SIS-2207

Presentation

TOXIC COMMENT

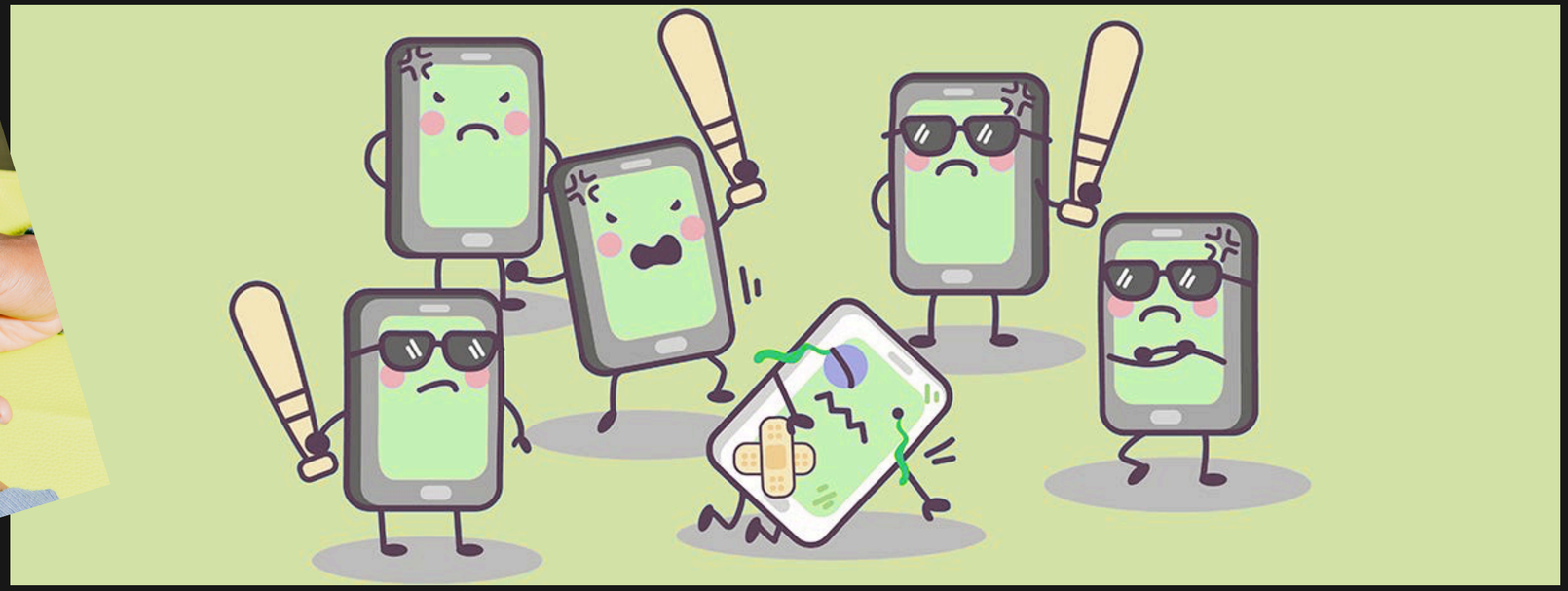
CLASSIFICATION DETECT HATE SPEECH AND ABUSIVE LANGUAGE ONLINE

Alexeyeva Alina
Yerbolova Amina
Tnasheva Zhanel

PROJECT AIM

Build a model that identifies:

- Toxic/abusive comments
- Hate speech & discrimination
- Threats and insults
- Obscene/offensive language



PROBLEM

- INCREASE OF TOXIC CONTENT IN SOCIAL MEDIA
- HARM TO MENTAL HEALTH & ONLINE SAFETY
- MANUAL MODERATION IS SLOW AND SUBJECTIVE
- NEED FOR AUTOMATED DETECTION SYSTEMS

Dataset

Overview

DATASET INFORMATION

/1 Dataset

Jigsaw Toxic Comment Classification (Kaggle)
Contains ~160,000 real user comments with assigned labels

/2 Categories include

toxic, severe toxic, obscene, threat, insult, identity hate

/3 Task type

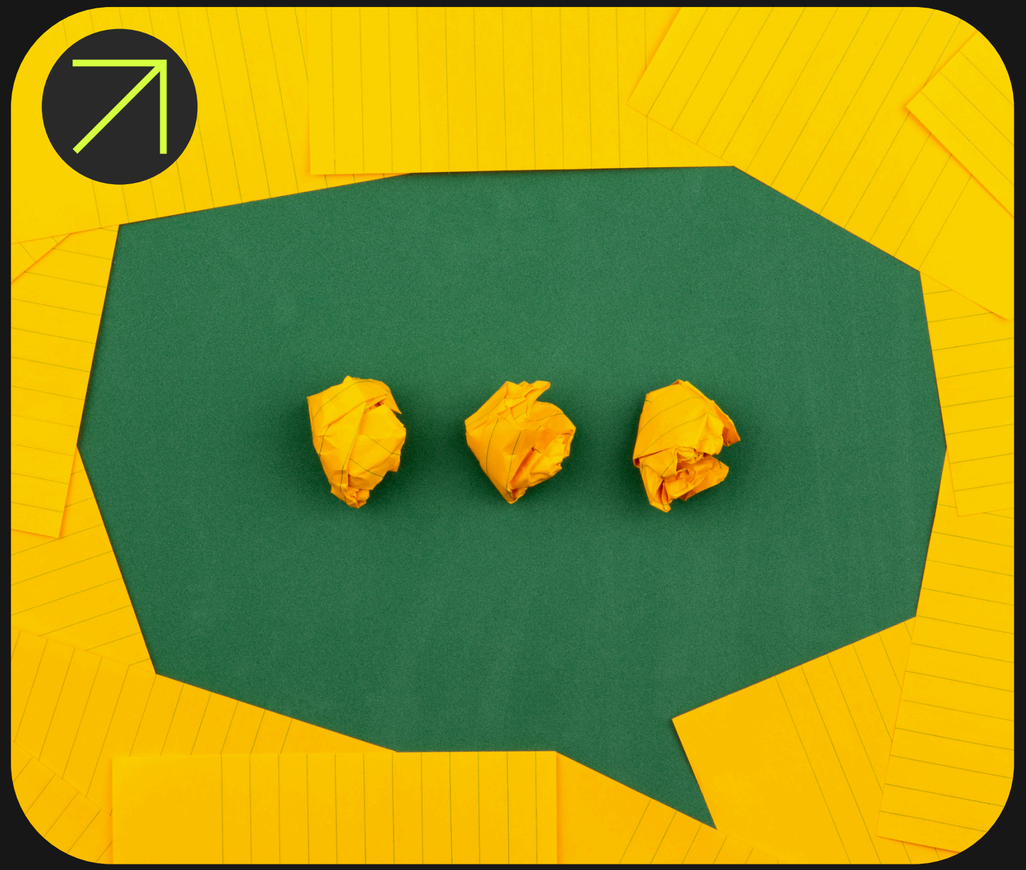
multi-label classification

PREPROCESSING STEPS

- Data cleaning and normalization of text.
- Lowercasing, removal of punctuation and unnecessary symbols.
- Tokenization and lemmatization to standardize words.
- Preparation of the dataset for model training.

EXPLORATORY DATA ANALYSIS

- Distribution of labels shows class imbalance.
- Some categories appear much more often than others.
- The dataset contains comments with multiple toxic labels at once.
- Insights help to choose suitable modeling techniques.

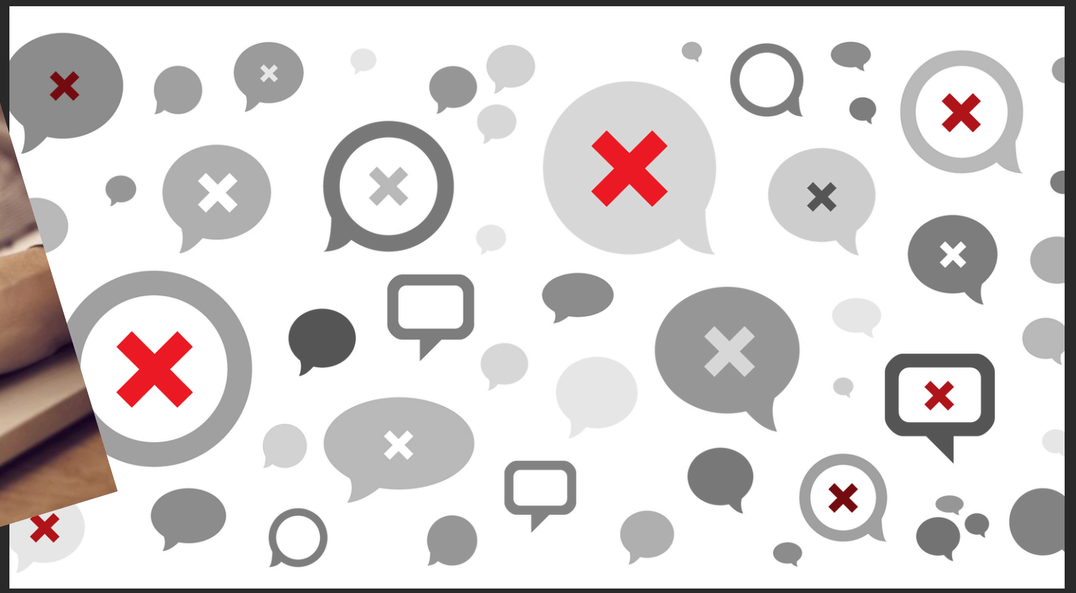


FEATURE REPRESENTATION

TF-IDF applied for traditional machine learning methods.
Tokenized word sequences and embeddings used for deep learning.
These representations allow models to interpret text meaningfully.

MODELS USED

/1 Baseline model: Logistic Regression with TF-IDF features.
/2 Advanced model: LSTM neural network for sequence understanding.

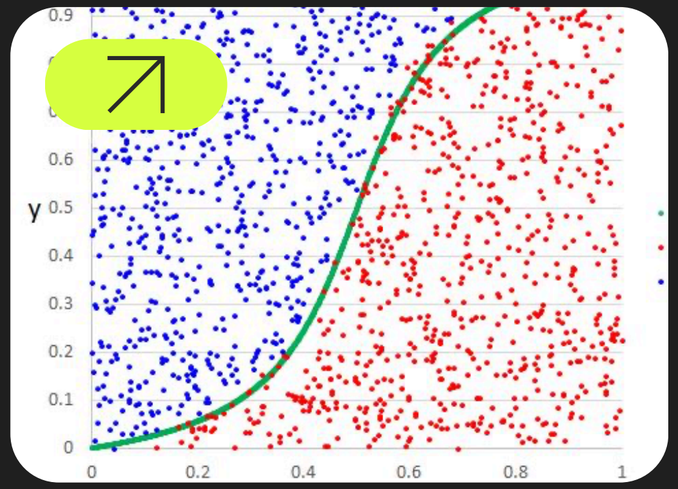


The comparison shows differences between classical ML and deep learning performance.

Evaluation

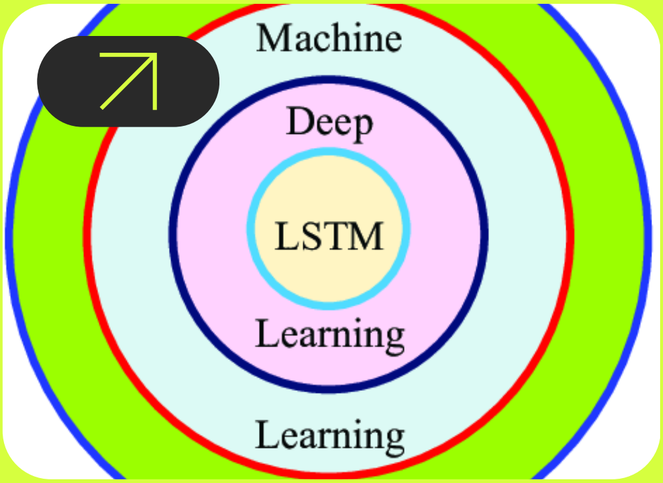
Results

EVALUATION & RESULTS



Logistic Regression

good for common labels,
limited context understanding



LSTM

better detection of rare and subtle toxic comments, higher overall accuracy

Challenges & Tools

Challenges

- Class imbalance impacts rare categories
- Sarcasm and implicit toxicity are hard to detect
- Dataset is English-only → limits multilingual application
- Deep learning requires more computational resources

Tools & Technologies

- Python, Pandas, NumPy — data processing
- Scikit-Learn — baseline ML models
- TensorFlow/Keras — LSTM and deep learning
- NLTK — text preprocessing

FUTURE WORK & CONCLUSION

/1 Future Improvements

- Use transformer models (BERT) for better context understanding
- Balance data and augment rare classes
- Extend to multilingual toxic comment detection
- Develop web interface/API for real-time moderation

/2 Conclusion

- Toxic comment classification system successfully developed
- LSTM outperformed the baseline model
- NLP and ML are effective tools for improving online safety

THANK YOU