Comparative Analysis of Text Summarization and Classification Techniques

A Text Mining Project

Presented by: Francesco Beretta, Riccardo Borserini and Federico Cesare Cattò

Strategic Overview

Context:

- Increasing demand for efficient information extraction from massive textual data.
- Need to address the gap in inter-task validation across different datasets.

Objective:

- Verify the solidity and generalizability of methodological assumptions.
- Evaluation of models within a coherent and comparative framework.

Selected Corpora: NEWSROOM and AGNews

Summarization

NEWSROOM (1.3M articles) featuring diverse extractive strategies

Classification

AGNews, a consolidated benchmark for news topic classification

Methodological choice

Distinct datasets to respect original application domains

NEWSROOM Sampling

Management of computational constraints for NEWSROOM

Experimental Framework and Implementation

Replication of Approaches

Replication of state-of-the-art methods following original paper guidelines

Consistent Preprocessing

Uniform preprocessing to ensure fair comparison between diverse models

Baseline Inclusion

Integration of baseline models to isolate the impact of architectural complexity

Reproducibility Focus

Focus on transparency and reproducibility in the absence of official code

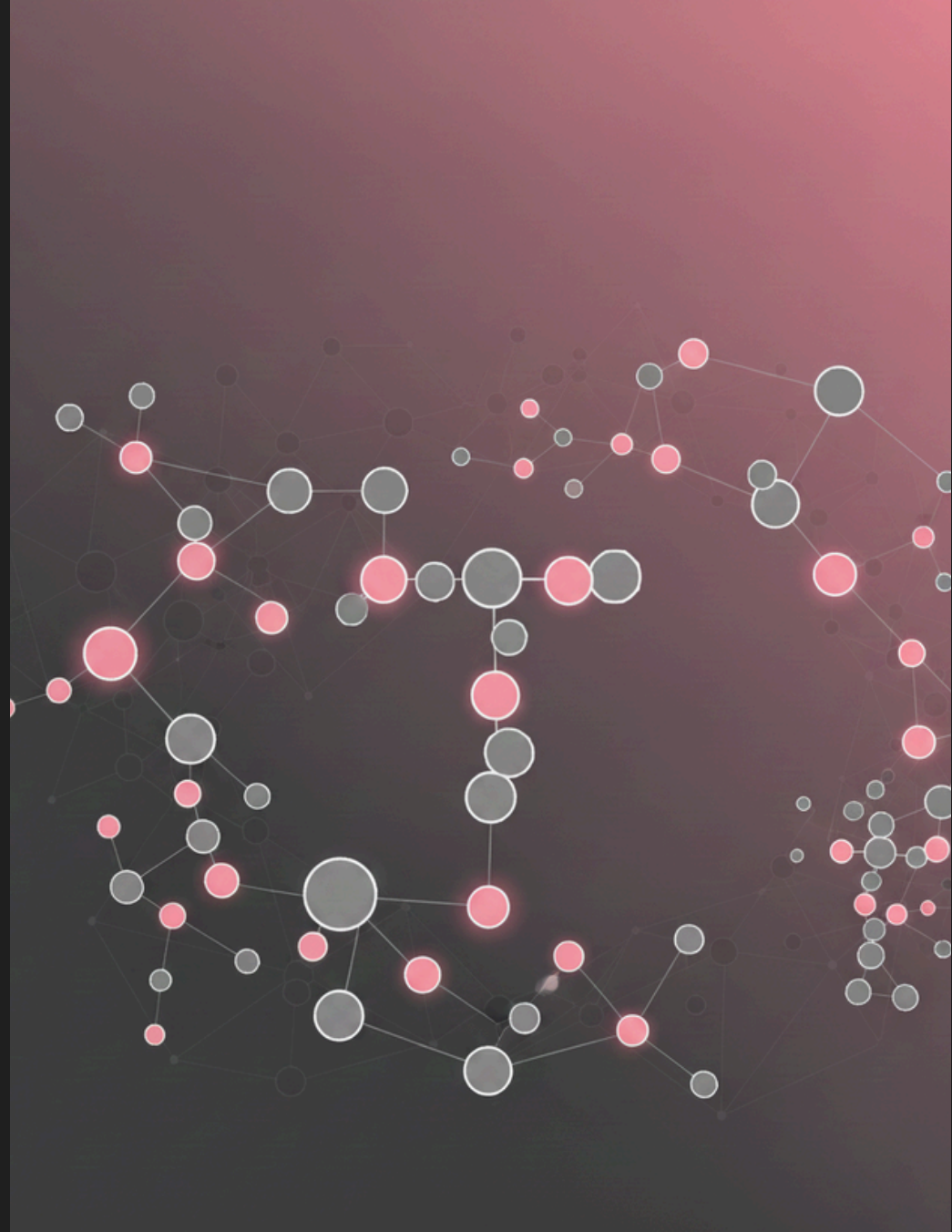
Graph-Based Summarization: Enhanced TextRank

Baseline Model
Traditional TextRank relying on lexical frequency

Core Model
TextRank enriched with BERT contextual embeddings

Hybrid approach
Combining BERT semantics with TF-IDF weighting

Unsupervised Selection
Based on sentence connectivity and centrality



Classification Strategies

Random Forest Miner

- Ensemble model optimized for high-dimensional textual data
- Parameter optimization via systematic Grid Search for the RF-Miner

Baseline Models

- 1D Convolutional Neural Network to capture local patterns and n-gram structures
- Logistic Regression and Linear SVC for high-volume efficiency

Performance Indicators and Standards

Summarisation Evaluation

- Application of ROUGE-1, ROUGE-2, and ROUGE-L (F1-score variant).
- Measurement of lexical overlap between generated and reference summaries.

Classification Evaluation

- Application of Error Rate (1−Accuracy) for direct literature comparison.
- Metrics chosen to ensure comparability with original benchmark papers.

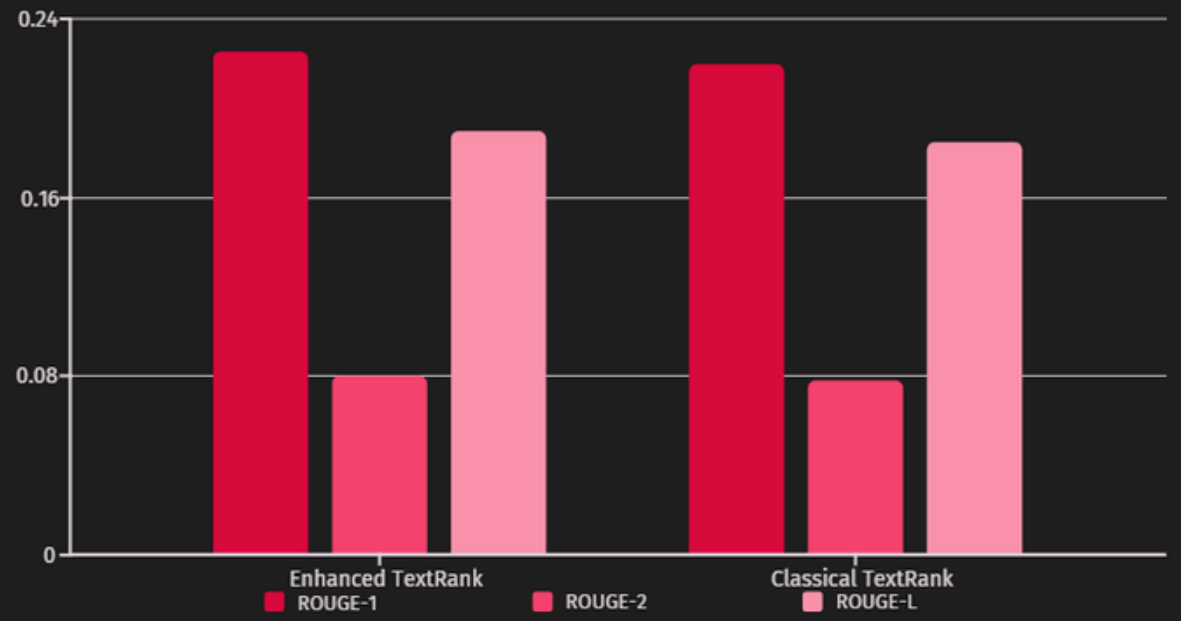
Results: Stability of Lexical Overlap

Similar ROUGE performance across summarisation models

ROUGE-1 ≈ 0.22 for both approaches

Contextual embeddings: only marginal improvements

Keyword frequency remains dominant in extractive news summarisation



Results: Impact of Architectural Complexity

1D CNN

Achieved the highest accuracy amongall classification models.

RF-Miner

Demonstrated improved performance overtraditional linear baselines.

Linear Models

Despite lower accuracy, they remain strong contenders due to their computational efficiency.

A clear performance gradient was observed, directly correlating with increased model complexity.

Critical Discussion: Complexity vs. Utility

Summarization

Lexical regularity dominates distributional semantics

Classification

Neural structures significantly reduce systematic error

Methodological choice

Complexity is necessary for discrimination but optional for extraction

NEWSROOM Sampling

Prioritizing methodological coherence over extreme individual optimization

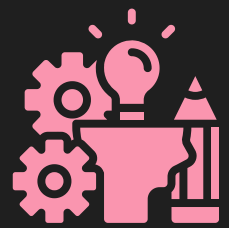
Constraints and Scope

- Feature space limited to 200 terms (VSM constraints)
- Reduced NEWSROOM subset due to RAM limitations
- BERT input capped at 512 tokens (mean pooling)
- Extractive approach: no semantic paraphrasing

Future Developments

Expanding feature sets by removing the 200-term constraint.

Implementation of Full Transformer models for the classification task.



Transition to Abstractive Summarization using generative models.

Qualitative evaluation of logic, coherence, and readability.

Conclusions

Validation of distinct text mining paradigms across diverse tasks

Contextual embeddings are not a "silver bullet" for extractive summaries

Neural architectures provide measurable gains for complex discrimination

Preprocessing and feature selection remain critical performance drivers