Classification of Companies using Graph Neural Networks

Jovan Manchev [1], Mirsolav Mirchev [1,2], and Igor Mishkovski [1]
[1] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
[2] Complexity Science Hub, Vienna, Austria
jovan.manchev@students.finki.ukim.mk, {miroslav.mirchev, igor.mishkovski}@finki.ukim.mk

***Abstract* — Classification of companies into GICS categories can be addressed using Graph Neural Networks (GNN), by utilizing the different types of relationship between companies such as customer, supplier, partner, competitor, and investor. We use the Relato business graph data and compare the performances of several GNNs and a large language model like BERT that is trained only on the descriptions of the companies. Our goal is company classification into its corresponding category within the four tiers of the GICS hierarchy. Several architectures of GNNs are explored such as GCN, GraphSAGE and GAT, but also RGCN and RGAT that consider the edge type, or relationship between the companies. The main purpose is to reveal what kind of relationship between the companies is most valuable when determining the category of the company. The findings indicate that Graph Neural Networks (GNNs) enhance both classification performance and the understanding of collaboration patterns among companies, providing valuable insights for determining the industry in which these companies operate. This contrasts with the classification based solely on company descriptions using BERT.**

Keywords - graph neural networks, relato business graph, companies' classification

I. INTRODUCTION

Classifying companies into specific categories according to the Global industry categorization standard (GICS) [1], provides valuable insights into their economic activities. Accurate classification is essential for financial professionals because it helps in various ways. Firstly, different types of companies have distinct risk profiles and financial characteristics. Secondly, various industries face unique economic and market risks. Thirdly, comparing companies' performance within the same industries allows for effective benchmarking. Additionally, regulatory requirements often align with specific industries. Furthermore, strategic decisions and resource allocations are often industry dependent. This classification benefits investors and financial professionals by enhancing investment decision-making, risk assessment and management, performance evaluation, and strategic planning. Economists and financial analysts benefit from improved forecasting and market analysis. Lenders can better assess the creditworthiness of companies, and legal counsels can ensure compliance with financial regulations and reporting standards.

Artificial intelligence (AI) plays a pivotal role in streamlining the categorization of companies by automating the process. Leveraging advanced algorithms, AI analyzes extensive datasets to discern patterns and relationships. This automation simplifies and accelerates the classification of businesses, allowing for a more efficient and accurate assessment of their economic and operational activities.

Researchers and scientists have approached the company classification problem using various machine learning techniques. For instance, in [2] Doc2vec is used for embedding information from corporate disclosures and after that Ward's hierarchical clustering method is applied for categorizing securities. Clustering together companies into their industry fields using word embeddings generated from general news text using word2vec is examined in [3]. Different solutions for classifying companies are explored in [4] using short textual descriptions of companies and their economic activities. In [5] it is shown that using time series data for the companies achieves better results than using static data. Classifying companies using articles about companies and their economic activities using different text classification techniques, using both deep learning and classical vector-space models is considered in [6]. In [7], different models are used to generate word embeddings for text-based industry classification like BERT, word2vec, doc2vec and latent semantic indexing, and after getting the embeddings different approaches for clustering are used such as k-means, Gaussian mixture model and greedy cosine similarity. Another work [8], builds graphs using supply chain network data and after that GNN is used for company classification. Fine-tuning BERT on annual reports for companies for getting vector representations of companies is tried in [9]. An unsupervised approach on financial data using t-distributed stochastic neighbor embedding (t-SNE) in combination with spectral clustering is proposed in [10]. Two methods, dynamic industry representation and hierarchical assignment were proposed in [11], where the first represents industry as a sequence of time-specific vectors by integrating definition-based, structure-based, and assignment-based knowledge, while the second takes industry and firm representations as inputs, computes probability and assigns the firm to the industry with highest probability. Generating embeddings using companies' website data as input to Word2Vec and after that using unsupervised approach for clustering is examined in [12]. Some authors also tried to build knowledge graphs based on business reports, and then use large language models for enhancing the word representations with external knowledge [13].

Graph neural networks (GNNs) are powerful methods for various deep learning tasks such as node classification, link prediction, generating node and graph embeddings, discovering network communities and other tasks. By

enhancing company data using company interrelationships on top of company descriptions we demonstrate superior classification performance compared to relying solely on a natural language processing (NLP) model based only on the descriptions. In this work, we apply five architectures of GNNs for the problem of companies' classification, namely graph sample and aggregated GraphSAGE [14], graph convolutional neural network GCN [15], graph attention networks GAT [16], relational graph convolutional networks RGCN [17], and relational graph attention networks RGAT [18]. To the best of our knowledge, we are the first to address the problem of companies' classification using various GNN architectures by not only utilizing the company business background description, but also to integrate and evaluate the different types of relationship between companies such as customer, supplier, partner, competitor, and investor.

The paper is organized in the following way. Section II describes the categories in the GICS hierarchy and the dataset on which the GNNs were trained on. Section III describes the different types of GNN models used for the classification. In Section IV we provide the results, while Section V gives final thoughts about the problem and its solutions described in the paper.

II. DATA AND GICS HIERARCHY

We use a subset of companies extracted from the Relato Bussiness Graph [19] which contains different relations between companies. In this subset there are 7000 companies distributed in 7 sectors, 12 industry groups, 44 industries and 96 subindustries. Not all categories from each tier of the GICS hierarchy are present in the dataset [19]. The types and number of relationships can be seen in Table I.

Due to lack of data for the companies, the dataset is enriched using data augmentation. The initial descriptions and company categories were created using GPT models, and the training process was applied to this data. Different prompts were used for generating the descriptions and categories. Although the data is generated, these models can be applied to real data if it is available.

TABLE I. NUMBER OF RELATIONSHIPS FOR EACH TYPE

Relationship	***Number of relationships***
Partnership	16876
Customer	3065
Supplier	177
Investment	1423
Competitor	719

There are four tiers in GICS, and they are sectors, industry groups, industries, and subindustries. The sector is the highest level in the GICS hierarchy and there are 11 sectors, each representing a broad segment of the economy. The second tier is industry group, providing a more specific classification within a second. There are 24 industry groups, offering a finer breakdown of economic activities. For example, within the Information Technology sector there are industry groups such as Software & Services and Technology Hardware & Equipment. The third tier is further refining the classification to capture more specific business activities. There are 69 industries, offering a more detailed view of the company's operations. In the Software & Services industry group that was mentioned above, there are for example industries like Application Software or Internet Software & Services. The fourth and most detailed tier in the GICS hierarchy is subindustry. There are 158 subindustries, representing highly specific business activities. For example, in the Application Software industry there are subindustries like Enterprise Software or Systems Software.

The challenge in GICS classification usually lies in determining the appropriate subindustry for a particular company, which typically requires a thorough understanding of the company's core business activities and operations. However, besides classifying companies into their respective subindustries within the GICS hierarchy, we also consider the other classes in the GICS hierarchy.

III. GNN MODELS ARCHITECTURE AND CONFIGURATION

First, we have used different techniques to obtain high-quality initial node embeddings derived from the company descriptions. The initial node embeddings are vectors of 384 numbers generated using paraphrase-MiniLM-L6-v2 [20] and these vectors are inputs to the GNNs. Several models of GNNs were applied to the dataset, where the initial node embedding is the same for every graph and only the edges are different.

A graph is created using the customer, supplier, investment, partnership, and competitor relationships between the companies. We have employed the following models of GNNs, namely GraphSAGE, GCN, GAT, RGCN and RGAT. The first three architectures GraphSAGE, GCN and GAT are similar. The fourth and fifth architecture RGCN and RGAT differ from the first three because they take additional information such as edge type and edge attributes, while the first three do not distinguish the edges. The first layers of the architectures are learning vector representations of the nodes, while the final layer makes the prediction about the class. Some post processing techniques are also used, such as passing the embeddings learned from the graph neural network to a multi layered perceptron and applying traditional methods such as correct & smooth. Three GNN layers were used for this task where the first layer is an input layer that receives a vector of size 384 which is text embedding of the company description. The second layer transforms these input vectors into vectors of size 128, based on the graph structure.

In simpler words GNNs learn two weight matrices, one for the neighboring nodes and one for the node from which GNN is built. After multiplying the neighborhood matrix with the current embeddings of the neighborhood nodes and multiplying the node matrix with the current embedding of the given node, they aggregate the result, and a new embedding is established for the node. The third and final layer is the output layer where the dimension depends on how many classes there are to predict in each class, and after that softmax is applied to the output of the final layer. The difference between GraphSAGE, GCN and GAT is in the aggregating method used to extract information from the neighboring nodes, while RGAT and RGCN use the edge types as additional information.

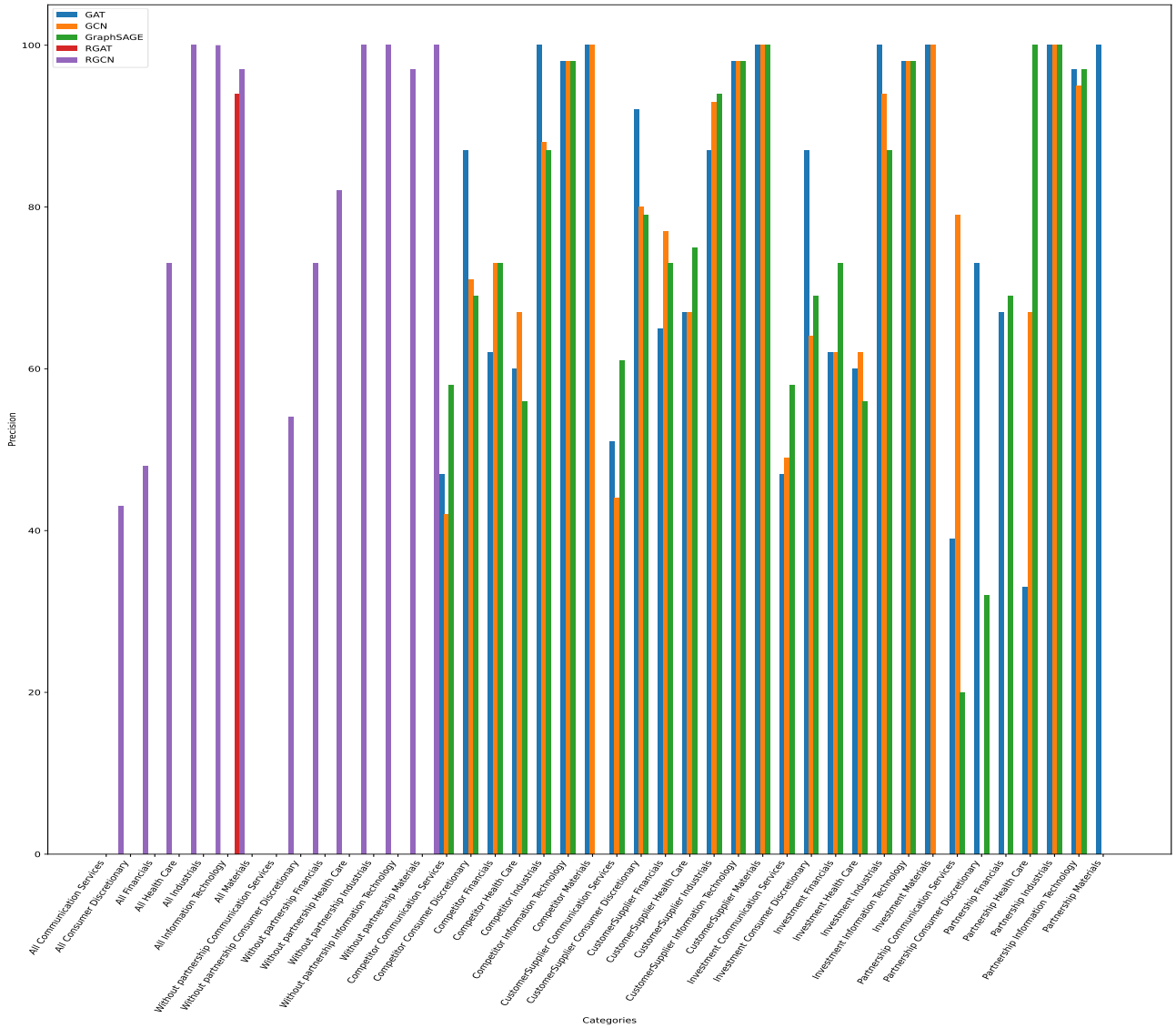

Figure 1. Models' precision across GICS sector categories. All relations are used for the first seven cases, all relations except partnership in the following seven, while single relationship types are used in the other cases. The first 7 cases are addressed with RGAT and RGCN, the second 7 with RGCN, while the other cases are addressed with GAT, GCN and GraphSAGE.

During the training of GCN, GraphSAGE, and GAT, we considered a single edge type or relationship, whereas in the training of RGCN and RGAT, all edges are considered, and the edge type is provided to the graph neural network as supplementary information. The main objective is to highlight the most valuable relationship between companies when determining their category. In both the training and the test datasets there are companies from every category, more precisely 80 % of the companies from each category are in the training and 20 % are in the test dataset.

IV. Results

In this section we will provide the evaluation of all the models for the different relationship types in terms of accuracy, as well as macro-averaged precision, recall and F1 score, which they achieve on the test set. Additionally, for the Sector tier the precision of each model for each Sector category is also shown. For the Industry Group tier, just the best model with the corresponding relation from which the graph was built is shown. Because the number of categories grows rapidly in the next two tiers, the models were not evaluated in deeper details with the precision metric. Generally, in all cases the accuracy is better than the precision, which is due to the unbalanced class distribution of the categories.

A. Sector

In Figure 2, the distribution of the companies across the first tier is shown, where the Information Technology sector dominates. In Table II the results from each of the models with corresponding type of relation are shown, where the GAT model in combination with the customer & supplier relationship achieves the best results. The BERT model which uses just companies' descriptions achieves good accuracy, but it performs poorly on the other metrics. Figure 1 shows the precision of each model in combination with relation types for each category in the Sector tier. The

RGCN model is examined with all the relations, as RGAT, but it also it is also tested without the partnership relation, because the GCN architecture achieved poor results on a dense graph, like the one with partnership relationships. When this was also implied using RGAT architecture the results were not improved so they are not shown. The RGAT architecture achieved worse performance, as can be seen in Figure 1, for most of the categories the precision is around 0. It has already been reported in other studies that RGAT can often completely fail in the learning process. Also, it can be seen in Figure 1 how GNNs trained on the graph built using partnership relation achieved worse performance than the GNNs trained on graphs using other types of relations, due to the fact that this relation type is far more present than the other. It can be concluded that as the graph gets denser it is harder to determine the sector in which the company operates.

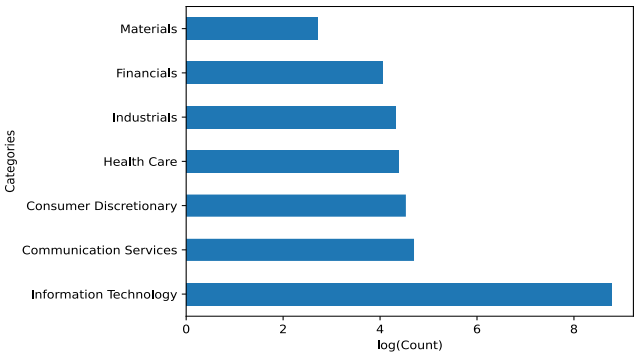

Figure 2. Distribution of each sector

TABLE II. RESULTS FOR GICS SECTOR CATEGORY

Model	*Relationship*	*Accuracy*	*Precision*	*Recall*	*F1*
GCN	Competitor	95	72	78	75
GraphSAGE	Competitor	96	79	79	79
GAT	Competitor	96	77	77	77
GCN	CustomerSupplier	96	77	77	77
GraphSAGE	CustomerSupplier	96	79	75	77
GAT	CustomerSupplier	96	78	80	79
GCN	Partnernship	94	38	29	33
GraphSAGE	Partnernship	90	41	53	46
GAT	Partnership	94	65	61	63
GCN	Investment	95	74	83	78
GraphSAGE	Investment	96	61	63	62
GAT	Investment	96	76	82	79
RGAT	All	93	14	14	14
RGCN	All	94	52	52	52
RGCN	Without Partnership	96	74	65	69
BERT	/	90	7	9	8

B. *Industry group*

In Figure 3, the distribution of each industry group is shown where the Software & Services industry group is most represented across the companies. In Table III, the performance of each model with the corresponding relation type from which the graph was built is shown, and for the Industry Group tier the GCN architecture on the graph built using just investment relation showed best performance. In Table IV, it is shown which architecture in combination with which relation achieved best performance according to precision for each category in the industry group tier. The BERT model achieved zero precision for each Industry group, except for Software & Services for which achieved precision of 91 percents.

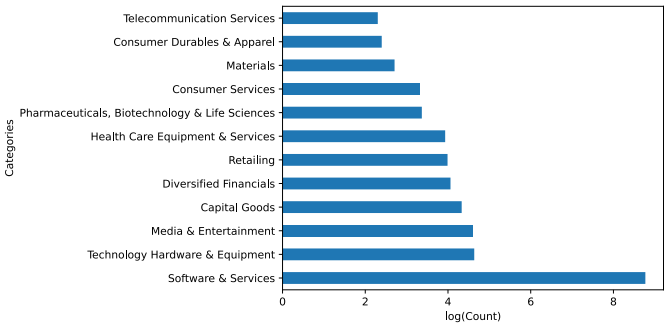

Figure 3. Distribution of each industry group

TABLE III. RESULTS FOR GICS INDUSTRY GROUP CATEGORY

Model	*Relationship*	*Accuracy*	*Precision*	*Recall*	*F1*
GCN	Competitor	94	61	55	58
GraphSAGE	Competitor	95	53	44	48
GAT	Competitor	95	62	53	57
GCN	CustomerSupplier	94	54	54	54
GraphSAGE	CustomerSupplier	94	46	44	45
GAT	CustomerSupplier	94	55	49	52
GCN	Partnernship	92	16	16	16
GraphSAGE	Partnernship	93	26	28	27
GAT	Partnership	91	9	9	9
GCN	Investment	95	67	58	62
GraphSAGE	Investment	94	44	44	44
GAT	Investment	94	57	51	54
RGAT	All	91	7	9	8
RGCN	All	93	36	32	34
RGCN	Without partnership	94	50	39	44
BERT	/	92	10	10	10

TABLE IV. BEST MODEL FOR EACH INDUSTRY GROUP CATEGORY ACCORDING TO PRECISION

Model	*Relationship*	*Industry Group*	*Precision*
GCN	CustomerSupplier	Capital Goods	88
GCN	CustomerSupplier	Consumer Drables & Apparel	100
GCN	CustomerSupplier	Consumer Services	100
GCN	Partnernship	Diversified Financials	64
GCN	Partnernship	Health Care Equipment & Services	50
GCN	Partnernship	Pharmaceuticals, Biotechnology&Life Sciences	83
GCN	Investment	Reatailing	96
GCN	Investment	Software & Services	44
GCN	Investment	Technology Hardware	50

C. *Industry*

On the third tier where there are 44 industries in the dataset, and as can be seen in Table V the models dropped in performance, but they are still better than the BERT model. Again, the partnership relationship is worse than the other types of relationship. The partnership type is most present in the dataset, and it is shown that as the graph is getting denser the performance of the model drops. Here the GCN architecture on the graph built using competitor relation achieved best performance.

TABLE V. RESULTS FOR GICS INDUSTRY CATEGORY

Model	*Relationship*	*Accuracy*	*Precision*	*Recall*	*F1*
GCN	Competitor	74	46	46	46
GraphSAGE	Competitor	73	42	40	41
GAT	Competitor	74	43	41	42
GCN	CustomerSupplier	73	42	40	41
GraphSAGE	CustomerSupplier	72	36	36	36
GAT	CustomerSupplier	73	43	41	42
GCN	Partnership	72	23	25	24
GraphSAGE	Partnership	71	22	27	24
GAT	Partnership	71	32	32	32
GCN	Investment	74	41	41	41
GraphSAGE	Investment	73	38	40	39
GAT	Investment	76	40	38	39
RGAT	All	59	1	2	1
RGCN	All	72	33	29	31
RGCN	Without partnership	75	47	38	42
BERT	/	65	1	1	1

D. Subindustry

In the fourth and final tier, subindustry, which is the most challenging, because it has the highest number of categories, the GraphSAGE architecture using customer & supplier relations achieved best performance, as shown in Table VI.

TABLE VI. RESULTS FOR GICS SUBINDUSTRY CATEGORY

Model	*Relationship*	*Accuracy*	*Precision*	*Recall*	*F1*
GCN	Competitor	49	27	31	29
GraphSAGE	Competitor	50	32	34	33
GAT	Competitor	49	31	33	32
GCN	CustomerSupplier	51	31	35	33
GraphSAGE	CustomerSupplier	52	35	35	35
GAT	CustomerSupplier	51	31	33	32
GCN	Partnership	46	20	25	22
GraphSAGE	Partnership	44	22	27	24
GAT	Partnership	42	20	25	22
GCN	Investment	50	30	32	31
GraphSAGE	Investment	49	27	31	29
GAT	Investment	50	28	32	30
RGAT	All	11	1	1	1
RGCN	All	43	31	24	27
RGCN	Without partnership	50	37	35	36
BERT	/	40	1	1	1

V. CONCLUSION

Although the classification performance of the models is not very high it needs to be taken into consideration that the number of classes is very large, the dataset is not big, and the distribution of the labels is imbalanced, with some labels having a significantly low number of samples associated with them. Nevertheless, the models are better than using just a fine-tuned large language model that uses just the description of the company for classification. Also using the individual types of relationships with the simpler models showed that better results are achieved than using all the relations and the multi-relational models. The RGAT architecture that uses all the relations achieved very poor performance, similar to the fine-tuned BERT. It also has been observed that as the number of edges in the graph grows, the performance of the GNNs drops. The RGCN architecture that uses all the relationships is far better than RGAT. Also, when the partnership relationship between the companies is removed the performance is improved by nearly ten percent. The same removal was also tried using RGAT, but the performance was not improved. The graph built using partnership relationship that is far more present than the other type of relationship is dense and the GNNs trained on this graph achieved much worse results, the recall score dropped for more than 10 percent compared to using competitor relationship type that is more than twenty times less present in the graph. By using graph neural networks, we showed that companies can be better classified if relationships between them are available compared to classification based only on their description. Therefore, we can conclude that the different types of collaboration between companies bring valuable information about the industry in which the companies are operating. There is still space for improvement of the results because the models were trained on a limited dataset, so the performance can be enhanced by further model refinements and data enrichment.

REFERENCES

[1] Global Industry Classification Standard (GICS), Available: https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook_2018_v3_letter_digitalspreads.pdf

[2] H. Yang, H.J. Lee, S. Cho, E. Cho, "Automatic classification of securities using hierarchical clustering of the 10-Ks", In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016, pp. 3936–3943.

[3] M. Lamby, and D. Isemann, "Classifying companies by industry using word embeddings", In Proceedings of the International Conference on Applications of Natural Language to Information Systems, Springer, 2018, pp. 377–388.

[4] S. Slavov, A. Tagarev, N. Tulechki, and S. Boytcheva, "Company industry classification with neural and attention-based learning models", In Proceedings of the 2019 Big Data, Knowledge and Control Systems Engineering (BdKCSE), IEEE, 2019, pp. 1–7.

[5] D. Katselas, B.K. Sidhu, and C. Yu, "Know your industry: The implications of using static GICS classifications in financial research", Accounting & Finance, vol .59, no. 2, pp. 1131–1162, 2019.

[6] A. Tagarev, N. Tulechki, and S. Boytcheva, "Comparison of machine learning approaches for industry classification based on textual descriptions of companies", In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 2019, pp. 1169–1175

[7] J. He, and K. Chen, "Exploring machine learning techniques for text-based industry classification", 2020. Available at SSRN: https://ssrn.com/abstract=3640205

[8] D. Wu, Q. Wang, and D.L. Olson, "Industry classification based on supply chain network information using graph neural networks", Applied Soft Computing, vol. 132, p. 109849, 2023.

[9] T. Ito, J. Camacho-Collados, H. Sakaji, and S. Schockaert, "Learning company embeddings from annual reports for fine-grained industry characterization", In Proceedings of the 2nd workshop on financial technology and natural language processing, 2020, pp. 27–33.

[10] S. Husmann, A. Shivarova, and R. Steinert, "Company classification using machine learning" Expert Systems with Applications, vol 195, p. 116598, 2022.

[11] X. Zhao, X. Fang, J. He, and L. Huang, "Exploiting expert knowledge for assigning firms to industries: a novel deep learning method", arXiv:2209.05943, 2022.

[12] C. Gerling, "Company2Vec–German company embeddings based on corporate websites", arXiv:2307.09332, 2023.

[13] S. Wang, Y. Pan, Z. Xu, B. Hu, and X. Wang, "Enriching BERT with knowledge graph embedding for industry classification", In Proceedings of the International Conference on Neural Information Processing. Springer, 2021, pp. 709–717.

[14] W.L. Hamilton, R. Ying and J. Leskovec, “Inductive reprentation learning on large graphs”, arXiv:1706.02216, 2018.

[15] T.N. Kipf, and M. Welling, “Semi-supervised classification with graph convolutional networks, arXiv:1609.02907, 2017.

[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks, arXiv:1710.10903, 2017.

[17] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling “Modeling relational data with graph convolutional networks”, arXiv:1703.06103, 2017.

[18] D. Busbridge, D. Sherburn, P. Cavallo and N. Y. Hammerla, Relational graph attention networks, arXiv:1904.05811, 2019.

[19] Data.world, Relato buisiness graph database, Available: https://data.world/datasyndrome/relato-business-graph-database

[20] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, arXiv:2002.10957, 2020.