Optimizing Email Classification in the Construction Industry through a Multimodal NLP Approach

Yu Zha, (yu.zha@ruhr-uni-bochum.de)
Faculty of Computer Science, Ruhr-Universität Bochum, Germany

Sherief Ali, (sherief.ali@ruhr-uni-bochum.de)
Chair of Computing in Engineering, Ruhr-Universität Bochum, Germany

Sebastian Schumacher, (s.schumacher@jaeger-gruppe.de)
Jaeger Zentrale Dienste GmbH + Co KG, Germany

Michael Schulte , (m.schulte@jaeger-gruppe.de)
Jaeger Zentrale Dienste GmbH + Co KG, Germany

Markus König (corresponding author), (koenig@inf.bi.rub.de)
Chair of Computing in Engineering, Ruhr-Universität Bochum, Germany

Keywords: Multimodal Classification, Natural Language Processing, Hierarchical Classification, Email Classification

Abstract

The construction industry is actively pursuing the digitalization of its workflow processes, with email communication being a routinely utilized element. Construction companies receive a significant number of daily emails containing documents, photos, technical drawings, etc. Employees invest substantial time reviewing these emails, and the absence of categorization results in some emails lacking urgency receiving high attention, while others requiring immediate responses may be overlooked. Email categorization is resource-intensive, and reliance on manual processes makes the procedure ineffective and error-prone. This study proposes an approach that combines multimodal techniques to interpret the textual information of emails and process the information in attachments, thereby improving classification accuracy. The process involves using Bidirectional Encoder Representations from Transformers to generate contextual embeddings that effectively capture the semantics of words within their respective contexts. Furthermore, a Bidirectional Long Short-Term Memory layer is employed to further process these embeddings, enabling the extraction of forward and backward dependencies across the text. The classification of emails also utilizes an attention mechanism to attain varied levels of feature focus. Distinct attention layers are employed for major and minor classification tasks to capture relevant features at different hierarchical levels within the email text. The model was evaluated under both artificially balanced and naturally imbalanced class distributions across two classification tasks. It achieved F1 scores of 0.8374 for the major class and 0.859 for the minor class. This study's innovative integration of multimodal techniques and natural language processing highlights its potential for broader applications in handling complex, unstructured data.

1. Introduction

The process of digitization has led to significant changes in the way many construction companies work, particularly in enhancing project management efficiency, reducing operational costs, and improving communication flows across various departments. This transformation has been crucial in adapting to the rapid changes in market demands and environmental standards, as highlighted by (Nývlt & Kubecka, 2023) and the broader impacts discussed by (Chen, 2023). The majority of information, including documents, photos, design drawings, etc., is now transmitted digitally, i.e. via e-mail. Such a transformation has had a huge impact on the daily workflow associated with project management, architectural planning, and among others, which have all moved to digital platforms. But until now there is no way to automate manage and categorize these emails effectively, which is a resource and labor-intensive process (Jiang et al., 2022). Moreover, the dependency on manual interpretation makes the entire process labor intensive and error-prone (Acemoglu & Restrepo, 2019).

Existing Large language models (LLMs) are intelligent and capable of many basic tasks, but they are still a long way from true general-purpose AI, and they have many limitations (Bender et al., 2021). For instance, they underperform in specific domains. Besides, they are not open-sourced and need huge computing resources. Thus, local deployment is impractical it requires transferring data to a third party for processing. For highly sensitive data, companies prefer to process it locally.

Moreover, traditional email classification studies (Q. Li et al., 2020; Salloum et al., 2022) primarily rely on text analysis using natural language processing (NLP) to analyze only the textual content of the email. Email, as a hybrid data type, also contains a wealth of information in their attachments. In this paper, the proposed approach can fully utilize this part of the information to automatically classify the emails.

The proposed method leverages innovative multimodal information as input, combined with advanced NLP techniques. This enables the model to 'understand' the information in the text used in emails and process and interpret the complex, unstructured data in attachments. This is very comprehensive regarding the analysis of data and, by and large, boosts accuracy in the classification. Our approach further amplifies the efficacy in email management with the use of Bidirectional Encoder Representations (BERT) (Devlin et al., 2019). Bidirectional Long Short-Term Memory (Bi-LSTM) (Hochreiter & Schmidhuber, 1997), which generates contextual embeddings and captures both forward and backward dependencies across email text to have a better understanding of the content. This also harnesses the unique attention mechanisms found in the different nature of email data our approach uses. In this paper, these are combined to present a novel benchmark study that, for the first time, solves the problem existing in the construction industry regarding the challenges of digital workflow and also how complex real-world unstructured data can be managed in a new business context.

Section 2 presents a review of previous related work and earlier research techniques. Section 3 outlines the methods used and the architecture proposed for Multimodal and Hierarchical. Section 4 details the comparative study and discusses the experimental findings. Finally, section 5 concludes with a summary of the main outcomes of our research.

2. Related works

Due to the data imbalance and the fact that a large part of the valuable information is transmitted through attachments, but they are underutilized, email classification in real production environments presents unique challenges and opportunities. In this section, we will focus on reviewing the literature covering email and text classification methods using multimodal data and hierarchical approaches, suiting our stated problems. For instance, (Yang et al., 2016) proposed a new hierarchical attention model called Hierarchical Attention Networks (HANs) to learn the finer-grained feature interactions in text data. This model utilizes word and sentence-level attention mechanisms to process textual data in a hierarchical manner. Although Yang et al. (2016) primarily focus on document classification, the hierarchical attention mechanisms they discuss are directly applicable to email classification, where knowledge about the context and relevance of parts of the email to each other might aid in improving the accuracy of email classification. For addressing data imbalance, their hierarchical design can effectively capture the relative importance of different sections of a document, which can be analogous to different parts of an email. This method ensures that important parts of the data are represented well and contribute meaningfully to the classification decision. The technique could be considered as an answer to the problems faced due to class imbalance that exists in industry-specific datasets.

Another line of research has shown promising results by integrating BERT with BiLSTM models in text classification tasks. (Khan et al., 2021) introduced a hierarchical BERT and a BiLSTM-BERT ensemble model using this deep learning model to solve the task of efficient detection and classification of advertisement slogans. The proposed model performed better than classical models and showcased the ability to combine different deep-learning architectures for doing complex classification tasks.

(Hnini et al., 2021) proposed a deep multimodal feature-level fusion architecture for hybrid spam detection. The architecture includes the Paragraph Vector Distributed Bag of Words and the Convolutional Neural Network (CNN). This will be put into practical use for the extraction of text and image features and post-fusion purposes of the two components from emails. Such integration further enhances the email classification features with a very representative way of presenting emails that boosts the accuracy in the classification of spam. Their experiments are conducted on datasets such as Enron (Klimt & Yang, 2004), Dredze (Dredze, 2007), and TREC 2007 (Cormack & Lynam, 2007). The authors stated that their practice model was remarkably efficient, as confirmed by the five cross-validation datasets. The outcomes of the models were very high—98.91% to 99.16% in accuracy, 98.92% to 99% in F1 scores, and 99.33% to 99.83% in recall, and from 98.20% to 98.84% in precision.

In conclusion, the current literature is a rich source of insight and methodology used to optimize the email classification model in the construction industry.

3. Methodology

Emails remain the most popular forms of communication in the construction industry, and they contain many attachments with critical information in various forms. Traditional techniques for email classification are based on text classification methods and are hence insufficient to deal with the multimodal data for the attachments. This research applies the state-of-the-art methods of multimodal machine learning for information extraction to such critical data. Eventually, these pieces of information considerably increase the model's ability to classify emails with better effectiveness (Mu et al., 2020).

3.1. Problem definition

This research primarily aims to address an effective classification and analysis of multimodal email data in the construction industry. Traditional text-based processing methods in most cases leave out a lot of key information, especially that contained within non-textual attachments, such as images, videos, and scanned documents, therefore leading to inefficient use of data (Altulaihan et al., 2023).

3.2. Problem analysis

The analysis of multimodal data presents numerous challenges, particularly in effectively combining different types of modalities, such as text, images, and videos while managing the disparities in the quantity and quality of information across these sources. This complexity is further compounded by the hierarchical nature of document types, necessitating a sophisticated multimodal approach to model building that leverages both textual and visual data for accurate classification. Unlike traditional rule-based methods, which rely on predefined heuristics and often struggle with the variability inherent in multimodal data, this approach dynamically integrates diverse data sources to adapt to the content's complexity. Moreover, while LLMs excel in processing and understanding text, they can fall short when visual or other non-textual data is critical to the task.

3.3. Data preparation

To apply our approach, we relied on a dataset provided by Jaeger Gruppe, which was collected through web-scraping in January 2024 for three subsidiaries of the company. The dataset reflects a diversified scope of email data, and it is a typical communication method for a construction company. The Dataset includes over 40,000 records from the different classes, each of which represents a different kind of communication or document type commonly used within the subsidiaries.

The document types range from regular correspondence ('Schriftverkehr') and contractual agreements ('Vertrag') to very specific documents like photos ('Bilder') and legal consultations ('Rechtsanwalt/Gutachter'). The category of incoming correspondence, 'Schriftverkehr AG Eingang,' predominates, highlighting the company's extensive external communications. **Figure** 1 demonstrates a graphical representation of classes distribution in the dataset.

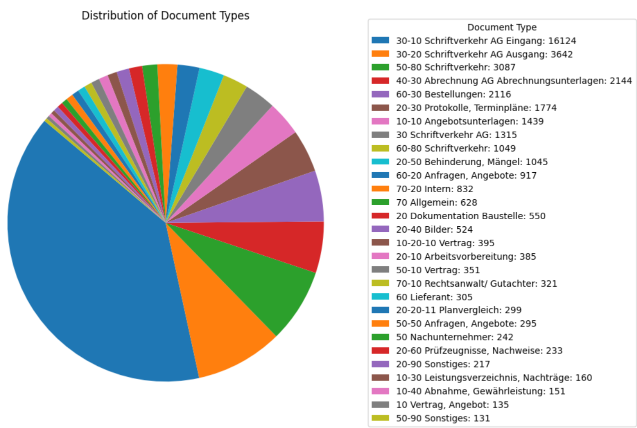

Figure 1: Dataset distribution

3.4. Data preprocessing

The dataset requires cleansing preparation for the next steps, thus we initiated preprocessing to convert the email content into plain text. This process also involves extracting and cleaning the email text data to ready it for further analysis and modeling within this study. The text of an email often comprises three: the subject line, cc (carbon copy) fields, and the body of the content. It is first cleaned to remove special characters among other noise elements, including HTML tags and even email signatures that may interfere during analysis.

Each document type undergoes a different preprocessing technique, thus images and videos are processed using multimodal models like BLIP (J. Li et al., 2022) for image captioning, which helps derive textual descriptions of the visual content. This is crucial for integrating visual data into the predominantly textual analysis framework of this study.

PDF attachments have diverse content types, each requiring a specific preprocessing approach. Whether it is a scanned document, design drawing, or machine-readable text. Design drawings are content described through image captioning techniques used to convert their visual data into descriptive text, similar to image and video processing. For machine-readable documents, the information will be extracted directly. In the case of scanned documents, a multimodal non-structured documents understanding model LayoutLM (Huang et al., 2022) will be used to extract key information from those documents.

Then the text extracted from the body of the email or from attachments is merged to form comprehensive inputs for further modelling. The whole process is represented in **Figure** 2. It ensures the efficient use of text and non-text data, and therefore, the outcomes drawn from the email dataset will be richer and more accurate.

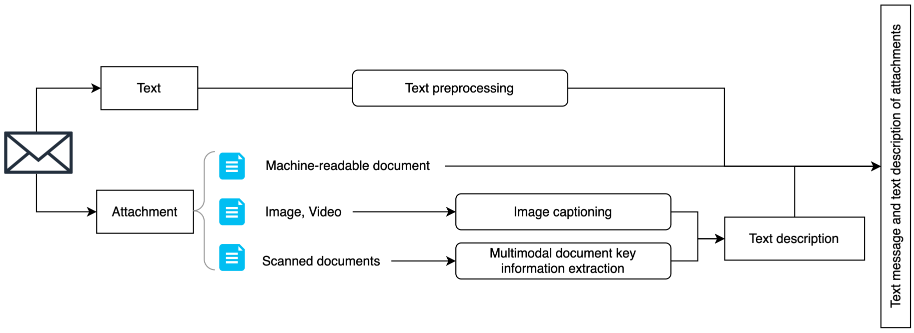

Figure 2: Data preprocessing

3.4.1. Long text segmentation using sliding windows

To cope with the constraint of the maximum sequence length that BERT has, this research proposes using an algorithm to handle longer text beyond the input limits of BERT effectively. Like sliding windows in image processing. This algorithm divides longer text into shorter and more manageable segments, while at the same time ensuring contextual continuity through an overlap of specified regions in the adjoining segment. Then, each segment is processed one after the other, and the resultant of each single segment will be summarized into one and passed to the next stage of the process as input.

3.4.2. Feature extraction and Sequential processing

Feature extraction is the next step in the pipeline, as shown in **Figure** 3. This process proceeds by segmenting smaller pieces of the large input text into manageable chunks, largely following the sliding window approach described above. After successfully segmenting the text, the next step involves processing every text segment through BERT, the leader in a line of deep contextual transformer models Each of these segments is processed by BERT to produce the feature representations. These representations of every such segment include the contextual features of every such segment, information content related to it, and information from the areas of overlap. This is a crucial step as it maintains the semantics contained within the information over many segments. The outputs from BERT are then pooled (pool-out) and the sequence representations for

each segment are extracted. These sequence features encapsulate crucial attributes of each segment and are subsequently used as inputs to a BiLSTM network.
The BiLSTM layer (Hochreiter & Schmidhuber, 1997) of the developed architecture is powerful in processing the sequence data and hence increases the performance of the model to understand and integrate the temporal dynamics, and dependencies concerning different texts.
Thus, feeding into the BiLSTM the BERT-extracted features of every segment would make these features get synthesized for the model to build a complete representation at the document level effectively (doc_repr). This representation captures the details within individual segments and amalgamates the contextual nuances across the whole length of the input text. This innovative use of BERT for initial feature extraction followed by sequential processing through BiLSTM can allow for conducting a robust analysis of extended texts. Because it is typical for email communications where critical information might be placed deep inside attachments and lengthy correspondences.
This methodological approach ensures the model can leverage both the local and global contextual features of the emails, thus providing for more effective and accurate classification and analysis within the construction industry email dataset.

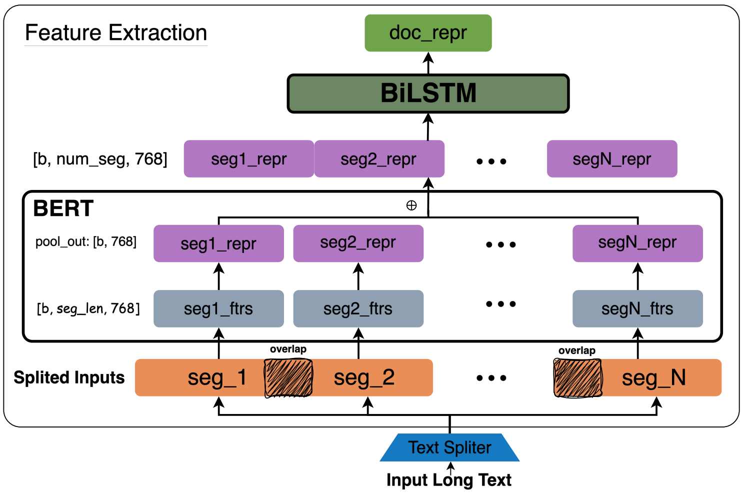

Figure 3: Feature extraction

3.5. Hierarchical classifier

Before diving into the detailed architecture of the hierarchical structure of the classifiers. it is imperative to understand the definition of the hierarchical structure. In this paper, the major classes represent general categories that cover broad themes or types of email communications in the construction industry. These major classes are further subdivided into minor classes to further differentiate among the major categories in great detail. This hierarchy of classifiers significantly extends the capability of the model in dealing properly with complicated email data, achieving refined analysis, and improved classification accuracy.
After the acquisition of the comprehensive document representation (doc_repr), we designed a hierarchical classifier, as shown in **Figure** 4 due to the hierarchy of the labels. Moreover, attention mechanisms, feature fusion, and cross-layer feature transmission were used in the classifier to meet the complexities of the dataset. These techniques are important components for the classifier to improve accuracy and efficiency.
The attention mechanisms decide what amount of weight to give the different parts of

the text (Brauwers & Frasincar, 2023). The model will focus on the most relevant parts of the text through different weights.
Fusing the selected features from different layers of the model enables the classifier to use a rich set of features so that the model can learn information properly. Putting all these diverse features together in a single classifier can capture a more holistic view of the text, which is a critical aspect of clearly distinguishing the often subtle differences among the classes. Cross-layer feature propagation is when the feature of interest passes through different layers of the classifier, adding to the feature set used at every layer of Decision-making (Brauwers & Frasincar, 2023). This approach will help preserve valuable information across the depths of classifier architecture, to lead toward more informed and refined classification decisions. It guarantees that subtle cues that are captured through the earlier layers are not lost but put to use in deeper layers where final decisions take place.
The design of classifiers follows the divide-and-conquer policy, particularly in handling the highly imbalanced data distribution across classes. It doesn't classify all sub-classes directly at a time, which would be more problematic in the case of data imbalance. First, major class classifiers will identify and help narrow down the variance in the data distribution for every major category in an email. Then, the minor classifier could handle finer classification within every major category. This structural approach eases the model learning task, making it possible for the model to attain better classification accuracy through a separate focus on specific label hierarchies. This method is effective in dealing with large and complex datasets, like those email communications of the construction industry, since precision in classification has to be maintained.

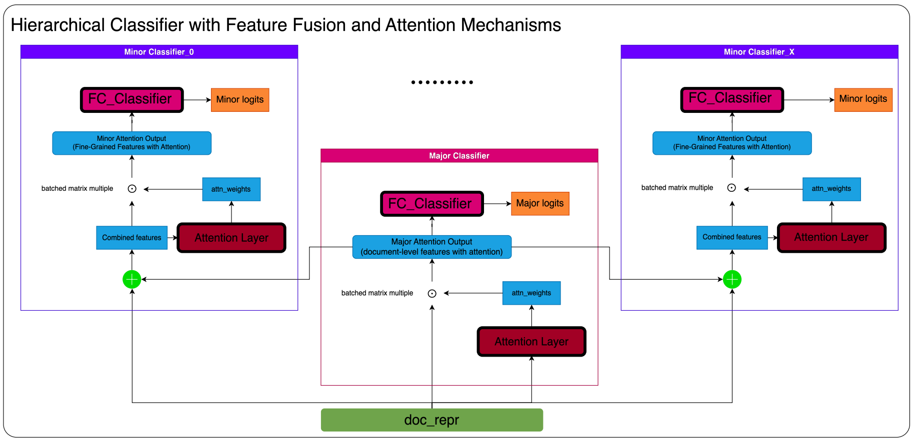

Figure 4: Hierarchical classifier with feature fusion and attention mechanisms

4. Experiments

4.1. Evaluation metrics

We evaluated our hierarchical multimodal email classification model with performance evaluation based on F1 Score, Accuracy, and Top-3 Accuracy (Margherita Grandini, 2020). Accuracy measures the proportion of correct predictions, providing a quick overall effectiveness snapshot but can mislead in imbalanced datasets. F1 Score, the harmonic mean of Precision and Recall, is crucial for imbalanced classes, ensuring a balance between precision and recall, and enhancing model assessment. Top-3 Accuracy evaluates if the true label is among the three most probable predicted labels, useful for multi-class problems where approximate correctness is acceptable. These metrics help us understand both major and minor class predictions in our imbalanced dataset. Given the dataset's significant imbalance, we selected the F1 Score as the more appropriate metric, balancing precision and recall for a more accurate measure of model performance under uneven class distribution.

This study has also benchmarked the solution with the available commercial solution using the same dataset for comparison purposes. The market solution only disclosed their weighted accuracies as the result by using the same dataset. Since the weighted accuracies are only included in the market solutions. Furthermore, we listed the weighted accuracies so that direct comparisons are conducted under the same conditions.

In addition, concerning future practical implementations for construction companies, the top three most likely predictions are also recommended for them by our model. Top-3 Accuracy here is a metric describing how well the model performs, suggesting potential classifications with practical utility— extra information for end-users to classify email efficiently.

4.2. Dataset

There have been challenges with the Jaeger Cloud dataset due to the uneven distribution of different email types. To address this, two sampling strategies are implemented to create two datasets for our analysis.

Since the division of the e-mails was not even in all classes, it was considered to use stratified sampling to ensure that every stratum was proportionally represented. In cases where counting samples in a class exceeded the average across all classes, they were all limited to this average count. By doing so, data become more evenly distributed and more representative of each category. This avoids the model learning only from the classes with the most samples.

The second set was prepared using random sampling with the original dataset. According to (Cochran, 1977), this is a way to select entries at random to ensure that each item in the dataset has an equal chance of being selected, thereby ensuring the structure and statistical integrity of the data. This method will keep all the original distributional characteristics, particularly the proportion of different types of emails, to ensure the real-world heterogeneity of the dataset. It will retain the representativeness of the dataset. However, it also poses training challenges, mainly because the distribution of data types is skewed. (Thompson, 2012) also discussed random sampling in statistical analysis as a double-edged sword, particularly when data distributions are very much skewed.

Such sample methods have allowed us to fit our models and test them under various conditions, thus giving a glimpse of how well our approach really could cope with complexity coming from an imbalanced dataset in a practice-specific industrial context.

4.3. Experimental results

The results of our experiments, as shown in **Table** 1, clearly demonstrate the effectiveness of our hierarchical and multimodal email classification employed under different sampling strategies and token lengths.

Initially, our model was evaluated on the artificially balanced class distribution stratified sample dataset achieving an F1 score of 0.8099 for the major classes and 0.8529 for the minor classes with a 600 token configuration. These results indicate the model's potential efficiency under controlled and balanced conditions, giving us a perspective on what further improvements are needed.

The model demonstrated superior performance under the naturally imbalanced conditions of the original dataset, particularly through random sampling. Using a 600-token configuration of random sampling, the F1-score increased to 0.8374 for major classes and 0.859 for minor classes. This is evidence not only of the robustness of our model but also of its ability to well serve data with real-world imbalanced distributions, which is an indispensable feature of any practical application. Compared to existing market solutions, our model significantly outperforms the benchmarks. The market solution showed considerably lower performance metrics, with an accuracy of 0.798 for major classes and 0.752 for minor classes. Our model demonstrates a superior capability to meet the nuanced requirements of construction industry email communications. Finally, these results validate the effectiveness of our hierarchical and multimodal approach, setting a new standard for email classification models within the construction industry and promising significant improvements in data management and operational efficiency.

Table 1: Evaluation metrics of the model

	F1 Major	Acc Major	Top 3 Major	F1 Minor	Acc Minor	Top 3 Minor	Both_Correct
Strat. samp., 600 tok.	0.8099	0.8095	0.9634	0.8529	0.8523	0.9407	0.6408
Rand. samp., 600 tok	0.8374	0.8349	0.9772	0.859	0.856	0.9737	0.6977
Rand. samp., 1200 tok	0.8569	0.8578	0.9793	0.9121	0.9134	0.9981	0.7456
Rand. samp., 1600 tok	0.923	0.9231	0.9661	0.9661	0.9134	0.9899	0.7992
Market solution	N/A	0.798	N/A	N/A	0.752	N/A	0.6451

5. Conclusion and Future work

This study proposed an email classification model for the construction industry, making email management in the construction industry much more efficient. The most advanced multimodal models, such as BLIP, LayoutLM, and NLP technologies, including BERT with BiLSTM, have been used to treat massive amounts of information from data. Our results demonstrate a significant improvement over traditional methods, due to consideration given to handling imbalances in the data and extracting useful insights from multifarious inputs to facilitate better data management and operational efficiency.

While the model is effective, particularly with the pre-trained language model, its major setback lies in the scalability and computational demands posed by big data. In today's digital age, enormous volumes of data require real-time processing for practical applications. This approach requires high-quality labeled data for the model, which limits its adaptability in dynamic environments. The constant evolution of email formats may further impact performance, necessitating frequent updates to maintain accuracy.

Future research aims to develop and enhance computational efficiency and adaptability, deriving more effective algorithms using semi-supervised learning techniques. This approach will reduce the need for manual interpretation.

References

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, *33*(2), 3–30.

Altulaihan, E., Alismail, A., Rahman, M. M. H., & Ibrahim, A. A. (2023). Email security issues, tools, and techniques used in investigation. *Sustainability*, *15*(13), 10612. https://doi.org/https://doi.org/10.3390/su151310612

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT*.

Brauwers, G., & Frasincar, F. (2023). A general survey on attention mechanisms in deep learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3279–3298. https://doi.org/10.1109/tkde.2021.3126456

Chen, S. (2023). Digital transformation in the construction industry: A systematic literature review. *Proceedings of the 2023 9th International Conference on Industrial and Business Engineering*, 235–244. https://doi.org/10.1145/3629378.3629431

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.

Cormack, G. V., & Lynam, T. R. (2007). Overview of the trec 2007 spam track. *Sixteenth Text REtrieval Conference*. https://trec.nist.gov/pubs/trec16/papers/SPAM.OVERVIEW16.pdf.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/n19-1423

Dredze, M. (2007). Spam email dataset for research.

Hnini, G., Riffi, J., Mahraz, M. A., Yahyaouy, A., & Tairi, H. (2021). Mmpc-rf: A deep multimodal feature-level fusion architecture for hybrid spam e-mail detection. *Applied Sciences*, *11*(24). https://doi.org/10.3390/app112411968

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). Layoutlmv3: Pre-training for document ai with unified text and image masking.

Jiang, S., Hu, J., Magee, C. L., & Luo, J. (2022). Deep learning for technical document classification. *IEEE Transactions on Engineering Management*.

Khan, M. A., Parvez, S., Rahman, M., & Anwar, M. (2021, February). Efficient advertisement slogan detection and classification using a hierarchical bert and bilstm-bert ensemble model. https://doi.org/10.1007/978-3-030-68154-8_81

Klimt, B., & Yang, Y. (2004). The enron corpus.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning*. https://doi.org/null

Li, Q., Cheng, M., Cheng, M., Wang, J., Wang, J., Wang, J., Sun, B., & Sun, B. (2020). Lstm based phishing detection for big email data. *IEEE Transactions on Big Data*. https://doi.org/10.1109/tbdata.2020.2978915

Margherita Grandini, G. V., Enrico Bagli. (2020). Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756*. https://arxiv.org/abs/2008.05756

Mu, S., Cui, M., & Huang, X. (2020). Multimodal data fusion in learning analytics: A systematic review. *Sensors*, *20*(23). https://doi.org/10.3390/s20236856

Nývlt, V., & Kubecka, K. (2023). Challenges and opportunities of digitization in the construction industry. *10th SWS International Scientific Conferences on SOCIAL SCIENCES - ISCSS Proceedings 2023*.

Salloum, S. A., Gaber, T., Vadera, S., & Shaalan, K. (2022). A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access*. https://doi.org/10.1109/access.2022.3183083

Thompson, S. K. (2012). *Sampling* (3rd ed.). Wiley.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 1480–1489). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1174