BERTopic Classification Workflow

This document outlines the steps and findings from applying BERTopic, a topic modeling approach based on BERT embeddings, to patent abstracts related to cyberbullying and cyber safety. Unlike LDA, which relies on bag-of-words representations, BERTopic leverages contextual embeddings to generate more meaningful clusters, particularly effective for short, technical texts like patents.

Why BERTopic?

Because our LDA models produced relatively low coherence scores, we adopted BERTopic for richer semantic analysis. BERTopic combines BERT embeddings, UMAP for dimensionality reduction, and HDBSCAN for clustering, followed by class-based TF-IDF to assign interpretable topics. Compared to LDA:

- Embedding Type: Bag-of-Words (LDA) vs. Contextual (BERT)
- Semantic Handling: Limited (LDA) vs. Strong (BERTopic)
- Output Quality: Moderate for long texts (LDA) vs. High for short texts (BERTopic)

Step 1: Install Dependencies

We installed required Python packages including BERTopic, transformers (for BERT embeddings), UMAP, and HDBSCAN.

```
!pip install bertopic
!pip install openpyxl
!pip install umap-learn
!pip install nltk
```

Step 2: Load Dataset

Patent abstracts were loaded from an Excel file into a pandas DataFrame for analysis.

```
from google.colab import files
uploaded = files.upload()  # upload your 'Abstracts Only.xlsx'

import pandas as pd
df = pd.read_excel("Abstracts Only.xlsx", engine="openpyxl")
df = df.dropna(subset=["Abstract"])
texts = df["Abstract"].tolist()
```

Step 3: Text Cleaning

We preprocessed abstracts by removing numbers, punctuation, and stopwords, converting text to lowercase, tokenizing, and lemmatizing. The cleaned text was added as a new column in the dataset.

```
stop_words = set(stopwords.words("english"))

lemmatizer = WordNetLemmatizer()


def clean_text(text):
    text = re.sub(r'\d+', '', text)                  # remove numbers
    text = re.sub(r'[^\w\s]', '', text)              # remove punctuation
    text = text.lower()                              # lowercase
    tokens = nltk.word_tokenize(text)                # tokenize
    tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in stop_words and len(word) > 2]
    return " ".join(tokens)


df['cleaned_abstract'] = df['Abstract'].astype(str).apply(clean_text)

df[['Abstract', 'cleaned_abstract']].head()
```

Step 4: Fit BERTopic

BERTopic was applied on the cleaned abstracts. It generated embeddings using BERT, reduced them with UMAP, clustered with HDBSCAN, and assigned topics using class-based TF-IDF.

```
from bertopic import BERTopic
from sklearn.feature_extraction.text import CountVectorizer

texts = df['cleaned_abstract'].tolist()

vectorizer_model = CountVectorizer(stop_words="english")
topic_model = BERTopic(vectorizer_model=vectorizer_model, min_topic_size=5)
topics, probs = topic_model.fit_transform(texts)
```

Step 5: Review Results

The model produced multiple clusters of patents, each characterized by distinctive keywords. These clusters were then interpreted for relevance to cyberbullying and cyber safety innovations.

```
# Topic assignments for each abstract
df['Topic'] = topics
df[['Abstract', 'Topic']].head()

# Summary of discovered topics
topic_model.get_topic_info()
```

Step 6: Save Results

Cluster assignments and associated keywords were saved into an Excel file for further analysis and reference.

```
df.to_csv("bertopic_output.csv", index=False)
```

Topic Analysis and Interpretation

BERTopic identified several coherent clusters:

• Topic 0 (38 patents): Focused on NLP-based methods for text/content abuse classification, particularly cyberbullying detection. Strong alignment with social platform safety.

• Topic 1 (32 patents): Abuse detection systems integrating hardware and software (e.g., IoT or mobile devices). Moderate relevance to cyber safety, with context-specific applications.

• Topic 2 (28 patents): Digital threat detection and cybersecurity systems leveraging machine learning and server-side data. High relevance at infrastructure level, though less focused on specific user groups.

• Topic -1 (miscellaneous): Generic or less interpretable patents, often containing broad keywords like 'user,' 'data,' and 'system.'

Insights Summary

BERTopic offered clearer and semantically richer clusters than LDA, making it especially effective for short patent abstracts. Results highlighted three major themes:

1. Text/content-based abuse classification (high relevance)
2. Device-based abuse detection systems (medium relevance)
3. Digital threat and ML-based cybersecurity (high relevance)

This approach provided robust insights into how AI-driven techniques are applied to safeguard online environments, particularly in contexts tied to marginalized user groups.

RESULTS APPENDIX

We clustered 4 clusters:

TOPIC 0 CLUSTER (38 patents):

These are the keywords:

TOPIC 0: [('content', np.float64(0.0839897355580018)), ('method', np.float64(0.057349442460392225)), ('one', np.float64(0.056093718625452815)), ('model', np.float64(0.054685475425638126)), ('sentence', np.float64(0.0541735602949227)), ('text', np.float64(0.04976874137085101)), ('cyberbullying', np.float64(0.04908585848146668)), ('data', np.float64(0.04660558010889914)), ('using', np.float64(0.04460444888202159)), ('set', np.float64(0.0444798061266541))]

Number of patents: 38

Interpretation of Cluster 0

Topic 0 (38 patents) centers on **NLP-based methods** for identifying harmful or abusive content, particularly **cyberbullying**. It includes innovations that use machine learning models to classify sentences or online text. This cluster likely covers patents on automated moderation, cyberbullying detection, and AI-based analysis of online communication — all of which are highly relevant to your research question on software-based cyber safety for marginalized communities.

TOPIC 1 CLUSTER (32 patents)

THESE ATR the keywords for topic 1

: [('abuse', np.float64(0.10054381744260915)), ('system', np.float64(0.06636400235038381)), ('device', np.float64(0.058857901926456416)), ('detection', np.float64(0.05629193390810848)), ('data', np.float64(0.05558446043515982)), ('method', np.float64(0.04686543371578028)), ('electronic', np.float64(0.045437627998194625)), ('application', np.float64(0.04255445887119683)), ('user', np.float64(0.03875362396505096)), ('module', np.float64(0.036573485945162146))]

Thematic Summary for Topic 1

Topic 1 (32 patents) is focused on **abuse detection systems** involving both **hardware (devices, modules)** and **software applications**. While "abuse" is a key term, it may not exclusively refer to **cyberbullying or online abuse** — it could also include fraud, device misuse, or inappropriate content. However, the consistent presence of "user", "data", "application", and "detection" suggests a **technological approach to monitoring and mitigating harmful user behavior**, possibly in online or mobile environments.

TOPIC 2 CLUSTER (32 patents)

here are the keywords for topic 2: [('threat', np.float64(0.09783836797308708)), ('digital', np.float64(0.08170756968748116)), ('data', np.float64(0.06745566160949659)), ('security', np.float64(0.06176139260263769)), ('online', np.float64(0.056459351490327055)), ('event', np.float64(0.05584211423913612)), ('machine', np.float64(0.05531987240438722)), ('information', np.float64(0.04996860287609146)), ('server', np.float64(0.049482940627315536)), ('based', np.float64(0.04902996950679488))]

Thematic Summary for Topic 2

Topic 2 (28 patents) focuses on **AI-powered digital threat detection**, with strong emphasis on cybersecurity, data protection, and real-time event monitoring. The use of terms like “machine,” “server,” “event,” and “based” points to software engineering solutions for **automated threat detection**, potentially involving behavioral analysis or anomaly detection on servers or networks. This topic is **highly relevant** to cyber safety, though not necessarily centered on *marginalized communities* unless specific user groups are addressed in individual patents.

Topic Comparison:

Topic	Key Theme	Relevance to Cyber Safety for Marginalized Users
0	Text/content-based abuse classification	High (esp. for social platforms)
1	Device-based abuse detection systems	Medium (context-specific)
2	Digital threat and ML-based security	High (infrastructure level, less focused on user groups)

Topic -1_user_data_system_device (remaining patents)

- **Keywords:** user, data, system, device, may, one, method
- **Interpretation:** Likely a **generic cluster** — could be anything from network tools to generic user interfaces.
- Often the -1 topic is **miscellaneous**