Applying NLP

Lecture 5: Text Classification

Harito

September 15, 2025

What is Text Classification?

Definition

Text classification is the task of assigning predefined categories or labels to text documents.

Examples:

- **Sentiment Analysis**: Positive/Negative/Neutral review.
- **Spam Detection**: Spam/Not Spam email.
- **Topic Labeling**: Assigning news articles to categories (Sports, Politics, Tech).
- **Intent Recognition**: Identifying user's intent in a chatbot (e.g., "book flight", "check balance").

Supervised Learning for Text Classification

We learn from examples!

- We need a dataset of text documents, each manually labeled with the correct category.
- The model learns patterns from these labeled examples to predict labels for new, unseen texts.

The Text Classification Pipeline

This brings together everything we've learned so far!

- **Tokenization**: Break text into words/tokens (Lab 1).
- **Vectorization**: Convert tokens into numerical features (Lab 2: Count, Lab 3: TF-IDF, Lab 4: Embeddings).
- **ML Model**: A machine learning algorithm that learns to map features to labels.

Machine Learning Model: Logistic Regression

Logistic Regression is a simple, yet powerful and widely used linear model for binary classification.

- Despite its name, it's a classification algorithm.
- It models the probability that a given input belongs to a particular class.
- It's a good baseline model due to its interpretability and efficiency.

Training and Testing

How do we know if our model is good?

Data Split

We split our labeled dataset into two parts:

- **Training Set**: Used to train the model (e.g., 80
- **Test Set**: Used to evaluate the model's performance on unseen data (e.g., 20

Why? To ensure the model generalizes well to new data and doesn't just memorize the training examples (overfitting).

Evaluation Metrics

Beyond simple accuracy, we use several metrics to understand model performance, especially for imbalanced datasets.

- **Accuracy**: $\frac{\text{Correct Predictions}}{\text{Total Predictions}}$ - Overall correctness.
- **Precision**: $\frac{\text{True Positives}}{\text{True Positives+False Positives}}$ - Of all predicted positives, how many were actually positive?
- **Recall**: $\frac{\text{True Positives}}{\text{True Positives+False Negatives}}$ - Of all actual positives, how many did we correctly identify?
- **F1-score**: Harmonic mean of Precision and Recall. Good for balancing both concerns.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Next Steps

Time for Lab 5!

Objective:

- Implement a ‘TextClassifier‘ using ‘LogisticRegression‘.
- Build a full text classification pipeline.
- Evaluate its performance using standard metrics.