DR. B. R. AMBEDKAR NATIONAL INSTITUTE OF TECHNOLOGY, JALANDHAR

EVENT TREND DETECTION IN EXTREMIST DISCUSSIONS ON DARK WEB

Minor Project under supervision of

Dr. Jaspal Kaur Saini (Assistant Professor - IT Dept.)

PRESENTED BY:

Vidur Dua (23124119)
Shivansh Dutta (23124102)
Vaibhav Kanda (23124113)
Karman Singh (22124054)

INDEX

INTRODUCTION

- **A Digital World is Full of Hidden Threats: The Dark Web.**

  The anonymity provided by encrypted networks like Tor has created a critical challenge. This concealed environment is now the primary ground for extremist coordination and large-scale cybercrime.

- **Why Traditional Monitoring is Inadequate.**

  1. **Manual Monitoring**: Impossible due to the volume, multi-lingual nature, and transient state of Dark Web forums.
  2. **Surface Web Tools**: Cannot access the hidden .onion domain structure, leaving 99% of criminal planning unseen.
  3. **Keyword Search**: Ineffective, as threat actors use coded language and constantly shift channels to evade detection.

To develop an intelligent framework using **Machine Learning (NLP)** to automatically detect and analyze threat patterns and discussion trends on the Dark Web.

WHAT IS DARK WEB?

- **Definition:** The Dark Web is a secluded and intentionally concealed segment of the Deep Web, deliberately excluded from standard search engine indexing.
- **Access Requirement:** Requires mandatory use of specialized anonymizing software, primarily Tor (The Onion Router).
- **Anonymity Mechanism:** Relies on multi-layered encryption (Onion Routing), ensuring high anonymity for users and site operators.
- **Key Function:** Serves as a primary hub for illicit activities and secure, confidential communication.

SURFACE VS DARK WEB?

Surface Web

- Surface Web (Clearnet)
- Indexed and searchable by all engines.
- Low; user activity and IP are easily traceable
- (≈ 4% to 5% of the total web)
- Public news, retail, social media

Dark Web

- Dark Web (Darknet)
- Not Indexed; intentionally hidden
- High traffic is encrypted and relayed.
- A tiny fraction (0.1% of the total web).
- Black markets, cybercrime forums, protected content.

WORKING OF THE DARK WEB

Core Principle (The "Onion"): User data is encapsulated in multiple, nested layers of encryption before transmission.

Circuit Construction: The data travels through a randomly selected, decentralized path (Circuit) of at least three volunteer servers (nodes):

- Entry Node: Sees the user's IP, but only the next node's identity.
- Middle Node: Strips one layer of encryption, but knows neither the source nor the final destination.
- Exit Node: Sees the final destination's address, but only knows the Middle Node's identity.

Decentralized Anonymity: Each node only decrypts its dedicated layer, ensuring no single server knows both the origin and the final destination, guaranteeing strong anonymity

ACTIVITIES/USES ON DW

- **Terrorism & Radicalization**: Used as a "virtual safe haven" for encrypted planning, coordination, and the secure distribution of radical propaganda and fundraising efforts.

- **Organized Cybercrime**: The primary marketplace for exploiting data breaches, trading stolen PII, financial credentials, and corporate access. This includes Cyber-Services (CaaS) like Ransomware kits, hacking tools, and botnets.

- **Illegal Market Trade**: Facilitates global transactions via Darknet Markets for physical illicit goods, including narcotics, illegal firearms, and counterfeit identification documents.

- **Legitimate Applications**: The same anonymity is critical for legal entities:

- **Media & Activism**: Provides secure communication for whistleblowers and journalists to evade censorship.

- **Government & Law Enforcement**: Used by intelligence agencies and cyber defense teams for threat intelligence gathering and conducting covert surveillance and sting operations.

PROBLEM STATEMENT

- The Dark Web serves as a hub for extremist and criminal activities, including illegal weapon procurement, terrorist coordination, and violent propaganda.

- Monitoring such content manually is extremely difficult due to its anonymous, multilingual, and unstructured nature.

- There is a need for an automated system that can:
    1. Detect and classify violent or illegal discussions, and
    2. Analyze their temporal patterns to identify spikes or emerging threats.

- **Goal**: Develop a framework that detects, classifies, and tracks violent activities and discussion trends over the Dark Web.

APPROACH

- Data Collection and Preprocessing
  - Extract multilingual text data (English, Arabic) from Dark Web forum datasets.
  - Perform cleaning, language detection, and translation using deep-translator.
  - Extract timestamps for trend analysis
- Detection and Classification
  - Use hybrid NLP and deep learning models:
    - CNN and BiLSTM for sequence modeling
    - Transformer-based models (BERT or mBERT) for multilingual understanding
  - Classify posts as:
    - Normal discussions
    - Violent/extremist content
    - Illegal weapon trade
- Trend Analysis and Interpretation
  - Aggregate posts over time and detect spikes or anomalies using:
    - Statistical models (Z-score, Prophet)
    - Anomaly detection (Isolation Forest)
  - Visualize keyword trends and spikes using Streamlit dashboards.

DATASET – DARK WEB FORUM DATA (UNIVERSITY OF ARIZONA)

- **Dataset Source: Dark Web Project – University of Arizona**
- **Developed under the Artificial Intelligence Lab, containing data from:**
  - Jihadist and extremist forums
  - Dark Web discussion boards
  - Archived forum posts related to terrorism, recruitment, propaganda, and weapons
- **Data includes:**
  - Post content (text)
  - User details (anonymized)
  - Timestamps
  - Thread and topic metadata
- **Supports multilingual analysis (mainly English and Arabic)**

APPLICATIONS & ADVANTAGES

- **Applications:**
  - Cybersecurity & Intelligence: Detect threats, illegal weapon trading, and extremist planning.
  - Law Enforcement: Monitor suspicious discussions for early intervention.
  - Research: Analyze online radicalization trends and communication behavior.
  - Policy Making: Provide insights to counter propaganda and misinformation.
- **Advantages:**
  - Handles multilingual content (Arabic + English)
  - Provides real-time trend insights using temporal analysis
  - Scalable and adaptable to different forums or social media platforms
  - Visual dashboard for easy interpretation by non-technical users

CHALLENGES

- **Data Accessibility:** Difficult due to ethical and legal restrictions.
- **Multilingual Complexity:** Handling different languages, dialects, and mixed scripts.
- **Labeling Difficulty:** Manual annotation is time-consuming.
- **Model Interpretability:** Deep learning models can be hard to explain.
- **Data Imbalance:** Violent content is rare compared to normal posts.
- **Real-Time Analysis:** Processing large, continuous streams is challenging.

CONCLUSION & FUTURE WORK

- Conclusion:
  - Built a framework to detect and track violent or illegal activities.
  - Integrated classification and temporal trend analysis.
  - Demonstrated multilingual capability and visualization.
- Future Work:
  - Integrate real-time Dark Web crawlers.
  - Extend analysis to multimedia content (images, videos, voice).
  - Use graph-based social network analysis for influencer detection.
  - Deploy as a cloud-based alert system for real-time threat detection

REFERENCES

- **Dark Web Forums Portal:** Searching and Analyzing Jihadist Forums – University of Arizona.

- **Sentiment Analysis in Multiple Languages:** Feature Selection for Opinion Classification in Web Forums – Khalid et al.

- **Dark Web Geo-Web Research Project:** Artificial Intelligence Lab, University of Arizona (Dataset Source).

- **LSTM based deep learning approach to detect online violent activities over dark web** - Jaspal Kaur Saini
- **A Comparative Study and Automated Detection of Illegal Weapon Procurement over Dark Web** - Jaspal Kaur Saini & Divya Bansal