ISSN 2007-9737

Enhancing Text Classification Using BERT: A Transfer Learning Approach

Haider Zaman-Khan[1], Muddasar Naeem[3], Raffaele Guarasci[2,*], Umamah Bint-Khalid[1], Massimo Esposito[2], Francesco Gargiulo[2]

[1] Quaid-i-Azam University, Pakistan

[2] Institute for High Performance Computing and Networking, National Research Council, Italy

[3] Giustino Fortunato University, Italy

{haiderzkhan, umamahkhalid}@ele.qau.edu.pk, m.naeem@unifortunato.eu, {raffaele.guarasci, massimo.esposito, francesco.gargiulo}@icar.cnr.it

Abstract. This paper investigates the application of Natural Language Processing (NLP) techniques for enhancing the performance of document-level classification tasks. The study focuses on leveraging a Transformer-based Neural Language Model (NLM), particularly BERT, combined with cross-validation to exploit transfer learning algorithms for classification tasks. To address the challenges, the approach has been tested on the two different types of the widely-known 20 Newsgroups benchmark dataset using pre-trained BERT models refined through cross-validation, resulting in notable accuracy rates of 92.29% for the pre-processed dataset without noise and 90.08% for the raw filtered dataset. These encouraging results confirm the effectiveness of combining transfer learning, cross-validation, and NLMs in NLP, with a particular focus on the state-of-the-art performance achieved by pre-trained BERT models in real-world text classification tasks.

Keywords. NLMs, transfer learning, text classification, BERT.

1 Introduction

In recent years, there has been an exponential growth in Natural Language Processing (NLP), mainly due to the emergence of Neural Language Models (NLMs). The introduction of Transformer-based models, initiated by the release of pioneering models such as BERT [10], has guided a new era in NLP methodologies.

These advances have steadily boosted achievable performance in a plethora of tasks in different domains across different languages [15, 19], ranging from general-purpose tasks such as sentiment analysis [27], information extraction [37, 18] or anaphora and coreference resolution [34, 43, 20], to very domain-specific verticalized approaches [47, 21, 29, 25, 36, 46]. Concurrently, resources, corpora, and collections of benchmark datasets suitable for different applications [5, 38] have flourished as the models have progressed in complexity, requiring a massive amount of annotated data in training. Although text classification has always occupied a priority place among the ”classic” NLP tasks, this enormous data availability has given it a crucial role [24].

In the NLP field, text classification is usually formalised as follows: automatically classifying and arranging massive amounts of textual input into predetermined categories or classes. An

algorithm is trained on a labelled dataset of examples and their pre-tagged classes for text classification or categorisation.

The system uses this training data to generalise and categorise new, unseen text into one of the pre-defined groups. The primary issue in text categorisation is finding pertinent details or patterns that distinguish the various groups. Naive Bayes, Support Vector Machines, Decision Trees, Random Forests, and Neural Networks are the few well-known machine learning techniques for text categorisation [24, 35]. CNN and RNN have recently exhibited outstanding results in text categorisation tests [4].

These models can develop meaningful representations of the text data and capture complicated linkages and dependencies in the text [54]. However, text classification faces various limitations hindering its effectiveness. Some are inherent limitations to natural language processing, such as ambiguity, multi-word expressions, specialised lexicons, and language-dependent phenomena, which have always been challenging for NLMs.

Another type of problem stems from extrinsic limitations, such as unbalanced datasets, models not trained on a specific domain, onerous computational costs, or ineffective evaluation metrics. Starting from these premises, this work aims to enhance the accuracy of multi-class text classification by leveraging pre-training methodologies and exploiting the language abilities of BERT.

While this NLM has demonstrated its effectiveness in various language understanding (NLU) tasks, its adaptation for multi-class categorisation remains a debated topic with many open issues. Besides using BERT as NLM, the present approach benefits from the transfer learning algorithm and integration of cross-validation techniques; this hybrid integration helps overcome traditional models in terms of accuracy and resource efficiency. The annotated dataset chosen to experiment is the 20 Newsgroups collection [2], a *de facto* benchmark dataset for multi-class text classification in NLP.

The dataset has been cleaned up and prepared to be optimised for the fine-tuning phase of BERT, to which a classification layer has been added. Concerning evaluation, comparative analyses with existing models estimate the methodology's effectiveness in content classification and sentiment analysis.

The paper is structured as follows. Section 2 provides an overview of the recent related works. Section 3 describes the research methodology, including the NLM and dataset details. Section 4 describes the experimental assessment and then presents and discusses the results. Finally, Section 5 summarises the paper and hints at future developments.

2 Related Work

Rehman et al. [4] proposed a filter-based feature selection algorithm called Normalized Difference Measure (NDM) for text classification tasks. The study compared the performance of NDM with seven other feature selection algorithms (ODDS, CHI, IG, DFS, GINI, ACC2, POISON) using Support Vector Machine (SVM) and Naive Bayes (NB) classifiers.

The research aimed to demonstrate how removing irrelevant and redundant features through NDM could enhance the performance of text classification models. The experimentation conducted by Rehman et al. showed that NDM significantly improved the classification accuracy on the 20 Newsgroups dataset.

The study highlighted the importance of feature selection in optimizing text classification models. It showcased the effectiveness of NDM in enhancing classification performance compared to other feature selection algorithms when used in conjunction with SVM and NB classifiers. Another study claims that to improve on their earlier Normalized Difference Measure (NDM) algorithm, Rehman et al. [4] developed a new version of the filter-based feature selection algorithm called Maximum Margin Ranking (MMR). The study combined MMR's effectiveness for text classification tasks with SVM and NB classifiers to assess MMR's effectiveness for text classification tasks.

By comparing MMR with NDM, the researcher aimed to demonstrate the improvements in

Table 1. 20news-18828.tar.gz(Dataset 1) and 20news-19997.tar.gz(Dataset 2) dataset types, changes and their samples

Dataset name	Changes	Samples
20news-18828.tar.gz	Duplicates removed, only ”From” and ”Subject” headers	18828
20news-19997.tar.gz	Original 20 Newsgroups data set	19997

classification accuracy achieved through the enhanced feature selection algorithm. The experimentation conducted by Rehman et al. showed that MMR outperformed NDM, achieving a significant performance improvement.

The study highlighted the importance of continuous refinement and development of feature selection algorithms to enhance the effectiveness of text classification models, showcasing the advancements made by MMR in improving classification accuracy when integrated with SVM and NB classifiers. Lai et al. [4] introduced a Modified CNN (Convolutional Neural Network) approach for text classification.

Their model utilized a multi-channel CNN architecture with specific modifications to enhance performance. The study focused on classifying text documents into four classes: comp, politics, rec, and religion. The Modified CNN model achieved an impressive accuracy of 96.49% on these four classes, showcasing the effectiveness of their approach in text classification tasks. Aziguli et al. [4] proposed an Autoencoder-based approach for text classification.

Their methodology utilized a denoising deep neural network (DDNN) that incorporated a restricted Boltzmann machine (RBM) and denoising autoencoder (DAE). By leveraging the DDNN model, the researchers aimed to reduce noise in the data and improve feature extraction for text classification tasks. The Autoencoder approach successfully enhanced feature extraction performance, showcasing its potential in text classification applications.

Jiang et al. [4] proposed a hybrid text classification model that combined a Deep Belief Network (DBN) with Softmax regression for text classification tasks. The DBN was utilized for feature extraction, while Softmax regression was employed to classify textual data. This hybrid approach addressed the challenge of computing high-dimensional sparse matrices in text classification.

The researchers reported that their hybrid methodology outperformed traditional classification methods on benchmark datasets, highlighting the effectiveness of combining DBN with Softmax regression for text classification. Liu et al. [4] presented an attentional framework based on deep linguistics that incorporated concept information from meta-thesauri into neural network-based classification models.

The researchers utilized MetaMap and WordNet to annotate biomedical and general text, respectively, enhancing the understanding of text content for classification tasks. By leveraging meta-thesauri and deep learning techniques, Liu et al. aimed to improve the performance of text classification models by incorporating rich concept information into the classification process.

Shih et al. [4] researched using Siamese Long Short-Term Memory (LSTM) networks for text categorization. The researchers proposed a deep learning methodology based on Siamese LSTM networks to enhance the learning of document representations for text classification tasks. By leveraging LSTM networks, Shih et al. aimed to improve the performance of text classification models by effectively capturing the sequential dependencies and context within textual data. Their study demonstrated promising results, achieving a performance of 86% on the 20 newsgroup dataset, showcasing the effectiveness of LSTM-based approaches in text classification. Shirsat et al. [4] researched sentiment identification at the sentence level using positive and negative word lists from the Bing Liu dictionary.

The study used machine learning techniques for sentiment analysis tasks on news articles, specifically Support Vector Machine (SVM) and Naive Bayes (NB) classifiers. Shirsat et al.

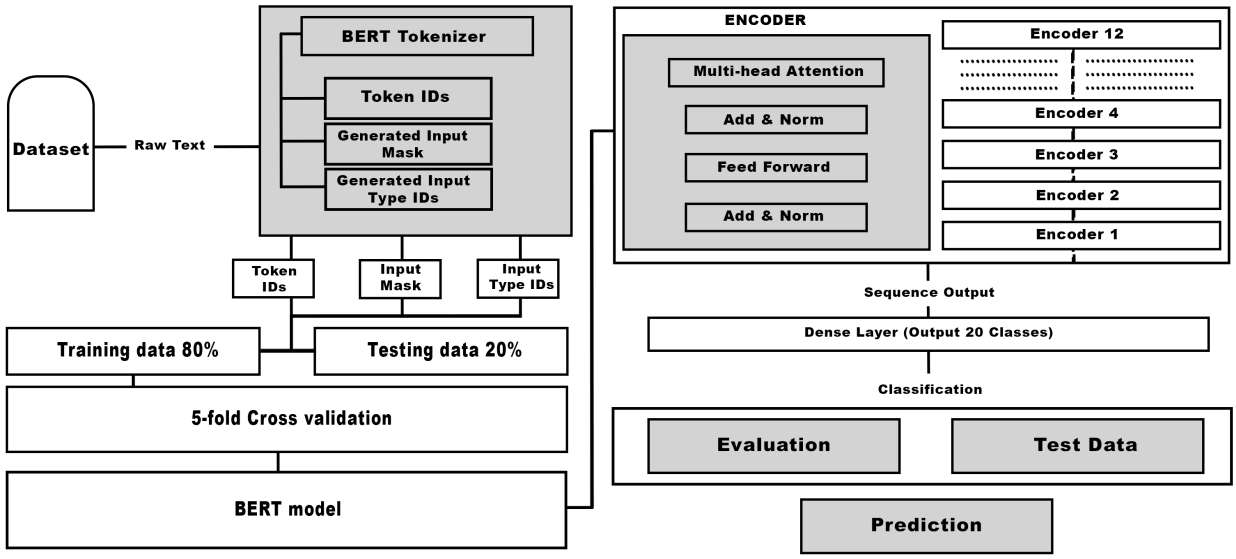


Fig. 1. Proposed pre-trained BERT models with strategic five-fold cross-validation methodology structure

reported a performance of 96% using the SVM classifier on the BBC news dataset, highlighting the effectiveness of SVM in sentiment identification at the sentence level. The study showcased the application of traditional machine learning algorithms like SVM and NB in sentiment analysis tasks, emphasizing their performance in text classification.

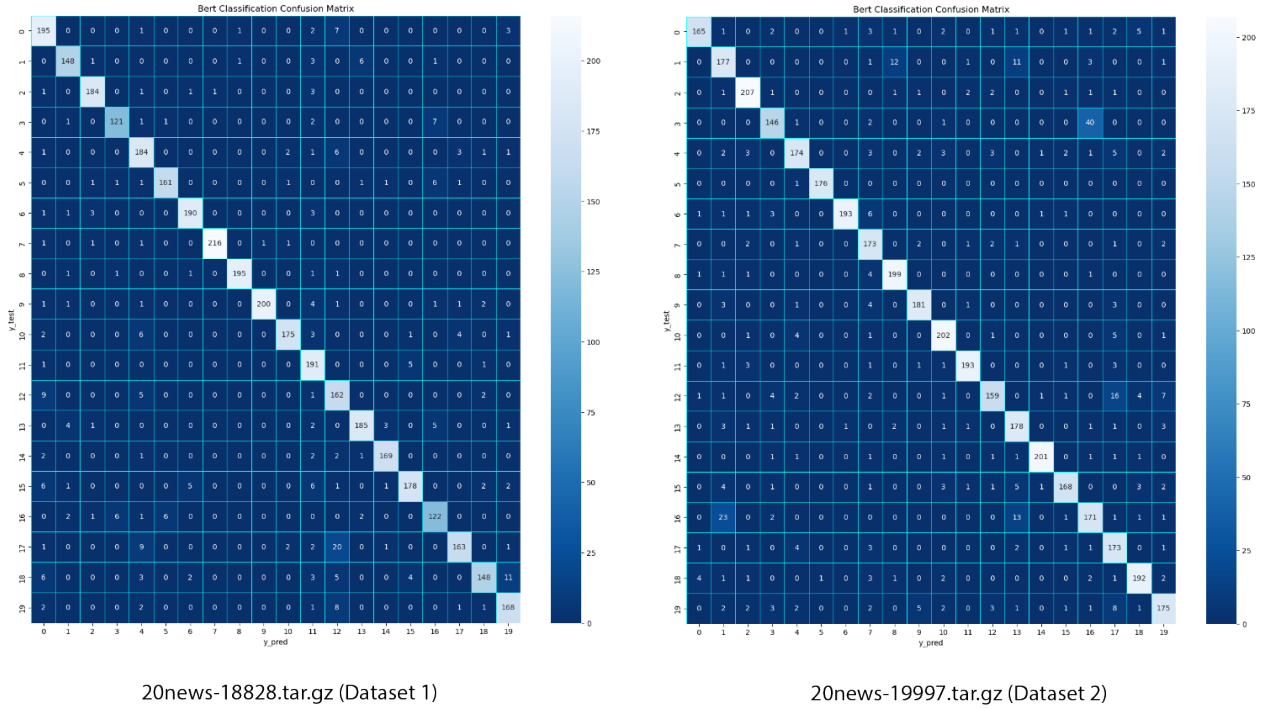
Camacho-Collados and Pilehvar [4] conducted a study on the role of text preprocessing in neural network architectures for text categorization and sentiment analysis tasks. The researchers evaluated different preprocessing practices to train word embeddings for text classification. They experimented with two versions of Convolutional Neural Networks (CNN): a standard CNN with Rectified Linear Unit (ReLU) activation function and a standard CNN with the addition of a recurrent layer (LSTM).

Their study aimed to enhance the accuracy of text classification models by optimizing text preprocessing techniques and neural network architectures. The experimentation achieved an accuracy of 97% on the BBC dataset and 90% on the 20-newsgroup dataset using six classes, demonstrating the impact of effective preprocessing on improving the performance of CNN and LSTM-based models for text classification.

Pradhan et al. [4] compared different machine learning classifiers for text classification tasks on news articles. The researchers evaluated the performance of various machine learning algorithms on topic categorization. Pradhan et al. compared the effectiveness of different classifiers in accurately categorizing news articles into relevant topics.

The study aimed to identify the most suitable machine learning classifier for text classification based on the performance metrics evaluated. Their research provided insights into the comparative analysis of machine learning classifiers for text classification tasks, contributing to understanding the strengths and limitations of different classification algorithms in handling textual data.

Elghannam et al. [4] focused on text representation and classification based on a bi-gram alphabet approach. The study proposed a methodology that utilized bi-gram frequencies for representing documents in a typical machine learning-based framework. By leveraging bi-gram features, Elghannam aimed



20news-18828.tar.gz (Dataset 1)

20news-19997.tar.gz (Dataset 2)

Fig. 2. Confusion matrices of pre-trained BERT model for two different 20Newsgroup dataset

to address the challenge of data sparsity and improve the representation of textual data for classification tasks.

The research did not rely on Natural Language Processing (NLP) tools and demonstrated significant improvements in alleviating data sparsity. Elghannam reported an F1 score of 92% on the BBC news dataset, highlighting the effectiveness of the bi-gram alphabet approach in feature representation for text classification tasks.

The study contributed to enhancing the performance of machine learning classifiers by optimizing the representation of textual data through bi-gram features. Wang et al. [4] presented a transfer learning method for text classification in cross-domain scenarios.

The study addressed the challenge of classifying text data from different domains by leveraging transfer learning techniques. Wang et al. conducted experiments on six classes of the 20 newsgroup dataset to evaluate the performance of their transfer learning approach.

By transferring knowledge from one domain to another, the researchers aimed to improve the classification accuracy in cross-domain text classification tasks.

The methodology proposed by Wang et al. demonstrated promising results, achieving a performance of 95% on the 20 newsgroup dataset.

The study highlighted the effectiveness of transfer learning in enhancing text classification models' performance across different domains, showcasing the potential of transfer learning techniques in handling cross-domain text classification tasks.

Asim et al. [4] proposed a two-stage text document classification methodology that combines traditional feature engineering with automatic feature engineering using deep learning. The methodology employs a filter-based feature selection algorithm to develop a noiseless vocabulary fed into a multi-channel Convolutional Neural Network (CNN).

Each CNN channel consists of two filters of different sizes and two dense layers. By utilising wide convolutional layers, the methodology aims to address the issue of unequal feature convolution in traditional CNN models. Experimental results showed that feeding only the most discriminative features of the vocabulary to the CNN model improved performance significantly compared to using the entire vocabulary. The study also discussed the potential for further assessment of the methodology using Recurrent Neural Networks (RNN) and other hybrid deep learning approaches.

In another research work, Asim et al. [4] proposed a robust hybrid approach for textual document classification, which combines traditional feature engineering with automatic feature engineering using deep learning techniques. The methodology involves a two-stage classification process: the first stage focuses on feature selection using a filter-based algorithm to develop a noiseless vocabulary. In contrast, the second stage utilises a multi-channel Convolutional Neural Network (CNN) model for classification.

By integrating traditional and deep learning approaches, the proposed methodology aims to improve the classification accuracy of textual documents by addressing issues such as data sparsity and feature representation. The study demonstrates the effectiveness of the hybrid approach in outperforming state-of-the-art machine learning and deep learning-based text classification methodologies on public datasets.

In Natural Language Processing (NLP), text classification is a common task that has been studied for a long time. One popular model used for this task is BERT, which is based on transformers. Researchers often use BERT to see how well it works compared to other techniques. However, they face challenges like overfitting (when the model fits too closely to the training data), imbalanced classes (when some classes have too few examples), low performance, and issues with computational resources.

Many research papers have been done to tackle these problems and improve text classification performance using models like BERT. Yusuf et al. [3] proposed evaluating pre-trained language models (PLMs) for multi-class text classification in finance. They followed BERT's methodology, testing model performance at 1, 3, and 5 epochs using the Adam optimizer. Challenges like convergence issues or overfitting may have impacted model effectiveness, emphasising the need for careful parameter selection. Their study aims to compare PLMs' effectiveness in financial text classification and address potential evaluation challenges.

Vedangi et al. [49] proposed using the BERT model to enhance the accuracy of long document classification tasks, particularly focusing on the 20 Newsgroups (20NG) dataset. To optimise the model's performance, the authors implemented the Adam optimizer during the model training process, alongside gradient descent and a cross-entropy loss function. The primary purpose of employing the BERT model was to leverage its advanced capabilities in capturing complex relationships within text data, thereby improving the overall classification accuracy on the dataset of interest.

Wang et al. [50] proposed training downstream models with an Adam optimizer using a learning rate of 0.001. The models were trained for 140 epochs with early stopping based on accuracy evaluated on the development set. The study highlighted the advantages of BERT over ELMo in datasets such as 20NewsGroup, Reuters, and AAPD, although specific accuracy values were not provided in the excerpts.

The research aimed to conduct a comprehensive comparative analysis of word embeddings using CNN and BiLSTM as downstream encoders for text classification, aiming to offer evidence-based guidance for practitioners selecting word embeddings for deep learning models in text classification tasks . Taneja et al. [45] explored transfer learning and traditional machine learning for text classification, comparing BERT and DistilBERT with TF-IDF. They fine-tuned these models on the 20 Newsgroups dataset, achieving 96% accuracy across five classes.

The study's limitations include dataset dependency and the computational demands of large models. Overall, it aimed to shed light on the effectiveness of different approaches in text classification tasks. The task of text classification has a long history in NLP, and it can be canonically

divided into four different levels based on their scope and granularity of categories:

1. Document-level classification: This level involves categorising entire documents or individual texts into predefined groups or classes (i.e., categorising news into sports, entertainment, or politics). This classification is usually best suited for content filtering, document management, and aggregation tasks [48].
2. Sentence-level classification [23]: Individual sentences contained in a document are classified. This type of classification has been very successful in tasks such as sentiment analysis [12], opinion mining [30] or chat-bots.
3. Entity-level classification: It focuses on the categories, i.e., named entities. These entities can be general-purpose (people, organisations, and locations) or domain-tailored. This approach is typically used in tasks such as in information extraction [16, 17] or knowledge management [26].
4. Aspect-level classification: This level involves identifying and categorising specific aspects or attributes of a product, service, or topic within a document. It is commonly used in opinion mining, customer feedback analysis, and product recommendation systems[51].

The emergence of word embedding [32, 31] and neural language models (NLMs) [10, 1] marked a pivotal shift in the field, raising the bar in all NLP tasks. In particular, BERT was one of the first NLMs to rank tasks. Its ability to capture bidirectional contextual information makes it particularly effective in classification tasks once fine-tuned on specific classification objectives [44]. In addition, the model achieved excellent performance in topic-specific identification.

In BERT, the potential to enhance topic modelling methodologies and facilitate more profound insights into large-scale text datasets has been demonstrated. Subsequently, a BERT-derived model, RoBERTa [28], characterised by an augmented pre-training using larger datasets combining Common Crawl and BooksCorpus, dynamic masking, longer sequences, and excluding next sentence prediction, has achieved superior performance in text classification, highlighting the impact of optimised training strategies.

RoBERTa has also been adapted to perform the task in several languages [52]. [41] has proposed a different approach. This approach uses pre-trained word embedding to capture local n-gram features through CNNs and sequential dependencies through LSTMs. This combination achieves state-of-the-art performance on benchmark datasets like SST-2, MR, and TREC, demonstrating the potential to combine deep learning architectures for improved text classification accuracy.

The methodology utilises a joint training strategy with Adam optimiser and learning rate decay to optimise the hybrid CNN-LSTM network effectively. In [22], transfer learning using the ULMFiT model is exploited for text classification tasks. Significant improvements can be noted by fine-tuning pre-trained models on specific tasks compared to training from scratch. The approach adopted in this study entailed refining the ULMFiT model through fine-tuning text classification datasets such as AG News and SST-2, employing the Adam optimiser with learning rate decay.

Results confirm the effectiveness of transfer learning methodologies in enhancing NLP tasks. A hybrid approach is instead presented in [4], combining traditional feature engineering and deep learning. Using the 20 Newsgroups dataset, this approach uses a filter-based feature selection algorithm followed by a deep convolutional neural network. This two-stage methodology achieves significant accuracy improvements compared to traditional and deep learning-based approaches, paving the way for combining multiple techniques for text classification tasks.

Among more recent works using BERT or BERT-based models for text classification in [9], a text classification system for academic papers is proposed, leveraging a hybrid BERT and Bidirectional Gated Recurrent Unit (BiGRU) model. BERT extracts semantic features from

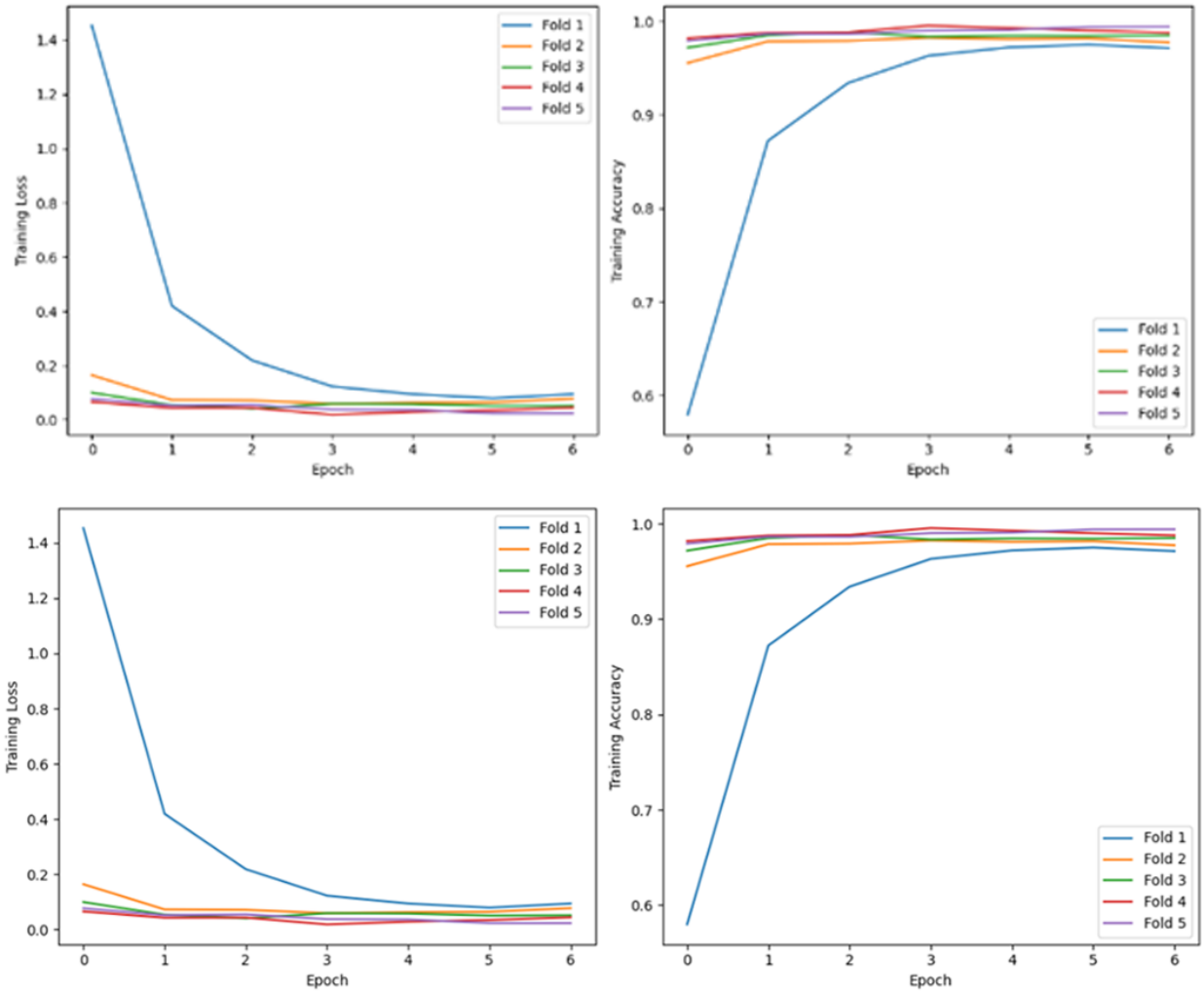


Fig. 3. Visualisation of model performance: the first row represents outcomes derived from (Dataset 1), while the second row illustrates results from (Dataset 2). Within each row, the first and second graphs correspondingly portray the training loss and training accuracy over the training epochs across the five chosen cross-validation folds

paper abstracts, while BiGRU captures sequential information.

The model is evaluated on a dataset comprising 10,000 academic papers from four disciplines, demonstrating superior performance compared to several baselines.

Furthermore, [42] introduces a BERT-based hybrid recurrent neural network (RNN) model for multi-class text classification, focusing on the impact of pre-trained word embedding. BERT is employed to acquire contextualised word embedding, fed into an RNN with an attention mechanism for classification.

Experimental evaluations are conducted on three datasets: IMDb movie reviews, AG news, and Yelp reviews, highlighting the effectiveness of their proposed model. Finally, in [33], an extensive review of over 150 deep learning-based models and more than 40 widely-used datasets for text classification is presented.

	precision	recall	f1-score	support
0	0.85	0.93	0.89	209
1	0.93	0.93	0.93	160
2	0.96	0.96	0.96	191
3	0.94	0.91	0.92	133
4	0.85	0.92	0.88	199
5	0.96	0.93	0.94	174
6	0.95	0.96	0.96	198
7	1.00	0.98	0.99	221
8	0.99	0.97	0.98	200
9	1.00	0.94	0.97	212
10	0.97	0.91	0.94	192
11	0.83	0.96	0.89	198
12	0.76	0.91	0.83	179
13	0.95	0.92	0.93	201
14	0.97	0.95	0.96	177
15	0.95	0.88	0.91	202
16	0.86	0.87	0.87	140
17	0.94	0.82	0.88	199
18	0.94	0.81	0.87	182
19	0.89	0.92	0.91	183
accuracy			0.92	3750
macro avg	0.92	0.92	0.92	3750
weighted avg	0.93	0.92	0.92	3750

Dataset 1

	precision	recall	f1-score	support
0	0.95	0.88	0.92	187
1	0.80	0.86	0.83	206
2	0.93	0.95	0.94	218
3	0.89	0.77	0.82	190
4	0.91	0.87	0.89	201
5	0.99	0.99	0.99	177
6	0.99	0.93	0.96	207
7	0.82	0.94	0.88	185
8	0.92	0.96	0.94	207
9	0.94	0.94	0.94	193
10	0.92	0.94	0.93	215
11	0.96	0.95	0.96	204
12	0.92	0.80	0.85	199
13	0.84	0.92	0.88	193
14	0.98	0.96	0.97	210
15	0.94	0.88	0.91	190
16	0.76	0.80	0.78	213
17	0.78	0.93	0.85	187
18	0.93	0.91	0.92	210
19	0.88	0.84	0.86	208
accuracy			0.90	4000
macro avg	0.90	0.90	0.90	4000
weighted avg	0.90	0.90	0.90	4000

Dataset 2

Fig. 4. Classification report of 20news-18828.tar.gz(Dataset 1) and 20news-19997.tar.gz(Dataset 2)

3 Materials and Methods

This section describes the dataset and NLM used in the experiment in detail. In particular, the configuration and parameters used considering the BERT model are specified; further word embedding is described, and the 20 Newsgroups collection is introduced, specifying the underlying reasons for this choice.

The dataset chosen in this research is the 20 Newsgroups collection [2], a widely recognised benchmark dataset in NLP and text classification. The dataset spans different topics, including posts extracted from various newsgroups related to sports, politics, technology, and news.

This dataset offers a rich repository of textual content suitable for multi-class text classification tasks. A preprocessing step was necessary to prepare the dataset before it could be used in the experiments.

First, documents were cleaned up by removing duplicates to mitigate potential biases and redundancy, enhancing the accuracy and reliability of model training and evaluation. Subsequently, only each document's "From" and "Subject" headers contain pertinent information for predicting the corresponding category or topic. After that, classical preprocessing operations, widely known in the literature, were performed.

The WordPiece tokenization technique has been chosen, facilitating the decomposition of words into sub-word units. Furthermore, the tokenized data is shuffled and split into training and validation subsets, ensuring a balanced representation across both sets. This preprocessing phase helps optimise the dataset for fine-tuning BERT models tailored for multi-class text classification tasks.

Concerning evaluation, classical metrics like accuracy, precision, recall, and F1 score are taken into account to estimate performance using two independent subsets: the pre-processed *20news-18828.tar.gz* dataset with 18,828 unique documents and the raw *20news-19997* dataset with the original samples.

Removing duplicates from 20news-18828 ensures reliable training and evaluation by

eliminating biases and redundancy, while 20news-19997 retains the raw data (see Table 2).

3.1 Model

The paper presents a novel methodology focusing on fine-tuning a BERT model for text classification, considered a standard baseline model in contemporary natural language processing tasks. Recent studies, including experiments with multi-text classification using the 20 Newsgroups dataset, have encountered significant challenges such as model forgetting unseen data, leading to overfitting, class imbalance, low performance, and high computational demands, ultimately resulting in lower accuracies.

To address these challenges comprehensively, the primary objective of this research is to optimize the fine-tuned BERT model using advanced cross-validation techniques. This optimization strategy includes leveraging the Radam optimizer to enhance the model's learning capabilities and improve its performance.

The methodology involves training the fine-tuned BERT model with seven epochs on five-folds, encompassing all available data, including previously unseen data samples. The results of this approach have demonstrated significant improvements, achieving an impressive 92% accuracy on pre-processed data and maintaining a solid 90% accuracy on original data, even with noisy elements.

Notably, these improvements were achieved relatively quickly and with limited computational resources. Comparative analysis with baseline models highlights the superior performance of the proposed pre-trained BERT model, showcasing its ability to outperform traditional methods in text classification tasks.

Furthermore, the research aims to explore additional optimization strategies in future directions. These strategies may include fine-tuning model parameters or investigating ensemble techniques to enhance further the robustness and generalizability of the fine-tuned BERT model, thereby addressing key limitations observed in current approaches. As depicted in Figure 1, the proposed methodology involves several key steps. First, a large-scale labelled dataset for fine-tuning regarding multi-class text classification tasks is collected. Next, an appropriate BERT model, already pre-trained on natural language data to perform NLU, is selected and acquired. The BERT is a pre-trained deep learning model for NLP developed in 2018 by Devlin *et all* at Google [39].

It uses the transformer architecture. The key attribute of BERT is its ability to pre-train huge amounts of text data using a distributed computing system and memorise the underlying patterns and relationships in the language syntax and semantics. This pre-training is done through masked language modelling (MLM) and next sentence prediction (NSP).

A pre-trained BERT can be easily fine-tuned on specific NLP tasks. For fine-tuning, we add a task-specific layer on top of the pre-trained BERT and then train the network on a smaller dataset related to the specific task. The effectiveness of BERT in NLP has been demonstrated through various benchmark datasets [14], wherein it has achieved advanced performance. The availability of pre-trained BERT models has also facilitated researchers and developers to apply them to various NLP tasks and achieve high levels of accuracy with less data and computing resources.

3.2 Technical Details

In our approach, we employ wordpiece tokenization for word embedding, breaking down words into sub-word units. This enhances the model's capacity to handle less common terms and ensures more accurate data representation. Wordpiece tokenization effectively dissects words into smaller units, aiding the model in understanding complex structures and dealing with rare or unknown words. Leveraging pre-trained wordpiece embedding from the BERT model, we obtain dense vector representations for each sub-word. These embeddings capture essential semantic and syntactic information, playing a vital role in the success of our multi-class text classification methodology.

In our proposed algorithm, the hidden layers of BERT include the attention layers, renowned for

Table 2. Comparison of pre-trained BERT model with state-of-the-art machine and deep learning methodologies on 20 newsgroup dataset regarding Advantages and disadvantages. More details are in related work

Model	Size of dataset	No of Models	Architecture	Pre-trained	Accuracy	Disadvantage	Performance	TC task	Classification
	20	3	Simple	No	71.10%	Not as accurate as other models	Poor	No	binary
MMR, SVM, NB	20	3	Simple	No	84.00%	Not as accurate as other models with fine-tuning	Fair	No	binary
Modified CNN	4 Classes only	2	Complex	No	96.49%	Requires large dataset to achieve high accuracy	Good	No	binary
Auto Encoder	20	2	Complex	No	773.78%	Not as accurate as other models	Poor	No	binary
DBN+Softmax	20	2	Complex	No	85.57%	Not as accurate as other models with fine-tuning	Fair	No	binary
DL and meta-thesaurus	20	2	Complex	No	69.82%	Not as accurate as other models	Poor	No	binary
LSTM	20	2	Complex	No	86.00%	Requires large dataset to achieve high accuracy	Good	Yes	binary
SVM+NB	9 classes only	2	Complex	No	82.20%	Not as accurate as other models with fine-tuning	Fair	No	binary
CNN+LSTM	Topic categorization	2	Complex	No	90.00%	Requires large dataset to achieve high accuracy	Good	Yes	binary
ML Classifier comparison	20	4	Complex	No	86.00%	Not as accurate as other models with fine-tuning	Fair	No	binary
Feature representation, ML classifiers	20	2	Complex	No	68.00%	Not as accurate as other models	Poor	No	binary
Cross-domain Transfer learning	6 classes only	3	Complex	Yes	95.62%	Requires large dataset to train the initial model	Good	Yes	binary
Multi-Channel CNN	20	2	Complex	Yes	82.76%	Requires large dataset to get better accuracy	Fair	No	binary
Feature Engineering, Multi-Channel CNN	20	2	Complex	Yes	91.72%	Requires large dataset to get better accuracy	Good	No	binary
Pre-Trained BERT	**20**	**1**	Deep bidirectional transformer	**Yes**	**92.13%**	Already trained and better accuracy	**Excellent**	**Yes**	**binary**

capturing contextual relationships between words. These attention layers are the initial hidden layers, effectively encoding input text and capturing crucial semantic information.

Additionally, we augment these layers with one or more dense layers comprising fully connected neurons. These dense layers enable the model to learn non-linear input representations, which is pivotal in extracting meaningful features for accurate predictions in multi-class text classification tasks.

As essential parts of BERT's pre-training phase, Next Sentence Prediction (NSP) and Masked Language Modeling (MLM) use the model's hidden layers to improve comprehension of textual context and semantic linkages. NSP involves training the model to predict whether two sentences are consecutive, enhancing its understanding of sentence-level relationships and coherence. On the other hand, MLM tasks the model with predicting masked words in a sentence based on the surrounding context, improving its grasp of contextualized word representations. Both NSP and MLM leverage the hidden layers of BERT to refine word embeddings and contextual embeddings, ultimately enhancing the model's performance across various natural language processing tasks.

Activation functions introduce non-linearity, which is crucial for learning intricate connections and capturing non-linear dependencies in textual data. Our approach applies the softmax activation function in the final layer.

This function normalises output probabilities across multiple classes, creating a reliable probability distribution. It ensures predicted class probabilities sum up to 1, facilitating trustworthy and interpretable predictions in multi-class classification. Utilising the softmax activation function empowers the model to assign probabilities to each class, making confident predictions based on these probabilities:

$$S(x^i) = \frac{e^{xi}}{\sum_{j=1}^{n} e^{xj}}. \tag{1}$$

The RAdam optimiser has been chosen for the optimisation algorithm. RAdam combines Adam optimiser benefits with rectified linear units, enhancing convergence speed and performance. Adam's predefined learning rate allows faster convergence and improved sparse gradient handling. RAdam's integration of rectified linear units introduces non-linearity, enabling the model to learn complex patterns and enhance generalisation. Our choice of the RAdam optimiser aims to optimise training and enhance fine-tuning for multi-class text classification.

Evaluation has been assessed considering diverse evaluation metrics, including a generated confusion matrix. This matrix provides a comprehensive overview of predictions aligned with true labels. The model's performance is scrutinised by examining true positives, true negatives, false positives, and false negatives for each class. Computed metrics like F1 score, recall, accuracy, and precision offer insights into overall prediction accuracy, false positive reduction, and minimisation of false negatives. These metrics aid in understanding the model's performance, identifying strengths and weaknesses, and guiding further optimisation.

4 Results

This section presents the results from experiments that indicate the performance of a fully trained (pre-trained + fine-tuned) BERT model for multi-class text classification by implementing the cross-validation technique. This section provides a detailed analysis of the results and discusses their implications.

The details of the experimental results are directly presented, highlighting key findings and the insights gained. The impact of different components of the proposed methodology, such as Wordpiece word embedding, BERT's attention layers, the dense layer, the soft-max activation function, and the RAdam optimiser, is analysed.

The effect of hyperparameters, parameter tuning, and architecture on the model's performance is investigated. Starting with the pre-processed dataset (Dataset 1) and original dataset (Dataset 2), evaluating metrics, including accuracy, precision, recall, and F1 score, provide insights into the model's capability to classify text into multiple classes correctly. These metrics are calculated using the test dataset for the *20news-18828.tar.gz* and *20news-19997.tar.gz dataset*. The performance of a fully trained BERT model is then examined.

The confusion matrix in the following figure lists the proportions of true positives, true negatives, false positives, and false negatives for each class. This analysis helps spot correct or incorrect behaviours during the classification process and the model's advantages and disadvantages for various classes. The data in the figure below shows that the numbers that move along the diagonal are those that were correctly derived from the whole data. The chart also shows no class imbalance issue, indicating that the model has been trained well and is in stable mode.

Figure 3 visually represents the model's weaknesses and strengths. The x-axis of the graph shows the epoch number, and the y-axis gives the accuracy. A bar plot is generated using the Matplotlib library with the help of validation data. This plot displays the number of folds, allowing visualisation of any overfitting exhibited by the model, which can impact its performance and provide insights for further analysis.

The plot showed that the BERT pretrained model had trained well on the 20newsgroup dataset. The training accuracy curves are fairly close, and there is no indication that the model is starting to plateau or decreasing the accuracy. This suggests that the model can generalise well to new unseen data as intended and that it is not simply memorising the training data.

The image shows the training accuracy of a model trained with cross-validation. Cross-validation is a widely used technique for machine learning model assessment, which involves dividing the dataset into subsets or "folds".

The model is trained on all but one fold, serving as the training set, and evaluated on the excluded validation set. This process iterates through all folds, ensuring comprehensive training and testing on the entire dataset. Its key advantages include mitigating overfitting by exposing the model to diverse data partitions

reducing the impact of data-specific nuances. This approach optimally utilises available data, enhancing the model's reliability, robustness, and generalizability. The proposed methodology employs 5-fold cross-validation. Figure 4 also shows the metrics for each class in the dataset and the overall metrics for the model. The overall metrics are calculated by averaging the metrics for each class. It also shows the number of instances in different classes.

This information can be used to calculate the class imbalance, which is the difference in the number of instances in different classes. Overall, the image shows the performance of the BERT model on a multi-class classification task. The metrics in the table can be used to evaluate performance and to identify any potential problems.

Table 2 shows different text classification models compared to each other. Each model is described by several features in the related work section, including the size of the dataset it was trained on, the type of models used, the architecture of the models, whether the models were pre-trained, the accuracy of the models, the disadvantages of the models, the performance of the models, and whether the models are good for text classification tasks and lastly Classification tab.

The classification tab includes Binary classification, which means models can only predict two classes, while multi-class classification models can predict more than two classes. In this chart, the performance of different text classification models is investigated. The accuracy, disadvantages, performance, and suitability for text classification tasks are also compared using the Pre-trained BERT model with different other models [4]. This comparison helps us assess the effectiveness and superiority of the proposed methodology in achieving higher accuracy and better performance in multi-class text classification tasks.

5 Discussions and Conclusions

This paper has presented a BERT-based approach exploiting cross-validation to enhance text classification, achieving state-of-the-art results on a widely used benchmark dataset. The presented approach reaches an accuracy score of 92.29% for the preprocessed cleaned dataset and 90.08% for the raw dataset (with noise). These outcomes confirm the positive impact of transfer learning in NLP tasks, emphasising the advantage of using pre-trained BERT models in real-world text multi-class classification contexts. As expected, these performances highlight the significance of clean data in maximising the model's success.

Future improvements identified include incorporating domain-specific knowledge, exploring ensemble methods, addressing imbalanced data, and enhancing model interpretability to increase robustness and achieve even better performance metrics. Concerning other future developments, given the enormous proliferation of increasingly refined NLMs and LLMs in this field of study [7], it would be interesting to extend the experiments using additional models.

Some of them [8] proved remarkably versatile, outperforming BERT with lower computational cost in several tasks [53, 13]. Additionally, new perspectives have emerged with the advent of Quantum NLP [11, 19], a sub-field of NLP employing techniques derived from quantum theory to enhance performance. Although still limited by the youth of the research field and currently available hardware, quantum-based approaches have shown tremendous potential, especially in classification tasks [40, 15, 6], so future developments will be oriented in this direction.

References

1. **Acheampong, F. A., Nunoo-Mensah, H., Chen, W. (2021).** Transformer models for text-based emotion detection: a review of bert-based approaches. Artificial Intelligence Review, pp. 1–41. DOI: 10.1007/s10462-021-09958-2.
2. **Albishre, K., Albathan, M., Li, Y. (2015).** Effective 20 newsgroups dataset cleaning. 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent

Agent Technology (WI-IAT), Vol. 3, pp. 98–101. DOI: 10.1109/WI-IAT.2015.90.

3. **Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., Goujon, A. (2021).** A comparison of pre-trained language models for multi-class text classification in the financial domain. Companion Proceedings of the Web Conference 2021, pp. 260–268. DOI: 10.1145/3442442.3451375.
4. **Asim, M. N., Khan, M. U. G., Malik, M. I., Dengel, A., Ahmed, S. (2019).** A robust hybrid approach for textual document classification. 2019 International conference on document analysis and recognition, pp. 1390–1396. DOI: 10.1109/ICDAR.2019.00224.
5. **Bonetti, F., Leonardelli, E., Trotta, D., Guarasci, R., Tonelli, S. (2022).** Work hard, play hard: Collecting acceptability annotations through a 3D game. 2022 Language Resources and Evaluation Conference, pp. 1740–1750.
6. **Buonaiuto, G., Guarasci, R., Minutolo, A., de-Pietro, G., Esposito, M. (2024).** Quantum transfer learning for acceptability judgements. Quantum Machine Intelligence, Vol. 6, No. 1, pp. 13. DOI: 10.1007/s42484-024-00141-8.
7. **Chu, Z., Ni, S., Wang, Z., Feng, X., Li, C., Hu, X., Xu, R., Yang, M., Zhang, W. (2024).** History, development, and principles of large language models-an introductory survey. AI and Ethics, pp. 1–17. DOI: 10.1007/s43681-024-00583-7.
8. **Clark, K., Luong, M. T., Le, Q. V., Manning, C. D. (2020).** Electra: Pre-training text encoders as discriminators rather than generators. Proceedings of the International Conference on Learning Representations, pp. 1–18. DOI: 10.48550/arXiv.2003.10555.
9. **Dai, J., Chen, C. (2020).** Text classification system of academic papers based on hybrid Bert-BiGRU model. 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics, Vol. 2, pp. 40–44. DOI: 10.1109/IHMSC49165.2020.10088.
10. **Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018).** BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT.
11. **Di-Sipio, R., Huang, J. H., Chen, S. Y. C., Mangini, S., Worring, M. (2022).** The dawn of quantum natural language processing. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8612–8616. DOI: 10.1109/ICASSP43922.2022.9747675.
12. **Elia, A., Pelosi, S., Maisto, A., Guarasci, R. (2015).** Towards a lexicon-grammar based framework for NLP an opinion mining application. International Conference Recent Advances in Natural Language Processing, RANLP, pp. 160–167.
13. **Gargiulo, F., Minutolo, A., Guarasci, R., Damiano, E., de-Pietro, G., Fujita, H., Esposito, M. (2022).** An electra-based model for neural coreference resolution. IEEE Access, Vol. 10, pp. 75144–75157. DOI: 10.1109/ACCESS.2022.3189956.
14. **Garrido-Merchan, E. C., Gozalo-Brizuela, R., Gonzalez-Carvajal, S. (2023).** Comparing BERT against traditional machine learning models in text classification. Journal of Computational and Cognitive Engineering, Vol. 2, No. 4, pp. 352–356. DOI: 10.47852/bonviewjcce3202838.
15. **Guarasci, R., Buonaiuto, G., de-Pietro, G., Esposito, M. (2023).** Applying variational quantum classifier on acceptability judgements: a QNLP experiment. Numerical Computations: Theory and Algorithms NUMTA 2023, pp. 116.
16. **Guarasci, R., Damiano, E., Minutolo, A., Esposito, M. (2019).** Towards a gold standard dataset for open information extraction in italian. 2019 6th International Conference on Social Networks Analysis, Management and Security, pp. 447–453. DOI: 10.1109/SNAMS.2019.8931822.
17. **Guarasci, R., Damiano, E., Minutolo, A., Esposito, M. (2019).** When lexicon-grammar

meets open information extraction: A computational experiment for italian sentences. CEUR Workshop Proceedings, Vol. 2481.

18. **Guarasci, R., Damiano, E., Minutolo, A., Esposito, M., de-Pietro, G. (2020).** Lexicon-grammar based open information extraction from natural language sentences in italian. Expert Systems with Applications, Vol. 143. DOI: 10.1016/j.eswa.2019.112954.
19. **Guarasci, R., de-Pietro, G., Esposito, M. (2022).** Quantum natural language processing: Challenges and opportunities. Applied Sciences (Switzerland), Vol. 12, No. 11. DOI: 10.3390/app12115651.
20. **Guarasci, R., Minutolo, A., Damiano, E., de-Pietro, G., Fujita, H., Esposito, M. (2021).** Electra for neural coreference resolution in italian. IEEE Access, Vol. 9, pp. 115643–115654. DOI: 10.1109/ACCESS.2021.3105278.
21. **Gupta, A., Furniturewala, S., Kumari, V., Sharma, Y. (2023).** Steno AI at SemEval-2023 task 6: Rhetorical role labelling of legal documents using transformers and graph neural networks. Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pp. 1858–1862. DOI: 10.18653/v1/2023.semeval-1.256.
22. **Howard, J., Ruder, S. (2018).** Universal language model fine-tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 328–339. DOI: 10.18653/v1/P18-1031.
23. **Iqbal, F., Hashmi, J. M., Fung, B. C., Batool, R., Khattak, A. M., Aleem, S., Hung, P. C. (2019).** A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. IEEE Access, Vol. 7, pp. 14637–14652. DOI: 10.1109/ACCESS.2019.2892852.
24. **Kowsari, K., Jafari-Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D. (2019).** Text classification algorithms: A survey. Information, Vol. 10, No. 4, pp. 150.
25. **Kumarage, T., Liu, H. (2023).** Neural authorship attribution: Stylometric analysis on large language models. International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 51–54. DOI: 10.1109/CyberC58899.2023.00019.
26. **Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C. (2016).** Neural architectures for named entity recognition. Proceedings of NAACL-HLT, pp. 260–270.
27. **Li, W., Zhu, L., Shi, Y., Guo, K., Cambria, E. (2020).** User reviews: Sentiment analysis using lexicon integrated two-channel CNN-LSTM family models. Applied Soft Computing, Vol. 94, pp. 106435. DOI: 10.1016/j.asoc.2020.106435.
28. **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019).** ROBERTA: A robustly optimized bert pretraining approach. Proceedings of the International Conference on Learning Representations, pp. 1–15. DOI: 10.48550/arXiv.1907.11692.
29. **Maisto, A., Guarasci, R. (2016).** Morpheme-based recognition and translation of medical terms. Communications in Computer and Information Science, Vol. 607, pp. 172–181. DOI: 10.1007/978-3-319-42471-2_15.
30. **Maisto, A., Pelosi, S., Guarasci, R., Vitale, P. (2018).** Text analysis on user generated content: The rap lyrics challenge. Vol. 2018-January, pp. 657–662. DOI: 10.1109/WAINA.2018.00163.
31. **Marulli, F., Pota, M., Esposito, M., Maisto, A., Guarasci, R. (2018).** Tuning syntaxnet for POS tagging italian sentences. Lecture Notes on Data Engineering and Communications Technologies, Vol. 13, pp. 314 – 324. DOI: 10.1007/978-3-319-69835-9_30.
32. **Mikolov, T., Yih, W. T., Zweig, G. (2013).** Linguistic regularities in continuous space

word representations. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 746–751.

33. **Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J. (2021).** Deep learning–based text classification: a comprehensive review. ACM computing surveys (CSUR), Vol. 54, No. 3, pp. 1–40. DOI: 10.1145/3439726.

34. **Minutolo, A., Guarasci, R., Damiano, E., de-Pietro, G., Fujita, H., Esposito, M. (2022).** A multi-level methodology for the automated translation of a coreference resolution dataset: an application to the italian language. Neural Computing and Applications, Vol. 34, No. 24, pp. 22493–22518. DOI: 10.1007/s00521-022-07641-3.

35. **Moe, Z. H., San, T., Khin, M. M., Tin, H. M. (2018).** Comparison of Naive Bayes and support vector machine classifiers on document classification. 2018 IEEE 7th global conference on consumer electronics, pp. 466–467. DOI: 10.1109/GCCE.2018.8574785.

36. **Monteleone, M., Guarasci, R., Maisto, A. (2018).** NooJ morphological grammars for stenotype writing. Communications in Computer and Information Science, Vol. 811, pp. 200–212. DOI: 10.1007/978-3-319-73420-0_17.

37. **Mukund, S., Srihari, R., Peterson, E. (2010).** An information-extraction system for urdu—a resource-poor language. ACM Transactions on Asian Language Information Processing, Vol. 9, No. 4, pp. 1–43. DOI: 10.1145/1838751.183875.

38. **Polignano, M., Basile, P., Degemmis, M., Semeraro, G., Basile, V. (2019).** ALBERTo: Italian bert language understanding model for NLP challenging tasks based on tweets. CEUR Workshop Proceedings, Vol. 2481, pp. 1–6.

39. **Qasim, R., Bangyal, W. H., Alqarni, M. A., Ali-Almazroi, A. (2022).** A fine-tuned BERT-based transfer learning approach for text classification. Journal of healthcare engineering, Vol. 2022, No. 1. DOI: 10.1155/2022/3498123.

40. **Ruskanda, F. Z., Abiwardani, M. R., Mulyawan, R., Syafalni, I., Larasati, H. T. (2023).** Quantum-enhanced support vector machine for sentiment classification. IEEE Access, Vol. 11. DOI: 10.1109/ACCESS.2023.3304990.

41. **She, X., Zhang, D. (2018).** Text classification based on hybrid CNN-LSTM hybrid model. Proceedings of the 11th International Symposium on Computational Intelligence and Design, Vol. 2, pp. 185–189. DOI: 10.1109/ISCID.2018.10144.

42. **Shreyashree, S., Sunagar, P., Rajarajeswari, S., Kanavalli, A. (2022).** BERT-based hybrid RNN model for multi-class text classification to study the effect of pre-trained word embeddings. International Journal of Advanced Computer Science and Applications, Vol. 13, No. 9.

43. **Sukthanker, R., Poria, S., Cambria, E., Thirunavukarasu, R. (2020).** Anaphora and coreference resolution: A review. Information Fusion, Vol. 59, pp. 139–162. DOI: 10.1016/j.inffus.2020.01.010.

44. **Sun, C., Qiu, X., Xu, Y., Huang, X. (2019).** How to Fine-Tune BERT for Text Classification? Springer International Publishing, pp. 194–206. DOI: 10.1007/978-3-030-32381-3_16.

45. **Taneja, K., Vashishtha, J. (2022).** Comparison of transfer learning and traditional machine learning approach for text classification. 9th International Conference on Computing for Sustainable Global Development, pp. 195–200. DOI: 10.23919/INDIACom54597.2022.9763279.

46. **Trotta, D., Guarasci, R., Leonardelli, E., Tonelli, S. (2021).** Monolingual and cross-lingual acceptability judgments with the

italian cola corpus. Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021, pp. 2929–2940.

47. **Trotta, D., Stingo, M., Guarasci, R., Elia, A., Albanese, T. (2018).** Multi-word expressions in spoken language: Polisdict. Computational Linguistics, Vol. 2253.
48. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017).** Attention is all you need. Advances in neural information processing systems, pp. 5998–6008.
49. **Wagh, V., Khandve, S., Joshi, I., Wani, A., Kale, G., Joshi, R. (2021).** Comparative study of long document classification. TENCON 2021-2021 IEEE Region 10 Conference (TENCON), pp. 732–737. DOI: 10.1109/TENCON54134.2021.9707465.
50. **Wang, C., Nulty, P., Lillis, D. (2020).** A comparative study on word embeddings in deep learning for text classification. Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval, pp. 37–46.
51. **Wang, Z., Hamza, W., Florian, R., Carbonell, J. (2021).** ABSA-KG: Enhancing aspect-based sentiment analysis with knowledge graph. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 5821–5832.
52. **Xu, Z. (2021).** RoBERTa-wwm-ext fine-tuning for chinese text classification. DOI: 10.48550/ARXIV.2103.00492.
53. **Zhang, S., Yu, H., Zhu, G. (2022).** An emotional classification method of chinese short comment text based on ELECTRA. Connection Science, Vol. 34, No. 1, pp. 254–273. DOI: 10.1080/09540091.2021.1985968.
54. **Zhang, S., Zhang, X., Chan, J. (2017).** A word-character convolutional neural network for language-agnostic twitter sentiment analysis. Proceedings of the 22nd Australasian Document Computing Symposium, pp. 1–4. DOI: 10.1145/3166072.3166082.

Article received on 13/04/2024; accepted on 27/07/2024.
**Corresponding author is Raffaele Guarasci.*