Social Media Mining to Detect Online Violent Extremism using Machine Learning Techniques

Shynar Mussiraliyeva, Kalamkas Bagitova, Daniyar Sultan
Al-Farabi Kazakh National University, Almaty, Kazakhstan

***Abstract*—In this paper, we explore the challenging domain of detecting online extremism in user-generated content on social media platforms, leveraging the power of Machine Learning (ML). We employ six distinct ML and present a comparative analysis of their performance. Recognizing the diverse and complex nature of social media content, we probe how ML can discern extremist sentiments hidden in the vast sea of digital communication. Our study is unique, situated at the intersection of linguistics, computer science, and sociology, shedding light on how coded language and intricate networks of online communication contribute to the propagation of extremist ideologies. The goal is twofold: not only to perfect detection strategies, but also to increase our understanding of how extremism proliferates in digital spaces. We argue that equipping machine learning algorithms with the ability to analyze online content with high accuracy is crucial in the ongoing fight against digital extremism. In conclusion, our findings offer a new perspective on online extremism detection and contribute to the broader discourse on the responsible use of ML in society.**

***Keywords*—NLP; machine learning; social networks; extremism detection; textual contents**

I. Introduction

In the burgeoning digital age, the ubiquity of social media as the world increasingly navigates towards digitization, the rise of social media and the vast landscape of user-generated content it produces have opened up new vistas of communication and social interaction [1]. However, this digital evolution has also given rise to formidable challenges, one of the most pressing being the proliferation of online extremism. The cloak of anonymity provided by the internet, along with the unprecedented reach of social media, has exacerbated the spread of extremist ideologies, thereby necessitating effective detection mechanisms. This paper is dedicated to exploring the application of Machine Learning (ML) techniques for the detection of online extremism on social media platforms [2].

The cornerstone of our study revolves around the deployment of six specific Machine Learning algorithms. The comparative analysis of these methodologies forms a significant part of our research, enabling us to discern the relative strengths and weaknesses of each in the context of online extremism detection.

The nature of online extremism is as complex as it is harmful [3]. To address it effectively, we need to delve into the intricacies of online communication patterns, exploring how coded language and digital interaction networks serve as conduits for the spread of extremist ideologies [4]. In this context, our study is not merely a technical exploration of machine learning techniques but also a sociolinguistic inquiry into the nature of online extremism itself [5].

Machine Learning has an ability to sift through vast datasets and identify patterns that may be invisible to the human eye [6]. By training ML algorithms to recognize and flag extremist sentiments, we aim to create an effective line of defense against the spread of dangerous ideologies.

However, the application of Machine Learning in such a sensitive domain also raises ethical considerations. With the power to scrutinize digital communication comes a responsibility to use it wisely and fairly. Therefore, we also dedicate a portion of our research to discussing ethical considerations surrounding the use of ML for online extremism detection.

In essence, our exploration is multifaceted, combining a technical examination of Machine Learning methodologies with an investigation into the sociolinguistic phenomena that characterize online extremism [7]. We conclude with a discussion on the ethical implications of applying these techniques, thereby providing a holistic view of the challenge at hand.

In this paper, we investigate the challenge of identifying extremist views and appeals for violence on social media platforms. More specifically, our emphasis is on comprehending and identifying extremist ideas in the information posted by online users. In order to comprehend the appeals made by extremist groups through a data extraction point of view, we first undertake an in-depth study of the material, including vocabulary and subject descriptors [8]. In order to detect extremist views included in the data, six distinct sets of relevant traits were uncovered, and multiple training methods were evaluated against one another. This is an original use for the automated identification of violent extremism in material, and it uses a mix of the efficient attribute architecture and classifiers that we have provided.

The following are some of the ways in which this paper significantly contributes to the body of knowledge and pioneers new ground in the field:

1) The use of information extraction as well as knowledge mining to identify the distinctive characteristics of violent extremism and appeals to conduct violent actions that are included in the material created by internet users. This technique exposes details regarding extremist ideologies through the lens of data analytics.

2) Corpus: this paper introduces the social network and compiles a fresh collection of information for the purpose of identifying extremist communications and appeals to extremism [22], which is now the most widespread social network among young people in Kazakhstan [9]. Psychologists divide the dataset into two groups based on whether or not it contains extremist statements or appeals to violence. The dataset was acquired from a social network that is extensively utilized in the republics of the Commonwealth of Independent States (CIS).

3) Applying machine learning: There, we applied six machine learning algorithms to violent extremism detection problem. The results are given using different evaluation parameters of machine learning methods.

This paper provides an in-depth analysis of the use of ML in online extremism detection, offering insights into both the technical and ethical aspects of this issue. Our findings contribute to the broader conversation on online safety, the responsible use of AI, and the role of digital platforms in our society. This research underlines the necessity of a multidisciplinary approach to address the escalating issue of online extremism and exemplifies the crucial role of machine learning in curbing this growing menace. Our paper contributes to the broader discourse on online safety, the responsible use of artificial intelligence technologies, and the role of digital platforms in maintaining the fabric of our social structure. Through our research, we strive to shed light on how a multidisciplinary approach can effectively address the escalating issue of online extremism and underline the pivotal role of Machine Learning in combating this digital menace.

II. Related Works

Machine learning models are used in different applied tasks as smart city and smart energy [10], security-related problems [11], and text processing [12]. The issue of detecting and mitigating online extremism has evolved over time, calling for more sophisticated, adaptable, and scalable solutions. This review delves into the transition from traditional methodologies to machine learning (ML) techniques and offers a comparative analysis of these two broad categories.

A. Conventional Methods for Online Extremism Detection

The initial responses to online extremism have largely been traditional in nature, utilizing manual review processes and rule-based algorithms [13]. These methods involve human moderators, who, based on their interpretation of the content, decide on its appropriateness or otherwise. Similarly, rule-based algorithms operate by matching specific patterns, keywords, or blacklisting certain types of content or users known to promote extremist views [14].

While these traditional methods have proven effective to some extent, they are not without significant limitations. The major constraint is the issue of scalability, given the exponential growth of user-generated content on social media [15]. Manual review processes are inherently time-consuming and labor-intensive, making them less practical for large scale operations [16]. Rule-based algorithms, despite being automated, are often rigid, unable to adapt to the evolving nuances of online extremism [17]. Furthermore, both methods carry the risk of inherent bias due to the subjective nature of interpretation, which can lead to both over-censorship and under-censorship.

Traditional methods of addressing online extremism have relied predominantly on manual and rule-based approaches [18]. While these methods have provided a starting point, they grapple with issues related to scalability, adaptability, and inherent bias. As user-generated content has grown exponentially, the limitations of these methods, in terms of time, resources, and subjectivity, have become increasingly pronounced [19].

B. Machine Learning Methods for Online Extremism Detection

To overcome these limitations, researchers have turned to ML techniques that can handle vast amounts of data and adapt to evolving online communication patterns. These techniques rely heavily on feature extraction methods, which transform raw text data into a structured format that ML algorithms can understand and analyze.

The limitations of traditional methods have driven the exploration of more advanced solutions, leading to the adoption of Machine Learning (ML) techniques. These methods offer key advantages including scalability, accuracy, adaptability, and the potential for real-time detection [20]. This study focuses on six ML algorithms: Support Vector Machine (SVM), Decision Tree, Random Forest, K Nearest Neighbors (KNN), Naive Bayes, and Logistic Regression. Each algorithm exhibits unique strengths and weaknesses, offering a versatile toolkit for addressing the multifaceted challenge of online extremism detection [21].

A critical component of ML success in detecting extremist content lies in feature extraction, where raw data is transformed into a format that these algorithms can utilize. Techniques such as Term Frequency-Inverse Document Frequency (tf-idf) [22], Bag of Words (BoW) [23], and Word2Vec have been commonly employed for this purpose [24]. Tf-idf emphasizes the importance of words in a document, BoW assesses the frequency of words independent of the order or grammar, and Word2Vec encapsulates semantic relationships between words, by mapping them as vectors in a multidimensional space.

In a seminal work, authors employed the tf-idf method in text categorization, highlighting its effectiveness in weighing the importance of words within a given document [25]. By measuring the frequency of a term adjusted by its rarity in the entire corpus, tf-idf helps identify key terms that might indicate extremist content.

The BoW method, despite its simplicity, has been extensively used due to its effectiveness and interpretability. In this approach, text is reduced to a 'bag' of its words, disregarding grammar and word order but preserving frequency. Researchers have shown how BoW can be powerful when combined with ML techniques, particularly in topic modeling [26].

Part of Speech (PoS) tagging has also been used to improve the performance of ML algorithms in detecting extremist

content. Next study demonstrated that PoS features, when used in conjunction with SVM, significantly improved the detection of hate speech [27].

Word2Vec, goes beyond simple frequency-based methods and captures the context and semantic relationships between words [28]. By representing words as vectors in a high-dimensional space, it enables the detection of patterns and associations that can be indicative of extremist ideologies.

.When comparing traditional methods with ML techniques, it's evident that each offers distinct advantages. Traditional methods are relatively straightforward to implement and their outputs are easily interpretable. However, their scalability issues and inability to evolve with the changing landscape of online extremism significantly limit their efficacy.

Conversely, ML methods present greater adaptability and scalability. Their ability to learn from data patterns, adapt to new information, and handle extensive and diverse content make them a promising solution for online extremism detection. However, the complexity of ML models can pose challenges, specifically in interpretability and transparency [29]. Additionally, the effectiveness of ML methods is inherently tied to the quality of input data and the relevance of the features extracted.

In conclusion, despite the efficacy of traditional methods in certain contexts, ML techniques appear to be a more potent solution for detecting online extremism on a large scale. However, careful consideration must be given to ethical implications, such as potential privacy infringements and biases, as we harness the power of these advanced technologies in our quest to maintain safe and respectful online environments.

III. DATASET

We first wanted to establish the hazard criterion before we could decide if a text was connected to extremism or not. Putting together a list of phrases is one such method. For the purpose of defining the term, a collection of key terms was compiled and used to conduct an investigation into the data contained inside the social networking site [30]. The software program deduces that the text should be further investigated because it contains the stated keywords, and this conclusion is based on the fact that these keywords are present in the content. The whole data collection, an analysis of the postings, and a categorization of the texts are shown in Fig. 1.

The achievement of data collecting might be carried out differently depending on the source of information, but the fundamental idea of its framework should be maintained throughout. The component of the program that is required for the extraction knowledge from publicly accessible sources has as its primary objective the successful completion of tasks in a timely and efficient manner. It is vital to make advantage of the built-in techniques for receiving data from sources (API) in order to achieve a high level of effectiveness [31]. In the event that such techniques are not available, it will be essential to collect the relevant data via making HTTP queries.

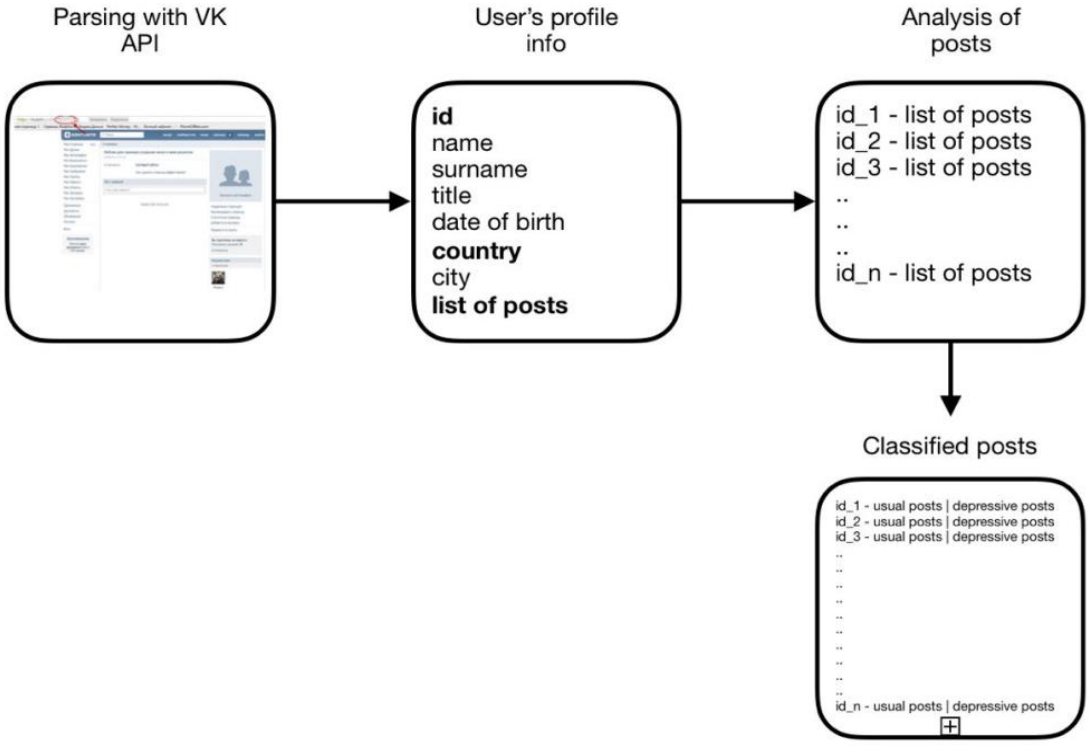

Fig. 1. Architecture of the data collection unit.

The application is comprised of three separate components as data collection unit, keyword scan unit, and machine learning unit which are as follows:

1) The Data Collection Unit is in charge of gathering data from publicly available channels and sending it on for further processing; In order to process the data coming from the VKontakte social network, an application written in Python programming language has been created. We partly parsed open access profiles using the publicly available VK API [32], which we accessed.

2) The Keyword scan unit is liable for detecting keywords in a vast quantity of data; because we had earlier a list of keywords that are often discovered in communications connected to violent extremism, we used a linear search for terms in each written content, and then partitioned the text into tokens. Specialists were consulted and had a role in the development of keywords that may be utilized when searching for potentially harmful content;

3) The determination of whether or not the data is connected to violent extremism falls within the purview of the textual content classification module. In this stage, we apply different machine learning algorithms for online extremism detection problem.

Data collection unit is the initial stage of the proposed framework. For the purpose of data collection, we make use of the Vkontakte social network. For the purpose of data collecting, a parser is developed in Python version 3.6. The query was used in order to interact with the application programming interface (API) of the social networking site. It was decided that the program package Pycharm will serve as an experimental platform.

IV. Materials and Methods

A. Feature Engineering

It is vital to specify the criterion of "hazard" before assigning the material to being associated to violent extremism in any way. Defining a list of keywords is one possible solution to the problem. The produced application made use of this strategy for identifying the different kinds of information that may be found. For the purpose of defining the term, a list of keywords was developed, and those keywords were utilized to conduct an analysis of the material contained inside the social network Vkontakte. The conclusion reached by the software application about the text's suitability for more investigation is based on whether or not the text contains the keywords that were provided. In our research, we used statistics, POS features, features based on LIWC features, and features based on TF-IDF word frequency.

In order to get an understanding of the informativeness of these feature sets, we use principal component analysis (PCA) [33] in Fig. 2 to create a visual representation of the features on the gathered corpus in a two-dimensional space. When we look at Fig. 2, we can see the polarity and subjectivity of the explored dataset to detect violent extremism. This suggests that our classifier should have an easier time distinguishing between the two categories.

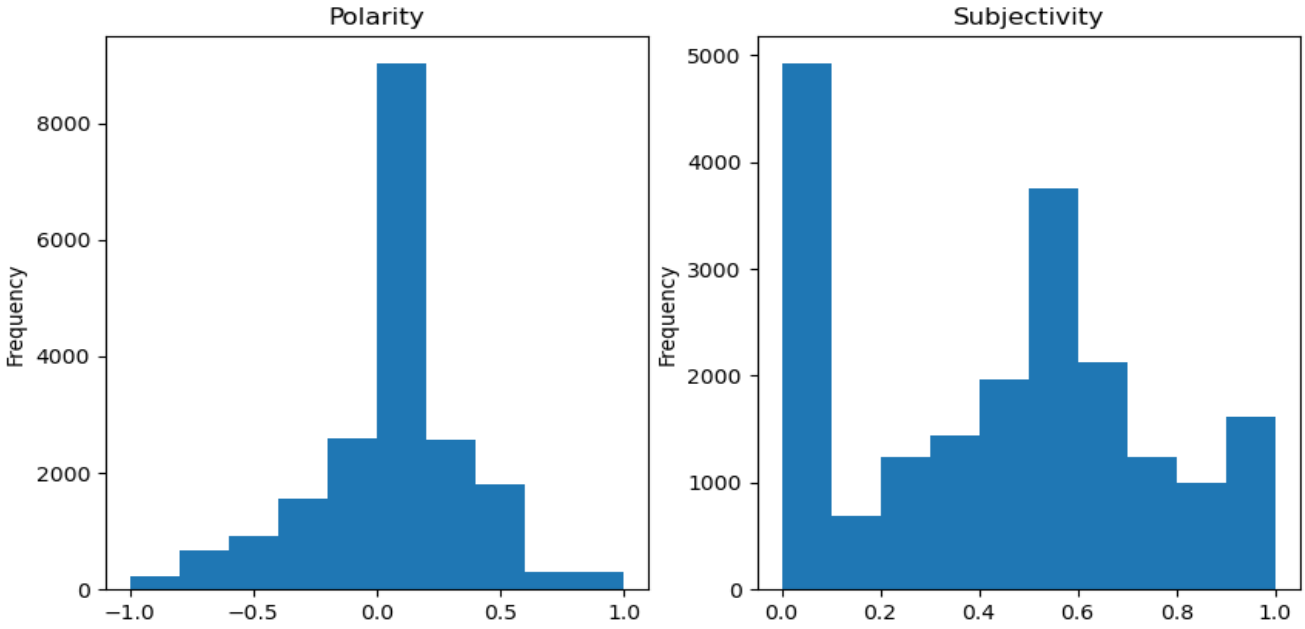

Fig. 2. Extracted features.

1) Classification Models: Messages relating to extremism that are found within the content of social networks may be detected using the normal supervised learning based classification problem. We created classifiers to train data couples of input objects $\{x_i, y_i\}_i^n$ and supervisory signals $\{x_i\}_i^n$ [34], taking into consideration a corpus that consisted of texts with tags $\{y_i\}_i^n$.

$$Y_i = F(x_i) \tag{1}$$

If yi=1 indicates that the text in question is "extremist intended text," then yi=0 indicates that the text in question is "not extremist intended text." The goal of training phase of the classification is to reduce the amount of incorrect classifications made in the data used for training. The inaccuracy in the prediction is going to be presented in the form of a loss function called L(y, F(x)), where y will represent the actual label and F(x) will represent the anticipated label. In broad strokes, the purpose of training is to arrive at the best possible prediction model F(x) by finding solutions to the following optimization problems:

$$\hat{F} = \arg\min_{F} E_{x,y}\left[L\left(y, F(x)\right)\right] \tag{2}$$

The categorization of extremism-related writings is shown in Fig. 3, which illustrates the schema. The methods of oversampling and undersampling, in addition to statistics, LIWC, POS, and TF-IDF, are included in the features. These approaches are used to manage unbalanced data. The machine learning models were given access to all of the retrieved characteristics.

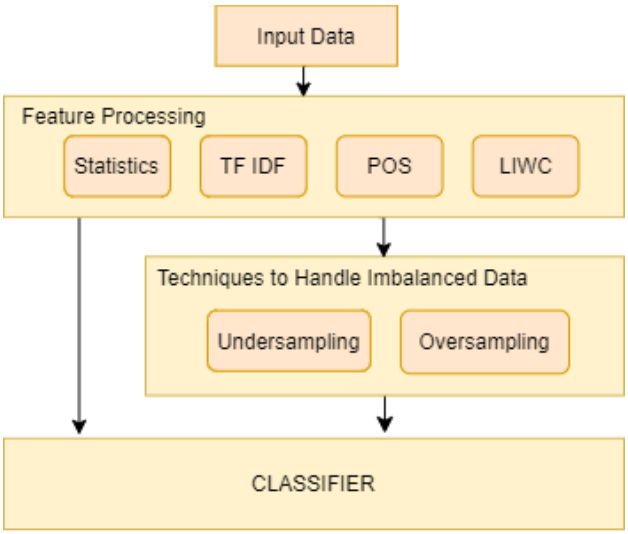

Fig. 3. Architecture of the proposed framework.

Our primary objective is to detect and isolate any content within the chosen dataset that exhibits ties to violent extremism, sourced from any participating users. The initiation of our text classification process involves engaging with the entirety of the domain, which consists of multidimensional objects procured from the dataset under scrutiny.

Our fundamental feature extraction mechanisms stem from a diverse array of methodologies, namely, possibilities offered by N-gram [34], Linguistic Inquiry and Word Count (LIWC) category functions [35], the Latent Dirichlet Allocation (LDA) model [36], and an assortment of their combinational elements. These characteristic traits, constituting the foundation of our analytical tools, are solely derived from the training data that we have meticulously collated.

This holistic approach facilitates a comprehensive understanding of the data, subsequently enabling the successful identification of any material that may be associated with violent extremism. It is through the deployment of these robust methods and their respective combinatory factors that we are able to achieve our mission in an effective and efficient manner.

The confusion matrix is an instrumental method for collating an encapsulated summary of the results stemming from a classification process. Relying on accuracy as a solitary metric can potentially lead to misleading interpretations, particularly in scenarios where the volume of observations in individual classes is imbalanced. This tool offers a comprehensive insight into our methodology's effectiveness in discerning correct classifications from erroneous ones, and its proficiency in successfully obtaining correct outcomes.

It becomes manifestly clear through the confusion matrix that correct classification occurrences for classes with low extremity are fewer, which is predominantly responsible for their dismal accuracy and recall rates. This phenomenon underlines the importance of considering multiple performance metrics beyond mere accuracy, especially when dealing with data that is inherently imbalanced. It underscores the necessity to implement more nuanced approaches that can accurately reflect the capabilities of the classification model in diverse scenarios, thereby contributing to a more informed interpretation of its performance.

Accuracy is sometimes referred to as positive predictive value, while precision is the ability to remember information. This is the fraction of comparable examples that were retrieved from the total number of instances. The sensitivity is measured by the recall, which is the percentage of the total number of appropriate cases that is equal to the number of relevant examples that have been retrieved. When it comes to classification, accuracy is determined by dividing the number of true positives (TP) by the total number of labeled members (TP + FP) that belong to this class. It is important to keep in mind that the total number of true positives (TP) in classification is split into the number of instances that do in fact belong to the class (TP+FN), so, in this research we used accuracy [37], precision [38], recall, and F1-score [39].

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{6}$$

In this part, we evaluate the outcomes of using several ML techniques for the categorization of violent extremism using a variety of distinct feature sets.

As can be seen in Table I, the general efficacy of each approach increases when more characteristics are combined

into one cohesive whole. This finding provides further evidence that the traits that were obtained are both informative and useful. However, the role of each characteristic fluctuates quite a bit, which shows that there are oscillations in the outputs of the different techniques. When using all groupings of characteristics as input data, the SVM and LR techniques showed the highest levels of productivity compared to the other methods that were utilized. Both Random Forest and Naive Bayes have shown impressive performance in F1.

TABLE I. RESULTS OF APPLYING MACHINE LEARNING IN ONLINE EXTREMISM DETECTION

Approach	Applied Feature	Accuracy	Precision	Recall	F-measure	AUC-ROC
SVM	Statistics	78.64%	78.17%	77.29%	74.16%	74.72%
	Statistics&TF-IDF	79.35%	79.08%	79.75%	79.43%	78.17%
	Statistics&TF-IDF&POS	81.31%	81.15%	81.68%	81.37%	81.07%
	Statistics&TF-IDF&POS&LIWC	84.97%	84.28%	83.21%	83.08%	83.01%
Decision Tree	Statistics	58.64%	58.17%	57.29%	54.16%	54.72%
	Statistics&TF-IDF	61.35%	61.08%	60.75%	60.43%	60.17%
	Statistics&TF-IDF&POS	62.31%	62.15%	61.68%	61.37%	61.07%
	Statistics&TF-IDF&POS&LIWC	64.97%	64.28%	63.21%	63.08%	63.01%
RF	Statistics	60.64%	60.17%	59.29%	56.16%	56.72%
	Statistics&TF-IDF	63.35%	63.08%	62.75%	62.43%	62.17%
	Statistics&TF-IDF&POS	64.31%	64.15%	64.68%	64.37%	64.07%
	Statistics&TF-IDF&POS&LIWC	66.97%	66.28%	65.21%	65.08%	65.01%
KNN	Statistics	62.64%	62.17%	61.29%	58.16%	58.72%
	Statistics& TF-IDF	65.35%	65.08%	64.75%	64.43%	64.17%
	Statistics&TF-IDF&POS	66.31%	66.15%	65.68%	65.37%	65.07%
	Statistics&TF-IDF&POS&LIWC	68.97%	68.28%	67.21%	67.08%	67.01%
Naïve Bayes	Statistics	56.64%	56.17%	55.29%	52.16%	52.72%
	Statistics&TF-IDF	59.35%	59.08%	58.75%	58.43%	58.17%
	Statistics&TF-IDF&POS	60.31%	60.15%	59.68%	59.37%	59.07%
	Statistics&TF-IDF&POS&LIWC	62.97%	61.28%	61.21%	61.08%	61.01%
LR	Statistics	79.64%	79.17%	78.29%	77.16%	77.72%
	Statistics&TF-IDF	82.35%	82.08%	81.75%	81.43%	81.17%
	Statistics&TF-IDF&POS	83.31%	83.15%	82.68%	82.37%	82.07%
	Statistics&TF-IDF&POS&LIWC	85.97%	85.28%	84.21%	84.08%	84.01%

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC), encompassing all of the extracted features, is a crucial metric employed for evaluating the performance of each classification task [40]. This measure offers a comprehensive overview of the effectiveness of our classification model across different thresholds, serving as a critical instrument for performance evaluation.

Based on our empirical findings, it was observed that an incremental enhancement in the AUC-ROC performance was intimately linked with an increase in the quantity of incorporated features. This positive correlation signifies the importance of feature richness in enhancing classification performance and illuminates the consequential role these characteristics play in fine-tuning the efficacy of our model.

This discovery substantiates the concept that extending the complexity of our feature space, through the addition of more discriminative characteristics, is likely to augment the accuracy and reliability of our classifier. Therefore, this observation can inform future developments and refinements to optimize the model's predictive capabilities.

Employing the Logistic Regression methodology yielded an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score of 0.893. This value surpasses the AUC generated by any other method, signifying its superior performance in the classification task. Remarkably, a considerable portion of the alternate strategies we explored also achieved AUC values exceeding 0.9, an indication of their substantial classification prowess.

In an attempt to critically evaluate the performance of our textual classification model and its capability to discern content affiliated with extremism from a diverse set of online communities, we embarked on a systematic enhancement of

our textual corpus. The purpose of this extensive augmentation was to ensure a broad array of data sources, providing a more diverse and encompassing data set for the model to learn from.

The expanded corpus, characterized by a heterogeneous amalgamation of data, was deployed to test the efficiency of our algorithms across a wide spectrum of contexts. These contexts comprised news articles, which offer a formal representation of language, content identified as toxic that typically involves aggressive or harmful language, spam that usually entails repetitive or irrelevant content, promotional material characterized by persuasive language, and humoristic entries, encapsulating a different style and tone of language.

Upon examination of the results, it was observed that our models exhibited a remarkable precision exceeding 90% in successfully identifying and distinguishing text related to extremism from the various other domains tested. Thus, these findings indicate that the utilization of our chosen methodology to extract distinctive features was indeed efficacious in categorizing instances of extremist rhetoric emanating from diverse sources.

This suggests a significant potential of our models to detect and isolate such extremist content from a wide array of online communities, thereby reaffirming their reliability and precision. The success of our approach opens a new path in text classification, particularly in areas related to security and online community management, demonstrating the power of advanced artificial intelligence in dealing with complex, real-world challenges.

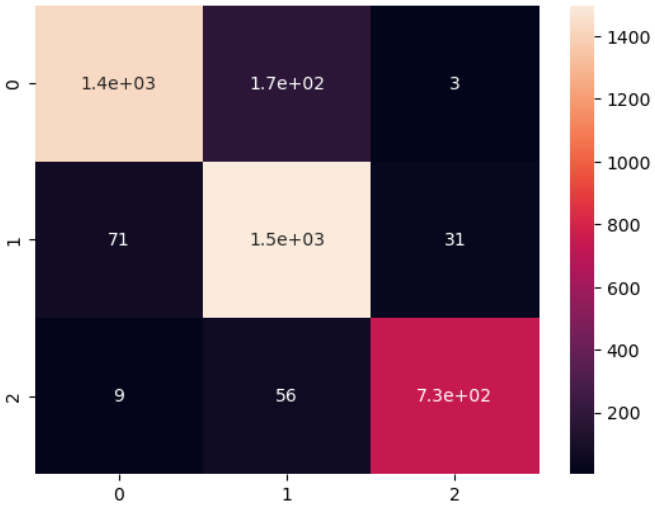

Fig. 4. Confusion matrix of three class classification.

Fig. 4 provides a visual representation of a confusion matrix, a potent tool adopted in the context of discerning violent extremism through textual analysis. This specific matrix employs a tripartite classification scheme. Class 0 constitutes texts which display no correlation with violent extremism, thereby serving as a benchmark of non-extremist discourse. Class 1 comprises neutral texts which, although not explicitly extremist, contain terminology and phrases associated with violent extremism. Consequently, these texts present a subtler form of discourse that necessitates nuanced understanding. Finally, Class 2 envelops texts that are unequivocally linked to violent extremism, highlighting the most explicit and overt instances of such discourse.

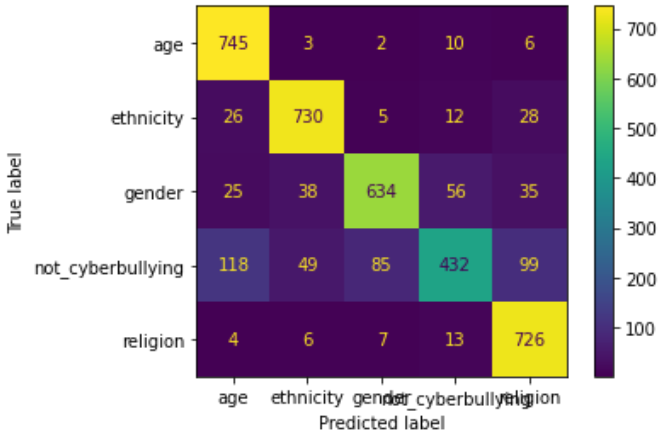

Fig. 5. Confusion matrix of five class classification.

The empirical evidence gleaned from the outcomes of this classification scheme underscores the effectiveness of the proposed framework. Not only does it showcase a capacity for binary classification - differentiating between extremist and non-extremist content - but it also adeptly navigates the intricate landscapes of multi-classification challenges. This includes distinguishing between varying degrees and types of extremist discourse. Thus, the framework's success in such a complex task accentuates its potential for broader application in the realm of text classification.

Fig. 5 provides an illustrative exposition of the application of a five-category classification scheme, implemented within the conceptual framework of the proposed model. The quintuple classification strategy organizes the data into five unique classes, namely: age, ethnicity, gender, non-cyberbullying incidents, and religious considerations. This categorization delineates a wide-ranging spectrum of potential data types, offering a comprehensive perspective on the classification abilities of the model across diverse contexts.

The findings derived from the subsequent results suggest the occurrence of a minimal number of false negatives. In classification tasks, false negatives represent instances where a particular data point has been incorrectly labeled, signifying a discrepancy between the anticipated and actual classification. The paucity of false negatives in the results implies a considerable degree of precision in the classification process undertaken by the proposed model.

This high level of accuracy not only provides testimony to the model's proficiency in accurately classifying data but also underpins its potential for trustworthy deployment in scenarios necessitating precise data categorization. Therefore, this finding bolsters the robustness and versatility of the proposed model in undertaking a variety of data classification tasks.

V. Discussion

As we navigate the complexities of detecting and mitigating online extremism, the power of Machine Learning (ML) stands out as a promising tool in this ongoing struggle. This

discussion delves into the practical applications, limitations, and future perspectives associated with the use of ML for online extremism detection.

A. Practical Use

The emergence of ML techniques in the realm of online extremism detection opens up myriad practical applications. As a technology that can process large volumes of data swiftly and accurately, ML algorithms provide a feasible solution for monitoring the vast and rapidly expanding universe of user-generated content on social media platforms.

In the context of public safety and national security, ML algorithms can help law enforcement agencies to proactively identify potential threats by detecting extremist narratives or recruitment attempts within the vast amount of social media content [41]. On a similar note, social media companies can leverage these technologies to maintain community standards, flagging and removing harmful content, thereby preserving the platform's integrity and ensuring the safety of their users.

Furthermore, the use of feature extraction methods like tf-idf, Bag of Words, Part of Speech, and Word2Vec facilitates the identification of underlying themes, patterns, and trends that may not be apparent to human observers [42]. This assists not only in the identification of extremist content but also in understanding its context, evolution, and influence.

B. Limitations

Despite its promise, employing ML in this arena is not without its challenges. First and foremost is the issue of accuracy. False positives (non-extremist content wrongly flagged as extremist) and false negatives (extremist content that goes undetected) can both have severe implications [43]. The former risks infringing upon freedom of speech, while the latter fails to stem the spread of harmful content. Balancing sensitivity and specificity in model performance is an ongoing challenge.

Secondly, the effectiveness of ML algorithms heavily depends on the quality of the training data. Gathering extensive, representative, and accurately labeled data for training purposes is a significant challenge due to the sensitive and dynamic nature of extremist content.

Thirdly, ML algorithms often suffer from a 'black box' problem, where the decision-making process is opaque and hard to interpret. This can lead to difficulties in understanding why certain content was flagged, hindering improvements and adjustments to the system.

Lastly, there are ethical considerations. While using ML to detect online extremism has clear security benefits, it may also raise concerns about user privacy and data misuse. Achieving a balance between security needs and user privacy is a delicate and complex task.

C. Future Perspectives

In this research we consider detection of violent extremism in online user contents. Nowadays, different methods are used to teach students in low school and high school to give children and students good knowledge ethics [44] and righteousness [45]. In our study, we used a machine learning approach that is the one of the state-of-the-art methods in this area. Looking forward, the application of ML for online extremism detection presents a vibrant research area with numerous exciting prospects [46]. The development of more sophisticated algorithms and feature extraction methods can further improve the accuracy and efficiency of detection systems.

Further exploration into hybrid models combining multiple ML algorithms or integrating ML with traditional rule-based methods could leverage the strengths of both approaches. Moreover, the integration of ML with Natural Language Processing (NLP) and sentiment analysis techniques could offer a more nuanced understanding of extremist narratives and rhetoric [47].

Additionally, interpretability of ML algorithms is a critical area for future work. Developing techniques to enhance the transparency of these models will not only increase trust in their decisions but also provide more insight into the underlying patterns of extremist content [48].

Finally, research should also focus on ethical and privacy-preserving ML methodologies. This includes exploring how to minimize data requirements, anonymize data used, and ensure that the application of these technologies respects user rights and societal norms.

In conclusion, the adoption of Machine Learning for online extremism detection presents a potent tool with numerous practical applications. However, addressing its limitations and considering future directions is crucial for the responsible and effective use of this technology in ensuring a safer digital landscape.

VI. Conclusion

Online extremism poses a significant challenge in today's digital society, with its rapid dissemination and evolving nature causing widespread concern. This research has examined the use of Machine Learning (ML) methods as a valuable solution for detecting such extremist content within the vast landscape of user-generated social media content. The use of ML techniques, particularly in conjunction with feature extraction methods such as tf-idf, Bag of Words, Part of Speech, and Word2Vec, has been demonstrated to offer a scalable, adaptable, and effective approach.

However, as we progress in the application of ML techniques for online extremism detection, it is crucial to address the inherent limitations. These include issues of accuracy, dependency on the quality of training data, the interpretability of ML models, and the ethical implications related to user privacy and data usage.

Looking forward, there is vast potential for further development and refinement of ML algorithms, particularly in enhancing interpretability, improving data collection and labeling, integrating with other computational techniques, and considering ethical and privacy-preserving strategies. We must strive to balance security needs with preserving user rights and societal norms.

Ultimately, the objective of this research and the broader field is to contribute towards a safer and more respectful online

environment. ML has demonstrated significant promise in this regard, but its application must be conscientiously guided, ethically aware, and continually adapting to the evolving challenges of online extremism. It's a powerful tool in our arsenal, but like any tool, its effectiveness will depend on how we wield it.

ACKNOWLEDGMENT

This research has been/was/is funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP15473408)

REFERENCES

[1] Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., Shah, J., Sentiment Analysis of Extremism in Social Media from Textual Information, Telematics and Informatics (2020), doi: https://doi.org/10.1016/j.tele.2020.101345

[2] Mohammad Fraiwan, Identification of markers and artificial intelligence-based classification of radical twitter data, Applied Computing and Informatics, 2020, ISSN 2210-8327, https://doi.org/10.1016/j.aci.2020.04.001

[3] Ferreira, M. L. D. A., Graciano, P. F., Leal, S. R., & Costa, M. F. D. (2019). Night of terror in the city of light: terrorist acts in Paris and Brazilian tourists' assessment of destination image. Revista Brasileira de Pesquisa em Turismo, 13(1), 19-39.

[4] Al-Zewairi, M., & Naymat, G. (2017). Spotting the Islamist Radical within: Religious Extremists Profiling in the United State. Procedia computer science, 113, 162-169.

[5] Lestari, N. I., Hussain, W., Merigo, J. M., & Bekhit, M. (2023, January). A Survey of Trendy Financial Sector Applications of Machine and Deep Learning. In Application of Big Data, Blockchain, and Internet of Things for Education Informatization: Second EAI International Conference, BigIoT-EDU 2022, Virtual Event, July 29–31, 2022, Proceedings, Part III (pp. 619-633). Cham: Springer Nature Switzerland.

[6] Narynov, S., Mukhtarkhanuly, D., & Omarov, B. (2020). Dataset of depressive posts in Russian language collected from social media. Data in brief, 29, 105195.

[7] Hannah Ritchie, Joe Hasell, Cameron Appel and Max Roser. Terrorism. Our world in data. https://ourworldindata.org/terrorism

[8] Rashida, U., & Suresh Kumar, K. (2023). Social Media Mining to Detect Mental Health Disorders Using Machine Learning. In Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022 (pp. 923-930). Singapore: Springer Nature Singapore.

[9] Gautam, A. K., & Bansal, A. (2022). Effect of features extraction techniques on cyberstalking detection using machine learning framework. Journal of Advances in Information Technology Vol, 13(5).

[10] Altayeva, A., Omarov, B., Suleimenov, Z., & Im Cho, Y. (2017, June). Application of multi-agent control systems in energy-efficient intelligent building. In 2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS) (pp. 1-5). IEEE.

[11] Omarov, B., Suliman, A., & Tsoy, A. (2016). Parallel backpropagation neural network training for face recognition. Far East Journal of Electronics and Communications, 16(4), 801-808.Tsilingeridis, O., Moustaka, V., & Vakali, A. (2023). Design and development of a forecasting tool for the identification of new target markets by open time-series data and deep learning methods. Applied Soft Computing, 132, 109843.

[12] Mursi, K. T., Alahmadi, M. D., Alsubaei, F. S., & Alghamdi, A. S. (2022). Detecting Islamic radicalism Arabic tweets using natural language processing. IEEE Access, 10, 72526-72534.

[13] Berhoum, A., Meftah, M. C. E., Laouid, A., & Hammoudeh, M. (2023). An Intelligent Approach Based on Cleaning up of Inutile Contents for Extremism Detection and Classification in Social Networks. ACM Transactions on Asian and Low-Resource Language Information Processing.

[14] Koehler, D. (2017). How and why we should take deradicalization seriously. Nature Human Behaviour, 1(6), 1-3.

[15] Borum, R. (2017). The etiology of radicalization. The handbook of the criminology of terrorism, 218-219.

[16] Scrivens, R., Windisch, S., & Simi, P. (2020). Former Extremists in Radicalization and Counter-Radicalization Research. In Radicalization and Counter-Radicalization. Emerald Publishing Limited.

[17] Saleh, H., Alhothali, A., & Moria, K. (2023). Detection of hate speech using bert and hate speech word embedding with deep model. Applied Artificial Intelligence, 37(1), 2166719.

[18] Suliman, A., Shakil, A., Sulaiman, M. N., Othman, M., & Wirza, R. (2008, August). Hybrid of HMM and Fuzzy Logic for handwritten character recognition. In 2008 International Symposium on Information Technology (Vol. 2, pp. 1-7). IEEE.

[19] Scrivens, R., Wojciechowski, T. W., Freilich, J. D., Chermak, S. M., & Frank, R. (2023). Comparing the online posting behaviors of violent and non-violent right-wing extremists. Terrorism and political violence, 35(1), 192-209.

[20] Berhoum, A., Meftah, M. C. E., Laouid, A., & Hammoudeh, M. (2023). An Intelligent Approach Based on Cleaning up of Inutile Contents for Extremism Detection and Classification in Social Networks. ACM Transactions on Asian and Low-Resource Language Information Processing.

[21] Adraoui, M. A. (2017). Borders and sovereignty in Islamist and jihadist thought: past and present. International affairs, 93(4), 917-935.

[22] Sahu, A. K., Umachandran, K., Biradar, V. D., Comfort, O., Sri Vigna Hema, V., Odimegwu, F., & Saifullah, M. A. (2023). A Study on Content Tampering in Multimedia Watermarking. SN Computer Science, 4(3), 222.

[23] Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., & Khassanova, M. (2022). A Skeleton-based Approach for Campus Violence Detection. Computers, Materials & Continua, 72(1).

[24] Hart, G., & Huber, A. R. (2023). Five Things We Need to Learn About Incel Extremism: Issues, Challenges and Avenues for Fresh Research. Studies in Conflict & Terrorism, 1-17.

[25] Bamsey, O., & Montasari, R. (2023). The Role of the Internet in Radicalisation to Violent Extremism. In Digital Transformation in Policing: The Promise, Perils and Solutions (pp. 119-135). Cham: Springer International Publishing.

[26] Jahan, M. S., & Oussalah, M. (2023). A systematic review of Hate Speech automatic detection using Natural Language Processing. Neurocomputing, 126232.

[27] Bamsey, O., & Montasari, R. (2023). The Role of the Internet in Radicalisation to Violent Extremism. In Digital Transformation in Policing: The Promise, Perils and Solutions (pp. 119-135). Cham: Springer International Publishing.

[28] Ige, T., Kolade, A., & Kolade, O. (2023). Enhancing Border Security and Countering Terrorism Through Computer Vision: A Field of Artificial Intelligence. In Data Science and Algorithms in Systems: Proceedings of 6th Computational Methods in Systems and Software 2022, Vol. 2 (pp. 656-666). Cham: Springer International Publishing.

[29] Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., & Shah, J. (2020). Sentiment analysis of extremism in social media from textual information. Telematics and Informatics, 48, 101345.

[30] Ahmad, S., Asghar, M. Z., Alotaibi, F. M., & Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Human-centric Computing and Information Sciences, 9(1), 24.

[31] Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining, 13(1), 30.

[32] Z. Ul Rehman, S. Abbas, M. Adnan Khan, G. Mustafa, H. Fayyaz et al., "Understanding the language of isis: an empirical approach to detect radical content on twitter using machine learning," Computers, Materials & Continua, vol. 66, no.2, pp. 1075–1090, 2021.

[33] Sowmya, B. J., Hanumantharaju, R., Kumar, D. P., & Srinivasa, K. G. (2023). Identification of authorship and prevention of fraudulent transactions/cybercrime using efficient high performance machine

learning techniques. International Journal of Business Intelligence and Data Mining, 22(1-2), 144-169.

[34] Marinho, R., & Holanda, R. (2023). Automated Emerging Cyber Threat Identification and Profiling Based on Natural Language Processing. IEEE Access.

[35] Ferrara, E. (2017). Contagion dynamics of extremist propaganda in social networks. Information Sciences, 418, 1-12.

[36] Sharif, W., Mumtaz, S., Shafiq, Z., Riaz, O., Ali, T., Husnain, M., & Choi, G. S. (2019). An Empirical Approach for Extreme Behavior Identification through Tweets Using Machine Learning. Applied Sciences, 9(18), 3723.

[37] Salleh, N. S. M., Suliman, A., & Ahmad, A. R. (2011, November). Parallel execution of distributed SVM using MPI (CoDLib). In ICIMU 2011: Proceedings of the 5th international Conference on Information Technology & Multimedia (pp. 1-4). IEEE.

[38] Salleh, N. S. M., Suliman, A., & Jørgensen, B. N. (2020, August). A systematic literature review of machine learning methods for short-term electricity forecasting. In 2020 8th International conference on information technology and multimedia (ICIMU) (pp. 409-414). IEEE.

[39] Ahmad Sh., Asghar M., Alotaibi F., Awan I. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. 2019

[40] Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. Big data and cognitive computing, 7(1), 15.

[41] Scrivens, R., & Frank, R. (2016, August). Sentiment-based classification of radical text on the web. In 2016 European Intelligence and Security Informatics Conference (EISIC) (pp. 104-107). IEEE.

[42] A. A. Fahoum and T. A. Ghobon, "Accurate machine learning predictions of sci-fi film performance," Journal of New Media, vol. 5, no.1, pp. 1–22, 2023.

[43] Ji, S., Yu, C. P., Fung, S. F., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. Complexity, 2018.

[44] Sultanovich, O. B., Ergeshovich, S. E., Duisenbekovich, O. E., Balabekovna, K. B., Nagashbek, K. Z., & Nurlakovich, K. A. (2016). National Sports in the Sphere of Physical Culture as a Means of Forming Professional Competence of Future Coach Instructors. Indian Journal of Science and Technology, 9(5), 87605-87605.

[45] Kaldarova, B., Omarov, B., Zhaidakbayeva, L., Tursynbayev, A., Beissenova, G., Kurmanbayev, B., & Anarbayev, A. (2023). Applying Game-based Learning to a Primary School Class in Computer Science Terminology Learning. In Frontiers in Education (Vol. 8, p. 26). Frontiers.

[46] Gaikwad, M., Ahirrao, S., Kotecha, K., & Abraham, A. (2022). Multi-Ideology Multi-Class Extremism Classification Using Deep Learning Techniques. IEEE Access, 10, 104829-104843.

[47] Devyatkin D., Smirnov I., Ananyeva M., Kobozeva M. Exploring linguistic features for extremist texts detectyion (on the material of Russian-speaking illegal texts). 2016

[48] N. Mahmood and M. Usman Ghani Khan, "Prediction of extremist behaviour and suicide bombing from terrorism contents using supervised learning," Computers, Materials & Continua, vol. 70, no.3, pp. 4411–4428, 2022.