NSM NORDIC STATISTICAL MEETING 2026

Applied text analysis to extract ownership relationships from annual reports

Ivan Elmehagen, Masters Student at Umeå University

Petter Ehn Wingårdh, Statistics Sweden, petter.ehn.wingardh@scb.se

Olof Löf, Statistics Sweden, olof.lof@scb.se

Abstract.

In a context of increasing regulatory demands, public authorities need to find new ways to comply with requirements, such as those arising from the European Group Register and European Profiling. Large language models offer new opportunities for identifying ownership relationships in text-based data sources that have previously been difficult to analyse automatically. This work examines the feasibility of automatically extracting structured information on ownership relationships from digital annual reports to enhance internal data availability, improve the quality of official statistics, and meet growing legal requirements. We fine tune BERT-based models for both named entity recognition and sequence classification, focusing on the detection of ownership-related statements. The analysis shows that ownership links between companies can be extracted from digital annual reports with high precision. While not yet implemented in production, the approach is expected to enable earlier validation of existing information and the identification of previously unrecorded relationships, thereby improving the quality of business registers and official statistics while reducing reliance on manual processing.

Keywords: Digital annual reports, AI/ML, text analysis, BERT

1. Background

Within the European Union, the production of business statistics is governed by common regulations that define statistical units, including enterprises and enterprise groups, in order to ensure the comparability of statistics across member states. Since enterprises are constructed from legal units within enterprise groups, reliable

information on ownership relations is essential for producing a wide range of official statistics.

At present, ownership information is primarily derived from a corporate group register based on data purchased from a commercial provider. In statistical production, this information is complemented with data from the European Group Register for multinational enterprise groups within the EU and EFTA. For the largest Swedish enterprise groups, ownership structures are additionally maintained through manual work at Statistics Sweden, a process that is resource intensive and difficult to scale as the enterprise group population continues to grow.

At the same time, Statistics Sweden has access to annual reports submitted to the Swedish Companies Registration Office, which constitute an administrative data source containing information on ownership relations. To make systematic use of this information, Statistics Sweden has established an internal register of digital annual reports. The increasing availability of annual reports in digital form enables methodological exploration of automated approaches to deriving ownership relationships from these documents.

This paper describes an exploratory development project at Statistics Sweden that investigated methods for extracting ownership relations from digital annual reports using large language models. The work is based on a master's thesis conducted at Statistics Sweden in 2025 by Ivan Elmehagen and has been adapted and condensed for the purposes of this paper.

2. Digital annual reports

Statistics Sweden receives annual reports from the Swedish Companies Registration Office daily. For companies applying K2 and K3 accounting rules, approximately half of the reports are currently received in digital form, with the remainder submitted as scanned documents (TIFF/PDF). The share of digitally submitted reports has increased steadily over recent years and is expected to continue to do so. This development is further supported by a government memorandum proposing mandatory digital submission of annual reports for these companies in the coming years (Regeringskansliet, 2023).

Digital annual reports contain ownership information at two distinct levels of structure.

Structured ownership information

The structured ownership disclosures represent the parts of the annual report where companies are expected to report ownership relations in predefined structured fields. These disclosures include information about:

1. the immediate parent company
2. parent companies at higher levels of the corporate group
3. subsidiaries of the reporting company

In principle, this structured information provides a strong foundation for constructing corporate group structures without the use of machine learning methods. In practice, however, the structured disclosures are incomplete. Experience from annual report processing shows that companies do not always report ownership relations in the intended fields, even though the same information may be described elsewhere in the report.

Ownership relations in unstructured text

In addition to structured disclosures, ownership relations are frequently described in free-text sections of the annual report. Such descriptions may appear in sections intended to describe the company's operations, organisational structure, or general context.

These textual descriptions vary in wording, detail, and placement within the report. Ownership relations may be expressed explicitly ("The company is a subsidiary of…") or implicitly ("The group consists of…"). As a result, this information is not necessarily accessible using rule-based methods and may require more advanced text analysis techniques.

3. Methodology

This study explores the use of supervised machine learning to extract ownership relations from unstructured text in annual reports.

The methodological choice is motivated by the observation that companies which report ownership relations in structured fields often describe the same relationships elsewhere in the annual report using free text. The structured ownership information was therefore used to generate initial labels, making it possible to train models without manually labelling all texts from scratch.

A key challenge is that extracting ownership relations requires both identifying company mentions and determining the relationship between them. Treating this as a single task can lead to ambiguous training signals, as the presence of company names in a text does not necessarily imply an ownership relation. This may cause models to conflate entity detection with relation classification.

To address this, a two-stage approach was developed. In the first stage, company names are identified using named entity recognition (NER). In the second stage, ownership relations are classified between pairs of companies based on the surrounding text.

Both stages use the National Library of Sweden's BERT models, KB-BERT, which have been pretrained to be representative for Swedish text (Malmsten, M., Börjeson, L. & Haffenden, C. 2020).

Stage 1: Named entity recognition

As a first step, company names were identified using named entity recognition. The objective was to detect all company names mentioned in the documents, independent of whether a specific ownership relation was expressed.

The text data were preprocessed and converted into a token-level representation, where each token was assigned a label according to the BIO2 tagging scheme (B-ORG, I-ORG, O), as shown in Table 1. The annotation was based on known company names derived from structured ownership information, which were matched to the text and used to assign labels to corresponding token spans. Generic references to the reporting company (e.g. ‘Bolaget’ or ‘Företaget’) were replaced with the company’s name to ensure that all company mentions were represented explicitly.

A pretrained Swedish BERT model (‘KB/bert-base-swedish-cased-ner’) was fine-tuned on the annotated dataset using a token classification setup. To obtain a realistic evaluation, the data were split at the company level, ensuring that texts from the same reporting company did not appear in both training and validation sets. Longer texts were divided into overlapping segments to accommodate model input constraints.

Table 1 – How tokens were labeled, according to BIO2 tagging.

Tokens	[‘company1’, ‘AB’, ‘is’, ‘a’, ‘subsidiary’, ‘to’, ‘company2’, ’AB’, ‘.’]
Labels	*[‘B-ORG’, ‘I-ORG’, ‘O’, ‘O’, ‘O’, ‘O’, ‘B-ORG’, ‘I-ORG’,’O’]*

Stage 2: Relation extraction

Ownership relation extraction was formulated as a relation classification task between pairs of companies mentioned within the same text.

Training data were constructed by generating pairwise combinations of company names identified in each text. Each pair, together with the full text, was used as input to a model that predicts the ownership relationship between the two companies (parent-to-child, child-to-parent, or no relation). Labels were derived from structured ownership information.

A pretrained Swedish BERT model (‘KB/bert-base-swedish-cased’) was fine-tuned for this task using a sequence classification framework. The input consists of the company pair followed by the document text, encoded as separate segments using BERT’s segment embeddings. This input structure separates the entity pair from the textual context, allowing the model to base its prediction on how the relationship between the two companies is described in the document.

Model predictions were reviewed to identify potentially incorrect labels, enabling iterative refinement of the training data.

In an operational setting, the company pairs used as input would be derived from the output of the entity recognition step.

Table 2 - The necessary information to fine-tune a BERT model for the task of relation extraction, via sequence classification. It shows information about the company that wrote the text and the relation between two companies based on what is says in the corresponding text.

id	name	company1	company2	text	labels
Reporting company id	Reporting company name	Company	Company	Text from annual report	OWNS, OWNED, NO_RELATION

4. Results

Stage 1: Named entity recognition

Table 3 summarizes the performance of a pretrained baseline model and the fine-tuned Stage 1 model on the validation set. The baseline model refers to the out-of-the-box ('KB/bert-base-swedish-cased-ner') model applied directly to the validation data without additional training, while the fine-tuned model corresponds to the same model after training on the annotated annual report training data.

Performance is evaluated at the entity level, meaning that a prediction is counted as correct only if the full company name is correctly identified. Precision measures the proportion of predicted entities that are correct, recall measures the proportion of true entities that are detected, and F1 is the harmonic mean of precision and recall.

Table 1 – The metric results on the entity level for both the fine-tuned model and the baseline model (KB/bert-base-swedish-cased-ner).

Stage 1	Baseline model	Fine-tuned model
F1	0.768	0.931
Precision	0.679	0.913
Recall	0.883	0.951

The results show a clear improvement compared to the baseline model, with the fine-tuned Stage 1 model achieving substantially higher precision while also increasing recall, leading to a much higher F1-score.

Table 4 provides a detailed breakdown of the fine-tuned model's performance by showing the number of correctly extracted, missed, and extra company names.

Table 4 – The number of companies that were expected, extracted, correct (true positives), missed (false negatives), and extra (false positives). Evaluation is performed at the entity level, meaning that the full company name must be correctly extracted to count as correct.

Stage 1: Fine-tuned model	Expected	Extracted	Correct	Missed	Extra
	3773	3931	3588	185	331

Stage 2: Relation extraction

Table 5 presents the performance of the Stage 2 model on the validation set. Here, each example consists of a pair of companies and the associated document text, and the model predicts a single label describing the ownership relationship between the pair. Evaluation is therefore conducted at the pair level.

Table 5 – The pair level metrics per class for Stage 2.

Stage 2	OWNED	OWNS	NO_RELATION
F1	0.984	0.981	0.974
Precision	0.977	0.981	0.984
Recall	0.992	0.981	0.965

The model achieves high performance across all classes, with F1-scores above 0.97. This indicates that the model can reliably determine whether an ownership relationship exists between two companies, and correctly identify which company is the parent and which is the subsidiary.

5. Discussion and conclusions

The results demonstrate that ownership information can be extracted from unstructured text in digital annual reports with high precision and recall.

The named entity recognition model achieves strong performance, with an F1-score of 0.93, indicating reliable identification of company names. Most company names in the

validation data are correctly identified, although some false positives and negatives remain, as shown in Table 4.

The relation extraction model further shows that relationships can be identified between pairs of companies, achieving F1-scores above 0.97 across the three classes.

The evaluation of the NER model is affected by incomplete annotations, meaning that some correct predictions may be counted as false positives. Improving the quality of the annotations and evaluating the full approach on new data are therefore important next steps.

Further work is also needed to assess how information from structured and unstructured parts of the reports can be combined in practice. In general, the approach needs to be developed to work in a production setting, with a clear end-to-end pipeline and a plan for continuous evaluation.

6. References

Elmehagen, I. (2026). Use of Swedish BERT for Extracting Company Relationships. Master's thesis, Master of Science in Industrial Engineering & Management – Data Science, Umeå University.

Malmsten, M., Börjeson, L. & Haffenden, C. (2020). *Playing with Words at the National Library of Sweden: Making a Swedish BERT*. arXiv preprint arXiv:2007.01658

Regeringskansliet (2023). Digital ingivning av årsredovisningar. Available at: https://www.regeringen.se/contentassets/d80917262ab14de5bbff617fcbda2b2c/digital-ingivning-av-arsredovisningar.pdf. 2026-06-01