Text Mining and Search Project

Francesco Beretta – 880032 Riccardo Borserini – 880911
Federico Cesare Cattò – 880670

Contents

1 Introduction

The growing availability of textual data and the need to extract information efficiently have made text mining techniques a central element in numerous fields of application, from digital communication to editorial content analysis. Among the most important tasks are automatic text summarization and classification, both of which are characterised by specific methodological challenges and continuous evolution of the models proposed in the literature. However, reference works often focus on individual tasks and datasets, leaving open the problem of comparative inter-task validation to verify the robustness and generalisability of methodological assumptions in different contexts.

This work addresses this gap by comparatively analysing different summarization and classification techniques, evaluating their performance within each task according to methodologies consistent with the relevant field, and comparing these results with those reported in the reference corpus papers. Since the corpora used were originally developed for different tasks, the decision to adopt two distinct datasets responds to a specific methodological need: each dataset is accompanied by studies that evaluate its performance only in its own field of application, using specific metrics and experimental configurations that are not comparable with each other. To this end, the work uses "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies" as the reference corpus for the summarisation task and AGNews[1] for the classification task.

For the summarization task, the NEWSROOM corpus, which collects over 1.3 million articles from 38 international newspapers, is characterised by the richness and stylistic variety of its summaries, which adopt extractive strategies and make it a suitable test bed for evaluating TextRank-based summarisation algorithms. This project has adopted the approach proposed in "Enhanced TextRank using weighted word embedding for text summarization", which introduces a variant of the TextRank algorithm enriched with distributional representations based on BERT models. This method was compared with a version of TextRank without BERT components, a choice that allows the information contribution provided by contextual embeddings to be isolated and evaluated in a targeted manner with respect to the basic algorithmic structure.

For the classification task, the use of the AGNews dataset is justified by the fact that it is one of the most established benchmarks in the field of text classification. The approach described in "An improved text classifier based on random forest algorithm - comparative studies on multiple text classifiers" was implemented on this dataset, which proposes a variant of Random Forest specifically adapted to text classification using VSM vectorisation and CART trees without pruning. To support the classification analysis, additional comparative models were considered, selected for their complementarity with the characteristics of the method described above. In particular, the Support Vector Classifier (SVC) was included for its proven effectiveness in text classification in high-dimensional vector spaces; Logistic Regression serves as a linear baseline, providing a reference for assessing the impact of architectural complexity; finally, a Convolutional Neural Network (CNN) based on word-level representations was adopted to evaluate the effectiveness of models capable of capturing local structures in texts, offering a comparison with traditional approaches. The models described and implemented in this study will be compared in terms of performance with those reported in "Character-level Convolutional Networks for Text Classification", recognised as one of the most established benchmarks for the AGNews corpus, in order to contextualise the results obtained with respect to the standards in the literature.

In summary, the methodological choices adopted, from the main methods to the comparative models and selected datasets, define the framework of the analysis, allowing for a systematic evaluation of the models' performance and verification of their robustness in the contexts considered. The following sections will follow the outlined structure, systematically illustrating the methods taken from the literature, justifying the choice of datasets, describing the comparison

[1] http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

models and presenting the metrics and experimental setup used. This approach will provide a clear and verifiable framework that will allow for a critical interpretation of the results reported in the final part.

2 Background and Related Work

This section contextualises the two tasks considered in the project, text summarization and text classification, with the aim of illustrating the importance of the works selected as reference models, before describing their structure in detail (see Methodology). The two models chosen, respectively the extension of the TextRank algorithm with the BERT model for summarization and the Random Forest miner for classification, represent established approaches in their respective lines of research and constitute the main models of the analysis.

Alongside the two reference models, additional approaches were considered for comparative purposes, in order to contextualise the results within a broader methodological framework. For summarization, the focus is on a traditional TextRank algorithm without semantic components, useful as a basis for comparison with the extension proposed with BERT, while for classification, both established methods, such as Support Vector Classifier with linear kernel and Logistic Regression, and a solution based on a neural model, such as 1D Convolutional Neural Network, are used. This approach allows the main models to be evaluated not in isolation, but within a broader methodological context, enriching the interpretation of the results and placing them in relation to other relevant approaches in the literature.

2.1 Text Summarization

Automatic text summarization is a well-established field of study within Natural Language Processing, with the aim of producing concise and coherent summaries that facilitate access to large volumes of information. In this work, the focus is on extractive summarization, chosen both for the transparency of the underlying mechanisms and for the continuing relevance of these techniques in real-world applications. In a context characterised by exponential growth in textual sources, the ability to reduce the information load without compromising content quality is a crucial requirement, and extractive approaches continue to be a methodologically sound solution.

The main model adopted in this study is based on the approach presented in "Enhanced TextRank using weighted word embedding for text summarization", considered a significant contribution to the evolution of extractive methods. The idea behind this framework is to enrich the TextRank architecture with semantic information learned through BERT, combining distributional representations and TF-IDF weights within a graph ranking mechanism. This integration allows the construction of phrase vectors that reflect not only statistical co-occurrences but also deeper conceptual relationships, thus overcoming the well-known limitations of models based purely on lexical similarity. In the absence of an implementation released by the authors, the entire method was faithfully replicated following the instructions in the paper. The only change concerns the linguistic model used: while the original version uses IndoBERT, bert-base-uncased was adopted here; it represents a necessary choice to maintain consistency with the English nature of the dataset and to ensure a semantic representation appropriate to the domain analyzed.

For comparison purposes, a traditional version of TextRank based exclusively on TF-IDF and graph ranking was adopted as a baseline. The approach uses the similarity matrix between sentences to construct a graph, to which the PageRank algorithm is then applied, assigning each sentence a score that reflects its relevance within the text. This configuration allows the most significant information units to be identified without resorting to distributional representations or pre-trained linguistic models.

The combined use of a model enriched with distributional representations and a purely statistical baseline makes it possible to highlight more clearly the benefits of the approach inspired by the original work, not only in terms of the quality of the selected content, but also in terms of the internal consistency of the summaries generated. In this way, the analysis is not limited to verifying a single implementation, but helps to outline the relative strengths and weaknesses of extraction strategies in the contemporary context of text summarisation.

2.2 Text Classification

Text classification is one of the central tasks of text mining and is used in areas such as content filtering and automatic news organization. Over time, various approaches have emerged: from methods based on sparse representations of text, such as bag-of-words and TF-IDF, to neural models that automatically learn patterns and distributional representations. This evolution has made classifiers more accurate and better suited to handling increasingly large datasets and increasingly complex tasks.

Within this framework, the Random Forest model enhanced by the miner approach, proposed by Luo in "An improved text classifier based on random forest algorithm - comparative studies on multiple text classifiers", occupies a significant position in the literature dedicated to ensemble methods for text classification. The contribution of the paper lies in the targeted adaptation of a widely established technique, Random Forest, to a domain in which sparse representations and the high dimensionality of text make the informative selection of attributes and appropriate splitting strategies particularly relevant. The model highlights how classical tools can be extended in a targeted manner to improve discrimination on textual data, achieving competitive results while maintaining a high degree of interpretability. Although no official implementation is available, this method has been adopted as the main reference point as it represents a recognised and solid line of research, capable of combining effectiveness, robustness and a transparent modelling structure.

To complement this approach, comparative models belonging to different methodological families were selected to allow for a more comprehensive performance evaluation. These include classic linear models, such as Logistic Regression and Linear SVC, which represent established baselines in high-dimensional text classification contexts, and a 1D Convolutional Neural Network based on word embedding, which allows the traditional approach to be compared with the ability of neural models to capture local patterns in text. The selection of comparative models is dictated by the desire to cover the main strategies used today in text classification. Logistic Regression and Linear SVC represent established linear baselines, allowing the improved Random Forest to be compared with simple but effective approaches in high-dimensional contexts. The CNN based on word embedding, on the other hand, provides a benchmark with neural models capable of learning local patterns in texts, showing the impact of deep learning architectures compared to traditional ensemble methods.

In this way, the analysis is not limited to testing a single model, but allows the performance of the improved Random Forest to be evaluated against different paradigms, clarifying the relative strengths and limitations of each approach within the state of the art of text classification.

3 Datasets

The choice of datasets plays a crucial role in designing a comparative study on summarization and classification tasks. An approach based on a single dataset would not be adequate, as each corpus has been designed, annotated and validated specifically for a particular task: data structures, evaluation metrics and collection methods reflect different requirements depending on whether the aim is to generate summaries or assign class labels to texts. Furthermore, the possibility of comparing the results reported in the original papers requires the use of datasets for which relevant and comparable metrics are already available. The adoption of separate datasets for the two tasks therefore ensures methodological consistency, allowing the performance of the models to be evaluated in the context for which the corpora were designed and the results to be interpreted in relation to the standards established in the literature.

This section illustrates the reasons behind the choice of NEWSROOM for summarization and AGNews for classification, highlighting how these decisions support the validity and relevance of the methodological comparison.

3.1 Newsroom

The decision to adopt the Newsroom corpus for the summarization task is motivated by the distinctive characteristics of this corpus. It comprises over 1.3 million articles and related summaries written by authors and editors from 38 major newspapers between 1998 and 2017. The summaries show considerable stylistic variety and varying degrees of lexical extraction from the original texts, reflecting different summarization approaches used in real journalistic contexts. This diversity, together with the broad time span and coverage of multiple subject areas, makes Newsroom a particularly suitable dataset for evaluating extractive summarization algorithms such as TextRank. Comparative studies with other historical datasets, such as DUC 2003-2004, CNN / Daily Mail and New York Times Corpus, highlight how Newsroom offers a wider range of summarization styles, differing in information density, content coverage and text compression level.

From an operational point of view, given the absence of classes to stratify, random sampling of 500,000 observations was applied to the training set, a choice dictated by memory constraints: above this threshold, the RAM available on Colab does not allow the entire dataset to be managed. For the testing phase, for similar reasons, a subset of 1,000 examples was selected, sufficient to evaluate the performance of the models without exceeding computational limits. This configuration ensures a compromise between the representativeness of the dataset and the feasibility of training.

3.2 AGNews

The choice of the AGNews corpus for the text classification task is motivated both by its widespread use as an established benchmark in the literature and by the intrinsic characteristics of the dataset. AGNews collects over a million articles from more than 2,000 news sources, aggregated by the academic news search engine ComeToMyHead over a period of more than a year. The dataset is made available for academic research purposes, supporting studies in areas such as data mining, information retrieval, and text classification. In order to ensure comparability with the results reported in “Character-level Convolutional Networks for Text Classification”, the study considered the same portion of the dataset defined by the authors, known as AG’s news topic classification dataset, consisting of the four most representative classes of the corpus (world, sports, business, and science) and using only the title and description fields of the articles. In this configuration, the training set comprises 30,000 examples for each class, while the test set contains 1,900.

4 Methodology

This section presents the methodologies adopted in this study, detailing the text preprocessing steps and their application across the different models.

For text classification, Logistic Regression, Linear SVC, and a 1D Convolutional Neural Network (CNN) are considered as comparative models, while the Random Forest Miner is adopted as the reference model. For each model, the preprocessing pipeline and model architecture are described.

For text summarization, two TextRank-based approaches are presented: an enhanced version incorporating BERT embeddings, used as the reference model, and a baseline version without embeddings, employed for comparative analysis. For both approaches, the text preprocessing steps and algorithmic implementation are detailed.

4.1 Summarization

4.1.1 Enhanced TextRank using BERT embedder

This sub-section describes the key phrase extraction methodology adopted for the summarization system, drawing on the approach proposed in the reference paper and implementing an enhanced variant of the TextRank algorithm. The central methodological choice lies in the integration of BERT's contextualised embeddings with TF–IDF (Term Frequency–Inverse Document Frequency) statistical weights. Despite the apparent contrast between a statistical indicator and the semantic representations generated by an attention model such as BERT, this combination responds to the need to balance the deep semantics captured by BERT with the statistical relevance of terms within the document. BERT, in fact, tends to distribute information relatively evenly; the introduction of TF–IDF allows the most representative terms to be explicitly highlighted, modulating their influence on the overall vector representation.

The process consists of three stages. Initially, the TF–IDF weight is calculated for all terms in the reference corpus. To this end, a procedure has been developed to analyze and standardize the terms, converting them to lowercase and removing textual noise, in order to ensure consistent frequency estimates. The TF–IDF value, as the product of the local frequency of the term and its global importance, is then used as a coefficient to weight the vector representations.

Next, embeddings are generated using the pre-trained bert-base-uncased model. The adoption of a bidirectional architecture based on the self-attention mechanism allows for contextualized representations that overcome the limitations of static techniques, capturing complex syntactic and semantic relationships. The text is tokenized and adapted to the model's requirements, respecting the 512 token limit, and embeddings are generated from the last hidden layer. To stabilize word-level representations, token vectors derived from subword units are aggregated using averaging. The core of the method consists of applying weighting: each word vector generated by BERT is multiplied by the corresponding TF–IDF weight, amplifying the importance of significant terms. The final vector representation of the sentence is obtained as the average of the weighted vectors, offering an effective compromise between numerical stability and statistical discrimination, resulting in a dense and informative representation.

The third phase is extractive selection, which employs an enhanced version of TextRank. The effectiveness of TextRank lies in its unsupervised nature, which exploits a network of similarity relationships between sentences. The premise is that the relevance of a sentence is directly proportional to its connectivity with other significant sentences in the text. To construct the graph, the similarity between sentences is calculated using the cosine similarity of vectors generated by BERT and previously weighted with TF-IDF. This integration ensures that the weight of the connections jointly reflects the semantic component and statistical importance, facilitating the identification of conceptual nuclei. The TextRank algorithm then applies a weighted version of PageRank (using a damping coefficient of 0.85), which converges on a distribution of centrality scores for each sentence. The final summary is extracted by selecting the sentences with the highest scores and reordering them according to their original position in the text to ensure consistency and readability. The integration of contextualized embeddings, TF-IDF weighting and graph-based ranking produces a system capable of generating coherent, informative summaries that faithfully reflect the semantic structure of the document. This architecture also avoids the need for supervised models, maintaining high generalizability across different types of text and sustainable computational complexity.

4.1.2 TextRank without BERT embedder

This sub-section describes the methodology adopted for extracting phrases using the TextRank algorithm in its classic formulation, based exclusively on TF-IDF representations, and used as a baseline for comparison with the enhanced variant that integrates contextualised BERT embeddings, described above. The methodological choice is aimed at isolating the impact of deep semantic representations, while keeping the preprocessing, ranking method and experimental context unchanged. In this way, the differences observed during the evaluation phase can be attributed solely to the method of sentence representation. The text preprocessing is kept as consistent as possible with that adopted in the BERT version. In particular, the Newsroom corpus undergoes a normalization procedure that includes converting characters to lowercase, removing stopwords specific to the journalistic domain, and reducing textual noise. This uniformity in the preprocessing phase is essential to ensure the comparability of results, as it allows us to operate on a common lexical basis and obtain consistent estimates of term frequencies.

The representation of sentences is based exclusively on TF-IDF weights, calculated from the training corpus. In a first step, a global IDF dictionary is estimated using the entire training set in order to model the informational importance of terms with respect to the domain under consideration. This choice avoids the introduction of information leakage from the test set and ensures a correct evaluation of the model's performance. The TF-IDF value is the only source of information for the construction of vector representations. At the sentence level, each text unit is represented as a vector in TF-IDF space using a weighted bag-of-words approach. Vectorization is performed by learning the vocabulary directly from the sentences in the document under consideration. Unlike the BERT-based approach, this representation does not incorporate contextual information or latent semantic relationships, but reflects only the statistical relevance of terms and their distribution within the text. This limitation is deliberate and functional for comparative analysis, as it allows the effectiveness of TextRank to be evaluated in the absence of deep semantic information.

The next step is to construct the similarity graph between sentences. Each node in the graph represents a sentence in the document, while the weight of the edges is determined by the cosine similarity between the corresponding TF-IDF vectors. This measure provides a direct estimate of the information overlap between sentences, based on the sharing of relevant terms. The resulting similarity matrix therefore encodes the lexical affinity relationships between all pairs of sentences in the document. The TextRank algorithm, implemented as a weighted variant of PageRank, is applied to this graph. The score assigned to each phrase is calculated recursively based on the scores of the phrases connected to it, according to the principle that a phrase is more relevant the more similar it is to other phrases centrally located in the graph. The algorithm adopts a damping factor of 0.85, a value widely used in PageRank studies, as highlighted in the work "PageRank as a function of the damping factor", as it ensures the convergence of the algorithm and stabilizes the behaviour of the iterative process.

The final summary is selected using a purely extractive approach. Sentences are sorted according to the centrality score produced by TextRank, and the first k are selected, where k is dynamically determined as the number of sentences in the reference summary or, in the absence of ground truth, as a predetermined fraction of the total number of sentences in the document. In order to preserve the textual coherence and readability of the summary produced, the selected sentences are finally reordered according to their original position in the text. The entire system remains completely unsupervised and has limited computational complexity, making it a solid, interpretable and easily replicable baseline. In this context, comparison with the variant based on contextualized embeddings allows us to rigorously quantify the contribution of deep semantic representations compared to a model based exclusively on frequency statistics and lexical similarity relations.

4.2 Classification

4.2.1 Logistic Regression & Linear SVC

In an initial experimental phase, before proceeding with more complex models such as the Random Forest Miner version used as a reference model for the classification task, we evaluated the performance of two linear classifiers widely used in the field of text classification: Logistic Regression and Linear Support Vector Classifier (LinearSVC). The decision to start with these models is motivated by the fact that, being linear models, they are suitable for large corpora such as AGNews, allowing competitive results to be obtained at low computational cost and providing a solid reference point for the evaluation of subsequent models. To ensure a fair and rigorous comparison, the two linear models were trained using the same preprocessing pipeline, as they share similar operating principles and analogous methods of text feature processing. In both cases, the text was obtained by combining the title and description, converted to lowercase and cleaned up by removing punctuation and non-alphabetic characters in order to standardise the lexical information and reduce noise. Subsequently, the textual representation was transformed into TF-IDF vectors using identical parameters for both models: a maximum limit of 200 features, the inclusion of uni- and bigrams, a minimum frequency threshold of two occurrences, a maximum frequency limit of 90%, and a sublinearly scaled TF weighting, so as to highlight the most informative words and ensure a balanced representation of the terms in the dataset. In this way, in accordance with the reference model, the textual representation is constructed according to a controlled-dimensionality Vector Space Model approach: the choice of a maximum of 200 features, indicative of the total number of attributes, is implemented with the TF-IDF vectorizer, limiting the feature space to the most significant terms (or n-grams) according to TF-IDF weighting. This approach ensures consistency with the reference paper and a fair comparison between the different classification models, keeping the dimensionality of the textual input constant. As for modelling, Logistic Regression was configured with the lbfgs solver, recommended for multinomial logistic regression problems, a maximum number of iterations equal to 2000 and a regularization strength of 1, chosen to balance the model's generalization capacity without introducing excessive penalization on the coefficients. Linear SVC was also set with a regularization strength of 1 and trained for a maximum of 5000 iterations to ensure the convergence of the linear solver.

The initial use of Logistic Regression and LinearSVC therefore allowed us to obtain a reliable baseline of the performance achievable with linear models, providing a useful starting point for more critically evaluating the effectiveness of subsequent models and the impact of the methodological choices made in the later stages of the analysis.

4.2.2 1D Convolution Neural Network

In a second experimental phase, in order to explore non-linear models capable of capturing local relationships between consecutive terms, a one-dimensional convolutional neural network (1D CNN) was used for the classification task. Again, text preprocessing was consistent with that applied to linear models, combining title and description into a single sequence, standardizing the text to lowercase and removing non-alphabetic characters and multiple spaces in order to reduce noise and standardize lexical information.

For numerical representation, text sequences were encoded using embedding based on a vocabulary limited to the 200 most frequent terms, in line with the dimensionality used in the reference model. The sequences were then normalized to a fixed length using padding, ensuring a uniform input size for the network. The CNN architecture includes an embedding layer to transform each token into a continuous feature space, followed by a one-dimensional convolution to extract local n-gram patterns, a global pooling operation to aggregate the most relevant features, and one or more fully connected layers with dropout to control overfitting. The final output is a softmax layer that returns the probability of belonging to each class.

Training was conducted by minimizing categorical cross-entropy, using a stochastic optimization algorithm with learning rate adaptation and validation loss monitoring to stop training at the first sign of convergence. This configuration allows the network to capture both local patterns and global information within the texts, offering a comparison with the TF-IDF vector-based representations used in linear models and providing a more comprehensive analysis of the effectiveness of different text classification approaches.

4.2.3 Random Forest - Miner

Finally, the Random Forest Miner reference model is adopted as the reference model for the classification task. Text preprocessing, which was not specified in the original work, is defined in a manner consistent with the other models used in the experiment, combining the title and description into a single sequence, standardizing the text in lowercase and removing non-alphabetic characters and multiple spaces, in order to ensure a fair comparison between different approaches.

For numerical representation, the texts are transformed into a controlled-dimensionality Vector Space Model, limited to the 200 most informative terms, according to the same logic applied to linear models and CNN. This choice ensures methodological consistency, respecting the constraint indicated in the paper, and limits the feature space to a set of significant terms, combining uni- and bigrams and adopting a sublinearly scaled TF-IDF weighting.

After obtaining the textual representation in the vector space model, the next step involves constructing the Random Forest classifier according to the methodological guidelines proposed in the reference document. In particular, the procedure adopted faithfully reproduces the steps described in the paper: the training set is subjected to bagging, generating n_{tree} bootstrap subsets of the same size as the original dataset, and for each subset a CART classification tree is constructed without any pruning phase. At each internal node, $mtry$ attributes are randomly selected from the 200 available in the VSM, and the choice of the best split attribute is made by minimizing the Gini index. Each tree is grown to the maximum depth allowed by the data, thus producing a set of complete trees that constitute the final forest.

Although the paper provides indicative values for the main parameters, such as the number of trees (n_{tree}), the number of attributes considered at each split ($mtry$) and the minimum size of terminal nodes, these values are intrinsically linked to the specific characteristics of the original dataset used by the authors. For this reason, while maintaining the methodological architecture proposed by the authors unchanged, we choose to determine the most effective parameter values through a systematic optimization process in order to find those most suited to the AGNews dataset. The selection is made using Grid Search with cross-validation, using "accuracy" as the reference metric, a particularly appropriate criterion in the presence of a balanced dataset, as it provides a direct and comprehensive measure of the model's classification capacity. The procedure explores different configurations of the key model parameters described in the paper, with the aim of identifying the combination that would guarantee the best performance on the training dataset. The optimal values found, using AGNews data as a reference, are: `max_features` = 14, `min_samples_leaf` = 1 and `n_estimators` = 200, corresponding respectively to the number of attributes considered at each split ($mtry$), the minimum size of terminal nodes and the number of trees (n_{tree}) in the set.

5 Evaluation Metrics

To ensure result comparability, the same evaluation metrics adopted in the benchmark studies associated with the datasets used in this work were applied to each task. This approach allows for a direct comparison between the performance of the models developed and those reported in the literature, minimizing any discrepancies arising from different metric choices.

For the summarization task, the metrics used are the ROUGE-1, ROUGE-2 and ROUGE-L scores, calculated in the F1-score variant. The ROUGE scores measure the lexical overlap between the summaries generated by the models and the reference summaries, and the F1 variants allow for differences in text length to be taken into account. The use of the error rate also provides a direct measure of the deviation from human summaries, ensuring consistency with what is reported in the paper "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies".
For the classification task, the accuracy is adopted as the main indicator. This choice is supported by the even distribution of the four classes considered in the AGNews dataset, a condition in which accuracy represents a reliable measure of overall performance. The use of the error rate allows for immediate comparison of the model results with the established benchmarks set out in the reference paper "Character-level Convolutional Networks for Text Classification".

6 Results

This section illustrates the results obtained from the analyses conducted, with the aim of providing a clear and systematic overview of the performance of the methods applied. The main experimental results will be presented, accompanied by a critical interpretation highlighting their strengths and limitations, in order to contextualize the observations with respect to the methodological choices discussed in the previous sections.

6.1 Summarization Results

The following table summarises the results obtained by comparing the classic version of TextRank, based exclusively on TF-IDF weights, with the enhanced variant using contextualized BERT embeddings, also including the scores reported in the Newsroom paper as a baseline reference. The values reported correspond to the average ROUGE-1, ROUGE-2 and ROUGE-L scores calculated on the same test dataset.

Method	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
TextRank (w/o BERT)	**0.2288**	0.0856	**0.1760**
TextRank (with BERT)	0.2269	**0.0892**	0.1745
TextRank (paper)	0.2230	0.7870	0.1775

Table 1: Average scores for ROUGE-1, ROUGE-2, and ROUGE-L obtained by classic TextRank, TextRank with BERT, and the Newsroom paper baseline on the test dataset.

The results of Table 1 show that the two versions of TextRank have comparable performance on average ROUGE scores. The classic version, based exclusively on TF-IDF weights, achieves an average ROUGE-1 of 0.2288, an average ROUGE-2 of 0.0856, and an average ROUGE-L of 0.1760. The variant integrated with contextualised BERT embeddings shows similar values, with an average ROUGE-1 of 0.2269, an average ROUGE-2 of 0.0892, and an average ROUGE-L of 0.1745.
These results indicate that, at an aggregate level, the differences between the two sentence representation modes are limited. Despite introducing deeper semantic information through BERT, the average ROUGE scores remain close to those obtained by the classic version of TextRank, suggesting an overall stable performance on this test set.

6.2 Classification Results

The following table shows the results obtained from the text classification models described above, including Logistic Regression, Linear SVC, RF-Miner, and CNN. For completeness, the model described in the reference paper is also reported, used as a baseline for comparison, although it was not directly implemented in this work. The values indicate the average accuracies calculated on the test dataset, allowing us to compare the performance of the different approaches and observe the differences in performance between the various experimental configurations.

Model	**Accuracy**
Logistic Regression	0.7268
Linear SVC	0.7247
Random Forest (RF–Miner)	0.7395
CNN	**0.7616**
Reference paper (ngrams TF–IDF)	0.9236

Table 2: Comparison of classification performance on the AGNews dataset. The accuracy of the reference paper is derived from the error rate (1 - accuracy) of 7.64%.

The results show that the linear models, Logistic Regression and Linear SVC, perform very similarly, with an accuracy of around 72.5%, confirming their role as robust and efficient baselines. The RF-Miner model slightly improves performance, reaching around 74%, thanks to the ensemble's ability to capture non-linear interactions between features. CNN is the best model among those implemented, with an accuracy of around 76%, highlighting the advantages of neural models in treating text as a sequence and learning local patterns and word combinations not captured by bag-of-words-based methods.

7 Discussion

7.1 Interpretation of results and justification of models

The comparative analysis conducted highlights distinct dynamics between the two text mining tasks considered. In the context of summarization, the integration of contextualized BERT embeddings into the TextRank algorithm does not produce the expected performance gap compared to the statistical baseline based exclusively on TF-IDF. This phenomenon suggests that, in journalistic contexts such as those of the Newsroom corpus, informational relevance is strongly anchored to lexical recurrence and sentence position, making the deep semantic component less decisive for calculating centrality in the graph than terminological overlap.

With regard to classification, the results show a performance hierarchy consistent with the architectural complexity of the models. Linear models (Logistic Regression and Linear SVC) prove to be robust baselines for high-dimensional data, although limited by their ability to capture only linear relationships. The Random Forest Miner model shows increased effectiveness due to its ensemble nature, which allows it to handle non-linear interactions between terms through unpruned decision trees, optimizing discrimination between classes. Finally, the 1D Convolutional Neural Network achieves the best accuracy among the models implemented. The superiority of CNN lies in its intrinsic ability to extract local patterns and sequential relationships (n-grams) through convolutional filters, overcoming the limitations of bag-of-words-based models that ignore word order.

7.2 Critical comparison and evaluation metrics

The adoption of standardized metrics allows for an objective evaluation of the practical value of the models. In summarization, the use of the ROUGE score highlights how both TextRank approaches are effective in producing summaries with good lexical overlap compared to human references, although limited by their purely extractive nature, which does not allow for paraphrasing. In classification, the use of the error rate on a balanced dataset such as AGNews

provided a direct measure of generalization ability, confirming that neural models significantly reduce systematic error compared to linear classifiers.
However, a substantial divergence emerged from the results reported in the reference literature for the classification task. While the benchmark of the original paper achieves excellent performance, the models implemented here perform at lower levels. This discrepancy is not attributable to an intrinsic ineffectiveness of the models, but to a specific methodological choice: the adoption of a uniform and simplified preprocessing pipeline is necessary due to computational constraints. The aim of the study was to ensure a fair comparison between different paradigms within a shared framework, prioritizing methodological consistency over the extreme optimization of the individual classifier.

7.3 Study Limitations

The study presents several methodological limitations, primarily related to computational constraints and model assumptions:

- **Feature dimensionality**: The restriction to 200 terms for the Vector Space Model, required to ensure the feasibility of training on the AGNews dataset, reduced the available informational space for the models. This limitation negatively affected classification accuracy compared to approaches that rely on full vocabularies or extended n-gram representations.
- **Data sampling**: For the summarization task, memory constraints necessitated random sampling of the observations (500,000 instances for training and 1,000 for validation), potentially reducing the statistical representativeness of the stylistic diversity present in the million-article Newsroom corpus.
- **BERT architecture**: The use of `bert-base-uncased` with a maximum sequence length of 512 tokens, combined with mean pooling for vector aggregation, may have flattened subtle semantic nuances required to effectively distinguish highly similar sentences in the extractive ranking process.

8 Conclusion and Future Developments

The present study enables a systematic evaluation of the effectiveness of different text mining paradigms applied to the tasks of summarization and classification. With respect to text summarization on the Newsroom corpus, the results indicate a substantial equivalence in performance between the traditional formulation of the TextRank algorithm and its BERT-enriched variant, as evidenced by nearly identical ROUGE scores. This finding suggests that, within the considered journalistic domain, informational relevance is often encoded in the explicit lexical structure of the text, rendering the additional contribution of deep distributional semantics marginal. In the context of classification on the AGNews dataset, the analysis instead revealed a clearer performance hierarchy among the adopted models: the 1D convolutional neural network achieves the highest performance, followed by the Random Forest Miner model, while linear models provide a solid but less flexible baseline. This trend confirms that, for complex discriminative tasks, the use of architectures capable of capturing local patterns or nonlinear interactions yields measurable gains in accuracy compared to simpler approaches.
The comparison among the different models also provides relevant methodological insights. In particular, the analysis of alternative textual representations highlighted that the introduction of contextual embeddings does not necessarily guarantee performance improvements in extractive summarization techniques, at least in settings characterized by strong lexical regularity. Conversely, in the classification task, increased model complexity translates into tangible benefits, revealing a clear trade-off between model simplicity and expressive capacity. An additional

aspect concerns the crucial role of preprocessing: the observed gap with respect to reference benchmarks reported in the literature, which achieve significantly higher accuracy values, can be attributed both to the adoption of a uniform preprocessing pipeline and to the intentional limitation of the feature space, fixed at 200 informative terms to ensure comparability across models. Although this choice constitutes a bottleneck for absolute performance, it strengthens the robustness of the experimental comparison.

In light of the identified limitations, several future developments can be envisaged to enhance the adopted analytical framework. A first direction concerns architectural evolution, through the integration of full Transformer-based models also for the classification task, in order to overcome the limitations of CNNs in modeling long-range dependencies. In parallel, expanding the feature set by removing the 200-term constraint and introducing higher-order n-grams could help bridge the performance gap with the reference literature. Further improvements may arise from the adoption of more advanced tuning strategies, including task-specific fine-tuning of pre-trained language models and a more exhaustive hyperparameter optimization compared to the Grid Search procedure employed. Finally, complementing quantitative metrics with qualitative evaluation would allow the analysis of aspects such as logical coherence and readability, which metrics based on lexical overlap are unable to fully capture.

The study also lends itself to several extensions aimed at assessing its generalizability. Applying the same methodologies to more recent datasets or to different domains, such as social media or legal texts, would enable an evaluation of model robustness in less structured contexts than the journalistic domain. Another promising research direction involves the exploration of abstractive summarization, to assess whether generative models can offer significant advantages over the purely extractive approaches analyzed. Finally, cross-task validation could investigate the impact of summarization outputs on the downstream performance of a classifier, paving the way for the development of integrated and multi-stage text mining pipelines.

References

Boldi, Paolo, Paolo Ferragina, Santini, Massimo, Vigna, and Sebastiano (Jan. 2005). "PageRank as a function of the damping factor". In: DOI: 10.1145/1060745.1060827.

Grusky, Max, Mor Naaman, and Yoav Artzi (2018). "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 708–719. DOI: 10.18653/v1/N18-1065. URL: https://aclanthology.org/N18-1065/.

Luo, Xin (Jan. 2018). "An improved text classifier based on random forest algorithm - comparative studies on multiple text classifiers". In: DOI: 10.2991/macmc-17.2018.39.

Yulianti, Evi, Nicholas Pangestu, and Meganingrum Jiwanggi (Oct. 2023). "Enhanced TextRank using weighted word embedding for text summarization". In: *International Journal of Electrical and Computer Engineering (IJECE)* 13, p. 5472. DOI: 10.11591/ijece.v13i5.pp5472-5482.

Zhang, Xiang, Junbo Zhao, and Yann LeCun (2015). "Character-level Convolutional Networks for Text Classification". In: *Advances in Neural Information Processing Systems*. Vol. 28. NIPS 2015. Curran Associates, Inc., pp. 649–657. URL: https://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf.