summary & mcq

summaty

* **Text Classification Defined:** The process of assigning predefined categories (e.g., spam/ham, topic, sentiment) to text data. Examples include spam detection, sentiment analysis, and authorship identification.

* **Types of Text Classification:** Binary (two categories) and multi-class (more than two categories).

* **Text Classification Process:**
  * **Data Collection:** Labeled data is needed.
  * **Text Preprocessing:** Cleaning the text (removing numbers, special characters, converting to lowercase, removing stop words, normalization).
  * **Feature Extraction:** Transforming text into numerical features using methods like Bag-of-Words (BoW) or TF-IDF. Choosing relevant features is crucial.
  * **Algorithm Selection:** Choosing a machine learning algorithm (e.g., Logistic Regression, Naive Bayes, Decision Trees).
  * **Evaluation:** Measuring performance using metrics like precision, recall, and F1-score.

* **Challenges in Text Classification:**
  * **Imbalanced Data:** One category having significantly more samples than others.
  * **Ambiguity in Language:** Words having multiple meanings.
  * **Preprocessing Decisions:** The choices made during text cleaning significantly impact results.

* **Examples and Datasets:** The text mentions using the IMDB movie review dataset for sentiment analysis, the BBC News dataset for topic classification, and the AG News dataset for general text classification.

mcqs

Here are 5 multiple-choice questions based on the provided text:

1. **What are the two main types of text classifiers mentioned in the text?**
   a) Single-class and Multi-class
   b) Binary and Multi-class
   c) Positive and Negative
   d) Spam and Ham

2. **Which of the following is NOT a step in the text classification process described?**
   a) Data Collection
   b) Model Deployment
   c) Feature Extraction
   d) Classification Evaluation

3. **What is a common problem encountered when dealing with imbalanced datasets in text classification?**
   a) The model becomes overly complex.
   b) The model might become biased towards the majority class.

c) Feature extraction becomes impossible.
d) The dataset becomes too large to process.

4. **The text mentions several techniques for text pre-processing. Which of the following is explicitly mentioned?**
a) Removing punctuation
b) Lemmatization
c) Stemming
d) Stop word removal

5. **Which of the following datasets is mentioned as an example for sentiment analysis?**
a) BBC News Dataset
b) AG News Dataset
c) Kaggle IMDB Dataset
d) FastText Dataset

Answer Key:
1. b
2. b
3. b
4. d
5. c