Master Thesis

Evaluating pre-trained language models on partially unlabeled multilingual economic corpora

---

Author

Jacopo Rizzo

Supervisors

Prof. Dr. Christian Heumann, Dr. Matthias Aßenmacher

Advisors

Prof. Dr. Ralf Elsas, M.Sc. Moritz Scherrmann

Department of Statistics

Ludwig-Maximilians-Universität München

Munich, 26th of April 2022

Declaration of Authenticity

The work contained in this thesis is original and has not been previously submitted for examination which has led to the award of a degree.

To the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made. This applies also to all graphics, drawing, maps, tables and images included in this thesis.

Place and Date	Jacopo Rizzo

Abstract

This master thesis analyses economic text documents of public companies and aims to set up a transformer-based NLP model, which is able to automatically classify such documents in one or more self-defined classes. Specifically, the goal of this thesis is to fine-tune a pre-trained BERT model on documents of German public companies, and use this model to classify documents of American public companies. Therefore, we evaluate the model on two different, but related, data sources, which is comparable to evaluating a transfer learning problem. These documents can be divided into three main categories: labeled German documents from German public companies, unlabeled English documents from German companies and unlabeled English documents from American companies. As we are primarily interested in classifying the American documents, we make use of the labeled German data containing the same content of the English documents of the German companies to automatically label the latter, in order to have a labeled English dataset that can be used for fine-tuning. The results show that our model partly outperforms the chosen benchmark model by about 4 percentage points on the F1-score.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
BaFin	Bundesanstalt für Finanzdienstleistungsaufsicht
BERT	Bidirectional Encoder Representations from Transformers
BPE	Byte Pair Encoding
CLS	Classification Token
CNN	Convolutional Neural Networks
DL	Deep Learning
GRU	Gated Recurrent Unit
LM	Language Model
LSTM	Long Short-term Memory
ML	Machine Learning
MLM	Masked Language Modelling
NLP	Natural Language Processing
NN	Neural Networks
NSP	Next Sentence Prediction
RNN	Recurrent Neural Networks
SBERT	Sentence Bidirectional Encoder Represenations from Transformers
SEC	U.S. Securities and Exchange Commission
SEP	Separation Token
SGD	Stochastic Gradient Descent
STS	Semantic Textual Similarity
TL	Transfer Learning
WP	WordPiece

Chapter 1

Introduction

In economics, as in many other fields, it is essential to create the conditions to ensure the fairness for all parties involved. This becomes particularly important when dealing with listed companies, as it is not unusual for them to have a large number of stakeholders, who must be informed simultaneously whenever the companies take important decisions. This transparency is fundamental for a fair market. Thus, listed companies are obliged to publish specific documents reporting all the relevant information about important corporate events or valuable disclosures every time these occur (SEC, 2022). In Germany these kind of reports are best known as **Ad-Hocs** and are regularised and controlled by the Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin), which is the German Federal Financial Supervisory Authority, while in the United States these are named **forms 8-K** and the U.S. Securities and Exchange Commission (SEC), the American counterpart of the BaFin, clearly defines in which occasions these are required. These documents contain information that is considered *material* (Kenton, 2022), which means that there is a high probability that investors will take them into great consideration when making an investment decision. For example, a company publishes information concerning internal changes in the Executive Board. Or a pharmaceutical company announces the results of tests on a new drug that they plan to market. Normally a company releases a variable amount of such documents during a year. Taking all listed companies into account, this results in a large amount of documents. It therefore becomes very difficult for an investor to have a constant overview of various companies and to *filter* the documents that may be relevant for him. Automating this process would bring numerous advantages and of course would eliminate the likelihood of human errors. The automation of processes like this is becoming more and more important, especially when dealing with a big amount of data. In this regard, the advent of Artificial Intelligence (AI) models has provided incredible improvements in terms of results. And the branch of AI that deals with analysing and processing text data, and more generally the natural language used by human beings, is

better known as Natural Language Processing (NLP). In particular, in this field research is carried out to *teach* machines to understand and analyse natural language in different manners. And as with most modern AI models, *artificial neural networks* often form the basis of many NLP models. Although the idea of this type of model was already presented in the mid-1990s by Rosenblatt (1958), they have gained notoriety and only found a significant and regular application in recent decades. This is mainly due to the constantly increasing computational powers of CPUs and GPUs in recent years, with which it is nowadays possible to train and use AI models capable of tackling different NLP tasks with remarkable results, as for example summarising newspaper articles, answering messages automatically, as in the case of chat-bots, or classifying documents into defined categories. In this sense, a significant increase in performance in this AI area was achieved with the introduction of the **transformer**-based models by Vaswani et al. (2017). Transformers have partially revolutionised the world of NLP by succeeding in solving mathematical and technical problems present in previous models. They also paved the way for *pre-trained* models, i.e. models trained from scratch on a vast quantity of data that can be generally used as a starting point for various challenges. And among the first models with a transformer-based architecture, **BERT** is certainly one of the most important, being capable of achieving outstanding results in several NLP tasks.

The goal of this thesis is to implement an NLP model, which is able to classify economic documents in pre-defined classes. More precisely we want to *fine-tune* a pre-trained BERT model using the Ad-Hocs, which we will then use to classify the forms 8-K. Fine-tuning a model means taking the pre-trained version of it and training it further on a specific task using the own data. The problem or task for which we optimise the model with this process is also called the *downstream task*. In our case this amounts to optimising the model to perform a multi-label classification. This work is structured in the following way: In the next chapter we will introduce the problem setting in detail and will provide some related works. In chapter 3 all the models and methods used in this work will be explained, with a particular attention on transformers and BERT. Chapter 4 provides some descriptive statistics and describes the pre-processing steps we have done on the data, while in chapter 5 we analyse and interpret the results of our classification model. These will be further discussed in chapter 6, where we also present some ideas on what can be done to improve our results. Chapter 7 will draw the final conclusions of this thesis. Note that this work was done in cooperation with the Chair of Finance and Banking of the LMU Munich, whom we will refer to as the project partner and who provided us with the topic and the data.

Chapter 2

Problem Description and Related Works

In this thesis we want to set up a classification method that automatically classifies economic documents, i.e. the forms 8-K, into pre-defined classes. We will use a total of 22 classes, defined by the project partner, which we will look at more in detail in chapter 4. Each document can be classified in, and thus belong to, more than a single class. This means, that we face a so-called *multi-label* classification problem, i.e. a single instance can be assigned to more classes[1]. For example, a document reporting the quarterly financial results and informing that a new CEO will replace the current one would then belong to the classes *Earnings* and *Management* simultaneously. Bit since the forms 8-K are not labeled with our classes, they cannot be used for the fine-tuning process. Instead, labeled Ad-Hocs are available for the scope of this thesis. We will make use of these to create a suitable dataset for fine-tuning a BERT model. We are going to use the latter in turn to classify the forms 8-K. This process of using a classification model trained on some specific data to classify other data, which is related to some extent to the former one, can be seen as a kind of **Transfer Learning (TL)** problem. As Bengio et al. (2003, p. 526) defines it, TL "refer to the situation where what has been learned in one setting is exploited to improve generalisation in another setting". With respect to this, the language of the labeled Ad-Hocs represents the first major obstacle we have to deal with. The labeled Ad-Hocs that we have are only in German language. Fine-tuning a BERT model in one language and using it to classify text in another language would not be very reasonable, because one, it was not optimised for a multilingual context and two, most languages do not share much of their vocabularies, thus the shared knowledge would be limited. Consequently, a BERT model trained on a single language is more reliable than a model trained on more languages. Fortunately, many of the German

[1]This is in contrast to a *multi-class* classification problem, where a single instance can belong to only one class, i.e. the classes are mutually exclusive

public companies, obligated to publish such documents, do so in both languages, i.e. German and English, with both versions containing exactly the same information. We are therefore confronted with a multilingual context, where we have *partially labeled* data, i.e. the German Ad-Hocs, and *partially unlabeled* data, i.e. the English Ad-Hocs. But we can make use of the fact that both language versions of a given Ad-Hoc contain the same semantic information, to label the English one according to the German one. This will set up a labeled English dataset, which can be used to fine-tune BERT. With this we can then finally classify the forms 8-K in our classes. A first classification for specific categories defined by the SEC is actually already supplied with the forms 8-K. These categories are called *items* and the SEC defines 31 different ones. Each company filing such reports, needs to specify to which items the single sections of the form belong to. The main reason, why we do not use these already defined classes for our problem is that the definitions of these classes are really vague and shallow. The two most frequent items for example, are defined as *Regulations* and *Other events* and account alone for a third of the forms 8-K, despite the fact that these contain information which belongs to other categories as well. Hence, we want our model to be able to allocate categories to the disclosures in a more fine-grained manner. While fine-tuning such a model seems feasible, a second major problem arises when evaluating it. The inconsistency of the number and definitions between our classes and the US items makes it difficult for us to judge whether our model is able to correctly classify the single documents or not. In order to overcome this issue, we will propose an allocation between our classes and the items that yields some stimulating results.

In this work we will use the terms *label* and *class*, as well as their plural forms, indistinctly. Trivially in both cases we will refer to one or more of the 22 classes used in our classification problem. Since we are in a Deep Learning (DL) context, the technical terminology of this field will be used, and refer to a single data observation as an *instance* and to a variable in statistical terms as a *feature* (Google, 2021). Moreover, we will call the process of using our fine-tuned BERT on the forms 8-K the *TL task*.

2.1 Related works

Multi-label classification for text data using DL methods is an active and wide research field. Fine-tuning a pre-trained LM on a specific downstream task, as in our case multi-label classification, and on a specific semantic field, as in our case only on economic documents, has become a common practice in recent years and has significantly increased the performance of the various models. In this respect, the advent of *transformers* model by Vaswani et al. (2017) has brought considerable

improvements in the field of NLP, and consequently also for this specific downstream task. Sarwar et al. (2020) for example implement a **DistilBERT** model in order to automatic classify commit messages of software developers in multiple defined classes to improve the development process of the applications. More closely related to our topic Arslan et al. (2021) present a comparison between different LMs, for the classification of financial documents. Unlike us, they face a multi-class classification problem, but among the various models used, they also fine-tune a BERT model. Their experiments show that BERT performs slightly worse or in some cases the same as **RoBERTa** (Liu et al., 2019), which turned out in their case to give the best results. Very interestingly, and partially in contrast with the assumption that domain-specific fine-tuned models perform better, is the comparison with **FinBERT** (Yang et al., 2020), a domain-specific BERT version which is further pre-trained and fine-tuned only on financial text data. Indeed, in one of the scenarios the author experimented, RoBERTa was able to outperform FinBERT and generally it achieved similar results. Regarding a domain-specific multi-label classification problem the results of Chalkidis et al. (2019) on the classification of EU legislation documents in a vast amount of classes, show how their fine-tuned BERT yields the best results for different problem set-ups, compared to all the other models. Moreover earlier methods, which did not involve transformers models, such as the one presented by Du et al. (2019) supports the idea of fine-tuning a domain-specific model. The authors demonstrate how theoretically it is sufficient to fine-tune a bidirectional model such as **ELMO** (Peters et al., 2018) on medical text data to obtain good results with respect to a multi-label classification task.

In the pre-processing phase, we are going to label the English text data by looking at which part of the English text has the same semantic meaning of the German labeled one. To accomplish this we need to create sentence embeddings. LASER by Artetxe and Schwenk (2018) was among the first model to present remarkable results for this purpose. This model uses an encode-decoder architecture in the training process and discards the decoder when used for inference. The classification results presented in the paper showed a very good performance. Nonetheless, using a siamese network architecture as done in SBERT by Reimers and Gurevych (2019), is better if we want to find semantically similar sentences among different languages, while LASER performs better when the task is to find the exact translation in a cross-lingual context.

Chapter 3

Methods

In this section we will go through and explain all the methods used in this thesis. For overview and time reasons we will only dive into the details of the used methods. Nonetheless for the interested reader we provide references for all those mentioned methods that are not directly part of this work. In particular we focus our attention on *Word embeddings*, *Transformers* and *BERT*.

(Deep) Feedforward neural networks, or just Neural Networks (NN)[1], form the core of the topics we cover in this chapter. Apart from providing great results most of the time, the great advantage of NNs is their versatility. It is possible to use them to analyse both structured and unstructured data, for supervised (i.e. with labeled data) or unsupervised (i.e. with unlabeled data) learning tasks, as well as for hybrids of the latter, f.e. semi-supervised learning. The introduction of the *back-propagation* algorithm by Rumelhart et al. (1986)[2] for updating the parameters in combination with mathematical optimisation techniques, like the *Stochastic Gradient Descent (SGD)* in all its variations (Bottou, 1998), and the constantly increasing calculation power of computers rapidly boosted the fame and use of NNs since begin of the century. An example of a very simple fully-connected NN architecture, consisting of a single hidden layer, an input and an output layer is illustrated in figure 3.1. The parameters we want the model to learn are the individual connections between all the neurons, here represented as black lines between the layers, which are called *weights*, and for which we will use the matrix notation $\mathbf{W}$. For example here, each connection between the neurons of the input layer (i.e. the blue ones) and the hidden units (i.e. the green ones) and between these latter and those of the output layer (i.e. the yellow ones) correspond to some $w \in W$. When passing data to the model, each input (i.e. feature) is multiplied by the weight of the corresponding

[1]The correct term is Artificial Neural Networks, but to avoid annoying repetitions we refer to them just as NN

[2]They are the first to propose the use of this technique in the field of NNs. Backpropagation as idea already existed

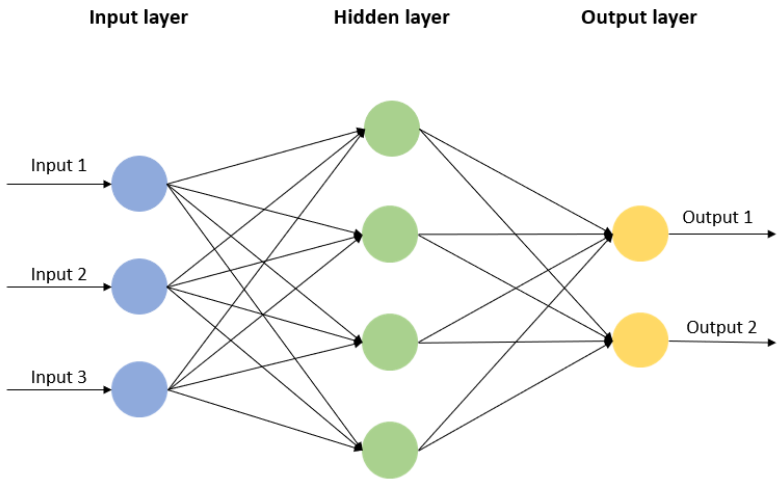

Fig. 3.1: Simple fully-connected neural network architecture with a single hidden layer consisting of four neurons (or *hidden units*), an input layer with three neurons, corresponding to the single features, and an output layer with two neurons. Usually a bias term, which we omit here, is added to each layer, except for the output one.

connection between two units and a bias term is added. The outcome of this operation is then passed to some function denoted as *activation function*, which activates the neuron towards which the connection is pointed. This happens between each layer in our network. The input of the hidden layer in our figure, which is passed to the activation function, can be expressed in mathematical notation as

$$W^T x + b \tag{3.1}$$

with W^T being the transposed weights matrix, x the vector of inputs, i.e. the features, and b the vector of the bias terms. We then use another function to classify or process the output of the last layer. In a multi-label classification problem we use a *sigmoid* function defined as

$$sigmoid(x)_c = \frac{1}{1 + e^{-x}} \tag{3.2}$$

to compute the probability for the output vector x (i.e. outcome vector or values of output units) to belong to class $c \in C$, with C denoting the set of all the classes of our problem. During training the model's output is compared to the true class using a *loss function*, or simply loss. In order to minimise this loss, which is the mathematical way of expressing the *task* of our model, we iteratively update the weights[3] via backpropagation, i.e. computing the gradient of the loss. Typical used

[3]In general in the beginning the weights (and biases) are initialised randomly, for example using **He initialisation** by He et al. (2015)

losses are the *MSE*-loss for regression problems and the *BCE*-loss, short for binary cross-entropy, for classification problems, which can be defined as

$$L(\hat{y}_n, y_n) = -\sum_{c=1}^{C} y_{n,c} * log(\hat{y}_{n,c}) + (1 - y_{n,c}) * log(1 - \hat{y}_{n,c}) \tag{3.3}$$

for a multi-label classification problem, with $\hat{y}_n$ being the model's predicted probability for instance $n = 1, ..., N$ (i.e. probability for an instance to belong to a class c computed with the sigmoid function), y_n being the true label for that instance and $c = 1, ..., C$ denoting the single classes of our problem. Usually, the true labels are binarized and set to 0 if the instance does not belong to that class and 1 if it does, yielding a binary vector of length equal to the number of classes of the problem. So, one of the two terms of the sum of equation 3.3 is always multiplied by 0. Moreover, normally more instances are inputted at the same time in the model to speed up the training or fine-tuning process. A group of input instances is also called a *batch*. In this case the loss is computed singularly for each instance in the batch and then the final loss, used for the weights' update, is computed by averaging all these losses by $\frac{1}{\mathcal{B}}$, with $\mathcal{B}$ denoting the number of instances in the batch.

NNs are called *feedforward* since all the information is evaluated by flowing from the input to the output layer through the intermediates hidden layers, without any feedback connection. This is a problem when modelling sequences of data, where each instance is related in some form to the others. For example time series or as in our case natural language in form of texts. Extending NNs to the case where we allow these connections leads to the creation of **Recurrent Neural Networks (RNN)**, which are the precursors of the transformers, the family of models we use in this thesis. We will look more in detail at the structure of these latter in section 3.2. For a more detailed overview of NNs in general (and in particular their classes, like RNNs and Convolutional Neural Networks (CNN)) we refer to Goodfellow et al. (2016, Ch. 6) and Hastie et al. (2017, p. 389-416). Details and computation of gradient-based optimisation, including backpropagation and SGD (*stochastic gradient descent*), can be found in Goodfellow et al. (2016, Ch. 4-6).

3.0.1 Neural Networks in NLP

Since in NLP we analyse and process language in a statistical way, each model is also called a **Language Model (LM)**. Basically a LM is nothing more than a probability distribution over a sequence of words. In other words, we want to model the conditional probability for a word i to be predicted given all or just some of the

other ones in the sequence, i.e

$$P(w_i^J) = \prod_{j=1}^{J} P(w_j | w_i^{j-1}) \tag{3.4}$$

with w_i being the $i-$th word and $w_i^j = (w_i, w_{i+1}, ..., w_{j-1}, w_j)$ the entire sequence in which the word appears. Parallel to this, the final target is to compute meaningful vector representations for words in a corpus, i.e. **word embeddings** (more on this in section 3.1). In this regard a first approach, implementing a simple feed-forward NN, is presented by Bengio et al. (2003). Many variations have been presented during the years, which included the use of RNNs, in combination with Long Short-term Memory (LSTM) (Graves, 2013) and Gated Recurrent Unit (GRU) (Cho et al., 2014) in order to solve the *vanishing gradient* problem (Hochreiter et al., 2001). But a first small revolution in that field is presented by the *Word2Vec* framework (see section 3.1.1) by Mikolov et al. (2013a,b), which laid the foundations for pre-trained word embeddings based on TL. In this regard, *ULMFiT* by Howard and Ruder (2018) is the first model implementing a unidirectional TL architecture, while *ELMo* (Peters et al., 2018) is among the first to use bidirectionality in the model. During the same year an improvement in the field was carried out by *GPT* (Radford et al., 2018) (also being the first model based on parts of the transformers architecture, see section 3.2) and by the Bidirectional Encoder Representations from Transformers (BERT) architecture, which was called from then on the NLP state-of-the-art model, introduced by Devlin et al. (2018). We are going to use BERT for the main analysis in this work and a detailed introduction is given in section 3.3.

3.1 Word Embeddings

Before jumping to the models used in this thesis we first need to define the important concept of *word embedding*. It is well known that computers, and hence in our specific case NLP models, only understand numerical representations (single numbers, vectors, matrices...). But in the natural language context we deal with sequences of letters and/or symbols. For example a text of an economic document can be seen as a sequence of semantically related words[4]. And a single word is nothing more than a sequence of letters. So, to be understood and analysed by the computers we need to convert words, or as we will see part of them, i.e. **tokens** (see section 3.1.2), into numbers, or more specifically vectors. A word embedding can therefore be defined as the numerical vector representation of a word. Thus, what we try

[4]The **semantic field** of a word is a set of words that refers to a specific subject and are related to each other (Faber and Usón, 1999, p. 67)

to achieve with an NLP model is to create the best possible word embeddings[5]. The outputs of most of the modern NLP models are in fact a multi-dimensional vector representation of the the input tokens. And a good starting point in order to construct these in a meaningful way is the **distributional hypothesis** by Firth (1957). This states that "*a word is characterised by the company it keeps*", hence words that appear in similar contexts tend to have similar meanings. Building on this we can assert that the best embedding includes all the necessary information of an input token, in particular its contextual meaning. We therefore seek to build *contextual embeddings*, i.e. context dependent representations of words.

3.1.1 Static Embeddings

The idea of using vector space models for representing text data is firstly introduced by Salton et al. (1975). First approaches make use of the intuitive one-hot-encoding. Assume we have a finite vocabulary V with 5 words, for instance $V = \{cat, dog, house, computer, bottle\}$, we can then simply create binary embeddings, such as

$$cat = \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix} \quad dog = \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix} \quad house = \begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix} \quad computer = \begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix} \quad bottle = \begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}$$

and use them for our downstream tasks. However this method has many disadvantages. First of all, words that are not present in the vocabulary are not represented. Second, the dimension of each vector depends on the dimension of the vocabulary, which for very large vocabularies leads to the *curse of dimensionality* problem (Hastie et al., 2017, p. 22-26). Third, a word will always get the same representation, i.e. context-independent[6]. Fourth, the output embeddings would be orthogonal to each other, making it impossible to compute a notion of word similarity based on some distance metric, since we would get the same distance for each couple of (orthogonal) vectors.

But what we actually want are dense, trainable, continuous vectors of a fix dimension that allows us to calculate similarity scores between them. These reasons, in combination with all the above-mentioned problems, paved the way to NN-based embeddings. Among the first approaches in this sense are the *CBOW*[7] and *Skip-Gram*

[5]From this point on, we will just use the term "embedding" to refer to a "word embedding"

[6]Think for example at the word *left*, which might refer to the direction itself or to the past simple of *leave*

[7]Short for Continuous bag of words

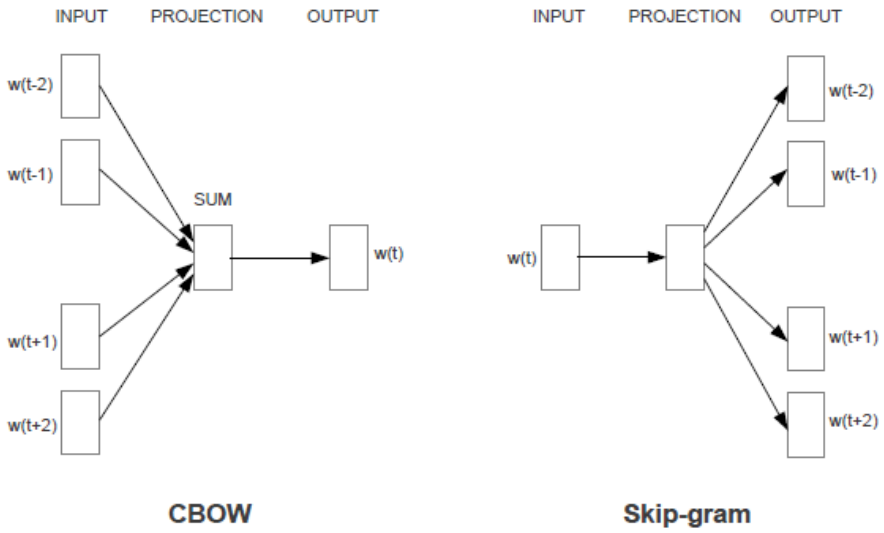

Fig. 3.2: CBOW (left) and Skip-gram (right) models' architecture. Image source: Mikolov et al. (2013a)

model proposed by Mikolov et al. (2013a). The architecture of these is represented in figure 3.2. Both consist of an input, an output and a single hidden layer. Let us stick to an example, in order to understand both models. Let us take the sentence *"I will eat pizza tonight for dinner."*. The task of CBOW is to predict a word given its context words in a fix window size (in the figure set to 2). This means that given the words *will*, *eat*, *tonight* and *for*, we want the model to predict the word *pizza*. While the task of Skip-gram is to predict the context words given the centre word (see Appendix A.1 for the mathematical notations). In general starting from a corpus $\mathcal{C}$ we compute one-hot-encoded embeddings for each word $w \in \mathcal{C}$. These vectors form the input to both models that are passed to the hidden units without any activation function (i.e. only computing the dot product between W and input vector(s)), which in turn are directly passed to the output layer. The probability of predicting a word is then computed by passing the hidden layers' outputs to a softmax function (Appendix A.2). Fine, but wait a second. We are looking for some low-dimensional, continuous embeddings, and what we do here is just computing a probability for one-hot-encoded embeddings. This is not very useful for our purposes. Indeed the embeddings are represented by the matrix of weights of the hidden layer and not by the model's output. These weights are learned using the usual techniques described at the beginning of this chapter. The dimension of the hidden layer and consequently of the features we are going to train is an hyper-parameter that can be tuned. The authors state that CBOW is faster, while Skip-gram is better for infrequent words.

We now have a method (actually two) that computes continuous embeddings of

a reasonable dimension, i.e. of a dimension $d < |V|$, with V being the number of words in our vocabulary. Using these methods we can also compute a similarity score between embeddings using for example the *cosine similarity*, defined as

$$cossim(emb_i, emb_j) = \frac{emb_i^T emb_j}{\|emb_i\|_2 \|emb_j\|_2} \tag{3.5}$$

where emb_i and emb_j respectively denotes the embedding for word i and j and $\|emb\|_2$ being the Euclidean norm of an embedding. The cosine similarity lies in the interval [-1, 1]. The higher it is, the more similar two embeddings are. For example we expect the cosine similarity of the embeddings for the words *earnings* and *income*, to be $cossim(emb_{earnings}, emb_{income}) \approx 1$. And that is great, since this solves our similarity problem. Nonetheless, both models only compute static embeddings for each single word, meaning that we still need to find a way to compute context-based representations. Transformers are going to help us solving this latter issue.

Skip-gram and CBOW are both part of the **Word2Vec** algorithm presented by Mikolov et al. (2013a,b). Alternatively Pennington et al. (2014) propose the **GloVE** algorithm, which unlike Word2Vec, implements methods that focus on words co-occurrences over a corpus. We refer to Goldberg (2019) for an overview of other methods like the *Bag-of-words*[8] model (in particular Ch. 2, 6), as well as for insights about the optimisation objectives of Skip-gram and CBOW (Ch. 10, keywords: *Negative-Sampling* and *Hierarchical Softmax*).

3.1.2 Tokenization

A problem related to the methods discussed so far is that if a word is not present in the corpus we use to train the model, then it is impossible to compute an embedding for it. We can solve this issue by using *tokens*. Manning et al. (2009, p.22-24) defines these as "sequences of characters in some particular document that are grouped together as a useful semantic unit for processing". So the process of tokenization means to chunk a text into smaller sequences of characters, called tokens. In this sense the separation of words by means of white-spaces in a text is in itself a type of tokenization. But we already explained why using single words as tokens might not be a good idea. Part of them or generally short sequences of letters, called *n-grams*[9] (Cavnar and Trenkle, 1994), where n can represent any $\mathbb{N}$, turn out to solve the problem carried by using words.

Various tokenization techniques have been developed. *Byte Pair Encoding (BPE)* (Gage, 1994) looks for the most common pair of consecutive bytes, i.e. letters/symbols, in a document and replaces this pair with a new single unused character (i.e. byte).

[8] This is to some extent the precursor of CBOW

[9] With $n = 1$ we call it a uni-gram, $n = 2$ a bi-gram and so on

The process is then repeated until no further compression is possible. More recent version of this algorithm (Sennrich et al., 2016) adapt this process to vocabularies, instead of documents. Other tokenizers, like *SentencePiece* (Kudo and Richardson, 2018) are also partially based on BPE. The tokenization technique we use in this work for preprocessing our data is called **WordPiece (WP)**, proposed by Schuster and Nakajima (2012) and Wu et al. (2016). The learning strategy of WP is similar to the one of BPE, but differs in the way the score for each candidate token is calculated. Figure 3.3 shows an example of WP tokenization on a small corpus. Starting from this, we first take all the single words and split them in *uni-grams*, i.e.

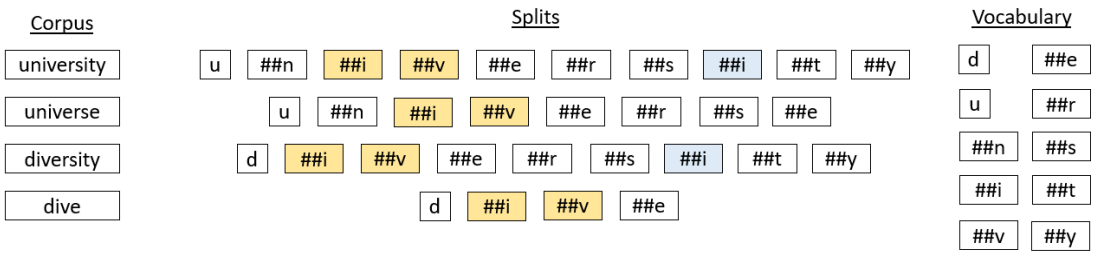

Fig. 3.3: Example of WordPiece tokenization. The hashtags in front of the characters means the word does not begin with these tokens.

single characters. These form our vocabulary, which contain all the single uni-grams present at least once in the corpus. We then compute the score for each pair of uni-grams using the following function

$$score = \frac{freq\ of\ pair}{freq\ of\ first\ element \times freq\ of\ second\ element}. \tag{3.6}$$

In our example the score for the pair "iv" would be $\frac{4}{6*4} = 0.1667$. Within a single iteration the score is computed for all possible pairs present in our corpus. We then replace the pair of uni-grams with the highest score with the bi-gram formed from these two and also add this new token to our vocabulary. We then repeat the process by also taking into account the new formed tokens added to the vocabulary. As a stopping criteria for the procedure the authors propose either reaching the desired vocabulary size, or the incremental increase of the likelihood. Once the vocabulary is set we assign to each token a unique value. In order to tokenize a text we then start by looking at the longest possible token present at the beginning of the text and tokenize it accordingly to our vocabulary. A great advantage of WP is that it is language independent, hence useful for many tasks.

3.2 Transformers

Towards the end of 2017, a new type of architecture, which do not include the use of RNNs or CNNs and therefore solves the problems associated with them, revolutionised the world of NLP. Transformer-based models introduced by Vaswani et al. (2017)[10] have led to significant improvements in the models' performances, in addition to a decreased training time, thanks to their higher parallelization compared to RNNs or CNNs. This is possible thanks to the **Attention** mechanism (more on this in section 3.2.1). Figure 3.4 shows the basic architecture of a transformer. We can distinguish two main components, the *encoder* (left part of the image) and

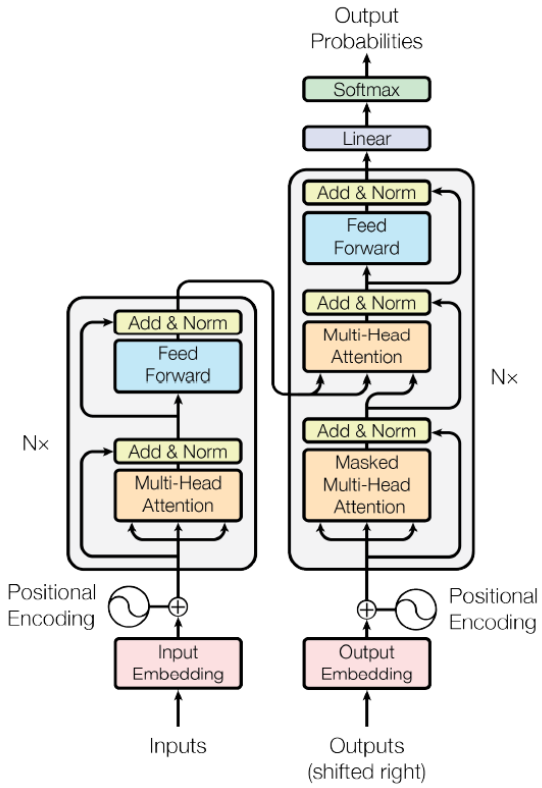

Fig. 3.4: Transformers architecture. Image source: Vaswani et al. (2017)

the *decoder* (right part of the image). Before passing data to the encoder, this is preprocessed by an *Input Embedding* layer, which converts it to numerical values (i.e. vectors), for example via a (pre-trained) BPE or WP tokenization (section 3.1.2). The lack of recurrences and convolutions makes it necessary to supply the model with some more information about the order of the various elements in the sequence. Otherwise we would always get the same representation for a given embedding, regardless of its contextual meaning (see Appendix A.3). This is done by adding a *positional encoding* (i.e. a vector of the same dimension) to the embeddings. The

[10] All the topics covered in this section refers to this paper

authors propose a sine and cosine function to compute these, respectively defined as

$$\begin{aligned} PE_{pos,2i} &= sin(pos/10000^{2i/d_{model}}) \\ PE_{pos,2i+1} &= cos(pos/10000^{2i/d_{model}}) \end{aligned} \tag{3.7}$$

with *pos* being the position and i the dimension of the input token[11] and d_{model} being the dimension of the output produced by the model. The choice of these two functions in tandem is motivated by the linear properties they carry, which makes it easier for the model to learn which tokens to attend to. But in principle computing the positional encodings can be done using other methods. The output of the preprocessing step is passed to the encoder that is composed of a stack of N identical layers (in the figure only one is represented), whose job is to map its input to an abstract, continuous sequence that captures all the learned information for that input. A single layer consists of two main sub-layers, a *Multi-Head Attention* layer followed by a fully-connected NN, both of them being normalised, before passing them to the next (sub-)layer. There is also a residual connection around each sub-layer, in order to provide the model with more information. The output of the encoder is passed to the Multi-head Attention sub-layer of the decoder, which also consists of a stack of N identical layers[12] and whose job is to generate text sequences. The only architectural difference between a decoder's layer and an encoder's one is represented by an additional *Masked* Multi-Head Attention[13] sub-layer placed before the Multi-Head Attention sub-layer of the former. The decoder's output is passed to a linear layer that acts like a classifier and then to a softmax that computes the probabilities for each word. The weights are then updated in the usual way via backpropagation after computing the loss. Decoders are *autoregressive*, which means that their outputs are fed back into them as an additional input (after being preprocessed as for the encoder). A big advantage of transformers is that the encoder and the decoder can be separated and used as *independent* models. In this thesis, we will only use the encoders.

The experiments performed by the authors on different tasks and using different parameters demonstrate that attention-based models outperform all the previous NLP models (Vaswani et al., 2017, Tab.2). Moreover attention layers can also be trained significantly faster than other type of layers (Vaswani et al., 2017, Tab.1).

[11]This means that for tokens located at an even position in the sequence, the positional embedding is computed using the sine function, and for tokens at an odd position using the cosine function

[12]For their experiments the authors chooses $N = 6$ for both, the encoder and decoder

[13]The masking prevents the model to attend to words that are generated subsequently a word, hampering the model from "cheating" by simply looking at these

3.2.1 Attention!

The heart of transformers is the *Attention mechanism*, or just *attention*. This technique is firstly proposed by Bahdanau et al. (2014) and has three main components: a query q, a key k and a value v. We can define it as a function that maps queries in combination with pairs of keys and values to some output, with queries, keys, values and outputs being vectors. The attention mechanism used in encoders is also known as *Self-Attention*. This kind of attention allows a model to associate each individual word (token) in an input sequence to the other words of that sequence. The idea is to train the model to understand which previous tokens should be put attention on when processing the next one. This is done by using a scoring function that computes a "relevance" score for each query-key pair. Let us see step by step, how such a mechanism works. We will use the matrix notations Q, K, V for the matrices containing the queries, keys and values vectors.

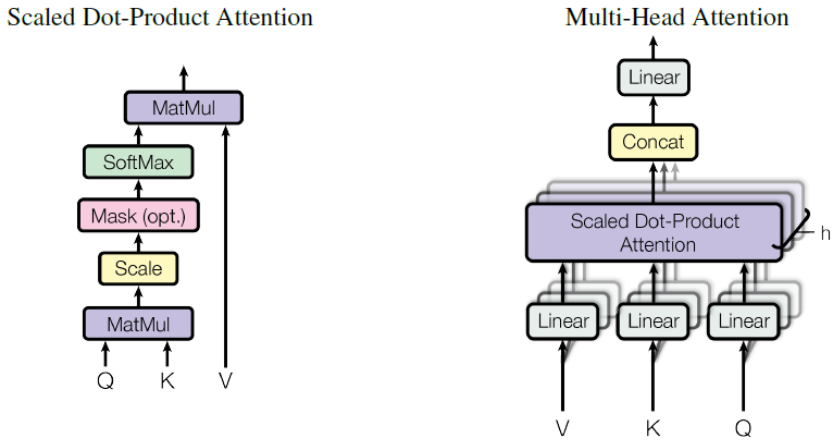

Fig. 3.5: Attention mechanism with the scaled dot-product attention (left) and the multi-head attention (right) that can be seen in figure 3.4. Image source: Vaswani et al. (2017)

Figure 3.5 shows the attention mechanism used within the transformers architecture. The embeddings are fed into three distinct linear fully connected layers in parallel. The output of these are the queries, keys and values matrices. We can think of them as three different abstractions of our embeddings. The weights (matrices) for computing these are respectively denoted W^Q for Q, W^K for K and W^V for V. These are normal parameters of the model and are therefore trained during training. We then compute the *Scaled Dot-Product Attention* (left part of figure 3.5). In this process Q and K undergo a dot-product matrix multiplication. The result of this produce a (score) matrix, whose entries (i.e. scores) determines how relevant the other tokens are for a given token. Hence this matrix quantifies how much *attention* should be put on the other tokens (the higher the score, the higher

the focus). This score matrix is then scaled by $\sqrt{d_k}$, with d_k being the dimension of the keys[14]. This allows for more stable gradients, since multiplying values can lead to exploding effects, thus resulting in the *exploding gradient* problem (Pascanu et al., 2012). We then compute the scaled final score for each query-key pair using a softmax, which yield probabilities[15], i.e.

$$\alpha_{ij} = \frac{exp(q_i^T k_j)}{\sum_{l=1}^{n} exp(q_i^T k_l)} \tag{3.8}$$

with α_{ij} denoting the attention score for the dot-product of the $i-$th query vector and the $j-$th key vector. This implies that higher scores get heightened and lower ones depressed, giving the model more confidence on which tokens to attend. The result of this operation is multiplied with V. The output of the softmax intrinsically decides which tokens of the values matrix are more important. In mathematical notations the Scaled Dot-Product Attention process can be summed up as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.9}$$

with K^T denoting the transposed K-matrix. All this process is repeated h times simultaneously by separate identical layers. A single layer is called a *head* and packed together we refer to the entire system as *Multi-Head Attention*, which is depicted on the right side of figure 3.5. So, we basically learn h different projections of Q, K and V, with a fix dimension, allowing the model to extrapolate information by looking at different (independent) representations of the same input at the same time. These are then concatenated together, before being further processed. We can therefore define Multi-head Attention as

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{3.10}$$

and a single attention head as

$$head_h = Attention(QW_h^Q, KW_h^K, VW_h^V) \tag{3.11}$$

with $W_h^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_h^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_h^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ being matrices of the parameters, and d_v being the dimension of the values vectors[16]. W^O is the weights matrix of the last liner layer after the concatenation, i.e. the

[14]This is equal to the dimension of the queries, i.e. $d_k = d_q$

[15]In figure 3.5 (left) we can notice a further step, called "Mask" before the softmax. We omit this here, since it is mainly used for the decoders, which are not used in this work. For more details about this we refer to Vaswani et al. (2017)

[16]Note that in this work $d_k = d_q = d_v = d_{model} = 768$, which is the dimension of a single embedding

outputs' weights. In BERT's base variant, which we use in this work, $h = 12$ and for each head $d_k = d_v = d_{model}/h = 768/12 = 64$.

There exist several variations of the attention mechanism. Besides the above described Self-Attention, another common choice is the *Cross-Attention* (sometimes referred to as *Multi-dimensional Attention*). This latter differs from the former in that it takes into account also the true targets/labels of a supervised learning task for example. Other variants like the *Hierarchical Attention* make use of the hierarchical lexical properties of semantics, f.e. $character \in word \in sentence \in document$. For a detailed overview of these and other variants we refer to Dichao (2020).

3.3 BERT

In this section we introduce **BERT**, the main model we are going to use in this work. BERT is proposed by Devlin et al. (2018), with the intent of pre-training deep bidirectional representations from unlabeled text and providing a pre-trained model that can be easily fine-tuned on specific downstream tasks. The rough architecture of our used version is depicted in figure 3.6. The authors propose two different version, a base one and a large one. The one we use is the base version, which is composed of a stack of 12 encoders. Each of these itself use 12 attention heads for computing the Multi-Head Attention, and outputting embeddings of dimension $d_{model} = 768$ for a total of about 110 million parameters. Before BERT the prediction of the next token for an input sequence was the commonly used LM objective. BERT instead is trained on two major tasks: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP).

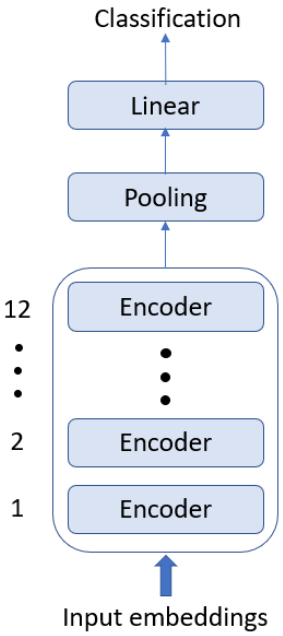

Fig. 3.6: Architecture of BERT used in this work

Masked Language Modelling

The main goal of this task is to predict the *masked* tokens of an input sequence. Given for example the sequence "*The child plays in the park.*", we randomly select and mask 15%[17] of the input tokens. This is done by replacing them with a `[MASK]` token, i.e. "*The child [MASK] in the park.*". We then train the model on predicting which token was replaced by `[MASK]`. This allows the model to have conditioning on context tokens from both sides (left and right) of the token to be predicted.

[17] These selected tokens have actually only 80% of chance of being really masked, since otherwise this would create a mismatch between pre-training and fine-tuning. For more details we refer to Devlin et al. (2018, Sec. 3.1)

Next Sentence Prediction

The other main task of BERT is the NSP task. This is particularly useful for all those downstream tasks based on understanding the relation between two sentences (f.e. Question Answering or Natural Language Inference). We can binarize this task, such that given a pair of sentences A and B, we train the model on understanding whether they are consecutive sentences or not. For example, ideally the model does not classify the sentences "*The child plays in the park.*" and "*The catcher in the rye is a great book.*" as consecutive sentences.

For both tasks BERT uses a separate cross-entropy loss. In pre-training the final loss is simply a linear combination of these two. But in order to accomplish these tasks we need to slightly modify the input embeddings generated for example with a pre-trained WP tokenizer. We do this by adding a position embedding (for the reason explained in the previous sections) and a segment embedding to each token embedding. This latter is done to provide BERT with the information whether a token belongs to the first or second sentence for the NSP task. Figure 3.7 shows

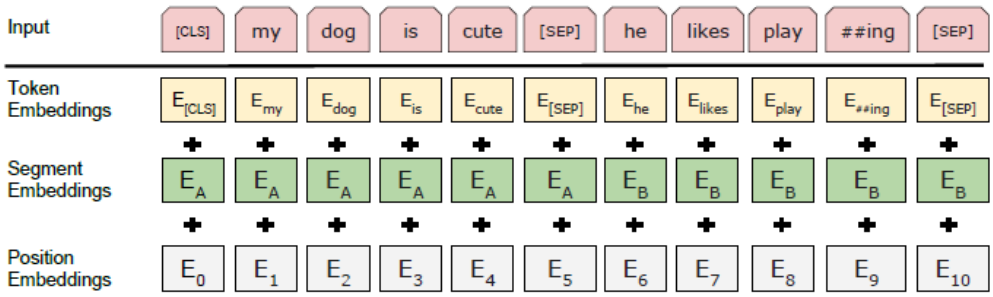

Fig. 3.7: Representation of BERT's input embeddings. These are the sum of the token, segment and position embeddings. Image source: Devlin et al. (2018)

BERT's input representation. Moreover, we supply our inputs with two special tokens: a `[CLS]` token indicating the beginning of the sequence and a `[SEP]` token that indicates the end of a sentence, i.e. placed between a pair of sentences and at the end of the entire sequence. The final embedding of the `[CLS]` token, short for *classification* token, as its name already states is used for the final classification within a classification task[18]. In fact, as the experiments of Clark et al. (2019) show, most of the information, i.e. the entropy, contained in an output sequence is included in this token embedding, which therefore can be seen as a sort of *sentence embedding*. So, the *pooling* layer of figure 3.6 is not to be intended as an usual pooling layer (like *max-* or *average*-pooling), but rather as a layer that pools out, i.e. extracts, only the final embedding of the `[CLS]`. This is then passed to a fully-

[18]In principle also using other tokens, like f.e. the average of all the output tokens is also possible. In this work we use only the `[CLS]` for the final classification

connected layer, which after being activated finally classifies the output using, in our multi-label context, a sigmoid function.

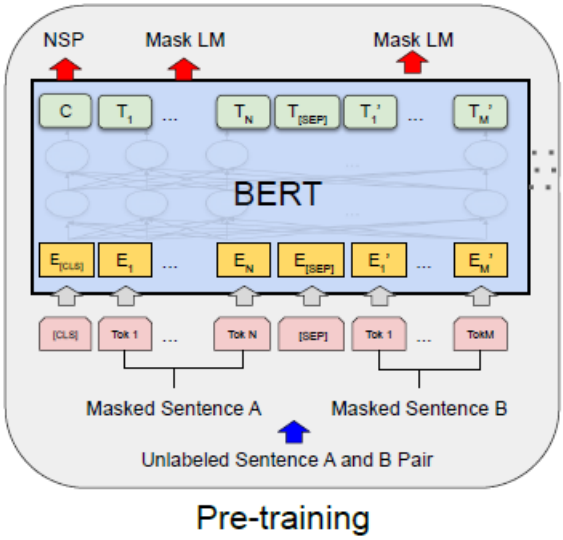

Fig. 3.8: Pre-training procedure for BERT, with all its peculiarities, like the `[CLS]` and the `[SEP]` token and both its training objectives, i.e. MLM and NSP. Image source: Devlin et al. (2018).

Figure 3.8 sums up the pre-training procedure, with all its main characteristics. In its original version BERT is trained on the BooksCorpus (about 800M words) (Zhu et al., 2015) and English Wikipedia (about 2500M words) for approximately 40 epochs. To the time it was presented, BERT[19] outperformed all of the preexisting LMs on various tasks, becoming de facto the state-of-the-art NLP model. More details on BERT can be found in Devlin et al. (2018) and Pilehvar and Camacho-Collados (2020, Ch.6), which we also take as reference for everything discussed in this section.

3.4 Multilingual Transformers

With BERT we now have a method that enables us to derive contextualised words and sentence[20] embeddings, showing great performance. Among the various fields of NLP, BERT also provides good results for **Semantic Textual Similarity (STS)** tasks. These kind of tasks include all those problems where we want to find out how *similar* two sentences are in terms of their semantic content. However, even though BERT seems to achieve these quite good, two major problems still remains, at least for the purpose of this thesis. The first one is a computational problem, since we have to provide BERT with all the sentences and compute some STS-score for each

[19]The same pre-trained model was separately fine-tuned on each task
[20]In this context sometimes it is also referred to such models as *Sentence transformers*

possible pair, in order to find the most similar one. Doing this for a collection of 10.000 sentences requires to do about 50M inference steps, which would take approximately 65 hours. Second, BERT is trained on a monolingual corpus. Thus, finding the most similar pair of sentences across two different languages is not a reasonable task. In order to solve the first problem Reimers and Gurevych (2019) propose the Sentence Bidirectional Encoder Represenations from Transformers (SBERT), a modified version of the pre-trained BERT, which uses a siamese network architecture, and which significantly speeds up computational time. Siamese NNs are first

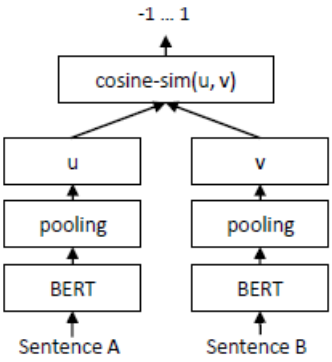

Fig. 3.9: Architecture of SBERT. Image source: Reimers and Gurevych (2019)

proposed by Bromley et al. (1994), who wanted to set up a model able to detect signature forgeries[21], by comparing the vector representations of two signatures using some distance metric. Such a structure lends itself very well to NLP problems, since being embeddings vectors one can directly and easily calculate the distance between them. The idea behind a siamese NN is to use two networks with the same architecture, f.e. two pre-trained BERTs, which share the same weights. Meaning, that we have to update a single set of parameters, albeit implementing two models. The two networks work in tandem. We input a sentence to each of them in parallel and each model will output a sentence embedding using some pooling operation. We can then apply a distance or similarity metric on them, in order to compute their semantic similarity. Figure 3.9 depicts the architecture of SBERT, which applies the cosine similarity[22] on the output vectors. The model is trained with a mean-squared-error loss. The authors experimented on this architecture with different setups and the results show a remarkable gain in performance compared to BERT with respect to STS tasks. For these and further details on SBERT we refer to Reimers and Gurevych (2019).

[21] Actually a similar idea is adopted earlier by Baldi and Chauvin (1993) to recognise fingerprints

[22] An high cosine similarity means that two sentences contain the same or very similar information

Knowledge distillation

Reimers and Gurevych (2020) also propose a method to solve the monolingual issue. They call this new approach *multilingual knowledge distillation*. The idea behind this, is that the translation of a sentence in another language should be mapped in the vector space to the same location as the original one, or at least very close to it. Figure 3.10 depicts the concept of knowledge distillation. Similarly to

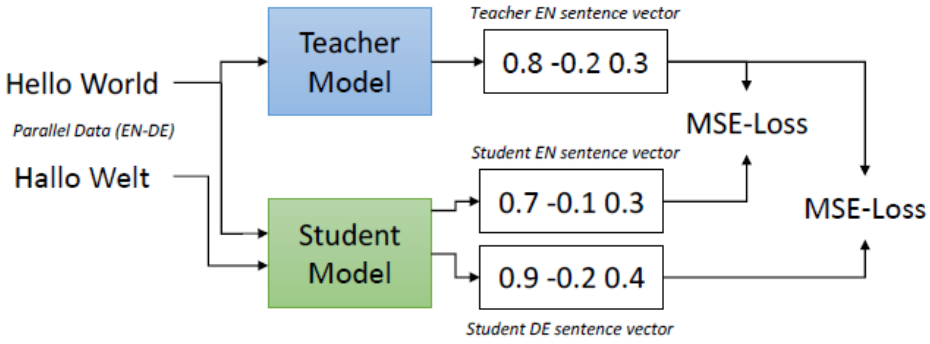

Fig. 3.10: *Multilingual knowledge distillation* training procedure. Image source: Reimers and Gurevych (2020)

a siamese NN, here we also have two models that work in tandem. But with two main important differences: both model do not necessarily have the same weights and architectures. We refer to the first model as the *teacher model* M and to the second as the *student model* $\hat{M}$. These definitions suit perfectly, since the idea is that the student $\hat{M}$ distills the knowledge of the teacher M. As input for training we need a set of parallel translated sentences $((s_1, t_1), ..., (s_j, t_j))$ with t_j being the translation of sentence s_j. We input s_j in both M and $\hat{M}$ and t_j only in the student. The teacher computes an embedding for sentence s_j and the student computes an embedding for each s_j and t_j. The weights are then updated in the usual way, by computing the mean-squared loss of the outputs. This can be mathematically summarised as

$$\frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}}[(M(s_j) - \hat{M}(s_j))^2 + (M(s_j) - \hat{M}(t_j))^2] \tag{3.12}$$

with $\mathcal{B}$ being a mini-batch, $M(s_j)$ the output embedding of the teacher for sentence j of the source language, $\hat{M}(s_j)$ the embedding of the student for the same sentence and $\hat{M}(t_j)$ the embedding of the student for the translation of sentence j. The goal is to train the model to mimic the relations $\hat{M}(s_i) \approx M(s_i)$ and $\hat{M}(t_i) \approx M(s_i)$. The model was trained using different datasets on 50 distinct languages. Once again, the authors carried out experiments with different setups and most of them outperformed the models used as benchmark. For these results and other details

about multilingual model distillation we refer to Reimers and Gurevych (2020).

The pair of most similar sentences in two different languages can then be found by passing, f.e. a German sentence to the teacher and a set of suitable candidates of English sentences to the student. With the output embeddings we can then compute the cosine similarity for all possible pairs and pick the one with the highest score. In our implementation we use as teacher model an MPNet (Song et al., 2020) and as student an XLM-RoBERTa (XLM-R) (Conneau et al., 2019). MPNet is a model based on the XLNet model (Yang et al., 2019) that unifies the MLM (section 3.3) training objective of BERT and the *Permuted language model* (PLM) training objective. In contrast to MLM, PLM uses various permutations of the input sequence during training, which intrinsically means that the model acts in a bidirectional manner without needing to corrupt the input as done in MLM with the `[MASK]` token. XLM-R instead is a transformer-based multilingual masked LM model, which is based on the cross-lingual XLM model (Lample and Conneau, 2019) and the RoBERTa model (Liu et al., 2019), with this latter being a robust optimised version of BERT trained without the NSP objective task. We will not cover any of these models here in detail for one, time reasons and two, since we are going to use only the already fine-tuned knowledge distillation model by Reimers and Gurevych (2020). For more details about the models we refer to their original papers, i.e. Song et al. (2020) for the MPNet and Conneau et al. (2019) for XLM-R. Albeit using this set up, which does not include a BERT model, for the sake of simplicity we are going to refer to this model, i.e. the one we use to find the most similar semantic sentences between the German and the English Ad-Hocs , simply as SBERT.

Chapter 4

Data

In this section we introduce the data used in this thesis. We present the various datasets used by providing some descriptive statistics, as well as all the preprocessing steps done in order to set up the final dataset utilised to fine-tune our final model. We will primarily focus only on the features used in our analysis. The interested reader can find an overview of all the available features in each dataset in Appendix B.1.

4.1 Original Data

Our original data is divided into three main parts, hence in three datasets. Throughout this thesis we will refer to each of them using a specific name. The *Origin* dataset contains the Ad-Hoc announcements. For the majority of the Ad-Hocs there is both a German and an English version. The *goldstandard* dataset (GS) contains the labeled sentences (or pair of sentences) of the German Ad-Hocs, while the *forms 8-K* dataset contains the forms 8-K data. These latter are used to evaluate our fine-tuned model on the TL task. Figue 4.1 shows the relation between the three datasets. The Origin and the GS share the same German data. While in the former

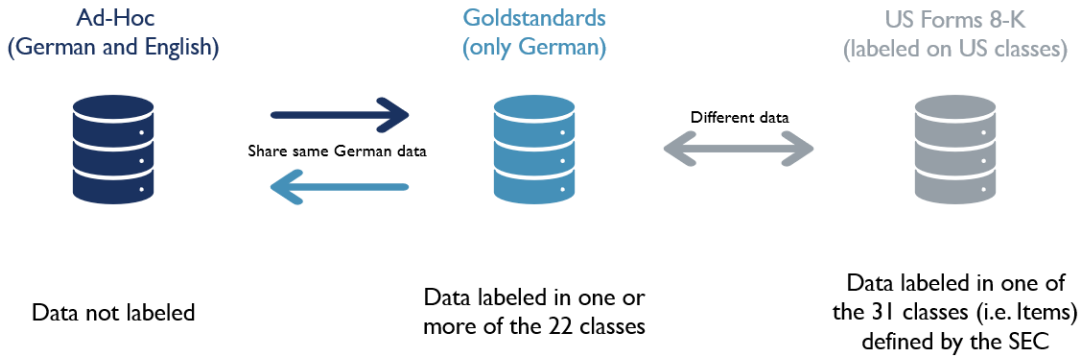

Fig. 4.1: Relation between the three different datasets.

we have the whole unlabeled Ad-Hocs as single instances, in the latter the individual instances are the single sentences or sentence pairs of some Ad-Hocs[1], each labeled separately. Each instance can be labeled with one or more classes and more details about these are given in the next section. The preprocessing of these two datasets will set up the data we are going to use to fine-tune our BERT model. The forms 8-K, on the other hand, does not share any data with the other two. The instances of this are the distinct sections of the 8-Ks, which are labeled in one of the 31 classes defined by the SEC. We refer to these classes as *items*. An overview of these can be found in Appendix C. The Ad-Hocs come from two sources, on the one hand from the DGAP[2], whose purpose is to provide a service for the disclosure obligations of listed companies in Germany, and on the other hand from the company register[3]. The forms 8-K are taken from the EDGAR database, directly provided by the SEC[4]. The labeling of the Ad-Hocs, hence the GS dataset, was performed by various researchers from the Chair of Finance and Banking of the LMU University of Munich.

The raw data contains many features. For our purposes we just need some of them. Figure 4.2 depicts the features we use for a generic instance of each dataset.

Origin	Goldstandard	Forms 8-K
- German Ad-Hoc document	- Single (German) sentence or pair of sentences	- Item text
- English Ad-Hoc document		- Item class
- Publication date and time	- Label or labels	- Unique hash of the form 8-K the item belongs to
- Company's name	- Unique hash of document the instance belongs to	
- Unique hash per document		

Fig. 4.2: Overview of the datasets and respective features used in this work

The unique hash of the Origin[5] and the GS are the same, since the instances of the GS are simply the (German) Ad-Hocs of the Origin split in sentences or sentence pairs. The GS contains 31.771 instances, the forms 8-K 418.596 instances, i.e. items, and the Origin contains a total of 127.395 instances, i.e. Ad-Hocs. Using the hash of the GS we first filter the Origin to get the German Ad-Hocs that have been labeled. We then use the publication date and time, and the company name to find the English Ad-Hoc counterpart in the Origin for each (labeled) German document. The reason for this, is that the vast majority of the German companies normally

[1]In order to avoid annoying repetitions from now on we will refer to an instance from this dataset, i.e. the GS, simply as a "goldstandard"

[2]Deutsche Gesellschaft für Ad-Hoc Publizität, https://www.eqs.com/ir-solutions/dgap/

[3]https://www.unternehmensregister.de/ureg/?submitaction=language&language=en

[4]https://www.sec.gov/edgar/search-and-access

[5]Of the German Ad-Hocs

publish an Ad-Hoc at the same time in both languages[6]. Note however, that a company listed in Germany is not obliged to publish a document in **both** languages, a German version is sufficient. That is why we do not always have an English version for a labeled German Ad-Hoc. Nevertheless, this process yields a parallel dataset containing 1.603 document pairs, i.e. two documents in two different languages that contain the same information and for which we have labels within the GS for the German documents. But since both documents contain the same content, we can transfer the classes to the English Ad-Hocs, which allows the training of an English model. In section 4.2 we explain how this transfer is performed. We call this dataset containing the document pairs as the *base* dataset. For the further processing we use the base and the filtered GS, with only the 17.575 goldstandards of the documents for which we have an English version.

4.1.1 Labels

In our set-up we have a total of 22 different classes. These are summed up in table 4.1. The classes were defined by the project partner and details about their

1. Earnings	7. Dividende	13. Insolvenzantrag	19. Real Invest
2. SEO	8. Restructuring	14. Insolvenzplan	20. Delisting
3. Management	9. Debt	15. Delay	21. Irrelevant
4. Guidance	10. Law	16. Split	22. Empty
5. Gewinnwarnung	11. Großauftrag	17. Pharma Good	
6. Beteiligung	12. Squeeze	18. Rückkauf	

Tab. 4.1: Classes used for the classification of the documents.

definitions can be find in Appendix B.2. Figure 4.3 shows the relative frequency of all the classes for the labeled goldstandards of the GS that belong to a German Ad-Hoc of the base dataset. This also implicitly applies to the English Ad-Hocs, since as we are going to see in the next section, the goal is to match each German goldstandard to an English sentence or sentence pairs of the equivalent English Ad-Hoc. As we can see the classes are quite imbalanced. In particular we have more than 40% of the data labeled with the class *Empty* and approximately 14% of the data labeled in the class *Earnings*, while each of the other classes covers something between 0.8% (*Real Invest*) and approximately 5% (*Guidance*) of the data. The high percentage of Empty cases is motivated by the fact that intuitively most of the sentences within a document do not contain any information that can be classified

[6]It can happen that the English version is published after the German one, but in our data we did not face this situation, it was therefore sufficient to use date, time and company name

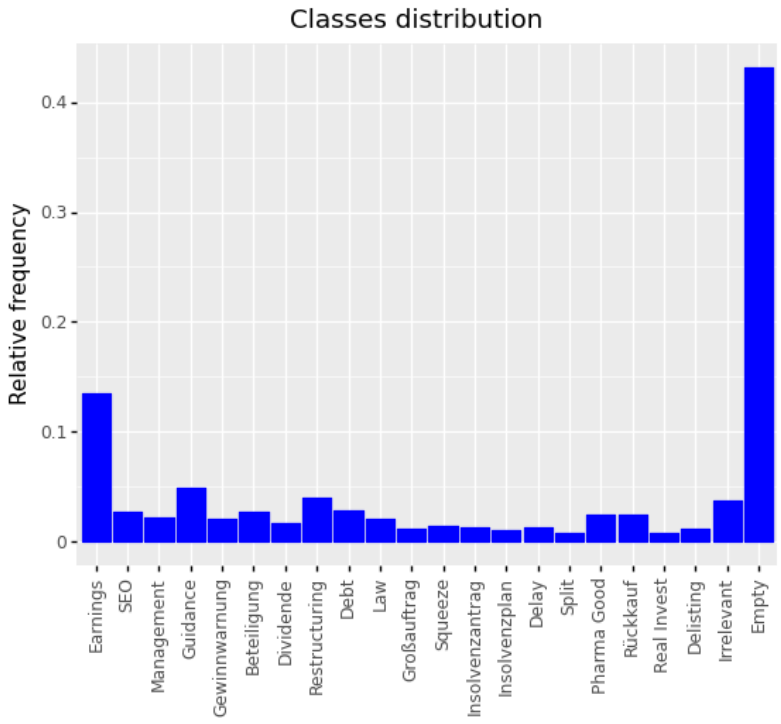

Fig. 4.3: Relative frequency of each label of the goldstandards for the base dataset

in one class. For example sentences like *"The position will be filled externally"* or *"We are now planning, in conjunction with Cytos, the next steps for this project"* are labeled as Empty, since they do not really carry any relevant information with them. Although this imbalance may seem problematic at first glance, we will see in chapter 5 that our model is still able to deliver good results. Moreover, we will also evaluate our model on the *document level* by aggregating the labels of all the goldstandards that belong to an Ad-Hoc together. Reasonably, we will no longer have the Empty class on this level, since if we consider the Ad-Hocs as whole documents instead of single sentences, then obviously each document will at least be classified in one class. This will further boost the performance of the model.

As stated in chapter 2 one of the major problems for us regards the inconsistency between the classes used for the classification task of our model and the US classes, i.e. items, provided by the SEC in which the forms 8-K are labeled. The two sets of classes differ not only in the total number of classes but also in their definitions. This makes it even more difficult to automatically assign one class to one or multiple items or vice versa. For this work, in accordance with the project partner, we decided to do the allocation between our defined classes and the forms 8-K items basing on the single definitions of both. Table 4.2 shows for which items we were able to find a reasonable class among all the 22 classes. Unfortunately, due to the divergences explained above, it was not possible to find a reasonable allocation for each of our

classes. Therefore, in order to have consistent and logically interpretable results, the classes for which we did not find an item will be omitted when evaluating the model on the TL task.

Class	Item (number)
Earnings	2.02
SEO	1.01, 6.03
Management	5.01, 5.02, 5.08
Guidance	2.02
Gewinnwarnung	2.02
Beteiligung	2.01
Dividende	7.01
Restructuring	2.05
Debt	1.01, 2.03
Insolvenzantrag	1.03
Insolvenzplan	1.03
Split	5.03
Rückkauf	8.01
Real Invest	2.01
Delisting	3.01

Tab. 4.2: Allocation of classes and items

4.2 Labels Transfer

As for now we have the base dataset with the document pairs and the filtered GS, where we can look up the labels of the German sentences for an Ad-Hoc. What we actually need, in order to fine-tune our model in English, is a dataset containing labeled English sentences. In order to create this we make use of the knowledge distillation method introduced in section 3.4, doing the following. For clarity reasons we explain the procedure only for a single Ad-Hoc, but of course this is applied to all the document pairs. As a first step we take all the available goldstandards from the GS dataset that belong to an Ad-Hoc. For each goldstandard we compute a sentence embedding with SBERT in the set-up explained in section 3.4. Using exactly the same model we also compute sentence embeddings for each single sentence of the English counterpart of this Ad-Hoc the goldstandards belong to. This process is best depicted in figure 4.4. Now we can compute the cosine similarity between the output embeddings of each goldstandard and every English sentence. Since some goldstandards are pairs of sentences we also compute the cosine similarity for each possible pair of consecutive English sentences. For example, given five sentences A, B, C, D and E, we also compute the sentence embeddings for the pairs AB, BC, CD, DE. But instead of computing the embeddings separately, we just

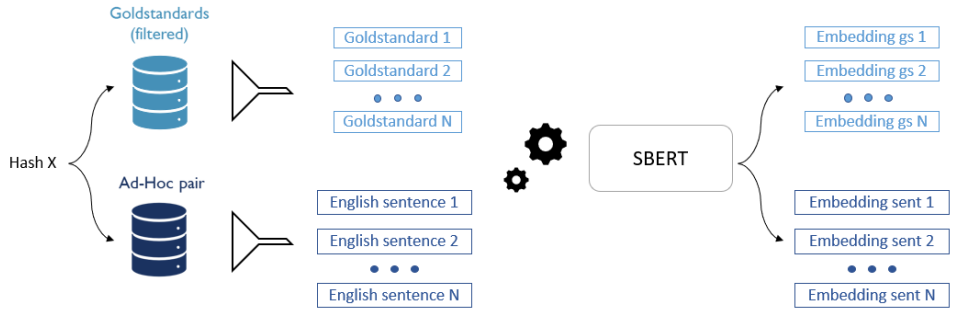

Fig. 4.4: Given an hash we filter the GS and the Ad-Hocs in order to find the English version of the German document the goldstandards belong to. We compute the embeddings for each goldstandard and each sentence of the English Ad-Hoc version.

add the already computed sentence embeddings of two sentences together in order to get the sentence embedding of that pair. The main reason why we do this is to reduce computation time, since a normal vector addition is simpler and faster to implement. Furthermore, since we do not modify the model by updating the weights, the sum of the embeddings of two sentences is a quite good approximation of the emebedding that would be computed by inputting the concatenation of these to the model. Equations 4.1 and 4.2 gives an overview how the cosine similarity for a single embedding for a goldstandard (Emb. gs 1) and all the single sentences (Emb. sent n) and pairs of the same consecutive sentences (Emb. sent n_1 + Emb. sent n_2) is computed using random numbers, with $n = 1, ..., N$ and N being the total number of sentences present in the English Ad-Hoc.

$$\begin{aligned} cossim(\textbf{Emb. gs 1}, &\textbf{Emb. sent 1}) = 0,2134 \\ cossim(\textbf{Emb. gs 1}, &\textbf{Emb. sent 2}) = 0,5678 \\ &\vdots \\ cossim(\textbf{Emb. gs 1}, &\textbf{Emb. sent N}) = 0,1118 \end{aligned} \tag{4.1}$$

$$\begin{aligned} cossim(\textbf{Emb. gs 1}, &\textbf{Emb. sent 1} + \textbf{Emb. sent 2}) = 0,8796 \\ cossim(\textbf{Emb. gs 1}, &\textbf{Emb. sent 2} + \textbf{Emb. sent 3}) = 0,6093 \\ &\vdots \end{aligned} \tag{4.2}$$

Once this is done we select the English sentence or pair of sentences with the highest cosine similarity and label this according to the labels of the goldstandard we are analysing. For example, if the goldstandard is labeled with the classes *Guidance* and *Gewinnwarnung*, we then label the English sentence or pair of sentences with the highest cosine score with these classes too. This yields a dataset of the same dimension of the filtered GS (i.e. 17.757) with labeled English text. So we just

performed our *labels transfer*. It is important to note, that we do not fine-tune our SBERT. We simply use the pre-trained model by Reimers and Gurevych (2020) and compute the embeddings with it. Fine-tuning would also not be possible, as we have no data that we can use to optimise the model, i.e. data containing the most semantic similar English text for a given German goldstandard. Moreover, this is another reason why we chose to use SBERT instead of other models like LASER introduced by Artetxe and Schwenk (2018). In fact this latter model works better when the task is to find a perfect translation across two languages, as stated in (Reimers and Gurevych, 2020, Tab. 3). However, the same authors also show that SBERT performs definitely better when it comes to finding the most semantic similar sentences across different languages. We decide to use SBERT, since we are interested in parts of the English Ad-Hocs that contain the same information of the goldstandards, rather than exact translations.

4.3 Transfer Evaluation

In order to evaluate the output of the transfer labeling process we select 15 random pairs of documents from the data. For each pair we take the goldstandards for the German Ad-Hoc and manually assign to these the sentence or pair of sentences of the corresponding English Ad-Hoc, which contain the most semantic similar information. The assignments of this manual procedure for each goldstandard of the selected documents we use for the evaluation are reported in Appendix D. We then compare our "choice" to the output of SBERT and evaluate the results using a confusion matrix. Confusion matrices are typically used in binary classification problems. This is not exactly our case, since the output of SBERT is some kind of text[7] and not a binary variable. Therefore, we define the different entries of the confusion matrix the following way:

1. **True Positive**: If the output of SBERT equals the manual assignment, then this is a **TP** case and a 1 is assigned to both (SBERT's output and manual assignment)

2. **False Positive**: If the output of SBERT contains the manual assignment's entry and some more text/information, then this is a **FP** case. The SBERT's output is then assigned a 1 and the manual assignment a 0

3. **False Negative**: If the output of SBERT contains just a part of the manual assignment's entry, then this is a **FN** case. The SBERT's output is then assigned a 0 and the manual assignment a 1

[7] Well, the embeddings of this text

4. **True Negative**: If the output of SBERT and the manual assignment are completely different, then this is a **TN** case and a 0 is assigned to both (SBERT's output and manual assignment)

The idea behind these definitions is the following. For the TP case the intuition is straightforward. If both entries are the same SBERT is able to correctly "classify" them, i.e. to find the English sentences that contain the same semantic information. Trivially we can use the same argument the other way around (two entries that contain completely different information) to explain the TN case's definition. For the FP and the FN cases we adopt a different way of thinking. Even if we do not have an exact result, in the case of a FP SBERT is still able to find the same information of a goldstandard somewhere in the English Ad-hoc, even though in combination with a superfluous sentence. This means that the SBERT's output still contains all the necessary information present in the German goldstandard. Implying that we can assume, that the labels for that goldstandard applies also to the English output. This assumption, on the contrary, does not apply in the case of a FN. In this case, we face the opposite problem as for the FP, meaning that we cannot ensure, that

Case	Manual assignment	SBERT's output
True Positive	The proceeds from the transaction will be used to further increase the capital structure and for general corporate purposes.	The proceeds from the transaction will be used to further increase the capital structure and for general corporate purposes.
False Positive	The size of the order is significantly over 1 million Euro.	The size of the order is significantly over 1 million Euro. The FPD Glass Business Unit is one of the important strategic sectors for the ISRA Group's expansion.
False Negative	Super Airbus to take off with CENIT EADS Airbus GmbH places major order worth DM 3m with CENIT. In the development of the new A380 jumbo jet and the military transporter A400M, EADS Airbus GmbH is counting on co- operation with CENIT AG Systemhaus, Stuttgart. Stuttgart, February 6th 2001	Super Airbus to take off with CENIT EADS Airbus GmbH places major order worth DM 3m with CENIT Stuttgart, February 6th 2001
True Negative	Information missing	On July 19, 2001, MorphoSys and GPC Biotech announced in vivo efficacy of an antibody in their cancer collaboration.

Tab. 4.3: Example for the entries' definitions of the confusion matrix

the output of SBERT actually contains all the necessary information, in order to belong to the goldstandards' labels. Just because part of the information contained in the goldstandard is missing in the model's result. And we cannot guarantee that the labels refer to the non-missing part. That is why we tag it as (false) negative. Table 4.3 shows an example for all the four cases. According to our definitions, we get the confusion matrix of table 4.4. And basing on this we can compute some performance metrics like *precision* and *recall* defined as

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \tag{4.3}$$

and the *F1-score* defined as

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{4.4}$$

The precision tells us how much of all the data labeled in a class, do actually belong to that class. Recall instead, tells us how many of the data that belong to a given class, were also predicted in that class. While the F1-score can be interpreted as an arithmetic mean of the two metrics. Precision can be therefore seen as a measure of quality and recall as a measure of quantity. On the other hand it would not be very meaningful to use the accuracy as a performance metric here. This is because this latter also considers the TNs as correctly classified instances, which is in contrast to our definitions, since we asses them as incorrect. So, in our case we get a precision of 88,49% a recall of 99,19% and an F1 of 93,54%, which can be considered as very good results. In particular we can interpret the lower precision in comparison to the higher recall as a "safety" measure by the model. We can see it as SBERT being unsure whether a goldstandard has the same semantic meaning of an English sentence or not and therefore prefers to "add" some more information to it. But this uncertainty at the same time ensures that at least all the necessary information for an instance to be classified in a class is also present within the English text, which is better than omitting (maybe) important information as in the FN case. As a result, we have more false positives than false negatives, hence an higher recall than the precision. These satisfactory results therefore provide us with a dataset containing labeled English data with which we can fine-tune our model.

		SBERT's output	
		0	1
Manual assignment	0	4	16
	1	1	123

Tab. 4.4: Confusion matrix for the evaluation of SBERT's output

4.4 Forms 8-K

A single form 8-K document can belong to multiple items. In fact, a document normally consists of several distinct sections containing different types of information about a company and each of these sections must be *labeled* in a specific item. For that reason the single forms can be easily split according to the different items they contain and we use this property to set up the forms 8-K dataset. Thus, each section of a form 8-K represents a unique instance of the dataset, which is labeled according to its (single) item. So if we stay on an item level we actually have a multi-class problem, whereas on a document level we keep the multi-label property. Indeed, what we are really interested in, is the classification of entire documents, rather than the single sections of them. The problem, however, is that many documents are quite long. But BERT accepts tokenized inputs of up to 512 tokens in length[8]. A single tokenized form 8-K almost certainly exceeds this length. And the same

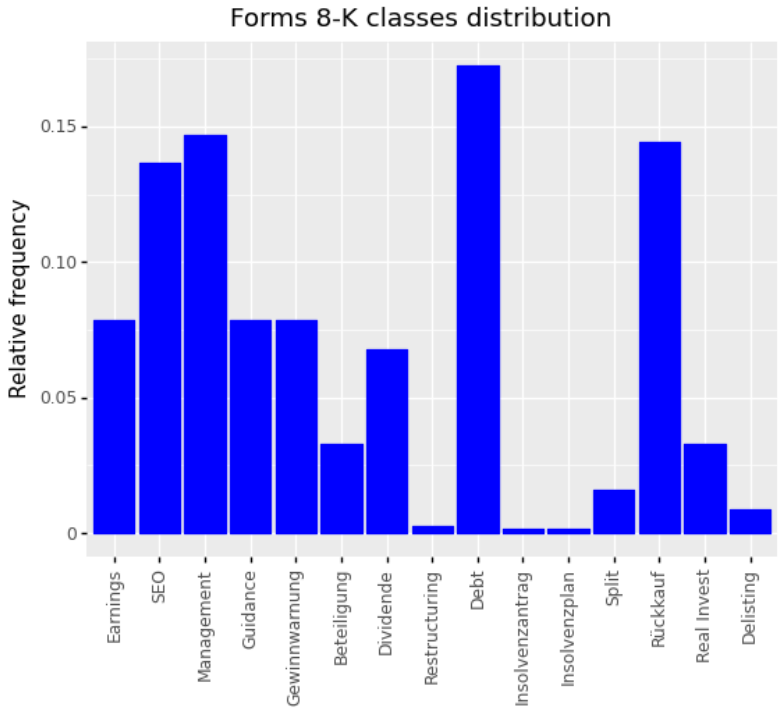

Fig. 4.5: Relative frequency of each linked class of the items

holds also if we take the single sections, i.e. items, of a document. Therefore, we additionally need to split up each item in single sentences and each of these is then labeled as belonging to that item. This procedure automatically introduces some *bias* in the data, since we know that not all the sentences within an item carry information about that class. We therefore expect, and as we will see in the next chapter this is also the case, to have low recall scores on the TL task. In this work we

[8]In theory, the input length can be extended up to 2.048 tokens, but since our training data rarely exceeds 512 tokens, we decided to set the input to this length

first labeled the items of the forms 8-K with our classes according to the allocation presented in table 4.2. The labeling process to "our" classes, although not exactly precise, is necessary, otherwise we could not evaluate the model on the TL task. We then removed from the data all those instances of the items for which we could not find a reasonable link to our classes. This yields a dataset containing 364.354 items belonging to 305.888 distinct forms 8-K. And each of these items is then further split into sentences. The final dataset we will use for the TL task contains 3.262.695 labeled instances, i.e. single sentences labeled with the class corresponding to the item they belong to according to table 4.2. The classes distribution of this dataset according to the single labels is depicted in figure 4.5. Trivially, classes that have been linked to the same item, also have the same frequency, since we labeled the entire item (thus, each sentence of it) to belong to more classes. For example this is the case with the classes *Earnings*, *Guidance* and *Management*, that we all linked to the item 2.02. As for the test dataset of the Ad-Hocs, we are also going to evaluate the results on sentence and document level. In addition, the data belonging to the items 7.01 and 8.01[9], which we respectively linked to the classes *Dividende* and *Rückkauf*, are evaluated separately. Reason for this is that, conforming to the definition of the items by the SEC, these classes act as a melting pot for the forms 8-K. This means that all those documents, which contain information that cannot be classified in one of the other items, find their allocation (hence, are classified) in one of these two items. But in principle both items can contain information, which can belong to any of our classes. These two classes alone account for approximately 30% of the instances in our dataset (1.029.175 instances). Even if the allocation of these items to our two classes is not really exact, this is necessary for the evaluation.

[9]From now on simply referred to as items 7 and 8

Chapter 5

Fine-Tuning and Results

In this chapter we will evaluate our fine-tuned BERT model. We will evaluate both the results of the fine-tuning process and the results of the TL task on the forms 8-K. We will also compare the results of the fine-tuning process with the results obtained by the project partner on fine-tuning the German version of the same model[1]. All the results obtained are rounded to 4 digits and then multiplied by 100, in order to have the same scale for each metric in the bounded interval [0,100]. Each value must be intended as a percentage, although for overview reasons we omit the percentage sign. Trivially, the higher the metric, the better the result. Moreover, we will evaluate the model on two different levels, the *sentence* and the *document* level. In both cases we input the same data, i.e. the labeled goldstandards of the English Ad-Hocs or the labeled forms 8-K items' sentences, and get the same outputs, i.e. the models' classification for the input instances. Evaluating these outputs is equal to evaluating the results on the sentence level, since these are indeed the classification for the single sentences. Parallel to this we also group the sentences (and their labels) belonging to a single document together and compare all the predicted labels for a document with the true labels of the same document. In addition to this, we also make a distinction between a *local* level and a *global* level. In the former we evaluate the results and compute the used metrics for each single class, while in the latter this is done globally over all classes. To compute this, we simply average the metrics' outcomes of the single classes using *macro* averaging, meaning that we treat each class equally, hence giving them the same importance[2]. It goes without saying, that whenever we present results on a global level we use macro averaging.

5.1 Fine-Tuning

In this thesis we use the pre-trained version of BERT[3] (cased) and fine-tune it, with regard to our multi-label downstream task. We already presented the architecture characteristics of our model in chapter 3. For the fine-tuning process, and the evaluation of this, we use the English labeled Ad-Hocs data, which are pre-processed as explained in chapter 4. For the tokenization of the data, we use a WP tokenizer and the final tokenized inputs are computed as explained in section 3.3. We will use batches of the data for each optimisation step. But the model only accepts equally long input data, and since the single English goldstandards have different lengths, we need to bring them to the same size. This process is also known as *padding*. Basically, we append a sequence of zeros to each tokenized input, in order to have inputs of the same length. To speed up the processing operations we decide to use a dynamic padding instead of a fix padding. Given a batch we look for the longest tokenized sequence i in it and pad all the other sequences present in that batch with zeros to be of the same length of i. This procedure halves the computation time of the fine-tuning process. Furthermore, we set the maximum length of an input sequence to 512 tokens. Longer sequences are truncated, i.e. cut off, after the 512th position (in our case this happens very rarely). We also divide the data into

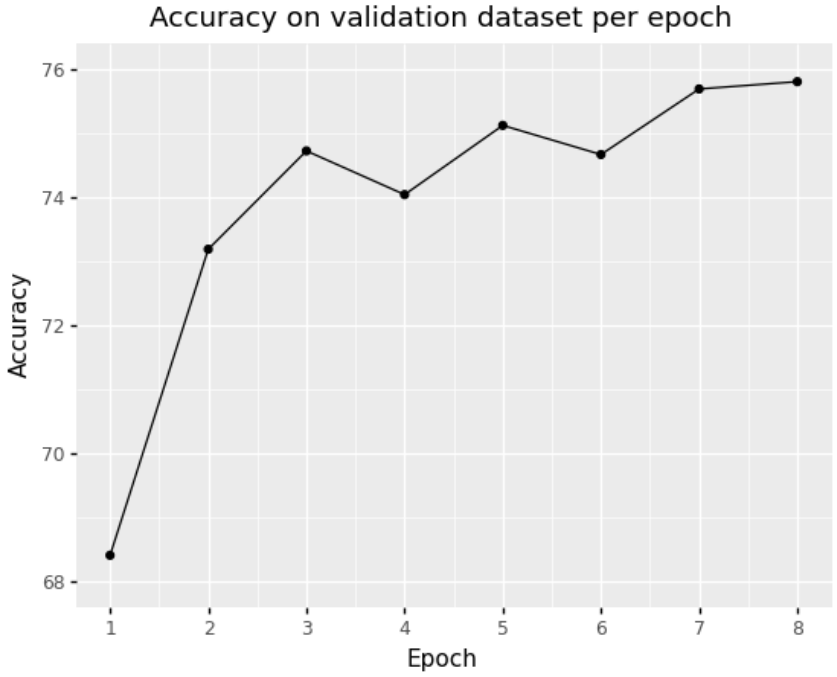

Fig. 5.1: Accuracy of validation set per epoch.

[1]The model trained on the German goldstandards. We will refer to this model as **GerBERT**

[2]This is also the reason why in some cases we get an F1-score, which is not between precision and recall

[3]https://github.com/huggingface/

a *train*, a *test* and a *validation*[4] set using stratified sampling, i.e. dividing the data in such a manner that we have more or less the same percentage of data belonging to each class in each set. For this we decided to do an 80/10/10 split of the English goldstandards, yielding respectively 14.060, 1.758 and 1.757 instances for the train, test and dev set[5]. For fine-tuning we decided to use the following hyperparameters: a learning rate of 4e-5, a training and evaluation batch size of 16, a dropout of 0,5 for the last layer and of 0,1 for each encoder. Dropout is a regularisation technique in NNs for preventing overfitting of the model. This method "drops" randomly selected neurons during training with probability equal to the selcted dropout rate, i.e. in our case with 50% and 10%, so that a reduced network is left. The weights associated with that neurons are then not updated during the backpropagation step. We trained our model for 8 epochs using the AdamW (Loshchilov and Hutter, 2019) optimiser algorithm. For hardware reasons we could not increase the batch size or the number of epochs. Fine-tuning the model on a Tesla V100 PCIe 16GB GPU took approximately 18 minutes.

For the training process we use the accuracy[6] as the performance measure and evaluate the model on the dev set after each epoch. This is best depicted in figure 5.1. As expected, we have a significant increase in accuracy within the first epochs with a gain of around 6 points in performance in the first three epochs. After the third epoch the accuracy stabilises in the interval between 74% and 76%, with a slight upward trend towards the end. So the model learns the most in the early stages of the fine-tuning process and less in the final stages. This is further confirmed when looking at the loss of the train and dev set per epoch, as shown in figure 5.2. The loss for the train set is continuous monotonously falling. In this case too, we have the steepest descend, hence the greatest gain in performance, in the first three to four epochs. But overall the trend of the train loss and the one of the accuracy indicate a steady improvement of the model. This would speak in favour of using the fine-tuned model after eight epochs. However, if we take a closer look at the validation loss in figure 5.2, then it is evident that after an initial downward phase it starts to increase after the third epoch. Between the third and fourth epoch, a slight increase can be seen, which becomes constant until the end of the training process. This means the model begins to **overfit** from that point on. It is therefore not reasonable to use the eight epochs trained model for the TL task. As a consequence, basing on the results of both losses and the accuracy of the validation set, we decide to use the model fine-tuned **until the third epoch** for the further analysis. All the results presented in this chapter are computed using this model.

[4]Also called *dev* set

[5]The fact that the dev set has one instance less than the test set is motivated by the odd total number of instances

[6]The number of correctly classified instances divided by the total number of instances

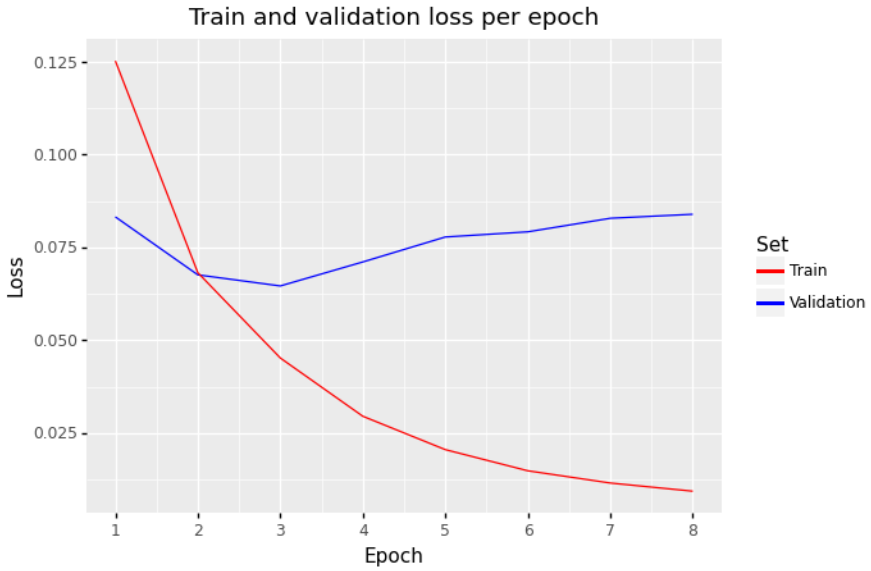

Fig. 5.2: Training set and validation set loss per epoch.

5.2 Test Data and Threshold Decision

We now have a fine-tuned model that we can evaluate and use for the TL task. But it still remains to understand in which case an instance should be classified in a class. Since we face a multi-label problem and use a sigmoid function for the classification in the last layer of the model, we need to set a threshold above which an instance will be classified in a class, i.e. above which output's score should the instance be labeled with a class. In order to pick the best threshold we compute the global performance metrics on both sentence and document level for different thresholds between 0 and 1 in 0,05 steps for the test set. The results of these are reported in table 5.1. Based on these results, and since precision and recall are equally important in our work, 0,45 seems to be the best trade-off for the choice of our threshold. Even though for other thresholds we have single performance metrics that are slightly higher than the ones for 0,45, the differences between precision and recall are lower for this threshold on both levels. Thus, we set our classification threshold to 0,45 and even if not mentioned all the following results are computed using this threshold.

With a fixed threshold, we can now evaluate our model on the test set. From table 5.1 we can read out the global performance metrics (highlighted for our picked threshold), while table 5.2 reports the local performance metrics of the test set. The global and local metrics show a big increment in performance when dealing with entire documents instead of single sentences. In this case we have a global increment of 13 to 14 points for all three metrics. Here we have an F1-score of 83,23%, a precision of 84,45% and a recall of 84,01%, which can be seen as satisfying results. Moreover, the local documents metrics always outperform the sentences ones or are at least

Threshold	Sentence			Document		
	F1-score	Precision	Recall	F1-score	Precision	Recall
0,05	62,92	51,88	86,05	73,13	63,47	91,20
0,10	67,51	59,31	82,24	78,50	71,78	89,50
0,15	69,16	62,91	79,64	80,17	75,55	87,90
0,20	70,00	65,45	77,35	81,66	78,82	86,94
0,25	70,40	67,11	75,76	82,19	80,25	86,20
0,30	70,46	68,29	74,39	82,76	81,23	86,18
0,35	69,88	68,76	72,67	82,18	81,15	85,20
0,40	69,28	69,33	70,97	82,17	82,08	84,35
0,45	69,71	71,27	69,82	83,23	84,45	84,01
0,50	69,40	73,10	67,41	83,35	85,56	82,63
0,55	67,09	74,38	63,76	80,86	88,75	78,76
0,60	65,08	73,17	60,34	80,31	85,03	78,08
0,65	61,35	72,97	54,99	78,74	85,02	76,07
0,70	58,40	76,04	50,88	78,68	86,79	74,53
0,75	53,14	70,37	45,73	71,28	80,85	67,29
0,80	47,78	68,71	39,96	65,93	76,12	62,63
0,85	39,45	61,94	30,56	63,66	68,14	60,26
0,90	22,50	53,66	16,24	44,54	55,21	39,99
0,95	4,23	21,33	2,66	15,54	18,61	13,83

Tab. 5.1: Global performance metrics for different thresholds on test set

equal to these. As we are mainly interested in the classification of entire documents, we can state that our model is able to accomplish this task quite good. These outcomes are even more interesting if we benchmark our model with GerBERT. The global performance metrics for this latter are reported in table 5.3. Surprisingly our model outperforms GerBERT on document level by about 4 percentage points on the F1-score for example, while clearly underperforming on sentence level. These results add more value to our model, if we take into consideration that GerBERT was fine-tuned on about twice the data (26.793 instances). One possible explanation for this behaviour could be represented by the number of training epochs used for fine-tuning the model. In contrast to our model, GerBERT was fine-tuned for 10 epochs. And this version of the model was also used for testing, while we decided to use our 3-epochs old version. We saw how our model started to overfit after three epochs, and in general this is a usual behaviour in NLP, with many fine-tuned models starting to overfit after few epochs. Therefore, we suppose that GerBERT fine-tuned for 10 epochs has a similar behaviour and overfit on the train data, hence performing worst on the test set. In addition, GerBERT implemented a higher dropout rate of 0,35, while we decided to use a lower dropout, due to the less amount of available training data. This might be another point that could be investigated, although it certainly has a less impact than using an overfitted model. For GerBERT the

Label	Sentence			Document		
	F1-score	Precision	Recall	F1-score	Precision	Recall
Earnings	80,33	81,67	79,03	93,38	95,48	91,36
SEO	70,37	66,67	74,51	88,57	81,58	96,88
Management	73,17	69,77	76,92	93,55	93,55	93,55
Guidance	72,28	68,87	76,04	83,44	80,95	86,08
Gewinnwarnung	37,74	47,62	31,25	45,45	55,56	38,46
Beteiligung	59,79	59,18	60,42	81,16	77,78	84,85
Dividende	81,97	86,21	78,12	89,36	91,30	87,50
Restructuring	57,89	58,67	57,14	75,47	80,00	71,43
Debt	76,92	78,43	75,47	88,10	88,10	88,10
Law	77,33	78,38	76,32	91,23	96,30	86,67
Großauftrag	66,67	76,47	59,09	81,48	91,67	73,33
Squeeze	91,53	87,10	96,43	94,55	92,86	96,30
Insolvenzantrag	65,52	55,88	79,17	72,34	58,62	94,44
Insolvenzplan	50,00	41,38	63,16	57,89	44,00	84,62
Delay	77,55	73,08	82,61	94,44	94,44	94,44
Split	64,29	90,00	50,00	78,26	100	64,29
Pharma Good	71,26	73,81	68,89	93,62	95,65	91,67
Rückkauf	79,55	87,50	72,92	98,36	100	96,77
Real Invest	57,14	53,33	61,54	80,00	88,89	72,73
Delisting	84,21	100	72,73	92,86	100	86,67
Irrelevant	62,89	56,18	71,43	74,34	66,67	84,00
Empty	75,23	77,66	72,95	–	–	–

Tab. 5.2: Local evaluation metrics for the test set on both levels. The class *Empty* is not present on document level, as explained in section 4.1.1

threshold was set to 0,8, because it yielded the best results. This could also be motivated by the larger amount of data, which increase the model's confidence in classifying the data, and thus is able to predict an instance's classes with an higher accuracy. Nevertheless, we strong believe that our model could perform even better, and outperform GerBERT also on the sentence level, if it were trained with more data.

	F1-score	Precision	Recall
Sentence	73,85	78,19	70,87
Document	79,65	78,03	81,89

Tab. 5.3: GerBERT global performances (macro averaged) on test set on the data used to fine-tune it.

5.3 Transfer learning on forms 8-K

For the TL task on the forms 8-K we already pre-processed the data as explained in section 4.4 and in general in chapter 4. In order to evaluate the results we needed to classify the data in *our* classes and this is done as shown in table 4.2 for all those classes for which we could find a plausible match to the US items. We also explained the problem related to items 7 and 8, and we are going to evaluate the results for these two classes separately in the next sub-section. The local results for the other forms 8-K are reported in table 5.4, from which we derive the global metrics in table 5.5. In both cases of course, we use 0,45 as the classification threshold. It is clear

Label	Sentence			Document		
	F1-Score	Precision	Recall	F1-score	Precision	Recall
Earnings	12,01	87,21	6,45	43,38	93,24	28,27
SEO	0,92	35,42	0,46	10,91	44,49	6,22
Management	13,38	97,09	7,19	68,18	86,80	56,14
Guidance	1,37	73,89	0,69	4,26	67,89	2,20
Gewinnwarnung	0,15	63,03	0,07	0,60	57,91	0,30
Beteiligung	11,86	21,20	8,23	32,24	21,86	61,39
Restructuring	25,65	21,28	32,27	13,81	7,43	97,65
Debt	37,80	96,28	23,52	66,80	91,13	52,73
Insolvenzantrag	5,25	46,70	2,78	31,93	31,49	32,38
Insolvenzplan	2,78	51,42	1,43	21,39	31,03	16,32
Split	2,77	81,87	1,41	12,05	61,16	6,69
Real Invest	1,16	27,18	0,60	9,83	14,58	7,42
Delisting	34,00	98,33	20,55	81,51	72,12	93,72

Tab. 5.4: Local performance metrics for 8-K predictions on both level

that, based on these results, we can say that our model has a poor performance, yielding a global F1-score of 11,47% on sentence and of 30,53% on document level, which are much lower scores compared to the results on the test set. Despite showing good results in isolated cases. It can be seen that for example we get a recall of 93,72% for the *Delisting* class on document level, which even outperforms the same metric on the test set. Indicating the model correctly identifies the majority of documents we labeled in that class. In general, these results are in line with those of the test set, with higher performance metrics for classes showing better scores in the test set. Also in this case, we can state that the model performs better on document than on sentence level, with a difference of about 19 percentage points on the F1-score between the two. Interestingly, the global precision on sentence level is better than the one on document level by about 9 points but at the expense of a significantly lower recall. This trend can be best depicted on the local level, where especially on sentence level we have many classes showing a big discrepancy between

the precision and the very low recall, as in the classes *Gewinnwarnung* and *Guidance* for example. This means that our model returns only very few instances belonging to that class (quantified by the recall), but most of them correctly predicted (quantified by the precision). In fact these results, and particularly the low recall scores, are comprehensible. As explained in section 4.4 we labeled the entire items, hence all the single sentences in them, with our classes that we have linked to them. Intuitively, as for the Ad-Hocs, we also assume that only few sentences of the documents contain the relevant information, while the majority could be classified as *Empty*. For example we label all the 10 sentences of an item according the linked class to that item, even though only one or two sentences contain relevant information about that class, while the rest is theoretically *Empty*. So, we assume in advance that we have this bias in our data. Considering this aspect, we can state that in many classes on sentence level our model clearly distinguish a minority of sentences (recall score), which it is also able to correctly classify (precision score). Of course, basing on these metrics we cannot say anything about the fact, whether these few sentences are really those that carry the relevant information with them or not. But on the other hand, we can say that the link between the US items and our classes is partially correct, at least for all those classes with an high precision score, as in the *Management* or *Debt* case. Our interpretation is further supported by the results on the document level, where we gain in recall in each class, but at the expense of lower precision scores (except for three classes). Classes having a low precision on document level should probably be further investigated, whether the single items of the forms 8-K that we labeled in those classes should also be labeled in other classes or not.

	F1-score	Precision	Recall
Sentence	11,47	61,61	8,13
Document	30,53	52,39	35,49

Tab. 5.5: Global performance metrics for 8-K predictions on both level

5.3.1 TL on items 7 and 8

Due to the reasons explained at the end of chapter 4, we evaluate the documents belonging to the items 7 and 8 separately. Table 5.6 reports the local performance metrics for the two classes. Remember, that the items 7 and 8 are not really equivalent to the classes *Dividende* and *Rückkauf* respectively, since both contain documents with information that can theoretically belong to any class. But since we needed reference classes for the evaluation, we linked them to these two. Despite that, we get similar performances as for the other classes, with very low recall on both levels for both classes, but with an high precision for the items 8, while having also a

quite low precision for the items 7. Once again, at least for the class Rückkauf, our model seems to be very picky, retrieving only few documents or sentences it thinks belong to this class. Nonetheless, the good precision scores indicates that our model considers those few retrieved data as actually belonging to this class. This is a quite interesting result, since we assumed these classes to contain documents, which did not belong to a particular class. In fact on the contrary, the results of the Dividende

Label	Sentence			Document		
	F1-Score	Precision	Recall	F1-score	Precision	Recall
Dividende	2,99	21,31	1,61	11,49	27,59	7,26
Rückkauf	2,28	88,50	1,16	8,24	79,03	4,34

Tab. 5.6: Local performance for items' 7 and 8 predictions on both level

class seem to support our assumptions. The lower precision on both levels suggests that our model considers only a small fraction of the items 7 as belonging to this class. In this regard, figure 5.3 gives an idea of the classes, according to which our model, these items contain information. The image depicts a bar plot for each class, with the relative frequency of instances predicted in each class on document level, i.e. for each form 8-K, which contains either the item 7 or 8, or both. Note, that we omit the *Irrelevant* class. The reason is that we are actually interested in finding out, which of our topics do the forms 8-K of these two classes cover. And the Irrelevant class can be seen as a class, which includes all those documents that contain information unrelated to our classes, in which we are not interested. Conversely the Empty class points out to all those forms 8-K, which in general do not contain any relevant information. That said, we can see that more than 50% of both items

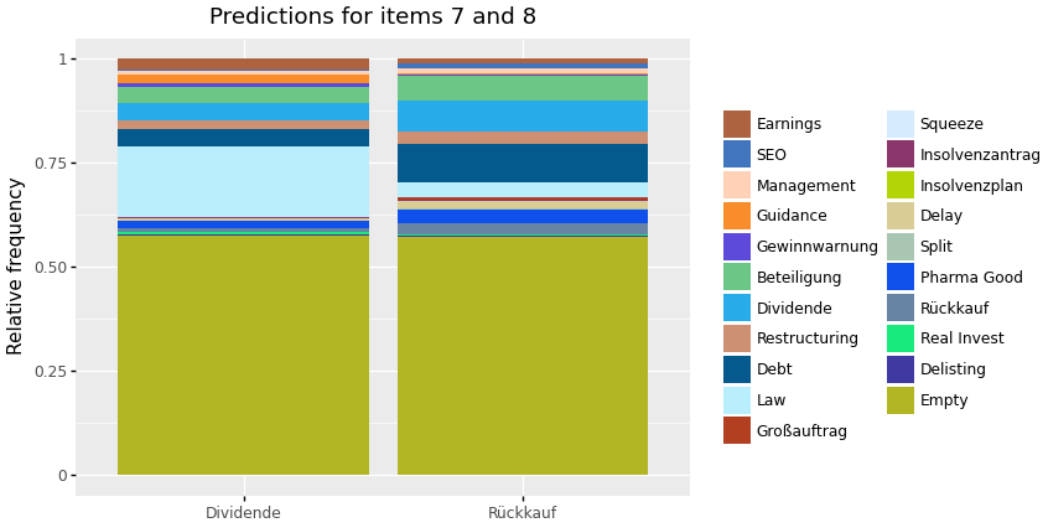

Fig. 5.3: Relative frequency of predicted classes for items 7 (Dividende) and 8 (Rückkauf) on document, i.e. forms 8-K, level

are indeed classified as Empty by our model. This supports the fact that the two items are mainly used as melting pot for various companies' disclosures. However, one can note how about 16% of the items 7 are predicted in the class *Law*. We can also notice that the classes *Beteiligung*, *Debt* and the actual class *Dividende*, are predicted more or less all three the same amount of times. The predictions in these four classes together account for about 28% of the data. While according to our model the items 7 contain information, which would classify them in the other classes, only to a minor extent. Similarly, we can note that about 23% of the items 8 are predicted in the classes *Dividende*, *Beteiligung* and *Debt*, while only about 2,5% of the documents are predicted in the actual class *Rückkauf*. Hence, without taking into account the instances classified as *Empty*, we can see how actually the items of both classes contain information that belong to many of our classes. Nevertheless, according to our model, most of the content of both items seems to mainly belong to three to four precise classes for each item.

Chapter 6

Discussion and Outlook

Very surprisingly our model outperforms GerBERT by about 4 percentage points on the F1-score on document level. Since we want to classify entirely documents, we can just focus on this level and omit the sentence one, which performs worse in our model. Despite good results on the test set, our model did not perform as well with respect to the TL task. The evaluation on this task, performed using the classes we linked to the single items, shows in many cases a low performance. In particular we often have low recall values. This was predictable, for the reasons explained in the chapters 4 and 5 concerning the labeling process of the items. But overall we can be pleased with the high precision scores for some classes. This suggests that our model is still able to recognise relevant information within the items that belong to one or more classes. Nonetheless, a more reliable and accurate evaluation of the TL task can be obtained by labeling the single items of the forms 8-K in a more precise manner. In this regard, manually classifying the single items or just a part of them in our classes is the safest way, in order to have a labeled data to use for the evaluation. However, this is possible with a solid economic background, which is beyond our reach. In a more *statistical* way this could be approached as a *self-learning* problem. The experiments carried out by Dong and de Melo (2019), which similar to us face a text classification problem in a multilingual context, show that a self-learning approach can lead to very good results. The basic idea is to let the model find patterns within the data to classify them and then use only those instances for which the model has an higher confidence level to fine-tune it further. With this approach it would be not necessary to manually find an allocation between the items and our classes. As a downside, we would have no control over the labeling process, thus we could not be sure whether the self-learning approach correctly classified the instances or not.

Undoubtedly, our model and our results can be additionally improved. Trivially of course, as already mentioned in chapter 5, we strong believe that simply increasing the training data already leads to considerable improvements. But we also realise

that labeled data is difficult to find and often expensive. However, as shown by the satisfying results of the transfer process in section 4.3, it seems sufficient to label German Ad-Hocs and then process the data as we did in order to get labeled English text. It does not seem necessary to label the English data directly. Even though the output of the labels transfer process is already good, these might be further improved if we also compute the embeddings for each pair of sentences using SBERT, instead of simply adding the two output embeddings of the two single sentences. Moreover, the output dataset of this process contains German and English data with the same content. This can be used to eventually fine-tune a version of SBERT and use this model on new data in future. There would be the advantage here of using a fine-tuned model on purely economic text data and therefore even more accurate if used on new domain-specific data.

Due to lack of time we only fine-tuned a BERT model. We suggest to try to fine-tune different LMs on the same data and compare the results to ours. It might be interesting to see to what extent do the results of a fine-tuned version of BERT-large or of models with a slightly different architecture, such as ALBERT or DistilBERT, differ from ours. With regard to our model instead, we suggest to further analyse the classification process on the `[CLS]`-token. We found the results on the test set acceptable, especially on document level and therefore, we did not experimented with different classification methods. But as for example the experiments on different tasks using BERT by Choi et al. (2021) show, taking the averaged pooled token embeddings instead of the `[CLS]`-token output can lead to improved results, since it provides a better representation of the input. Besides averaging all the output embeddings, which is equal to *mean-pooling*, also other techniques, like applying a *max-pooling* on the embeddings can be taken into consideration.

Chapter 7

Conclusion

The main goal of this thesis was to implement an NLP model, which could classify economic documents of American public companies in self-defined classes. We faced this problem as a transfer learning process, since we fine-tuned a BERT model using economical text documents from German public companies. In order to do this, we had to label the English version of the Ad-Hocs and did this using a multilingual siamese neural network. This network made use of the knowledge distillation approach proposed by Reimers and Gurevych (2020). As we saw in chapter 4 the process showed great results. This step provided us a stable dataset that we could use to fine-tune a BERT model, originally proposed by Devlin et al. (2018), on our multi-label classification problem. We evaluated our model on two different levels and were able to outperform on the document level the German version of the same model, which we took as a benchmark model. This is a great result considering that our model was fine-tuned with less data. As a final step in chapter 5 we performed the TL task on the forms 8-K and used our model to classify these in our classes. Despite a general global low performance on the TL task, in chapter 5 we were able to retrieve some good and interesting results on the local level for some of the classes. In any case, especially on the global level, these results are probably mainly due to the different definitions between the American classes, i.e. items, and our self-defined classes, and to the allocation we made between them, rather than a powerless model. In conclusion, we can say that we have succeeded to set up a classifier, in form of an NLP model, able to classify the forms 8-K in our classes. Nevertheless, despite having some good results on a local level, achieving a better or complete accuracy involves major challenges that must be overcome.

Bibliography

Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T., Klein, J., and Goujon, A. (2021). A Comparison of Pre-Trained Language Models for Multi-Class Text Classification in the Financial Domain. In *Companion Proceedings of the Web Conference 2021*. Association for Computing Machinery.

Artetxe, M. and Schwenk, H. (2018). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *CoRR*, abs/1812.10464.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*.

Baldi, P. and Chauvin, Y. (1993). Neural Networks for Fingerprint Recognition. *Neural Computation*, 5:402–418.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155.

Bottou, L. (1998). Online Learning and Stochastic Approximations. In *On-line Learning in Neural Networks*, pages 9–43. Cambridge University Press.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature Verification using a "Siamese" Time Delay Neural Network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7.

Cavnar, W. and Trenkle, J. (1994). N-Gram-Based Text Categorization. *In Proc. SDAIR*, pages 161–175.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., and Androutsopoulos, I. (2019). Large-Scale Multi-Label Text Classification on EU Legislation. *CoRR*.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *CoRR*, abs/1409.1259.

Choi, H., Kim, J., Joe, S., and Gwon, Y. (2021). Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks. *CoRR*.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. (2019). What does BERT look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–278.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Dichao, H. (2020). An Introductory Survey on Attention Mechanisms in NLP Problems. In *Intelligent Systems and Applications*, pages 432–448. Springer International Publishing.

Dong, X. and de Melo, G. (2019). A robust self-learning framework for cross-lingual text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6306–6310, Hong Kong, China. Association for Computational Linguistics.

Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., and Lu, Z. (2019). Ml-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26:1279–1285.

Faber, P. and Usón, R. (1999). *Constructing a Lexicon of English Verbs*. De Gruyter Inc.

Firth, J. (1957). A Synopsis of Linguistic Theory. *Studies in Linguistic Analysis 1930-55*, 1952-59:1–32.

Gage, P. (1994). A New Algorithm for Data Compression. *The C Users Journal*.

Goldberg, Y. (2019). *Neural Networks Methods for Natural Language Processing*. Morgan & Claypool Publishers.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Google (2021). Machine learning glossary. https://developers.google.com/machine-learning/glossary.

Graves, A. (2013). Generating Sequences With Recurrent Neural Networks. *CoRR*, abs/1308.0850.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, USA.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet classification. *CoRR*, abs/1502.01852.

Heumann, C., Aßenmacher, M., Poerner, N., Schütze, H., and Weißweiler, L. (Winter term 2020/21). Lecture notes in Deep Learning for Natural Language Processing.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. *IEEE Press*.

Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339.

Kenton, W. (2022). 8-k (form 8k). https://www.investopedia.com/terms/1/8-k.asp.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *CoRR*, abs/1808.06226.

Lample, G. and Conneau, A. (2019). Cross-lingual Language Model Pretraining. *CoRR*, abs/1901.07291.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization. *CoRR*, abs/1711.05101.

Manning, C., Raghavan, P., and Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *CoRR*, abs/1310.4546.

Pascanu, R., Mikolov, T., and Bengio, Y. (2012). On the difficulty of training recurrent neural networks. *CoRR*, abs/1211.5063.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.

Pilehvar, M. and Camacho-Collados, J. (2020). *Embeddings in Natural Language Processing: Theory and Advances in Vector Representation of Meaning*. Morgan & Claypool Publishers.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-training.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR*, abs/1908.10084.

Reimers, N. and Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *CoRR*, abs/2004.09813.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Review*, pages 386–408.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature 323*, 5.

Salton, G., Wong, A., and Yang, C. (1975). A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620.

Sarwar, M., Zafar, S., Mkaouer, M., Walia, G., and Malik, M. (2020). Multi-label Classification of Commit Messages using Transfer Learning. In *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 37–42.

Schuster, M. and Nakajima, K. (2012). Japanese and Korean Voice Search. *ICASSP 2012*, pages 5149–5152.

SEC (2022). Form 8-k. https://www.sec.gov/fast-answers/answersform8khtm.html.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2020). MPNEt: Masked and Permuted Pre-training for Language Understanding. *CoRR*, abs/2004.09297.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need.

Wu, Y., Schuster, M., Chen, Z., Le, Q., and M., N. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

Yang, Y., Uy, M., and Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. *CoRR*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *CoRR*, abs/1906.08237.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *CoRR*, abs/1506.06724.

Appendix A

Mathematics

A.1 CBOW and Skip-gram

CBOW

The task of CBOW is to maximise the likelihood of a word w_i given its context $C_i = (w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2})$ (assuming a window size of 2), i.e.

$$argmax \prod_{w_i, C_i} P(w_i|C_i)$$

or expressed as a negative log likelihood loss

$$L = \sum_{w_i, C_i} -log(P(w_i|C_i))$$

Skip-gram

The task of Skip-gram is to maximise the likelihood of the context words $C_i = (w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2})$ given their centre word w_i (assuming a window size of 2), i.e.

$$argmax \prod_{w_i, C_i} \prod_{w_{i'} \in C_i} P(w_{i'}|w_i)$$

with $w_{i'}$ denoting the $i-th$ word in C_i. Or expressed as a negative log likelihood loss

$$L = \sum_{w_i, C_I} \sum_{w_{i'} \in C_i} -log(P(w_{i'}|w_i))$$

A.2 Softmax function

The standard *softmax* function $\sigma : \mathbb{R}^J \to (0,1)^J$ is defined as

$$\sigma(\mathbf{x})_i = \frac{exp(x_i)}{\sum_{j=1}^{J} exp(x_j)}$$

for $i = 1, ..., J$ and $\mathbf{x} = (x_1, ..., x_J) \in \mathbb{R}^J$. In other words we divide the exponential of a single input by the sum of the exponentials of all the inputs. This can be seen as a normalisation step, hence with the results lying in in the bounded interval (0,1).

A.3 Word order in self-attention

This proof is taken from Heumann et al. (2021) and demonstrate that self-attention cannot model the order of the words, implying the need for positional embedding. Let $Q = XW^Q$, $K = XW^K$ and $V = XW^V$ be the query, key and value matrices. Let $g \in \{1, ..., J\}^J$ be the permutation of the word order between $X^{(1)}$ and $X^{(2)}$. With this definition $j \to g_j$ is bijective, i.e.

$$x_j^{(1)} = x_{g_j}^{(2)} \quad \forall j \in \{1, ..., J\}$$

This is trivially true, since the word embedding lookup layer represents x_j as $\mathbf{x}_j = \mathbf{w}_{\mathcal{I}_{(x_j)}}$, with $\mathcal{I}$ a bijective indexing function, i.e.

$$x_j^{(1)} = x_{g_j}^{(2)} \implies \mathbf{w}_{\mathcal{I}_{(x_j^{(1)})}} = \mathbf{w}_{\mathcal{I}_{(x_j^{(2)})}} \implies x_j^{(1)} = x_{g_j}^{(2)}.$$

We can define the output of the self-attention for a single j as

$$\mathbf{o}_j = \sum_{j'=1}^{J} \alpha_{j,j'} \mathbf{v}_{j'}$$

with $\alpha_{j,j'}$ defined as in section 3.2.1 and $\mathbf{v}_{j'} \in V$. We want to show that $\mathbf{o}_j^{(1)} = \mathbf{o}_{g_j}^{(2)} \quad \forall j \in \{1, ..., J\}$. But since addition is commutative and the permutation bijective, it is sufficient to show that

$$\alpha_{j,j'}^{(1)} \mathbf{v}_{j'}^{(1)} = \alpha_{g_j,g'_j}^{(2)} \mathbf{v}_{g'_j}^{(2)} \quad \forall j \in \{1, ..., J\}, j' \in \{1, ..., J\}$$

Proof

We show that $\mathbf{v}_j^{(1)} = \mathbf{v}_{g_j}^{(2)} \ \forall j$. We know that $\mathbf{v}_j = W^{(v)T}\mathbf{x}_j$. Therefore,

$$\mathbf{x}_j^{(1)} = \mathbf{x}_{g_j}^{(2)} \implies \mathbf{W}^{(v)T}\mathbf{x}_j^{(1)} = \mathbf{W}^{(v)T}\mathbf{x}_{g_j}^{(2)} \implies \mathbf{v}_j^{(1)} = \mathbf{v}_{g_j}^{(2)}.$$

It only remains to prove that $\alpha_{j,j'}^{(1)} = \alpha_{g_j,g_j'}^{(2)} \ \forall j \in \{1,...,J\}, j' \in \{1,...,J\}$. We know that

$$\alpha_{j,j'} = \frac{exp(e_{j,j'}}{\sum_{j''=1}^{J} exp(e_{j,j''}}$$

with $e_{j,j'}$ defined as

$$e_{j,j'} = \frac{1}{\sqrt{d_k}}\mathbf{q}_j^T\mathbf{k}_{j'} = \frac{1}{\sqrt{d_k}}(\mathbf{W}^{(q)T}\mathbf{x}_j)^T(\mathbf{W}^{(k)T}\mathbf{x}_{j'}).$$

Since the sum in the denominator of $\alpha_{j,j'}$ is commutative and the permutation bijective, we just need to show that $e_{j,j'}^{(1)} = e_{g_j,g_j'}^{(2)} \ \forall j \in \{1,...,J\}, j' \in \{1,...,J\}$. Hence,

$$\begin{aligned}
&\mathbf{x}_j^{(1)} = \mathbf{x}_{g_j}^{(2)} \ \wedge \ \mathbf{x}_{j'}^{(1)} = \mathbf{x}_{g_{j'}}^{(2)} \\
\implies &\mathbf{W}^{(q)T}\mathbf{x}_j^{(1)} = \mathbf{W}^{(q)T}\mathbf{x}_{g_j}^{(2)} \ \wedge \ \mathbf{W}^{(k)T}\mathbf{x}_{j'}^{(1)} = \mathbf{W}^{(k)T}\mathbf{x}_{g_{j'}}^{(2)} \\
\implies &\mathbf{q}_j^{(1)} = \mathbf{q}_{g_j}^{(2)} \ \wedge \ \mathbf{k}_{j'}^{(1)} = \mathbf{k}_{g_{j'}}^{(2)} \\
\implies &\mathbf{q}_j^{(1)}\mathbf{k}_{j'}^{(1)} = \mathbf{q}_{g_j}^{(2)}\mathbf{k}_{g_{j'}}^{(2)} \\
\implies &\frac{1}{\sqrt{d_k}}\mathbf{q}_j^{(1)}\mathbf{k}_{j'}^{(1)} = \frac{1}{\sqrt{d_k}}\mathbf{q}_{g_j}^{(2)}\mathbf{k}_{g_{j'}}^{(2)} \\
\implies &e_{j,j'}^{(1)} = e_{g_j,g_{j'}}^{(2)}
\end{aligned}$$

q.e.d

So $\mathbf{o}_j^{(1)} = \mathbf{o}_{g_j}^{(2)} \ \forall j \in \{1,...,J\}$, or in other words, the representation of a word is always the same, regardless of its meaning. Thus, a positional embedding is necessary.

Appendix B

Data Overview

B.1 Available features in each dataset

Origin
- Ad-Hoc's title
- Publication date
- Publication time
- Source
- Ad-Hoc's text split in sentences
- Unique hash
- Company's ISIN code
- Company's WKN code
- Company's name
- Company's address
- Company's mail
- Company's web address
- Ad-Hoc's language
- Ad-Hoc's text complete

Goldstandard
- Single German sentence or pair of sentences
- Numerical position of gs in document
- Unique hash of document the gs belongs to
- Labels as string
- Boolean value for each single label

Forms 8-K
- Publication date
- Form 8-K's unique hash the item belongs to
- Company's name
- Company's CIK code
- Item's number
- Item's text

B.2 Labels' definitions

Definitions of the single classes used in this work

Class	Definition
Earnings	Earnings announcements; regular reporting of the financial results or disclosure of key performance indicators
SEO	Capital increase or reduction through the issue of additional shares
Management	All kind of changes in the Management (executive/supervisory board etc.)
Guidance	Company's forecast of its own profit or loss in the near future
Gewinnwarnung	Surprising deterioration of the financial result or the result of the forecast
Beteiligung	New or expanding participation in company or own participation in other company, incl. takeover
Dividende	Announcement of dividends or amount of them, incl. corrections and expectations
Restructuring	Restructuring measures concerning processes, organisation, capital structure, f.e. debt-equity-swap, operational restructuring, separation of business/subsidiary etc.. Usually occurs when the company is in crisis
Debt	Company issues loan/bond or repatriates

Law	Company involved in court cases or under investigation (proceedings opened/closes, provisions for litigation, sued)
Großauftrag	Completion of major project/order for the company
Squeeze	Majority shareholder applies for squeeze, incl. progress of proceedings
Insolvenzantrag	Company or third party has filed/will file for insolvency
Insolvenzplan	Information on concrete progress of the insolvency proceedings is published
Delay	Report or general meeting is postponed or not published at all/does not take place or if audit firm needs time
Split	Company undertakes share split
Pharma Good	Drug approval/announcement/study success
Rückkauf	Repurchase/Buyback of own shares
Real Invest	Purchase or sale of assets such as land, factories, machinery, etc.
Delisting	Permanent delisting of company
Irrelevant	Sentence/section does not belong to the core of the message
Empty	Sentence/section does not contain any important information concerning our classes

Appendix C

Forms 8-k items

Items for the Forms 8-k as defined by the SEC[1]

Section 1	**Registrant's Business and Operations**
Item 1.01	Entry into a Material Definitive Agreement
Item 1.02	Termination of a Material Definitive Agreement
Item 1.03	Bankruptcy or Receivership
Item 1.04	Mine Safety - Reporting of Shutdowns and Patterns of Violations

Section 2	**Financial Information**
Item 2.01	Completion of Acquisition or Disposition of Assets
Item 2.02	Results of Operations and Financial Condition
Item 2.03	Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Registrant
Item 2.04	Triggering Events That Accelerate or Increase a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement
Item 2.05	Costs Associated with Exit or Disposal Activities
Item 2.06	Material Impairments

Section 3	**Securities and Trading Markets**
Item 3.01	Notice of Delisting or Failure to Satisfy a Continued Listing Rule or Standard; Transfer of Listing
Item 3.02	Unregistered Sales of Equity Securities
Item 3.03	Material Modification to Rights of Security Holders

[1]Source: https://www.sec.gov/fast-answers/answersform8khtm.html

Section 4	Matters Related to Accountants and Financial Statements
Item 4.01	Changes in Registrant's Certifying Accountant
Item 4.02	Non-Reliance on Previously Issued Financial Statements or a Related Audit Report or Completed Interim Review

Section 5	Corporate Governance and Management
Item 5.01	Changes in Control of Registrant
Item 5.02	Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensatory Arrangements of Certain Officers
Item 5.03	Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year
Item 5.04	Temporary Suspension of Trading Under Registrant's Employee Benefit Plans
Item 5.05	Amendment to Registrant's Code of Ethics, or Waiver of a Provision of the Code of Ethics
Item 5.06	Change in Shell Company Status
Item 5.07	Submission of Matters to a Vote of Security Holders
Item 5.08	Shareholder Director Nominations

Section 6	Asset-Backed Securities
Item 6.01	ABS Informational and Computational Material
Item 6.02	Change of Servicer or Trustee
Item 6.03	Change in Credit Enhancement or Other External Support
Item 6.04	Failure to Make a Required Distribution
Item 6.05	Securities Act Updating Disclosure

Section 7	Regulation FD
Item 7.01	Regulation FD Disclosure

Section 8	Other Events
Item 8.01	Other Events (The registrant can use this Item to report events that are not specifically called for by Form 8-K, that the registrant considers to be of importance to security holders.)

Section 9	Financial Statements and Exhibits
Item 9.01	Financial Statements and Exhibits

Appendix D

Labels Transfer Evaluation

German goldstandard	Best English counterpart in the English Ad-Hoc
Sunways AG: Sunways AG stellt Antrag auf Eröffnung eines Insolvenzverfahrens.	Sunways AG: Sunways AG to file for the opening of insolvency proceedings
20. März 2014 - Der Vorstand der Sunways AG (SWW : GR, SWWG.DE ISIN DE0007332207) hat heute wegen Zahlungsunfähigkeit des Unternehmens den Beschluss gefasst, morgen beim Amtsgericht Konstanz die Eröffnung eines Insolvenzverfahrens zu beantragen.	Due to illiquidity of the company, the Management Board of Sunways AG (SWW:GR, SWWG.DE, ISIN DE0007332207) has today taken the decision to file with the Konstanz local court tomorrow for the opening of insolvency proceedings.
Der Antrag wird das Vermögen der deutschen Konzerngesellschaften, also der Sunways AG mit Sitz in Konstanz und ihrer hunderprozentigen Tochtergesellschaft, der Sunways Production GmbH mit Sitz in Arnstadt, betreffen.	The application will concern the assets of all German Group companies, i.e. Sunways AG with registered office in Konstanz and its wholly owned subsidiary, Sunways Production GmbH with registered office in Arnstadt, Germany.
Angestrebtes Ziel ist ein Insolvenzplanverfahren, das den Erhalt der Sunways AG als börsennotierter Gesellschaft auf der Grundlage eines tragfähigen Sanierungskonzeptes und nach einem Vergleich mit den Gläubigern des Unternehmens ermöglicht.	Ultimate objective is an insolvency plan procedure that allows the preservation of Sunways AG as an exchange-listed stock corporation on the basis of a viable restructuring plan and an arrangement with the creditors of the company.
Der Vorstand hat bereits Gespräche mit potenziellen Investoren aufgenommen und wird diese auch nach Antragstellung fortführen.	The Board is already in talks with potential investors and will continue these talks regardless of the filing.

Tab. D.1: Document 1

German goldstandard	Best English counterpart in the English Ad-Hoc
DEUTZ AG veräußert Standort Köln-Deutz.	DEUTZ AG sells its Cologne-Deutz site
Die DEUTZ AG hat heute die Grundstücke ihres bisherigen Standorts Köln-Deutz an den Düsseldorfer Projektentwickler GERCHGROUP veräußert.	Today, DEUTZ AG has sold the land occupied by its former Cologne-Deutz site to the Düsseldorf-based project developer GERCHGROUP.
Die Absicht zur Veräußerung dieses Standorts hatte DEUTZ bereits Mitte Februar bekannt gegeben.	DEUTZ had announced its intention to sell the site back in mid-February.
Der bisherige Standort Köln-Deutz mit einem Areal von rund 160.000 qm wird nach der erfolgten Verlagerung dieses Standorts nach Köln-Porz nicht mehr benötigt.	The premises in Cologne-Deutz, which cover an area of around 160,000 square metres, are no longer required following the site's relocation to Cologne-Porz.
Die GERCHGROUP beabsichtigt, die bisherige Industriefläche in den kommenden Jahren in ein urbanes Stadtquartier mit hohem Wohnanteil in Nähe des Rheins zu konvertieren.	GERCHGROUP intends to redevelop this former industrial site, which is close to the Rhine, to create a new city district with a high proportion of housing.
Aus der Veräußerung erwartet DEUTZ im laufenden Jahr den Zufluss eines Kaufpreises von rund 125 Mio. EUR. In Abhängigkeit vom Abschluss des laufenden Bebauungsplanverfahrens rechnet DEUTZ für die kommenden Jahre noch mit einer weiteren finalen Kaufpreisrate , deren Höhe variabel ist und die im Erfolgsfall bis in den mittleren zweistelligen Millionen Euro-Bereich reicht.	DEUTZ expects to receive a sum around EUR125 million as purchase consideration this year. Depending on completion of the ongoing planning process, DEUTZ anticipates a further, final instalment of the purchase consideration in the coming years. The exact amount is not yet known and, provided the planning application is successful, will reach into the mid double-digit million euros.
Aus dieser Transaktion erwartet DEUTZ im laufenden Jahr einen positiven Ergebnisbeitrag im hohen zweistelligen Millionen Euro Bereich (nach Steuern), der als Sondereffekt ausgewiesen wird.	In the current year DEUTZ expects this transaction to deliver a positive contribution to earnings in the high double-digit million euros (after taxes) that will be recognised as an exceptional item.

Tab. D.2: Document 2

German goldstandard	Best English counterpart in the English Ad-Hoc
PNE WIND AG veräußert 80prozentige Beteiligung an der PNE WIND YieldCo Deutschland GmbH.	PNE WIND AG sells 80% majority stake in PNE WIND YieldCo Deutschland GmbH.
Die PNE WIND AG hat heute mit einer Tochtergesellschaft der AREF II Renewables Investment Holding S.à.r.l. einen Vertrag über den Verkauf von 80 Prozent der Geschäftsanteile an der PNE WIND YieldCo Deutschland GmbH mit einem Unternehmenswert (Enterprisevalue) von mehr als 330 Mio. Euro unterzeichnet.	Cuxhaven, December 9, 2016 - Today PNE WIND AG signed a share purchase agreement with a subsidiary of AREF II Renewables Investment Holding S.á.r.l. concerning the sale of an 80% shareholding in PNE WIND YieldCo Deutschland GmbH, which has an enterprise value of more than EUR 330 million.
Die PNE WIND AG bleibt mit 20 Prozent an der PNE WIND YieldCo Deutschland GmbH beteiligt und übernimmt das Betriebsmanagement YieldCo und der darin enthaltenen Windparks.	PNE WIND AG retains a share of 20% in the company and will be responsible for the operational management of the YieldCo and its wind farm projects.
Der Vollzug des Kaufvertrags steht noch unter aufschiebenden Bedingungen, u.a. der Freigabe des Bundeskartellamts.	The closing of the share purchase agreement is subject to conditions precedent, including the approval of the transaction by the Federal Cartel Authority.
Der Kaufpreis, dessen Zahlung noch in diesem Jahr erwartet wird, beträgt rund 103 Mio. Euro.	The payment of the purchase price in the amount of approximately EUR 103 million is expected before this year's end.
Bei Vollzug des Vertrages noch in diesem Jahr ist der nunmehr vereinbarte Anteilsverkauf ein wesentlicher Schritt zum Erreichen des Ergebnisziels eines Konzern-EBIT im Bereich von bis zu 100 Mio. Euro für das Geschäftsjahr 2016.	The share purchase agreement now signed is a major step towards attaining the earnings target of a consolidated EBIT up to EUR 100 million for the financial year 2016, provided that its closing takes place this year.
Bei dem Käufer handelt es sich um eine Enkelgesellschaft des Energie- und Infrastrukturfonds Allianz Renewable Energy Fund II, der von Allianz Global Investors GmbH verwaltet wird.	The purchaser, a sub-subsidiary company of the energy- and infrastructure- fund Allianz Renewable Energy Fund II, is managed by Allianz Global Investors GmbH.
In der PNE WIND YieldCo Deutschland GmbH hat die PNE WIND AG bislang Windparkprojekte mit insgesamt 142,5 MW gebündelt, von denen sich 6 MW noch in Bau befinden.	PNE WIND AG has bundled in PNE WIND YieldCo Deutschland GmbH wind farm projects with a total of 142.5 MW, including wind farm projects under construction with a total of 6.6 MW.
Es besteht eine Option auf den Erwerb weiterer 9,9 MW von der PNE WIND Gruppe.	Furthermore, PNE WIND YieldCo Deutschland GmbH has an option to purchase a further 9.9 MW from PNE WIND Group.

Tab. D.3: Document 3

German goldstandard	Best English counterpart in the English Ad-Hoc
GK Software AG - Verschiebung der Hauptversammlung wegen Änderung des Geschäftsberichtes 2013 und des Quartalsabschlusses 2014.	GK Software AG - Postponement of Annual Shareholders' Meeting due to Changes of Annual Report 2013 and Quarterly Report 2014
Die Hauptversammlung der GK SOFTWARE AG wird nicht, wie am 9 Mai des Jahres im Bundesanzeiger angekündigt, am 18 Juni 2014 stattfinden.	The annual shareholders' meeting at GK SOFTWARE AG will not take place on 18 June 2014, as announced in the Federal Gazette on 9 May this year.
Die Ursache dafür liegt in einer notwendigen Änderung des Geschäftsberichtes auf Grund der neuen Würdigung eines Sachverhaltes, die in Zusammenarbeit mit den Wirtschaftsprüfern der Gesellschaft vorgenommen worden ist.	The reason for this is a change that is necessary in the financial statement because of the new emphasis on one accounting issue, this has been made in conjunction with the company's auditors.
Der Sachverhalt wurde bei der prüferischen Durchsicht des ersten Quartalsberichtes für das Geschäftsjahr 2014 festgestellt.	The issue came to light during the auditor's checks on the report for the first quarter of the 2014 business year.
Dem folgend muss die Zuordnung einer Aufwandsrechnung in das Vorjahr erfolgen	As a result, it is necessary to transfer the allocation of one invoice for expenses to the previous year.
Nach vorläufigen Zahlen ändern sich durch diesen Sachverhalt einzelne Positionen im Einzelabschluss (nach HGB) und im Konzernabschluss (nach IFRS).	According to preliminary figures, particular items in the individual accounts (in line with the German Commercial Code) and in the consolidated accounts (in line with IFRS) will change as a result of this issue.

Im geänderten Konzernabschluss 2013 wird der Konzernüberschuss voraussichtlich um 0,19 Mio. Euro geringer ausfallen.	In the amended consolidated accounts for 2013, the Group's annual profits are estimated to be by EUR 0.19 million lower than previously reported.
Die Änderungen werden auch Einfluss auf die entsprechenden Überträge in den Quartalsbericht 2014 haben, der ebenfalls geändert wird.	The changes will have an effect on the relevant carrying amounts into the quarterly report for 2014, this is also being amended.
Diese Änderungen werden nicht die Gewinn- und Verlustrechnung des Quartalsberichtes betreffen	These changes will not affect the profit and loss statement in the quarterly report.
Der Vorstand der GK SOFTWARE AG wird die Änderungen in Einzelabschluss und Konzernlagebericht zügig vornehmen und nach der Feststellung des Jahresabschlusses und der Billigung des Konzernjahresabschlusses durch den Aufsichtsrat die Hauptversammlung umgehend erneut einberufen.	The Management Board at GK SOFTWARE AG will speedily introduce the changes in the individual accounts and the consolidated accounts and promptly convene the annual shareholders' meeting again once the annual accounts have been settled and the Supervisory Board has approved the consolidated annual accounts.
Dabei wird es seitens der Gesellschaft keine Änderungen der Tagesordnung oder der Beschlussvorschläge der Verwaltung, einschließlich des Vorschlages des Vorstandes eine Dividende von 0,25 Euro pro Aktie auszuschütten, geben	The company will not make any changes to the agenda or the suggestions for decisions made by management, including the proposal by the Management Board to pay a dividend of EUR 0.25 per share.

Tab. D.4: Document 4

German goldstandard	Best English counterpart in the English Ad-Hoc
Allgeier platziert eigene Aktien.	ALLGEIER sells treasury shares
Der Vorstand der Allgeier SE (ISIN DE0005086300, WKN 508630), hat heute mit Zustimmung des Aufsichtsrats vom selben Tag auf Grundlage der Ermächtigung der Hauptversammlung vom 17. Juni 2010 beschlossen, bis zu 450.000 eigene Aktien (entsprechend bis zu 5 % Prozent des Grundkapitals) zu verkaufen.	The Management Board of ALLGEIER SE (ISIN DE0005086300, WKN 508630) with consent of the Supervisory Board today resolved to sell up to 450,000 treasury shares (corresponding to up to 5% of the Company's share capital) on the basis of the authorisation granted by the Annual General Meeting on 17 June 2010.
Der Verkauf erfolgt über ein beschleunigtes Platzierungsverfahren (Accelerated Bookbuilding-Verfahren), in dem die Aktien im Rahmen einer Privatplatzierung qualifizierten institutionellen Anlegern in Deutschland und im europäischen Ausland angeboten werden.	The sale will be carried out through an 'Accelerated Bookbuilding' process, in which the shares are offered in a private placement to qualified institutional investors in Germany and Europe.
Im Rahmen eines Rückkaufprogramms in den Jahren 2009 bis 2013 hatte die Allgeier SE insgesamt 760.493 Aktien, bzw rund 8 % des Grundkapitals erworben.	As part of share buyback programmes during the period of 2009 through 2013, ALLGEIER SE has purchased a total of 760,493 shares corresponding to app 8% of the share capital.
Durch die Aktienplatzierung wird sich der Streubesitz der Allgeier SE auf bis zu 56,4 % erhöhen.	Through the placement, the free float of ALLGEIER SE will increase to up to 56.4%
Der Erlös aus der Transaktion dient der weiteren Stärkung der Kapitalstruktur sowie allgemeinen Unternehmenszwecken.	The proceeds from the transaction will be used to further increase the capital structure and for general corporate purposes.
Die Baader Bank begleitet die Transaktion als Sole Lead Manager und Sole Bookrunner.	INFORMATION MISSING

Tab. D.5: Document 5

German goldstandard	Best English counterpart in the English Ad-Hoc
Medisana AG: MEDISANA übernimmt Mehrheit an der Gimelli Laboratories Co. Ltd., Hong Kong.	Medisana AG: MEDISANA acquires majority of Gimelli Laboratories Co. Ltd., Hong Kong
Auslandsexpansion wird weiter voran getrieben.	Next step in continued overseas expansion announced.
Erweiterung der Wertschöpfungskette und des Produktportfolios sowie Entfall des Besserungsscheins stärken Umsatz- und Ertragsperspektiven erheblich	Extension of the value-creation chain and product portfolio, as well as liquidation of the debtor warrant, significantly enhance revenue and earnings prospects.
Die MEDISANA AG hat heute 51% der Anteile an der Gimelli Laboratories Co. Ltd., Hong Kong, (Gimelli) übernommen.	Today, MEDISANA AG has acquired 51% of the shares in Gimelli Laboratories Co. Ltd., Hong Kong, (Gimelli).
Das Unternehmen mit 450 Mitarbeitern, gegründet 1945 in der Schweiz, ist seit Anfang der 1990er Jahre in Hong Kong und China als ISO-zertifizierter Hersteller von Medizinprodukten in den Bereichen Dental Care sowie Cosmetic Products und Qualifizierte Körperpflege tätig.	The company, which employs 450 staff members and was founded in Switzerland in 1945, has operated since the early 1990s in Hong Kong and China as an ISO-certified manufacturer of medicine products in the areas of dental care, as well as cosmetic products and qualified personal hygiene.
Finanziert wird diese Akquisition zu einem überwiegenden Teil durch die Ausgabe von 630.000 MEDISANA Aktien im Rahmen einer Sachkapitalerhöhung.	The greater portion of this acquisition will be financed by issuing 630,000 MEDISANA shares as part of a capital increase through contributions in kind.
Darüber hinaus wird der MEDISANA AG im heute unterzeichneten Vertrag das Recht eingeräumt, über eine Kaufoption innerhalb von 4 Jahren weitere 49% der Unternehmensanteile zu erwerben	The agreement that has been signed today also entitles MEDISANA AG to purchase a further 49% of the company's shares within a four-year period by way of a purchase option.
Mit dem Erwerb der Gimelli Laboratories Co. Ltd. verbreitert die MEDISANA AG ihre Wertschöpfungskette und verfügt nun auch über eine moderne integrierte Fertigung.	With the acquisition of Gimelli Laboratories Co. Ltd., MEDISANA AG is not only extending its value-creation chain, but now also has access to modern integrated manufacturing capabilites.

Zugleich wird das Produktportfolio der MEDISANA AG um den Bereich Dental Care erweitert.

Für die Bereiche Cosmetic Products und Qualifizierte Körperpflege entstehen darüber hinaus erhebliche Synergieeffekte.

Auch dem bereits deutlich wachsenden Auslandsgeschäft der MEDISANA AG kommt die Akquisition zugute.

So verfügt Gimelli unter anderem über eine starke Position im US-Markt

Der Kauf von Gimelli Laboratories Co. Ltd. bewirkt nicht nur spürbare Umsatz- und Ergebnisbeiträge, sondern auch den Wegfall eines Besserungsscheins gegenüber Gimelli International Ltd, Hong Kong, demgemäß die MEDISANA AG bisher nur ein Ergebnis vor Steuern (EBT) von 2 Mio. EUR p.a. einbehalten durfte.

Ein darüber hinaus gehendes Jahresergebnis hätte an Gimelli International Ltd, Hong Kong gezahlt werden müssen, wofür das Unternehmen einen Forderungsverzicht erklärte.

Im Rahmen des heutigen Erwerbs entfällt dieser Besserungsschein, womit das zukünftige Ergebnis der MEDISANA AG nicht mehr belastet ist und voll dem Unternehmen zur Verfügung steht

Ausblick: Viertes Rekordjahr in Folge avisiert

At the same time, MEDISANA AG's product portfolio is being extended to include the dental care area.

Above and beyond this, significant synergy effects arise from the cosmetic products and qualified personal hygiene areas.

The acquisition also benefits the international business of MEDISANA AG that is already recording marked growth.

Gimelli commands a strong position in the US market, among other areas.

The purchase of Gimelli Laboratories Co. Ltd. not only realises considerable revenue and earnings contributions, but also results in the cancellation of a debtor warrant to Gimelli International Ltd, Hong Kong, as a result of which MEDISANA AG has been restricted to retaining earnings before tax (EBT) of only EUR 2 million per annum to date.

Any annual earnings above this level would have had to have been disbursed to Gimelli International Ltd, Hong Kong, for which the company issued a waiver of debts outstanding.

This debtor warrant lapses as part of today's acquisition, as a consequence of which MEDISANA AG's future earnings are no longer subject to deductions, and are fully available to the company.

Outlook: notification of fourth consecutive record year

Konsolidierungseffekte mit spürbar positiven Umsatz- und Ergebnisbeiträgen aus dieser Akquisition ergeben sich für den MEDISANA-Konzern ab dem 1 Januar 2010.	The MEDISANA Group will realise consolidation effects from this acquisition with significantly positive revenue and earnings contributions from January 1, 2010.
Die Übernahme von Gimelli wird das Unternehmenswachstum 2010 deutlich beschleunigen .	The acquisition of Gimelli will result in a marked acceleration of corporate growth in 2010.
Für das auslaufende Geschäftsjahr 2009 geht der Vorstand unverändert von einem Umsatzplus von mindestens 5% gegenüber dem Rekordwert des Vorjahres von 30,2 Mio. EUR und einem Erreichen des Ergebnisses aus 2008 von 1,0 Mio. EUR aus.	As far as the 2009 financial year that has just ended is concerned, the Management Board continues to expect revenue growth of at least 5% compared with the prior year's record level of EUR 30.2 million, and that the 2008 earnings level of EUR 1.0 million will be attained.
Dementgegen wurde für das Geschäftsjahr 2010 bisher lediglich ein weiteres Unternehmenswachstum avisiert.	By contrast, with a look to the 2010 financial year, the management had only forecast the continuation of company growth to date.
Nunmehr erwartet der Vorstand ein deutliches Wachstum von Umsatz und Ergebnis und damit das vierte Rekordjahr in Folge	The Management Board now anticipates considerable revenue and earnings growth, and consequently the fourth consecutive record year.

Tab. D.6: Document 6

German goldstandard	Best English counterpart in the English Ad-Hoc
Constantin Film AG: Highlight Communications AG legt Barabfindung auf EUR 17,64 je Aktie der Constantin Film AG fest.	Constantin Film AG: cash payment by Highlight Communications AG of EUR 17,64 per Constantin Film AG share
Die Highlight Communications AG, Pratteln/Schweiz, hat mit Schreiben vom 02.03.2009 ihr Verlangen nach §327a Abs 1 Satz 1 AktG an die Constantin Film AG (ISIN DE0005800809) vom 02.12.2008 konkretisiert.	In a letter dated March 2, 2009, Highlight Communications AG, Pratteln/Switzerland, has announced further details of the request it made to Constantin Film AG (ISIN DE0005800809) on December 2, 2008 in accordance with § 327a paragraph 1 sentence 1 of the German Companies Act (AktG).
Im Schreiben vom 02.12.2008 hatte die Highlight Communications AG als Hauptaktionär ein Verlangen an die Constantin Film AG gerichtet, auf der nächsten Hauptversammlung die Übertragung der Aktien der übrigen Aktionäre (Minderheitsaktionäre) auf den Hauptaktionär Highlight Communications AG gegen Gewährung einer angemessenen Barabfindung zu beschließen.	In the letter of December 2, 2008, Highlight Communications AG as the main shareholder asked Constantin Film AG to have a resolution passed at the next shareholders' meeting to transfer the shares of the other shareholders (minority shareholders) to the main shareholder Highlight Communications AG in return for an appropriate cash payment.
Siehe insoweit auch unsere Ad-hoc-Mitteilung vom 02.12.2008	Reference is made in this context to our ad hoc release of December 2, 2008.
Mit dem heutigen Schreiben vom 02.03.2009 hat die Highlight Communications AG ihr Verlangen vom 02.12.2008 konkretisiert und die von ihr festgelegte und den Minderheitsaktionären für die Übertragung ihrer Aktien zu zahlende Barabfindung in Höhe von EUR 17,64 je auf den Inhaber lautender Stückaktie der Constantin Film AG mitgeteilt	In the letter today dated March 2, 2009, Highlight Communications AG has announced further details about its request of December 2, 2008 and has disclosed the cash payment of EUR 17,64 per Constantin Film AG bearer share with no par value that it has specified and that is to be paid to the minority shareholders for the transfer of their shares.
Die Constantin Film AG beabsichtigt, im Rahmen der ordentlichen Hauptversammlung dem Verlangen der Highlight Communications AG entsprechend dem Beschluss über die Übertragung der Aktien der Minderheitsaktionäre auf den Hauptaktionär Highlight Communications AG gegen Gewährung der genannten Barabfindung durch die Highlight Communications AG zur Abstimmung zu stellen	Constantin Film AG intends to put the motion requested by Highlight Communications AG about the transfer of the shares of the minority shareholders to the main shareholder Highlight Communications AG in return for the above-mentioned cash payment by Highlight Communications AG to a vote at the Annual Shareholders' Meeting.

Tab. D.7: Document 7

German goldstandard	Best English counterpart in the English Ad-Hoc
E.ON AG: E.ON führt Wertberichtigung durch - Ergebnisplus für 2008 von 7-8 Prozent - höhere Dividende.	E.ON AG: E.ON to record impairment charge, expects 7-8 percent increase in 2008 earnings and higher per-share dividend
E.ON AG/Jahresergebnis/Dividende	E.ON AG / Final Results/Dividend
Im Rahmen der Erstellung des Konzernabschlusses ist die E.ON AG nach IAS 36 verpflichtet, regelmäßig Impairmenttests durchzuführen.	In preparing its Consolidated Financial Statements, E.ON AG is required under IAS 36 to perform impairment tests on a regular basis.
Für den Konzernabschluss 2008 werden die Bewertungen zu einem Goodwill-Impairmentbedarf bei der Market Unit US Midwest in Höhe von 1,5 Mrd € sowie zu einem Impairmentbedarf auf den Unterschiedsbetrag der von Enel/Acciona und Endesa erworbenen Aktivitäten in Italien, Spanien und Frankreich in einer Größenordnung von ungefähr 1,8 Mrd € führen.	In E.ON's Consolidated Financial Statements for 2008, the impairment tests will result in an impairment charge of €1.5 billion on goodwill for the company's U.S. Midwest market unit and an impairment charge of roughly €1.8 billion on the difference between the book value and the fair value of the operations in Italy, Spain, and France that it acquired from Enel/Acciona and Endesa.
Gründe für die Anpassung bei US Midwest sind vor allem ein Anstieg der market-unit-spezifischen Kapitalkosten sowie niedrigere langfristige Wachstumsraten aufgrund des generellen Marktumfeldes	The main reasons for the U.S. Midwest impairment charge are an increase in the market-unit-specific cost of capital and lower long-term growth rates due to the deterioration of the overall economic situation.
Bei den von Enel/Acciona und Endesa erworbenen Beteiligungen und Kraftwerken wirkte sich vor allem die in Italien vorgenommene Erhöhung des Unternehmenssteuersatzes für Energieunternehmen, Banken und Versicherungen von 27,5 auf 33 Prozent aus.	The increase in Italy's corporate tax rate from 27.5 percent to 33 percent for companies in the energy, banking, and insurance industries is a main factor in the impairment charge regarding the shareholdings and power stations acquired from Enel/Acciona and Endesa.
Ferner hat sich die Perspektive auf dem italienischen Energiemarkt im Herbst 2008 u.a. aufgrund regulatorischer Eingriffe in die Großhandelsmärkte sowie durch derzeit verminderte Produktionsmengen aufgrund der zeitlich verzögerten Inbetriebnahme von Kraftwerken eingetrübt.	In addition, the outlook for the Italian energy market became gloomier in the fall of 2008, in part due to regulatory intervention in wholesale markets and to the current reduction in power production resulting from a delay in the start of operations at certain power plants.

Unter anderem diese Sachverhalte waren bei Abschluss der Transaktion nicht bzw. nicht vollumfänglich bekannt	At the time the transaction closed, these factors, among others, were not fully apparent.
Die Wertberichtigungen mindern den Konzernüberschuss.	The impairment charges will reduce E.ON's consolidated net income.
Auswirkungen auf das Adjusted EBIT und den Bereinigten Konzernüberschuss (Bemessungsgrundlage der Dividendenausschüttung) ergeben sich dadurch aber nicht	But they will not affect the company's adjusted EBIT or its adjusted net income, the key figure E.ON uses to determine its dividend payout.
Nach vorläufigen Zahlen wird E.ON für 2008 ein Adjusted EBIT erzielen, das 7-8 Prozent über Vorjahr liegt.	Based on preliminary numbers, E.ON expects its adjusted EBIT for 2008 to surpass the prior-year figure by 7 to 8 percent.
Der Anstieg des Bereinigten Konzernüberschusses wird in einer vergleichbaren Größenordnung liegen.	The company anticipates a similar increase in adjusted net income.
Vor diesem Hintergrund wird der Vorstand dem Aufsichtsrat eine Dividende von voraussichtlich 1,50 € vorschlagen.	Based on this figure, the E.ON Board of Management anticipates that it would propose to the E.ON Supervisory Board that the company pay a dividend of €1.50 per share.
Unter Berücksichtigung des Aktiensplits entspricht dies einer Steigerung der Dividende um 9,5 Prozent gegenüber Vorjahr	Adjusted for E.ON's stock split, this represents an increase of roughly 9.5 percent on the dividend for the 2007 financial year.
Die vorgenannten Werte basieren auf dem noch nicht final erstellten und geprüften Konzernabschluss der E.ON AG und können somit noch Änderungen unterliegen	These figures are based on E.ON AG's preliminary and unaudited Consolidated Financial Statements and are therefore subject to change.

Tab. D.8: Document 8

German goldstandard	Best English counterpart in the English Ad-Hoc
Deutsche Wohnen AG: Restrukturierungsprogramm vor Umsetzung.	Deutsche Wohnen AG: Implementation of restructuring plan
Der Vorstand der Deutsche Wohnen AG hat heute die Neustrukturierung der Deutsche Wohnen Gruppe beschlossen.	Today, the Management Board of Deutsche Wohnen AG agreed on a restructuring plan for Deutsche Wohnen Group.
Die Folgen sind Standortschließungen und Personalabbau	As a consequence, several branches will be shut down and personnel will be reduced.
Bis 2009 werden konzernweit im Segment Wohnen von derzeit rund 490 Arbeitsplätzen insgesamt ca. 140 Arbeitsplätze eingespart und jährlich insgesamt Personaleinsparungseffekte in Höhe von ca. 10 Mio. € realisiert	Group wide, approximately 140 out of 490 jobs will be cut in the residential business division, resulting in annual cost savings of approximately EUR 10 million by 2009.
Mit dem beschlossenen Restrukturierungsprogramm lassen sich umfangreiche Synergien realisieren, die die Wettbewerbsfähigkeit der Unternehmensgruppe insgesamt nachhaltig sicherstellen und Voraussetzungen für weiteres Wachstum schaffen	The agreed restructuring plan will enable synergies, which ensure the sustainable competitiveness of Deutsche Wohnen Group and prepare the company for further growth.

Tab. D.9: Document 9

German goldstandard	Best English counterpart in the English Ad-Hoc
ISRA VISION AG: Strategisch wichtiger Auftrag im Expansionsbereich FPD-Glas.	ISRA VISION AG: Strategically important contract in the FPD glass field of expansion
ISRA VISION: Strategische Kooperation für den FDP-Markt vereinbart	ISRA VISION: Strategic cooperation in the FPD business
Die ISRA VISION AG (ISIN: DE 0005488100), einer der globalen Top 10 Anbieter für industrielle Bildverarbeitung (Machine Vision), und der Weltmarktführer für Oberflächen-Inspektionssysteme hat mit dem japanischen Anlagenhersteller für FDP-Produktion, Nakan, eine Kooperation vereinbart.	Strategically important contract in the FPD glass field of expansion Darmstadt, February 6, 2008 – ISRA VISION AG (ISIN: DE 0005488100) – one of the top ten suppliers of industrial image processing (Machine Vision) and the world's market leader for surface inspection systems – has signed a cooperation agreement with Nakan, the Japanese line integrator for FPD (flat panel display) production.
ISRA wird mehrere Inspektionsanlagen (Surface Vision) für eine neue Produktions-Linie für Flat Panel Displays (FPD-Glas) liefern, die Nakan errichten wird.	ISRA will be delivering multiple inspection systems (Surface Vision) for a new production line for flat panel displays that Nakan is building.
Der Auftrag hat ein Volumen im siebenstelliges Euro-Bereich	The size of the order is significantly over 1 million Euro.
Die Business Unit FPD-Glas ist einer der wichtigen strategischen Expansionsbereiche für den ISRA-Konzer.	The FPD Glass Business Unit is one of the important strategic sectors for the ISRA Group's expansion.
Mit dem jüngsten Projekt stellt ISRA erneut die herausragende Marktposition als führendes Unternehmen in der Oberflächeninspektion unter Beweis.	With this most recent project, ISRA is once again proving its prominent position on the market as the leading company in surface inspection.
Für einen bedeutenden Hersteller in China wird Nakan eine Beschichtungsanlage für automatische FPD-Glas-Produktion aufbauen.	Nakan will be building a coating for automated FPD glass production for a producer in China.
ISRA wird dafür eine Reihe von unterschiedlichen Inspektionsanlagen liefern.	ISRA will be providing a series of various inspection systems for this project.
Dieser Auftrag kennzeichnet den Beginn einer fruchtbaren Partnerschaft zwischen dem bedeutenden japanischen Automatisierungshersteller für FPD-Produktion Nakan und dem führenden Obeflächen-Inspektions-Systemanbieter ISRA', erklärt Enis Ersü, Vorstandsvorsitzender der ISRA Vision AG.	'This order signifies the beginning of a fruitful partnership between Nakan, the prominent Japanese line integrator for automated FPD production and ISRA, the leading supplier of surface inspection systems,' said Enis Ersü, Chairman of the Executive Board of ISRA VISION AG.

Tab. D.10: Document 10

German goldstandard	Best English counterpart in the English Ad-Hoc
Business Media China AG: Halbjahrergebnis 2007 - kräftige Umsatzsteigerung.	Business Media China AG: Half year results 2007 - strong increase in sales
Die Business Media China AG gibt anlässlich ihrer heutigen Hauptversammlung die Resultate des ersten Halbjahres 2007 bekannt.	On the occasion of its Annual General Meeting, Business Media China AG (WKN 525040) reports half-year results 2007 today.
Die Gruppe hat ihre Aufbauphase abgeschlossen und zeigt jetzt erstmals größere Umsatzsteigerungen.	The Group has successfully completed the business expansion phase and can show now for the first time a substantial increase in revenues.
In der Berichtsperiode wurden Umsatzerlöse von EUR 8,0 Mio. erzielt, davon über zwei Drittel im zweiten Quartal.	A turnover of EUR 8.0 million, of which more than two-thirds in the second quarter, has been achieved in the reporting period.
Das entspricht einer Steigerung von 260% gegenüber der Vorjahresperiode und liegt damit auch bereits über dem Gesamtjahres-Umsatz 2006 (EUR 7,5 Mio.).	This is already exceeding the total revenue accomplished in 2006 and represents an increase of 260% compared to the same period last year.
Das Ergebnis ist wie erwartet aufgrund der enormen Vorleistungen im Geschäftsbereich Werbemedien noch negativ und betrug EUR -3,9 Mio. (Vorjahr EUR -1,3 Mio.)	As expected and due to the fact that enormous intermediate investments have been undertaken in the travel media division, earnings still remain negative and amounted to EUR -3.9 million (2006: EUR -1.3 million).
Die sich jetzt rasch beschleunigende Umsatzentwicklung sollte in der zweiten Jahreshälfte 2007 zu den ersten positiven Quartalsabschlüssen führen.	Turnover is rapidly increasing from now onwards and should lead to the first positive quarterly closings in the second half of 2007 - and will confirm our strategy also in terms of profitability.
Damit wird auch beim Ergebnis die eingeschlagene Strategie bestätigt werden.	Turnover is rapidly increasing from now onwards and should lead to the first positive quarterly closings in the second half of 2007 - and will confirm our strategy also in terms of profitability.
Für das Geschäftsjahr 2007 wird weiterhin ein Gesamtumsatz in Höhe von ca. 26 bis 28 Millionen Euro erwartet	Total revenue for the business year 2007 is still expected to reach between 26 and 28 million Euro.
Um die Strategie, die Geschäftseinheiten und das Marktumfeld der Gesellschaft in China den Anlegern besser verständlich zu machen, wurde zudem ein ausführliches Investoren Handbuch erstellt, das ab heute auf der Website verfügbar ist und regelmäßig aktualisiert werden wird	Additionally, a detailed investor's handbook will be available on the company's website as of today. The document will be regularly updated and illustrates the strategy, business units and market environment in China to interested investors.

Tab. D.11: Document 11

German goldstandard	Best English counterpart in the English Ad-Hoc
ERGO hebt Ergebnisausblick an.	ERGO raises profit outlook
Die ERGO Versicherungsgruppe hebt ihren Ergebnisausblick für das Gesamtjahr 2006 von ursprünglich 450 bis 500 Mio. Euro auf über 600 Mio. Euro an.	The ERGO Insurance Group raises its profit outlook for the financial year 2006 from the previously predicted level of EUR 450 to 500 million to over EUR 600 million according to the latest forecast for the overall year 2006 which has just been submitted to the Board.
Dies sieht die jetzt dem Vorstand vorgelegte aktuelle Hochrechnung für das Gesamtjahr vor.	The ERGO Insurance Group raises its profit outlook for the financial year 2006 from the previously predicted level of EUR 450 to 500 million to over EUR 600 million according to the latest forecast for the overall year 2006 which has just been submitted to the Board.
Diese geht vom Ausbleiben unerwarteter Entwicklungen auf der Schadenseite und den Kapitalmärkten aus	The forecast is based on the assumption that no unexpected developments in claims and on capital markets occur.
Im dritten Quartal 2006 hatten sich die Kapitalmärkte besser entwickelt als angenommen.	During the third quarter in 2006, capital markets performed better than expected.
Auf der Schaden- und der Kostenseite waren weiterhin sehr günstige Entwicklungen zu verzeichnen.	Both claims and costs continued to record positive trends.
Nicht zuletzt deshalb legt die ERGO per 30.9.2006 ein um 36,0% auf 566 (416) Mio. Euro verbessertes Konzernergebnis vor.	Not least due to these effects, ERGO presents a consolidated result of EUR 566 (416) million, an increase of 36.0 percent, as at 30 September 2006.
Die gebuchten Bruttobeiträge blieben stabil bei 11,78 (11,80) Mrd. Euro (-0,2%)	Gross premiums written remained stable at EUR 11.78 (11.80) billion (-0.2 percent).

Tab. D.12: Document 12

German goldstandard	Best English counterpart in the English Ad-Hoc
MorphoSys AG deutsch.	MorphoSys AG english
MorphoSys erreicht zweiten Meilenstein in der Kooperation mit GPC Biotech MorphoSys (Neuer Markt: MOR) und GPC Biotech (Neuer Markt: GPC) gaben heute das Erreichen eines präklinischen Meilensteins im GPC Biotech Immunologie Programm zur Behandlung von Transplantatabstoßung und Graft versus Host Disease (GvHD) bekannt.	MorphoSys achieves Second Milestone in GPC Biotech Collaboration MorphoSys (Neuer Markt: MOR) and GPC Biotech (Neuer Markt: GPC) today announced achievement of a preclinical milestone in GPC Biotechs immunology antibody program for the treatment of transplant rejection and Graft versus Host Disease (GvHD).
GPC Biotech konnte in einem transgenen Tiermodell die signifikante Wirksamkeit des Antikörpers, der von MorphoSys aus der firmeneigenen HuCAL Bibliothek isoliert und optimiert wurde, zeigen.	GPC Biotech showed that a human antibody generated and optimised by MorphoSys from its proprietary HuCAL library had significant in vivo efficacy in a transgenic animal model.
Dies ist nach der Veröffentlichung eines präklinischen Meilen-steins auf dem Gebiet der Krebstherapie vor wenigen Tagen der zweite Erfolg innerhalb kurzer Zeit, den die beiden Firmen in ihrer Kooperation auf verschiedenen Therapie-Gebieten vorweisen können.	INFORMATION MISSING
MorphoSys erhält von GPC Biotech eine weitere Meilensteinzahlung, über deren Höhe keine Angaben gemacht wurde.	The milestone triggers an additional payment from GPC Biotech to MorphoSys.
Der humane HuCAL Antikörper zeigt hohe Affinität zum Zielmolekül MHC Klasse II und wurde für die in vivo Studien ausgewählt, da er sehr effizient die humane T- Helferzellen-Reaktion in vitro hemmen konnte.	The fully human antibody generated by MorphoSys showed high affinity for the MHC class II target and was selected for in vivo studies because it efficiently suppressed human T-helper cell responses in vitro.
Eine einzige Verabreichung des Antikörpers in MHC-transgene Mäuse unterdrückt bestimmte Immunreaktionen in der Haut dieser Tiere.	A single injection of the antibody, administered to transgenic mice carrying the human MHC class II target molecule, was shown to efficiently suppress skin hypersensitivity reactions in these mice.
Im Moment testet GPC Biotech den Antikörper auch in einem transgenen Transplantationsmodell	GPC Biotech is currently assessing this antibody in transgenic mouse-based transplantation models.

Tab. D.13: Document 13

German goldstandard	Best English counterpart in the English Ad-Hoc
Eurofins Scientific S.A. deutsch.	Ad hoc-Service: Eurofins Scientific S.A. english
Im Geschäftsjahr 2000 übertrifft Eurofins Scientific, ein führender globaler Anbieter von bioanalytischen Dienstleistungen, die anlässlich des Secondary Public Offering (SPO) im Oktober letzten Jahres von führenden Analysten getroffenen Prognosen.	In 2000 Eurofins Scientific, a leading global bioanalytics service provider, exceeded the forecasts which were set by lead analysts upon the company's Secondary Public Offering (SPO) in October last year.
Die konsolidierten Umsätze erreichten 50,9 Mio EUR	Consolidated sales reached EUR 50.9 million.
Die EBITDA von 4,5 Mio EUR übertrafen die Analystenschätzungen um 10% und die EBIT von 2,1 Mio EUR lagen sogar um 21% über den Analystenschätzungen	The EBITDA of EUR 4.5 million exceeded analyst expectations by 10% and the EBIT of EUR 2.1 million was even 21% above analyst estimates.
Verglichen mit 1999 stiegen die konsolidierten Umsätze um 58% (1999: 32,2 Mio EUR).	Compared to 1999 the consolidated sales of Eurofins Scientific increased by 58% (1999: EUR 32.2 million).
Dies spiegelt die internationale Expansion der Gruppen-Infrastruktur durch die Akquisition mehrerer Laboratorien wider.	This reflects the expansion of the group's infrastructure internationally by acquisition of several laboratories.
Als Resultat der Investitionen in ein Expansionsprogramm, das darauf ausgerichtet ist, Eurofins Scientific zu einem weltweiten Führer ihres Industriezweiges zu machen, blieben die EBITDA in 2000 fast stabil bei 4,5 Mio EUR (1999: 4,6 Mio EUR).	As a result of the investments into a scale-up program which aims at making Eurofins Scientific a global leader of its industry the 2000 EBITDA remained almost stable at EUR 4.5 million (1999: EUR 4.6 million).

Der Jahresüberschuss vor Abschreibungen von Firmenwerten in Höhe von +0,8 Mio EUR, der Jahresüberschuss/- verlust nach Abschreibungen von Firmenwerten in Höhe von -0,2 Mio EUR, der Gewinn pro Aktie vor Abschreibungen von Firmenwerten in Höhe von 0,06 EUR sowie der Gewinn/Verlust pro Aktie nach Abschreibungen von Firmenwerten in Höhe von - 0,02 EUR übertrafen ebenfalls die Schätzungen führender Analysten	Net profit before goodwill amortization of EUR 0.8 million, net profit after goodwill amortisation of EUR - 0.2 million, EPS before goodwill amortisation of 0.06 EUR and EPS after goodwill amortisation of -0.02 EUR also exceeded lead analyst estimates.
Es sei darauf hingewiesen, dass die im November und Dezember 2000 akquirierten Laboratorien nicht im Jahr 2000 konsolidiert worden sind	It should be noted that the laboratories acquired in November and December 2000 have not been consolidated in 2000.
Die nach IAS konsolidierten Finanzergebnisse geben nur teilweise die substanzielle Expansion der Gruppe wider, weil sie u.a. nicht diejenigen Firmen enthalten, welche im November und Dezember 2000 akquiriert wurden.	The IAS consolidated financial results only partly reflect the substantial expansion of the group, because they i.e. do not include the companies which were acquired in November and December 2000.
So zeigt die Pro-forma Berechnung, die u.a. auch Firmen wie Miljo Kemi enthalten, dass die konsolidierten Umsätze 82,4 Mio EUR erreicht hätten, wenn alle per Jahresende zur Gruppe gehörenden Unternehmen zu 100% für 12 Monate des Jahres 2000 konsolidiert worden wären	However, a pro-forma calculation including also companies such as Miljo Kemi shows that the consolidated sales would have reached EUR 82.4 million, if all companies in the current group's perimeter had been consolidated at 100% for 12 months in 2000.

Tab. D.14: Document 14

German goldstandard	Best English counterpart in the English Ad-Hoc
CENIT AG Systemhaus deutsch.	Ad hoc-Service: CENIT AG Systemhaus english
Super Airbus geht mit CENIT an den Start EADS Airbus GmbH vergibt Großauftrag über 3 Millionen DM an CENIT Stuttgart, 6 Februar 2001. Bei der Entwicklung des neuen A380 Großraumpassagierflugzeuges und des Militärtransporters A400M setzt die EADS Airbus GmbH auf die Zusammenarbeit mit der CENIT AG Systemhaus, Stuttgart	Super Airbus to take off with CENIT EADS Airbus GmbH places major order worth DM 3m with CENIT Stuttgart, February 6th 2001. In the development of the new A380 jumbo jet and the military transporter A400M, EADS Airbus GmbH is counting on co- operation with CENIT AG Systemhaus, Stuttgart.
Die A380 wird mit 555 Sitzplätzen auf zwei Decks das größte Passagierflugzeug der Welt sein.	The A380, with 555 seats on two decks, will be the world's biggest passenger plane.
Die am Neuen Markt notierte CENIT AG übernimmt die Software Schulung für CATIA, die technische Unterstützung sowie die Migration bestehender Daten auf das neue System.	CENIT AG - a company listed on the Neuer Markt is to be responsible for software training in CATIA, technical support and the migration of existing data to the new system.
Mit Hilfe der eingesetzten Software CATIA lässt sich die A380 vollständig digital entwickeln.	The use of CATIA software permits development of the A380 to be performed entirely in digital form.
CATIA ermöglicht es, das Flugzeug virtuell am Computerbildschirm zu bauen und Funktionalitäten zu simulieren.	CATIA makes it possible to build the aircraft virtually on the computer screen, and simulate its functions.
So lassen sich mögliche Konstruktionsfehler bereits im Vorfeld ausschließen.	This allows design errors to be eliminated in advance.
Damit der Leistungsumfang der anspruchsvollen Anwendungen voll ausgeschöpft wird, übernimmt CENIT die Schulung von rund 800 Konstrukteuren und steht den Fachabteilungen mit technischer Beratung zur Seite.	In order to exploit the full potential of the sophisticated applications, CENIT will take over the training of about 800 designers, and stands by the specialist departments with technical support.
CENIT entwickelt mit ihnen zusammen auch die speziellen Konstruktionsmethoden, um den Entwicklungsprozess noch effizienter zu gestalten.	With these, CENIT will also develop the special design methods required to make the development process even more efficient.
Insgesamt brachte das Projekt im vergangenen Jahr bereits 1,4 Millionen Mark Umsatz für CENIT.	Altogether, this project contributed DM 1.4m to CENIT's sales already last year.
Für das Jahr 2001 ist ein Auftragsvolumen in Höhe von über drei Millionen Mark geplant	For 2001, orders with a volume of over three million D-Mark are planned.

Tab. D.15: Document 15