Ivy Professional School

HATE SPEECH DETECTION

PROBLEM STATEMENT, VARIABLE DESCRIPTION & DELIVERABLES

Hate Speech Detection – Objective & Deliverables

Problem Description

SocialSafe, a rapidly growing social media platform dedicated to fostering safe and inclusive interactions, faces significant challenges in moderating harmful content, particularly hate speech, as its user base expands. Manual moderation is increasingly inefficient, inconsistent, and unsalable with the growing volume of user-generated content, posing risks to community safety and regulatory compliance. To address this, SocialSafe seeks to implement an automated machine learning pipeline to classify text into "hate speech" and "non-hate speech," enabling real-time detection and flagging of harmful content. This solution will streamline moderation efforts, ensure a positive user experience, and uphold the platform's commitment to creating a secure digital environment.

Steps: End-to-End Workflow

Data Cleaning:

Remove duplicates, missing entries, and irrelevant data.

Text Preprocessing:

- **Remove Special Characters and Numbers:** Strip unnecessary symbols to retain only text content.
- **Convert to Lowercase:** Standardize text for uniformity.
- **Tokenization:** Split text into individual words or tokens.
- **Remove Stopwords:** Eliminate common words (e.g., "and," "the") that do not contribute to the text's meaning.
- **Stemming/Lemmatization:** Reduce words to their root forms to ensure consistency (e.g., "running" -> "run").
- **Handle Spelling Variations:** Correct common misspellings or variations often used in hate speech.

IVY Professional School

© - ALL RIGHTS RESERVED - IVY PROFESSIONAL SCHOOL

Feature Engineering:

- Word Counts and Frequency: Analyze word frequency to understand common terms in hate speech.
- N-grams: Extract sequences of words (e.g., bigrams, trigrams) to capture contextual information.
- TF-IDF Vectorization: Transform the text data into numerical vectors based on Term Frequency-Inverse Document Frequency.

Model Building:

- Split the dataset into training and testing sets.
- Train and evaluate at least five models to classify hate speech:
  - Logistic Regression
  - Naive Bayes
  - Support Vector Machines (SVM)
  - Random Forest
  - XGBoost
- Optimize hyperparameters using techniques like Grid Search or Random Search.

Model Evaluation:

- Use metrics like precision, recall, F1-score, and accuracy to assess model performance.
- Analyze confusion matrices to understand false positives and false negatives.
- Select the best-performing model for deployment.

Ivy Professional School

© - ALL RIGHTS RESERVED - IVY PROFESSIONAL SCHOOL