISSN 22473631 doi:10.1109/ICIRCA54612.2022.9985683**This paper addresses the growing concern of fake profiles on online social networks (OSN) by proposing a machine learning-based approach for their detection. The authors highlight the increasing threats to individual privacy and security on OSNs due to the exploitation of vulnerabilities. They present a method to collect and preprocess a dataset for training their model, using techniques like data cleaning and validation. Various machine learning algorithms, including Support Vector Machine (SVM), Deep Neural Network (DNN), and Random Forest, are employed to distinguish between trusted and fake profiles. The results are evaluated through metrics like accuracy, precision, recall, and F1 score, demonstrating the effectiveness of the proposed approach.

**[9] AlAjlan, Shatha AbdulAziz and A. Khader, “Machine learning approach for threat detection on social media posts containing Arabic text,” Evolutionary Intelligence, vol. 14, no. 2, pp. 811–822, 2021, doi: https://doi.org/10.1007/s1206502000458w.**

The research's primary objective is the development of a Convolutional Neural Network (CNN) model for classifying Instagram content, encompassing images and Arabic comments, with a focus on threat detection, addressing a gap in research regarding Arabic social media threat monitoring. The Inception v3 model, a deep CNN pretrained on ImageNet, is fine-tuned using transfer learning, while a separate CNN model is utilized for comment classification. The dataset includes 1792 training samples, 384 validation samples, and 384 testing samples for comment classification, and 700 training, 150 validation, and 150 testing samples for image classification. Data is sourced from Instagram through the Instagram API and search engine, manually labeled for threat categorization, and cleaned to remove non-Arabic comments, ensure sufficient word count, eliminate @mentions, punctuation, and tokenization, and remove text-only or self-images.

The research, unique in its focus on Instagram, may require further investigation to determine the method's applicability to other social media platforms. Feature extraction techniques include Histogram Oriented Gradient (HOG) and Local Binary Pattern (LBP) for image analysis, and Haar-like features for Viola-Jones face detection. The methodology consists of three phases: data collection, data preprocessing, and classification, with a CNN model employed for both image and comment classification.

The model's performance is exceptional, achieving a 96% accuracy for images and an outstanding 99% for comments, demonstrating the potential of CNN models in threat tracking on Instagram. This research holds promise for future applications in social media monitoring, showcasing the significance of this work in the field.

**[10] Z. Abbass, Z. Ali, M. Ali, B. Akbar and A. Saleem, "A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning," 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 2020, pp. 363-368, doi: 10.1109/ICSC.2020.00073.**

The research paper explores a framework to predict different types of social media crimes using Twitter data, employing machine learning algorithms like Multinomial Naive Bayes, K-Nearest Neighbors, and Support Vector Machine. The framework successfully categorizes crimes like Cyber-Stalking, Cyber-Harassment, Cyber-Bullying, Cyber-Scam, and Cyber-Hacking with a notable 92% accuracy when using Bag-of-Words.

Despite these achievements, the paper raises concerns about the practicality of applying this framework to real-time or temporal data, as it primarily focuses on static data. The researchers find it important to note that traditional machine learning and data mining approaches that deal with such static data already exist, which, while effective, may not be well-suited for real-time detection or time series data. Nevertheless, the research highlights the significance of social media data for various investigative purposes and provides a foundation for potential improvements in the future, including the addition of new crime categories, real-time Twitter data analysis, and system enhancements. Ultimately, this framework holds promise for law enforcement agencies in reducing and preventing social media-related crimes.