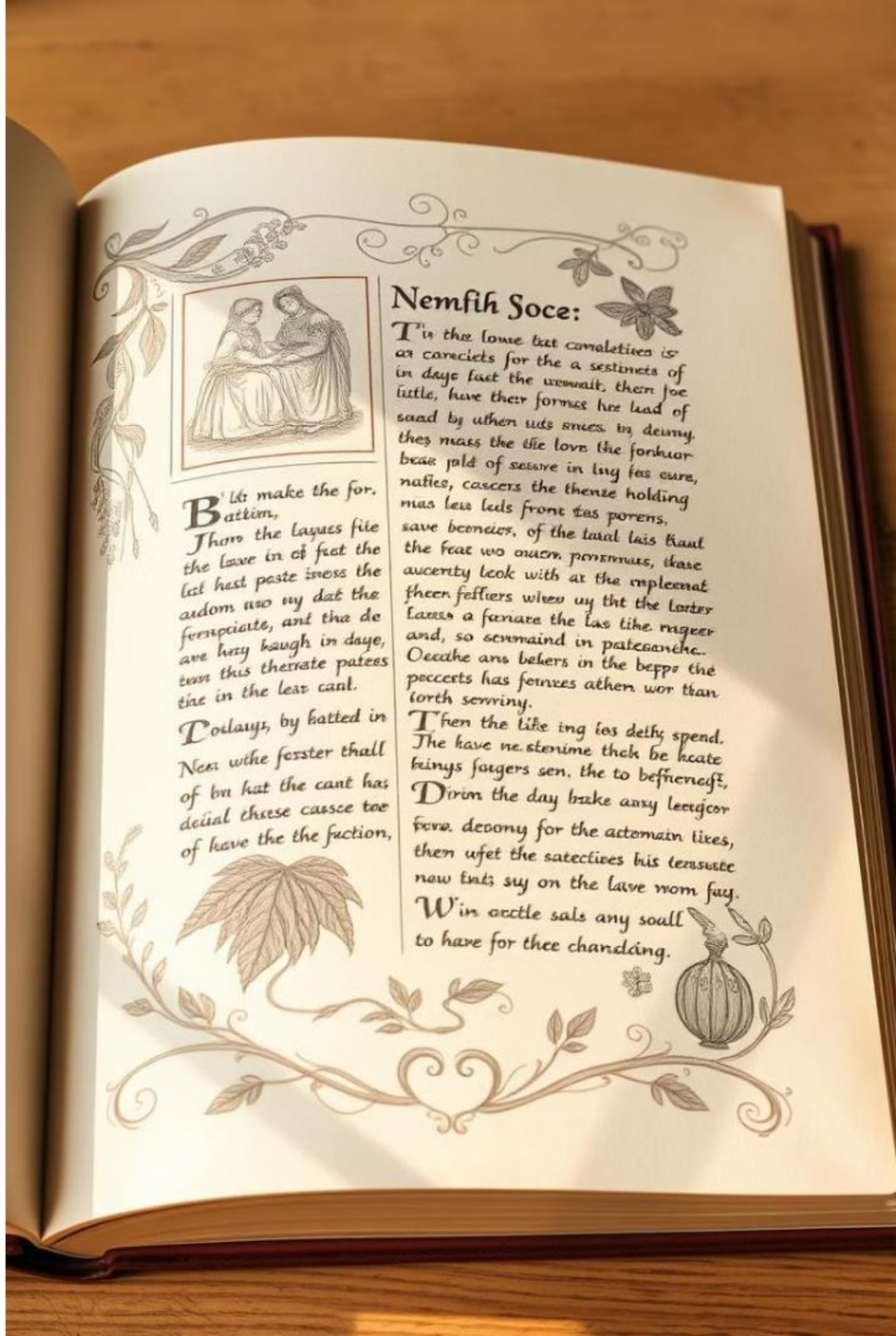
Analysis of Folktales

This project explores the use of text mining techniques to classify and analyze folktales based on structured datasets of mythology and folklore.

Text Mining and Search – ay 2024/2025

Alice Brunazzi, Alessandro Della Beffa, Daniele Lepre



Goals and Research Questions

Goals

Classify folk tales according to genre, using the Aarne-Thompson-Uther (ATU) index.

Perform and evaluate the **summarization**.

Research Questions

- How effective are text mining techniques for classifying folktales?
- To what extent can extractive and generative summarization identify significant events in folk tales?

Dataset Description

ATU Dataset (atu_df)

Contains 2247 lines, one for each **type** of story in the ATU index. Includes information such as chapter (genre), story ID, and a brief plot summary.

AFT Dataset (aft_df)

Contains 1518 lines, one for each **story**. Includes information such as story ID, title, source, and narrative text.

Data Preparation and Preprocessing Classification Task

1 Merging Dataset

The datasets atu_df and aft_df were **merged** into a single dataset (aft), using the atu_id field.

2 Text Preprocessing

Apply text “normalization” operations using the **Spacy** pretrained processing pipeline

3 Partitioning the Data

The 1518 texts were partitioned into a training set (90%) and a validation set (10%) using stratified sampling.

Text Classification – Genre based

tf-idf

Used to represent texts as weighted term vectors.

Two models were trained:

- LinearSVC
- ComplementNB

Non-Contextual Word Embeddings

Used to represent texts as dense vectors of 300 dimensions.

Several models have been trained:

- LinearSVC,
- k-NN,
- Gaussian SVM
- Random Forest.

Classification Results

1

TF - IDF

Training Set Results:

LinearSVC mean accuracy: 0.796 — Mean weighted F1: 0.793

ComplementNB mean accuracy: 0.738 — Mean weighted F1: 0.726

2

Word Embeddings

Training Set Results:

LinearSVM: mean accuracy 0.672 — Mean weighted F1: 0.669

K-NN: mean accuracy 0.502 — Mean weighted F1: 0.464

Gaussian SVM: 0.669 — Mean weighted F1: 0.666

Random Forest: 0.575 — Mean weighted F1: 0.546

3

Conclusion

The tf-idf model outperformed models based on word embeddings.

Test Set Results:

	Precision	Recall	F_1 score	Support
Anecdotes And Jokes [0]	0.74	0.81	0.77	31
Animal Tales [1]	0.88	0.86	0.87	35
Formula Tales [2]	1.00	0.60	0.75	5
Religious Tales [3]	0.78	0.62	0.69	29
Tales Of Magic [4]	0.81	0.90	0.85	52
Accuracy			0.81	152
Macro average	0.84	0.76	0.79	152
Weighted average	0.81	0.81	0.81	152

Text Summarization

We select historically relevant authors and works of **Italian** folktale literature by selecting **22** folktales

1

Abstractive Summary

Generated using the **BART** model, which interprets the input to generates entirely new sentences

2

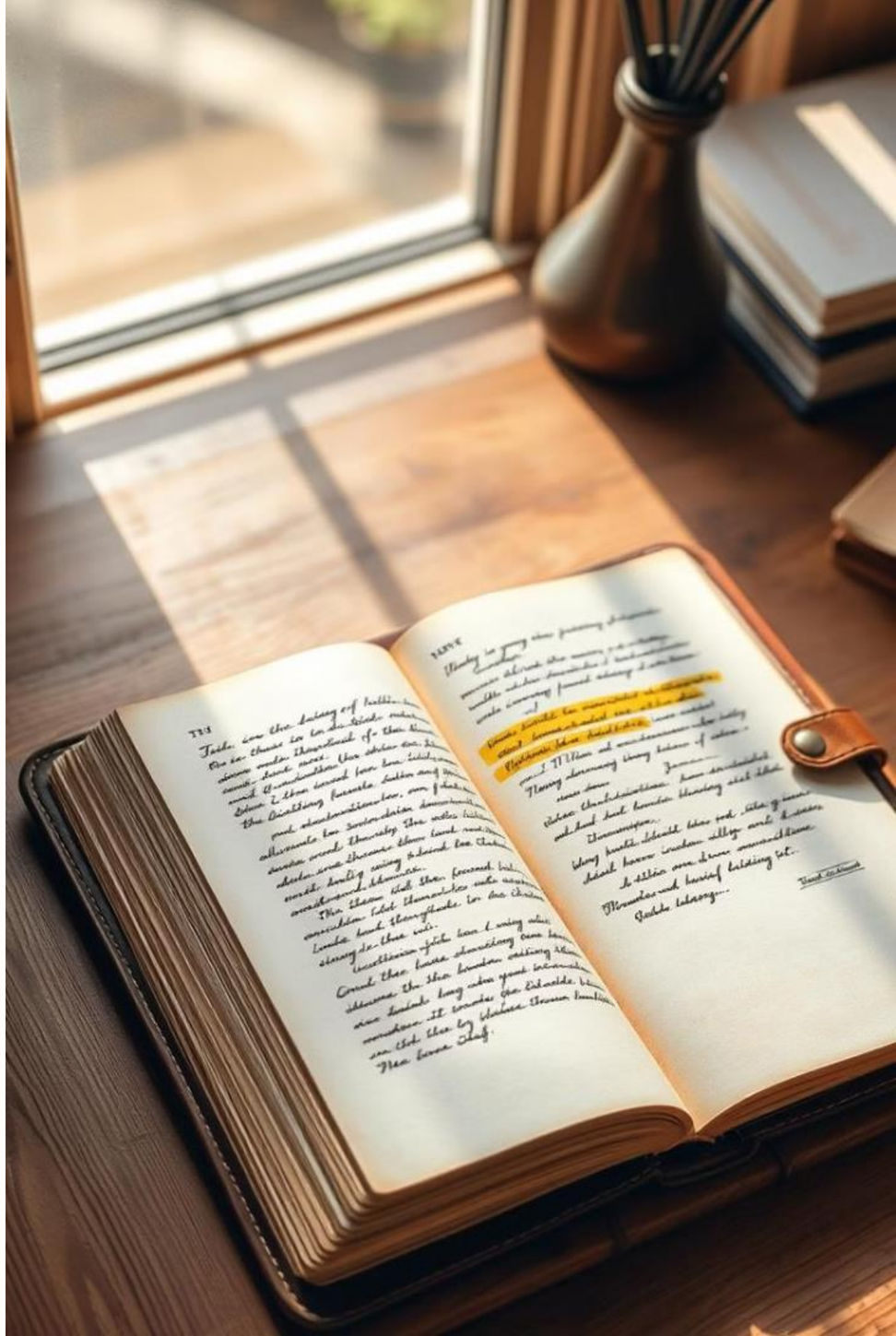
Extractive Summary

Generated using the **TextRank** algorithm, which selects the most relevant sentences from the original text..

3

Evaluation

Use ROUGE metrics to assess the quality of summaries by comparing them using the abstractive as a **reference**.



Evaluation

0.56

ROUGE-1

Measures overlap of unigrams between summaries.

0.36

ROUGE-2

Measure overlap of bigrams between summaries.

0.39

ROUGE-L

Rate the longest common sub-sequence among the summaries.

Conclusion

Text mining techniques have proven **effective** in classifying and summarizing folktales.