Evaluating Text Classifiers: Beyond Accuracy

> Abhinav Kumar - 2301AI50
> Swagatam Pati - 2301AI28
> Shobhit Raj - 2301AI21
> Sravya Reddy - 2301AI24

... what is text classification?

understanding its definition and applications

Text classification refers to the automated process of categorizing text into predefined groups based on content.

> applications

- **spam detection:** identifying unwanted emails
- **sentiment analysis:** gauging public opinion in social media
- **topic labeling:** assigning relevant tags to articles or papers
- **language detection:** recognizing the language of a text

> common text classification algorithms

- **naive Bayes:** effective for large datasets, particularly in spam detection
- **support vector machines (SVMs):** known for high accuracy in text classification tasks
- **decision trees:** useful for understanding decision-making processes
- **deep learning models:** such as recurrent neural networks (RNNs) and transformers, which excel in capturing contextual nuances

... importance of model evaluation

why evaluation metrics matter greatly?

> avoiding over-fitting

proper evaluation metrics help in identifying model over-fitting, ensuring that the model generalizes well to unseen data, leading to more reliable predictions in real-world applications

... understanding accuracy

importance of a clear definition

> definition of accuracy

accuracy is defined as the **ratio of correct predictions to total predictions**

>> provides a straightforward measure of model performance

>> **however ...** it may not tell the whole story in complex datasets

... limitations of accuracy

why accuracy can mislead evaluations?

> class imbalance

accuracy alone can be deceptive, especially in **imbalanced datasets** where one class dominates

... leads to misleadingly high accuracy while failing to capture the model's true performance on minority classes

Example

95% non-spam emails → a model predicting all as non-spam gives 95% accuracy!

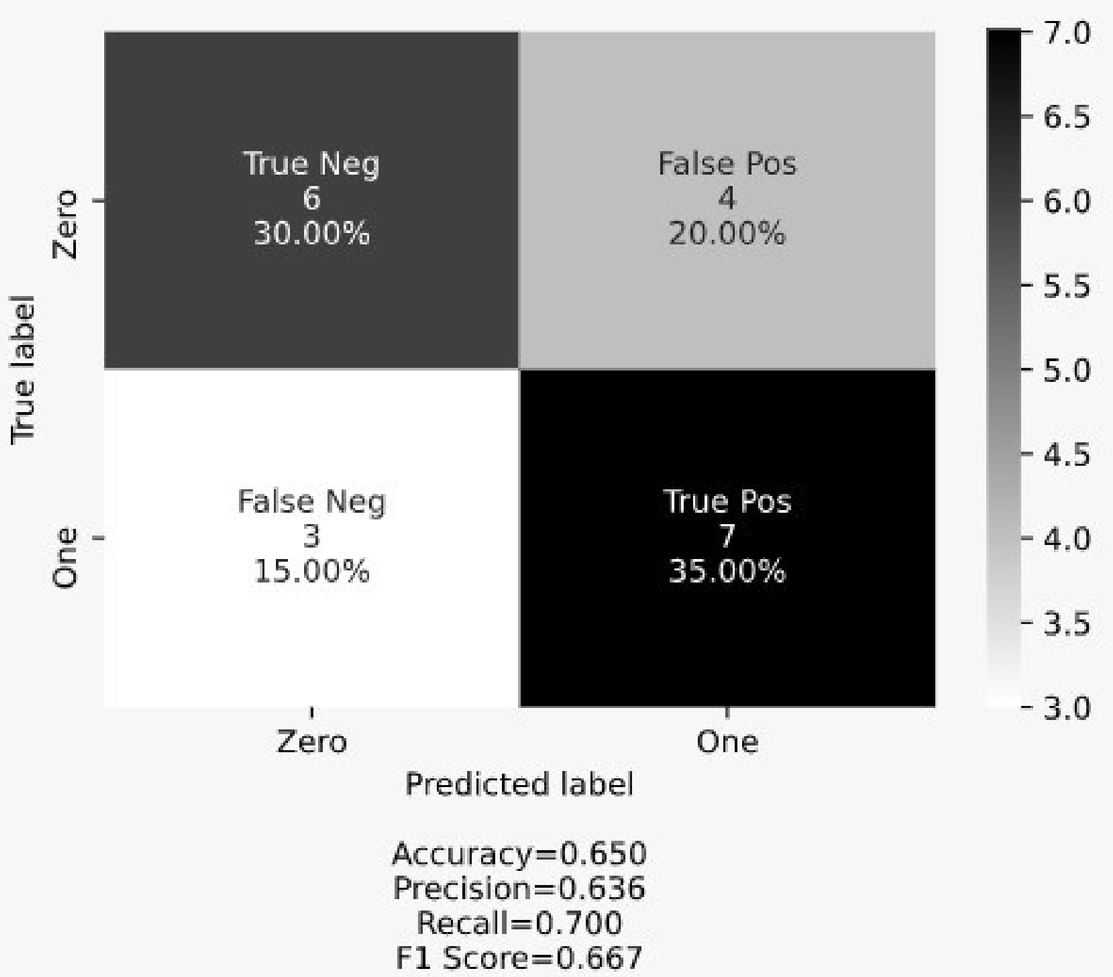
... confusion matrix overview

shows how many predictions were correct and where the model made mistakes

> four main terms
- true positives (TP)
- true negatives (TN)
- false positives (FP)
- false negatives (FN)

> allows us to derive metrics like
- precision
- recall
- accuracy
- F1-score

> helps identify whether the model is biased toward a particular class



... what is precision?

understanding a crucial evaluation metric

> definition of precision

precision is calculated as:
TP / (TP + FP)

> reflects the proportion of true positive predictions among all positive predictions

> indicates the model's accuracy in identifying relevant instances

> high precision ⇒ fewer false alarms

... what is recall?

understanding a crucial performance metric

> definition of recall

recall is calculated as:
  TP / (TP + FN)

> reflects the proportion of true positive predictions among all positive classes

> high recall indicates that the model has a low rate of missed positives

... precision-recall trade-off

understanding the inverse relationship dynamics

> competing metrics

increasing precision often leads to decreased recall

as the model becomes more selective, resulting in fewer positive predictions overall

> threshold tuning

adjusting classification thresholds can effectively balance precision and recall, allowing models to adapt based on the specific priorities of the task at hand

... understanding F1-score

the balance between precision and recall

> harmonic mean

the F1-score is calculated as the harmonic mean of precision and recall

2 x (precision x recall) / (precision + recall)

provides a single metric that balances both, making it particularly useful for imbalanced datasets

precision score

indicates a high correctness in predictions made

recall score

shows the model's efficiency in identifying positives

... averaging methods

understanding evaluation in multi-class tasks

> macro averaging

calculates metrics for each class separately,
then takes the average

treats all classes equally, providing a balanced
view of performance across different categories

> micro averaging

aggregates contributions from all classes before
calculating metrics

emphasizes the overall performance, making it useful
in scenarios with class imbalances or varying class
sizes

... real-world applications

impact of metrics in various fields

> healthcare diagnostics

- precision is crucial to ensure accurate diagnosis

- minimizing false positives that can lead to unnecessary treatments and patient anxiety

> fraud detection

- high recall is essential to catch as many fraudulent activities as possible

... interpreting metrics

aligning evaluation with objectives

> match metric

choosing the right metric for evaluation is crucial:

- it ensures alignment with business goals
- guides decision-making and model fine-tuning
- enhance model effectiveness in real-world applications

... importance of precision

real-world examples

- > email spam detection
- precision determines how many flagged emails are truly spam

- > healthcare
- precision ensures tests rarely give false positives, reducing unnecessary procedures

... importance of recall

real-world examples

- > fraud detection
- recall measures how many fraudulent transactions are caught

- > cancer detection
- high recall ensures no cases are missed, even if some false positives occur

... interpreting F1 score

- > the F1-score balances precision and recall, offering a single measure of effectiveness

- > it’s vital in imbalanced datasets where accuracy fails to reflect true model power

... conclusion

> understanding metrics for better evaluation

- relying solely on accuracy is misleading

- balancing precision and recall offers a comprehensive evaluation of model performance, leading to improved decision-making and insightful analytics