News Article Classification: Health & Wellness

Introduction & Problem Statement

What's the Problem?

- Automatically classifying news articles.
- Specifically, identifying "Health & Wellness" articles from a large dataset.

Why is this Important?

- Content filtering for users.
- Personalized news feeds.
- Analyzing trends in health and wellness reporting.
- Efficient content organization for publishers.

Our Goal: Develop a machine learning model to accurately categorize news articles.

Data Overview

Dataset Details

- **Source:** `training_data.json` (news article collection).
- **Size:** Over 200,000 entries.

Target Variable: `healthy_wellness`

- Binary label: `1` for 'HEALTHY LIVING' or 'WELLNESS', `0` otherwise.
- **Challenge:** Initial dataset was imbalanced (only ~12% Health & Wellness).

Data Preparation

Combined Text Feature

- Merged `headline` and `short_description` into `combined_text`.
- Provides comprehensive input for the model.

Addressing Data Imbalance

- Sampled 20,000 non-health/wellness articles.
- Combined with all 24,521 health/wellness articles.
- Result: A more balanced dataset of 44,521 entries for training.

Methodology - Model & Training

Model Choice

- **FastText classifier** from the `ktrain` library.
- **Why FastText?** Efficient and performs well in text classification, especially with large vocabularies.

Training Process

- **Preprocessing:** Tokenization, padding.
- **Parameters:**
  - Max sequence length: 32 words.
  - Vocabulary: 7000 features.
- **Training:** 16 epochs, triangular learning rate policy, with early stopping to prevent overfitting.

Results & Conclusion

Model Performance

- **Validation Accuracy:** Approximately 87%.
- **Metrics:** Precision, Recall, and F1-score for both classes (0 and 1) were consistently high (0.86 - 0.89).
- **Interpretation:** The FastText model effectively distinguishes "Health & Wellness" articles.

Future work

- Explore other advanced transformer models (e.g., BERT, RoBERTa) for potential accuracy gains.
- Implement real-time article classification for live news feeds.
- Expand to classify other specific news categories.

Thank you