Research Paper Topic Clustering

Jaworski Norbert - 474891

# BERTopic

### Custom HDBSCAN

A screen shot of a computer screen

Description automatically generated

### Ultimate Old Stop Words

A black screen with white text

Description automatically generated

### New stop words

A blurry image of a computer screen

Description automatically generated

    # Custom HDBSCAN model

    hdbscan\_model = HDBSCAN(min\_cluster\_size=15, min\_samples=10, metric='euclidean', cluster\_selection\_method='eom')

    # Apply BERTopic

    topic\_model = BERTopic(hdbscan\_model=hdbscan\_model)

    topics, probs = topic\_model.fit\_transform(combined\_texts, combined\_embeddings)

    topic\_model.update\_topics(combined\_texts, top\_n\_words=7)

### UMAP + HDBSCAN

    vectorizer = TfidfVectorizer(ngram\_range=(1, 2), min\_df=0.02, max\_df=0.8, max\_features=10000, stop\_words="english")

    umap = UMAP(n\_neighbors=n\_neighbors, min\_dist=min\_dist, n\_components=n\_components, random\_state=42)

    labels, tfidf\_matrix, vectorizer, reduced\_data = cluster\_with\_hdbscan(data, n\_neighbors=10, min\_dist=0.1, n\_components=50)