Multimodal Transformer for Risk Classification: Analyzing the Impact of Different Data Modalities

Niklas Holtz[1] and Jorge Marx Gómez[2]

[1]Future Research, Volkswagen, Wolfsburg, Germany
niklas.holtz@volkswagen.de
[2]Very Large Business Applications, Carl von Ossietzky Universität, Oldenburg, Germany
jorge.marx.gomez@uol.de

Abstract

Risk classification plays a critical role in domains such as finance, insurance, and healthcare. However, identifying risks can be a challenging task when dealing with different types of data. In this paper, we present a novel approach using the Multimodal Transformer for risk classification, and we investigate the use of data augmentation for risk data through automated retrieval of news articles. We achieved this through keyword extraction based on the title and descriptions of risks and using various selection metrics. We evaluate our approach using a real-world dataset containing numerical, categorical, and textual data. Our results demonstrate that the use of the Multimodal Transformer for risk classification outperforms other models that only utilize textual data. We show that the inclusion of numerical and categorical data improves the performance of the model, particularly for risks that are difficult to classify based on textual data alone. Additionally, our research indicates that the utilization of data augmentation techniques yields enhanced performance outcomes in models. This methodology presents a promising avenue for enterprises to effectively mitigate risks and make well-informed decisions.

Keywords

Risk classification, Multimodal Transformer, Data augmentation.

1. Introduction

Risk classification is a crucial task in many domains, such as finance, insurance, and healthcare. It involves identifying potential risks and estimating the likelihood of their occurrence, which is essential for businesses and organizations to make informed decisions. Traditionally, risk classification has been performed by analyzing numerical and categorical data. However, the increasing availability of textual data, such as news articles and social media posts, has opened up new opportunities to improve risk classification performance.

In recent years, machine learning techniques have shown promising results in risk classification tasks. One popular approach is to use deep learning models such as the Transformer architecture [1], which has achieved state-of-the-art performance in natural language processing tasks. However, the Transformer model has mainly been applied to textual data, and its performance when used in conjunction with numerical and categorical risk data remains largely unexplored.

In this paper, we propose a novel approach that uses the Multimodal Transformer [2] for risk classification. The Multimodal Transformer is an extension of the Transformer architecture

David C. Wyld et al. (Eds): NLPML, AIAP, SIGL, CRIS, SEC, COSIT, DMA -2023
pp. 29-39, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130803

that can handle multiple modalities of data, including numerical, categorical, and textual data. We investigate the impact of using different data modalities on risk classification performance and explore the use of data augmentation through automated retrieval of news articles. To perform data augmentation, we use keyword extraction techniques based on the title and descriptions of risks to retrieve relevant news articles. We employ various selection metrics to ensure the quality of the retrieved articles and integrate them into the training data to improve the model's performance.

In summary, our proposed approach provides an opportunity for businesses and organizations to better manage risks and make informed decisions. The remainder of this paper is organized as follows: Section 2 provides a brief overview of related work in risk classification and an introduction to multimodal transformer architectures. Subsequently, Section 3 presents the real world dataset as well as training and comparing different multimodal transformers. In Section 4, we present our approach to risk data augmentation, which automatically searches for news data that is evaluated and added to the training dataset as a textual feature. This chapter also trains other multimodal transformers and compares their performance with those from Section 3 to identify the added value of data augmentation. Finally, Section 5 discusses the results and presents the limitations. Section 6 summarises the results and suggests future work in this field.

2. Literature Review

The use of artificial intelligence has already proven useful in various areas of risk research. Especially in the financial or healthcare sector, the calculation of risk scores is of interest [3]. This quantitative measure is applicable in various contexts and can be used, for instance, in medicine to assess the risk of diseases and complications. The larger the score, the more likely the outcome or threat will end up happening. Numerous studies have investigated the use of machine learning algorithms to predict such risk scores or similar metrics. Recent advances in natural language processing (NLP) have also led to the development of language models such as BERT [4] and GPT [5], which have shown significant improvements in text classification tasks. A number of studies have already investigated the use of NLP in a wide range of risk fields. As an example, NLP has been used to create a rating based on electronic mental health records to assess patients' risk of harm to self and others [6].

One of the significant challenges, however, is the processing of different types of data, which can vary greatly depending on the risk area. Beyond risk scores and other metrics, categorical information and especially textual data can contain essential information. Traditionally, this diversity of data is used by human experts to make decisions and classify risk using their domain knowledge. However, several studies already demonstrate that modern machine learning algorithms can be used to automate this classification process.

2.1. Related work on risk classification

Several studies have been conducted on the use of machine learning algorithms for risk classification. For instance, a study by Zhou et al. [7] investigated operational risk classification in the financial industry, focusing mainly on text classification. By implementing a semi-supervised text classification framework evaluated with real-world data, baseline methods for operational risk classification were outperformed. A particular focus was on the handling of only partially labelled data.

In another study, different methods were used to extract and classify sentences from securities reports that describe the risks taken by companies [8]. The aim of this approach is to allow early risk management by identifying potential risks in such securities reports. A total of 2494 risk sentences were used to train different classification models, which either resemble an expert system or use machine or deep learning techniques. The most successful model turned out to be a BERT model. For future work, it was suggested to look at different document types and also to consider new dimensions (such as time).

Other medical studies have already addressed the issue of taking different data modalities into account in risk classification. Especially in the healthcare sector, the variety of data is comparatively high with regard to numerous different sensor technologies used in patient examinations. For instance, electronic medical records, radiological images and genetic repositories have been combined to train machine learning models that detect risks for cardiovascular diseases [9]. However, it is precisely the abundance of data that makes the practical use of such models in a clinic environment difficult. The authors point to the increasing complexity of the models and the resulting slow processing time of input data.

Overall, these studies demonstrate the importance of utilizing different types of data modalities and machine learning techniques for accurate risk classification. However, there is still a need for further research to improve the performance of risk classification models, particularly for complex and heterogeneous data.

2.2. Multimodal Transformers

Multimodal Transformers have become an increasingly popular approach for processing and modeling data that comes from different modalities. They have shown great success in various fields such as natural language processing, computer vision, and multimodal fusion. In this literature review, we discuss the recent advancements in Multimodal Transformers and their applications.

The Transformer architecture was first introduced by Vaswani et al. [1] for natural language processing tasks. The Transformer model consists of an encoder and a decoder that utilize self-attention mechanisms to capture dependencies between the input tokens. The model's ability to capture long-term dependencies and handle variable-length sequences has made it the state-of-the-art in various NLP tasks.

BERT [4] is a multilingual pre-trained Transformer-based language model algorithm developed by Google. The model uses a bidirectional architecture to learn contextual word representations in a large corpus of unannotated text. Unlike previous models that were trained only in a left-to-right or right-to-left fashion, BERT is trained using a masked language modeling objective that enables it to better understand the relationships between different words in a sentence. As a result, BERT has been shown to achieve state-of-the-art performance on a wide range of natural language processing tasks, including sentence classification, question answering, and named entity recognition. Its success has led to the development of a range of BERT-based models for various NLP applications, including the Multimodal-Transformers that can combine BERT with other modalities such as images and videos to enable more complex and versatile natural language understanding.

The challenges and opportunities of using a multimodal transformer in the literature have already been summarised [10]. Above all, the choice of a suitable architecture is essential for the successful implementation of a model with regard to the characteristics of different data modalities. The advantages of transformer models for multimodal learning are, among others, their flexibility resulting from the encoding of implicit knowledge. Numerous transformer models have been presented in the literature that can process other modalities besides textual data and are mostly based on BERT. Thus, MMBT [11] is a model for which only a fine-tuning for new data modalities is made, in order to ultimately enable the processing of image and text data. VLBERT [12] , on the other hand, is able to process image data as an additional input token, so the entire model must be pre-trained on a suitable set of data. Moreover, another publication presents the HG-BERT model [13] and deals with the direct optimisation of BERT in order to use the model in multimodal sentiment analysis. For this purpose, hierarchical multi-head self-attention mechanism and gate channels were used, among other things, to extract features more selectively and to realise noise filtering. Overall, the use of multimodal transformers is an open problem and therefore requires further investigation.

With the Multimodal Toolkit [14], an open-source Python package was made available with which text and tabular data (categorical or numerical) can be processed. It also allows the download of various pre-translated models. The core of the toolkit is a combining module, which receives as input textual features of the preprocessed transformer as well as the preprocessed numerical and categorical features. In the paper, different methods are investigated on how to best combine these input variables to output a combined multimodal representation to the final Fully Connected Layer, which ultimately performs the classification. The choice of the best combination method in the experimental evaluation depended mainly on the number of different features and varied. Due to the high accessibility and flexibility of the toolkit, different models are trained on its basis in the course of this paper.

In conclusion, Multimodal Transformers have shown great potential for modeling data from different modalities and achieving state-of-the-art performance on various tasks. Their ability to handle multiple modalities and perform fusion in a principled manner makes them a promising approach for risk classification in various domains.

3. Multimodal Risk Transformer

This section describes the application of the Multimodal Transformer to classify risks using different measures expressed in different modalities. Using the Multimodal Toolkit [14] and Python, three different models were trained on a real world dataset from the corporate environment. Each of the models was trained on a subset of different modalities from the original dataset to explore how these compositions affect the final models. In Section 3.1, an insight into the real dataset is given for a better understanding. Section 3.2 presents the accuracies achieved and the training behaviour of the different models.

3.1. Training data

The training data is based on actual corporate risks that occurred in the period from 2016 to 2022 and were evaluated through various filter measures. On the one hand, these are internal corporate risks that, for instance, address specific processes and are thus often very individual to the company. On the other hand, they can also be external risks that result, perhaps, from changes in the law or similar. The collected risks originate from reports that may have been sent in from any area of the company, so that a very high variance within the data is noticeable with regard to their domain and the level of detail. In total, the data set contains 5305 risks. There are a total of 6 different categories for the classification of risks, such as "financial risks" or "research and development risks". Textual data is available for all risks in the form of a title and a description. Risk descriptions can be either brief or multi-line reports, whereby they are always generally comprehensible and, for example, technical details have been generalised as far as possible. In addition to textual data, the data set also contains two further categorical values for the quarter in which the risk was reported and for the probability of occurrence, which describes the chance of a risk

Table 1: Comparison of the F1 scores with different data modalities. T is the number of text features. C is the number of categorical features, and N is the number of numerical features.

Model	**Epochs**	**T**	**C**	**N**	**F1-score**
TextOnly	10	2	0	0	0.7815
TitleOnly	10	1	2	6	0.7645
FullyModal	10	2	2	6	0.7947

happening in rough categories. Finally, there are 6 further numerical scores for each risk, which describe, among others, the reputational damage or the financial extent of damage to the company. In summary, there are two text features, 3 categorical features and 6 numerical features for each risk.

3.2. Method and Results

Three different models were trained for a total of 10 epochs, each of which had to process different data modalities. In order to best combine the different features within the model, a gating mechanism [15] was used, which has already proven successful in investigating different combination methods for a similar dataset [14]. Table 1 shows an overview of the three trained models, the number of features used and the F1 score achieved. The *TextOnly* model was trained exclusively on the basis of text features and thus only uses risk titles and descriptions. The second model *TitleOnly*, on the other hand, uses all categorical and numerical features, but limits the text features to the risk titles in order to examine the effects of the risk descriptions more closely. The *FullyModal* model uses all available features.

All models achieve a high F1 score, which shows that the existing real risk data set is appropriate for training multimodal transformers. The *FullyModal* model achieves the highest F1 score. Considering the *TitleOnly* model, it is notable that only the absence of the risk description reduces the F1-score by about 0.03. The absence of numerical and categorical features in the *TextOnly* model, on the other hand, leads to a reduction of about 0.013 compared to the *FullyModal* model.

4. Risk Data Augmentation

This chapter describes the augmentation of the existing training data using different methods with the purpose of examining how the larger amount of data affects the Multimodal Risk Transformer. In Section 4.1, we first present our approach to data augmentation based on the existing risk training data. Next, Section 4.2 presents different metrics used for data augmentation and demonstrates how they produce different datasets. Finally, an experimental evaluation of the approach is conducted in Section 4.3, comparing models trained on the basis of the augmented data with the models presented in Section 3.2.

4.1. Our Approach

Our approach is based on the extraction of keywords from the title and descriptions of the risk dataset already described in Section 3.1. After extraction, several metrics are applied to obtain a set of strings containing single or combinations of the selected keywords. Using these strings, news articles can then be retrieved using an API that match the risk data and ideally contain further information that facilitates the subsequent classification of the risks.

Finally, the data is aggregated with the existing reports and brought into the Multimodal Transformer model. An overview of the entire processing pipeline can be seen in Figure 1. In the following, the most important points of this pipeline are explained in more detail.

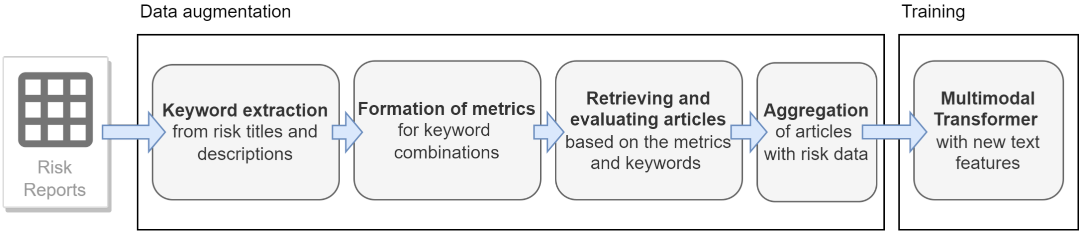
Data augmentation
Training
Risk Reports
Keyword extraction from risk titles and descriptions
Formation of metrics for keyword combinations
Retrieving and evaluating articles based on the metrics and keywords
Aggregation of articles with risk data
Multimodal Transformer with new text features

Figure 1: Overview of the data augmentation process.

4.1.1. Keyword extraction.

KeyBERT [16] is used to create minimal keywords for the titles and descriptions of the risks. This technique, first introduced by Grootendorst, leverages BERT embeddings to extract keywords or keyphrases that are as similar as possible to the input text. The number of keywords extracted in the training data set varies between 0 and 5, depending on the length of the title and descriptions, whereby keyphrases can be included as well. In addition, this technique also outputs accuracies for each keyword, which are taken into account in the further course of the pipeline.

4.1.2. Application of metrics for combining keywords.

Keywords are combined and selected for later retrieval of news articles in order to formulate search queries that are as precise as possible. For this purpose, various metrics are used that select and, if necessary, concatenate the numerous keywords in different ways based on the associated output accuracies. Section 4.2 presents the metrics in more detail. The results of this step are one string per metric containing either single or multiple combined keywords.

4.1.3. Retrieving and evaluating news articles.

The selected combinations of keywords and individual keywords are used to retrieve news articles through several search queries with the help of an API. The search queries contain the keywords as search terms on the one hand and a time period on the other, which is always limited to the retrospective quarter of the underlying risk. We used the GDELT DOC 2.0 API [17] for this purpose, through which a wide variety of news articles from 2014 can be retrieved. The articles are retrieved in German, as the risk data is also in German. After retrieving the news articles, a number of considerations, which are described in more detail in the following section, are used to evaluate which of the metrics could produce the most relevant articles. Finally, the most appropriate 5 articles of the best performing metric are selected and aggregated with the remaining risk data to finally introduce them into the Multimodal Risk Transformer model.

4.2. Metrics

Metrics in our approach have the task of evaluating the previously extracted keywords from risk titles and descriptions and selecting them according to certain specifications in order to retrieve news data in the further course. Each metric ultimately selects title keywords, with the description keywords being used to explore the content of the articles. For this purpose, 7 metrics were constructed, an overview of which can be seen in Table 2. The metric *Most Accuracy* selects the keyword that has achieved the highest accuracy in the extraction. This can be either a single word or a keyphrase consisting of several connected words. *Best Single Word* selects in a similar way, but only keywords that contain only a single word are considered here. Similarly, *Best Two Words* selects those keywords that consist of exactly two words. The two *Best Separated Two Words* metrics first select the keyword that has achieved the highest accuracy with exactly two words, but then the two words are separated into a string. This results in two metrics, each addressing one of the two words. Finally, the metrics *Best Combined Single Words* and *Best OR'd Single Words* first select the two keywords with the highest accuracy, both of which consist of only one single word. Then, the former metric concatenates both keywords into a string so that only data that exactly matches this string is considered in subsequent data retrieval. In contrast, the second metric combines the two keywords in the form of an OR search, so that at least one of the two keywords must be contained in the subsequent news articles, but not necessarily both.

Figure 2 shows a graph presenting how many articles were found for each risk by the different metrics. It is noticeable that the *Best Combined Single Words* metric in particular could only find a few articles with a total of 4,618 hits, whereas the *Best Single Word* metric could find the most articles with a total of 330,655 hits. Nonetheless, it is imperative to acknowledge that the quantitative analysis of hits per metric provides little insight into the semantic quality of the extracted title keywords. This is particularly relevant for metrics that primarily generate generic single-word strings.

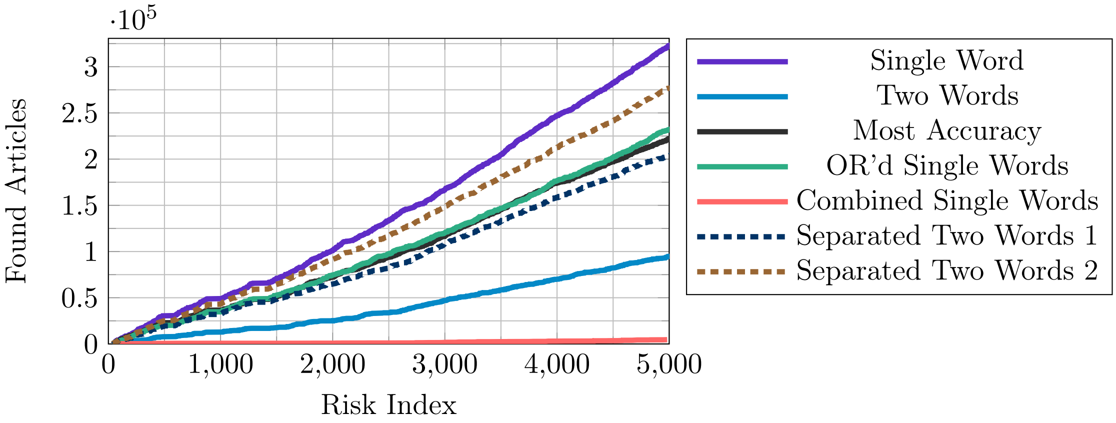

Figure 2: Course of articles found for each risk by using the different metrics.

For a better assessment of the quality of the metrics and thus also of the hits found, an evaluation method was programmed to provide information about the quality of the hits with the help of simple equations. For each risk, all metrics are evaluated iteratively, whereby the best-performing metric is ultimately selected for the further processing in the pipeline. As a measure for the quality of a title keyword k, the quotient q_k is considered, for which the following equation applies

$$q_k = \frac{p(k) * 100}{c_k} \tag{1}$$

Table 2: Comparison of the different metrics with the name of the metric, the articles found, the articles that matched extracted risk keywords, the average quotient as well as the percentage distribution. The latter indicates how often the metric was selected for data retrieval.

Metric	Articles	Matched Articles	Quotient	Chosen
Most Accuracy	225,638	1,901	8.71	53.9%
Best Single Words	330,655	2,520	5.29	12.8%
Best Two Words	95,936	749	7.34	4.7%
Best Separated Two Words 1	208,614	937	3.08	11.3%
Best Separated Two Words 2	280,621	1,293	1.79	7.6%
Best Combined Single Words	4,618	19	20.67	2.8%
Best OR'd Single Words	236,493	1,629	2.53	6.9%

The accuracy of the keyword $p(k)$ is put in relation to the total number of hits c_k when trying to retrieve data. In practice, it has been shown that a high number of hits often results from generic keywords. By dividing with the number of hits c_k, the quotient for such generic keywords should be reduced.

As a second evaluation measure, it is also analysed how many keywords of the risk descriptions are contained in the articles found. With the number of articles m_k found for which both the keywords of the titles and the keywords of the descriptions apply, the quotient is adjusted as follows:

$$q_k = \frac{q_k}{1 + \frac{m_k}{2}} \tag{2}$$

Table 2 shows the results of the evaluation for the different metrics. It is notable that *Most Accuracy* was ultimately used most frequently with 53.9%, although it neither has the highest quotient nor was able to produce the most articles. *Best Combined Single Words*, on the other hand, was selected the least, found the fewest articles, but has the highest quotient. This illustrates that this metric was only able to retrieve articles comparatively rarely, but the quality of these articles is high. It can be assumed that whenever data was retrieved using this metric, it was ultimately selected for further processing in the pipeline. *Best Single Words* was the most retrieved article, but the quotient is comparatively low, suggesting that this metric generated mostly generic keywords.

4.3. Experimental Evaluation

This section describes the experimental evaluation of the newly generated data. As in Section 3, three different models were trained using different data modalities, whereby for the risks there is now an additional textual feature with the top 5 news articles, which is processed as a single string. Table 3 shows the results of this evaluation, whereby the F1 scores achieved by the models trained on the basis of the original data set and their differences are also listed for comparison.

The augmented data produced an increase in all models, with *TitleOnly* showing the most significant improvement. As the only one of the three models that does not have access to the risk descriptions, the textual features in the form of news articles appear to have a similar influence on the performance of the model. The *TextOnly* model was improved the least by the new data set. The model with the highest F1 score *FullyModal* was slightly

Table 3: Comparison of the three models described in Section 3 with the original and augmented training data. *F1 (new)* describes the F1-score that was obtained with the augmented training data.

Model	F1	F1 (new)	Difference	Improvement
TextOnly	0.7815	0.7831	+ 0.0016	0.2%
TitleOnly	0.7645	0.7735	+ 0.009	1.17%
FullyModal	0.7947	0.7983	+ 0.0036	0.45%

improved, but the new dataset seems to have little impact on the models when all textual features are present. Comparing the latter two models, it is noticeable that the additional numerical and categorical features have an influence on the new training data, as only by taking them into account was it possible to achieve more than twice as much improvement.

5. Discussion and Limitation

In this section, the results of this paper will be discussed. While the *TextOnly* model was able to achieve high accuracy in predicting risk classes on the basis of textual features alone, further consideration of numerical and categorical features ultimately led to the best-performing model. Consequently, using the real world dataset presented here, a multimodal transformer is superior to the simple transformer, even though the F1 scores only differ by 0.013 in the end. The way in which transformer ouput and numerical and categorical features were combined was taken from the Multimodal Toolkit, which uses only a simple Combining Module. Other studies are explicitly concerned with avoiding the bottleneck that arises at this point, so that even more advanced multimodal transformers have potential to achieve even stronger results.

Through augmentation of the training dataset, minor improvements in the models were achieved, particularly when the textual feature describing the risk was missing. However, analysis of the generated keywords and the resulting news articles revealed that certain articles did not align with the context of the risk, leading to some noise. This may be attributed to our approach, which cannot generate search strings for the GDELT API that are specific enough. The extracted keywords were not always the most obvious choice in retrospect for capturing the actual information of the risk, and the presented metrics and evaluation criteria only partially aid in increasing the informational content of the generated search strings. Another potential reason is the content of the underlying training data, which exhibits significant variability with respect to internal and external risks. Understandably, for specific internal risks that describe a particular process, external news articles that substantiate the information content of the risk cannot be immediately found. Furthermore, in certain quarters, no data was available through the GDELT API, and the API's current form does not provide sufficient filtering capabilities to precisely adjust search queries to the content of the risk.

6. Conclusion

In this study, we have considered the potential of implementing multimodal transformers for real-world risk data to incorporate different data modalities. We found that the different features had different influences on the model, with the inclusion of all of them producing a model that achieved the highest F1 score of 0.7983. Furthermore, we considered how keyword extraction and the use of different metrics can build search strings for the GDELT

DOC 2.0 API. Specifically, we explored how different metrics can find different numbers of articles and how their information content can be evaluated. The combination of these metrics eventually led to an augmented dataset, whereby further multimodal transformers were trained and their performance compared to the previously trained models. By using the augmented dataset, a slight increase in the F1 scores of the models has been achieved.

In future research, the impact of different model architectures for distinct data modalities on Transformer application should be explored. The Combining Module utilized in this paper should be specifically addressed in further studies. Along with designing more complex metrics, investigating the utilization of an API to search for news articles more specifically could also be valuable in data augmentation. By reducing generic keywords and generating more concrete queries, a potential improvement in data quality could be achieved. In addition, the application of new data sources may also be explored, e.g. those that better cover internal risks or fit better with specific risk categories, such as financial data.

7. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[2] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. Gomez, S. Gouws, L. Jones, Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, and J. Uszkoreit, "Tensor2tensor for neural machine translation," 03 2018.

[3] J. Hegde and B. Rokseth, "Applications of machine learning methods for engineering risk assessment – a review," *Safety Science*, vol. 122, p. 104492, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925753519308835

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[5] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[6] D. V. Le, J. Montgomery, K. C. Kirkby, and J. Scanlan, "Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting," *Journal of Biomedical Informatics*, vol. 86, pp. 49–58, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S153204641830162X

[7] F. Zhou, S. Zhang, and Y. Yang, "Interpretable operational risk classification with semi-supervised variational autoencoder," 01 2020, pp. 846–852.

[8] M. Fujii, H. Sakaji, S. Masuyama, and H. Sasaki, "Extraction and classification of risk-related sentences from securities reports," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100096, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096822000398

[9] S. Amal, L. Safarnejad, J. Omiye, I. Ghanzouri, J. Cabot, and E. Ross, "Use of multimodal data and machine learning to improve cardiovascular disease care," *Frontiers in Cardiovascular Medicine*, vol. 9, 04 2022.

[10] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," 2022.

[11] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *CoRR*, vol. abs/1909.02950, 2019. [Online]. Available: http://arxiv.org/abs/1909.02950

[12] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," 2020.

[13] J. Wu, T. Zhu, J. Zhu, T. Li, and C. Wang, "A optimized bert for multimodal sentiment analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 2s, feb 2023. [Online]. Available: https://doi.org/10.1145/3566126

[14] K. Gu and A. Budhkar, "A package for learning on tabular and text data with transformers," 01 2021, pp. 69–73.

[15] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2359–2369. [Online]. Available: https://aclanthology.org/2020.acl-main.214

[16] M. Grootendorst, "Keybert: Minimal keyword extraction with bert." 2020.

[17] "The GDELT Project," accessed March 29, 2023. [Online]. Available: https://www.gdeltproject.org/