BullyShield

A Machine Learning Approach to Combat Online Bullying

Team Members:

1. Kazi Israrul Karim
2. Eshtiak Alam Shihab
3. Md Shamsur Shafi Nur E Aziz

Instructors:

1. Zadid Hasan
2. Labib Hasan Khan

Purpose of ***BullyShield:***

"BullyShield" employs AI and advanced NLP to swiftly detect and counteract online bullying. Utilizing cutting-edge machine learning models, it offers a faster and more accurate means of distinguishing harmful behavior, demonstrating transformative potential for safer and more positive digital interactions across diverse online communities.

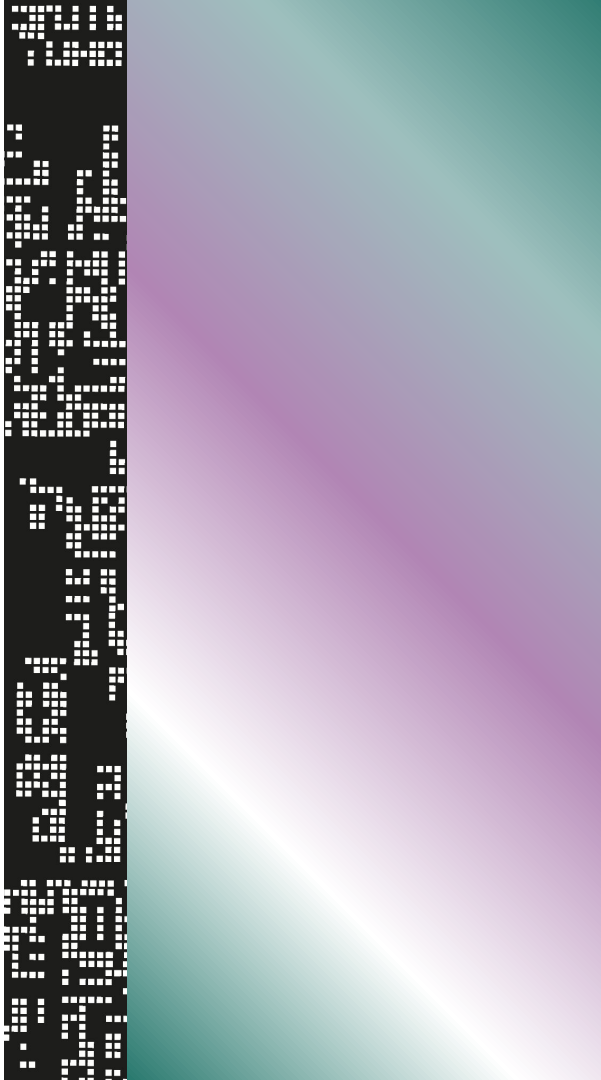


Table of contents

01 DATA PREPROCESSING

Essential for Accurate and Meaningful Natural Language Processing (NLP) Results.

Step: 01

Objective: Removing Duplicates and Null Values

- Reason: Duplicates and null values can distort analysis and lead to biased results.

Before:

ID	Text	Category	Rating
1	Hello, world!	A	4.5
2	Null text	B	3.0
3	Hello, world!	A	4.5
4	Another message	C	2.0
5	Null text	B	3.5

After:

ID	Text	Category	Rating
1	Hello, world!	A	4.5
4	Another message	C	2.0

Step: 02

Objective: Removing Punctuation and Special Characters

- Reason: Punctuation and special characters are often extraneous and can introduce noise in text analysis

Before:

ID	Text
1	Hello, world! This is an example.
2	Special characters can be problematic!@#$%^&*
3	Let's clean this text, shall we?

After:

ID	Text
1	Hello world This is an example
2	Special characters can be problematic
3	Lets clean this text shall we

Step: 03

Objective: Removing Stop words and Least Frequent Words

Stopwords:

- Excluding common words like "and" or "the" focuses on essential content by removing text elements with minimal contribution.

Least Frequent Words:

- By eliminating words with low frequencies, noise is reduced, and the significance of remaining terms is enhanced for better analysis..

Before:

ID	Text
1	Hello, world! This is an example of text data.
2	Common words like "and" can be removed.
3	Infrequent terms may not contribute much.
4	Repeating stopwords and uncommon words.

After:

ID	Text
1	Hello, world! This example text data
2	Common words removed
3	Infrequent terms may contribute much
4	Repeating stopwords uncommon words

Step: 04

Objective: Removing URLs and Emojis

URLs:

- Reason: URLs add noise and are often irrelevant to the text's meaning.

Emojis:

- Reason: Emojis can be subjective and lack standardized meanings.

Before:

ID	Text
1	Check out my latest blog post at https://example.com!
2	Having a great day! 😊
3	Exciting news on our website: 🎉 https://news.example.com
4	Emojis can add a fun touch to messages! 😍 👍

After:

ID	Text
1	Check out my latest blog post at
2	Having a great day!
3	Exciting news on our website:
4	Emojis can add a fun touch to messages!

Step: 05

Lemmatization:

- Objective: Reducing words to their base or dictionary form (lemma).
- Example: "running" -> "run," "better" -> "good."
- Result: Lemmatized words are valid words and provide a more meaningful representation

Before:

ID	Text
1	Running in the park is a great exercise.
2	The quick brown foxes are jumping.
3	Better late than never.

After:

ID	Text
1	Run in the park be a great exercise.
2	The quick brown fox be jump.
3	Good late than never.

Step: 05

Objective: Label Encoding

- Reason: Converting categorical labels to numerical format enables machine learning models to interpret and learn from the data

Before:

ID	Category
1	Apple
2	Banana
3	Orange
4	Banana
5	Apple

After:

ID	Category
1	0
2	1
3	2
4	1
5	0

Step: 06

Objective: Feature Extraction and Scaling (**TF/IDF** and **BoW**)

- Reason: Converting categorical labels to numerical format enables machine learning models to interpret and learn from the data

Before:

ID	Category
1	Apple
2	Banana
3	Orange
4	Banana
5	Apple

After:

ID	Category
1	0
2	1
3	2
4	1
5	0

03 Model Evaluation

Analyzing Accuracy, Recall, and Classification Metrics Across Multiple Models

Logistic Regression

- Utilizes a logistic function to model binary outcomes and can be extended to multiclass classification using techniques like one-vs-rest (OvR) or multinomial logistic regression.
- Coefficients can be interpreted to understand the importance of each feature.
- Can be regularized (L1, L2) to address overfitting and handle high-dimensionality.

Benefits: Efficiently handles large datasets and multiclass problems. Often applied in spam detection and sentiment analysis.

LogReg: Accuracy

TFIDF	Accuracy: 83.0%
BOW	Accuracy: 82.0%

LogReg: Evaluation

TFIDF

- **High accuracy and balanced performance** (83%): effective in correctly predicting labels; balanced precision, recall, and f1-score
- **Strength** in Identifying Age, Ethnicity, and Religion: precision and recall scores in "Age," "Ethnicity," and "Religion" categories (all above 0.93)
- **Balanced discrimination** in Gender and Not Cyberbullying categories: Balanced precision and recall in "Gender" (0.91 and 0.84) and "Not Cyberbullying" (0.60 and 0.56); false positives minimized
- **Effective handling** of Other Cyberbullying category: Achieved balanced precision and recall (0.59 and 0.65)
- **Consistent macro and weighted averages** (82%): indicating the model's reliability and balanced performance across diverse categories

LogReg: Evaluation

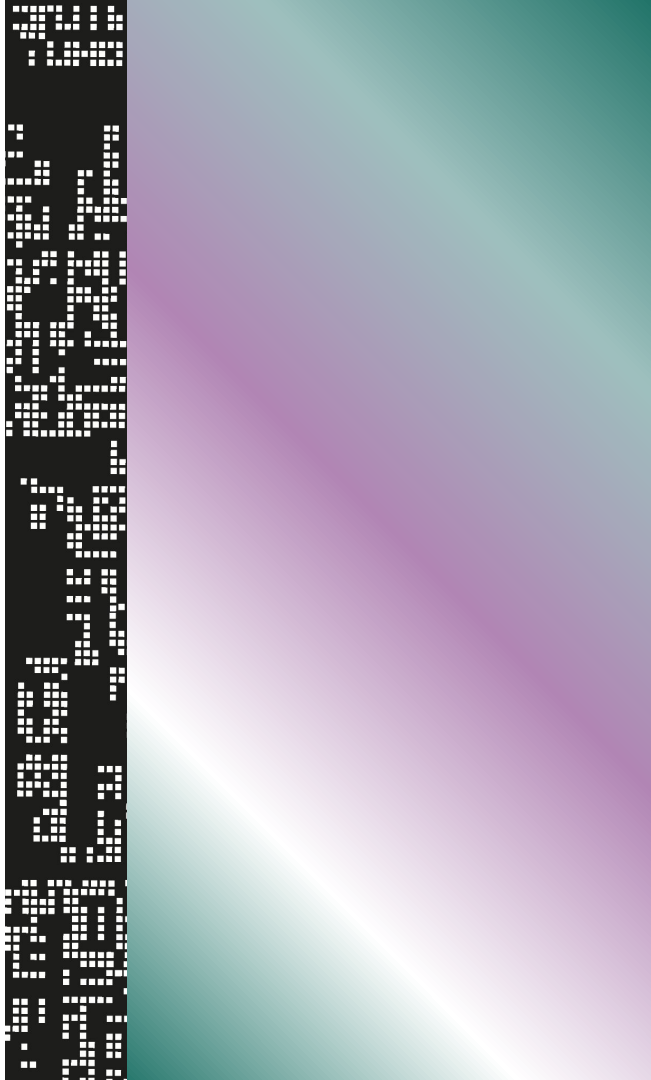
BOW

- **High Overall Accuracy** (82%)
- Consistent **high precision** in key categories: "Age" (97%), "Ethnicity" (98%), and "Religion" (97%)
- **Balanced discrimination** in "Gender" and "Not Cyberbullying": balanced precision and recall in "Gender" (91% and 85%) and "Not Cyberbullying" (57% and 57%); false positives minimized
- **Effective handling** of "Other Cyberbullying" category: balanced precision and recall (58% and 64%) in "Other Cyberbullying" category
- **Consistent macro and weighted averages** (83%): indicating the model's reliability and balanced performance across diverse categories.

KNN for Classification

- Predicts the class of a data point based on the majority class of its 'K' nearest neighbors.
- Distance metrics (e.g., Euclidean, Manhattan) measure proximity to neighbors.
- Chooses 'K' for the number of neighbors to consult for determining the class.
- Sensitive to the scale of data; preprocessing like normalization is often required.

Benefits: Simple and intuitive, no assumption on data distribution, effective for small datasets.



KNN: Accuracy

TFIDF

Accuracy: 27.2%

BOW

Accuracy: 70.4%

KNN: Evaluation

TFIDF

- **Challenges** in "Religion", "Age" and "Other Cyberbullying" categories: low recall in "religion" (14%), "age" (26%) and "other_cyberbullying" (5%)
- **Limited discrimination** in "Gender" and "Ethnicity" categories: precision-recall trade-off in "gender" (16% precision, 0.27 f1-score) and "ethnicity" (8% precision, 0.14 f1-score) indicates difficulty in distinguishing instances
- **Imbalanced prediction** for "Not Cyberbullying": High recall (96%) but low precision (18%) suggests a high rate of false positives
- **Overall** limited accuracy: model's overall accuracy is low (27%)

KNN: Evaluation

BOW

- **High precision** in "Religion" Category (94%)
- **Challenges** in "Age" and "Gender" Categories: Precision and recall are relatively high for "age"; recall for "gender" is 72%
- **Low precision** in "Not Cyberbullying" (38%) and "Ethnicity" Categories (41%): higher rate of false positives so needs improvement
- **Imbalanced performance** in "Other Cyberbullying": High precision (99%) but lower recall (71%) in "other_cyberbullying" suggests a need to balance model performance for this category
- **Overall** moderate performance: moderate overall accuracy (70%) and balanced macro and weighted averages for precision, recall, and f1-score (around 72-75%)

Naive Bayes

- Applies Bayes' theorem with the "naive" assumption of independence between every pair of features.
- Can handle discrete and continuous data by choosing appropriate distributions (e.g., Bernoulli, Gaussian).
- Works well with high-dimensional datasets and is computationally efficient.

Benefits: Efficiently handles large datasets and multiclass problems. Often applied in spam detection and sentiment analysis.

Naive Bayes: Accuracy

TFIDF	Accuracy: 76.0%
BOW	Accuracy: 77.0%

Naive Bayes: Evaluation

TFIDF

- **High Accuracy (76%)**
- **Challenges** in "Not Cyberbullying" category: Lower precision (67%) and recall (35%)
- **Balanced discrimination** in "Age," "Ethnicity," and "Gender" categories: balance between precision and recall in "Age" (70% precision, 99% recall), "Ethnicity" (84% precision, 90% recall), and "Gender" (83% precision, 86% recall)
- **Moderate performance** in "Other Cyberbullying" category: precision (65%) and recall (47%)
- **Consistent macro and weighted averages (73%)**: indicates the model's overall reliability and balanced performance across diverse categories

Naive Bayes: Evaluation

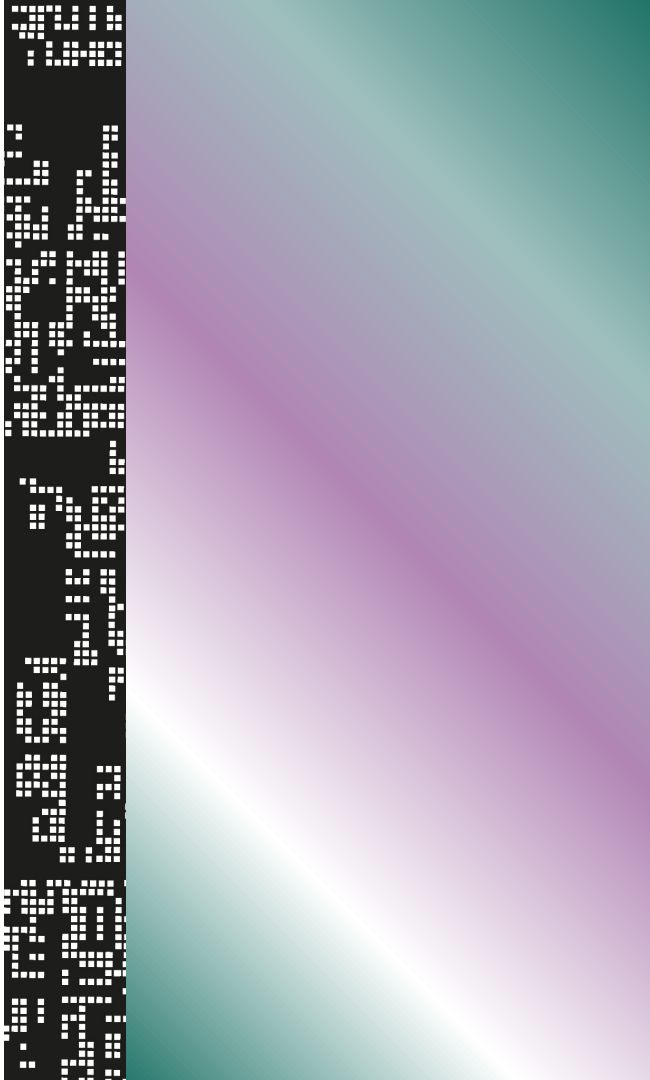
BOW

- **High accuracy (77%)**
- **Challenges** in "Not Cyberbullying" category: lower precision (68%) and recall (36%); challenges in effectively identifying instances in this group
- **Balanced performance** in key categories: precision and recall in key categories like "Age" (74% precision, 99% recall), "Ethnicity" (85% precision, 92% recall), and "Gender" (83% precision, 86% recall)
- **Moderate Improvement** in "Other Cyberbullying" Category:
    - Shows a slight improvement in precision (65%) and recall (51%) compared to the Naive Bayes (tf/idf)
- **Consistent macro and weighted Averages (75%)**: indicating the model's continued reliability and balanced performance across diverse categories.

SVM for Classification

- Seeks optimal hyperplane to separate classes with maximum margin.
- Uses support vectors to define the boundary.
- Employs kernel trick for non-linear data (e.g., Linear, RBF). Kernel functions handle non-linear text feature spaces.
- Kernel functions handle non-linear text feature spaces.

Benefits: Effective in high dimensions, memory efficient with kernel functions.



SVM: Evaluation

TFIDF

- SVM with TF-IDF achieves an **82% accuracy**, indicating a robust capability in identifying instances of cyberbullying.
- Noteworthy variations in precision and recall. High proficiency in discerning *age*, *ethnicity*, and *religion*-related bullying, but challenges in *not_cyberbullying* and *other_cyberbullying* categories.
- Macro and weighted average F1-scores, both at 0.82, underscore a balanced performance across diverse bullying categories.

SVM: Evaluation

BoW

- SVM with Bag-of-Words (BoW) achieves an 83% accuracy, indicating a slight improvement over the TF-IDF approach.
- Similar precision and recall trends. Strong performance in age, ethnicity, and religion-related bullying. Improvement in not_cyberbullying, but other_cyberbullying sees a notable boost.
- Macro and weighted average F1-scores maintain at 0.83, ensuring a consistent and balanced evaluation across diverse bullying categories.

Random Forest

- Ensemble method combining multiple decision trees.
- Each tree trained on random subsets of data with bootstrap aggregating.
- Handles high-dimensional text data effectively.
- Aggregates predictions to improve linguistic pattern recognition.

Benefits: Robust to overfitting and maintains accuracy with diverse data. Also useful for sentiment analysis and topic classification.

RF: Evaluation

TFIDF

- Random Forest with TF-IDF achieves an 82% accuracy, aligning closely with SVM with TF-IDF.
- Similar trends in precision and recall. Strong performance in *age*, *ethnicity*, and *religion*-related bullying. Challenges persist in *not_cyberbullying* and *other_cyberbullying* categories.
- Both F1-scores stand at 0.82, indicating a stable and balanced model performance across different bullying types.

RF: Evaluation

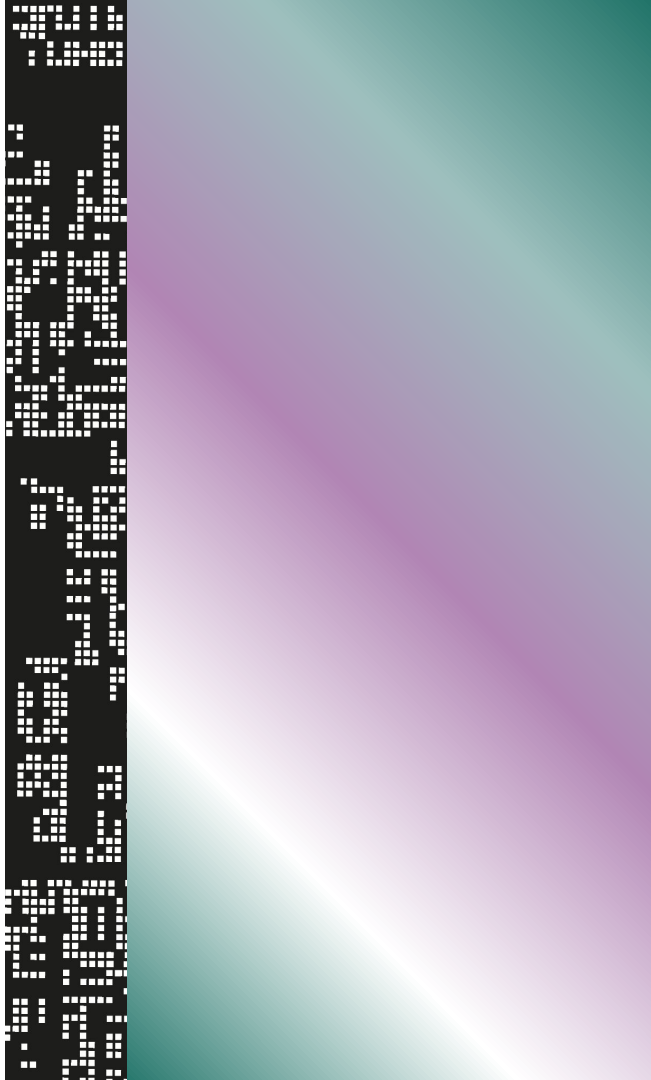
BoW

- Random Forest with Bag-of-Words (BoW) maintains an 81% accuracy, demonstrating a consistent performance across models.
- Similar precision and recall patterns. Strong identification of *age*, *ethnicity*, and *religion*-related bullying. Challenges persist in *not_cyberbullying* and *other_cyberbullying* categories, mirroring the TF-IDF counterpart.
- Macro and weighted average F1-scores remain at 0.81, indicating a reliable and balanced performance for diverse bullying types.

XGBOOST (eXtreme Gradient Boosting)

- Optimized gradient boosting library for performance and speed.
- Captures complex language patterns through sequential model refinement.
- Regularization controls overfitting, crucial for text data variance.
- Efficiently manages sparse data, common in text representations.

Benefits: Enhances performance in text classification and sentiment analysis tasks.



XGBoost: Evaluation

TFIDF

- XGBoost with TF-IDF achieves the highest accuracy at 84%, showcasing superior performance among the models.
- Strong precision and recall across various categories. Particularly noteworthy in accurately identifying *age*, *ethnicity*, and *religion*-related bullying.
- Macro and weighted average F1-scores stand at 0.84, emphasizing the consistent and superior performance of XGBoost in diverse bullying categories.

XGBoost: Evaluation

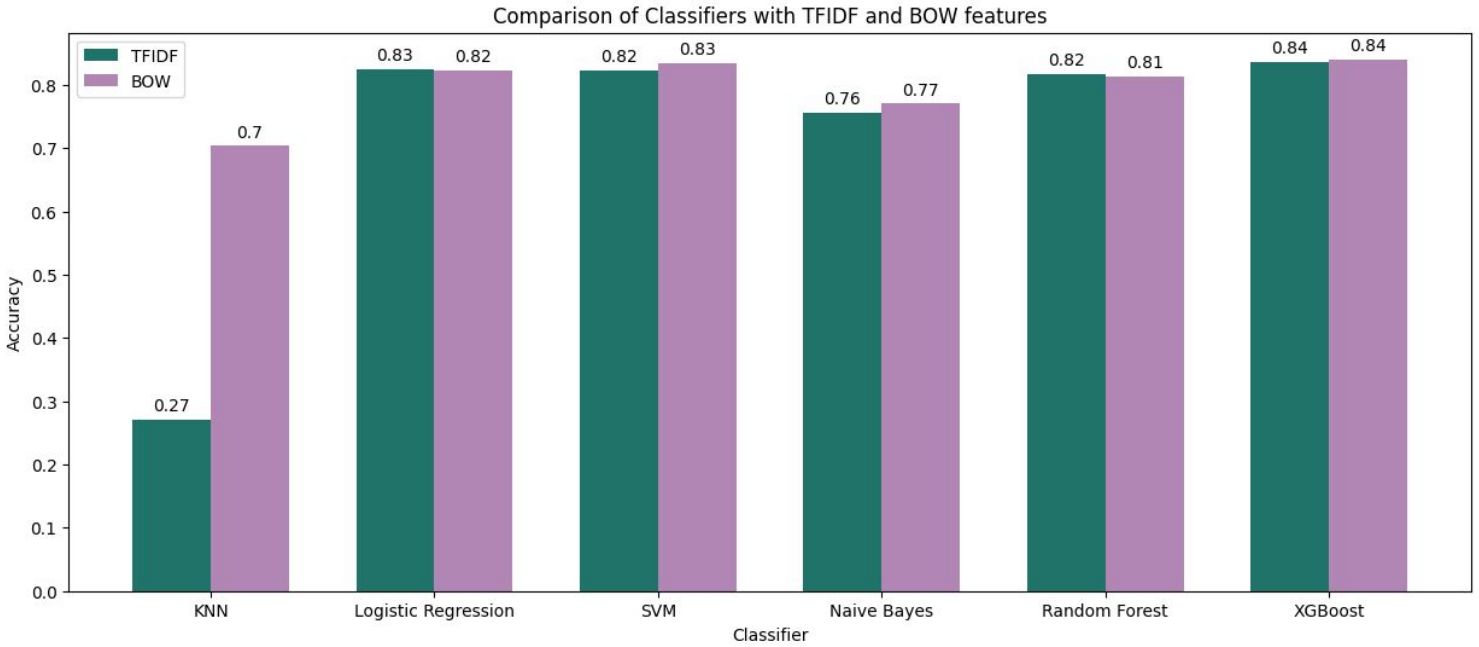
BoW

- XGBoost with Bag-of-Words (BoW) maintains a high accuracy of 84%, matching its TF-IDF counterpart.
- Strong precision and recall, particularly excelling in identifying age, ethnicity, and religion-related bullying. Challenges persist in not_cyberbullying, though an improvement is noted compared to the TF-IDF version.
- Both macro and weighted average F1-scores remain at 0.84, underlining the consistent and exceptional performance of XGBoost across diverse bullying categories.

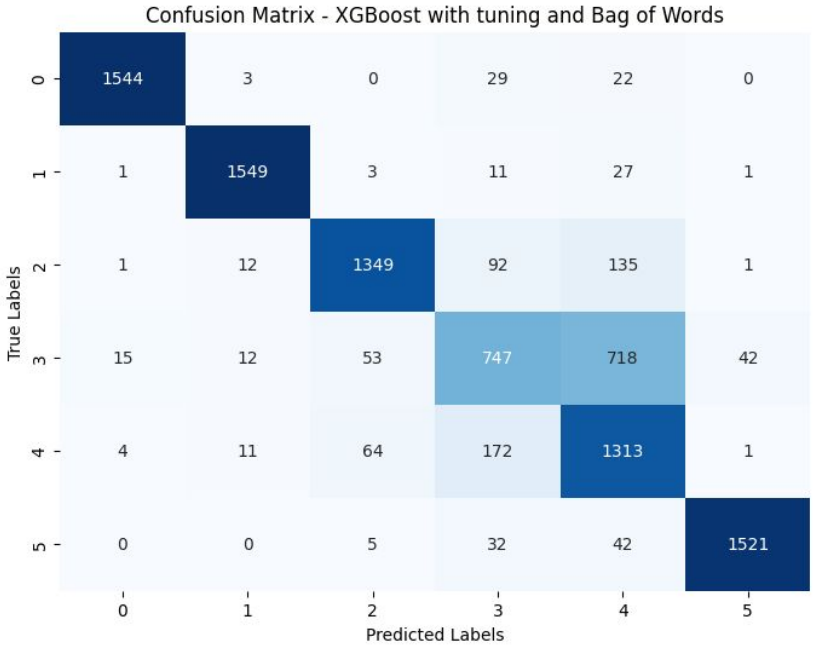
04 Model Selection and Decision Making

Informed Choices Based on Comparative Analysis

Comparison


Comparison of Classifiers with TFIDF and BOW features
TFIDF
BOW
Accuracy
Classifier
KNN
Logistic Regression
SVM
Naive Bayes
Random Forest
XGBoost
0.27
0.7
0.83
0.82
0.82
0.83
0.76
0.77
0.82
0.81
0.84
0.84
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Post Optimization for XGBoost with BoW



Thanks!

Presented By -
Kazi Israrul Karim
Md Shamsur Shafi Nur E Aziz
Eshtiak Alam Shihab