An Investigation of Transformer-based Event Type Classification Through the Linguistic Lens

Inaugural-Dissertation
zur Erlangung der Doktorwürde
der
Philosophischen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität
zu Bonn

vorgelegt von

Theresa Nindel
aus
Lich

Bonn 2026

Gedruckt mit der Genehmigung der Philosophischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Prof. Dr. Anke Grutschus
(*Vorsitzende*)
Prof. Dr. Svenja Kranich
(*Betreuerin und Gutachterin*)
Prof. Dr. Bernhard Schröder
(*Gutachter*)
Prof. Dr. Claudia Wich-Reif
(*weiteres prüfungsberechtigtes Mitglied*)

Tag der mündlichen Prüfung: 23.06.2026

Acknowledgments

I would like to take this opportunity to thank the people who have supported me in the process of completing this dissertation.

I wish to extend my sincere gratitude to my dissertation advisor, Prof. Dr. Svenja Kranich, for her consistently valuable guidance throughout the entire dissertation process. Her encouraging feedback over the years has helped me to continually advance this work. I also would like to thank my second dissertation advisor, Prof. Dr. Bernhard Schröder, for his thoughtful input and his support.

Furthermore, I am grateful to the other members of the examination committee, Prof. Dr. Anke Grutschus and Prof. Dr. Claudia Wich-Reif, for their willingness to be part of my dissertation defense and for the lively discussion.

I would like to express my great appreciation to my colleagues at Fraunhofer FKIE, both current and former. In particular, I would like to thank Prof. Dr. Ulrich Schade and Dr. Hans-Christian Schmitz for the roles they played in my dissertation process, from the discussions about the initial research direction to the final proofreading. I am also especially thankful to Samantha Kent and Sophie Decher for our many conversations and their unwavering support.

Finally, I want to thank my family and friends for their encouragement, and my husband, Hannes, for his enduring understanding and steadfast belief in me.

Contents

List of Figures

List of Tables

1 Introduction

Natural Language Processing (NLP) – implemented with the help of large language models (LLMs) – has proved to be a useful building block for the automatic extraction of events from text. Event extraction (EE) is described as "a crucial research task for promptly apprehending event information from massive textual data" (Li et al. 2021a). An event in the context of NLP is defined as "a particular form of information [...] [that] refers to a specific occurrence of something that happens in a certain time and a certain place involving one or more participants, which can usually be described as a change of state" (Li et al. 2021a).

To detect and classify events in texts reliably, several NLP subtasks can be implemented. These subtasks include, among others, binary classification to determine if a text includes an event (cf. Becker and Krumbiegel 2021), multiclass classification of event types (e.g., demonstration, natural disaster) (cf. Kent and Krumbiegel 2021), causal event classification (cf. Krumbiegel and Decher 2022) and event argument extraction and classification. Event arguments provide information about the location, time, and participants of a particular event (cf. Linguistic Data Consortium 2005).

All subtasks of EE facilitate the analysis of openly available textual data. Data from sources like newspapers and social media can be the subject of analysis as they present information on happenings in the physical world (cf. Krumbiegel et al. 2021). The automatic processing of these texts therefore results in answers to questions about the physical world such as: Did something (relevant) happen (binary event classification)? What happened (event type classification)? Why did it happen (causal event classification)? Where and when did it happen and who was involved (event argument classification)? EE enables the reduction of great amounts of data to essential information and therefore supports distant reading (cf. Krumbiegel et al. 2021).

Systems that use transformer-based large language models such as BERT

(Devlin et al. 2018) and RoBERTa (Liu et al. 2019) perform well on these subtasks. However, recent EE research skews heavily towards the computer science perspective, where model optimization is the focus. The influence of the data used for model training and testing as well as the presence of specific linguistic patterns that define events is often disregarded nowadays. Accordingly, valuable linguistic information falls to the wayside.

The decisive advantage that enables LLMs to perform well with regard to EE and a wide range of other NLP tasks is the way in which they represent and process language. LLMs supply so-called contextual embeddings. These embeddings are created automatically based on large amounts of texts and encode knowledge connected to different areas of linguistics (morphology, syntax, semantics). Contextual embeddings can be adapted to specific use cases. The general knowledge of language encoded by them is thereby leveraged to solve NLP tasks.

The explicit encoding of linguistic information as a way to solve EE tasks was focused on more strongly before the introduction of LLMs. This traditional approach was driven by necessity: Representations of language available to machine learning (ML) systems at this point in time lacked the ability of considering words in their context reliably. In order to mitigate this shortcoming, linguistic features that fit the task were constructed manually and used for model training.

The manual construction of linguistic features is time-consuming in contrast to simply building on an available LLM. Still, the usage of a set of features that fits the task provides a certain amount of control to the researcher. This control is normally not given for work with LLMs as they are black box models. This means that their inputs and outputs are observable, the decision-making process, including which linguistic patterns are focused on for classification, however, is mostly opaque.

This thesis proposes the combination of LLMs and manually created features as a way of approaching one subtask of EE, namely, the classification

of event types. Linguistic features are explicitly added to the database that is used for training and testing the chosen ML models. It is argued that this combination makes the most of the advantages of both LLMs and manually created features: On the one hand, contextual embeddings provide general linguistic knowledge, on the other hand, the explicit addition of features enforces relevant linguistic indicators in the texts that the LLM can then exploit. A model that is optimized in this way is assumed to display a higher degree of robustness than a model that does not use additional features. This means that the classification of more diverse data than was used during training is supported. To test this assumption, models will be trained with and without additional linguistic features. All model training and testing will be based on the same dataset to make the results comparable.

The rest of this work is structured in the following way:

In chapter 2, the task of text classification in general is defined. After that the evolution of language representations used for NLP tasks is introduced. To this end, first, two traditional frequency-based methods are introduced: bag-of-words and term frequency-inverse document frequency. A short overview of the construction of these representations is given. Additionally, the multinominal naive Bayes classifier as an older probabilistic ML model that works under the bag-of-words assumption is introduced. Moreover, the method of constructing features manually is elaborated on.

After the description of frequency-based language representations, static word embeddings are described. These kinds of embeddings resemble contextual embeddings insofar that they are also constructed automatically from text by means of a prediction based approach. They differ however in their ability to consider context.

Finally, contextual embeddings and the transformer architecture which underlies LLMs are detailed. These models are the basis for the work at hand. Therefore, two transformer models, BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019) are considered in more detail.

In chapter 3, the task of EE is introduced. The definitions of the individual subtasks provided by the Linguistic Data Consortium 2005 are drawn upon. Event type classification as a subtask of EE is described in its broader context. Lastly, an overview of approaches that are used for the task is given. This overview considers traditional as well as current methods.

After introducing text classification and language representations as such as well as the intended task of event type classification, chapter 4 describes the general experimental set up used for this work. This includes libraries that are used for model training and testing as well as for linguistic processing of text. Additionally, the corpus that is used for event type classification is described. Finally, a number of evaluation methods are defined.

In chapter 5 the first two experiments that were conducted are reported. Both experiments aim at the classification of event types with the help of transformer models. However, no linguistic features are used yet. The goal is to further the understanding of the automatic features that are selected and used by the models if they are not prompted to focus on specific linguistic indicators. This corresponds to the current trend of using only LLMs to solve NLP tasks.

The first experiment focuses on evaluating the relevance of specific word classes for the classification of event types. To gain first insights, three corpus modifications are created. For each of these modifications, words of a different class are masked in the texts (nouns, verbs, adjectives). A classifier is trained with each. The resulting classifiers are contrasted with a classifier that was trained on the original database. Differences in model performance are reported. The relevance of word classes is further evaluated with the help of the explainable AI library LIME (Ribeiro et al. 2016b).

For the second experiment, model training is repeated. In this case a cross-validation approach is implemented. The database is not modified. The focus of experiment two is the analysis of common errors made by models that use no additional linguistic features. The resulting models serve as baseline mod-

els. They will be used for the evaluation of the models that do use linguistic features later on.

The results of both experiments and the theoretical background that was introduced before lead to a number of research questions that are the basis for the remainder of the work. They are included in the last section of this chapter.

Chapter 6 introduces the linguistic features that were chosen. The features are: aspectual classes, named entities and sentiment. All three features are described. The chosen features need to be added to the data. In contrast to the named entity and sentiment features, a new classifier had to be trained for adding aspectual class information to the texts. Therefore, the chapter is mostly focused on the aspectual class feature. The description of aspectual classes starts with a short theoretical overview. This is followed by an introduction of automatic classification of aspectual classes as an NLP task. After that, all steps of training the aspectual class classifier are reported. This includes a description of the data, model training, results and an error analysis.

Chapter 7 includes the main experiment of this work. The main experiment again aims at the classification of event types. This time the described linguistic features are added to the data and used for model training. In a first step, the specific way in which the features are added to the data is described. As features are added to the data individually as well as in all possible combinations, seven data modifications arise. In a next step, specific assumptions about improvements that are expected by using the individual features are formulated.

Finally, the actual experiment is detailed. Similar to the second experiment in chapter 5, a cross-validation approach is implemented. Models trained with the different data modifications are evaluated individually. The results are contrasted with the baseline models, i.e., with the models that did not use the linguistic features during training. Errors that were identified for the baseline models are analyzed with regard to their existence for the modified

models.

Chapter 8 discusses the experimental results further. This time, the focus is shifted to also include a comparison of the modified models to each other, not only to the baseline models. Differences in model performance are analyzed. The research questions posed in chapter 5 are discussed in view of the experimental results.

The conclusion in chapter 9 finally summarizes this work and gives an outlook for future research directions.

2 Theoretical Background – Text Classification

In general, "[t]he goal of classification is to take a single observation, extract some useful features, and thereby **classify** [emphasis in the original] the observation into one of a set of discrete classes" (Jurafsky and Martin 2024: 61). In the context of text classification, an observation can for example be a single sentence, or a text consisting of a number of sentences. To reach this goal, a classification method is needed. A commonly used classification method in the field of NLP is supervised machine learning. For supervised machine learning, an input x and a set of output classes $Y = \{y_1, y_2, ..., y_M\}$ need to be provided. For the use case of event type classification which will be considered in this work, the input x are texts consisting of $1 - n$ sentences, the set of output classes Y includes all predefined event type labels. A set of N training documents is used for the implementation of supervised machine learning for text classification. Each document is labeled with a class. A trained classifier should ultimately be able to determine the correct class labels for unknown documents reliably.

Classification can be implemented in a binary or in a multiclass manner. As the name suggests, two output classes are included for binary classification. Classifying documents as either spam or not spam, for example, is a binary classification task. Multiclass classification considers more than two output classes. Event type classification is typically a multiclass classification task.

The way in which text classification could be approached in the last decades is highly dependent on available machine-readable representations of language. The method of how useful features are extracted and used by ML models changed along with these representations. Representations have evolved over time. This evolution resulted in what we know today as LLMs. A short overview of traditional approaches that are not predominantly used anymore but paved the way for the development of modern algorithms and

language representations is given in 2.1. An interim summary is provided in chapter 2.2. Chapter 2.3 introduces the transformer architecture and LLMs which are the basis for the work at hand.

If not stated otherwise, the explanations of text classification concepts and methodologies that are found in this chapter are based on the work of Jurafsky and Martin (2024).

2.1 Traditional Language Representations

Texts have to be available in a computer-readable form to be able to use them for automatic classification. In the cases that will be described here, texts are represented as vectors. The method with which language is represented as vectors has changed over time.

In the following chapter, first, two frequency-based approaches will be detailed. Frequency-based representations of language are highly transparent and can be implemented easily. Therefore, a minimal example will be realized to illustrate the approaches. Five fictitious sentences, each describing an event, are listed below. They will be used as a database.

1. The demonstration was held in Paris.
2. The demonstration was announced last week.
3. The participants demonstrated for two hours.
4. An earthquake struck Germany.
5. The earthquake caused several buildings to collapse.

Three sentences describe a demonstration event, two an earthquake event. The Scikit-learn library (Pedregosa et al. 2011) will be used to transform the texts.

Second, a prediction-based approach as the next evolutionary step of representing language as vectors is introduced. This approach results in so-called word embeddings. Word embeddings are derived automatically from large

amounts of texts. They represent linguistic properties that are present in the database (cf. Goldberg 2017). To illustrate the way in which these representations can be utilized, the Gensim library (Řehůřek and Sojka 2010) and word2vec word embeddings (cf. Mikolov et al. 2013a and Mikolov et al. 2013b) will be drawn upon.

Implications of using the representations will be mentioned.

2.1.1 Bag-of-Words

The first approach that focuses on frequencies in order to represent language is the bag-of-words approach. The bag-of-words approach results in vector representations for documents. To transform the documents (i.e., the event sentences) into vectors, the words included in all documents are first of all considered independent of their position. All unique words make up the vocabulary. For the example documents, the resulting vocabulary is shown below. The vocabulary size is 23.

> ['an', 'announced', 'buildings', 'caused', 'collapse', 'demonstrated', 'demonstration', 'earthquake', 'for', 'germany', 'held', 'hours', 'in', 'last', 'paris', 'participants', 'several', 'struck', 'the', 'to', 'two', 'was', 'week']

The crucial factor that is needed for the bag-of-words representations are the raw frequencies of words in the individual documents. A document is thus represented by a vector consisting of these word frequencies. The vector's length corresponds to the vocabulary size. The example documents given above are accordingly represented as vectors as depicted in table 1.
Words that are found in a document are denoted with their frequencies, words that are not found are denoted with 0. The seventh entry in the bag-of-words output, for example, corresponds to the word *demonstration*. The word occurs once in documents 1 and 2 respectively and is not present in documents 3, 4 and 5.

Words are examined in isolation. This means that the sequence of words is not considered (cf. Supriyono et al. 2024). The vectors resulting from the

Word	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
an	0	0	0	1	0
announced	0	1	0	0	0
buildings	0	0	0	0	1
caused	0	0	0	0	1
collapse	0	0	0	0	1
demonstrated	0	0	1	0	0
demonstration	1	1	0	0	0
earthquake	0	0	0	1	1
for	0	0	1	0	0
germany	0	0	0	1	0
held	1	0	0	0	0
hours	0	0	1	0	0
in	1	0	0	0	0
last	0	1	0	0	0
paris	1	0	0	0	0
participants	0	0	1	0	0
several	0	0	0	0	1
struck	0	0	0	1	0
the	1	1	1	0	1
to	0	0	0	0	1
two	0	0	1	0	0
was	1	1	0	0	0
week	0	1	0	0	0

Table 1: Bag-of-words example output

bag-of-words approach can be used by a model and thus as a representation of language. The frequencies of words constitute features for classification.

An example of a classifier that works under the bag-of-words assumption is the multinominal naive Bayes classifier. The naive Bayes algorithm was first introduced by Maron (1961). The original goal of the model was the classification of documents according to their content.

Multinominal naive Bayes classifiers are probabilistic classifiers. The goal of the classifier is to give back for a document d, the class with the maximum posterior probability of all classes, $\hat{c}$.

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) \tag{1}$$

For text classification, the probability of a class is determined by the product of two values: the prior probability of the class $P(c)$ and the likelihood of the

document $P(d|c)$. The most probable class $\hat{c}$ given some document d is the one that holds the highest product of these values. The following equation arises.

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(d|c)P(c) \tag{2}$$

The prior probability of the class and the likelihood of the document need to be ascertained. For the prior probability of class c, the number of documents annotated with the class in the entire database is considered. If the entire database holds for example five documents, as is the case for the minimal example above, three annotated for class 1 and two annotated for class 0, the prior probability of class 1 would be $P(1) = \frac{3}{5}$. To understand the intuition behind the value that eventually reflects the likelihood of a document, it has to be considered that a document d can be depicted as a set of features $f_1, f_2, ..., f_n$. These feature sets can be represented by the vectors resulting from the bag-of-words approach. Accordingly, a feature can be defined as the frequency of a word in a document.

For each feature, i.e., word, the probability $P(f_i|c)$ is calculated in view of how often it occurs in a specific class in relation to how many words are included in said class and how many word types can be found in the entire database. For calculating $P(f_1, f_2, ..., f_n|c)$, it is assumed that "the probabilities $P(f_i|c)$ are independent given the class c and can be 'naively' multiplied" (Jurafsky and Martin 2024: 63). This is called the naive Bayes assumption. These considerations finally arrive at the following equation which specifies the output class selected by a naive Bayes classifier:

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{f \in F} P(f|c) \tag{3}$$

Implementing text classification with the naive Bayes model and bag-of-words representation of language has several drawbacks. First, the language representations are build exclusively on word frequencies. As mentioned above,

word positions hold no importance. This means that no contextual information is encoded. Second, the naive Bayes classifier assumes conditional independence of words given the class. This, however, disregards relationships between specific words which might be helpful for classification.

Still, approaching text classification in this way has the advantage that the input (in form of word frequencies) as well as the output of the model (in form of probabilities) is highly interpretable.

2.1.2 Term Frequency-Inverse Document Frequency

Similar to the bag-of-words approach but with a decisive advantage that will become clear shortly, is the term frequency-inverse document frequency (tf-idf) representation. Tf-idf focuses on the co-occurrences of words with documents. The goal of tf-idf is to weight terms in view of their relevance instead of using raw frequencies of occurrences, as was seen for the bag-of-words approach. It builds on the following assumptions: "Words that occur nearby frequently [...] are more important than words that only appear once or twice. Yet words that are too frequent [...] are unimportant" (Jurafsky and Martin 2024: 116).

The tf-idf is constructed from the term frequency and the inverse document frequency. The term frequency, the frequency of word t in document d, is calculated as follows:

$$\mathrm{tf}_{t,d} = \mathrm{count}(t,d) \tag{4}$$

For calculating the inverse document frequency, the number of documents in the database, N, as well as the number of documents which include the term t, df_t, are considered:

$$\mathrm{idf}_t = \frac{N}{\mathrm{df}_t} \tag{5}$$

Both term frequency and inverse document frequency are squashed with a

log function. The final weighted tf-idf value that is assigned to a word t in document d, $w_{t,d}$, results from:

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t \tag{6}$$

The example documents that were introduced for the bag-of-words method can be used again to show the difference between the two approaches. The output of the tf-idf approach is included in table 2.

Word	**Doc 1**	**Doc 2**	**Doc 3**	**Doc 4**	**Doc 5**
an	0	0	0	0.52335825	0
announced	0	0.46528078	0	0	0
buildings	0	0	0	0	0.40933049
caused	0	0	0	0	0.40933049
collapse	0	0	0	0	0.40933049
demonstrated	0	0	0.43366097	0	0
demonstration	0.3753856	0.3753856	0	0	0
earthquake	0	0	0	0.42224214	0.33024526
for	0	0	0.43366097	0	0
germany	0	0	0	0.52335825	0
held	0.46528078	0	0	0	0
hours	0	0	0.43366097	0	0
in	0.46528078	0	0	0	0
last	0	0.46528078	0	0	0
paris	0.46528078	0	0	0	0
participants	0	0	0.43366097	0	0
several	0	0	0	0	0.40933049
struck	0	0	0	0.52335825	0
the	0.26213107	0.26213107	0.24431703	0	0.23060966
to	0	0	0	0	0.40933049
two	0	0	0.43366097	0	0
was	0.3753856	0.3753856	0	0	0
week	0	0.46528078	0	0	0

Table 2: Tf-idf example output

The main advantage of the tf-idf approach becomes clear. It can be seen that frequent words in the vocabulary, like *the*, are denoted with smaller values, thereby mitigating their importance. Words of the vocabulary that do not appear in a given document are represented with zeros. Words that are highly characterizing for the document, however, are denoted with higher values. For the sentences that are processed here, examples of such words are *Germany* and *struck* in document 4. They are denoted with higher values. In contrast to

the bag-of-words approach, the tf-idf method results in more meaningful representations of language as it does not use raw frequencies but also considers the distribution of terms in the entire database.

To summarize, both bag-of-words and tf-idf are frequency based methods. Vectors created according to these methods are characterized as sparse as they include many zeros. Sparsity in this case is the result of using the vocabulary size as dimensions for the vector representation. This makes the vectors high-dimensional. Using sparse vectors is computationally expensive.

A problem with both representations is that out of vocabulary words, i.e., words that were not seen during the construction of the vectors, can not be encoded. This means that the classification of texts that include such words is problematic.

Due to the limited linguistic information that is conveyed by the bag-of-words and tf-idf representations of texts as such, traditional classification methods use manually constructed linguistic features for model training. These features are meant to support classification. They are highly dependent on the given task. This means that for different tasks, different sets of features have to be created. This need becomes clear when considering for example the task of sentiment classification in contrast to the task of spam detection: For sentiment classification one might want to look at the numbers of positive and negative words in a text, while spam detection might focus on the occurrences of specific phrases like one hundred percent guaranteed (cf. Jurafsky and Martin 2024: 68).

For the classification of the example documents described in the beginning of this chapter, a feature set that includes event type specific words would have to be constructed. After considering the database, the words *demonstration* and *demonstrated* could be identified as indicative for a demonstration event, *earthquake* and *struck* for an earthquake event. This, however, assumes that specific words are present in all texts that belong to the individual event types.

The method of creating relevant features in this ways is called feature en-

gineering. Feature engineering has the advantage that the linguistic indicators that are focused on for specific classification tasks are interpretable for and controlled by the researchers. However, features are not universal and have to be constructed anew for each task.

2.1.3 Static Word Embeddings

The problem of sparsity seen for bag-of-words and tf-idf vectors can be solved by focusing on the creation of short and dense vectors, so-called embeddings. The main differences to the bag-of-words and tf-idf vectors are that embeddings (1) have a fixed number of dimensions and (2) include real-valued numbers that can also be negative. Vector size therefore no longer corresponds to the vocabulary size.

Representations of the meaning of words are learned automatically in the case of embeddings. The creation of word embeddings is an example of so-called representation learning which "constructs low-dimensional representations to summarize essential features of high-dimensional data" (Wang and Jordan 2021: 1). For the static embeddings that are described here, one representation is learned for each word type in the vocabulary.

One way to learn these fixed embeddings is the skip-gram with negative sampling method. The skip-gram method can be formulated as a binary classification task. The aim of the task is to determine if, given a target word w and a candidate context word c, the candidate context word is a real context word of the target word. This task is approached in an unsupervised manner, meaning that the data that is used for training is not annotated manually. Rather, a word and its neighbor are treated as a positive sample, i.e., target word and real context words, while the output of a random sampling of words in the vocabulary is treated as a negative sample, i.e., target word and not real context word. To create the embeddings, the logistic regression algorithm is used to solve the binary classification task. The general approach of logistic regression is to learn

> from a training set, a vector of weights [**w**] and a bias term [b]. Each weight w_i is a real number, and is associated with one of the input features x_i. The weight w_i represents how important that input feature is to the classification decision (Jurafsky and Martin (2024): 83).

The weights that are learned during this process will eventually constitute the embeddings. In the first step of the training process, each word is assigned a random embedding vector. These random vectors are then iteratively shifted in such a way that embeddings of words that appear close to each other in texts are similar to each other. Embeddings of words that do not co-occur are less similar. Embeddings can thus be seen to work under the assumption that "[y]ou shall know a word by the company it keeps" (Firth 1962: 11).

A simplified illustration that is often used to explain the intuition behind word embeddings and the vector space is shown in figure 1. In reality, word embeddings have a much higher number of dimensions than is depicted here. Most often the number of dimensions is 300.

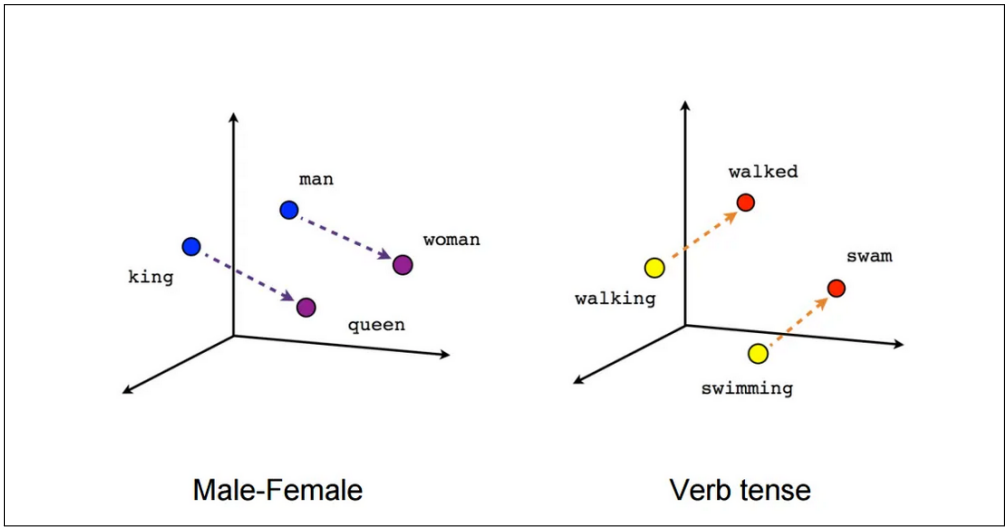

Figure 1: Vector space example

On the one hand, the figure shows the semantic relationship of the male-female designation, on the other hand, the syntactic relationship of verb tenses is depicted (cf. Liebman 2018).

By representing language in this way, the possibility of calculating the similarity between embeddings and therefore between words arises. The cosine similarity can be used for this:

$$\text{cosine}(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} \tag{7}$$

The cosine similarity can attain values between -1 and +1.

Words included in the sentences that were introduced in the beginning of this chapter are used as a database to illustrate the calculation of word similarities. A similarity of 0.460 for the related words *demonstration* and *demonstrated* and a similarity of 0.267 for *earthquake* and *collapse* are found. The similarity scores of unrelated words are consequently lower, e.g., 0.127 for *demonstration* and *earthquake* and -0.033 for *demonstrated* and *collapse*.

For the task of text classification, word embeddings are oftentimes utilized in combination with so-called neural networks. Neural networks are machine learning models that are capable of processing raw inputs and automatically learn relevant features during the training process. The basic architecture of a simplified two-layer neural network is shown in figure 2.

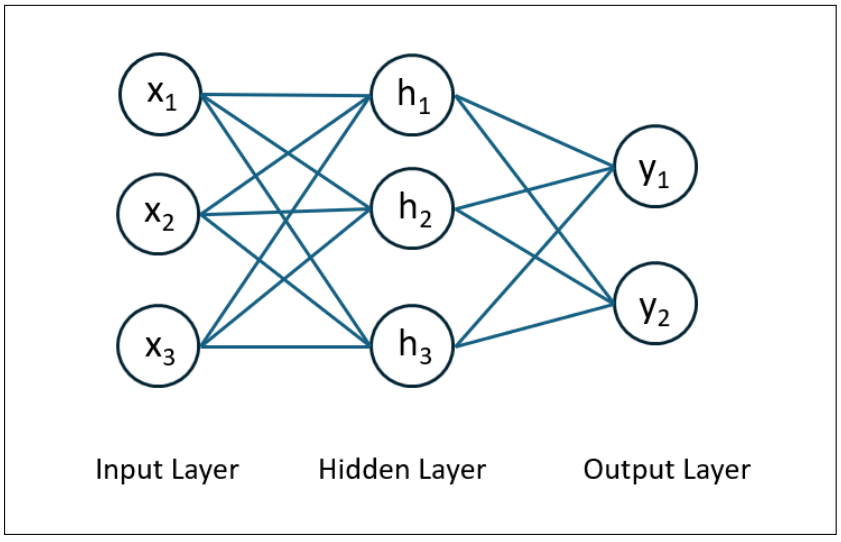

Figure 2: Two-layer neural network

Neural networks consist of an input layer, hidden layer(s) and an output layer. To use the just described word embeddings in combination with a neural network, the textual database has to be preprocessed. This preprocessing maps words included in the database to the corresponding vector representations. An embedding layer results from this which then constitutes the input layer

of a neural network. Embeddings therefore form the basis for processing language automatically and solving a given NLP task.

At the heart of a neural network are the hidden layers **h**. Neural networks typically have more hidden layers than represented in figure 2. The hidden layers are made up of hidden units $\mathbf{h}_i$. A singular neural unit is taking "a weighted sum [z] of its inputs, with one additional term in the sum called a bias term" (Jurafsky and Martin 2024: 137). This means that a unit has a set of inputs $x_1 ... x_n$, a set of weights $w_1 ... w_n$ that correspond to these inputs and a bias b. The set of inputs and the set of weights are vectors. z is then represented as:

$$z = \mathbf{w} \cdot \mathbf{x} + b \tag{8}$$

The output of a neural unit is obtained by applying a non-linear activation function f to z. This results in what is called the activation of the unit, a.

In the simplest form of a neural network, a feedforward neural network, outputs from the individual units contained in the layers are forwarded through the network. They are passed to units in the next layer. Other architectures, for example recurrent neural networks (RNNs), additionally allow for the outputs to be passed back.

The final layer in a neural network is the output layer. The number of nodes included in the output layer can correspond to the number of classes that should be classified. If the task aims at the classification of for example four distinct classes, four nodes are included in the final layer. The output layer of a neural network computes the network's final output by considering the representations supplied by the output of the hidden layer. The output layer of the network is a vector **y**. This vector supplies the results of the training process by means of probabilities that are assigned to the output nodes. This means that if four nodes are given in the output layer and each node corresponds to a class, each node and therefore each class receives an output

value that constitutes the probability of that class.

During training of a neural network, the weights and bias for each layer are learned in such a way that the prediction of the model, $\hat{y}$, is as close as possible to the ground truth y. The distance between $\hat{y}$ and y is determined using a so-called loss function and forms the basis for learning.

In chapter 2.1, the notion of feature engineering, i.e., manually creating a set of features that is meaningful for a specific task, was introduced. The use of word embeddings and neural networks can replace manual feature engineering. It is no longer necessary to determine specific words or phrases that are indicative for a certain class one wants to classify. Embeddings open up the possibility of using word similarity to facilitate classification. Neural networks utilize representation learning in the same manner as was already mentioned for word embeddings. Neural networks can draw upon existing embeddings, as was described here, or learn new embeddings during training. Learning representations in earlier layers of the network and using them in later layers makes such models powerful. Neural networks can be adapted to a wide range of tasks.

For the fictitious minimal example of event type classification that was introduced in the beginning of this chapter this means that the need to determine event type specific words as a feature for classification is made redundant. The semantic similarity between the input tokens is used to solve the task. Next to the words *demonstration* and *demonstrated*, which were present in the corpus and indicative for the presence of a demonstration event, word embeddings would also consider related words for a classification decision, for example *protested*. The cosine similarity of the word2vec embeddings of *protested* and *demonstration* is 0.341, the similarity of *protested* and *demonstrated* is 0.338.

In contrast to using feature engineering and traditional classifiers, decisions made by neural networks are more opaque. It therefore becomes more and more difficult for researchers to interpret how a model reached its output.

Word embeddings can be seen as a breakthrough in the way of how lan-

guage can be processed automatically. Still, they come with some restrictions. As mentioned in the beginning of this chapter, the word embeddings described here are static. There is one vector representation for each word type. This becomes problematic when encountering, for example, polysemes and homonyms as they are not disambiguated and are represented with the same vector.

2.2 Interim Summary

In the last chapters, the evolution of language representations typically used for NLP applications before the introduction of LLMs was described.

It was shown that traditional language representations were of a frequency-based nature. Manually created features were typically used to solve NLP tasks. Each task required the creation of a new set of relevant features. The introduction of static word embeddings and neural networks opened up the opportunity to solve these tasks without this manual effort. Word embeddings were no longer restricted in their applicability and could be used independent of the nature of the task: be it sentiment classification, spam detection or event type classification; word embeddings were drawn upon as a representation of language. The combination of embeddings and the possibility of learning feature interactions over many layers in a neural network made this approach powerful and relevant for NLP.

While models became more and more efficient to use, they also became more and more opaque. When creating features manually, the researcher is able to decide what should be considered as a feature for the model's decision. This ensures that all features are actually relevant to the task. For neural networks, and as will become clear also for transformers, it is not always known which features are used by a model to reach a decision. It is therefore possible that "models exploit[...] *spurious correlations*, or *shortcuts* between the training data and the task label" (Wang et al. 2022: 1719). This means that tokens that are not really indicative for a class but statistically relevant in the

training data are focused on which can lead to wrong predictions and biases for the classification of unseen data. In the context of event type classification, such a spurious correlation could for example be present for an event type and a specific location. This means that if an event predominately occurs in one location, this location name will be reported frequently in connection to it. The relationship between the event type and the location is detected in the data and wrongly learned to be indicative for the classification of the event type. If the location name occurs in texts belonging to other event types, classification might be negatively influenced by of it.

2.3 Transformers and Contextual Embeddings

The architecture underlying large language models that are commonly used for solving NLP tasks today, is the transformer architecture. Transformers are a neural architecture that support the processing of distant information.

The first transformer model was introduced by Vaswani et al. (2017) and focused on the task of machine translation. Vaswani et al. reported that their proposed architecture outperforms other ML models and can additionally be trained faster. The transformer described in their paper is based on the encoder-decoder architecture and utilizes self-attention. Broadly speaking, in this architecture the so-called encoder is responsible for converting a sequence of tokens into a sequence of embeddings (which is called the hidden state), the decoder generates an output sequence from this hidden state (cf. Tunstall et al. 2022). By means of the encoder and the decoder, the task of machine translation is realized.

Today's LLM landscape consists of not only encoder-decoder architectures but also of encoder-only and decoder-only ones. Two encoder-only transformer models, **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT, Devlin et al. 2018) and **R**obustly **o**ptimized **BERT a**pproach (RoBERTa, Liu et al. 2019), will be introduced in more detail later on as they are the models that were used for the classification tasks presented in this work. Well

known decoder-only models are the generative pre-trained transformer (GPT) models (Radford et al. 2018). They are not used in this work and will thus not be considered in more detail.

Neural architectures that work with encoders and decoders were already in use before the introduction of the transformer architecture. The main difference that made transformers successful is the usage of so-called self-attention as the only means of computing representations of the model's inputs and outputs (cf. Vaswani et al. 2017). The self-attention mechanism is used to "build contextual representations of a word's meaning that integrate information from surrounding words [across a series of layers], helping the model learn how words relate to each other over large spans of text" (Jurafsky and Martin 2024: 214).

Self-attention can be backward-looking, which means that it processes context left-to-right. In this case, only previous tokens are considered. Transformer models that work like this use so-called causal self-attention and can thus be called causal (left-to-right) transformers. In contrast to those causal transformers are models like BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019). They are based on the so-called bidirectional transformer encoder architecture. Bidirectional transformers differ from the just described causal transformers insofar that they are not exclusively backward-looking/left-to-right processing. This means that in the case of bidirectional encoders, the self-attention mechanism considers the entire context, therefore ranging over tokens not only to the left but also to the right. This includes future tokens.

Ultimately, by using self-attention the main advantage of transformer models is achieved which are "contextualized representations of the meanings of input words or tokens" (Jurafsky and Martin 2024: 215). These rich contextualized representations of language are constructed by a process called pre-training.

2.3.1 Pre-training of Transformers

As mentioned in the beginning of this chapter, the transformer architecture typically underlies LLMs. In its essence, an LLM is a pre-trained transformer model. During the process of pre-training, knowledge about language and the world as such is acquired. Pre-training is based on large amounts of texts. BERT for example was pre-trained on about 3.3 billion words. Pre-training can be implemented in several ways. In the following section, the process that was used to pre-train BERT and RoBERTa will be detailed.

The pre-training of BERT draws on the tasks of

1. Masked Language Modeling (MLM) and
2. Next Sentence Prediction (NSP).

For the task of MLM, Devlin et al. (2018) randomly mask 15% of all tokens in their experiments. The process of masking is done by either substituting a chosen token by the generic [MASK] token or by substituting it with another token that was randomly chosen. The model's goal is to predict what the masked token originally was.

The second pre-training objective, NSP, is formulated as a binary task. This means the possible outputs are 0/no and 1/yes. Each sample used for NSP consists of two sentences, sentence A and sentence B. The sentence pairs are generated from the training corpus. In 50% of all sentence pairs, sentence B actually follows sentence A. The pair is labeled accordingly. The remaining 50% of sentence pairs include sentences that do not follow each other in the corpus. Sentence B is chosen at random (cf. Devlin et al. 2018). The goal during pre-training is to correctly predict if sentence B follows sentence A.

RoBERTa builds on BERT's proposed methodology but modifies the tasks for pre-training. For MLM, the authors adapt dynamic masking. For the pre-training of BERT, the masking of tokens for MLM was implemented one time during the preprocessing of the data. For RoBERTa, masking is implemented

numerous times, namely every time a sequence is fed to the model (cf. Liu et al. 2019). With regard to NSP, Liu et al. (2019) find that removing the task actually improves performance slightly in contrast to BERT. Thus, NSP is not used for pre-training RoBERTa. Using these modifications in pre-training along with changes in the hyperparameters (among others: batch size, longer training, longer sequences), leads to RoBERTa outperforming BERT in numerous benchmark tasks.

Pre-trained transformer models supply contextual embeddings. Chapter 2.1.3 introduced static embeddings where one embedding is learned for each word type in a given vocabulary. Contextual embeddings do not represent the meaning of word types but word instances. This means that individual words are represented by considering the context that they appear in during pre-training. Contextual embeddings therefore grasp word meaning much more accurately than static embeddings which is essential for solving NLP tasks. Problems that were present for static embeddings, e.g. the processing of polysemes and homonyms, are dissolved. To make the difference between static and contextual embeddings even clearer, consider the following sentences as an example input:

> The participants **demonstrated** for two hours.
>
> She **demonstrated** her expertise during the presentation.

For a human reader, the different meanings of *demonstrated* are obvious. However, static embeddings were not able to encode this difference and *demonstrated* would be represented by the same vector in both cases. When processing large amounts of text, words are bound to appear in varying contexts. The possibility to factor in this variation makes contextual embeddings so powerful. Contextual embeddings advanced the automatic processing of natural language notably. Consequently, BERT and RoBERTa, and thus contextual embeddings, are used for the work at hand.

2.3.2 Fine-Tuning of Transformers

Pre-trained transformers are the models of choice for a wide range of tasks nowadays. To optimize them for a specific use case, for example event type classification, pre-trained models have to undergo the process of fine-tuning. Fine-tuning refers to "creat[ing] applications on top of pre-trained models by adding a small set of application specific parameters" (Jurafsky and Martin 2024: 254). This means that while the pre-trained model encodes general (deep) knowledge about language, the process of fine-tuning adapts this knowledge to a concrete task. This is done by adding a classification layer (classification head) on top of the pre-trained model body. The parameters of the pre-trained transformer model itself are, at most, changed minimally during this process.

Fine-tuning can be seen as an example of transfer learning. Transfer learning is "the method of acquiring knowledge from one task or domain, and then applying it (transferring it) to solve a new task" (Jurafsky and Martin 2024: 242). Fine-tuning is achieved through a training process. In the case of supervised learning, as is used in this work, this means that the pre-trained model is supplied with an annotated database and then fitted to this specific database in order to solve the corresponding task. Model training follows the principles that were already described for neural networks in chapter 2.1.3.

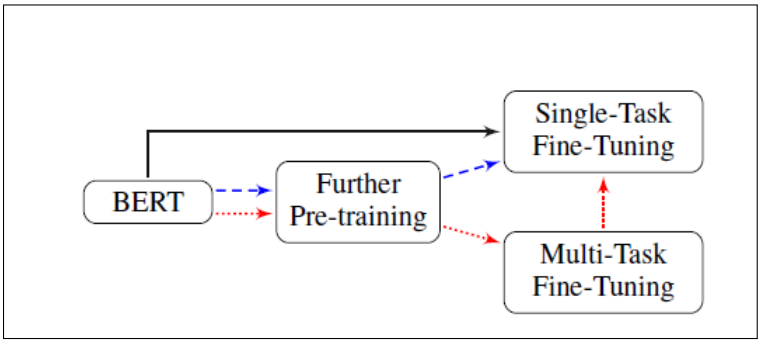

Figure 3: Fine-tuning strategies

A number of different fine-tuning strategies are available. This work implements single-task fine-tuning (cf. Sun et al. 2019). The model is optimized

solely for the task of event type classification. Further pre-training is not implemented. Therefore, the black arrow in figure 3 represents the chosen method.

3 Event Extraction

The automatic text classification task that this thesis aims to process with the help of transformer-based LLMs is event type classification. The classification of event types is a subtask of the broader task of EE. The following chapter will at first consider the definition of events in the sense of NLP. After that, the subtasks that are part of EE will be described along with the methods used to solve them.

3.1 (Task) Definition in NLP

As mentioned before, EE as such is described as "a crucial research task for promptly apprehending event information from massive textual data" (Li et al. 2021a). An event in the context of NLP is defined as

> a particular form of information [...] [that] refers to a specific occurrence of something that happens in a certain time and a certain place involving one or more participants, which can usually be described as a change of state (Li et al. 2021a).

Following Li et al., EE can be divided into four subtasks: event trigger identification, event trigger classification, event argument identification and event argument role classification (cf. 2021a: 3). The Linguistic Data Consortium defines an event trigger as follows:

> An Event's Trigger is the word (in its scope) that most clearly expresses its occurrence. In many cases, this will merely be the main verb in the part of the sentence [...] that most directly describes the Event (Linguistic Data Consortium 2005: 7).

The subtask of trigger identification therefore aims at determining which word in a given sentence introduces the event. In the example below, taken from the event annotation guidelines (Linguistic Data Consortium 2005: 7), the main verb die is indicative for the event.

(1) John **died** yesterday of renal failure.

While it is stated that an event trigger is most often represented by the main verb, the Linguistic Data Consortium also mentions the possibility of the trigger being represented by an adjective or a noun (cf. 2005: 7-8).

The classification of the event trigger, which follows its identification, is a multiclass classification task. The event trigger is classified according to its event type. Event types are commonly predefined, meaning they are fixed. Event type classification is a supervised classification task as described in chapter 2. This means that given an input text x and a predefined set of event types $Y = \{y_1, y_2, ..., y_M\}$, a predicted event type $y \in Y$ is returned. For the example above, the Automatic Content Extraction (ACE) Program's predefined main event type LIFE and the event subtype DIE would be expected as a classification result.

Which main and sub event types need to be considered depends on the database one is working with. The ACE dataset, for example, includes eight main event types (LIFE, MOVEMENT, TRANSACTION, BUSINESS, CONFLICT, CONTACT, PERSONELL, JUSTICE) each with $1-n$ subtypes. Other datasets are more focused with regard to the types they consider. The Armed Conflict Location and Event Data Project (ACLED) (Raleigh et al. 2010), for instance, focuses on crisis events exclusively and the dataset provided by Davani et al. (2019) only takes hate crime events into consideration.

Each event type holds a number of unique event arguments. Event arguments can be broadly split into event participants and event attributes (cf. Linguistic Data Consortium 2005: 48-49). Event participants are defined as "[e]ntities that are somehow involved in the Event" (Linguistic Data Consortium 2005: 48). For the sake of simplicity, consider again the DIE event example given earlier. The event participant in this case would be John. John is a named entity of type *person*.

Event attributes are divided into event-specific and general event attributes. Event-specific attributes are dependent on the event type and can be manifold, while general event attributes hold true for all event types. General event at-

tributes are event time (when the event takes place) and event location (where the event takes place) (cf. Linguistic Data Consortium 2005: 48-49).

The subtask of event argument identification consequently has the goal of identifying all arguments of an event. Event argument role classification then assigns a role to the identified arguments. To make this approach clear, a more complex example of a DIE event, taken from Li et al. (2021a), is presented below:

(2) Baghdad, a cameraman **died** when an American tank fired on the Palestine Hotel.

According to the Linguistic Data Consortium (2005) event annotation guidelines, event argument identification is supposed to find *Baghdad*, *cameraman* as well as *American tank* as arguments. In the next step of role classification, these arguments should then be classified as shown in table 3.

Event argument	Argument role
Baghdad	Place
cameraman	Victim
American tank	Instrument

Table 3: Event arguments and event argument roles

The four subtasks of EE can be processed sequentially in a pipeline as described, among others, in Krumbiegel et al. (2021). However, each subtask can also be worked on in isolation when a suitable database is available.

3.2 Document and Sentence-Level Event Extraction

EE can be implemented on the document and on the sentence level. As the task names suggest, sentence-level event extraction (SEE) focuses on processing individual sentences. Depending on the subtask that is worked on, all relevant annotations are given for each sentence, i.e., event trigger, event type, event arguments and event argument roles.

Document-level event extraction (DEE) considers documents that include more than one sentence. The term document in the context of EE can refer to, among others, Wikipedia articles (Li et al. 2021b), online newspaper articles (Tong et al. 2022) and Tweets (Petrovic et al. 2021). In contrast to SEE, event arguments can not all be found in a single sentence, rather, they are scattered over the document. This makes identifying all arguments challenging (cf. Li et al. 2021a). Additionally, documents can include more than one event depending on the database, which makes event type classification on the document level demanding.

3.3 Work on EE in NLP: Common Approaches

EE has been a task of interest in the research field of NLP for many years. How it was approached depended strongly on available models and language representations. Previous work on EE therefore step-by-step draws upon methods that were described in chapter 2. In the early 2000s and thus before word embeddings and transformer models were introduced, the focus was on feature engineering. Ahn (2006), for instance, manually creates sets of features for each EE subtask. For event trigger identification, lexical features like Part-of-Speech (POS) tags of words are considered. The reasoning behind this is that an event trigger can only belong to a limited subset of POS tags. Identifying relevant POS tags therefore supports the task of trigger identification in this case. For event argument extraction, information on entity types (e.g., person, organization) is included.

As already mentioned in chapter 2.1, manual feature engineering is said to be time consuming. Accordingly, with the introduction of embeddings and neural networks, feature engineering was widely laid to rest in the context of EE.

After the introduction of word embeddings, Chen et al. criticize feature engineering for EE to belong to a number of "traditional approaches [that] lack generalization, take a large amount of human effort and are prone to er-

ror propagation and data sparsity problems" (2015: 167). Instead, they use automatically learned features based on word embeddings and convolutional neural networks (CNNs) (cf. Chen et al. 2015: 167). Liu et al. (2016) also employ word embeddings for their work, however, they additionally consider FrameNet (Ruppenhofer et al. 2010: 5) as a linguistic resource to further support automatic event detection. FrameNet is a database that provides semantic frames for the English language. They argue that

> many frames in [FrameNet] actually express certain types of events. The above observations motivate us to explore whether there exists a good mapping from frames to event-types and if it is possible to improve event detection by using [FrameNet] (Liu et al. 2016: 2134).

They find that FrameNet as a resource does in fact support event detection (cf. Liu et al. 2016: 2142).

Given that it was state of the art in this period of time, numerous other works on EE also work with some kind of combination of word embeddings and neural networks (cf. among others, Nguyen and Grishman 2016, Feng et al. 2016, Sha et al. 2018).

The next evolutionary step in how language is processed in NLP, namely, with the help of transformer models and contextual embeddings (especially important here is the introduction of BERT (Devlin et al. 2018)), evoked another change in the landscape of EE research. Shortly after the introduction of BERT in 2018, Yang et al. exploit the newly available contextual embeddings to build an event extractor based on the features provided by BERT (cf. 2019: 5285). Also focusing on BERT's main advantage of capturing context well, Wadden et al. develop "a unified, multi-task framework for three information extraction tasks: named entity recognition, relation extraction, and event extraction" (2019: 5784). Both papers report improvements in scores when using BERT-based implementations.

With regard to transformer-based approaches to event extraction in general, Li et al. note that

> the method of identifying event types based on the full text has become

> mainstream. It is because BERT has outstanding contextual representation ability and performs well in text classification tasks (2021a: 9).

The consequence of this is that LLMs like BERT and RoBERTa are trusted in their decisions to classify event types correctly and according to relevant tokens in a given input text, even if their predictions are not interpretable for the researcher. Specific linguistic features that are supposed to be indicative for the differentiation of event types, for example, are no longer encoded explicitly in the form of features to support classification.

To gain an overview of approaches that are commonly used for EE in the context of NLP conferences today, the proceedings of the Workshop on the Future of Event Detection (FuturED) (Tetreault et al. 2024) as well as the proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) (Hürriyetoğlu et al. 2024a) are consulted. FuturED took place in November of 2024 and was co-located with the Empirical Methods in Natural Language Processing (EMNLP) conference. The CASE workshop was held in March of 2024 in connection to the Conference of the European Chapter of the Association for Computational Linguistics (EACL).

The organizers of FuturED state that

> a wide range of topics [is covered by the submitted papers]—from specialized [event detection] methods to broader discussions about the field's progress, from text-only data to multimodal approaches, and from static learning scenarios to dynamic social network data streams (Tetreault et al. 2024: ii).

The CASE workshop organizers report a "diverse range of papers and shared tasks presented at the workshop, from multimodal data analysis to fine-tuning language models for detecting hate speech and extracting event timelines [...]." (Hürriyetoğlu et al. 2024b: 252). The short summaries by the organizing teams show similarities in the content of the two workshops. When having a closer look at the individual papers included in the proceedings, trends arise with regard to model choices. On the one hand, it can be seen

that LLMs based on BERT are still dominating (cf. Loerakker et al. 2024, Tanev 2024, Uludoğan et al. 2024, Farsi et al. 2024, Boutaleb et al. 2024, Ray et al. 2024. On the other hand, generative models, such as models of the GPT family (Radford et al. 2018), seem to gain more prominence in EE research (cf. Fellman et al. 2024, Bakker et al. 2024).

While this short summary of the FuturED and CASE workshop does not assume to be completely generalizable to all research in EE, it shows that transformer models belonging to the BERT family still can be considered as state of the art for EE. Additionally, EE research also draws upon generative approaches. This is in accordance with the general rise of the use of generative models for NLP tasks.

3.4 Summary

The last chapters described the task of EE. The four subtasks of EE, event trigger identification and classification as well as event argument identification and classification, were introduced. The task that is considered in this work, event type classification, corresponds to the task of event trigger classification. It was shown that all subtasks can be implemented on the sentence- and on the document-level. The processing of whole documents is more demanding than focusing on sentences. This is due to relevant information being spread over the entire text in the case of document-level EE.

The literature review of approaches that are commonly used to solve tasks connected to EE revealed that before the widespread use of neural networks and transformer models the tasks were processed with a strong focus on linguistic indicators which were explicitly coded. With the introduction of these architectures, however, the usage of explicit linguistic features increasingly came under criticism. It was easier and more time efficient to rely on available representations of language (static as well as contextual embeddings) than creating complex sets of linguistic features.

As a result, the task of event type classification changed to an extent. The

identification of the event trigger, which was a necessary step to enable the classification of the event type, is no longer required. Nowadays, the event type is determined based on the entire input text. Features that facilitate the classification are learned automatically. BERT-based models are drawn upon most often when the tasks of EE are approached in a discriminative manner. In addition, generative approaches are becoming increasingly relevant for the task. It can be assumed that the adoption of generative strategies will evoke another change in how the task is processed in the future.

4 Methodology

In the following chapter, the general experimental set up that is used for the classification of event types in the remainder of this work is introduced. The chapter starts with a short introduction of models and libraries that are used for training the classification models. This is followed by a description of the event type data. The chapter ends with remarks on methods that were utilized for evaluation.

4.1 Implementation Libraries

All model training was implemented with the Simple Transformers library (Rajapakse et al. 2024). The Simple Transformers library supports all steps that are necessary for executing NLP tasks (cf. Rajapakse et al. 2024: 210). These steps are:

1. choosing a suitable model for the task at hand,
2. preprocessing the training data,
3. converting the text into input tokens that can be processed by a transformer (tokenization),
4. training the transformer on new data starting from the pre-trained weights (fine-tuning), and
5. testing the model on unseen data.

The library was chosen due to its general ease of use and, most importantly, its connection to the Hugging Face model hub [1]. Through the Hugging Face model hub, a large number of open-source transformer models are made available which can be directly accessed with the Simple Transformers library.

Basic linguistic processing and analyses were conducted with the help of spaCy (Honnibal et al. 2020). SpaCy is a Python library for NLP that offers

[1] https://huggingface.co/models

support for more than 70 languages (cf. Honnibal et al. 2020). The library provides a number of linguistic processing tools for the English language, those are, among others, Part-of-Speech (POS) Tagging, Lemmatization, Dependency Parsing and Named Entity Recognition (NER).

4.2 Data

A sufficiently large dataset is needed for the task of event type classification. Event type classification is processed as a supervised classification task. Therefore, the dataset additionally needs to be annotated with event type labels. Due to the required size of the corpus, manual annotation was not realizable. Instead, open source datasets available through Hugging Face's dataset repository[2] were considered.

The Document-level Event Extraction (DocEE) corpus (Tong et al. 2022) was identified as a suitable database for this work. The corpus was introduced by Tong et al. as a "Large-Scale and Fine-grained Benchmark for Document-level Event Extraction" (2022: 3970). In their paper, the authors describe the intended EE tasks that can be supported by DocEE: event type classification as well as event argument identification and classification.

The kinds of documents included in the corpus are twofold: On the one hand, a selection of Wikipedia articles describing historical events is found, on the other hand, news articles, obtained by consulting Wikipedia's *Portal:Current events* [3], are incorporated (Tong et al. 2022: 3973). In total, the corpus comprises 27485 annotated documents.

The authors state that the corpus is intended to support the classification of the main event included in a document (cf. Tong et al. 2022: 3970). Accordingly, every document is annotated with exactly one of the available event types. There are 59 event types in total. The DocEE corpus incorporates newspaper articles. Therefore, it is possible that one article reports on more

[2]https://huggingface.co/datasets
[3]https://en.wikipedia.org/wiki/Portal:Current_events

than one event. This is however disregarded in the annotation. There are no multi-label instances in the database. Figure 4 gives an overview of the distribution of the event types.

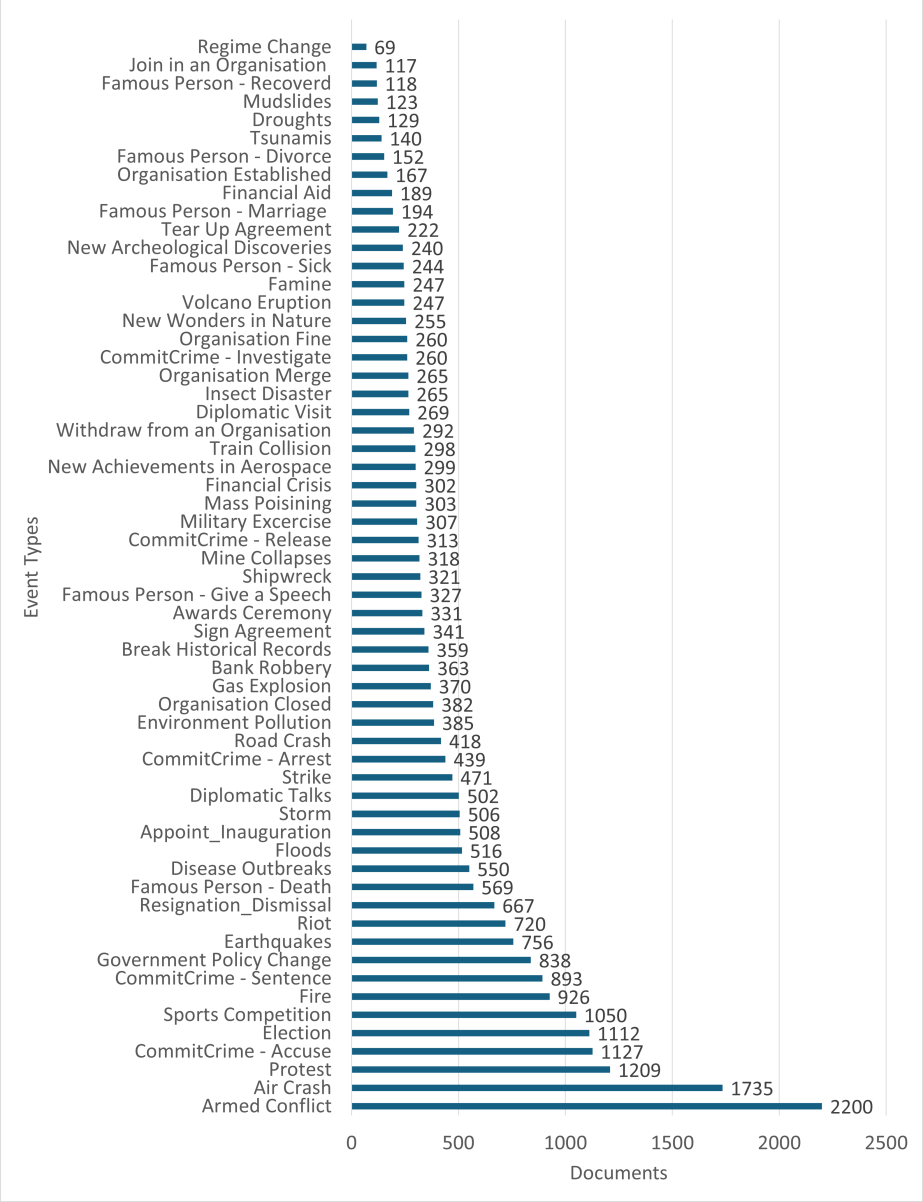

Figure 4: Distribution of event types in the DocEE corpus

Tong et al. (2022) differentiate between so-called hard and soft news when defining their event type annotation scheme. They base this differentiation on Lehman-Wilzig and Seletzky's (2010) framework. Following Lehman-Wilzig and Seletzky, soft news include:

- Reports on light or exotic topics and
- news that is of interest to a narrow segment of the public (2010: 45).

An additional characteristic of soft news is that due to their lack of actual significance there is no necessity to report them immediately or at all (Lehman-Wilzig and Seletzky 2010: 45). Event types included in the DocEE corpus that are examples of soft news are, among others, Famous Person – Divorce, Awards Ceremony and Sport Competitions

In contrast, the class of hard news include

- political, social, economic or serious environmental news of a highly significant nature,
- breaking, surprising events of great import on most of the public,
- new findings, discovery or reports regarding a continuing story of great significance and
- significant news on the national and the international plane (Lehman-Wilzig and Seletzky 2010: 46).

Hard news types included in the DocEE corpus are, for example, Tsunamis, Appoint_Inauguration and Financial Crisis. Tong et al. state that DocEE includes 31 event types that belong to the class of hard news and 28 types that can be seen as soft news (cf. 2022: 3972).

In addition to event type annotations, the authors provide event argument annotations. Each event type is assigned a distinct set of event arguments. In total, there are 356 argument types included in the corpus (cf. Tong et al. 2022: 3971). The complexity of the event argument annotation scheme is due to the author's decision to use the info boxes found on Wikipedia pages about events as the "prototype arguments of the event" (Tong et al. 2022: 3973). An overview of all possible event argument annotations per event type can be found in appendix A. Event arguments that are defined by the Linguistic Data Consortium as general event attributes are event time and event place (2005: 49). General attributes are expected to be given for all event types. In the case of the DocEE corpus, event time is included in the annotation scheme for all

event types. Event place is given for 49 of the 59 event types. Additionally, event participants are included for all event types. The specific labels that are used to denote the participants depend on the event type. In the case of the event type Commit Crime – Sentence, for example, participants that are involved in the event and therefore included in the event arguments are judge, victim as well as suspect. Most event arguments in the corpus can be seen as event-specific. This means that they only hold relevance for specific event types and can not be found for all. Examples for event-specific attributes are Maximum Wind Speed for the event type Storm and Marriage Duration for the event type Famous Person – Divorce.

Event types as well as event arguments were annotated manually with the help of crowdsourcing. The authors use Cohen's kappa coefficient to calculate the inter-annotator agreement. For the annotation of event types, the authors report an inter-annotator agreement of 94%, for the event arguments annotations an agreement of 81% (cf. Tong et al. 2022: 3974).

4.3 Data Preprocessing

The DocEE corpus was modified in several ways for the experiments conducted in this work. The 59 event types were summarized into eleven main event types to reduce the complexity of the annotation scheme. Additionally, a corpus consisting only of sentences that hold an event argument was created. Further, the corpus was analyzed with regard to the length of the individual samples and subsequently filtered to ensure correct processing by the chosen LLMs.

4.3.1 Construction of Main Event Types

In a first step, the event types included in DocEE were summarized into eleven main event type classes. In the remainder of this work, these newly constructed classes will be referred to as event types, while DocEE's original classes will be referred to as sub event types or simply subtypes.

1. Event type: **ACCIDENT**

    Sub event types:

    * Air Crash
    * Fire
    * Gas Explosion
    * Mine Collapses
    * Road Crash
    * Shipwreck
    * Train Collisions

2. Event type: **CRIME**

    Sub event types:

    * Bank Robbery
    * CommitCrime – Accuse
    * CommitCrime – Arrest
    * CommitCrime – Investigate
    * CommitCrime – Release
    * CommitCrime – Sentence
    * Mass Poisoning

3. Event type: **DIPLOMACY**

    Sub event types:

    * Diplomatic Talks
    * Diplomatic Visit
    * Sign Agreement
    * Tear Up Agreement

4. Event type: **DISCOVERY**

Sub event types:

* New Achievements in Aerospace
* New Archaeological Discoveries
* New Wonders in Nature
* Break Historical Records

5. Event type: **DISORDER**

Sub event types:

* Protest
* Riot
* Strike

6. Event type: **FINANCE**

Sub event types:

* Financial Aid
* Financial Crisis

7. Event type: **GOVERNMENT**

Sub event types:

* Appoint_Inauguration
* Election
* Government Policy Changes
* Regime Change
* Resignation_Dismissal

8. Event type: **LIFE**

Sub event types:

* Famous Person – Death
* Famous Person – Divorce

* Famous Person – Marriage
* Famous Person – Recovered
* Famous Person – Sick

9. Event type: **NATURAL DISASTER**

   Sub event types:

   * Disease Outbreaks
   * Droughts
   * Earthquakes
   * Environment Pollution
   * Famine
   * Floods
   * Insect Disaster
   * Mudslides
   * Storm
   * Tsunamis
   * Volcano Eruption

10. Event type: **ORGANISATION**

    Sub event types:

    * Join in an Organisation
    * Organisation Closed
    * Organisation Established
    * Organisation Fine
    * Organisation Merge
    * Withdraw from an Organisation

11. Event type: **WIN LOSE**

    Sub event types:

* Awards Ceremony
* Sports Competition

The event types were constructed by considering the general situations that can be assumed to be described by the given subtypes. Additionally, the differentiation between soft and hard news that was used by Tong et al. (2022) to build the annotation scheme was kept in mind. This means that when creating the main event types, care was taken to not mix subtypes belonging to the class of hard news with those belonging to the class of soft news. At this point, the actual content of the samples was not analyzed in detail to avoid subjective exclusion and inclusion of as well as a preferences for specific entries. Three event types that can be found in the original DocEE corpus are not considered further. Those are: Armed Conflict, Military Exercise and Famous Person – Give a speech.

4.3.2 Event Argument Corpus

Next to event type annotations, the DocEE corpus also includes event argument annotations. Sentences that are annotated with an event argument hold importance for the task of event type classification as they represent the most important information in the text. A corpus that only includes the texts of the event arguments was generated from the DocEE corpus. Sentences that hold an event argument were identified for each sample, concatenated and added to a new file. Below, an original sample taken from the DocEE corpus is given. Annotated event arguments are colored in blue. The event argument labels are not included in the original data but were added manually in parenthesis for better understanding.

> American agents operating inside Pakistan have been the cause of strained relations in the past and have fuelled anti-US feeling in the country. A spokeswoman for the US state department said he was on a temporary assignment with the US mission. US officials said the man had accidentally been carrying the ammunition in his luggage when he was detained on Monday night. "We are co-ordinating with Pakistani authorities to resolve this matter," she said. "We are working closely with them." The

> United States and Pakistan co-operate in the global fight against Islamist militants, angering many Pakistanis. An agent with the US FBI **[Event argument: police]** has been arrested under anti-terrorism **[Event argument: suspect]** laws in Pakistan for carrying ammunition while trying to board a flight **[Event argument: the charged crime]**. The US citizen **[Event argument: suspect]** was detained at Karachi airport after security staff found 15 bullets for a 9mm handgun in his luggage **[Event argument: criminal evidence]** ahead of a flight to Islamabad. A judge ruled he be held until Saturday while Pakistani officials investigate. The US embassy in Islamabad refused to reveal his identity but confirmed that a US citizen **[Event argument: suspect]** had been arrested and said the embassy was co-operating with local authorities. On Wednesday, US State Department spokeswoman Jen Psaki confirmed the man was an FBI employee "on temporary duty assignment to provide routine assistance to the legal attache at the US mission". In 2011 there was furore when CIA contractor Raymond Davis was arrested for shooting dead two men following what he said was an attempted armed robbery in Lahore. Mr Davis was later released after the families of the dead men accepted "blood money". **EVENT TYPE: CRIME**

The concatenated shortened version of the sample can be seen in the following:

> An agent with the US FBI has been arrested under anti-terrorism laws in Pakistan for carrying ammunition while trying to board a flight. The US citizen was detained at Karachi airport after security staff found 15 bullets for a 9mm handgun in his luggage ahead of a flight to Islamabad. The US embassy in Islamabad refused to reveal his identity but confirmed that a US citizen had been arrested and said the embassy was co-operating with local authorities. **EVENT TYPE: CRIME**

The entire corpus was modified in the described way. This corpus will be used for model training. The reason for summarizing the corpus in this way is that a difficulty that is detected for DEE is that event arguments are scattered over the entire document and that additionally other events than the main event can be reported on. This makes the concise representation of event types difficult as a document-level corpus includes noise. By summarizing the documents by means of their arguments, the textual representations of events are reduced to their essence. This then means that a classifier trained with only argument text data should be able to learn good and transferable representations of the individual types.

Additionally, the possibility arises to consider the robustness of the trained models by testing them on a part of the unreduced database. Noise that was

avoided during training is introduced back for testing. It can be evaluated if and how well the models are able to predict the event types when confronted with the unreduced database.

4.3.3 Data Reduction according to Length

The LLMs that were utilized for the work at hand are BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019). The models are able to process texts with a maximal sequence length of 512 tokens. A rule of thumb for equating tokens to words is that 1 token corresponds to approximately 0.75 words. Each text included in the DocEE corpus was analyzed with regard to its length to ensure correct processing. All texts exceeding the limit of 512 tokens (or 384 words) were dropped from the database.

4.4 Evaluation Methods

A number of commonly used methods is available for evaluating fine-tuned ML models. The evaluation methods selected for this work are described in the subsequent chapters. The first group of evaluation metrics that will be introduced are recall, precision and F1 score. These metrics focus on evaluating the performance of a fine-tuned model when it is used to predict new data. After that, cross-validation will be considered. Cross-validation is a data resampling method. It is used to create different versions of the corpus that is used for model training and testing. Lastly, paired bootstrap testing as a way to test for statistical significance between two models is outlined.

4.4.1 Precision, Recall and F1 Score

Precision, recall and F1 score are typically used methods for model evaluation in the area of NLP. Precision and recall are based on the distribution of false negatives, true negatives, false positives and true positives in a model's predictions.

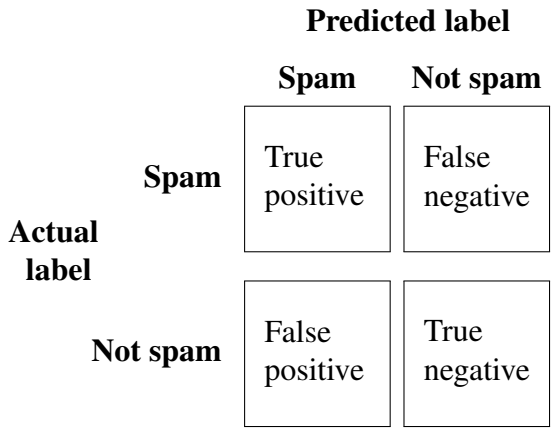

Figure 5: Recall and precision for a binary task

To illustrate the intuition behind these terms, the binary classification task of spam detection can be considered (cf. figure 5). The predefined classes in this case are spam and not spam. Consequently,

- a true positive is a text that is labeled as spam and also predicted as spam.
- A false negative is a text that is labeled as spam but predicted as not spam.
- A false positive is a text labeled as not spam but predicted as spam.
- A true negative is a text that is labeled as not spam and predicted as not spam.

Precision "measures the percentage of the items that the system detected [...] that are in fact positive" (Jurafsky and Martin 2024: 71). For the spam example this means that instances that are classified as spam are relevant for calculating the precision score. Precision focuses on evaluating how many of these samples actually are spam.

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives + false positives}} \tag{9}$$

Recall "measures the percentage of items actually present in the input that were correctly identified by the system" (Jurafsky and Martin 2024: 71). Therefore, recall focuses on all samples that, for the example above, should be classified as spam according to the actual labels.

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives + false negatives}} \tag{10}$$

Precision (P) and recall (R) are combined into a single score for model evaluation, the so-called F-measure. The F-measure is defined as:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{11}$$

The parameter β is responsible for weighing the importance of precision and recall. Thus, it can attain different values. When using $\beta > 1$, recall is favored, by choosing $\beta < 1$ precision is. For NLP tasks, precision and recall are typically equally important and balanced accordingly with $\beta = 1$ (cf. Jurafsky and Martin 2024: 71). This is then called F1 score and calculated as follows:

$$F_1 = \frac{2PR}{P + R} \tag{12}$$

The F1 score is used as the main evaluation score for all models in this work.

Many NLP tasks, as for example event type classification, do not aim at binary classification, but consider more than two classes. When using the described metrics for a multiclass classification task, one can choose between microaveraging or macroaveraging. In short, microaveraging

> collect[s] the decisions for all classes into a single confusion matrix, and then compute[s] precision and recall from that table [while macroaveraging] compute[s] the performance for each class, and then average[s] over classes (Jurafsky and Martin 2024: 72).

By first summarizing all predictions, as is the case for the microaverage, the score is "dominated by the more frequent class" (Jurafsky and Martin 2024:

72). The macroaverage, on the other hand, "better reflects the statistics of the smaller classes" (Jurafsky and Martin 2024:72) as all classes are in a first step evaluated in isolation with regard to their precision and recall before being averaged. In order to get the most meaningful score possible for model evaluation, results in this work will be reported with the macroaverage (called macro F1 score).

4.4.2 Cross-Validation

When training a model, a given dataset is split into three subsets: the training set, the development set and the test set. The training and development set are used for training and optimizing a model while the test set is not considered during this process. Instances included in the test set therefore constitute entirely new data for the model which is leveraged for testing the model's ability to handle unseen data. Performance scores like precision, recall and F1 score are calculated based on predictions made on this test set. Training, development and test set splits are oftentimes made at random. In consequence, a split might include unwanted effects as for example a random imbalance in class distribution or an over representation of certain textual characteristics. Accordingly, testing on only one fixed test set gives limited insights into a model's capability to solve a specified task.

An approach to mitigate this problem is the utilization of cross-validation (cf., among others, Stone 1974). Cross-validation is a data resampling method that is used to better assess the performance of trained models. When implementing cross-validation, the training, development and test sets are repeatedly sampled randomly from the corpus. The model is then trained and tested a given number of times, each time with a different partition of the data. The focus of evaluation thus shifts from rating a model's ability to correctly predict instances included in one specific test set to rating its ability to solve the specific task over a number of different test sets.

Cross-validation is especially helpful when comparing the performance

of different models to each other. By providing different splits for training and testing, the above mentioned random effects in the data or in the training process itself, which might give a specific model configuration an advantage (or disadvantage), are less likely to be overemphasized (cf. Qiu 2024).

4.4.3 Paired Bootstrap Testing

Bootstrap testing was first proposed by Efron (1979) and further described by Efron and Tibshirani (1993). Bootstrapping as such refers to "repeatedly drawing large numbers of samples with replacement (called bootstrap samples) from an original set [thus creating] many virtual test sets from an observed test set" (Jurafsky and Martin 2024: 75).

In the context of NLP, paired bootstrap testing is utilized to evaluate differences between the performance of two models with regard to statistical significance. Underlying is the hypothesis H_1: Model A is better than model B. The corresponding null hypothesis H_0, model A is not better than model B, is tested.

Berg-Kirkpatrick et al. described the process that is implemented to achieve this goal: "The bootstrap estimates p-value(x) th[r]ough a combination of simulation and approximation, drawing many simulated test sets $x^{(i)}$ and counting how often A sees an accidental advantage of $\delta(x)$ or greater" (2012: 996).

For the work at hand, paired bootstrap testing is implemented with the following steps:

1. The effect size between the two models that are to be contrasted, $\delta(x)$, is ascertained for the original test set. The basis are the respective macro F1 scores achieved by the models. Therefore, if model A reaches an F1 score of 0.9 and model B reaches a score of 0.8, $\delta(x)$ is 0.1.

2. A large number b of virtual test sets, $x^{(i)}$, is created from the original test set. In this case, the number of virtual test sets is 10000.

3. The predefined level of significance is set at 0.05.

4. The p-value is calculated by counting how often $\delta(x^{(i)})$, i.e., the effect size for a given virtual test set, is larger than the value of $\delta(x)$, i.e., the effect size observed for the original test set, by at least $\delta(x)$ (cf. Jurafsky and Martin 2024: 76):

$$\text{p-value}(x) = \frac{1}{b}\sum_{i=1}^{b} \mathbb{1}(\delta(x^{(i)}) \geq 2\delta(x)) \tag{13}$$

   The notation $\mathbb{1}$ is used by Jurafsky and Martin to express: "1 if x is true, and 0 otherwise" (2024: 76).

Paired bootstrap testing will be implemented for model evaluation in this work. The resulting p-values are used to assess the differences between models statistically.

5 Preliminary Experiments

This chapter describes the first two experiments of this work. The experiments aim at the classification of event types with transformer models. Classification is approached from an exploratory standpoint. The main aim is to gain an initial understanding of

1. the probable textual features that are automatically determined and focused on by transformer models in order to solve the task and

2. the observable consequences for classification that arise from this automatic selection.

5.1 Ablation Study Concerning Word Classes

As was seen in chapter 3.3, it is widely accepted that transformer models removed the need for feature engineering when solving a wide range of NLP tasks. In the context of event type classification, approaches that include feature engineering pay attention to the word class of the event trigger. The identification of the event trigger is necessary as it ultimately determines the event type. The event trigger is mostly encoded in the main verb, less often it is found as a noun or adjective. The goal of the first preliminary experiment is to gain insights into which word classes are most important for a transformer model when classifying event types. Event type classification with transformer models no longer extracts the event trigger explicitly. The event types are determined based on the entire input text. The question can be posed if transformer models still consider the same word classes as important without explicitly prompting them to do so.

To test this, a simple ablation experiment is implemented. For this experiment, the database for training and testing the model is modified. Specifically, three different modified corpora are created. For each corpus, all tokens of a

different word class are masked in the texts. It is reasoned that model performance will differ depending on which corpus version is used for training and testing. The corpus versions are:

- One corpus with masked verbs,
- one corpus with masked nouns and
- one corpus with masked adjectives.

Word classes are identified with spaCy's (Honnibal et al. 2020) Part-of-Speech (POS) tagger. The specific implementation used here is the en_core_web_md pipeline [4]. The POS tagger annotates tokens with universal POS (UPOS) tags (cf. de Marneffe et al. 2021:261). With regard to verbs, the UPOS annotation scheme differentiates between auxiliary verbs and main verbs. Auxiliary verbs were not considered for masking. Accordingly, tense information as expressed by auxiliary verbs remains in the database. Nouns are also split into two distinct classes: common and proper nouns. Only common nouns were masked.

In general, it has to be noted that masking specific word classes in a corpus also influences syntax, i.e., sentences are no longer syntactically complete, as well as contextual understanding. However, this preliminary experiment is meant to give only high-level insights into the role of the individual word classes for event type classification. The results will be refined further later on.

If a token corresponds to a given UPOS tag, it is masked with "XXX". An example text that belongs to the event type CRIME along with all three modifications is provided below.

> **Original**: Updated July 28, 2021 10:25 pm ET HONG KONG—A Chinese court sentenced an outspoken farming magnate to 18 years in jail for allegedly causing public disorder and a multitude of other offenses,

[4] https://github.com/explosion/spacy-models/releases/tag/en_core_web_md-3.8.0

dishing out heavy penalties in a case seen as a bellwether of the growing political risks that confront private businesses in China. Sun Dawu, 67 years old, was convicted Wednesday of crimes that included agitating crowds against state organs, illegal fundraising and unlawful occupation of farmland, a municipal court in the northern city of Gaobeidian said in a notice published on its website. He was also fined about 3.1 million yuan, the equivalent of about $405,500.

Modified masked verbs: XXX July 28, 2021 10:25 pm ET HONG KONG — A Chinese court XXX an outspoken farming magnate to 18 years in jail for allegedly XXX public disorder and a multitude of other offenses, XXX out heavy penalties in a case XXX as a bellwether of the XXX political risks that XXX private businesses in China. Sun Dawu, 67 years old, was XXX Wednesday of crimes that XXX XXX crowds against state organs, illegal fundraising and unlawful occupation of farmland, a municipal court in the northern city of Gaobeidian XXX in a notice XXX on its website. He was also XXX about 3.1 million yuan, the equivalent of about $ 405,500.

Modified masked nouns: Updated July 28, 2021 10:25 XXX ET HONG KONG — A Chinese XXX sentenced an outspoken XXX XXX to 18 XXX in XXX for allegedly causing public XXX and a XXX of other XXX, dishing out heavy XXX in a XXX seen as a XXX of the growing political XXX that confront private XXX in China. Sun Dawu, 67 XXX old, was convicted Wednesday of XXX that included agitating XXX against XXX XXX, illegal XXX and unlawful XXX of XXX, a municipal XXX in the northern XXX of Gaobeidian said in a XXX published on its XXX. He was also fined about 3.1 million XXX, the XXX of about $ 405,500.

Modified masked adjectives: Updated July 28, 2021 10:25 pm ET HONG KONG — A XXX court sentenced an XXX farming magnate to 18 years in jail for allegedly causing XXX disorder and a multitude of XXX offenses, dishing out XXX penalties in a case seen as a bellwether of the growing XXX risks that confront XXX businesses in China. Sun Dawu, 67 years XXX, was convicted Wednesday of crimes that included agitating crowds against state organs, XXX fundraising and XXX occupation of farmland, a XXX court in the XXX city of Gaobeidian said in a notice published on its website. He was also fined about 3.1 million yuan, the equivalent of about $ 405,500.

Table 4 summarizes the tokens that are omitted for each word class in the example text given above. Taking the event type annotation scheme into account, tokens that can be seen as indicative for the event type CRIME are colored.

The tokens *fined* and *penalties* could also be considered to be relevant for the event type CRIME. However, other event types in the corpus have to be considered. Thus, *fined* is assumed to also appear in connection to the event

Verbs	Nouns	Adjectives
updated	pm	chinese
sentenced	court	outspoken
causing	farming	public
dishing	magnate	other
seen	years	heavy
growing	jail	political
confront	disorder	private
convicted	multitude	old
included	offenses	illegal
agitating	penalties	unlawful
said	case	municipal
published	bellwether	northern
fined	risks	
	businesses	
	years	
	crimes	
	crowds	
	state	
	organs	
	fundraising	
	occupation	
	farmland	
	city	
	notice	
	website	
	yuan	
	equivalent	

Table 4: Masked tokens after word class preprocessing

type ORGANISATION because of the subtype Organisation Fine. *Penalties* can be connected to the event type WIN LOSE, specifically to the subtype Sports Competition.

The data used for the experiment is the event argument version of a subset of the DocEE corpus as provided by Tong et al. (2022). This dataset is randomly split into a training, development and test set with a 70/20/10 ratio. The training set includes 1577 samples, the development set includes 446 samples and the test set 213.

All four corpus versions are processed with the same model, RoBERTa, and the same hyperparameters: a learning rate of 3e-5, a maximum of 5 epochs and a batch size of 8.

5.1.1 Results

Table 5 shows the results of the first preliminary experiment.

Corpus version	**Macro F1 score**	**p-value**
Original	0.9068	
No verbs [POS TAG: VERB]	0.8985	0.2241
No nouns [POS TAG: NOUN]	0.8174	0.0173
No adjectives [POS TAG: ADJ]	0.8941	0.1002

Table 5: Results ablation study

The results indicate that the masking of nouns hinders the correct classification of event types the most. The models are tested for statistical significance by assuming the null hypothesis that the model that uses the original data is not better than the models that use the modified data. A significant difference with a p-value of 0.0173 is found for the model that was trained with masked nouns. No significance is reached for the other two modifications.

When looking at the example texts again, it has to be considered that the amount of tokens that is masked varies between word classes. Nouns occur more frequently in a text than verbs and adjectives. Therefore, the masking of nouns skews the database more heavily which leads to a greater loss of context. This means that the results obtained by training with masked data cannot be attributed exclusively to the importance of the word class as such. The effect of varying limitations with regard to context that result from the corpus modifications also has to be considered.

Consequently, to further investigate the role of the individual word classes, the Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al. 2016b) library is utilized to analyze the classification decision for the example text given above. The model trained with the original data is used. LIME provides local explanations. This means that by using LIME, the most important words for the classification of the specific sample that is processed at the moment are identified. These identified words, however, do not necessarily hold

importance for each sample of the class. Still, LIME's results give valuable insights into how a trained model behaves, if only locally.

Basically speaking, to find the most important words, LIME changes the input that is processed by a classification model and observes how the model's output changes. In more detail, LIME works as follows (cf. Ribeiro et al. 2016a and Ribeiro et al. 2016b):

1. At first, the input data is perturbed. In the case of textual data this means that words are removed from the samples. The input data used here is the test set.

2. Perturbed samples are assigned weights according to their proximity to the original sample. The more similar the perturbed sample is to the original sample, i.e., if only a few words are removed, the higher the weight is. Only samples that have a high enough weight and are therefore close enough to the original sample to still be meaningful are considered further.

3. The samples are then classified by the model whose decisions should be explained. The samples along with the model's predictions are used as the gold standard dataset for training the explanation model.

4. The perturbed dataset is then used to train a simple interpretable linear model. This model functions as the explanation model. Interpretability of the explanations is supported by representing the individual samples with binary vectors. In these vectors, words are represented by 0 if they were removed from the sample and by 1 of they are present. Individual words act as features.

5. During training of the explanation model, feature weights are learned to solve the task according to the given dataset and thus based on the original model's predictions.

6. The final word importance scores provided by LIME correspond to the feature weights that were learned by the explanation model.

Figure 6 shows the top 10 features that drive the classification of the example text provided earlier as identified by LIME. Tokens colored in green on the NOT 1, i.e., not the class CRIME, side show tokens that are seen by the model as indicators that another class than CRIME is present. Tokens in blue are the ones most important for the classification of the CRIME (and thus the correct) class.

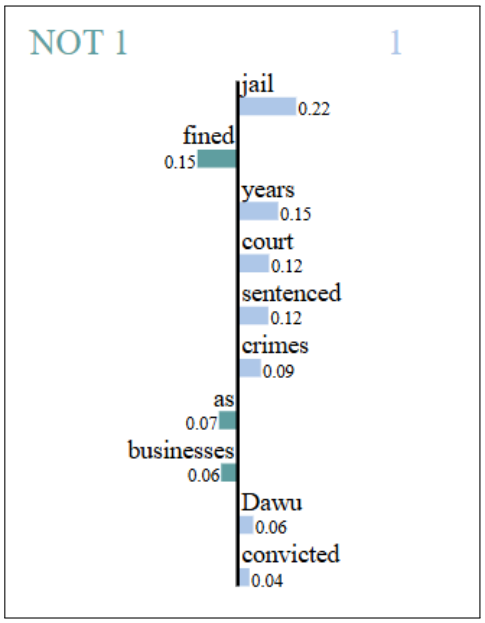

Figure 6: LIME output for model trained with original data

A similar picture to the one gained by masking entire word classes arises: Tokens belonging to the word class noun seem to hold the most importance for classifying the event type. Verbs that were determined to be important, namely *sentenced* and *convicted* are also found among the top 10 features, however, the nouns *jail*, *years* and *court* trump them in their relevance.

The previous manual analysis identified *fined* as a word that could potentially be relevant for the classification of the CRIME text. It was however assumed that *fined* also holds relevance for the event type ORGANISATION due to the subtype Organisation Fine. The results obtained by LIME support this assumption. *Fined* represents the strongest indicator that another class

than CRIME is given.

In order to evaluate the findings more broadly, the classification of the entire test set is evaluated with LIME in the just described way. The top 10 features are extracted for all texts included in the test set. Figure 7 shows the distribution of word classes in the positive features that were identified.

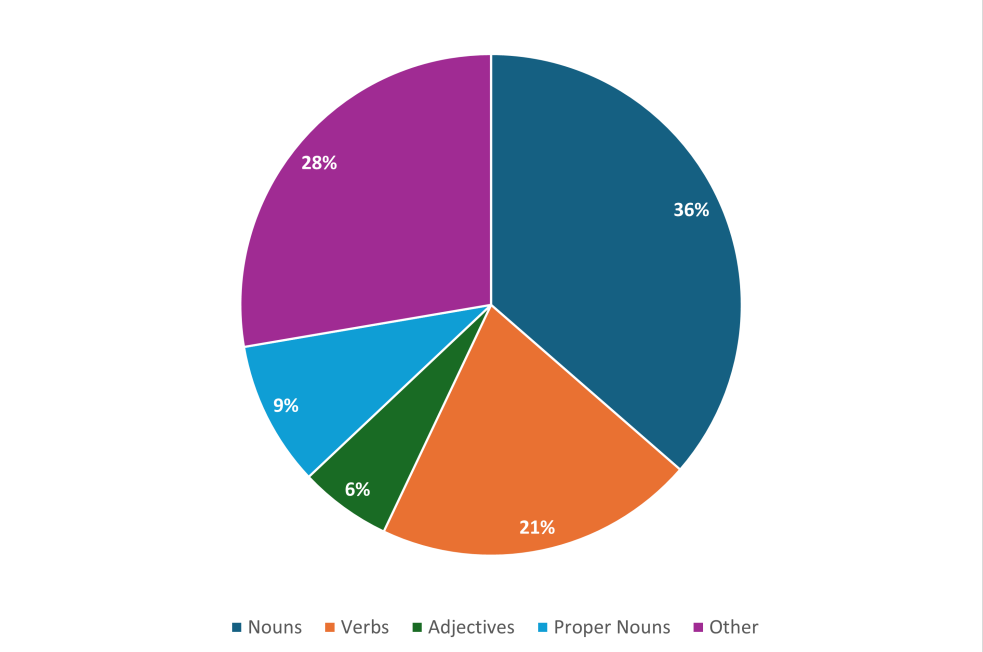

Figure 7: Word class distribution in LIME feature output

It can be seen that nouns constitute the largest word class with 36%. The second largest group, other, (28%) is composed of all word classes that are present in the extracted features but not specifically analyzed. Among others, those are conjunctions, pronouns and adpositions. Figure 8 gives an overview of the distribution of word classes included in the category other .

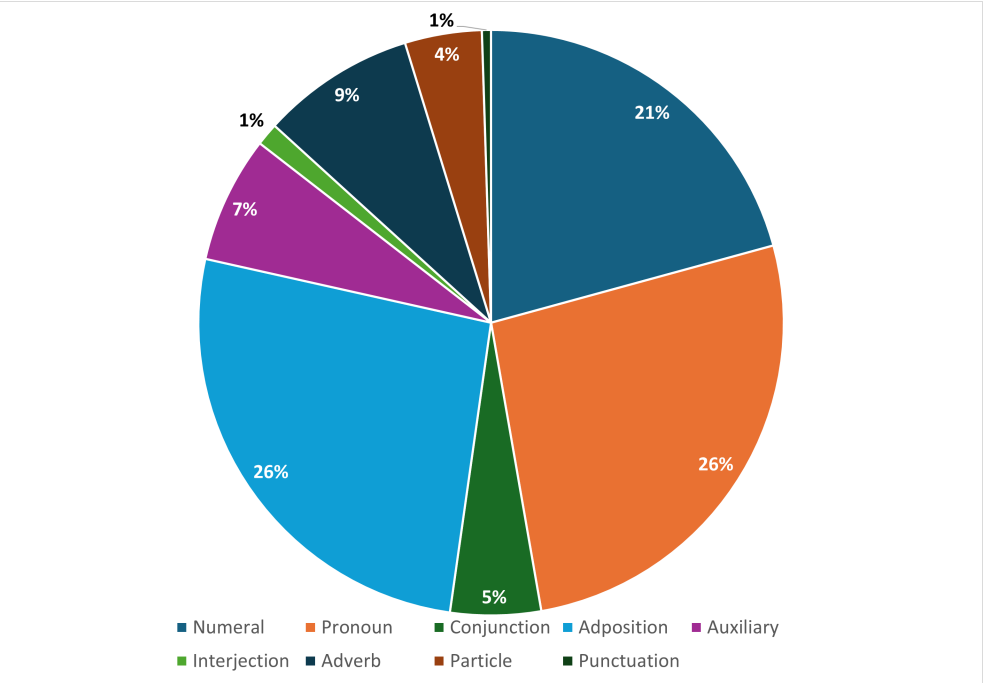

Figure 8: Word class distribution in the other category LIME output

Verbs amount to 21% of all features and follow nouns as the second most prevalent word class in the LIME output. With 6%, adjectives seem to be less important than nouns and verbs for the classifier. The analysis additionally reveals that proper nouns appear to hold importance for the classification to some extent. 9% of features fall into this word class. Proper nouns are, for example, names of persons and organisations as well as locations.

Table 6 gives an overview of specific words that are deemed important for the correct classification of the event types by LIME. The table includes the words in the same form as they are found in the data. They are not lemmatized. The focus is again on the word classes noun, verb, adjective and proper noun. Words are colored according to their word classes as identified by the POS tagger in the following way: common noun, verb, adjective, proper noun.

In order to focus on the most relevant words, only those that appear more than once in the output are included in this case. Exceptions are FINANCE and LIFE. For FINANCE, no words were seen more than once. Therefore, all words that hold importance for classification are listed. For LIFE, only one word was deemed important by LIME more than once, namely *has*. The

choice was made to also include all words of importance for the event type LIFE.

Event type	Words
Accident	fire, crashed, flight, crash, aircraft, injured, killed, passengers, injuries, driver, car, train, destroyed, accident, explosion, California, wildfire, including, local
Crime	police, arrested, custody, poisoning, monoxide, robbery, investigations, jail, suspicion, arrest, court, FBI
Diplomacy	talks, meeting, Accord
Discovery	time, record
Disorder	protests, protest, strike, demonstrations, riots, riot, United, people
Finance	UNHCR, development, system
Government	said, president, minister, has, vote, government, prime, law, country, new, presidential, day, party, elections, job, appointed, candidate, sworn
Life	has, hospital, hospitalized, recovering, Theroux, engagement, December, revealed, interview, producers, died, used, found, Rajesh, morning, Bollywood, kidney, had, passed, Gibb, committed, diagnosed, killed, Monday, reported, official, government, Justin, married, wed, Ed, marriage, singer, child, wife, proposal
Natural Disaster	earthquake, magnitude, eruption, volcano, high, rainfall, flooding, floods, mudslides, affected, conditions, famine, report, river, caused, depth, extreme, earthquakes, National, estimated, reported, weather, water, spill
Organisation	force, lives, fined
Win Lose	tournament, snooker, championships, world

Table 6: Overview LIME output

The analysis of the table reveals that in the case of the word *minister* which is found for the event type GOVERNMENT, two POS tags are present in the output: common and proper noun. The is due to the expression *Water Minister* which is tagged as proper noun.

A number of words are counter-intuitively tagged as proper nouns. Those are: *Accord* found for the event type DIPLOMACY, *United* for DISORDER and *National* for NATURAL DISASTER. The consultation of the correspond-

ing texts sheds light on this. *Accord* refers to the Copenhagen Accord and the Nuclear Accord, *United* refers to the United States and *National* to the National Weather Service.

The overview could be interpreted as follows: If the number of words that are included more than once is high, as for example seen for ACCIDENT, GOVERNMENT and NATURAL DISASTER, the data as well as the representations of the event types that were learned by the model can be seen to be focused. A low number of frequent words, as seen for example for LIFE and DISCOVERY, indicates that more important tokens found by LIME only appear once. The representation of the corresponding event types might therefore be more diverse. It also has to be considered that varying amounts of texts are available for the individual event types. This also influences the number of features that appear more than once.

At this point, it can be said that it seems as if the classification model at hand focuses primarily on nouns for solving the task of event type classification. This somewhat stands in contrast to the definition of an event type's defining tokens in the context of EE, where the event trigger, encoded (mostly) in the main verb, is seen to be crucial (cf. chapter 3.1).

To contextualize the findings, figure 9 depicts the distribution of word classes in the training and development set, i.e., the distribution present in the data that was seen during model training. The word classes that were focused on during the analysis of the LIME output (common nouns, verbs, adjectives and proper nouns) are also focused on here. The distribution of the word classes in the LIME output is included again to facilitate comparison.

In both cases the distributions are calculated based on the number of all tokens belonging to the word classes. This means that even though word types are presented in table 6, all occurrences of the individual words in the LIME output are considered for calculating the word class distribution. This also includes words that only appear once in the LIME output and are therefore not listed in table 6.

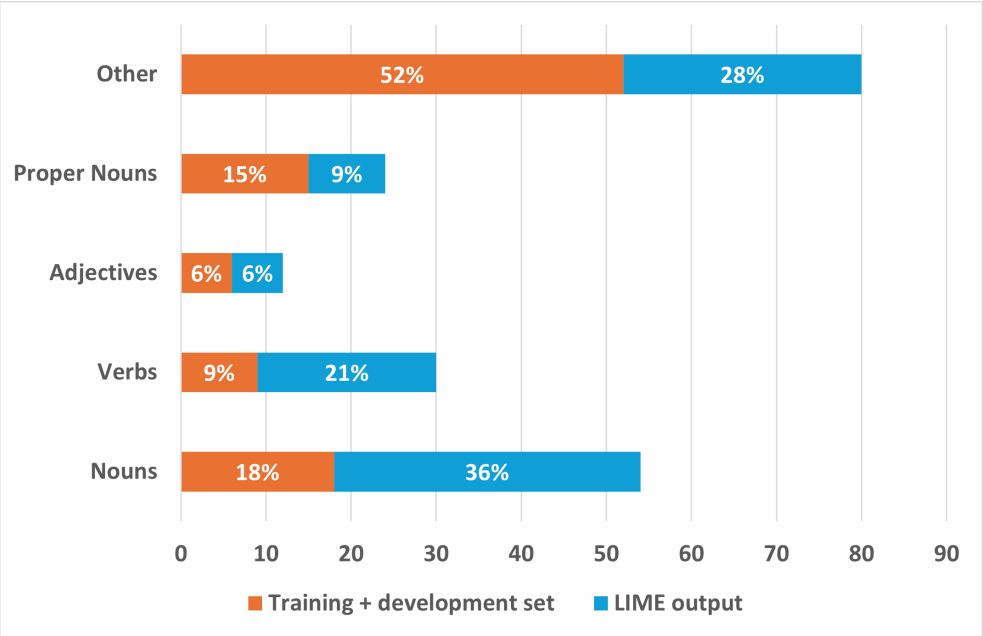

Figure 9: Word class distribution in training and development set

Most notably the category other is more dominant with 52% in comparison to the distribution found in the LIME output (28%). A large number of high-frequency function words are included in the classes that are part of the category other. Function words are seldom used as a deciding factor for content classification. Accordingly, the category other is less relevant in the LIME output. A fine-grained depiction of the specific word classes included in the other class is shown in figure 10. Some word classes that are found in the training and development data are not included in LIME's features. Those are determiner and symbol.

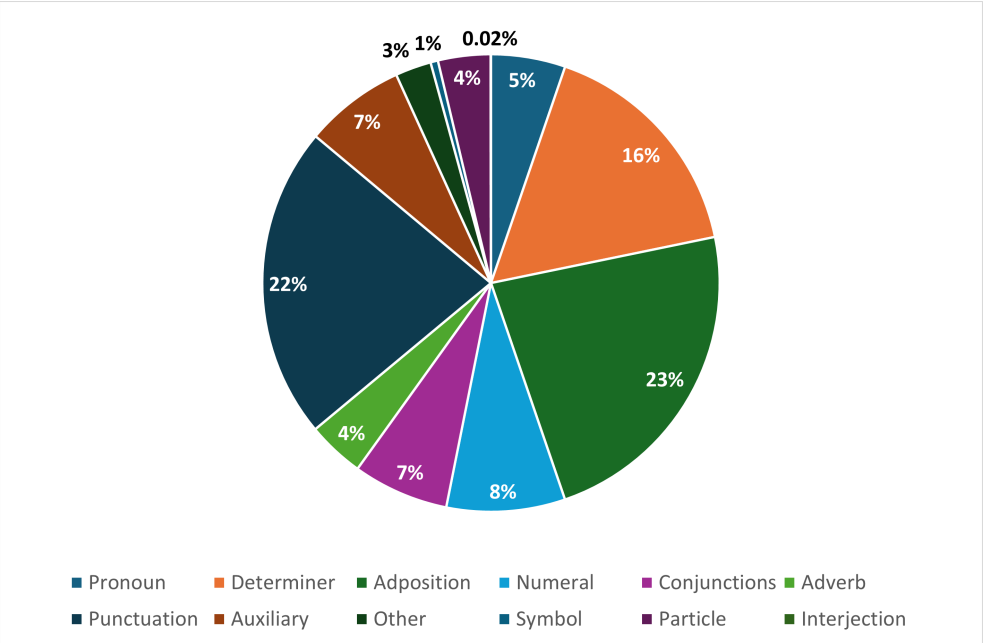

Figure 10: Word class distribution in the other category training and development set

The word class proper nouns is also found with a higher amount in the training and development data (15%) than in the features extracted with LIME (9%). The word class adjectives stays the same (6%). A notably higher proportion of common nouns and verbs is found in the LIME output in contrast to the database. The amount of common nouns increases by 18%, the amount of verbs by 12%. It can be reasoned that common nouns and verbs (as content words) convey the most relevant information for solving the task of event type classification. Therefore, they are focused on by the model.

The ablation study showed that for the model used here, i.e., RoBERTa, it can be assumed that nouns and verbs are focused on to reach a classification decision. Nouns hold a greater relevance than verbs. The analysis of the word class distribution of the training and development set revealed that this is most likely due to the higher count of available nouns.

In consequence, this means that today's research focus of "identifying event types based on the full text" (Li et al. 2021a) with the help of LLMs, instead of considering specific features to solve the task, shifts how the task of event type classification is approached to some extent. Existing definitions

of events state that an event is most distinctly characterized by its trigger (cf. Linguistic Data Consortium 2005). This trigger is most often the main verb. Earlier works aimed at extracting this trigger and then classifying it. LLMs necessarily consider entire input texts during the training process. The main verbs of sentences are still considered for classification as can be seen by the LIME results, however, the most sensible patterns for classification are found in the presence of specific nouns.

The question arises if the verbs that were identified by LIME as important tokens act as event triggers in the corresponding texts. A manual analysis revealed that this is the case a number of times. This means that the identified verbs can be seen as the tokens that most clearly express the event type of the text. A small selection of example texts that show this is listed below. The relevant verbs are colored.

> There were 102 passengers on board. The flight was led by Captain Nikolai Danilov, with 4 other pilots and two flight attendants. On 27 April 1974, an Aeroflot Il-18 airliner crashed while operating a charter flight from Leningrad (now St. Petersburg) to Zaporizhzhia, continuing to Krasnodar, Russia. None of the 109 people on board survived. The aircraft's number four engine caught fire two minutes after takeoff due to a faulty compressor disk. **ACCIDENT**
>
> COLOMBO: Sri Lanka's police arrested a top Muslim legislator Saturday in connection with the 2019 Easter Sunday attacks that killed 279 people, as pressure mounted to speed up the investigation. Detectives took Rishad Bathiudeen, leader of the All Ceylon Makkal Party (ACMP) and a former minister, into custody under the Prevention of Terrorism Act (PTA), police spokesman Ajith Rohana said. **CRIME**
>
> Mr Qandil was a member of the outgoing caretaker government of Kamal al-Ganzouri, who was appointed prime minister by the military last year. At 50, he is Egypt's youngest prime minister, reports say. Egyptian President Mohammed Mursi has appointed Water Minister Hisham Qandil as prime minister and asked him to form a new government, state TV says. **GOVERNMENT**
>
> This was his third hospitalisation since he was diagnosed with a kidney problem. He was 69. The yesteryear actor had been ailing for some time now, and had been discharged from hospital on Tuesday afternoon. Sources said that the actor was not responding to medicines, and had been put on IV and had been gasping for breath since morning. The first superstar of Bollywood Rajesh Khanna has passed away into the ages. He passed away at his

bungalow, Aashirwad at Carter Road, Mumbai. **LIFE**

It has to be noted, however, that there are also examples where the verbs that were determined to be relevant by LIME can not be seen as event triggers. Two of such instances are shown below.

> The president said the new law would not only apply to the AUC, but to all those guerrilla groups who commit to a peace process, EFE news agency said. In Colombia, Congresswoman Gina Parody said the law sent a clear message to Colombian society, namely that crime does pay. The AUC says it is committed to demobilising all its fighters by the end of the year. The legislation was passed primarily to demobilise the right-wing paramilitaries known as AUC - the United Self-Defence Forces of Colombia. **GOVERNMENT**
>
> The 2001 Geiyo earthquake (2001 Nisen-ichi-nen Gēyo Jishin) occurred with a moment magnitude of 6.7 on March 24 at 15:27 local time near Hiroshima, Japan. One person in Hiroshima and one person in Ehime were reported dead. About 3,700 buildings were damaged in the Hiroshima area. **NATURAL DISASTER**

With regard to these two texts, the event type is indicated more by nouns, namely by *law* and *legislation* in the first text and *earthquake* in the second text. These nouns were also identified by LIME to hold relevance.

5.2 Cross-Validation

To investigate the influence that the textual patterns that are focused on for classification by a transformer model might have with regard to model performance, model training is repeated. Both BERT and RoBERTa are tested. This time cross-validation as described in chapter 4.4.2 is implemented. Training, development and test sets are sampled randomly from the given dataset five times. This results in five different data splits. The splits hold the amounts of texts shown in table 7.

Figures 11 - 13 show the distribution of the individual event types in the training, development and test sets of all splits. Class sizes are depicted with their relative amounts in the respective datasets.

The distribution of event types is similar for all five data split. This means that even though different texts are included in the data for each split, class

Split	n Training	n Development	n Test
1	1518	430	204
2	1504	429	219
3	1513	430	209
4	1511	421	220
5	1530	420	202

Table 7: Number of instances in training, development and test sets

sizes are comparable. Especially for the training sets, there are no notable differences.

It becomes obvious that the event type distribution is unbalanced. Some event types are very dominant, e.g., ACCIDENT, others are consistently small, e.g., FINANCE. The unbalanced character of the corpus can be seen as imitating the real-world: Not all types of events occur and are reported on equally frequent. A car accident, for example, is more common than a financial crisis. Still, if vast amounts of texts are analyzed with regard to the event types that are reported on, it is equally important to be able to detect all event types. Accordingly, no event type classes were dropped from the database and class sizes were not manipulated. The influence of class size on classification success will be a point of analysis.

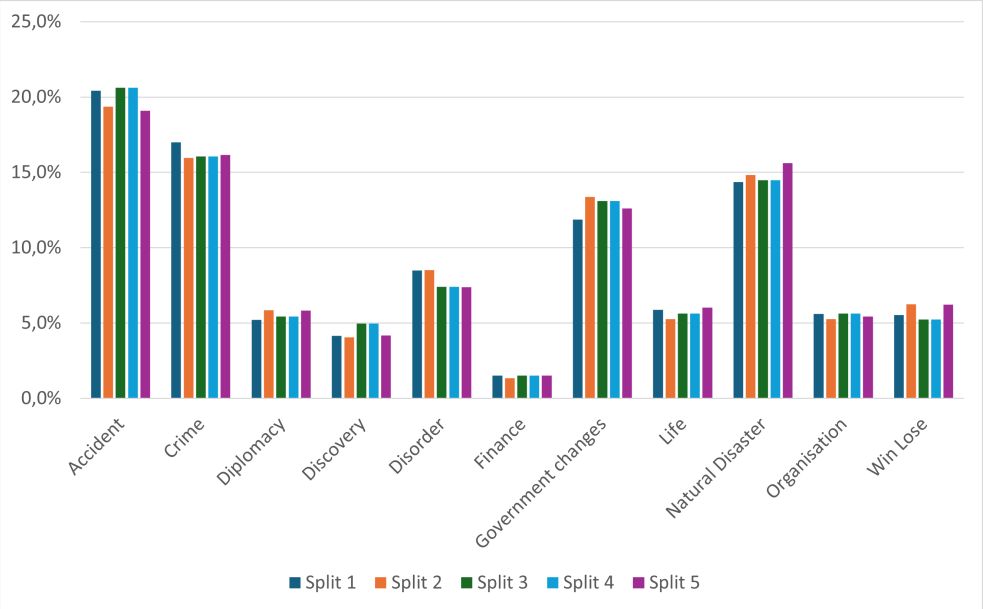

Figure 11: Event type distribution cross-validation training sets

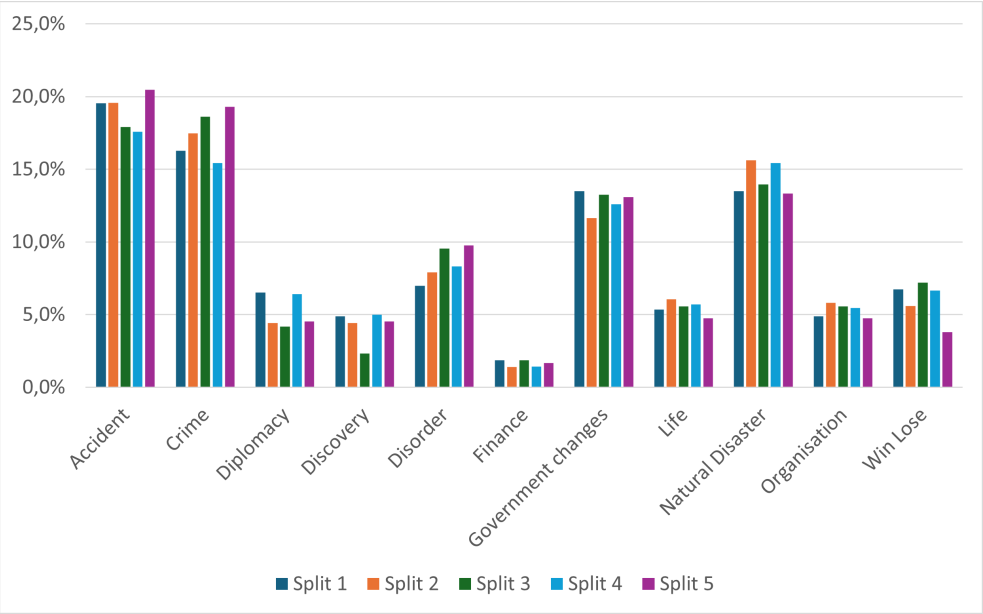

Figure 12: Event type distribution cross-validation development sets

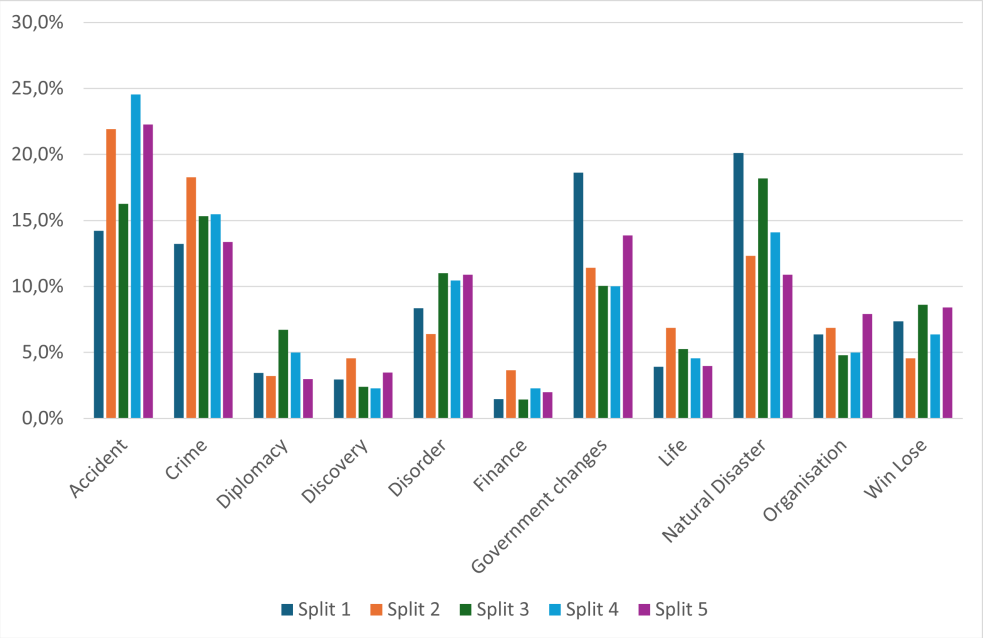

Figure 13: Event type distribution cross-validation test sets

In contrast to the ablation study, models are tested on two separate test sets in the cross-validation experiment. The first test set is taken from the same distribution as the training and development set and listed in table 7. This test set changes for each split along with the other data. The second test set includes 2000 instances from the DocEE corpus with texts that were not summarized to only include the event argument sentences, which arguably makes

it more diverse. This test set does not change between splits but remains the same. The decision was made to include this second test set in order to test the models' ability to transfer what was learned to a database that diverges in parts from the training data.

For simplicity, the test set that includes the event argument texts and changes between data splits will be referred to as the small test set. The second test set which includes the original data will be called the big test set. Figure 14 shows the distribution of event types for the big test set.

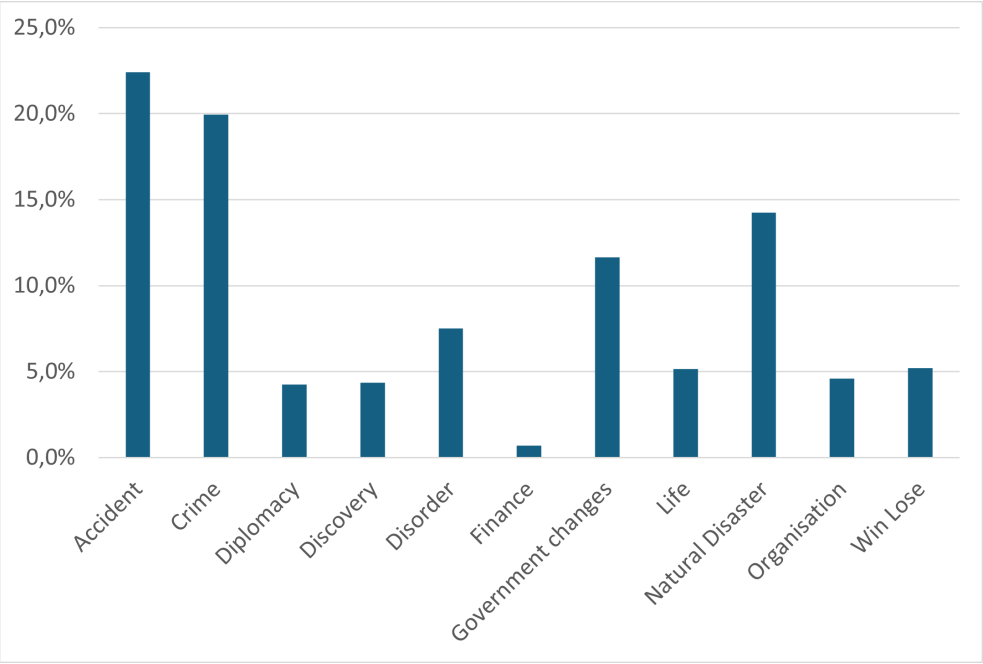

Figure 14: Event type distribution big test set

It can be seen that the distribution of event types in the big test set is comparable to the one found for the small tests sets as seen in figure 13. Relative class sizes are mostly similar.

During the training process of the models, hyperparameters are kept constant for all data splits with a maximum number of five epochs, a learning rate of 3e-5 and a batch size of eight. For testing, the best models are used. The value for determining the best model for each data split is the validation loss. The validation loss is calculated after every training epoch. The smallest validation loss can be reached before the maximum of five training epochs is concluded. This means that the best models which are used for testing and

analyzed in the following chapters are not necessarily trained for five epochs.

5.2.1 Results

The results that are obtained by the cross-validation experiment are presented and discussed with the help of a number of figures and tables that focus on different ways of evaluating the trained models.

First, model results are depicted as a bar graph. In the graph, the x-axis represents the five models that were trained with the five different data splits. The y-axis shows the macro F1 score. It ranges from 0.78 to 0.96 as all model scores lie in between these two values. The blue bar that is found in the graphs represents the small test set, the orange bar the big test set.

As a means of considering the scores of the individual event types next to the macro F1 scores, classification reports are drawn upon. A classification report includes precision, recall and F1 scores for each of the eleven event types. In the next chapters, classification reports will be utilized for the error analysis. Both test sets are reported on and included in the same table. For better readability, scores are rounded to two decimal places for these larger tables.

Further insights into the errors made by the models during testing can be gained by inspecting a confusion matrix. A confusion matrix shows interchanges between the correct event types (true label) and the predicted event types (predicted label). For this work, confusion matrices are used for a detailed analysis of a model's predictions. The matrices are normalized over the true condition, i.e., over row values. This means that they are focused on the recall scores. Recall and precision are connected to each other. Therefore, the confusion matrices also hint at precision related errors, even if they are not normalized accordingly.

Both BERT and RoBERTa were tested with regard to their suitability for the task of event type classification. It was found that RoBERTa on average performed better than BERT. The fine-tuned BERT models reached a mean

macro F1 score of 0.8772 for the small test set and a mean macro F1 score of 0.8179 for the big test set. It was decided to focus on the fine-tuned RoBERTa models. This means that all models that are discussed, analyzed and trained in the remainder of this work are based on RoBERTa. The reason for this is that RoBERTa is in general the more suitable model choice for the classification of event types.

Figure 15 shows the results that were reached for the classification of the small and the big test set by the five fine-tuned RoBERTa models. These models will be discussed in more detail in the following chapter.

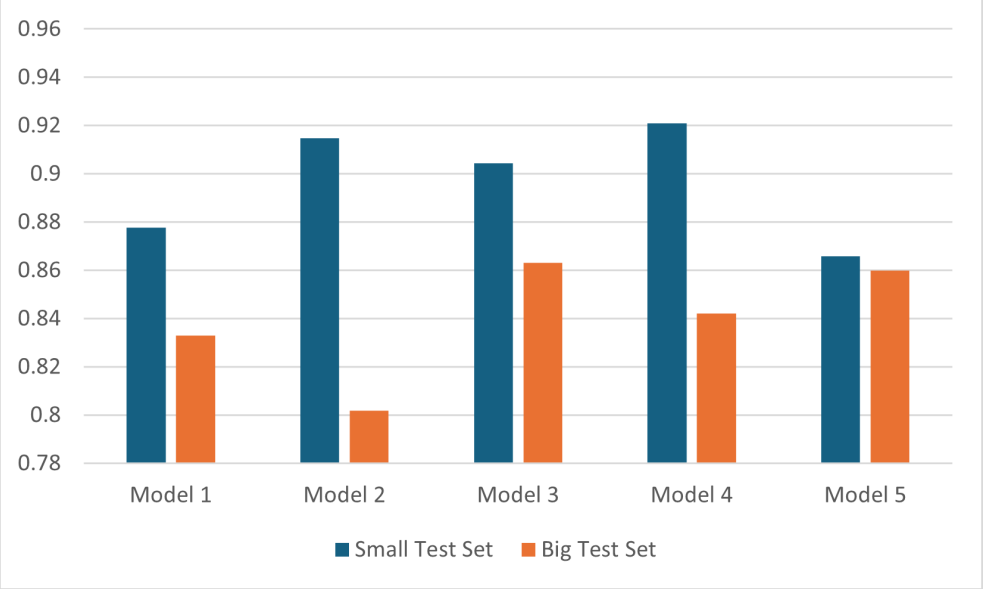

Figure 15: Results cross-validation

The key findings are:

- When testing on the small test set that includes the event argument texts, the macro F1 scores range between 0.8659 (model 5) and 0.921 (model 4).
- A mean macro F1 score of 0.8967 is reached.
- The standard deviation (SD) is 0.0239.
- As could be expected, macro F1 scores that were obtained when testing the models on the big test set were consistently lower, ranging from

0.8019 (model 2) to 0.86 (model 5).

- The mean macro F1 score over all five splits is 0.84.
- The SD amounts to 0.0247.
- A difference of 0.0567 between the mean macro F1 scores of the small and the big test set can be observed.

Figure 15 depicts noticeable discrepancies in model performance for model 2 and model 4. Models trained on split 2 and split 4 display the highest macro F1 scores for the small test set with 0.9147 and 0.921 respectively. However, testing those models on the big test set results in a stark decrease, with the lowest score among splits of 0.8019 for model 2 and a comparatively low score of 0.8421 for model 4. It can be argued that the models trained on split 2 and 4 exhibit problems in their ability to generalize what they learned when faced with a new and more diverse database for testing. This might be the result of overfitting which means that the models fitted too well to the data that was seen during training. In a way the models thus learn the representations of the different event types included in the training set by heart instead of learning a general representation.

The smallest difference between test sets can be observed for the model trained with split 5. Here, the lowest score of all five models was reached for the small test set (macro F1 score of 0.8659), however, this score is nearly the same when testing on the big test set (macro F1 score of 0.86). This might mean that the model learned a more shallow representation of the event types which in turn can be applied better to a new database.

In the following sections, an error analysis will be conducted for all five models. To this end, detailed classification reports for all test sets as well as confusion matrices focused on the performance on the big test set will be utilized. A reason for this analysis is to get an insight into difficulties that might be present for solving the task of event type classification as such.

Error Analysis Model 1 Table 8 shows the classification reports for predictions made by model 1 on both test sets. It gives insights into how well the individual event types are detected.

	Small test			**Big test**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Accident	0.91	1.00	0.95	0.91	0.98	0.95
Crime	0.93	1.00	0.96	0.87	0.95	0.91
Diplomacy	0.80	0.57	0.67	0.78	0.79	0.78
Discovery	0.86	1.00	0.92	0.80	0.94	0.86
Disorder	1.00	1.00	1.00	0.96	0.74	0.83
Finance	1.00	0.33	0.50	0.70	0.50	0.58
Government	0.95	0.92	0.93	0.86	0.85	0.86
Life	0.88	0.88	0.88	0.84	0.83	0.84
Natural Disaster	0.95	0.88	0.91	0.95	0.93	0.94
Organisation	0.87	1.00	0.93	0.86	0.70	0.77
Win Lose	1.00	1.00	1.00	0.96	0.73	0.83
Macro Average	**0.92**	**0.87**	**0.88**	**0.86**	**0.81**	**0.83**

Table 8: Model 1: classification reports

The classification of the small test set results in comparatively high F1 scores for eight event types, namely, ACCIDENT, CRIME, DISCOVERY, DISORDER, GOVERNMENT, NATURAL DISASTER, ORGANISATION and WIN LOSE. Out of these eight types, two reach an F1 score of 1: DISORDER and WIN LOSE. The event types that display the lowest scores are DIPLOMACY and FINANCE with 0.67 and 0.5 respectively. This entails that DISORDER and WIN LOSE have the greatest positive influence on the final score, DIPLOMACY and FINANCE have the greatest negative influence.

Model 1 exhibits a difference of 0.0447 between the macro F1 scores of the small and the big test set (0.8776 vs. 0.8329). However, the classification reports reveal that not all individual event type scores are lower with regard to the big test set.

Out of the eleven event types, three seem to contribute especially to the decrease in the final macro F1 score of the big test set. Those are DISORDER with a decrease in F1 score of 0.17, ORGANISATION with a decrease of

0.16 and WIN LOSE with 0.17. DISORDER as well as WIN LOSE held an F1 score of 1 for the small test set. This flawless performance could however not be transported to the prediction of the more diverse data included in the big test set. The remaining event types that exhibit losses regarding the big test set display differences between 0.04 and 0.08 in comparison to the scores reached for the small test set.

Three event types show an increase in F1 score for the big test set prediction. This mitigates the effect of the lower scores reached for the aforementioned event types on the final macro F1 score to an extent. The event types are DIPLOMACY with an increase of 0.11, FINANCE with an increase of 0.08 and NATURAL DISASTER with a less distinct increase of 0.03. ACCIDENT additionally shows no changes between test sets.

The confusion matrix in figure 16 uncovers more details about the errors made by the model during the classification of the big test set. For DISORDER, one of the event types with the greatest difference between test sets, the confusion with the class CRIME is detected as the main error source. 15% of texts belonging to DISORDER are predicted as CRIME by the model. The rest of the errors is distributed over a number of other event types. Among others, ACCIDENT with 7% and GOVERNMENT with 3%. The error rates, however, are smaller.

Errors concerning ORGANISATION, the second event type that shows a notable decrease with regard to the big test set, are less distinct in contrast to DISORDER. The main error source is the confusion with GOVERNMENT with 11%. This is followed by DIPLOMACY with another relatively high rate of 8% and NATURAL DISASTER with 4%. The remainder of ORGANISATION texts that were predicted incorrectly are found with negligible amounts between 2 and 1% for the event types CRIME, DISCOVERY, DISORDER and WIN LOSE.

The third event type holding a low score for the big test set, WIN LOSE, shows a clear picture regarding its errors. Even though confusions with five

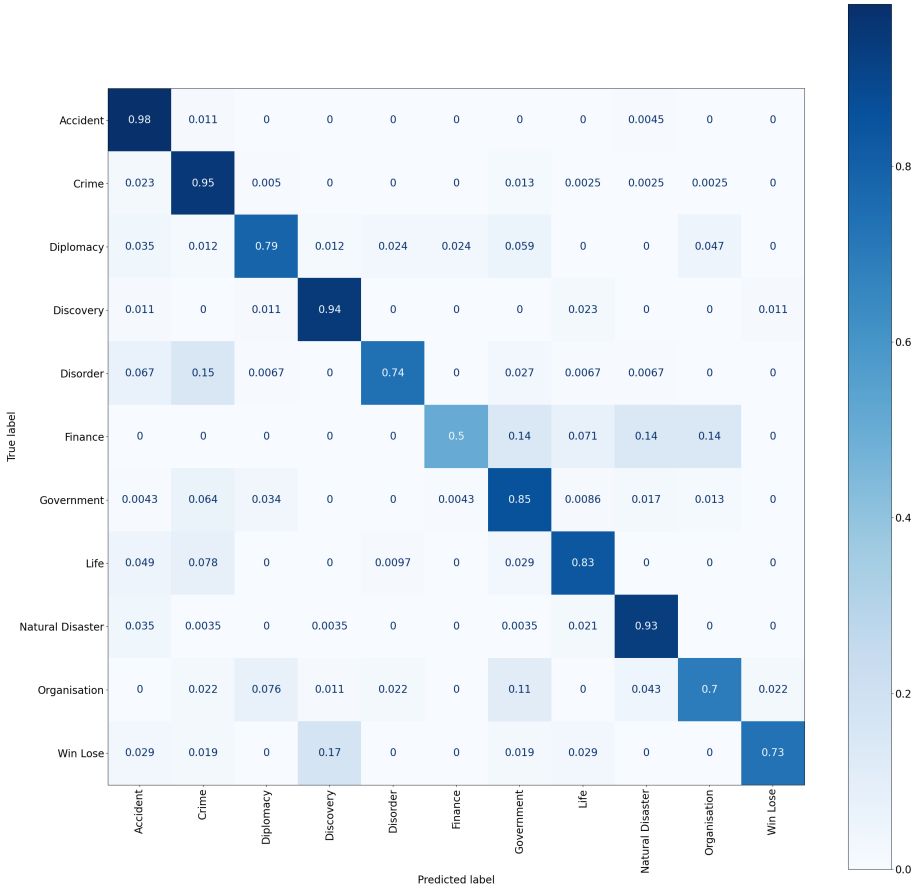

Figure 16: Model 1: confusion matrix big test set predictions

event types are observed, four of those show a percentage between 2 and 3. Only the confusion with DISCOVERY displays a higher error rate of 17% which makes the main challenge that was encountered during the prediction of WIN LOSE clear.

Error Analysis Model 2 The results of model 2 for the small test set show that six event types hold high F1 scores. These are ACCIDENT, CRIME, DIPLOMACY, DISCOVERY, DISORDER and WIN LOSE. Perfect F1 scores of 1 are observed for DISCOVERY, DISORDER and WIN LOSE. Of the remaining event types, FINANCE, GOVERNMENT and ORGANISATION display the lowest scores for the small test set (0.77, 0.81 and 0.85).

For model 2, the difference between the macro F1 scores of the test sets is most striking among all models with 0.1128 (0.9147 vs. 0.8019).

	Small test			**Big test**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Accident	0.96	1.00	0.98	0.96	0.95	0.95
Crime	0.90	0.95	0.93	0.89	0.93	0.91
Diplomacy	1.00	0.86	0.92	0.76	0.76	0.76
Discovery	1.00	1.00	1.00	0.82	0.93	0.87
Disorder	1.00	1.00	1.00	0.97	0.51	0.67
Finance	1.00	0.62	0.77	0.86	0.43	0.57
Government	0.78	0.84	0.81	0.66	0.92	0.77
Life	0.93	0.87	0.90	0.91	0.85	0.88
Natural Disaster	0.87	0.96	0.91	0.90	0.98	0.93
Organisation	1.00	0.73	0.85	0.83	0.65	0.73
Win Lose	1.00	1.00	1.00	1.00	0.63	0.78
Macro Average	**0.95**	**0.89**	**0.91**	**0.87**	**0.78**	**0.80**

Table 9: Model 2: classification reports

The scores reached for the event types that were predicted with an F1 score of 1 for the small test set (DISCOVERY, DISORDER and WIN LOSE) decrease with regard to the big test set. DISCOVERY is predicted with an F1 score of 0.87, DISORDER holds an F1 score of 0.67 and WIN LOSE is predicted with 0.78. Other event types that can be said to be unstable for model 2 are FINANCE and DIPLOMACY with a difference between F1 scores of 0.2 (0.77 vs. 0.57) and 0.16 (0.92 vs. 0.76) respectively.

The confusion matrix provided in figure 17 is used for further analysis of the errors.

For DISORDER it can be seen that for 51% of all texts the correct event type was predicted. However, it also becomes obvious that 27% of the texts were predicted as belonging to the event type GOVERNMENT. The main error source for DISORDER therefore lies in the confusion with the event type GOVERNMENT. A smaller but still noteworthy amount of errors (13%) can be seen for texts that are predicted as CRIME. The event types DISORDER and CRIME are arguably closely related in their content which might explain these mistakes. Problems regarding the classification of DISORDER were also seen for model 1. However, the main errors differ between the models.

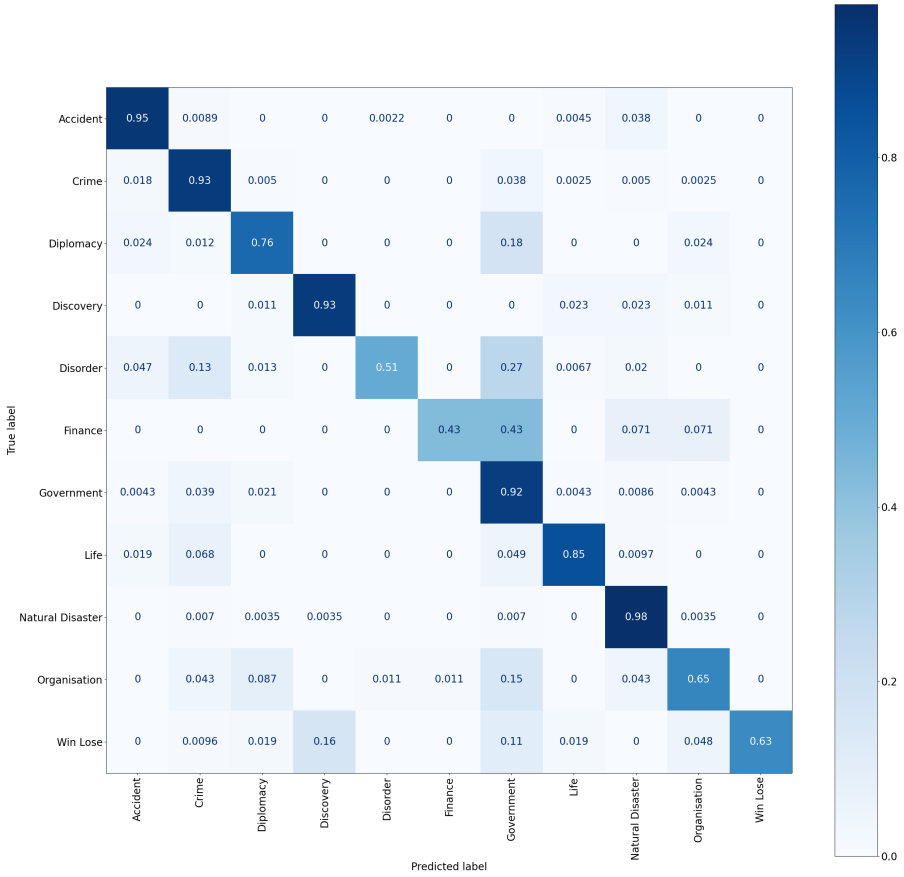

Figure 17: Model 2: confusion matrix big test set predictions

The confusion with GOVERNMENT is less frequent (3%) for model 1. The confusion with CRIME is the main error source.

A similar picture to DISORDER arises for the event type FINANCE. While 43% of texts are classified correctly, another 43% are wrongly predicted to be part of the GOVERNMENT class.

With regard to DIPLOMACY, the model also has the most problems in differentiating from GOVERNMENT. 18% of texts are classified as belonging to GOVERNMENT instead of DIPLOMACY.

For the event type WIN LOSE, 63% of texts are predicted correctly. Errors found for this event type are not as distinctly connected to one other event type as was seen for DISORDER and FINANCE. Rather, they can be found with regard to the event types DISCOVERY and GOVERNMENT in a (more or less) similar amount, 16% and 11% respectively.

A different picture arises with regard to errors connected to the event type

DISCOVERY. 93% of all texts belonging to the event type are classified as such. This is a notably higher rate than was observed for the other problematic event types. The main source of errors here is the just mentioned confusion of WIN LOSE texts with the DISCOVERY class which consequently decreases the precision value of DISCOVERY.

Error Analysis Model 3 In the case of model three, six event types are predicted with F1 scores that can be seen as high: ACCIDENT, CRIME, DISORDER, LIFE, NATURAL DISASTER and WIN LOSE. DISORDER and WIN LOSE reach a score of 1 which contributes positively to the macro F1 score. Event types that influence the final score negatively are DISCOVERY (0.83), FINANCE (0.67), GOVERNMENT (0.84) and ORGANISATION (0.86).

	Small test			**Big test**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Accident	0.97	0.97	0.97	0.96	0.94	0.95
Crime	0.97	1.00	0.98	0.92	0.91	0.92
Diplomacy	0.92	0.86	0.89	0.73	0.87	0.79
Discovery	0.71	1.00	0.83	0.98	0.91	0.94
Disorder	1.00	1.00	1.00	0.97	0.77	0.86
Finance	0.67	0.67	0.67	0.91	0.71	0.80
Government	0.94	0.76	0.84	0.82	0.90	0.86
Life	0.92	1.00	0.96	0.74	0.97	0.84
Natural Disaster	0.95	0.95	0.95	0.91	0.99	0.95
Organisation	0.82	0.90	0.86	0.82	0.64	0.72
Win Lose	1.00	1.00	1.00	0.98	0.80	0.88
Macro Average	**0.90**	**0.92**	**0.90**	**0.88**	**0.85**	**0.86**

Table 10: Model 3: classification reports

For model 3, a difference of 0.0411 between test sets exists (0.9043 vs. 0.8632). This difference is similar to the one observed for model 1.

In the case of model 3, most event types (seven out of eleven) show a decrease in F1 score for the classification of the big test set. This decrease ranges from 0.02 to 0.14. The biggest discrepancies are detected for DISORDER (0.14), LIFE (0.12) and WIN LOSE (0.12). Again, DISORDER and

WIN LOSE were originally classified with an F1 score of 1 for the small test set. The score reached for the event type NATURAL DISASTER stays the same between test sets with a comparatively high value of 0.95.

Three event types show an improvement for the big test set: DISCOVERY, FINANCE and GOVERNMENT. The improvement seen for GOVERNMENT is 0.02. DISCOVERY and FINANCE display a greater increase with 0.11 and 0.13 respectively.

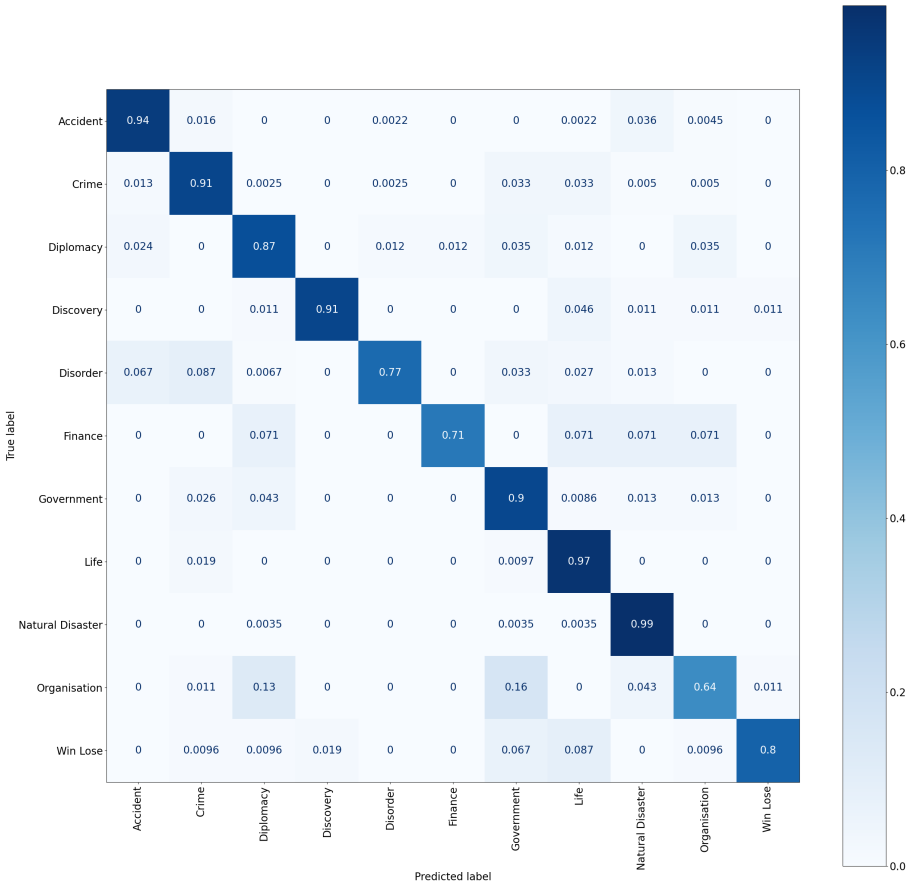

Figure 18: Model 3: confusion matrix big test set predictions

The confusion matrix in figure 18 shows that for DISORDER, errors are spread in similar amounts over a number of event types. No confusion of classes includes more than 10% of DISORDER texts. The highest error rates can be found for CRIME with 9% and ACCIDENT with 7%. This is followed by GOVERNMENT with 3%, LIFE with 2% and NATURAL DISASTER with 1%. This means that while a definite decline in the prediction of

DISORDER is observed, the errors seem to stem to an extent from random confusions instead of systematic problems. The main error sources are the confusion with CRIME as well as ACCIDENT.

Regarding LIFE, the confusion matrix indicates that 97% of texts belonging to the class are predicted as such. The errors for the event type therefore are dominantly found in connection to the precision not the recall score that is focused on here. This can also be seen in the classification report; while recall only decreases by 0.03 between test sets, precision shows a difference of 0.18. This means that texts that belong to other event types but are predicted as LIFE by the model (i.e., false positives) are mainly responsible for the decrease in F1 score. The confusion matrix still hints at this error type even if it is not normalized accordingly. Among others, texts that are part of the event types CRIME, DISCOVERY and FINANCE are classified as LIFE. A small amount of recall-focused errors (i.e., false negatives) can be seen for texts that are wrongly predicted to belong to the event type CRIME (2%).

The classification results of WIN LOSE instances reveal two main errors: the wrong prediction of GOVERNMENT (7%) and the wrong prediction of LIFE (9%). The rest of the errors is mostly smaller than 1%. The confusion with DISCOVERY, which constitutes the main error source for WIN LOSE predictions with regard to the other models, only occurs with 2%.

Error Analysis Model 4 Model 4 achieves comparatively high F1 scores for seven event types: ACCIDENT, CRIME, DIPLOMACY, DISORDER, LIFE, NATURAL DISASTER and WIN LOSE. This time, only one of these event types reaches an F1 score of 1, namely, WIN LOSE. Scores that are on the lower side in the context of the model's performance are connected to FINANCE (0.6) and ORGANISATION (0.84).

For baseline model 4, the test sets show a difference in macro F1 scores of 0.0789 (0.921 vs. 0.8421). As mentioned earlier, model 4 behaves similar to model 2, even though its difference between scores is not as stark.

	Small test			**Big test**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Accident	0.95	1.00	0.97	0.94	0.98	0.96
Crime	0.97	0.94	0.96	0.88	0.92	0.90
Diplomacy	0.92	1.00	0.96	0.73	0.74	0.74
Discovery	0.83	1.00	0.91	0.84	0.97	0.90
Disorder	0.96	0.96	0.96	0.97	0.66	0.79
Finance	1.00	0.60	0.75	0.80	0.57	0.67
Government	0.81	0.95	0.88	0.74	0.91	0.82
Life	1.00	0.90	0.95	0.94	0.84	0.89
Natural Disaster	1.00	0.94	0.97	0.97	0.95	0.96
Organisation	1.00	0.73	0.84	0.84	0.71	0.77
Win Lose	1.00	1.00	1.00	1.00	0.78	0.88
Macro Average	**0.95**	**0.91**	**0.92**	**0.88**	**0.82**	**0.84**

Table 11: Model 4: classification reports

The classification reports exhibit that in the case of model 4, all event types hold lower scores for the big test set prediction. The decrease in individual scores has a wide range from 0.01 to 0.22. The loss in F1 scores is most distinctly seen for DIPLOMACY (0.22), DISORDER (0.17) and WIN LOSE (0.12).

The main error source regarding wrong predictions of DIPLOMACY texts is the confusion with GOVERNMENT. 18% of all texts are wrongly classified to belong to this class. Another, however less striking, error is found for confusions with the event type ORGANISATION. 4% of texts fall into this error category.

For DISORDER, the confusion with CRIME again constitutes the majority of wrongly classified texts with 17%. Next to CRIME, 11% of errors are found for the wrong prediction of GOVERNMENT and another 5% for the confusion with ACCIDENT.

The analysis of the event type WIN LOSE shows that 15% of texts are classified by the model to belong to the event type DISCOVERY. This is by far the greatest source of errors for the class. Other error sources do not exceed 3%.

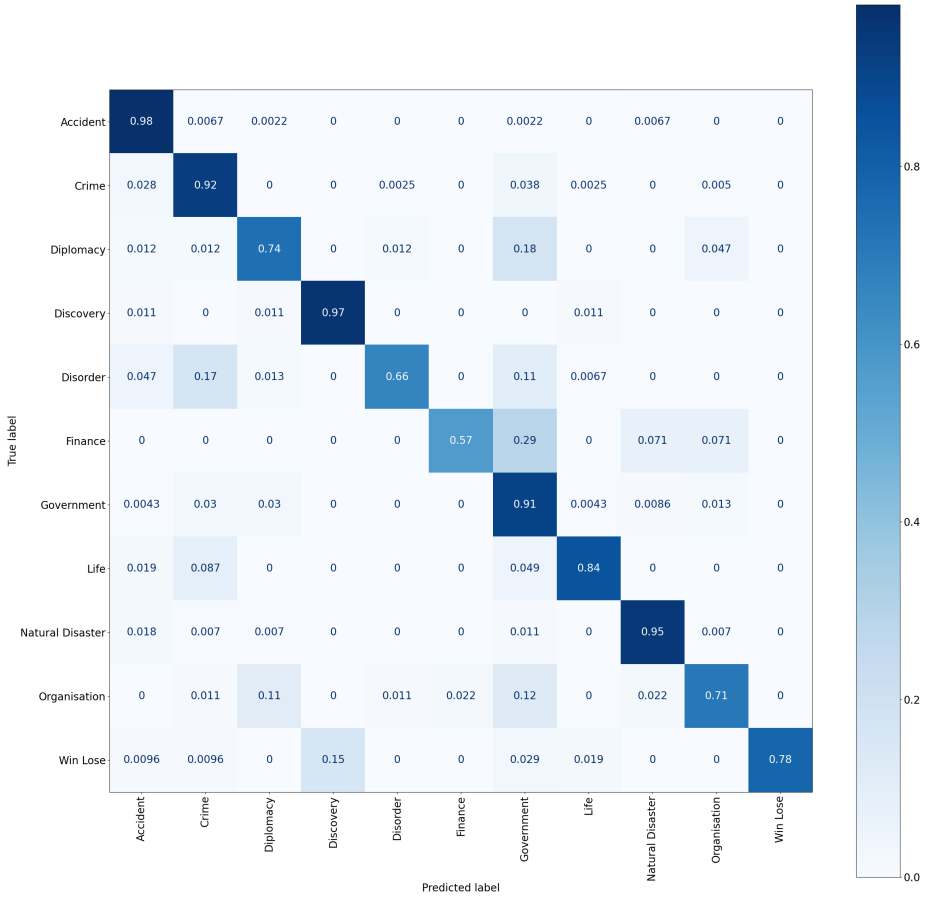

Figure 19: Model 4: confusion matrix big test set predictions

Error Analysis Model 5 For model 5, the classification report of the small test shows seven event types with high F1 scores: ACCIDENT, CRIME, DISCOVERY, DISORDER, GOVERNMENT, NATURAL DISASTER and WIN LOSE. Two event types achieve an F1 score of 1: ACCIDENT and DISORDER. The lowest scores are reached for FINANCE, LIFE and ORGANISATION (0.4, 0.77 and 0.82).

Model 5 shows a difference of 0.0059 (0.8659 vs. 0.86) between the small and the big test set. This is by far the smallest difference between test sets of all models.

For most event types (seven out of eleven) the discrepancies between test sets lie between 0.01 and 0.06. For the remaining 4 classes, namely DISORDER, FINANCE, GOVERNMENT and LIFE, greater differences can be observed. DISORDER shows a decrease of 0.12 when classifying the big test

	Small test			Big test		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Accident	1.00	1.00	1.00	0.96	0.96	0.96
Crime	0.96	0.96	0.96	0.91	0.93	0.92
Diplomacy	0.83	0.83	0.83	0.74	0.85	0.79
Discovery	0.88	1.00	0.93	0.92	0.95	0.94
Disorder	1.00	1.00	1.00	0.91	0.85	0.88
Finance	1.00	0.25	0.40	1.00	0.43	0.60
Government	0.93	0.93	0.93	0.80	0.89	0.84
Life	1.00	0.62	0.77	0.95	0.83	0.89
Natural Disaster	0.95	0.95	0.95	0.95	0.98	0.96
Organisation	0.78	0.88	0.82	0.86	0.67	0.76
Win Lose	0.85	1.00	0.92	0.97	0.89	0.93
Macro Average	**0.93**	**0.86**	**0.87**	**0.91**	**0.84**	**0.86**

Table 12: Model 5: classification reports

set. The confusion matrix in figure 20 shows that the model still predicts 85% of all DISORDER texts correctly. The main error source for the event type is the confusion with CRIME with 9%.

The second notable decrease can be seen for GOVERNMENT with a differences of 0.09 between test sets. There seem to be no distinct systematic errors with regard to the prediction of this event type. 89% of all texts that belong to GOVERNMENT are also predicted as such. Smaller amounts of 3% and 4% of errors can be found for wrong predictions of CRIME and DIPLOMACY respectively. It can be assumed that these errors stem from semantic similarities in the classes. Still, the model is in general capable of predicting texts that belong to the event type correctly. The event type's decrease in F1 score is therefore connected to its precision value (cf. table 12).

For the event types FINANCE and LIFE an increase in F1 scores can be observed when model 5 is used to predict the big test set. FINANCE shows an F1 score of 0.40 for the small test set and a score of 0.60 for the big test set. The difference can be found exclusively in the recall score (0.25 vs. 0.43). Precision is steady with a score of 1 which means that for both test sets no texts that belong to an event type other than FINANCE are predicted to be

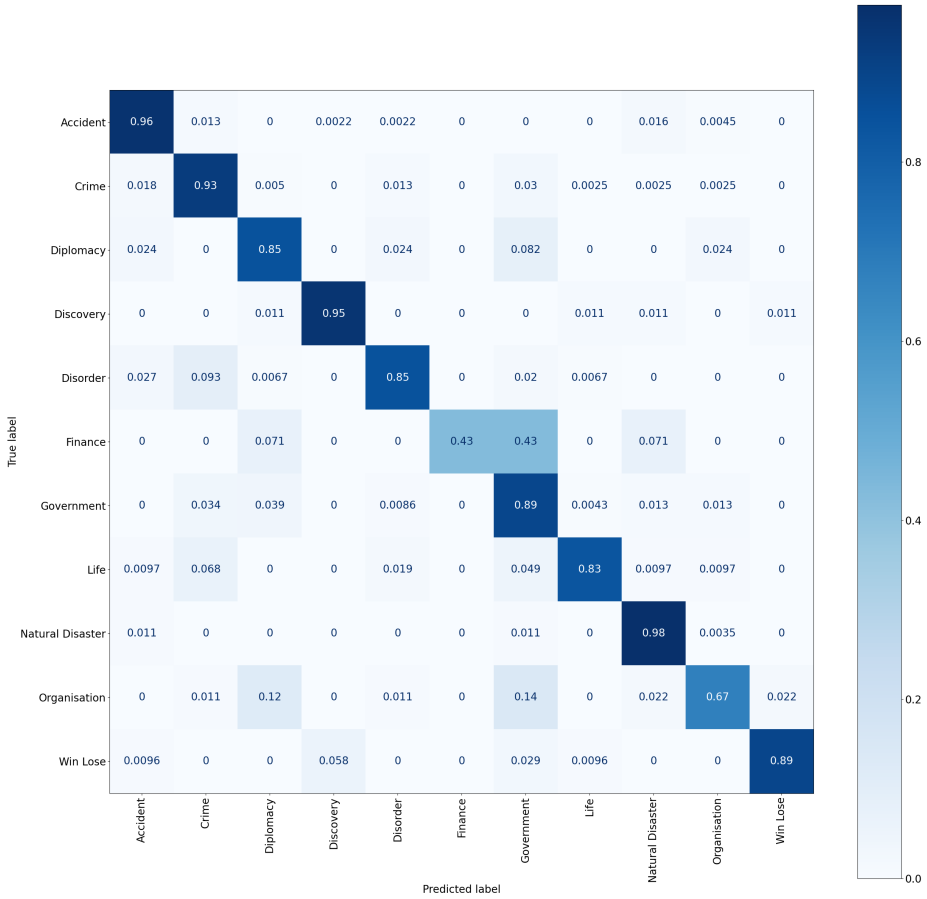

Figure 20: Model 5: confusion matrix big test set predictions

part of the FINANCE class. The increase in recall means that more texts that belong to the FINANCE class are predicted correctly. Still, an obvious problem can be seen in the confusion of the classes FINANCE and GOVERNMENT. This error was also seen for other models. Its recurrence hints at the already described difficulty concerning the overlap of the classes FINANCE and GOVERNMENT in the database.

The increase that is detected for LIFE shows a similar picture regarding precision and recall: The main improvement is seen for the recall scores (0.62 vs 0.83). Precision displays a decrease of 0.05. Errors that are still present for LIFE can be found with 7% and 5% for texts that were wrongly predicted as CRIME and GOVERNMENT respectively.

Model 5 shows comparatively small differences between all event types and in the case of FINANCE and LIFE improves when being used to predict

the big test set.

Summarizing Remarks The previous chapter reported the results of the cross-validation experiment. The performance of all five models was analyzed. As was stated in the beginning of this chapter, models 2 and 5 can be seen as two extremes on different ends of a scale. Model 2 is assumed to be overfitting, while model 5 learns the event type representations in a way that is better transferable to new and more diverse data (i.e. to the big test set). This difference is also detected when comparing the classification reports of models 2 and 5 in more detail.

The errors that are found with regard to the predictions of the two models are insofar similar that semantic similarity in the data leads to systematic confusions. However, the amount of such errors is smaller for model 5. Model 5 shows these smaller differences between all event types and in the case of the event types FINANCE and LIFE even improves when being used to classify the big test set.

This again supports the assumption that model 5 learns representations of the event types that are more transferable to new data as long as it still is of the same type. This however also leads to lower scores for the small test set in comparison to model 2 that appears to be fitted more closely to the training data.

The comparison of models 2 and 5 illustrates the importance of cross-validation and also the value of including a second, slightly more diverse test set. Cross-validation ensures that neither the results seen for model 2 nor the ones for model 5 are overinterpreted. Moreover, the big test set helps in understanding the ability of the models to generalize. Without testing on the big test set, model 2 would have been interpreted to be better than model 5. The notion of overfitting might have been overlooked or at least underestimated.

The analysis of the individual models aimed at understanding the influence that the automatically selected features have on the classification of event

types. To summarize the results, table 13 shows the mean F1 scores that were reached for the event types for both test sets.

	Mean F1 score small test	**Mean F1 score big test**
Accident	0.97	0.95
Crime	0.96	0.91
Diplomacy	0.85	0.77
Discovery	0.92	0.9
Disorder	0.99	0.8
Finance	0.62	0.64
Government	0.88	0.83
Life	0.89	0.87
Natural Disaster	0.94	0.95
Organisation	0.86	0.75
Win Lose	0.98	0.85

Table 13: Mean F1 scores of all event types

The analyses of the models already revealed patterns with regard to the classification of the individual event types. It was seen that a number of event types were consistently predicted with high scores by all five model with regard to the small test sets. These are: ACCIDENT, CRIME, DISORDER and WIN LOSE. The mean F1 scores show this pattern again and include scores ranging from 0.96 to 0.99 for the four types.

Next, the analyses pointed at difficulties with regard to the classification of the event types FINANCE and ORGANISATION for the small test set. All models revealed problems regarding the prediction of FINANCE, four out of five models additionally achieved comparatively low scores for ORGANISATION. The corresponding mean F1 scores are 0.62 for FINANCE and 0.86 for ORGANISATION.

Table 13 also shows a comparatively low score for DIPLOMACY (0.85) with regard to the small test set. This is however not the result of consistently low scores reached by all models. Rather, it is negatively influenced by the event type's F1 score obtained by model 1 (0.67). The score achieved by model 1 diverges from the scores of the other models which range from 0.89

to 0.96.

The detailed error analyses were focused on the classification of the big test set and therefore on the ability of the models to generalize to more diverse data. These analyses showed a systematic confusion of semantically similar event types for all models. Depending on the model and the data split, these confusions are more or less grave. Noteworthy class interchanges that are observed numerous times are

- DIPLOMACY texts classified as GOVERNMENT,
- DISORDER texts classified as CRIME,
- FINANCE texts classified as GOVERNMENT,
- ORGANISATION texts classified as either DIPLOMACY or GOVERNMENT and
- WIN LOSE texts classified as DISCOVERY.

The ability of the models to generalize to new data seems to be particularly difficult for DISORDER and WIN LOSE. Both event types were among those that were predicted with the highest scores for the small test set. However, these high scores were in no case transferred to the big test set. The different event type confusions are invoked again by means of example texts in the following section to support the understanding of the described errors.

The main error found for DIPLOMACY is the wrong prediction of the semantically similar event type GOVERNMENT. The following text shows an example of this confusion.

> The Federaly Administered Tribal Areas (FATA) include seven territorial agencies: Bajaur, Mohmand, Khyber, Orakzai, Kurram, North Waziristan, and South Waziristan. They are mainly rugged, mountainous areas along the border with Afghanistan. The legislation rids the northwestern tribal areas of what were seen as discriminatory laws under which those regions have been governed since the colonial rule of Britain. Pakistani President Mamnoon Hussain has signed legislation that merges the country's tribal regions with the Khyber Pakhtunkhwa Province and therefore grants some 5 million people in the regions the

> same rights as other Pakistanis. Hussain signed the law in Islamabad on May 31. The bill had previously been passed by the lower and upper houses of parliament as well as the assembly in Khyber Pakhtunkhwa with the necessary two-thirds majority. Pakistan was granted independence from London in 1947. Pakistan's government, led by Prime Minister Shahid Khaqan Abbasi, formally steps down on May 31, the end of its five-year term. **DIPLOMACY**

The main event that is described here is a Sign Agreement event which is a subtype of DIPLOMACY. However, information about the Pakistani government is also included in the text. This leads to similarities with the event type GOVERNMENT, specifically the subtype Regime Change. Words that were identified by LIME to be indicative for the classification of GOVERNMENT are present, e.g., *president* and *law* as well as the related word *bill*. All these factors influence the correct classification of the text. As mentioned before, the results obtained by LIME are not global. However, it is assumed that models that are trained on the same database also learn similar representations. Therefore, the LIME results can still be drawn upon.

A text that belongs to the class DISORDER but was classified as CRIME can be seen in the following paragraph.

> Many prisons are overcrowded and allegations of corruption are widespread. Matamoros, just across the border from Brownsville, Texas, has seen an increase in violence as rival drug cartels fight for control of the smuggling routes into the United States. The region is at the centre of a violent battle for territory between the Gulf cartel and the Zetas, who used to work as hitmen for the Gulf cartel but have since become their rivals. The prison system is struggling to cope with the large number of drug offenders jailed in President Felipe Calderon's battle against powerful drug cartels. Police spokesman Hector Walle told Reuters news agency the inmates in Matamoros used homemade weapons in the fight, which quickly spiralled out of control. The Mexican Centre for Research and National Security (Cisen), says more than 28,000 people have been killed in drug-related violence since December 2006. Local media say rival gangs fighting for control of the prison attacked each other with a variety of sharp weapons. Last month, prosecutors accused prison officials at a jail in Gomez Palacio in Durango state of allowing inmates to leave the prison at night to carry out revenge killings, even providing them with the weapons to carry out the attacks. Police in Mexico say at least 14 inmates have been killed in a prison riot in Matamoros, in Tamaulipas state, close to the US border. Soldiers and police eventually regained control of the penitentiary. **DISORDER**

The main event that is annotated in the text is a riot event located in a prison.

Riots are part of the event type DISORDER. However, tokens that indicate a CRIME event are also given in the text. Most notably, the location of the riot event, a prison, can be assumed to skew the classification in the direction of CRIME. As was seen in the LIME output of the first preliminary experiment, *jail* and *police* were identified to hold relevance for the classification of CRIME. The given text does not only mention *jail*, but also semantically similar words, namely *jailed*, *prison* and *penitentiary*. It follows that the context that is given for the riot event influences classification negatively. The classifier is not focused on the main event during prediction.

The event type FINANCE includes the subtypes financial aid as well as financial crisis. Thus, texts annotated with the event type FINANCE can also refer to governmental institutions and financial policies as is shown in the example text below. This might be a reason for the models' difficulties in differentiating the classes.

> Free tuition is great, and California excels at that compared to the rest of the country. But with rents sky high, affordable housing has become the chief expense for most students – and relief is harder to come by. The plan – part of a commitment of $2 billion over three years if the Legislature fully funds it – may seem like a massive sum, but the amount of housing the money can build is likely a rounding error in the total need for the state's students. "It's a drop in the bucket, but every drop counts," said Dana Cuff, a UCLA professor and director of CityLab, an urban design research center. The housing program that lawmakers approved last week is new – part of the heap of surplus cash in the state budget this year. A full-time schedule means graduating faster, but often students can't attend that many classes because of work obligations. Experts say it's "a drop in the bucket" compared to what's needed. That would require another act of the Legislature to change the terms of the housing program, but she doesn't rule out that the promise of cheaper rent may compel more community college students to enroll full-time if their campuses take up the money. Lawmakers have a plan for that: They've poured $500 million into this year's state budget so that public colleges and universities can build affordable housing or renovate existing property. Gov. Gavin Newsom and the Legislature reached a deal that will provide $500 million toward affordable student housing this year and possibly up to $2 billion in future years. Lea este artículo en español. The governor initially proposed $4 billion for student housing but it got halved during negotiations with the Legislature. The deal: If the full-time requirement slows down applications from community colleges, "we can adjust in the future," said Nancy Skinner, a Democrat and state senator from Oakland who chairs the Senate's budget commit-

tee. **FINANCE**

Next to the textual similarity of the event types FINANCE and GOVERNMENT, it also has to be considered that FINANCE constitutes the smallest class in the data. This means that only a small number of samples is available for training and testing. Using only these texts for training does not seem to be sufficient to learn a good representation of the class. This in turn makes its prediction less robust and enforces the problem of confusions with other event types. The comparatively low F1 scores reached for FINANCE underline this. The LIME results already hinted at problems with regard to the textual representation of the event type: Tokens that were determined to be indicative for the event type were *UNHCR*, *development* and *system*. These tokens can not be seen to be characteristic for texts describing FINANCE events.

The main errors connected to the classification of the event type ORGANISATION are the wrong prediction of DIPLOMACY and GOVERNMENT. Most of the time, organisations that are reported on in the dataset are businesses, e.g., Google, Cosco, Marriott, Sonic. The model is predominately confronted with these kinds of reports during training. However, texts dealing with political organisation as for example NATO and the EU are also included. In these texts, words occur that are more often connected to DIPLOMACY and GOVERNMENT which then leads to errors regarding their classification. Two example texts can be seen below.

ORGANISATION text that was classified as DIPLOMACY:

> Speaking in Brussels on Friday NATO Secretary General Jens Stoltenberg said, "North Macedonia is now part of the NATO family, a family of thirty nations and almost one billion people. A family based on the certainty that, no matter what challenges we face, we are all stronger and safer together." North Macedonia is a long-standing contributor to our Euro-Atlantic security, including by participating in NATO-led missions in Afghanistan and in Kosovo. A flag-raising ceremony for North Macedonia will take place at NATO Headquarters on 30 March 2020, in the presence of NATO Secretary General Jens Stoltenberg, Chairman of the NATO Military Committee Air Chief Marshal Sir Stuart Peach, and the Chargé d'Affaires of the Delegation of North Macedonia to NATO, Mr. Zoran Todorov. Today (27 March 2020), North Macedonia became NATO's newest member, upon depositing its instrument of

> accession to the North Atlantic Treaty with the US State Department in Washington DC. NATO Allies signed North Macedonia's Accession Protocol in February 2019, after which all 29 national parliaments voted to ratify the country's membership. The flag of North Macedonia will be simultaneously raised at Allied Command Operations in Mons (Belgium) and at Allied Command Transformation in Norfolk (US). **ORGANISATION**

ORGANISATION text that was classified as GOVERNMENT:

> And there was sadness in Brussels, where British flags were quietly removed from the bloc's many buildings. Thousands of enthusiastic Brexit supporters gathered outside the U.K. Parliament in London cheered as the hour struck. In a recorded message, Prime Minister Boris Johnson called Britain's departure "a moment of real national renewal and change" that would lead to a bolder, more energetic nation. But many Britons mourned the loss of their EU identity. LONDON, UK – Britain has left the European Union after 47 years of membership, taking a leap into the unknown in a historic blow to the bloc. Brexit became official Friday at 11 p.m. (2300GMT) – midnight in Brussels, where the EU is headquartered. **ORGANISATION**

The wrong prediction of DISCOVERY for WIN LOSE texts is mostly due to the subtypes Sports Competition and Break Historical Records. Both subtypes deal, among others, with the topic of sports. Interestingly, the error is not observed in the opposite direction: Texts of the event type DISCOVERY are seldom predicted to belong to the event type WIN LOSE, leading to error rates between 0 and 1%. An example text that was predicted wrong can be seen below.

> Carmen Small of the United States was third, 28.74 seconds behind Van Dijk. The powerful Van Dijk, who picked up gold in the team time trial on the opening day on Sunday, was one of the last to start. Ellen van Dijk of the Netherlands dominated the women's time trial to win her second gold medal at this year's road cycling world championships on Tuesday in Florence, Italy. The 26-year-old led from start to finish along the 21.8km course. The 26-year-old time trial specialist, who was the pre-race favorite, posted a time of 27 minutes, 48.18 seconds, beating Linda Villumsen of New Zealand by 24.10 seconds. Earlier, Igor Decraene of Belgium won the junior time trial, beating favorite Mathias Krigbaum of Denmark. **WIN LOSE**

The text describes a sports event in detail. It can be assumed that the event types WIN LOSE and DISCOVERY, which both include texts about sports, could be differentiated for the small test set because it includes more concise

descriptions of happenings. The prediction of instances belonging to the big test set, however, is a challenge for the models.

To further illustrate the problem, two training instances are included below, one belonging to the WIN LOSE event type, one to DISCOVERY.

> The World Figure Skating Championships is an annual figure skating competition sanctioned by the International Skating Union in which figure skaters compete for the title of World Champion. These were the last World Championships in figure skating before World War I. Men's competitions took place from February 21 to 22 in Helsingfors (Helsinki), Finland. Ladies' competitions took place from January 24 to 25 in St. Moritz, Switzerland. Pairs' competition took place on January 25 also in St. Moritz, Switzerland. **WIN LOSE**

> Covering the distance between hurdles 13 powerful steps at a time, Warholm never came close to breaking stride. Warholm responded a few weeks later by running the 46.70, and breaking the 29-year-old world record held by American Kevin Young since the Barcelona Olympics. Starting in Lane 6, Warholm flew out to the lead, and by the midpoint, he had drawn so far ahead of Benjamin that the real race appeared to be Warholm vs. the clock. Warholm's record came 24 hours before Americans Sydney McLaughlin and Dalilah Muhammad were scheduled to square off in the women's 400 hurdles — a race in which they've broken the world record the last three times they've squared off in a major competition. **DISCOVERY**

The subtle differentiation between the described situations, namely a common competition on the one hand and an outstanding new achievement on the other hand, seems to not be encoded fully. This leads to errors by the models for the classification of the big test set.

The difference between soft and hard news was considered for the construction of the event types (cf. chapter 4.2). The event types DISCOVERY and WIN LOSE represent this differentiation: Texts included in the DISCOVERY class are understood as hard news, i.e., "new findings, discovery or reports regarding a continuing story of great significance" (Lehman-Wilzig and Seletzky 2010: 46). The WIN LOSE class is an example of soft news, i.e. "reports on light or exotic topics [...] that [are] of interest to a narrow segment of the public" (Lehman-Wilzig and Seletzky 2010: 45). Therefore, the individual subtypes were not summarized into one event type even though the

included topics are at times similar to each other.

This summary focused on the main errors that were detected for the models during classification. Table 14 gives a broader overview of event type confusions found for the big test set predictions. Errors that on average exceed a rate of 1% are listed. For the event type NATURAL DISASTER, no mean error rate exceeded 1%. It is therefore not included. The main errors that were identified in this chapter are written in bold.

Event type	**Classified as**	**Mean error rate (%)**
ACCIDENT	NATURAL DISASTER	2
CRIME	ACCIDENT	2
CRIME	GOVERNMENT	3
DIPLOMACY	ACCIDENT	2
DIPLOMACY	**GOVERNMENT**	**11**
DIPLOMACY	ORGANISATION	4
DISCOVERY	LIFE	2
DISORDER	ACCIDENT	5
DISORDER	**CRIME**	**13**
DISORDER	GOVERNMENT	9
FINANCE	DIPLOMACY	3
FINANCE	**GOVERNMENT**	**26**
FINANCE	LIFE	3
FINANCE	NATURAL DISASTER	8
FINANCE	ORGANISATION	7
GOVERNMENT	CRIME	4
GOVERNMENT	DIPLOMACY	3
LIFE	ACCIDENT	2
LIFE	CRIME	7
LIFE	GOVERNMENT	4
ORGANISATION	CRIME	2
ORGANISATION	**DIPLOMACY**	**11**
ORGANISATION	**GOVERNMENT**	**14**
WIN LOSE	**DISCOVERY**	**11**
WIN LOSE	GOVERNMENT	5
WIN LOSE	LIFE	3

Table 14: Mean error rates

5.3 Research Questions and Hypothesis

The two preliminary experiments revealed insights into the probable textual features that are focused on by the fine-tuned RoBERTa models for the task of event type classification. In addition, problems that arise with regard to the classification of unknown data as such as well as the classification of more diverse data as included in the big test set were identified.

Systematic errors between semantically similar classes were observed for all models when the big test set was classified. These errors are more or less problematic, depending on the model. Model 2, which is assumed to be overfitting, encounters the most difficulties. The confusions of specific event types constitute the main problem for correct classification.

The assumptions arises that the given errors are a side effect of the automatic selection of features. Classification is mostly driven by nouns (cf. chapter 5.1.1), which have overlaps between event types, and to a lesser extent by verbs. This leads to the consideration that manually created features might be a helpful supplement for reaching more reliable classification results. The advantage of manually created features is that they are more controlled than the ones learned automatically and consequently ensured to fit the task.

The remainder of this work is focused on a number of research questions that arise on the basis of the preliminary experiments.

1. Do transformer models like RoBERTa benefit from the explicit addition of manually chosen features for the task of event type classification?

2. Which features are sensible for the task of event type classification?

3. Systematic errors represent shortcomings of the chosen model with regard to solving the task with the given data, i.e., how capable is the model as such to work with the database/task. Do the

added features target the identified main error sources or are other effects observed?

In contrast to traditional feature engineering as described in chapter 2.1, the additional features in this case are not supposed to be the only decisive factor for classification. Rather, the very strong contextual embeddings provided by transformer-based LLMs are exploited for the processing of the features. The features are meant to supply additional patterns in the texts that are indicative for the individual event types. This means that it is not assumed that the features play a greater role than the general content of the text, but that they can enrich the content in a targeted and meaningful way.

The H_1 and corresponding H_0 hypotheses that will be analyzed systemically later on with the help of paired bootstrap testing are:

H_1: A model that is supplied with additional linguistic features during training is better than a model that is trained without additional linguistic features.

H_0: A model that is supplied with additional linguistic features during training is not better than a model that is trained without additional linguistic features.

6 Feature Description

The preliminary experiments revealed a number of shortcomings for the classification of event types implemented with RoBERTa. These shortcomings need to be targeted. This chapter describes three linguistic features that are assumed to support event type classification. The way in which the linguistic features are made available for model training is adopted from Pritzkau (2023) and Pritzkau (2024).

Pritzkau combines transfomer-based classification with a set of manually constructed features derived from domain knowledge. The use case that is considered in his work is the classification of propaganda techniques in tweets (cf. Pritzkau 2023:1). The author identifies features that are meant to support this use case. Features are, on the one hand, focused on metadata available for the tweets, i.e., the tweet's country of origin as well as engagements with the tweet. On the other hand, emotion and sentiment information is extracted from the tweets and used as a feature (cf. Pritzkau 2023: 4).

After identifying the features, "[i]nstead of making technical changes to the architecture, the original message was augmented with the additional features in text form" (Pritzkau 2023: 4). Thereby, the features are made available for processing with the transformer model. The work at hand will utilize this method and add linguistic features to the event data explicitly in text form. In contrast to the work conducted by Pritzkau (2023), all features that are chosen for this work are extracted from the texts themselves. No features based on external data are included.

Domain knowledge which lends itself as a source of possible features for event type classification, can be found in EE research as described in chapter 3 as well as linguistic theory. Three features were selected for this work. Those are: aspectual classes as introduced by Vendler (1957), named entities as well as sentiment. All features are compiled automatically and added to the event data. The following chapter introduces the features in more detail.

At first, aspectual classes are described in the context of linguistic theory. In addition, their classification is formulated as a task in the context of NLP. Open source models that are capable of automatically classifying aspectual classes out of the box are rare and oftentimes not transparent with regard to their training data. Therefore, a new classifier is trained for the task. The process of data selection, training and evaluation will be included in chapter 6.2.

After that, the concepts of named entity recognition and sentiment analysis are introduced shortly in chapters 6.3 and 6.4. These two tasks are well known in the research area of NLP. Accordingly, established open source models are available. These models will be used for the annotation of the event data. This means that in contrast to the aspectual class feature, no new models need to be trained.

6.1 Aspectual Classes – Theoretical Background

The concept of sorting verbs into different time schemata, referred to as aspectual classes, was introduced by Vendler (1957). Vendler's work is situated in the area of lexical semantics and focuses on "the particular way in which [a] verb presupposes and involves the notion of time [implicitly]" (1957: 143). He establishes four distinct aspectual classes of verbs: state, activity, accomplishment and achievement. These classes can be differentiated with the help of several characteristics. Those characteristics are dynamicity, duration, and telicity.

- Dynamicity refers to a change of state.
- Duration refers to the amount of time passed (extended vs. punctual).
- Telicity refers to the presence of an inherent end point.

These properties, and thus the aspectual classes themselves, impart temporal information and can also be referred to as "inner aspect [which stands] in

contrast to outer aspect, which deals with temporal operators like Progressive, Perfect, and Habitual" (Mittwoch 2019: 30).

For the classification of event types, it is hypothesized that aspectual classes can be used as an additional linguistic features to improve model performance. The different event types include texts about situations that are similar to each other. It is assumed that situations that are similar to each other with regard to their textual description also exhibit similarities with regard to their internal temporal structure as expressed by aspectual classes. Texts belonging to the event type ACCIDENT, for example, likely focus on situations that are punctual and have an inherent end point:

> On January 8, 1959, about 2032 e.s.t., a DC-3, Southeast Airlines Flight 308, struck a mountainside during an ILS approach to the Tri-City Airport, Tennessee. **ACCIDENT – Air Crash**

Texts of the type NATURAL DISASTER, on the other hand, might be more often characterized by stative descriptions of the disaster:

> The storm has maximum sustained winds of 105 kilometers per hour near the center and gustiness of up to 135 kph. **NATURAL DISASTER – Storm**

While LLMs are capable of classifying event types based on content words (especially nouns), it is possible that rather implicit features such as dynamicity, duration and telicity are not considered (sufficiently) during training. Encoding these features in form of the distribution of aspectual classes in a text provides additional patterns for the classification of the respective event types.

6.1.1 Vendler's Aspectual Classes

Vendler's (1957) original attempt to sort verbs into so-called aspectual classes has since been criticized as well as modified by researchers. However, it still

finds application today. His work can be seen as a foundational study regarding aspectual classes. The classification scheme by Vendler is oftentimes the basis for research on aspectual classes in the context of NLP today, as will be seen later on. Therefore, it will also be used in this thesis and discussed in more detail in the following chapter.

Vendler's work deals with "the most important time schemata implied by the use of english verbs" (Vendler 1957: 144). In his article 'Verbs and Times' he proposes that "considerations involving the concept of time are relevant to the [...] use [of verbs]" and that next to the inference that can be made by considering grammatical tenses, "the use of a verb may also suggest the particular way in which that verb presupposes and involves the notion of time" (Vendler 1957: 143).

Formally, Vendler summarizes the four time schemata underlying his classes as follows (cf. 1957: 149):

> For activities: "*A* was running at time *t*" means that time instant *t* is on *a* time stretch throughout which *A* was running.
>
> For accomplishments: "*A* was drawing a circle at *t*" means that *t* is on *the* time stretch in which *A* drew that circle.
>
> For achievements: "*A* won a race between t_1 and t_2" means that *the* time instant at which *A* won that race is between t_1 and t_2.
>
> For states: "*A* loves somebody from t_1 to t_2" means that at *any* instant between t_1 and t_2 *A* loved that person.

As a first discriminating feature for sorting verbs into these categories, Vendler considers the ability of a verb to possess continuous tense (i.e. the progressive). This leads to a differentiation between verbs that can and verbs that cannot possess said tense. To illustrate the difference, the following exemplary answers to the question *What are you doing?* provided by Vendler (1957: 144) can be considered:

1. I am running (or writing, working, and so on).
    - Can possess continuous tense
2. I am knowing (or loving, recognizing, and so on).

- Can not possess continuous tense

Next, a differentiation in the class of verbs that can possess continuous tense is made by means of analyzing the presence or absence of an inherent terminal point. This corresponds to the feature of telicity. Vendler (1957: 145) illustrates this difference with the following examples:

3. Someone is running or pushing a cart.

    - The statement does not imply any assumption as to how long that running or pushing will go on.

4. Someone is running a mile [...] or drawing a circle.

    - The person who is running a mile will keep on running till he has covered the mile, the person drawing a circle will keep drawing till he has drawn the circle.

In the given examples, the difference between sentences that include an inherent terminal point and those that do not becomes apparent: "Running a mile and drawing a circle have to be finished, while it does not make sense to talk of finishing running or pushing a cart" (Vendler 1957: 145). Similarly to testing for continuous tense with a set of questions, Vendler introduces another set for examining the possible climax of a sentence. These are: *For how long [...]?* and *How long did it take to [...]?*.

For Vendler, these considerations lead to another characterizing difference, the one of homogeneity of time. While for the first class of verbs "any part of the process is of the same nature as the whole" (Vendler 1957: 146), the second class of verbs "proceed towards a terminus which is logically necessary to their being what they are" (Vendler 1957: 146). Consequently, for the second class of verbs, the single parts of the process do not correspond to the whole process.

Vendler calls the first type of verbs (running, pushing a cart) *activity terms* and the second type (running a mile, drawing a circle) *accomplishment terms* (cf. Vendler 1957: 146).

Turning to verbs that can not posses continuous tense according to Vendler, the discerning factor in this group is for how long they can be predicated (i.e. the expressed duration). Vendler states that "some of these verbs can be predicated only for single moments in time [...], while others can be predicated for shorter or longer periods of time" (1957: 146). The following examples provided by Vendler illustrate the difference along with a set of questions:

5. At what moment did you reach the top? At noon sharp.

    - A single moment in time

6. For how long did you love her? For three years.

    - A longer period of time

These two types of verbs are called *achievement terms* (reach the top) and *state terms* (love) (cf. Vendler 1957: 147).

It has to be noted that more recent research reveals that what Vendler calls *state terms* can, at times, also possess continuous tense. Used in this way, continuous tense can on the one hand dynamicize a situation (e.g., John is being polite) or, on the other hand, be used to describe states of affairs (e.g, John was sitting in the chair) (cf. Kranich 2010: 50).

Considering the parameters stated by Vendler, the four aspectual classes of verbs can be differentiated by three features: the ability of a verb to possess continuous tense, the presence or absence of a terminal point and for how long they can be predicated. As was already shown, however, Vendler's original approach evoked numerous reactions and has been analyzed, modified and criticized many times since its publication.

6.1.2 Dowty's Aspectual Classes

Dowty (1979), for example, criticizes Vendler's unit of annotation. While Vendler argues that verbs as such can be sorted into the given time schemata, Dowty states that "not just verbs but in fact whole verb phrases must be taken into account [and that] in a certain sense, even whole sentences are involved" (1979: 62). Dowty based this on considering the variability of a verb's affiliation with a specific aspectual class according to the context it appears in. This can already be seen in Vendler's examples: *Running* and *running a mile* elicit different aspectual class categories. Today's research on aspectual classes widely accepts Dowty's notes on considering at least the VP when considering aspectual classes.

Additionally, Dowty criticizes the progressive test that was described by Vendler as a means of differentiating activities and accomplishments from achievements. He describes that what Vendler calls achievements can also appear in the progressive, thereby "effectively undermin[ing] Vendler's key diagnostic test for the separation between achievements and accomplishments, which is one of the most criticized weaknesses of [Vendler's] classification" (Filip 2011: 1190). In connection to the progressive Dowty additionally describes the so-called imperfective paradox (cf. Dowty 1977) which is concerned with the class of accomplishment verbs. He states that when accomplishment verbs appear in the progressive, the result-state that is normally connected to them due to their inherent terminal point, can no longer be entailed (cf. Dowty 1977: 46).

Considering Dowty's and Vendler's work, figure 22 depicts the aspectual classes and their decisive features with the help of a decision tree.
Next to revising and extending Vendler's considerations, Dowty (1979: 60) presents a set of tests that can aid the annotation of aspectual classes (cf. figure 21)

Criterion	States	Activities	Accomplishments	Achievements
1. meets non-stative tests	no	yes	yes	?[7]
2. has habitual interpretation in simple present tense:	no	yes	yes	yes
3. *φ for an hour, spend an hour φing*:	OK	OK	OK	bad
4. *φ in an hour, take an hour to φ*:	bad	bad	OK	OK
5. *φ for an hour* entails *φ at all times in the hour*:	yes	yes	no	d.n.a.
6. *x is φing* entails *x has φed*:	d.n.a.	yes	no	d.n.a.[8]
7. complement of *stop*:	OK	OK	OK	bad
8. complement of *finish*:	bad	bad	OK	bad
9. ambiguity with *almost*:	no	no	yes	no
10. *x φed in an hour* entails *x was φing during that hour*:	d.n.a.	d.n.a.	yes	no
11. occurs with *studiously, attentively, carefully,* etc.	bad	OK	OK	bad

OK = the sentence is grammatical, semantically normal
bad = the sentence is ungrammatical, semantically anomalous
d.n.a. = the test does not apply to verbs of this class.

Figure 21: Dowty's tests for determining aspectual classes

Some of these tests were developed by Vendler, as can be seen in the previous chapter, others are build on Lakoff's (1966) (stative vs. non stative tests) and Kenny's (1963) work (entailments from the progressive to the non-progressive tense) and again others were developed by Dowty himself.

The tests do not necessarily result into one distinct category, but rather "distinguish subsets of the four categories" (Dowty 1979: 60). They can be drawn upon as a general aid for annotating aspectual classes.

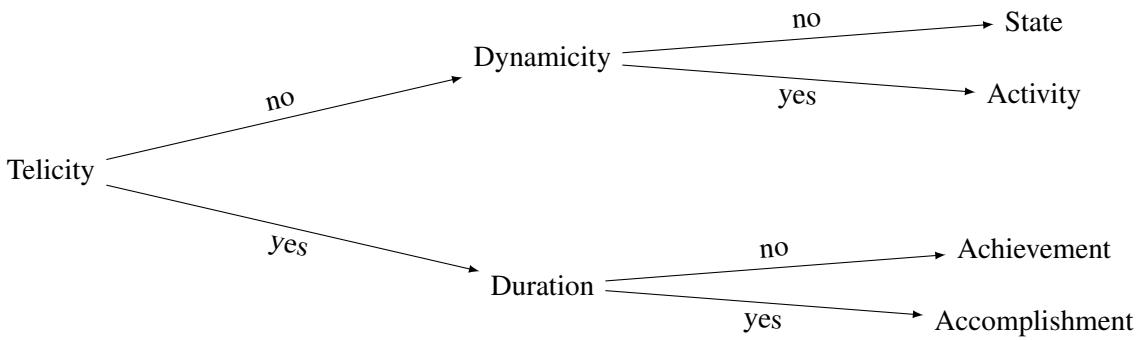

Figure 22: Aspectual classes decision tree

6.1.3 Further Additions/Revisions of the Annotation Scheme

Research on aspectual classes is ongoing in the area of linguistics. While this research furthered the understanding of aspectual classes as such, it will not be elaborated on in detail due to the scope of and relevance for this work. An overview of additional milestones with regard to the categorization of aspectual classes is included in table 15.

Author	Publication	Main contribution
Comrie (1976)	Aspect: An Introduction to the Study of Verbal Aspect and Related Problems	Description of an additional aspectual class, so-called semelfactives. Semelfactives can be seen as atelic achievements.
Smith (1997)	The Parameter of Aspect	
Carlson (1977a), Carlson (1977b)	Reference to Kinds in English, A unified analysis of the English bare plural	Introduction of two kinds of states expressed by the differentiation of individual- and stage-level predicates.
Moens and Steedman (1988)	Temporal Ontology and Temporal Reference	Proposition of another classification scheme based on "two dimensions, one concerned with the contrast between punctuality and temporal extension, the other with the association with a consequent state" (1988: 17). Description of an aspectual type transition network focused on possible aspectual class coercion.
Kranich (2010)	The Progressive in Modern English: A Corpus-Based Study of Grammaticalization and Related Changes	Analysis of the development of the progressive including a description of its role for the categorization of sentences according to aspectual classes.

Table 15: Revisions/additions of aspectual class categorization

6.1.4 Research in NLP

When considering the automatic classification of aspectual classes in the field of NLP, the focus tends to be on Vendler's classes as well as their characterizing features. Revisions like Dowty's notes on the ability of achievements to possess continuous tense are included. The characterizing features, as shown in table 16, can be adopted for automatic processing. The classification of aspectual classes is typically implemented on the sentence level.

	Dynamic	Durative	Telic
State	-	+	-
Activity	+	+	-
Accomplishment	+	+	+
Achievement	+	-	+

Table 16: Features for aspectual class classification

Research focuses on different levels when formulating the categorization of sentences according to aspectual classes as an automatic classification task. On the one hand, single characterizing features can be classified, on the other hand, aspectual classes as such are the object of classification.

Klavans and Chodorow (1992) and Brent (1991), for example, focus on the classification of stativity, Siegel and McKeown (1996) consider the classification of telicity. These earlier works use handcrafted rules that focus on linguistic markers for classification.

Klavans and Chodorow utilize "the tendency of a verb to occur in the progressive [in order to infer] [...] a value for stativity" (1992: 1129). This value is ascertained by "determining the ratio of total frequency of occurrence of a verb in a progressive usage, past or present, over its frequency as a verb in the same corpus" (Klavans and Chodorow 1992: 1129).

Brent analyses two linguistic cues for the detection of stativity. Those are: the percentage of a verb's occurrence in the progressive and the percentage of a verb's modification by rate adverbs (e.g. *quickly* and *slowly*) (1991: 223).

Finally, Siegel and McKeown (1996) use clausal similarity for the classi-

fication of telicity. They argue that the same aspectual class can be found for clauses that are similar to each other. Additionally, they consider a set of aspectual indicators, namely, *not progressive tense*, *perfect and not progressive* and *tense is past or present participle*.

Hermes et al. (2015), Friedrich et al. (2016) and Friedrich and Gateva (2017) base their work on manually constructed features but are able to use these features in combination with more powerful algorithms that were available at this point in time (support vector machine (SVM), conditional random field (CRF) and logistic regression, respectively). Hermes et al. (2015) aim at the classification of Vendler's classes for German verbs. To this end, they operationalize "'aspectivity' utilizing distributional information about nominal fillers in the argument positions of verbs in combination with aspectual features automatically derived from dependency information" (Hermes et al. 2015: 122). They find that the usage of explicit aspectual features based on Vendler's work aids classification the most. Aspectual features used here are, among others, time adverbials for durations (*minutenlang*), time units (*Jahrhundert*) and adverbials (*fast/beinahe*) (cf. Hermes et al. 2015).

Friedrich et al. (2016) and Friedrich and Gateva (2017) use the same set of features in their work. On the one hand, the authors incorporate features commonly found in NLP research, for example POS tags. On the other hand, they employ "syntactic-semantic properties of the main verb and main referent, as well as properties of the clause which indicate its aspectual nature" (Friedrich et al. 2016: 1760). Specific linguistic features that were incorporated are presence of adverbs/prepositional clauses, WordNet sense and supersense, dependency relations, voice and tense (cf. Friedrich et al. 2016: 1760).

In recent years, NLP research on aspectual classes focuses on the usage of transformer models and their capability to solve the task. Metheniti et al. (2022), and in more detail in her dissertation Metheniti (2023), analyze the question of classification capability for French and English with a focus

on BERT for the English data. Specifically, telicity and duration are classified in a binary manner. In this case, the contextual embeddings provided by transformer models were only supplemented with one additional feature: verb position. The authors find that in general "transformer models were very successful on [sic] the classification of aspect even when trained on small datasets" (Metheniti et al. 2022: 96). Additionally, they state that the inclusion of information on verb position can support classification. They also mention difficulties in classification, e.g., the processing of complex sentences (cf. Metheniti et al. 2022: 96). As Metheniti et al. (2022) and Metheniti (2023) use a similar methodology to the one proposed for this work, their results will be discussed in more detail in chapter 6.2.4.

6.2 Classification of Aspectual Classes

As mentioned before, the need for a custom aspectual class classifier was identified. Next to the general problem of availability of open source models, another reason for training a new classifier is that the data that was used to train models that are in principle available is engineered to fit the task. This means that simple sentences are used to train and test the models. Examples of such simple sentences are included in the next chapter which describes open source corpora that are annotated with aspectual class features.

These corpora will also be used for training and testing the aspectual class classifier in this work. However, in order to later on classify sentences incorporated in the event dataset, a small subset of sentences taken from online newspaper articles is annotated manually and used for model training as well. These sentences are complex in character.

6.2.1 Open Source Corpora

Open source corpora that can be used for the task at hand differ with regard to their annotations. Some are annotated for telicity and duration, others for telicity and dynamicity. To be able to use differently annotated datasets for

training, annotations need to be consolidated. Therefore, they are summarized into Vendler's aspectual classes. The classes can be constructed from different combinations of dynamicity, duration and telicity (cf. table 16).

Captions Corpus An open source corpus that is used for the work at hand is the Captions corpus[5] published by Alikhani and Stone (2018). As the name suggests, the texts found in the corpus are captions of images. The Captions corpus is annotated with telicity and duration. Data included in the corpus comes from five different collections of image captions. These are Microsoft Common Objects in Context (COCO) (Lin et al. 2015), Flickr30K (Young et al. 2014), Visual Storytelling (VIST) (Huang et al. 2016), Google's Conceptual Captions (CC) (Sharma et al. 2018) and RecipeQA (Yagcioglu et al. 2018). Additionally, instances from the American National Corpus (ANC) (Ide and Suderman 2004) are included. Alikhani and Stone (2018) use the ANC as a reference point for their analysis. All datasets are annotated in the same manner. The total number of entries in the corpus is 2946. After deleting empty rows, 2733 remain.

The data was analyzed before it was used for model training. A manual analysis regarding the data quality was conducted in a first step.

It became obvious that the RecipeQA subcorpus is not usable in large parts. This is due to the fact that numerous annotations were not done on the sentence level. This means that one annotation corresponds to two (or more) sentences. This is problematic as sentences contradict each other regarding their aspectual classes. Examples of this can be seen below. All following examples include their original annotations for telicity and duration.

(1) Use a spoon or butter knife to distribute jelly onto the opposite still plain bread slice. This ensures that when you put the sandwich together the jelly won't slip off, or if you put the jelly on the already topped slice of bread, it might slip off of that as well

[5] https://github.com/malihealikhani/Captions/tree/master

before you get the chance to put the other slice on top. **atelic durative**

(2) Easy Easy!! Are you in the mood for pizza? Don't want to or don't know how to make pizza dough? Just use a piece of bread as the pizza base and top with all your favourite pizza toppings! Bake it in the oven until the cheese is melted. **atelic durative**

Sentences that are not fitting for the task because they, for example, include no verb, are deleted from the corpus. These instances were already identified by the authors and annotated with n/a, however, they are still included in the Captions corpus.

(3) An abandon building. **n/a**

(4) A long corridor of trees. **n/a**

Next to problems regarding the quality of the data, problems regarding the annotations were identified. Faulty annotations were deleted along with the corresponding sentences.

(5) A boy is jumping into a lake. **atelic durative**

Example (5) is telic. The end point of the action, being in the lake, is explicitly coded.

Additionally, a pattern for the wrong annotation of state was found. The presence of *be* as an auxiliary verb seems to complicate the annotation process.

(6) The woman is ice skating and pulling the little girl in the cart across the ice. **stative**

(7) A man is carefully chiseling ice. **stative**

(8) This young man is rock climbing. **stative**

Even though the syntactic structure of *subject is*, as seen in example (9)

(9) The red panted cyclist is amongst nature. **stative**

is indicative for the presence of a state, this does not hold true if it is used for constructing the progressive tense as is the case in examples (6)-(8).

After the manual analysis, remaining duplicates were removed automatically. In summary, the following preprocessing steps were implemented for the Captions Corpus:

Preprocessing step	**Corpus size (sentences)**
Original Size	2946
Deletion of Empty Rows	2733
Manual Analysis	1191
Duplicate Deletion	1177

Table 17: Preprocessing steps Captions corpus

The goal is to use this corpus as training data for a classifier that classifies Vendler's aspectual classes. Therefore, the annotations of telicity and duration need to be combined into one class per sentence. The following distinct combinations arise:

- **Original annotations:** atelic + durative **Summarized annotation:** activity
- **Original annotations:** telic + durative **Summarized annotation:** accomplishment
- **Original annotations:** telic + punctual **Summarized annotation:** achievement
- **Original annotation:** state or stative **Summarized annotation:** state

Metheniti Telicity/Duration Another dataset was constructed by Metheniti et al. (2022). For their experiments regarding the classification of telicity and duration using transformers, they trained their models, among others, with the already introduced Captions dataset.

For testing model performance, Metheniti et al. (2022) created a small dataset of 160 sentences annotated for telicity. After duplicate removal, the final number of sentences is 147. In addition, they created an even smaller test set for dynamicity which includes 40 sentences. The authors use sentences that can be seen as typical examples of displaying the difference between telic and atelic as well as dynamic and stative respectively. The telicity dataset will be used for testing the models trained in this work in order to compare them to the literature.

6.2.2 Manually Annotated Data

A small number of sentences was annotated manually according to Vendler's classes. The database consists of sentences taken from the All the News Corpus (McKinley 2021). The corpus includes newspaper articles. The choice was made to use a newspaper corpus for the manual annotation of aspectual classes as the event type corpus that will be used for classification also includes data from this genre. It is reasoned that the classification of the event type data can benefit from the usage of a model that has already seen similar data during training.

Sentences were annotated by two researchers. Dowty's tests were used as an aid for annotation. After annotation, the inter annotator agreement was assessed. Sentences where the researches had diverging opinions regarding the fitting class, were discussed. If possible, an agreement was reached and a final label was assigned.

Summarizing the Captions corpus and the manually annotated data, 1346 sentences are included in the final corpus. The distribution of classes can be seen in table 18.

6.2.3 Implementation and Results

Aspectual class classification can be approached from two sides. On the one hand, a multiclass classification approach can be implemented which aims at

Aspectual class	Instances
State	326
Activity	714
Accomplishment	111
Achievement	195

Table 18: Final distribution of aspectual classes

classifying all four aspectual classes. On the other hand, a two-step binary approach can be drawn upon. In this case, a binary classifier is trained for each feature (dynamicity, duration, telicity; cf. table 16). The output of a binary classifier is either 1, which means that an instance belongs to the class that is considered at the moment, or 0 which means that it does not belong to the class. The unique combination of features per aspectual class can be understood as a kind of decision tree (cf. figure 22) that is used for the next steps. As can be seen in figure 22, the two-step binary approach starts with the classification of telicity. Depending on this prediction, either dynamicity or duration is classified next. In the end, all four aspectual classes can be determined by the output.

For both approaches, a combination of the Captions corpus and the data that was annotated manually is used for training and testing. To compare the results to the literature, the telicity test set provided by Metheniti et al. (2022) is also used for evaluation. BERT as well as RoBERTa are tested with regard to their suitability for the task.

A number of questions follow these theoretical considerations:

1. Are there notable differences in performance between BERT and RoBERTa?

2. Is there an advantage of one of the described approaches (multi-class classification vs. two-step binary classification)?

Multiclass Classification BERT as well as RoBERTa were trained for the multiclass classification of aspectual classes. For both models, the hyperpa-

rameters were kept the same:

- Epochs: 5
- Learning Rate: 3e-5
- Batch Size: 8

The results showed that BERT reached a macro F1 score of 0.82 while RoBERTa only reached a macro F1 of 0.74. This means that a difference of 0.08 between model performance was observed. This difference was tested for statistical significance by means of paired bootstrap testing. The p-value amounted to 0.0596. This means that the observed difference borders on significance. Accordingly, it was decided to move on with the BERT implementation for the remainder of this work.

Two-step Binary Classification For the two-step binary approach, three classifiers were trained with BERT and RoBERTa, one for each feature (dynamicity, duration, telicity). Hyperparameters stay the same. The results are shown in table 19.

	BERT	**RoBERTa**	**Effect size**	**p-value**
Dynamicity	0.89	0.87	0.02	0.22
Duration	0.89	0.86	0.03	0.33
Telicity	0.96	0.89	0.07	0.07
Two-step binary macro F1	0.81			

Table 19: Two-step binary approach results BERT and RoBERTa

Again, BERT reaches higher scores than RoBERTa for all features. However, no significance is reached. BERT was again chosen as the more suitable model. To eventually determine the aspectual class labels, the three binary models are used in a pipeline as depicted in figure 22. This results in a final macro F1 score of 0.81 for the aspectual class classification. Error propagation in the first step of the two-step binary classifier, i.e., in the telicity

classification was considered. 1.89% sentences were found to be classified wrong.

6.2.4 Analysis – Telicity Test Set

The models resulting from the described approaches are tested with the telicity test set provided by Metheniti et al. (2022). As this test set only holds a small amount of data (147 sentences), a manual error analysis on the sentence level was feasible. Of the 147 sentences, 76 are annotated as telic, 71 as atelic.

The dataset is annotated in a binary manner (0 and 1). Normally, an annotation of 0 represents the negative class, an annotation of 1 the positive class. For this dataset, however, 0 represents the positive class, i.e., a sentence is telic, 1 represents the negative class, i.e, a sentence is not telic. The predictions of the trained multiclass model are based on the four classes by Vendler. Therefore, the predicted labels and the gold standard labels are matched in the following way:

- **Gold standard label:** Telic **Trained model labels:** Accomplishment and Achievement
- **Gold standard label:** Atelic **Trained model labels:** Activity and State

127 sentences were classified correctly by the multiclass model using BERT, bringing the accuracy to 86.39%. 20 sentences were classified wrong. The accuracy score is in line with the results obtained by Metheniti (2023). She found that the same pre-trained BERT model as was used here reaches an accuracy of 81% with the unmodified training data and 87% after adding information about verb position (Metheniti 2023: 116).

When using the two-step binary approach, 121 sentences were classified correctly. This results in an accuracy of 82.31%. Out of the 27 sentences that were classified wrong, 15 are also among the mistakes made by the multiclass model.

Errors can be analyzed with regard to direction of error and type of error. Direction of error considers if a telic sentence is classfied as atelic or if an atelic sentence is classified as telic. The errors that are observed for the multi-class model are dominantly connected to atelic sentences that are classified as telic. This is the case for 18 sentences, i.e., 90% of all errors. The remaining two sentences (10%) are telic but are classified as atelic.

The distribution of direction of errors differs between the two approaches. For the two-step binary model, 14 sentences are predicted to be atelic even tough telicity is given (51.85%), and 13 are wrongly predicted to be telic (48.15%).

The type of error describes the probable source of misclassification. In most cases, the type of error can be connected to certain textual patterns that are found repeatedly for wrongly classified sentences. Types of errors are: Aspectual class change due to temporal modifiers, aspectual class change due to characteristics of the object and progressive bias. Two additional sentences are classified wrong by the multiclass model that convey the notion of telicity implicitly. In the following sections, a selection of sentences are presented along with the given direction and type of error.

The greatest amount of errors for both models occurs for the classification of sentences that include temporal modifiers. This was also found and discussed by Metheniti (2023). Metheniti "hypothesize[s] that the models give too much importance to the verb and not enough to the verb tense or the context, even if it contains prepositions and prepositional phrases of time." (2023: 107).

Direction of error: Atelic sentence classified as telic.

Type of error: Temporal modifier

(1) For months the Prime Minister made that declaration.

(2) On Fridays, I receive new stock.

(3) The workers painted the house since 8 am.

Direction of error: Telic sentence classified as atelic.

Type of error: Temporal modifier

(4) Louise made the biggest progress out of everyone this year.

(5) Over a four-week period, the classes lasted one hour and took place twice a week.

With regard to the object as an error source, it can be argued that a change in aspectual class conveyed by characteristics of the object is at times not correctly grasped by the model and consequently elicits a number of errors. Dowty already mentions the underlying problem by stating that "[a]ccomplishment verbs [in Vendler's sense] which take direct objects [...] behave like activities if an indefinite plural direct object or a mass-noun direct object is substituted for the definite (or indefinite singular) one" (1979: 62). This differentiation seems to be challenging for the models.

Direction of error: Atelic sentence classified as telic.

Type of error: Object

(6) I drank juice.

Direction of error: Telic sentence classified as atelic.

Type of error: Object

(7) The boy is eating an apple.

A bias towards the atelic label, probably due to the progressive, is seen once. This bias was also observed during the analysis of the annotations of the Captions corpus (cf. chapter 6.2.1). It can be argued that human annotators as well as the classification models are influenced by the progressive and tend to interpret sentences including it as atelic.

Direction of error: Telic sentence classified as atelic.

Type of error: Progressive bias

(8) The pond is freezing over.

Sentence (8) can be seen as an example of the so-called imperfective paradox (cf. Dowty 1977). This means that in this case the "reference to the reaching of a result-state is ‘excluded from view’ through the choice of the progressive" (Kranich 2010: 39).

Two sentences that are classified wrong include no explicit markers for telicity. This makes their classification challenging for the model.

Direction of error: Atelic sentences classified as telic.
Type of error: Implicit telicity

(9) The hunters chased the deer.

(10) They hunted the boar.

The short analysis of the telicity test set provided by Metheniti et al. (2022) gave insights into how well the trained model works in comparison to other examinations of transformer-based aspectual class classification. It is found that especially the multiclass model behaves very similar to what is described in the literature. Accuracy scores are comparable. Error sources that were identified were in parts also found by Metheniti (2023).

6.2.5 Analysis – Aspectual Classes Test Set

Next to the analysis of the test set provided by Metheniti et al. (2022), the self-compiled test set is analyzed. This test set comes from the same distribution as the training and development data and thus includes data points of the Captions corpus as well as manually annotated sentences. In contrast to the test set considered before, Vendler’s classes are included as annotations. Accordingly, the focus is not on a single feature but on all features and the general ability of the model to predict the classes. The test set holds 106 sentences in total. The predicted labels are included in the list of errors.

The multiclass approach results in an accuracy of 88.68%. This means that out of the 106 sentences, 94 were classified correctly, 12 were classified incorrectly. Specifically, one achievement sentence, one accomplishment sentence, seven activity sentences and three state sentences were predicted incorrectly. Considering the class distribution as depicted in table 18, the fact that there are few errors for the accomplishment and achievement classes is due to the classes being smaller than the state and activity classes.

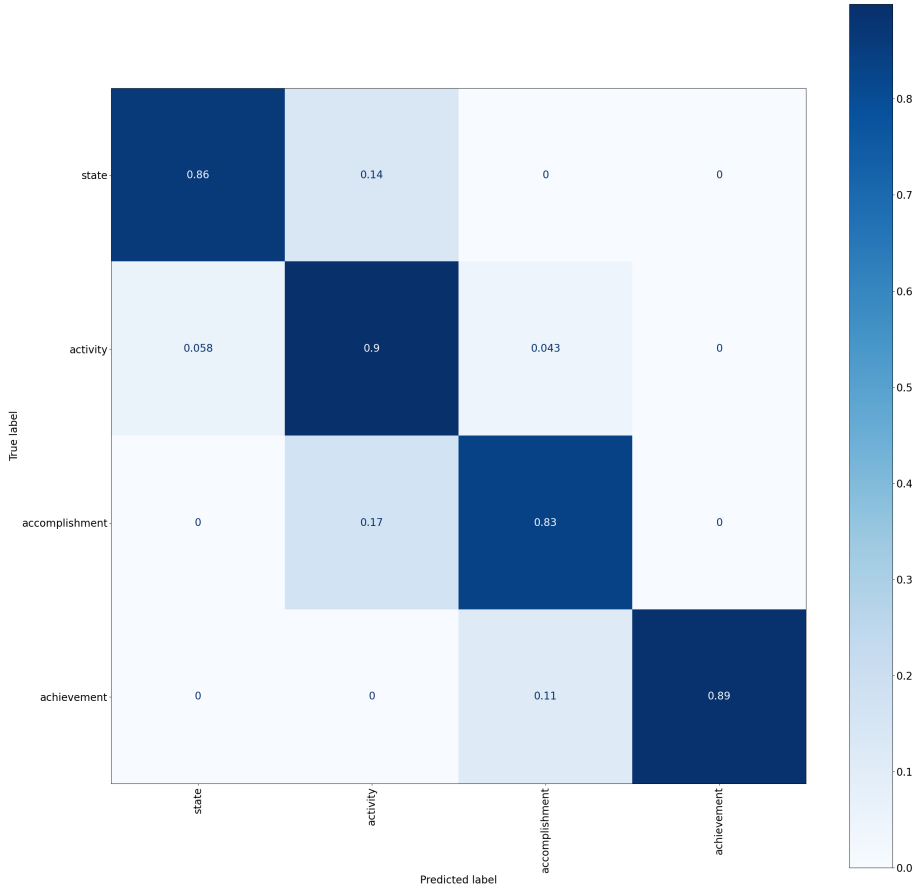

Figure 23: Confusion matrix multiclass approach

The confusion matrix in figure 23 reveals the following main error sources for the individual aspectual classes:

- State sentences classified as activity,
- activity sentences classified as state as well as accomplishment,
- accomplishment sentences classified as activity,

- and finally achievement sentences classified as accomplishment.

The two-step binary approach also reaches an accuracy of 88.68%. This means that again, 94 sentences are classified correctly, 12 are classified incorrectly. The errors made by the models differ to an extent. Nine of the 12 sentences were classified wrong by both models. Two achievement sentences and one activity sentence were classified wrong by the two-step binary approach only.

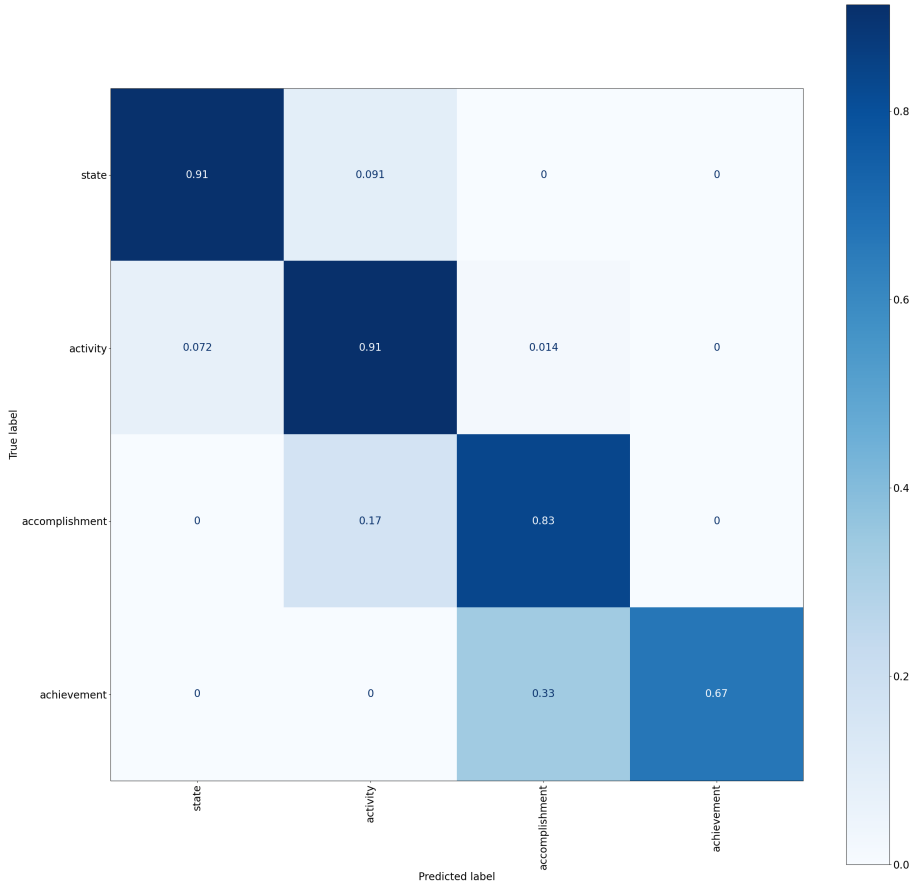

Figure 24: Confusion matrix two-step binary approach

The similarity of errors can also be seen in figure 24. The main error sources stay the same between models (cf. figure 23), however, more achievement sentences are classified as accomplishments in the case of the two-step binary approach.

It can be argued that the errors found for the individual classes are connected to different features (dynamicity, duration, telicity). The problematic feature

for the classification of state is dynamicity. Accordingly, state sentences that are inherently not dynamic are classified as activities which are dynamic.

(11) i have burning desire, let me stand next to this fire! [sic] Predicted Label: **Activity**

(12) worried: farmers were concerned the high rise in temperatures this year would mean a shorter season for biological genus but the recent rain might help it to keep growing until the end of july. Predicted Label: **Activity**

For examples (11) and (12) it has to be noted that more than one situation is described in both sentences. This might contribute to difficulties with regard to their classification.

The confusion of the classes state and activity is also observed in the opposite direction: Activity sentences, i.e., dynamic sentences, are wrongly classified as state.

(13) Maybe this is how you normally picture an Indian wedding. Predicted Label: **State**

(14) The market is bustling with people. Predicted Label: **State**

In the case of sentences (13) and (14), the manual analysis revealed that there might be a problem with regard to their annotations. Both sentences are included as activities in the dataset but should actually be annotated as state. Therefore, the classification of the model is correct.

The classification of activity sentences additionally displays errors with regard to the differentiation of telic and atelic sentences. This leads to the wrong prediction of accomplishment.

(15) Four kayakers paddle through the water. Predicted Label: **Accomplishment**

The same error can also be found for the accomplishment class. As mentioned earlier, only one accomplishment sentence was classified wrong by both models, namely as activity.

(16) A man eats a pastry. Predicted Label: **Activity**

The type of error that is seen for sentence (16) was already discussed during the analysis of the telicity test set (cf. examples (6) and (7)).

The classification of achievement reveals difficulties with regard to the duration feature. This results in the classification of accomplishment for the given punctual/non-durative sentence. In this case, the error is exclusively detected for the achievement class but not for the accomplishment class.

(17) WASHINGTON - The Trump administration proposed new rules on Wednesday to stabilize health insurance markets roiled by efforts to repeal the Affordable Care Act, by big increases in premiums and by the exodus of major insurers. Predicted Label: **Accomplishment**

6.2.6 Summarizing Remarks

The analysis shows that the models are in general capable of classifying the provided aspectual classes. BERT is more suitable for the task than RoBERTa. Metheniti (2023) also found that BERT outperforms RoBERTa when classifying aspectual classes. Metheniti conducted an experiment that focuses on explaining the possible cause of the differences between BERT and RoBERTa. She states that

> BERT's self-attention mechanism on [sic] earlier layers demonstrated a certain sensitivity to syntactic structure and better "focus" on individual tokens in early layers. [...] Other models did not show a specific focus on constituents [...] (2023: 130).

It can be argued that this gives BERT an advantage for the task.

The two approaches, multiclass and two-step binary, achieve similar F1 scores (0.82 vs. 0.81). The evaluation of the errors hint at the fact that in

some cases both models seem to draw upon a more superficial understanding of the categorization, mainly focused on the verb. Context is at times disregarded. This means that the main differences in the classes are grasped and used for classification, however, human annotators would potentially be more sensitive to smaller differences that can change the given aspectual class, e.g., the presence of specific temporal modifiers or certain characteristics of the object. This is in line with the findings obtained by Metheniti (2023).

It follows that some aspectual class annotations provided by the models can be seen to be faithful to Vendler's definition, thus conveying "the most important time schemata implied by the use of english verbs" (1957: 144), but diverge from the current definition that states that aspectual classes should be determined on the basis of the entire sentence (cf. Dowty 1979: 62).

Systematic errors can occur for all data points when the aspectual class classifier is used for the annotation of the event data. This means that errors are independent of the event type that is given. Therefore, elicited patterns between event types are still comparable and can be used as additional information.

For the remainder of this work, the multiclass model will be used. The reason for this is that the two approaches do not differ from each other in a meaningful way. However, the two-step binary approach faces the disadvantage of error-propagation. This means that errors that occur in the first step of the classification process, i.e., the telicity classification, directly influence the final classification. The event data consists of approximately 4000 texts in total (training set, development set and two test sets). Texts consist of $1-n$ sentences. All individual sentences need to be classified. Error-propagation is assumed to influence classification on this scale too negatively. By using the multiclass model this risk is avoided.

6.3 Named Entities

Named entities are used as the second feature in this work. The preliminary experiment using LIME (cf. chapter 5.1) revealed proper nouns to be one of the relevant word classes for the model's predictions. Proper nouns can be automatically detected and sorted into more semantically descriptive categories by means of named entity recognition algorithms. Jurafsky and Martin define named entities as:

> anything that can be referred to with a proper name: a person, a location, an organization [and continue to state that] [f]our entity tags are most common: PER (person), LOC (location), ORG (organisation), or GPE (geopolitical entity) (2024: 167).

The named entity classes include:

- people and characters (person),
- companies and sports teams (organisation),
- countries, cities and states (geopolitical entity) and
- regions, mountains and seas (location) (cf. Jurafsky and Martin 2024: 167).

Openly available NER taggers are especially optimized for the detection of these four (main) named entity classes. Therefore, they will also be focused on in this work. Next to the fact that named entity taggers are optimized for the four classes at hand, the decision to focus on person, organisation, location and geopolitical entities was strengthened on the basis that these named entities constitute event arguments in the DocEE corpus. Event arguments and consequently also named entities differ between event types. Therefore, ascertaining which of these named entities is dominant in a text and adding the information explicitly to the database is reasoned to support event type classification.

6.4 Sentiment

The last feature that will be utilized for event type classification is sentiment. Sentiment analysis as a task in NLP focuses on "assesses[ing] whether the author [of a text] has a negative, positive, or neutral attitude toward an item, administration, individual, or location" (Nandwani and Verma 2021: 1). Corresponding labels are assigned.

As mentioned in the beginning of this chapter, Pritzkau (2023), whose method of adding features to the data is used in this work, also uses sentiment as a feature for the classification of propaganda techniques. Additionally, emotions are employed as a second feature that focuses on affective content in Pritzkau's work. For the identification of propaganda techniques, sentiment and emotions are seen as sensible features as they are used as "a natural tool for communication and social influence" (Pritzkau 2023: 4).

Sentiment is not chosen as a feature due to the possibility of exerting social influence in this work. In the case of newspaper articles, sentiment is not expressed as explicitly as for example in reviews and social media posts (cf. Hamborg and Donnay 2021). Rather, it is assumed that the textual description of certain event types is inherently connected to different sentiment classes.

Thus, it is expected that the sentiment that is predominately given in a text varies between the individual event types. Texts describing a CRIME event, for example, are assumed to convey mostly negative sentiment. Reports on competitions and awards ceremonies as included in the WIN LOSE event class are believed to convey mostly positive sentiment.

7 Main Experiment

Considering the theoretical background of text classification again, it was stated that "[t]he goal of classification is to take a single observation, extract some useful features, and thereby **classify** [emphasis in the original] the observation into one of a set of discrete classes" (Jurafsky and Martin 2024: 61). As was shown in chapter 3.3, classification in the age of transformer models relies on the models' ability to extract these useful features automatically. While this works well for many applications, the thesis at hand wants to test whether the manual addition of relevant linguistic features can still support classification.

The following chapter will describe the main experiment of this work. The experiment approaches the task of event type classification by merging (manual) feature engineering with the usage of pre-trained transformer models. The experimental set up used for the main experiment is the same as was seen for the preliminary cross-validation experiment. The same data splits as described in chapter 5.2 are used. This was done to keep the results comparable. The models trained in the preliminary cross-validation experiment will serve as a point of comparison. They will be referred to as baseline models. The pre-trained model of choice is again RoBERTa (Liu et al. 2019).

This chapter at first describes the specific way in which the three features that were chosen for the present work (aspectual classes, named entities and sentiment) are added to the data. The general approach will be depicted by means of an example text belonging to the event type DIPLOMACY along with its annotations. Additionally, the distribution of the individual features in the data will be presented. Based on theses distributions and previous considerations regarding the features, concrete assumptions about the benefits the features might have for classification are introduced. After that, results will be presented and analyzed.

7.1 Data Augmentation

The method of data augmentation used in this work is based on Pritzkau (2023). Additional information is added to the training data in text form. This means that specific linguistic indicators are strengthened in the input text as they are mentioned explicitly for each sample.

7.1.1 Aspectual Classes

Aspectual class information is added to the data with the help of the previously trained multiclass model as described in chapter 6.2. The sentences that are classified according to their aspectual class in the case of the event data are oftentimes complex. This means that more than one situation can be described which results in a number of possible bases for the aspectual class annotations. With regard to such sentences, the multiclass model seems to mostly focus on the main verb to reach a classification decision, while also being influenced by some biases that were already described in chapter 6.2, e.g., a tendency to classify a sentence as activity if the progressive in included.

Aspectual information is added in form of the class distribution present in the text. This means that the percentage-wise occurrence of all classes is used as a feature. An example text belonging to the event type DIPLOMACY is displayed along with the aspectual class annotations provided by the multiclass model.

> Officials said the EU suggested holding the talks in Austria or Switzerland instead of Iran's proposal, Istanbul. Aspectual class: Activity
>
> Western powers fear Iran is building capacity to produce a nuclear bomb. Aspectual class: Activity
>
> Iran denies the charges, saying that it is pursuing a civilian atomic programme designed to meet its energy needs. Aspectual class: Activity
>
> Tehran's suggestion of Istanbul as a venue was seen as potentially irritating to the US, as Turkey set up a nuclear-swap deal with Tehran earlier this year just as Washington was bolstering sanctions against Iran. Aspectual class: State
>
> Last month, the New York Times said the Western side was drawing up a tougher offer for Iran than the one rejected by the country last year,

requiring it to send two-thirds more low-grade enriched uranium out of the country than the previous deal. Aspectual class: Activity

If Iran agrees, they would be the first formal talks in more than a year. Aspectual class: State

Pressure on Iran to return to the negotiating table has strengthened since the UN, US and the EU began imposing tighter sanctions on the country in June. Aspectual class: Accomplishment

The European Union has agreed to meet Iran to begin long-stalled talks on its controversial nuclear programme, EU officials have said. Aspectual class: Achievement

The EU has accepted the date of 5 December put forward by the Iranians earlier this week, but a venue has yet to be agreed. Aspectual class: Achievement

The EU's foreign affairs chief Catherine Ashton sent a letter proposing a date and venue for the talks in response to an offer from Iran's chief nuclear negotiator, Saeed Jalili. Aspectual class: Achievement

Lady Ashton - named EU foreign affairs chief last December - has been given a mandate by the five permanent members of the UN Security Council, plus Germany, to lead the negotiations with Iran. Aspectual class: Achievement

Iran nuclear talks set to resume IAEA reports on Iran. Aspectual class: Achievement

The aspectual class feature is added to all texts. The feature is build with the following template.

> The biggest category in this text is [ASPECTUAL CLASS] with [PERCENTAGE] percent, the second category is [ASPECTUAL CLASS] with [PERCENTAGE] percent, the third category is [ASPECTUAL CLASS] with [PERCENTAGE] percent, the smallest category is [ASPECTUAL CLASS] with [PERCENTAGE] percent.

For the DIPLOMACY text above, the classification consequently results in the textual feature:

> The biggest category in this text is achievement with 41.667 percent, the second category is activity with 33.333 percent, the third category is state with 16.667 percent, the smallest category is accomplishment with 8.333 percent.

The feature provides two patterns. First, the aspectual classes are sorted in the sentence according to their amount in the text. The words *biggest* and *smallest* describe the differences semantically. Second, concrete percentages are given to further denote the differences between categories.

The distribution of aspectual classes in the dataset is shown in figure 25. The classes are represented with their relative amounts (percentages) for the individual event types.

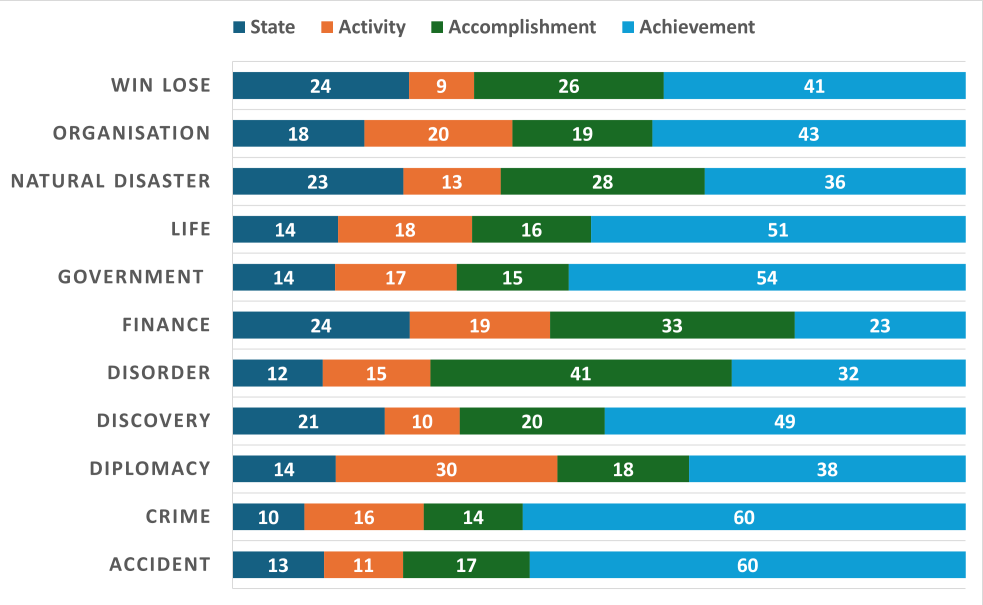

Figure 25: Distribution of aspectual classes per event type

It can be seen that for almost all event types, except DISORDER and FINANCE, achievement is the biggest aspectual class. However, the concrete distribution of the individual aspectual classes differ between event types. For CRIME, for example, the dominance of achievement is very clear while in the case of DIPLOMACY and NATURAL DISASTER, achievement only slightly exceeds the number of activities and accomplishments respectively. The inclusion of the percentages of all four classes therefore elicits different patterns for the event types.

7.1.2 Named Entities

Named entities are attained for all texts with SpaCy's (Honnibal et al. 2020) NER tagger. Not all tags that are provided by the NER tagger are used. As

mentioned before, the focus is on entities of the type person (PERSON), organisation (ORG), location (LOC) and geopolitical entity (GPE). The dominant named entity class in a text has to be identified. Consequently, after the automatic tagging of the individual texts, named entity classes are evaluated automatically with regard to the number of their occurrences. The most frequent named entity is added to the text with the following template:

> The dominant named entity in this text is [NAMED ENTITY TAG].

To allow for the possibility of two or more named entity classes existing in the exact same amount in a given text, a second template is provided during this comparison.

> The dominant named entities in this text are [NAMED ENTITY TAG], [NAMED ENTITY TAG], [NAMED ENTITY TAG].

An example output of the approach is given below. All detected named entities, i.e., PERSON, ORG and GPE are displayed. LOC was not found in the given text.

> Officials said the EU suggested holding the talks in Austria or Switzerland instead of Iran's proposal, Istanbul.
>
> Western powers fear Iran is building capacity to produce a nuclear bomb.
>
> Iran denies the charges, saying that it is pursuing a civilian atomic programme designed to meet its energy needs.
>
> Tehran's suggestion of Istanbul as a venue was seen as potentially irritating to the US, as Turkey set up a nuclear-swap deal with Tehran earlier this year just as Washington was bolstering sanctions against Iran.
>
> Last month, the New York Times said the Western side was drawing up a tougher offer for Iran than the one rejected by the country last year, requiring it to send two-thirds more low-grade enriched uranium out of the country than the previous deal.
>
> If Iran agrees, they would be the first formal talks in more than a year.
>
> Pressure on Iran to return to the negotiating table has strengthened since the UN, US and the EU began imposing tighter sanctions on the country in June.
>
> The European Union has agreed to meet Iran to begin long-stalled talks on its controversial nuclear programme, EU officials have said.

> The EU has accepted the date of 5 December put forward by the Iranians earlier this week, but a venue has yet to be agreed.
>
> The EU's foreign affairs chief Catherine Ashton sent a letter proposing a date and venue for the talks in response to an offer from Iran's chief nuclear negotiator, Saeed Jalili.
>
> Lady Ashton - named EU foreign affairs chief last December - has been given a mandate by the five permanent members of the UN Security Council, plus Germany, to lead the negotiations with Iran.
>
> Iran nuclear talks set to resume IAEA reports on Iran.

In this case, GPE is the dominant named entity. The text is modified by adding:

> The dominant named entity in this text is geopolitical entity.

Considering the entire corpus, the following distribution of named entities per event type arises.

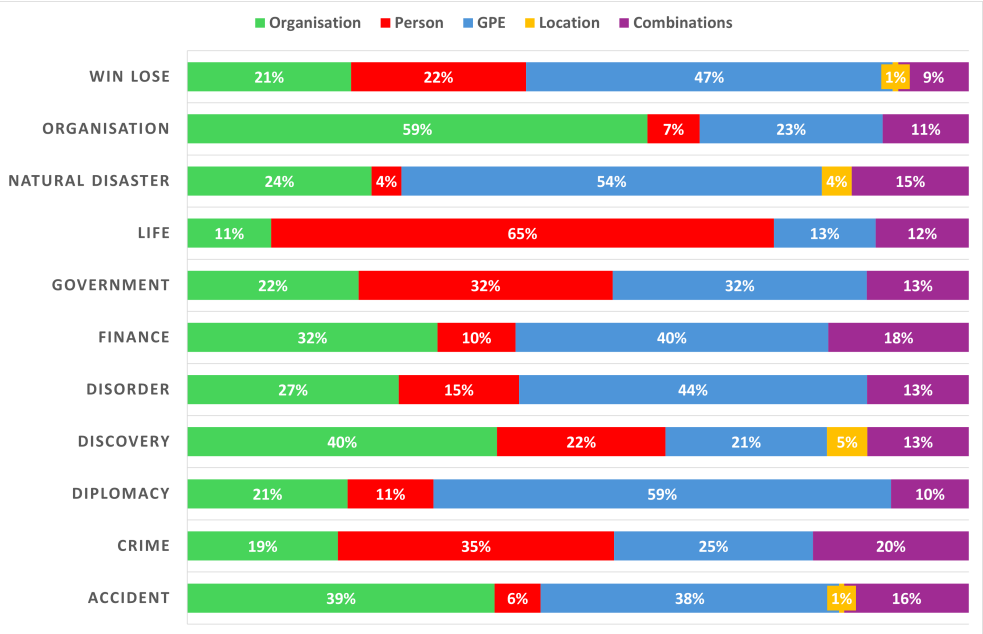

Figure 26: Distribution of named entity classes per event type

Figure 26 shows the percentages of texts that were annotated with a specific named entity per event type. It becomes obvious that location annotations do not hold relevance for the dataset.

The event types ACCIDENT, DISCOVERY and ORGANISATION mostly include named entities of the type organisation. GPE can be found as the dominant entity for DIPLOMACY, DISORDER, FINANCE, NATURAL DISAS-

TER and WIN LOSE. For CRIME and LIFE the named entity person is most common. The event type GOVERNMENT shows the same amount of occurrences for GPE and person.

Combinations of two (or more named entities) for a text occur with percentages between 9% (WIN LOSE) and 20% (CRIME).

Intuitively, the dominant named entity that is assigned to the individual event types seems sensible. ACCIDENT and WIN LOSE are the only types that might not be obviously linked to a high number of mentions of organisations and geopolitical entities, respectively. Consulting the data sheds light on this. For ACCIDENT, texts of the subtypes Air Crash and Road Crash oftentimes mention brands of cars and airplanes involved in the event. This results in a high number of organisation annotations. For WIN LOSE, the subtype Sports Competition includes references to athletes' home countries, leading to the detection of the class GPE.

7.1.3 Sentiment

Sentiment scores are calculated with the help of Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert 2014). For the work at hand, the so-called compound score is used. The compound score provided by VADER is calculated

> by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive) [...] [making] it a 'normalized, weighted composite score' (Hutto 2014).

A compound score >= 0.05 results in a positive sentiment tag, a score > -0.05 and < 0.05 in a neutral one and a score of <= -0.05 in a negative sentiment annotation. The compound score is calculated for each sentence in the text.

Similarly to adding named entities, the overarching sentiment in a text is then added as a feature as follows:

> The sentiment is [SENTIMENT].

The sentiments are [SENTIMENT], [SENTIMENT], [SENTIMENT].

Again, the possibility of all classes occurring in the exact same amount is provided for. For the example text, the following compound scores were calculated with the help VADER:

> Officials said the EU suggested holding the talks in Austria or Switzerland instead of Iran's proposal, Istanbul. Compound score: 0.0 Sentiment tag: neutral
>
> Western powers fear Iran is building capacity to produce a nuclear bomb. Compound score: -0.7506 Sentiment tag: negative
>
> Iran denies the charges, saying that it is pursuing a civilian atomic programme designed to meet its energy needs. Compound score: -0.4215 Sentiment tag: negative
>
> Tehran's suggestion of Istanbul as a venue was seen as potentially irritating to the US, as Turkey set up a nuclear-swap deal with Tehran earlier this year just as Washington was bolstering sanctions against Iran. Compound score: -0.4588 Sentiment tag: negative
>
> Last month, the New York Times said the Western side was drawing up a tougher offer for Iran than the one rejected by the country last year, requiring it to send two-thirds more low-grade enriched uranium out of the country than the previous deal. Compound score: -0.3818 Sentiment tag: negative
>
> If Iran agrees, they would be the first formal talks in more than a year. Compound score: 0.2023 sentiment tag: positive
>
> Pressure on Iran to return to the negotiating table has strengthened since the UN, US and the EU began imposing tighter sanctions on the country in June. Compound score: 0.0516 Sentiment tag: positive
>
> The European Union has agreed to meet Iran to begin long-stalled talks on its controversial nuclear programme, EU officials have said. Compound score: 0.0772 Sentiment tag: positive
>
> The EU has accepted the date of 5 December put forward by the Iranians earlier this week, but a venue has yet to be agreed. Compound score: 0.4939 Sentiment tag: positive
>
> The EU's foreign affairs chief Catherine Ashton sent a letter proposing a date and venue for the talks in response to an offer from Iran's chief nuclear negotiator, Saeed Jalili. Compound score: 0.0 Sentiment tag: neutral
>
> Lady Ashton - named EU foreign affairs chief last December - has been given a mandate by the five permanent members of the UN Security Council, plus Germany, to lead the negotiations with Iran. Compound score: 0.34 Sentiment tag: positive
>
> Iran nuclear talks set to resume IAEA reports on Iran. Compound score: 0.0 Sentiment tag: neutral

With four negative, five positive and three neutral instances, the sentiment feature would mark this text to have, in general, a positive sentiment. While

some individual annotations might not be in line with our understanding of positive, neutral and negative sentiment, the summary of all tags is expected to be true to the actual sentiment.

It is assumed that the overarching sentiment found in the texts differ between event types. Figure 27 shows the distribution of the different sentiment classes.

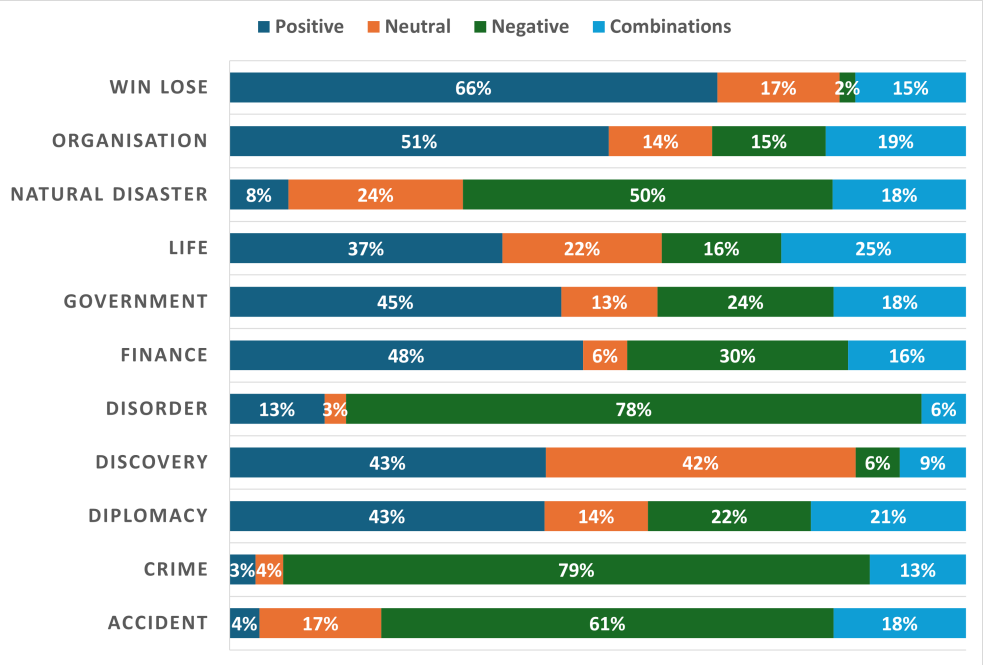

Figure 27: Distribution of sentiment classes per event type

The event types ACCIDENT, CRIME, DISORDER and NATURAL DISASTER are mostly connected to a negative sentiment. DIPLOMACY, FINANCE, GOVERNMENT, LIFE, ORGANISATION as well as WIN LOSE display positive sentiment most often. For DISCOVERY, the dominant positive sentiment is closely followed by neutral sentiment (43% vs. 42%). Combinations of sentiment annotations make up between 6% (DISORDER) and 25% (LIFE) of texts.

7.1.4 Final Features

In total, seven different features arise. There are three single features, three double features and one feature incorporating all elicited information. Fea-

tures are represented with the previously introduced templates. For the example text that was considered in the last chapters, the features look as follows.

1. Aspectual classes: [TEXT] The biggest category in this text is achievement with 41.667 percent, the second category is activity with 33.333 percent, the third category is state with 16.667 percent, the smallest category is accomplishment with 8.333 percent.

2. Named entities: [TEXT] The dominant named entity in this text is geopolitical entity.

3. Sentiment: [TEXT] The sentiment is positive.

4. Named entities and sentiment: [TEXT] The dominant named entity in this text is geopolitical entity. The sentiment is positive.

5. Named entities and aspectual classes: [TEXT] The dominant named entity in this text is geopolitical entity. The biggest category in this text is achievement with 41.667 percent, the second category is activity with 33.333 percent, the third category is state with 16.667 percent, the smallest category is accomplishment with 8.333 percent.

6. Sentiment and aspectual classes: [TEXT] The sentiment is positive. The biggest category in this text is achievement with 41.667 percent, the second category is activity with 33.333 percent, the third category is state with 16.667 percent, the smallest category is accomplishment with 8.333 percent.

7. Named entities, aspectual classes and sentiment: [TEXT] The dominant named entity in this text is geopolitical entity. The sentiment is positive. The biggest category in this text is achievement with 41.667 percent, the second category is activity with 33.333 percent, the third category is state with 16.667 percent, the smallest category is accomplishment with 8.333 percent.

All texts are modified in this way. A modified corpus arises for each feature. Separate models are trained with each corpus.

7.2 Remarks on Expected Improvements

Based on the findings of the preliminary experiment regarding common errors by the baseline models (cf. table 14), the distribution of the chosen features in the data and the general understanding of the event types, more concrete assumptions about the improvements that will probably be achieved by using the features can be made at this point. A general note can be made about the event type FINANCE. As the event type is underrepresented in the data, it can be assumed that all additional information that is provided during training will be helpful for classification.

The following main effects are expected for the features:

Aspectual Classes:

- DISORDER: The event type shows the highest number of accomplishments of all event types. It can be argued that this makes it differentiable from other event types. Especially the confusion with CRIME, which has a lower number of accomplishments, might be avoided in some cases. Due to the high number of accomplishments, the feature is assumed to also support the mitigation of smaller errors connected to the class.

- FINANCE: The event type displays a diverging distribution of aspectual classes in contrast to all other event types. The lowest number of achievements and a generally (more or less) balanced distribution of classes is observed. The feature is therefore assumed to support the event type's classification by providing an additional distinct pattern for model training and testing.

- DIPLOMACY: DIPLOMACY is characterized by the highest number

of activities of all event types. This makes the aspectual class feature promising for the event type's classification. It is reasoned that the feature can help distinguish DIPLOMACY texts from texts belonging to other event types.

Named Entities:

- DISORDER and CRIME: The difference between CRIME and DISORDER is assumed to be enhanced by the named entity feature. The distribution of named entities shows a higher number of person annotations for CRIME than for DISORDER. Reports on CRIME events, which include arrests, convictions and releases, are more concerned with individuals than reports on DISORDER events, leading to the higher number of person annotations. There are also differences regarding the named entity classes organisation and GPE, those are, however not as as great.

- WIN LOSE and DISCOVERY: The confusion of these two event types might be mitigated by the named entity feature. The reason for this is that DISCOVERY has notably more organisation annotations and WIN LOSE is more often connected to GPEs. Still, it has to be considered that both event types hold the same amount of person annotations which reinforces their similarity again.

- FINANCE: The named entity annotations are expected to aid the differentiation of FINANCE and GOVERNMENT. The named entity distribution shows that both event types hold a high number of GPE annotations. However, GOVERNMENT is more often connected to persons. Notably less person annotations are found for FINANCE. Additionally, more organisation mentions are found in connection to FINANCE. This might support the discrimination of the event types.

- ORGANISATION and DIPLOMACY as well as ORGANISATION and

GOVERNMENT: An improvement is expected with regard to the differentiation of these event types. As was seen in chapter 5.2.1, texts belonging to the event type ORGANISATION can be semantically similar to DIPLOMACY and GOVERNMENT depending on the kind of organisation that is reported on. However, texts that belong to the event type ORGANISATION dominantly include the corresponding named entity class while GOVERNMENT and DIPLOMACY are more often connected to geopolitical entities. This should support their differentiation.

Sentiment:

- WIN LOSE: The event type displays the highest number of positive sentiment and the lowest number of negative sentiment of all event types. The addition of the feature makes the event type more distinct and should support its classification in general. More specifically, it can also be assumed that the confusion of WIN LOSE and DISCOVERY is mitigated by the feature. Even though DISCOVERY also has positive as the biggest sentiment class, neutral sentiment is present in nearly the same amount for texts belonging to the event type.

- DISORDER and CRIME: Only small improvements are expected for the confusion of these two event types. They are semantically similar with regard to their general content. They are also similar with regard to the sentiment that is dominantly connected to them, i.e., negative sentiment. However, differences can be observed with regard to positive sentiment which is nearly non-existent for CRIME (3%) but more clearly detectable for DISORDER with 13%. Still, it is more probable that the sentiment feature will support the differentiation of CRIME and DISORDER from event types other than each other. Confusions of DISORDER with GOVERNMENT or CRIME with LIFE, for example, might be lessened.

7.3 Results

In the following chapters, the results of the main experiment will be detailed. Results that are obtained by the modified models, i.e., the models that use additional linguistic features, are contrasted with those reached by the baseline models. Five models are trained for each feature, one model of each feature configuration will be analyzed in more detail. Due to the scope of this work, it is not possible to consider all 35 models explicitly.

The results of the main experiment are presented in the same manner as was done for the cross-validation approach in chapter 5.2. For better readability, the classification reports will now include colored cells: Event types that are predicted with a higher F1 score than was seen for the baseline are colored green. Event types that are predicted with lower scores are colored red. Event types that hold the same F1 score for both models are not colored and remain white. To facilitate the comparison with the corresponding baseline model, the F1 scores reached by it are also included in the tables that hold the classification reports. The column is called F1 B which stands for F1 Baseline.

Next to the graphical depiction of F1 scores, classification reports and confusion matrices, the results of the paired bootstrap testing (i.e., statistical significance testing) are included in form of a table. Paired bootstrap testing evaluates the performance of the modified models against the baseline models. To reiterate, the H_0 that is tested is: A model that is supplied with additional features during training *does not* outperform a model that is trained without additional features. For all tables showing the statistical significance testing results, general model performance is depicted with macro F1 scores. Significant p-values with a predefined significance level of 0.05 are marked with a * in the table. Additionally, the models that perform better than the baseline according to the macro F1 score (regardless of whether statistical significance was reached or not) are colored green, models that reach significance are colored dark green, those that perform worse are colored red. Again, the small

as well as the big test set are considered. Scores are rounded to four decimal places.

Finally, mean F1 scores of the individual event types are presented in form of tables. These tables compare the modified models' results to those reached by the baseline models. Again, the cells are colored in accordance to differences and similarities that arise between the model configurations.

Figures 28 and 29 give a first overview of the results of the main experiment. Figure 28 shows the results of all 40 models with their macro F1 scores when tested on the small test set. This includes the five baseline models as well as all 35 modified models. Figure 29 displays the scores that were reached by the same models tested on the big test set.

Overall, the figures show that the scores reached for the small test set were higher than those reached for the big test set. That was expected as the small test set belongs to the same distribution as the training and development data. The big test set, however, includes entire newspaper articles, i.e. unpreprocessed data taken from the DocEE corpus, and consequently differs to the training and development data of the models to an extent. This finding is in line with the findings obtained for the preliminary cross-validation experiment.

Figure 28: Results main experiment small test set

Figure 29: Results main experiment big test set

7.3.1 Aspectual Classes

In figure 30 the results of the models that were trained with aspectual class information as a feature are shown for all five data splits.

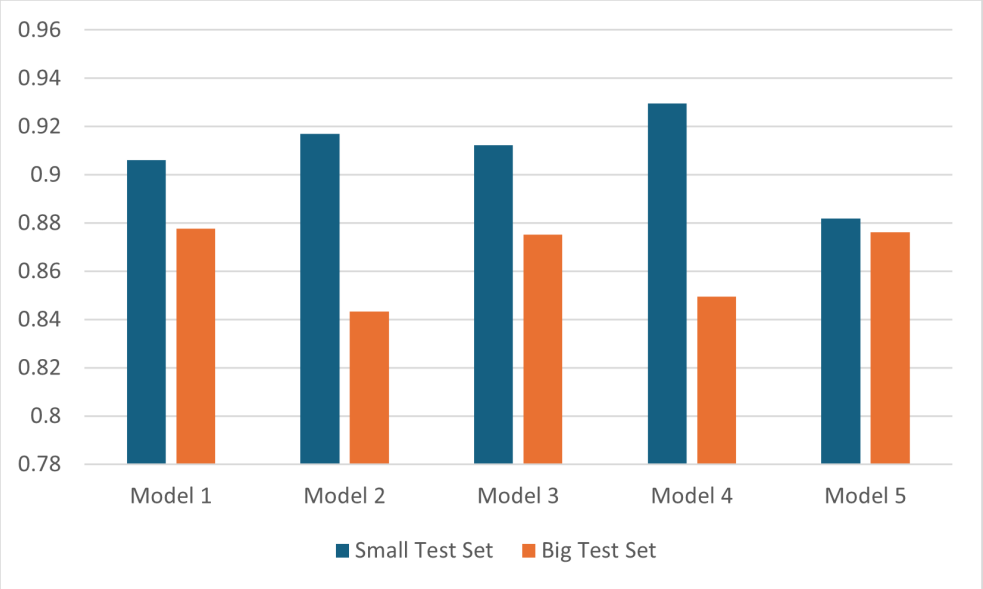

Figure 30: Feature aspectual classes: results main experiment

The key findings are:

- The macro F1 scores that were reached for the small test set range between 0.882 (model 5) and 0.9297 (model 4).
- The mean macro F1 score is 0.9094.
- The SD is 0.0176.
- For the big test set, the macro F1 scores show a general decrease and range from 0.8434 (model 2) to 0.8776 (model 1).
- A mean macro F1 score of 0.8644 is reached.
- The SD is 0.0166.
- A difference of 0.045 arises between model performance for the two test sets.

In comparison to figure 15 which includes the results achieved by the baseline models, it becomes apparent that the models closely follow the baseline. However, all modified models have a slight advantage over the baseline for both test sets. Additionally, the difference between the two test sets is on average smaller than was seen for the baseline. Less variance in scores is observed for the models trained with aspectual classes as is shown by the lower SD values. Furthermore, in contrast to the baseline, the SD that is observed for the big test is smaller than the one for the small test set, albeit only slightly.

The differences between the baseline models and the models using aspectual class information were tested for statistical significance with the help of paired bootstrap testing. Table 20 shows the results of the statistical significance testing.

	Small test			Big test		
Split	Baseline	Modified	p-value	Baseline	Modified	p-value
1	0.8776	0.9062	0.3158	0.8329	0.8776	0.0005*
2	0.9147	0.9169	0.2822	0.8019	0.8434	0.0022*
3	0.9043	0.9124	0.2682	0.8632	0.8753	0.105
4	0.921	0.9297	0.3747	0.8421	0.8495	0.2573
5	0.8659	0.882	0.2838	0.86	0.8763	0.1433

Table 20: Feature aspectual classes: results bootstrap testing

Bootstrap testing revealed no statistically significant differences between the models regarding the small test sets. However, p-values converge towards significance for the big test set. For models trained on split 1 and 2, statistical significance is reached.

To gain deeper insights into the differences and similarities of the modified and the baseline models, the individual event types are presented in table 21 with their respective mean F1 scores.

It is observed that the models using aspectual class information as a feature reach higher scores for almost all event types for both test sets. The differences are in parts small but, on average, constant. An exception is found

	Small test			Big test		
	Baseline	**Modified**	**Diff**	**Baseline**	**Modified**	**Diff**
Accident	0.97	0.98	0.01	0.95	0.96	0.01
Crime	0.96	0.97	0.01	0.91	0.92	0.01
Diplomacy	0.85	0.86	0.01	0.77	0.78	0.01
Discovery	0.92	0.94	0.02	0.9	0.91	0.01
Disorder	0.99	0.99	0	0.8	0.86	0.06
Finance	0.62	0.6	-0.02	0.64	0.77	0.13
Government	0.88	0.9	0.02	0.83	0.85	0.02
Life	0.89	0.95	0.06	0.87	0.88	0.01
Natural Disaster	0.94	0.94	0	0.95	0.95	0
Organisation	0.86	0.88	0.02	0.75	0.76	0.01
Win Lose	0.98	0.99	0.01	0.85	0.87	0.02

Table 21: Feature aspectual classes: mean F1 scores of all event types

for the prediction of FINANCE for the small test set. Here, the baseline models reach higher scores. The classification of the big test set, however, leads to a stark increase in the mean F1 score of FINANCE with regard to the modified model. Consequently, the modified models outperforms the baseline with a mean difference of 0.13.

No change is observed for the classification of the event type NATURAL DISASTER for both test sets. The prediction of the small test set also results in the same score for DISORDER for both the baseline and the modified models. DISORDER shows a mean F1 score of 0.99. This is close to a perfect score. The classification of the big test set does not result in equally high scores. Both model configurations display losses. However, the modified models again reach higher scores with a mean difference of 0.06.

Error Analysis Model 1 Model 1 was chosen for a more detailed error analysis. Model 1 shows a comparatively high p-value when contrasting it with the baseline model for the small test set. For the big test set, however, the difference becomes highly significant which makes it an interesting model for a more thorough error analysis.

The consultation of the classification reports listed in table 22 reveal sim-

	Small test				Big test			
	Precision	**Recall**	**F1**	**F1 B**	**Precision**	**Recall**	**F1**	**F1 B**
Accident	0.97	1.00	0.98	0.95	0.96	0.98	0.97	0.95
Crime	1.00	1.00	1.00	0.96	0.93	0.94	0.94	0.91
Diplomacy	0.71	0.71	0.71	0.67	0.77	0.84	0.80	0.78
Discovery	0.86	1.00	0.92	0.92	0.89	0.94	0.92	0.86
Disorder	1.00	1.00	1.00	1.00	0.97	0.81	0.88	0.83
Finance	1.00	0.33	0.50	0.50	0.77	0.71	0.74	0.58
Government	0.92	0.95	0.94	0.93	0.85	0.89	0.87	0.86
Life	1.00	1.00	1.00	0.88	0.89	0.89	0.89	0.84
Natural Disaster	0.97	0.93	0.95	0.91	0.96	0.96	0.96	0.94
Organisation	0.93	1.00	0.96	0.93	0.80	0.76	0.78	0.77
Win Lose	1.00	1.00	1.00	1.00	0.97	0.84	0.90	0.83
Macro average	**0.94**	**0.90**	**0.91**	**0.88**	**0.89**	**0.87**	**0.88**	**0.83**

Table 22: Feature aspectual classes: classification reports model 1

ilar results between the baseline and the modified model regarding the small test set (cf. table 8). The modified model reaches at least the same F1 score as the baseline model for all event types. Seven event types have higher scores than the baseline. The differences are mostly smaller or equal to 0.04. One notable discrepancy can be seen for LIFE with a difference of 0.12. It can be argued that the small differences indicate that both models process the textual data similarly. The addition of aspectual class information then contributes to the model's prediction of unseen data, making it more stable.

Moving on to the big test set, it can be seen that the differences between the models grow larger for the individual event types. While there were still four event types with the same F1 score for the small test set, the modified model now reaches higher scores for all types. The difference of 0.04 that was detected for the classification of nearly all of the event types in the small data set is in this case exceeded by five event types, namely, DISCOVERY, DISORDER, FINANCE, LIFE and WIN LOSE. By far the most striking difference of all event types can be observed for FINANCE (0.16). The confusion matrix depicted in figure 31 provides further insights into the behavior of these classes.

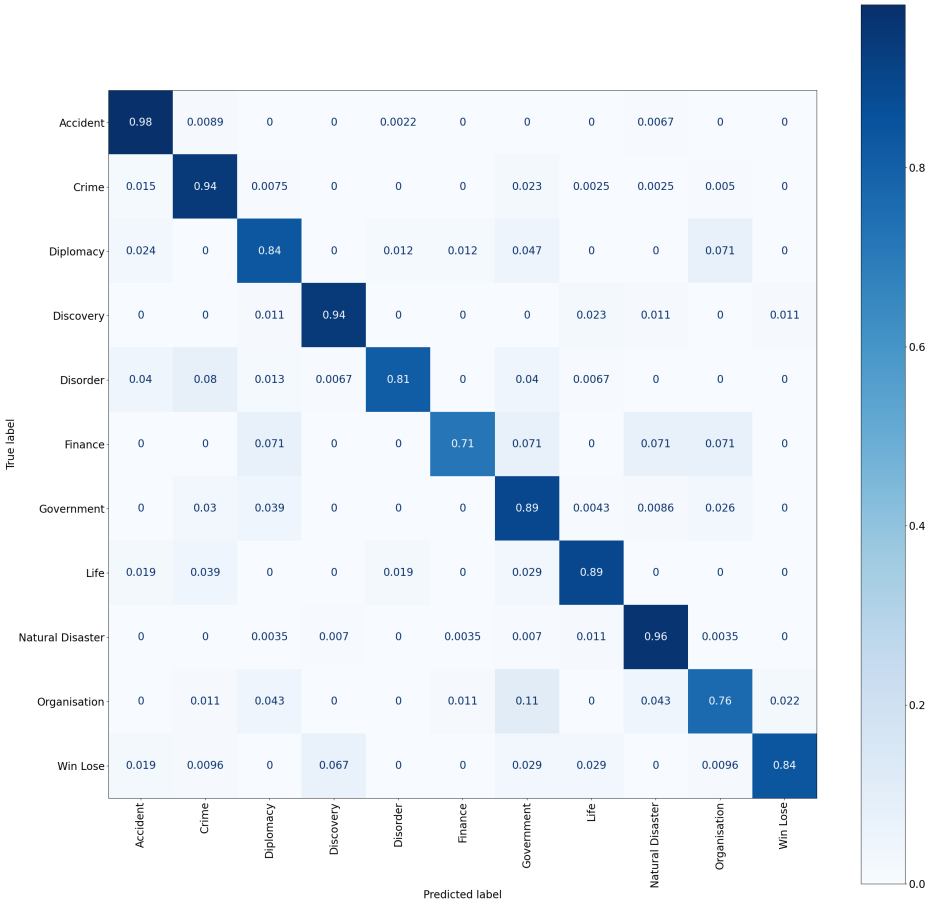

Figure 31: Feature aspectual classes: confusion matrix big test set predictions model 1

The confusion matrix shows that in contrast to the baseline model (cf. figure 16), errors that are rooted in the confusion of event types that are similar in their content are mitigated. Texts belonging to the class DISORDER, for example, were predicted to belong to CRIME in 15% of all cases when using the baseline model. This error is reduced to 8%.

Similarly, a decrease for WIN LOSE texts that are predicted as DISCOVERY is seen. The baseline model classified 17% of WIN LOSE texts to belong to the event type DISCOVERY. For the modified model this error only occurs for 7%.

For FINANCE, the event type with the greatest disparity between models, an improvement regarding the wrong classification of the event types GOVERNMENT, NATURAL DISASTER and ORGANISATION can be observed. As was already noted in chapter 5.2.1, FINANCE constitutes the smallest

class in the dataset. It could be argued that the usage of an additional feature contributes to a more distinct representation of the event type which in turn makes it better separable from others when it comes to the prediction of unseen data.

For texts of the event type DIPLOMACY, only a small improvement can be seen with regard to the main error source, i.e., the wrong prediction of GOVERNMENT. The error is reduced by 1%.

The systematic confusion of ORGANISATION texts with DIPLOMACY is also lessened; the error decreases by 4%. The confusion with GOVERNMENT, however, stays the same.

In general, an improvement can be observed for the main error sources that were present with regard to the baseline model. However, smaller errors, i.e., those making up as many as around 4% in their respective classes, partly persist. This can be seen for example for the WIN LOSE predictions: Confusions with the classes GOVERNMENT and LIFE still exist. For LIFE, the error stays the same with 3%, for GOVERNMENT it increases slightly for the modified model (2% becomes 3%).

It stands to reason that the addition of aspectual classes as a feature can not erase all errors. It has to be considered that, first, there is variation with regard to the distribution of aspectual classes for the individual event types, and second, that the model is still mainly trained on the original text data and the provided annotations. It can be assumed that the data necessarily includes inconsistencies either in textual content or labeling which influences classification.

In summary, the advantage of the modified model over the baseline is visible for all event types for the model at hand. This advantage becomes especially clear for the big test set. Systematic errors between classes that are similar content-wise are mitigated. Thus, the addition of aspectual class information seems to support event type classification.

7.3.2 Named Entities

The results that were obtained for the models trained with named entity information are depicted in figure 32.

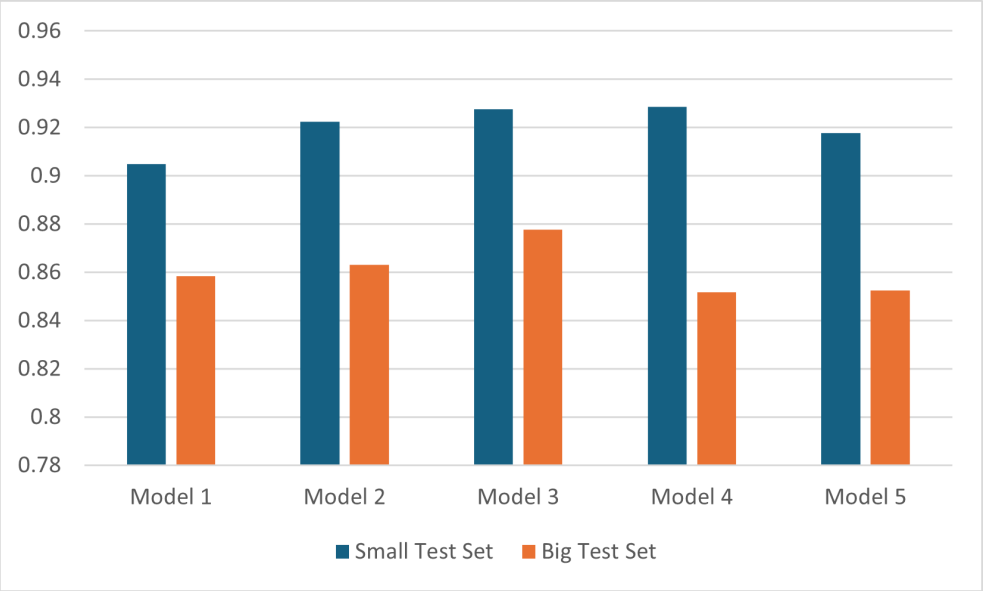

Figure 32: Feature named entities: results main experiment

The main results are:

- The range of macro F1 scores for the models at hand is 0.9040 (model 1) to 0.9275 (model 3) for the small test set.
- The mean macro F1 score results in 0.9203.
- The SD is 0.0096.
- Regarding the big test set, the highest macro F1 score is reached by model 3 with 0.8776, the lowest score is seen for model 5 with 0.8524.
- A mean macro F1 score of 0.8607 is reached.
- The SD amounts to 0.0105.
- The difference between test set scores is 0.05962.

No statistical significance is reached for the small test set as can be seen in table 23. For the big test set, model 1 and 2 significantly outperform the baseline.

	Small test			Big test		
Split	Baseline	Modified	p-value	Baseline	Modified	p-value
1	0.8776	0.9049	0.2244	0.8329	0.8585	0.0147*
2	0.9147	0.9225	0.1731	0.8019	0.863	0.0002*
3	0.9043	0.9275	0.1617	0.8632	0.8776	0.1168
4	0.921	0.9287	0.4115	0.8421	0.8518	0.2041
5	0.8659	0.9178	0.0808	0.86	0.8524	0.7175

Table 23: Feature named entities: results bootstrap testing

The models using named entity information behave in parts different to the baseline. This is especially noticeable for model 5. The baseline showed a low score for the model trained with data split 5 when testing on the small test set and a high score when testing on the big test set. This resulted in a very small difference between test set scores (0.0059). Similar results were seen for the model trained with aspectual class information for split 5. The model using named entities, however, shows the largest discrepancy between scores for this split (0.0654). Additionally, it is the only one to perform worse than the baseline model when classifying the big test set.

With regard to the mean F1 scores of the individual event types, table 24 shows mostly higher scores for the models using named entity information. Most notably this can be seen for the big test set classification of DISORDER and FINANCE. Mean differences between baseline and modified models are 0.06 and 0.14 respectively.

However, some event types do not seem to benefit from the feature. Those are ACCIDENT and NATURAL DISASTER. Both event types are predicted on average with the same scores for both test sets. In addition, there is also no difference between the models for the big test set predictions of ORGANISATION.

	Small test			Big test		
	Baseline	**Modified**	**Diff**	**Baseline**	**Modified**	**Diff**
Accident	0.97	0.97	0	0.95	0.95	0
Crime	0.96	0.97	0.01	0.91	0.92	0.01
Diplomacy	0.85	0.86	0.01	0.77	0.75	-0.02
Discovery	0.92	0.97	0.05	0.9	0.89	-0.01
Disorder	0.99	1	0.01	0.8	0.86	0.06
Finance	0.62	0.72	0.1	0.64	0.78	0.14
Government	0.88	0.87	-0.01	0.83	0.85	0.02
Life	0.89	0.94	0.05	0.87	0.88	0.01
Natural Disaster	0.94	0.94	0	0.95	0.95	0
Organisation	0.86	0.89	0.03	0.75	0.75	0
Win Lose	0.98	1	0.02	0.85	0.87	0.02

Table 24: Feature named entities: mean F1 scores of all event types

For three event types, GOVERNMENT, DIPLOMACY and DISCOVERY, the baseline models reach higher scores. In the case of GOVERNMENT, this is observed for the small test set predictions, for DIPLOMACY and DISCOVERY for the big test set.

Error Analysis Model 5 Due to the notable difference in scores seen for model 5 in contrast to the baseline, model 5 will be analyzed further in the following.

Comparing the classification reports in table 25 to the ones of the corresponding baseline model in table 12, it can be seen that for the small test set there are six event types with higher scores for the modified model, and two with higher scores for the baseline model. For the modified model those are CRIME with a difference of 0.02, DIPLOMACY (0.08), FINANCE (0.27), LIFE (0.17), ORGANISATION (0.02) and WIN LOSE (0.13). The baseline model has an advantage for GOVERNMENT (0.02) and NATURAL DISASTER (0.04). Event types that have a higher score for the modified model predictions show greater differences than the ones with a higher score for the baseline model which in turn results in a higher macro F1 score for the modified model. This then leads to a nearly significant difference in overall model

performance for the small test set with a p-value of 0.0808 (cf. table 23).

	Small test				**Big test**			
	Precision	**Recall**	**F1**	**F1 B**	**Precision**	**Recall**	**F1**	**F1 B**
Accident	1.00	1.00	1.00	1.00	0.95	0.98	0.96	0.96
Crime	1.00	0.96	0.98	0.96	0.94	0.90	0.92	0.92
Diplomacy	1.00	0.83	0.91	0.83	0.71	0.75	0.73	0.79
Discovery	0.88	1.00	0.93	0.93	0.86	0.94	0.90	0.94
Disorder	1.00	1.00	1.00	1.00	0.94	0.81	0.87	0.88
Finance	1.00	0.50	0.67	0.40	0.82	0.64	0.72	0.60
Government	0.90	0.93	0.91	0.93	0.75	0.93	0.83	0.84
Life	0.89	1.00	0.94	0.77	0.83	0.93	0.88	0.89
Natural Disaster	0.88	0.95	0.91	0.95	0.95	0.96	0.95	0.96
Organisation	0.87	0.81	0.84	0.82	0.93	0.60	0.73	0.76
Win Lose	1.00	1.00	1.00	0.92	0.99	0.79	0.88	0.93
Macro average	**0.95**	**0.91**	**0.92**	**0.87**	**0.88**	**0.84**	**0.85**	**0.86**

Table 25: Feature named entities: classification reports model 5

The results that were achieved for the big test set are clearly different regarding the advantages and disadvantages of the individual models. It becomes obvious that only the event type FINANCE is predicted with a higher score by the modified model. The difference is 0.12. Except for ACCIDENT and CRIME, all other event types now show higher scores for the baseline model. The differences lie between 0.01 and 0.06 and are therefore smaller than the ones seen for the small test set. This means that although the baseline model does not show stark increases in scores compared to the model using named entity information, it still performs consistently better.

To understand what might be the reason for the modified model's performance decline, the confusion matrix in figure 33 is consulted. In contrast to the baseline model (cf. figure 20), a number of interesting discrepancies can be observed.

With regard to texts of the event type DIPLOMACY, for example, the error of wrongly predicting GOVERNMENT increases for the modified model predictions. The baseline model displayed this error for 8% of DIPLOMACY texts, however, the modified model displays the error for 15%. As can be

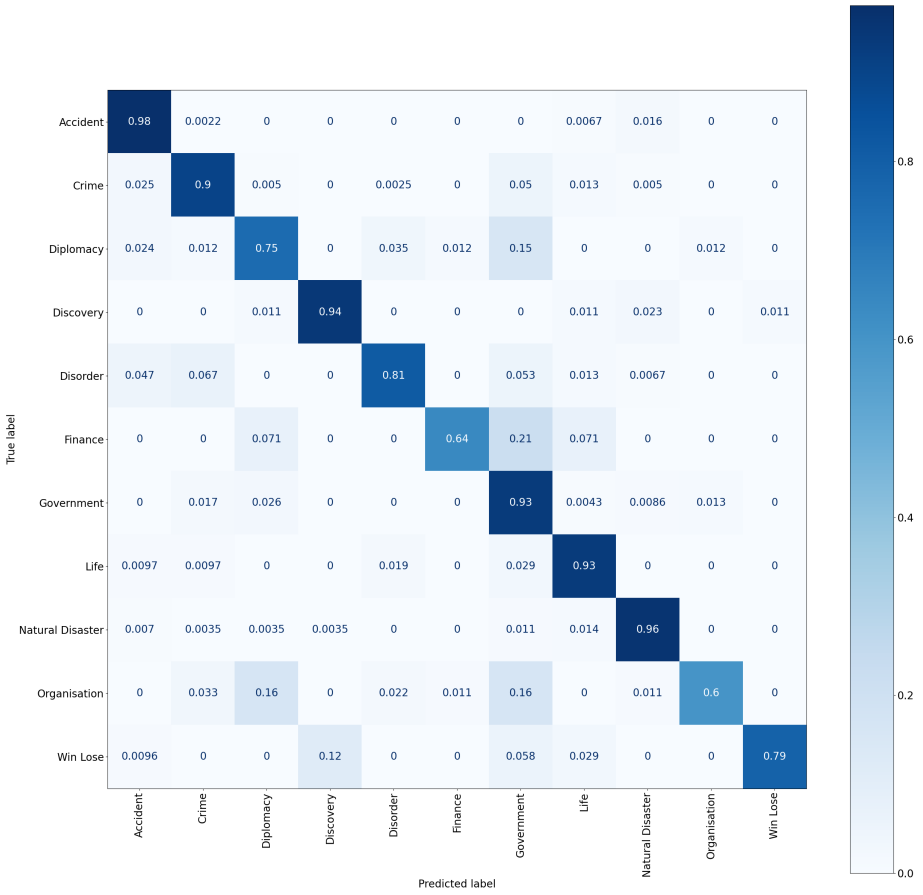

Figure 33: Feature named entities: confusion matrix big test set predictions model 5

seen in figure 26, the two event types have similar distributions with regard to the named entity classes. Especially person and organisation are very similar. Geopolitical entity, meanwhile, is more prevalent for DIPLOMACY. These similarities in feature distribution might contribute to the increase of the given error. Table 25 showed that DIPLOMACY has one of the greatest discrepancy between the small and big test set. This might mean that the class has been learned narrowly during training, focusing too much on the given named entity information. A transfer to the big test set then seems to be difficult.

Another apparent tendency of the modified model to strengthen systematic confusions of similar event types is present for the class WIN LOSE and its faulty classification of DISCOVERY. For the baseline model, this error constitutes 6% of WIN LOSE texts, for the modified model it increases to 12%. In the analysis of baseline model 5 it was seen that the model fits more loosely

to the training data which might explain why it is consequently less susceptible to the kinds of errors described for WIN LOSE and DIPLOMACY when it comes to the big test set predictions.

The only event type that seems to benefit notably from the added named entity information in this case is FINANCE. The event type, as noted before, is underrepresented in the database which could explain the special need for additional information. Here, the most dominant error that was seen for the baseline model, the wrong prediction of GOVERNMENT (43%), is reduced. Still, the error remains with 21% of texts falling into this category.

An additional error present for the baseline model predictions that is mitigated by the modified model is the confusion of DISORDER and CRIME. A change of 2% is observed.

In general, it can be stated that in the case of model 5, the named entity information might be focused on too closely by the model. This means that the additional information is leveraged for classification during training as it characterizes the event types well. For split 5, this approach seems to be effective for model optimization and leads to the model also focusing on it for the test set predictions. The high macro F1 score seen for the small test set and the nearly significant difference to the baseline model supports this interpretation. For the classification of the big test set, however, the event type representations that are used by the model are error-prone.

The named entity feature as such can still be seen as promising. Four out of five models outperform the baseline for both test sets.

7.3.3 Sentiment

The scores of the models that were trained with sentiment information can be seen in figure 34.

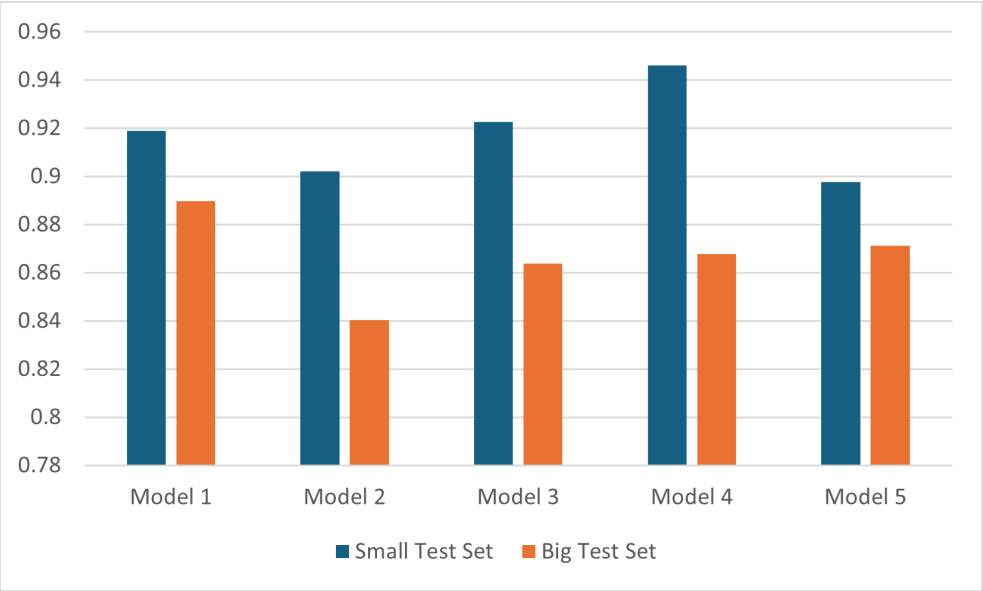

Figure 34: Feature sentiment: results main experiment

The evaluation of the models show the following results:

- The scores for the small test set range from 0.8976 (model 5) to 0.9461 (model 4).
- The mean macro F1 score is 0.9175.
- An SD of 0.0192 is observed.
- Testing the models on the big test set results in scores between 0.8402 (model 2) and 0.8896 (model 1).
- A mean F1 score of 0.8665 reached.
- The SD is 0.0177.
- A difference of 0.051 arises between the small and the big test set.

An advantage over the baseline is seen for almost all modified models when analyzing the results included in table 26. Model 2 tested on the small test set constitutes the only exception. There is again no statically significant difference for the classification results of the small test set. For the big test set, three models reach significance, models 1, 2 and 4. For the remaining models, namely models 3 and 5, a reverse effect can be observed. This means

	Small test			Big test		
Split	Baseline	Modified	p-value	Baseline	Modified	p-value
1	0.8776	0.9189	0.1404	0.8329	0.8896	0.0002*
2	0.9147	0.9021	0.7898	0.8019	0.8402	0.0065*
3	0.9043	0.9227	0.0962	0.8632	0.8637	0.477
4	0.921	0.9461	0.1973	0.8421	0.8678	0.0237*
5	0.8659	0.8976	0.1613	0.86	0.8711	0.2209

Table 26: Feature sentiment: results bootstrap testing

that while they are close to significance when contrasted with the baseline for the small test set, this apparent advantage seems to be lost for the big test set predictions. Model 3 shows this most strikingly: The p-value for the small test set is 0.0962, the one for the big test set increases notably to 0.477. A further analysis of model 3 will be conducted in the following section to gain insights into the possible causes.

The overview of the mean F1 scores of all event types in table 27 shows that the models supplied with sentiment information reach the same or higher scores as the baseline models for all event types and both test sets.

	Small test			**Big test**		
	Baseline	**Modified**	**Diff**	**Baseline**	**Modified**	**Diff**
Accident	0.97	0.98	0.01	0.95	0.96	0.01
Crime	0.96	0.96	0	0.91	0.92	0.01
Diplomacy	0.85	0.86	0.01	0.77	0.77	0
Discovery	0.92	0.94	0.02	0.9	0.9	0
Disorder	0.99	0.99	0	0.8	0.87	0.07
Finance	0.62	0.72	0.1	0.64	0.75	0.11
Government	0.88	0.9	0.02	0.83	0.85	0.02
Life	0.89	0.92	0.03	0.87	0.9	0.03
Natural Disaster	0.94	0.95	0.01	0.95	0.96	0.01
Organisation	0.86	0.88	0.02	0.75	0.77	0.02
Win Lose	0.98	0.99	0.01	0.85	0.88	0.03

Table 27: Feature sentiment: mean F1 scores of all event types

No change between scores is observed for CRIME and DISORDER with regard to the small test set and DIPLOMACY and DISCOVERY regarding the

big test set. The mean results for the small test set show a maximum difference of 0.03 between models. For the big test set, however, greater differences are reached for DISORDER and FINANCE with 0.07 and 0.11 respectively.

Error Analysis Model 3 The classification reports for model 3 are included in table 28. With regard to the small test set, the modified model has an advantage over the baseline for three event types. These are GOVERNMENT with a difference of 0.03, LIFE with a difference of 0.04 and ORGANISATION with a difference of 0.14.

	Small test				**Big test**			
	Precision	**Recall**	**F1**	**F1 B**	**Precision**	**Recall**	**F1**	**F1 B**
Accident	0.94	1.00	0.97	0.97	0.95	0.97	0.96	0.95
Crime	0.97	1.00	0.98	0.98	0.91	0.91	0.91	0.92
Diplomacy	0.92	0.86	0.89	0.89	0.73	0.80	0.76	0.79
Discovery	0.71	1.00	0.83	0.83	0.87	0.95	0.91	0.94
Disorder	1.00	1.00	1.00	1.00	0.95	0.81	0.87	0.86
Finance	0.67	0.67	0.67	0.67	0.73	0.79	0.76	0.80
Government	0.94	0.81	0.87	0.84	0.84	0.86	0.85	0.86
Life	1.00	1.00	1.00	0.96	0.88	0.90	0.89	0.84
Natural Disaster	0.95	0.92	0.93	0.95	0.93	0.99	0.96	0.95
Organisation	1.00	1.00	1.00	0.86	0.80	0.72	0.76	0.72
Win Lose	1.00	1.00	1.00	1.00	1.00	0.77	0.87	0.88
Macro average	**0.92**	**0.93**	**0.92**	**0.90**	**0.87**	**0.86**	**0.86**	**0.86**

Table 28: Feature sentiment: classification reports model 3

For one of the event types, NATURAL DISASTER, the baseline outperforms the modified model with a difference of 0.02. All other types are predicted with the same F1 scores by both models for the small test set. Consequently, this means that the reason for the modified model's higher macro F1 score is not a consistently better performing model, but a small number of event types with higher scores.

The analysis of the classification of the big test set reveals five event types with higher scores for the model using sentiment information: ACCIDENT (0.01), DISORDER (0.01), LIFE (0.05), NATURAL DISASTER (0.01) and ORGANISATION (0.04). In contrast to the small test set, the remaining event

types are predicted with higher scores by the baseline model. The differences in this case lie between 0.01 and 0.04. There are no event types with the same F1 score in this case. In both cases, the differences between models are comparatively small. This eventually leads to a nearly non-existent difference of 0.0005 between macro F1 scores for the big test set as seen in table 26.

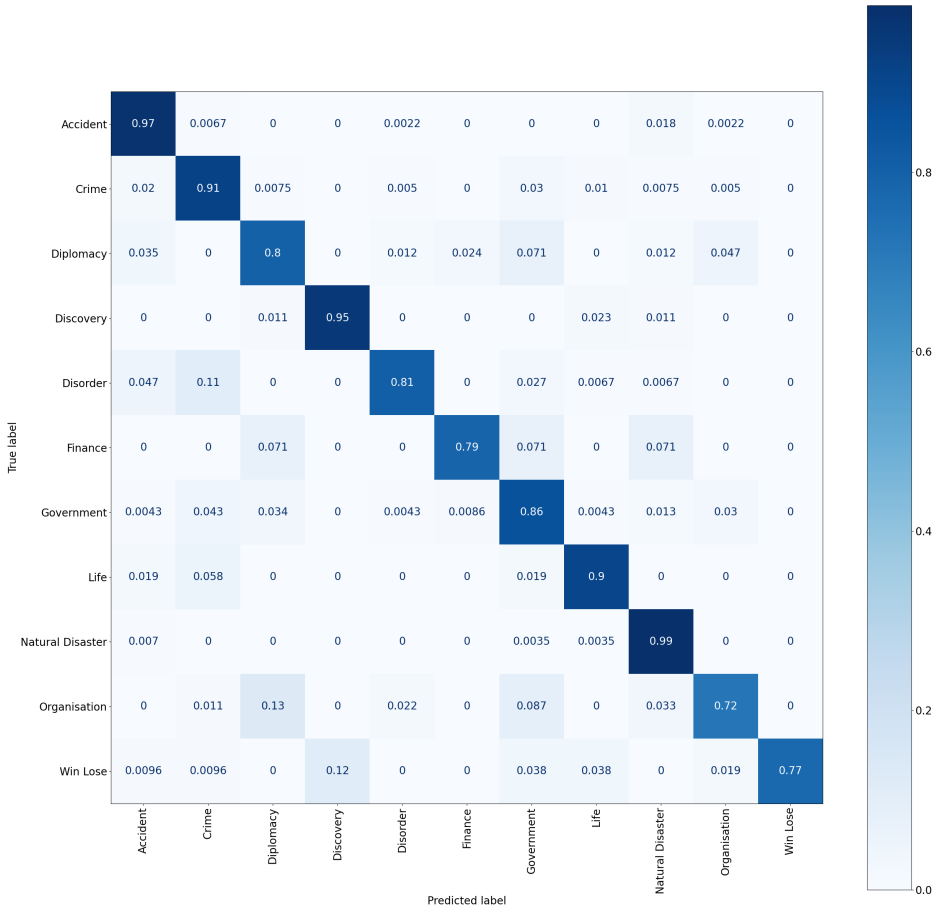

Figure 35: Feature sentiment: confusion matrix big test set predictions model 3

The comparison of the confusion matrices reveals similar errors for both models (cf. figure 18). This could already be expected due to the small differences between models. There are, however, three discrepancies in model performance that stand out.

First, for the model using sentiment information, the error of predicting GOVERNMENT instead of ORGANISATION decreases from 16% to 9%. This, however, constitutes the only notable improvement of errors for the modified model. This is in line with what was seen in the classification re-

ports: ORGANISATION has one of the greatest differences between models. The modified model reaches a higher score than the baseline model with a difference of 0.04.

Second, the systematic confusion of DISORDER texts with the event type CRIME increases for the modified model. The baseline exhibits 9% of DISORDER texts that were wrongly predicted to belong to CRIME. When using the modified model, this error can be observed for 11% of texts. The two event types are predominantly connected to negative sentiment (cf. figure 27). It follows that in this case, the feature does not contribute to the differentiation of the event types but rather enforces their similarity.

Third, the greatest change in performance is present for texts that belong to the event type WIN LOSE but are classified as DISCOVERY. This error was negligible for the baseline model. Only 2% of errors were due to this type of confusion. For the model that uses sentiment information, however, the error rate increases to 12%. At the same time, smaller errors that were observed for the prediction of WIN LOSE with regard to the baseline model, namely the wrong predictions of GOVERNMENT and LIFE, are mitigated by the modified model. As was noted in chapter 5.2.1, the low error rate of baseline model 3 regarding the wrong prediction of DISCOVERY for WIN LOSE texts is unique among all baseline models. All other baseline models include this systematic error to some extent. Baseline model 3 might therefore represent an outlier. While the addition of sentiment information seems to strengthen the error, it must be considered that this might not be because of the addition of the feature as such but rather because of the diverging representation of the event type that was learned by baseline model 3.

In general, the classification reports as well as the confusion matrix reveal that in the case of model 3, the training with sentiment information does not seem to support event type classification. However, the sentiment feature also does not hinder the general ability of the model to solve the task, meaning it conveys information that does not stand in contrast to the content of the texts.

This further strengthens the assumption that sentiment information can be a sensible feature for event type classification. This is supported by the results reached by models 1, 2 and 4 and the generally high macro F1 scores for the big test set.

7.3.4 Named Entities and Sentiment

Figure 36 presents the scores obtained by the models that were trained with a combination of named entity and sentiment information.

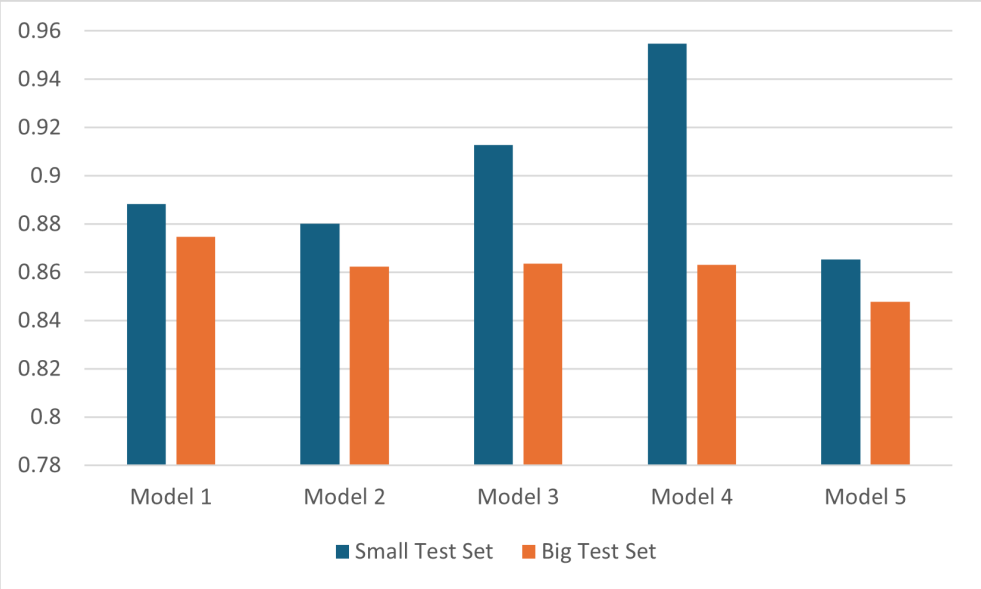

Figure 36: Feature named entities and sentiment: results main experiment

The following key findings are observed:

- The models at hand reached F1 scores between 0.8653 (model 5) and 0.9547 (model 4) for the classification of the small test set.
- A mean macro F1 score of 0.9003 was obtained.
- The SD is 0.035.
- For the big test set, the model trained with split 2 shows the lowest score with 0.8402, the model trained with split 1 the highest with 0.8896.
- The mean macro F1 score results in 0.8623.

- The SD is 0.0096.
- The difference between test sets stands at 0.038.

Three of the five models performed better than the baseline with regard to the small test set. For the big test set, four out of five models outperformed the baseline.

	Small test			Big test		
Split	Baseline	Modified	p-value	Baseline	Modified	p-value
1	0.8776	0.8883	0.4012	0.8329	0.8747	0.0004*
2	0.9147	0.8802	0.9524	0.8019	0.8623	0.0006*
3	0.9043	0.9128	0.3303	0.8632	0.8636	0.4879
4	0.921	0.9547	0.099	0.8421	0.8631	0.0591
5	0.8659	0.8653	0.5078	0.86	0.8477	0.779

Table 29: Feature named entities and sentiment: results bootstrap testing

Again, statistical significance was not reached for the small test set. Models 1 and 2, however, show a significant difference to the baseline for the big test set.

Considering figure 36, an obvious outlier can be detected for model 4's prediction of the small test set. This outlier contributes to the comparatively high SD of 0.035 of the small test set results. Other than model 4, scores between test sets are very similar which is represented by the small difference of 0.038 between mean macro F1 scores. The models supplied with named entity and sentiment information show the smallest difference between test sets of all model configurations.

Mean F1 scores of the individual event types are included in table 30. For the small test set, five event types are on average classified with higher scores by the models using named entity and sentiment information. Of the remaining six event types, five are classified with the same score by the models and one event type, DIPLOMACY, displays a higher score for the baseline.

The big test set results show a greater advantage for the modified models. Ten of the eleven event types are predicted with higher scores. Differences

	Small test			Big test		
	Baseline	**Modified**	**Diff**	**Baseline**	**Modified**	**Diff**
Accident	0.97	0.98	0.01	0.95	0.96	0.01
Crime	0.96	0.96	0	0.91	0.92	0.01
Diplomacy	0.85	0.77	-0.08	0.77	0.78	0.01
Discovery	0.92	0.95	0.03	0.9	0.9	0
Disorder	0.99	0.99	0	0.8	0.84	0.04
Finance	0.62	0.63	0.01	0.64	0.74	0.1
Government	0.88	0.88	0	0.83	0.85	0.02
Life	0.89	0.94	0.05	0.87	0.88	0.01
Natural Disaster	0.94	0.94	0	0.95	0.96	0.01
Organisation	0.86	0.86	0	0.75	0.78	0.03
Win Lose	0.98	1	0.02	0.85	0.88	0.03

Table 30: Feature named entities and sentiment: mean F1 scores of all event types

of the mean scores are between 0.01 and 0.04. The only event type that does not improve with the addition of the features is DISCOVERY which stays the same with regard to its mean F1 score.

Error Analysis Model 4 Model 4 is chosen for a more detailed error analysis. Model 4 nearly reaches significance for both test sets in comparison to the baseline model (p-value of 0.099 for the small test set and 0.0591 for the big test set).

For the small test set, the comparison of the classification reports in tables 11 and 31 show that seven event types reach higher F1 scores when the modified model is used for prediction. Of the remaining four event types, two are predicted with higher scores by the baseline model (CRIME and LIFE) and another two show the same F1 score for both models (DISORDER and WIN LOSE). This suggests that the difference of 0.0337 in macro F1 scores between models stems from an overall better learned representation of event types by the modified model and not from a small number of outliers skewing the score.

Interestingly, the amount of event types that are predicted with higher

	Small test				Big test			
	Precision	**Recall**	**F1**	**F1 B**	**Precision**	**Recall**	**F1**	**F1 B**
Accident	0.96	1.00	0.98	0.97	0.94	0.97	0.95	0.96
Crime	0.94	0.91	0.93	0.96	0.92	0.92	0.92	0.90
Diplomacy	1.00	1.00	1.00	0.96	0.76	0.76	0.76	0.74
Discovery	1.00	1.00	1.00	0.91	0.87	0.95	0.91	0.90
Disorder	0.96	0.96	0.96	0.96	0.97	0.75	0.84	0.79
Finance	1.00	0.80	0.89	0.75	0.83	0.71	0.77	0.67
Government	0.88	0.95	0.91	0.88	0.81	0.90	0.85	0.82
Life	0.90	0.90	0.90	0.95	0.89	0.90	0.89	0.89
Natural Disaster	1.00	0.97	0.98	0.97	0.93	0.99	0.96	0.96
Organisation	1.00	0.91	0.95	0.84	0.83	0.71	0.76	0.77
Win Lose	1.00	1.00	1.00	1.00	0.98	0.79	0.87	0.88
Macro average	**0.97**	**0.95**	**0.95**	**0.92**	**0.88**	**0.85**	**0.86**	**0.84**

Table 31: Feature named entities and sentiment: classification reports model 4

scores by the baseline model increases for the big test set. Three event types hold higher scores, namely, ACCIDENT, ORGANISATION and WIN LOSE. The difference between scores is 0.01 for all three types. Additionally, LIFE and NATURAL DISASTER reach the same F1 scores.

The other six event types are predicted with higher scores when the modified model is used. The differences that are observed between the models in this case are, in parts, more distinct. For DISCOVERY, the improvement also lies at 0.01. CRIME and DIPLOMACY display an increase of 0.02, GOVERNMENT of 0.03, DISORDER of 0.05 and, most notably, FINANCE of 0.1.

The greatest improvements achieved by the modified model are thus seen for the classification of DISORDER and FINANCE. The errors that are made by the modified model regarding DISORDER and FINANCE are contrasted with the ones made by the corresponding baseline model (figures 19 and 37) in the following section.

For the event type FINANCE, the baseline model showed a systematic error with regard to the confusion with the event type GOVERNMENT. 29% of FINANCE texts were classified as GOVERNMENT. In case of the modified

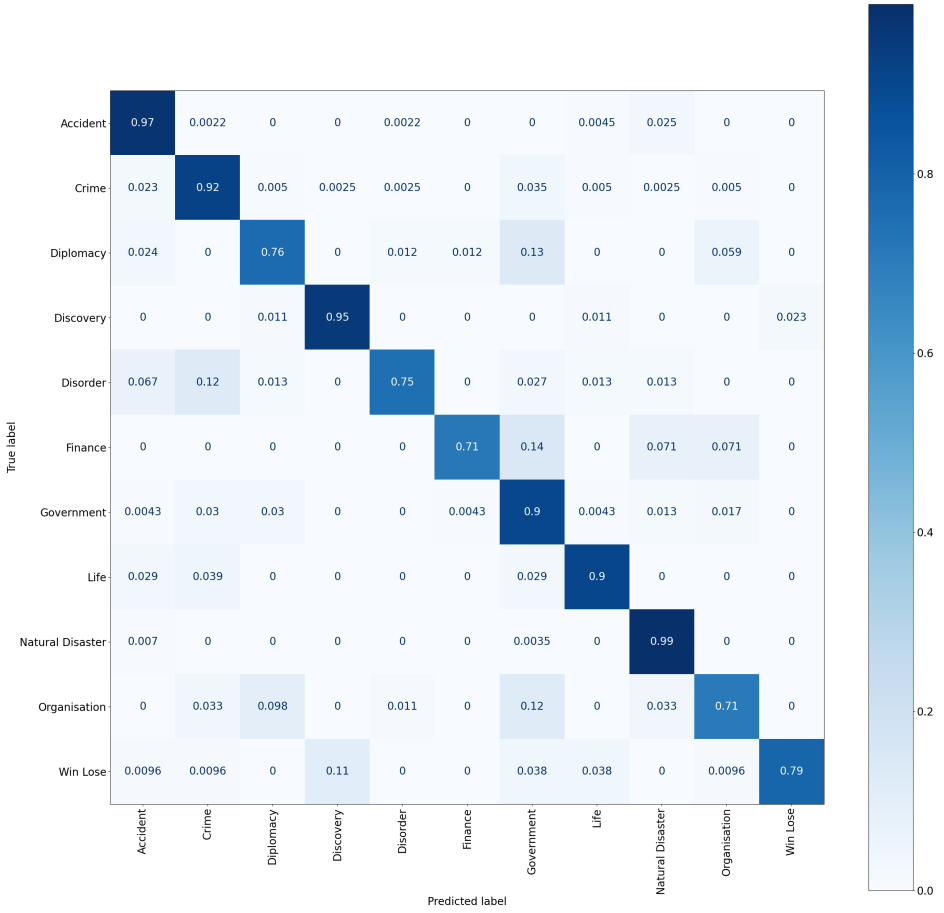

Figure 37: Feature named entities and sentiment: confusion matrix big test set predictions model 4

model, this error can also be detected. However, its severity is reduced. The amount of the errors now stands at 14%. It can be assumed that the feature encompassing named entity and sentiment information helps in differentiating the two event types. Still, the semantic overlap that is present for FINANCE and GOVERNMENT can not be compensated in its entirety.

With regard to the event type DISORDER, two notable improvements can be seen for the modified model. First, the error source represented by wrong CRIME predictions decreases by 5% (17% to 12%). Second, while the baseline model predicted GOVERNMENT for 11% of DISORDER texts, the modified model only shows this error 3% of the time.

It can be stated that the increase that was seen for the F1 scores of FINANCE and DISORDER for model 4 can be attributed to improvements regarding the event types' main error sources. The addition of the feature seems

to facilitate these rather difficult classification decisions.

This can also be observed for a third event type: WIN LOSE. Even though the baseline model outperforms the modified model with a small difference of 0.01 in this case, the modified model seems to be less susceptible with regard to the wrong prediction of the semantically similar event type DISCOVERY. The error is reduced from 15% to 11%.

The modified model analyzed here uses a combination of named entity and sentiment information. Considering the models that use only named entities and only sentiment information and were trained on data split 4, it becomes clear that they also exhibit the highest F1 scores for the small test set among all models using the respective feature. Additionally, model 4 using only sentiment information reaches significance compared to the baseline model for the big test set. This means that both features work well for split 4. The combination of features seems to result in an even greater advantage of the modified model for the classification of the small test set. Furthermore, the ability of the model to solve the task is transferable to the big test set.

It can be argued that the discriminatory power of the combined feature is higher than that of the single features. Named entities are added to the data by means of four distinct classes: person, organisation, GPE and location. The sentiment feature is represented by three classes: negative, neutral, positive. This means that there are inevitable overlaps between the event types with regard to the single features. Their combination results in more distinct additional linguistic information that can be leveraged by the model.

7.3.5 Named Entities and Aspectual Classes

The results of the feature combination named entities and aspectual classes can be seen in figure 38.

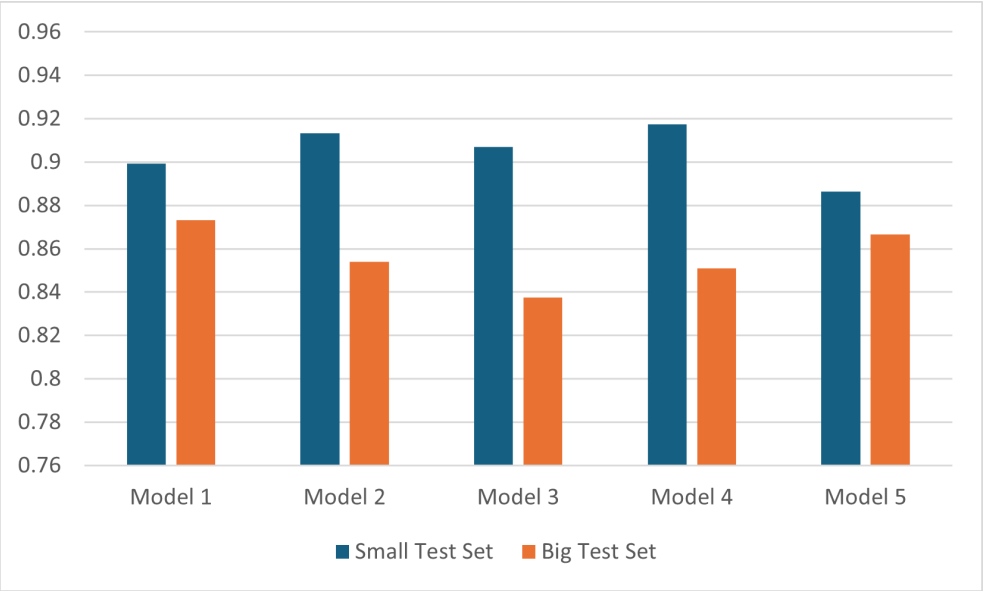

Figure 38: Feature named entities and aspectual classes: results main experiment

The detailed results are:

- The macro F1 scores lie between 0.8865 (model 5) and 0.9173 (model 4) for the small test set.
- The mean macro F1 score in this case is 0.9047.
- A SD of 0.0122 is observed.
- The scores for the big test set range from 0.8477 (model 5) to 0.8747 (model 1).
- A mean macro F1 score of 0.8565 is reached.
- The SD is 0.0139.
- Between the mean macro F1 scores of the respective test sets, a difference of 0.0482 is present.

The results in table 32 reveal that the modified models outperform the baseline in seven out of ten cases. However, for the small test set, models 2 and 4 reach slightly lower scores than the baseline model. Additionally, model 3 shows a lower score than the baseline for the big test set.

	Small test			Big test		
Split	Baseline	Modified	p-value	Baseline	Modified	p-value
1	0.8776	0.8994	0.3458	0.8329	0.8732	0.0024*
2	0.9147	0.9133	0.5554	0.8019	0.854	0.0017*
3	0.9043	0.907	0.414	0.8632	0.8375	0.9221
4	0.921	0.9173	0.5876	0.8421	0.851	0.2919
5	0.8659	0.8865	0.2662	0.86	0.8666	0.3041

Table 32: Feature named entities and aspectual classes: results bootstrap testing

Paired bootstrap testing showed no significance for the small test set. Significant p-values were reached for model 1 and interestingly also for model 2 for the big test set. This means that even though model 2 did not reach a higher macro F1 score than the baseline model for the small test set, the classification of the big test set reveals an advantage for the model. To consider this further, model 2 will be evaluated more closely in the next section.

The mean F1 scores of the event types in table 33 are mostly higher for the modified models.

	Small test			**Big test**		
	Baseline	**Modified**	**Diff**	**Baseline**	**Modified**	**Diff**
Accident	0.97	0.98	0.01	0.95	0.96	0.01
Crime	0.96	0.96	0	0.91	0.92	0.01
Diplomacy	0.85	0.83	-0.02	0.77	0.77	0
Discovery	0.92	0.94	0.02	0.9	0.89	-0.01
Disorder	0.99	0.99	0	0.8	0.87	0.07
Finance	0.62	0.62	0	0.64	0.7	0.06
Government	0.88	0.89	0.01	0.83	0.86	0.03
Life	0.89	0.93	0.04	0.87	0.88	0.01
Natural Disaster	0.94	0.93	-0.01	0.95	0.96	0.01
Organisation	0.86	0.87	0.01	0.75	0.77	0.02
Win Lose	0.98	1	0.02	0.85	0.85	0

Table 33: Feature named entities and aspectual classes: mean F1 scores of all event types

With regard to the small test set, six event types hold on average higher scores for the modified models' predictions, three are classified with the same scores

and two are predicted with higher scores by the baseline models.

The analysis of the results gained for the big test set reveal that higher scores are reached by the models using aspectual class and named entity information for eight event types. Additionally, two event types are predicted on average with the same score by the modified and the baseline models. Only one event type, DISCOVERY, is classified with higher scores by the baseline models.

Error Analysis Model 2 As mentioned above, model 2 was chosen for a detailed error analysis. The examination of the classification reports of the modified model in table 34 and the corresponding baseline model in table 9 shows two event types with distinctly higher scores for the baseline model regarding the small test set. These are DIPLOMACY and FINANCE, both with a difference of 0.06 between models.

	Small test				**Big test**			
	Precision	**Recall**	**F1**	**F1 B**	**Precision**	**Recall**	**F1**	**F1 B**
Accident	0.98	1.00	0.99	0.98	0.93	0.98	0.96	0.95
Crime	0.93	0.97	0.95	0.93	0.93	0.90	0.92	0.91
Diplomacy	0.86	0.86	0.86	0.92	0.76	0.78	0.77	0.76
Discovery	1.00	1.00	1.00	1.00	0.75	0.93	0.83	0.87
Disorder	1.00	1.00	1.00	1.00	0.90	0.82	0.86	0.67
Finance	0.83	0.62	0.71	0.77	0.83	0.71	0.77	0.57
Government	0.79	0.88	0.83	0.81	0.81	0.91	0.86	0.77
Life	1.00	0.87	0.93	0.90	0.92	0.84	0.88	0.88
Natural Disaster	0.90	0.96	0.93	0.91	0.94	0.98	0.96	0.93
Organisation	1.00	0.73	0.85	0.85	0.90	0.70	0.79	0.73
Win Lose	1.00	1.00	1.00	1.00	0.99	0.68	0.81	0.78
Macro average	**0.93**	**0.90**	**0.91**	**0.91**	**0.88**	**0.84**	**0.85**	**0.80**

Table 34: Feature named entities and aspectual classes: classification reports model 2

Five event types hold higher scores when classified by the modified model. The differences that are detected are, however, lower: ACCIDENT is classified with a difference of 0.01, CRIME with a difference of 0.02, GOVERNMENT with a difference of 0.02, LIFE with a difference of 0.03 and NATU-

RAL DISASTER with a difference of 0.02.

The listed differences in individual F1 scores lead to a minimally higher macro F1 score for the baseline model for the classification of the small test set (0.9147 vs. 0.9133). Even though the baseline model outperforms the modified model with regard to the macro F1 score, the modified model displays a greater number of event types with higher F1 scores.

This trend is strengthened further when the big test set results are considered. In this case, all event types except two are predicted with higher scores by the modified model. The baseline model reaches a higher score for DISCOVERY with a difference of 0.04. The same score is achieved for the event type LIFE by both models.

DIPLOMACY and FINANCE, the event types the baseline model predicted with higher scores regarding the small test set, now display lower scores in comparison to the modified model. For DIPLOMACY, the difference is 0.01. The prediction of FINANCE show a difference of 0.2 between the models.

A notable advantage of the modified model is also seen for the event types DISORDER as well as GOVERNMENT. The differences between the F1 scores are 0.19 and 0.09, respectively.

The error sources displayed in the confusion matrix in figure 39 stand in contrast to the ones detected for the baseline model in figure 17. For the two event types that are classified with markedly higher scores by the modified model, i.e., FINANCE and DISORDER, prevalent errors present for the baseline model are mitigated.

With regard to FINANCE, the confusion with GOVERNMENT is reduced by 36% (43% vs. 7%). However, the modified model's predictions include an error that was not seen for the baseline model, namely the wrong prediction of DISCOVERY (7%). This means that the model using named entities and aspectual class information is in general able to improve the classification of FINANCE. However, the addition of the features might also have created a

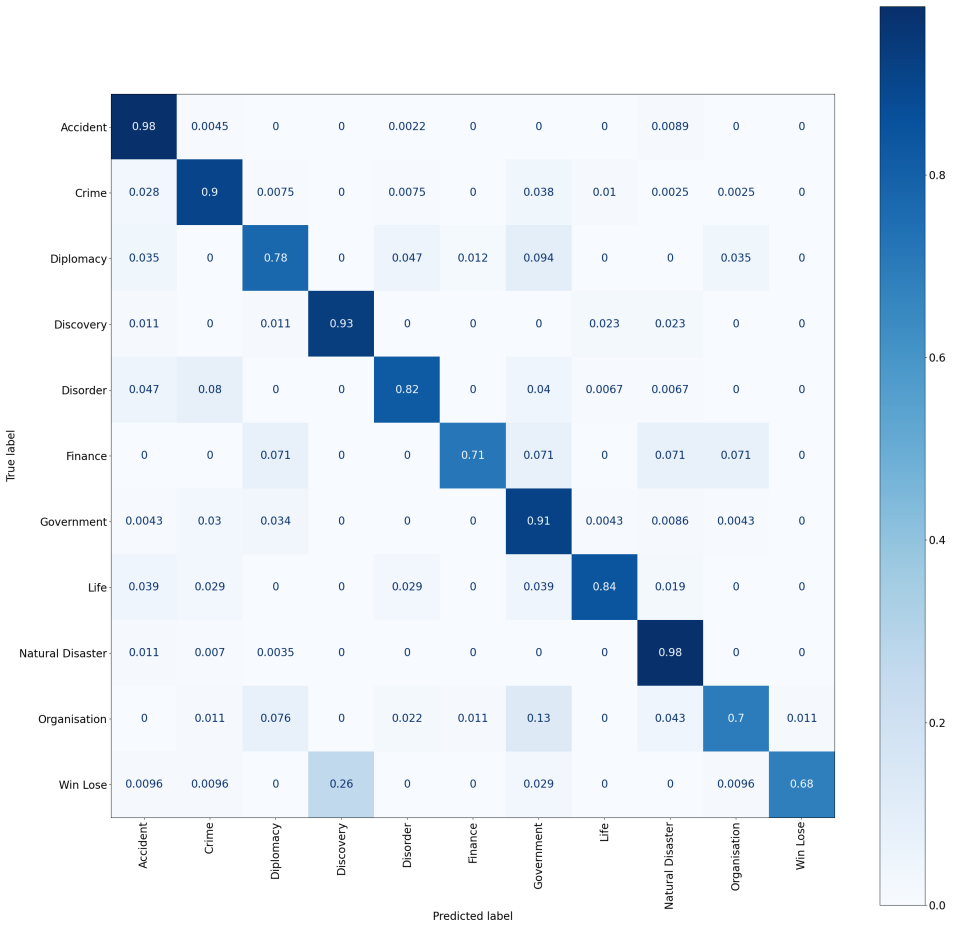

Figure 39: Feature named entities and aspectual classes: confusion matrix big test set predictions model 2

new, if definitely smaller, error source.

The greatest systematic error for DISORDER made by baseline model 2, the confusion with GOVERNMENT, decreases from 27% to 4%. Additionally, DISORDER texts that are wrongly predicted to belong to the event type CRIME are also less frequent (13% vs. 8 %).

For FINANCE and DISORDER, the additional information used for the training of the model at hand seems to improve systematic errors that were seen for the baseline model. However, for one event type, WIN LOSE, the added features seems to have the reverse effect.

As was mentioned in chapter 5.2.1, WIN LOSE is oftentimes confused with the event type DISCOVERY. This error can be observed for almost all models. In the case of model 2 using named entity and aspectual class information that is analyzed here, the error is reinforced. Baseline model 2 already

predicted 16% of WIN LOSE texts as DISCOVERY, the modified model increases this by 10%. This means that for 26% of WIN LOSE texts, the event type DISCOVERY is predicted.

The analysis of baseline model 2 (cf. chapter 5.2.1) hinted at the probable overfitting of the model. This overfitting has the effect that when comparing the modified model to the baseline model, the baseline model seems to have an advantage for the classification of the small test set. This advantage, however, disappears when the models are used to predict the more diverse big test set. The comparison of the models reveals higher scores for the modified model for this test set. This means that even though the baseline model reaches a higher macro F1 score than the modified model for the small test set, it is less robust. The additional features improve model performance. Overfitting is not observed for the modified model and thus, a better generalization capability is given.

The models that use only named entities as well as only aspectual class information and are trained with split 2 also outperform the baseline. In both cases, the main advantage is detected for the classification of the big test set. The combination of the features results in the same advantage over the baseline.

7.3.6 Aspectual Classes and Sentiment

The results of the models that were trained with aspectual classes and sentiment information are shown in figure 40.

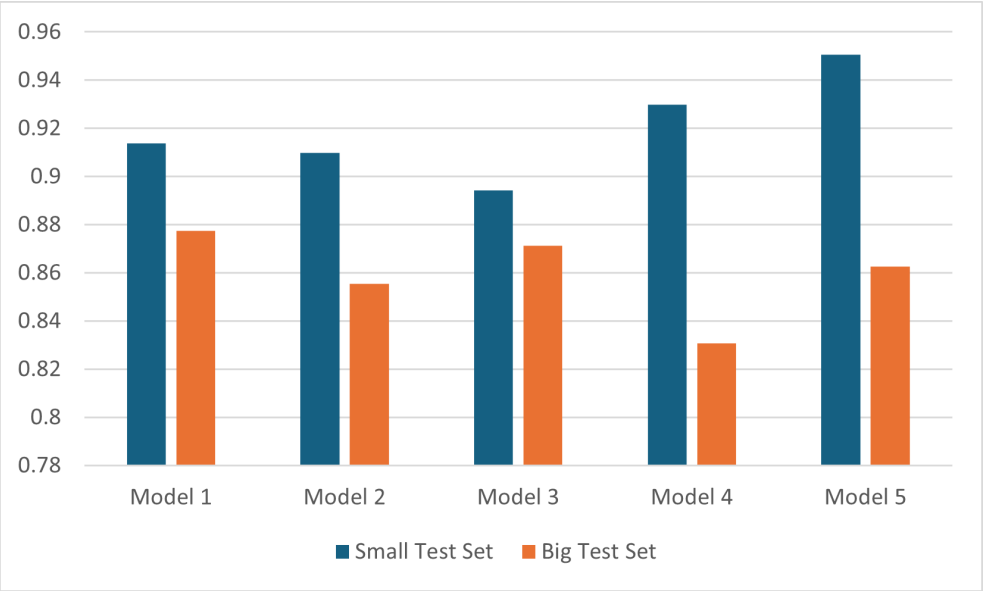

Figure 40: Feature aspectual classes and sentiment: results main experiment

Concretely, the following results were found:

- For the small test set, macro F1 scores between 0.8943 for model 3 and 0.9505 for model 5 are reached.
- The mean macro F1 score is 0.9196.
- The SD stands at 0.0214.
- For the big test set, macro F1 scores ranging from 0.8308 for model 4 and 0.8774 for model 1 are found.
- The mean macro F1 score in this case is 0.8595.
- The SD is 0.0181.
- Between the test sets, a difference of 0.0601 is observed.
- This difference between test sets is the largest of all model configurations.

The comparison of the modified models to the baseline models (cf. table 35) shows that for three of the ten test set predictions, the modified models reach lower scores than the baseline models. Those are models 2 and 3 with regard

	Small test			Big test		
Split	Baseline	Modified	p-value	Baseline	Modified	p-value
1	0.8776	0.9137	0.1646	0.8329	0.8774	0.0008*
2	0.9147	0.9098	0.6704	0.8019	0.8553	0.0008*
3	0.9043	0.8943	0.7423	0.8632	0.8713	0.2267
4	0.921	0.9297	0.3962	0.8421	0.8308	0.784
5	0.8659	0.9505	0.0069*	0.86	0.8626	0.437

Table 35: Feature aspectual classes and sentiment: results bootstrap testing

to the small test set and model 4 with regard to the big test set. Statistical significance is reached by modified models 1 and 2 when testing on the big test set as well as model 5 for the small test set classification. Model 5's significant advantage over the baseline model for the small test set is unique; no other model achieved significance for this test set prediction. Still, it can be seen that using model 5 for the classification of the big test set does not result in the same noteworthy performance. The difference between the baseline and the modified model becomes smaller and the corresponding p-value increases in contrast to the small test set. This might hint at model 5 being unstable to an extent. Still, the modified model generally reaches a higher score than the baseline model.

In table 36 the mean F1 scores of the event types that were reached by the models using aspectual class and sentiment information are listed. The modified models mostly outperform the baseline with regard to the small test set predictions. Seven event types have higher scores. The greatest improvement is seen for FINANCE with a difference of 0.15. Two event types are on average classified with the same scores, DISORDER and NATURAL DISASTER. The baseline models reach higher scores in the case of CRIME and DIPLOMACY.

The classification of the big test set results in the same scores for both model configurations for four event types, ACCIDENT, CRIME, DIPLOMACY and LIFE. Seven event types hold on average higher scores for the

	Small test			Big test		
	Baseline	**Modified**	**Diff**	**Baseline**	**Modified**	**Diff**
Accident	0.97	0.99	0.02	0.95	0.95	0
Crime	0.96	0.95	-0.01	0.91	0.91	0
Diplomacy	0.85	0.84	-0.01	0.77	0.77	0
Discovery	0.92	0.94	0.02	0.9	0.91	0.01
Disorder	0.99	0.99	0	0.8	0.85	0.05
Finance	0.62	0.77	0.15	0.64	0.72	0.08
Government	0.88	0.89	0.01	0.83	0.86	0.03
Life	0.89	0.92	0.03	0.87	0.87	0
Natural Disaster	0.94	0.94	0	0.95	0.96	0.01
Organisation	0.86	0.9	0.04	0.75	0.77	0.02
Win Lose	0.98	1	0.02	0.85	0.89	0.04

Table 36: Feature aspectual classes and sentiment: mean F1 scores of all event types

modified model. The biggest difference is again seen for FINANCE with 0.08.

Error Analysis Model 5 Model 5 is discussed in more detail in the following section. It was chosen due to its significant difference to the baseline for the small test set. It is the only model that reaches significance for this split and test set.

The classification reports for model 5 in table 37 show that for four classes an F1 score of 1 was reached for the small test set. These are ACCIDENT, DISORDER, FINANCE and WIN LOSE. Baseline model 5 also showed F1 scores of 1 for ACCIDENT and DISORDER. The greatest discrepancy between models can be observed for the event type FINANCE. The baseline model only reaches a score of 0.4 for the event type (cf. table 12).

The advantage of the modified model is, however, not present for all event types. Next to FINANCE, four other event types are predicted with higher scores than the baseline: DIPLOMACY with a difference of 0.08, LIFE with a difference of 0.11, ORGANISATION with a difference of 0.09 and WIN LOSE with 0.06. In the case of NATURAL DISASTER, the baseline shows a higher F1 score than the modified model (0.95 vs. 0.91). All other event

	Small test				Big test			
	Precision	**Recall**	**F1**	**F1 B**	**Precision**	**Recall**	**F1**	**F1 B**
Accident	1.00	1.00	1.00	1.00	0.95	0.98	0.96	0.96
Crime	0.96	0.96	0.96	0.96	0.91	0.92	0.91	0.92
Diplomacy	1.00	0.83	0.91	0.83	0.76	0.80	0.78	0.79
Discovery	0.88	1.00	0.93	0.93	0.82	0.97	0.88	0.94
Disorder	1.00	1.00	1.00	1.00	0.94	0.79	0.86	0.88
Finance	1.00	1.00	1.00	0.4	0.69	0.79	0.73	0.60
Government	0.93	0.93	0.93	0.93	0.81	0.90	0.85	0.84
Life	0.88	0.88	0.88	0.77	0.88	0.88	0.88	0.89
Natural Disaster	0.91	0.91	0.91	0.95	0.97	0.96	0.96	0.96
Organisation	0.94	0.94	0.94	0.82	0.87	0.74	0.80	0.76
Win Lose	1.00	1.00	1.00	0.92	0.98	0.77	0.86	0.93
Macro average	**0.95**	**0.95**	**0.95**	**0.87**	**0.87**	**0.86**	**0.86**	**0.86**

Table 37: Feature aspectual classes and sentiment: classification reports model 5

types hold the same scores for both models. This leads to the assumption that the significant difference that is observed between the performance of the modified model and the baseline model is mainly due to notably higher scores for a small number of event types.

This assumption is emphasized when considering the big test set. As mentioned above, the difference between models for the big test set no longer holds significance. Consulting the classification reports, it can be seen that only four event types are predicted with higher F1 scores by the modified model. The baseline model now reaches higher scores for five event types, namely CRIME, DIPLOMACY, DISCOVERY, DISORDER and WIN LOSE. For the modified model, the differences between scores range from 0.01 (GOVERNMENT) to 0.13 (FINANCE), for the baseline model from 0.01 (CRIME, DIPLOMACY, LIFE) to 0.07 (WIN LOSE). This means that even though the modified model worked notably better when testing on the small test set with a difference in macro F1 scores of 0.0846, this trend can not be observed for the big test set. The modified model's performance nearly aligns with the baseline, resulting in a difference of only 0.0026.

The confusion matrix in figure 41 gives further insights into the errors of

the modified model.

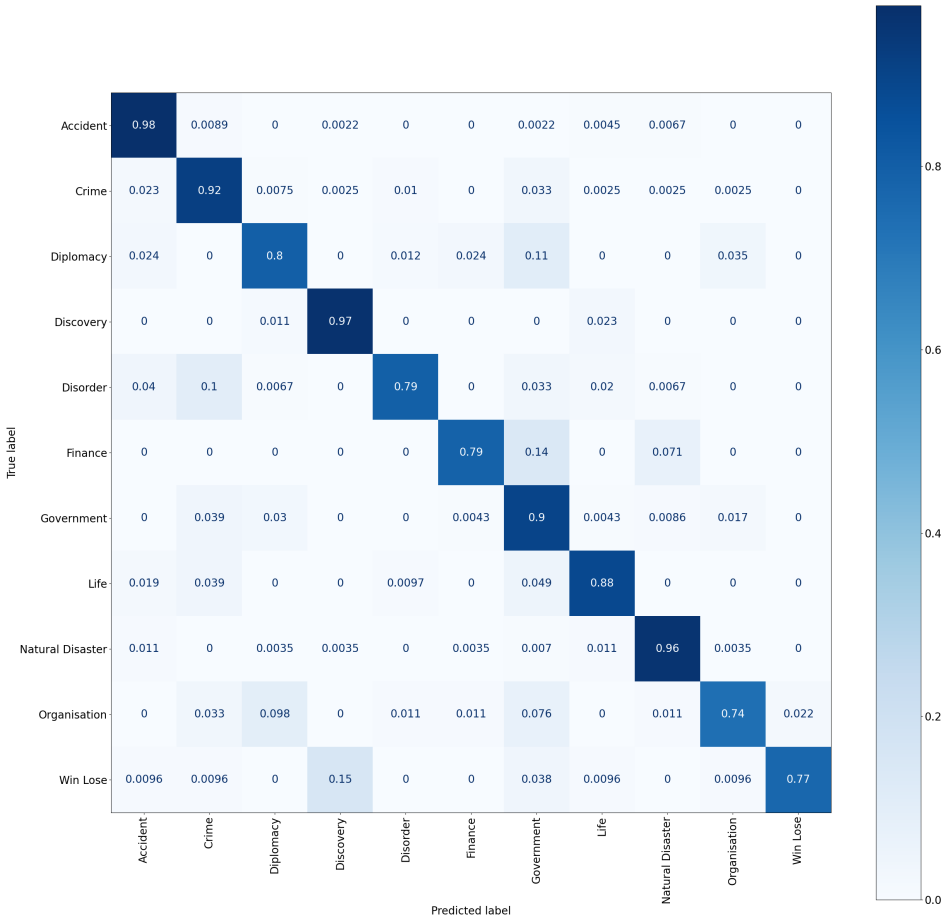

Figure 41: Feature aspectual classes and sentiment: confusion matrix big test set predictions model 5

The matrix shows that the mitigation of systematic errors that was seen for other modified models in contrast to the baseline, is less prevalent in this case. Merely FINANCE shows a considerable decrease regarding its main error source, the wrong prediction of GOVERNMENT (14% vs. 43%). Other than that, mainly smaller improvements can be observed.

Additionally, there are some instances where the confusion of similar event types is even promoted, e.g., the prediction of DISCOVERY for WIN LOSE texts increases from 6% to 15% and the prediction of GOVERNMENT for DIPLOMACY texts increases from 8% to 11%.

In summary it can be stated that in the case of split 5, the addition of the feature combination sentiment and aspectual classes does not greatly support the classification of more diverse data. The feature combination might be

focused on too much as it is indicative for the event types in the training data in this case. However, variance in the feature in the data included in the big test set leads to lower scores.

In general, the feature combination provides unstable results. This can be seen for the small test set where, on the one hand, two of the five models perform worse than the corresponding baseline model and, on the other hand, one model reaches significance. Additionally, p-values increase for the classification of the big test set in two cases (model 4 and 5). These models both achieved higher macro F1 scores than the baseline for the small test set. The apparent instability is further underlined by the comparatively high SDs for both test sets and the fact that the difference between test sets is the greatest of all model configurations.

7.3.7 Aspectual Classes, Named Entities and Sentiment

Figure 42 shows the results of the models trained with all three features, aspectual classes, named entities and sentiment.

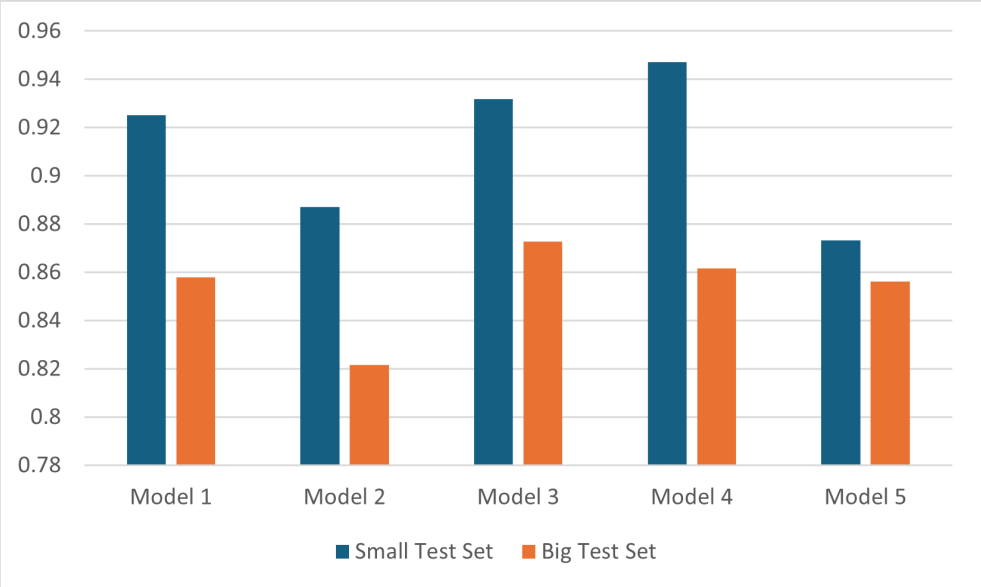

Figure 42: Feature aspectual classes, named entities and sentiment: results main experiment

For the models at hand, the key findings are:

- The macro F1 scores are between 0.887 (model 2) and 0.9471 (model 5) regarding the small test set.
- The mean macro F1 score results in 0.9128.
- The SD is 0.0313.
- For the big test set, macro F1 scores between 0.8215 (model 2) and 0.8726 (model 3) are reached.
- The mean macro F1 score amounts to 0.8539.
- A SD of 0.0192 is observed.
- The difference between the two test sets is 0.0589.

The modified models outperform the baseline most of the time (cf. table 38). Only model 2 for the small test set and model 5 for the big test set achieve lower scores than the respective baseline models. Significance is reached only once, by model 1 when tested on the big test set. However, models 2 and 4 nearly reach significance for the big test set with p-values of 0.0878 and 0.0761 respectively. Model 1 will be evaluated more closely in the following section.

	Small test			Big test		
Split	Baseline	Modified	p-value	Baseline	Modified	p-value
1	0.8776	0.925	0.1571	0.8329	0.8578	0.0122*
2	0.9147	0.887	0.8828	0.8019	0.8215	0.0878
3	0.9043	0.9319	0.1006	0.8632	0.8726	0.1489
4	0.921	0.9471	0.2171	0.8421	0.8616	0.0761
5	0.8659	0.8732	0.4234	0.86	0.8561	0.614

Table 38: Feature aspectual classes, named entities and sentiment: results bootstrap testing

The mean F1 scores of all event types that were reached by the models using all three features are shown in table 39.

	Small test			Big test		
	Baseline	**Modified**	**Diff**	**Baseline**	**Modified**	**Diff**
Accident	0.97	0.98	0.01	0.95	0.95	0
Crime	0.96	0.96	0	0.91	0.91	0
Diplomacy	0.85	0.83	-0.02	0.77	0.77	0
Discovery	0.92	0.98	0.06	0.9	0.92	0.02
Disorder	0.99	0.99	0	0.8	0.84	0.04
Finance	0.62	0.69	0.07	0.64	0.67	0.03
Government	0.88	0.9	0.02	0.83	0.84	0.01
Life	0.89	0.94	0.05	0.87	0.88	0.01
Natural Disaster	0.94	0.93	-0.01	0.95	0.95	0
Organisation	0.86	0.85	-0.01	0.75	0.76	0.01
Win Lose	0.98	0.99	0.01	0.85	0.89	0.04

Table 39: Feature aspectual classes, named entities and sentiment: mean F1 scores of all event types

The evaluation of the small test set predictions show that six event types have on average higher F1 scores when the modified model is used. The greatest increase in F1 score is observed for FINANCE with a difference of 0.07. The five remaining event types can be split into those that do not display a difference between models (CRIME and DISORDER) and those that are predicted with higher scores by the baseline models (DIPLOMACY, NATURAL DISASTER, ORGANISATION).

In contrast to the small test set, the modified models always reach at least the same F1 score as the baseline models for the big test set. Four event types are on average predicted with the same F1 score (ACCIDENT, CRIME, DIPLOMACY and NATURAL DISASTER). The seven remaining event types are classified with higher scores by the modified models. The differences in this case lie between 0.01 (GOVERNMENT, LIFE, ORGANISATION) and 0.04 (DISORDER, WIN LOSE).

Error Analysis Model 1 Model 1 is the only model that reaches a statistically significant difference to the baseline for the given feature configuration. Therefore, it will be analyzed in more detail.

For the small test set, model 1 using aspectual class, named entity and sentiment information generally outperforms the baseline. The classification report in figure 40 shows that in comparison to the baseline (cf. figure 8), the modified model reaches higher scores for six event types. Four of these event types, ACCIDENT, CRIME, GOVERNMENT and NATURAL DISASTER exhibit differences between 0.01 and 0.05. More notable differences can be observed for FINANCE (0.36) and LIFE (0.12). One event type, ORGANISATION, is predicted with a higher F1 score by the baseline model, displaying a difference of 0.1. Four event types remain that have the same scores for both models (DIPLOMACY, DISCOVERY, DISORDER and WIN LOSE).

	Small test				**Big test**			
	Precision	**Recall**	**F1**	**F1 B**	**Precision**	**Recall**	**F1**	**F1 B**
Accident	1.00	0.97	0.98	0.95	0.95	0.96	0.95	0.95
Crime	1.00	0.96	0.98	0.96	0.90	0.94	0.92	0.91
Diplomacy	0.80	0.57	0.67	0.67	0.74	0.76	0.75	0.78
Discovery	0.86	1.00	0.92	0.92	0.89	0.91	0.90	0.86
Disorder	1.00	1.00	1.00	1.00	0.92	0.81	0.86	0.83
Finance	0.75	1.00	0.86	0.50	0.80	0.57	0.67	0.58
Government	1.00	0.95	0.97	0.93	0.85	0.86	0.86	0.86
Life	1.00	1.00	1.00	0.88	0.92	0.89	0.91	0.84
Natural Disaster	0.95	0.98	0.96	0.91	0.92	0.99	0.95	0.94
Organisation	0.75	0.92	0.83	0.93	0.84	0.72	0.77	0.77
Win Lose	1.00	1.00	1.00	1.00	0.97	0.84	0.90	0.83
Macro average	**0.92**	**0.94**	**0.92**	**0.88**	**0.88**	**0.84**	**0.86**	**0.83**

Table 40: Feature aspectual classes, named entities and sentiment: classification reports model 1

The classification of the big test set makes the advantage of the modified model even clearer and results in a significant difference to the baseline model. Out of the eleven event types, seven are classified with higher scores by the modified model. Differences between 0.01 and 0.04 are observed for most of them: CRIME (0.01), DISCOVERY (0.04), DISORDER (0.03), NATURAL DISASTER (0.01). For FINANCE and LIFE, the difference between models exceeds 0.04. Still, the differences are not as pronounced as for the small test set. For FINANCE a difference of 0.09 and for LIFE a difference of 0.07

is reported. In contrast to the small test set, the modified model also has an advantage for WIN LOSE with a difference of 0.07.

While the baseline model outperformed the modified model with regard to the event type ORGANISATION for the small test set, the event type is now predicted with the same F1 score by both models. The event type DIPLOMACY is the only one that is predicted with a higher score by the baseline model. A difference of 0.03 to the modified model is observed.

The most striking event types regarding the comparison of the models are thus FINANCE, LIFE and WIN LOSE, as they display the greatest improvement when using the modified model, and DIPLOMACY as the only event type that was classified with a higher score by the baseline model.

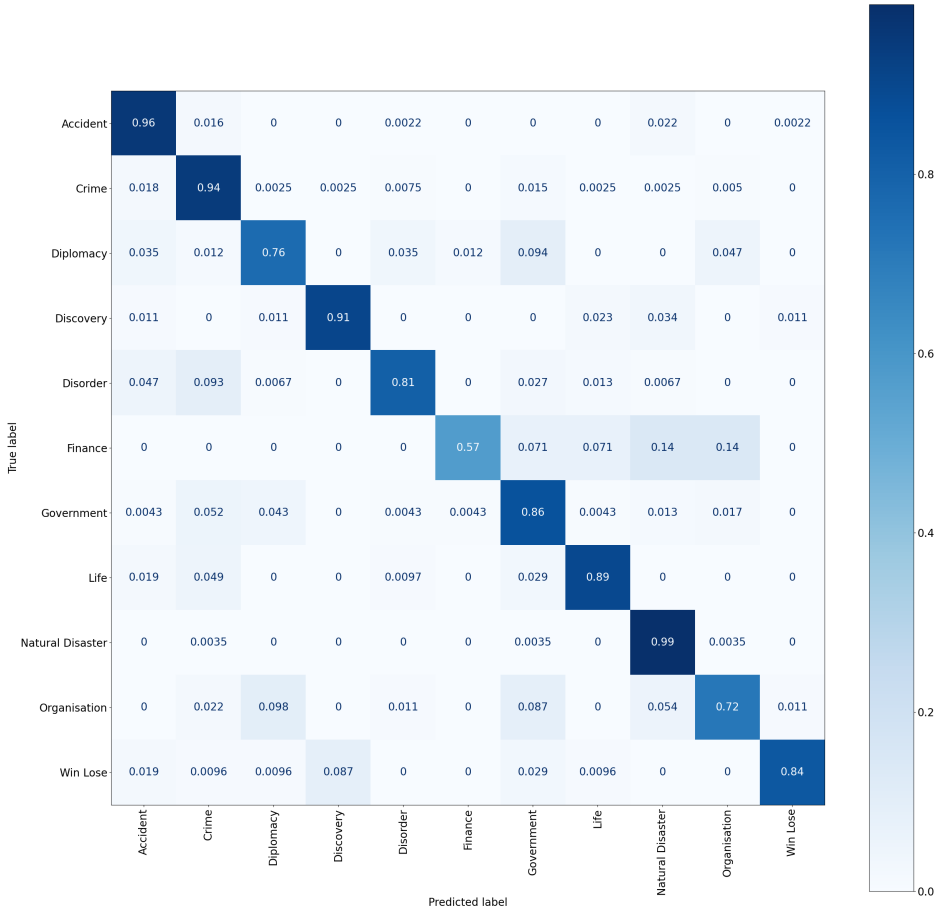

Figure 43: Feature aspectual classes, named entities and sentiment: confusion matrix big test set predictions model 1

The confusion matrix in 43 supports the further analysis of the mentioned event types. Starting with FINANCE, the class that holds the greatest differ-

ence in this case, it can be observed that the errors made by the two models are similar. The only difference lies in the confusion with GOVERNMENT. This error was also observed for numerous other models. In this case, the modified model predicts 7% of FINANCE texts to belong to GOVERNMENT. The baseline model predicted GOVERNMENT for 14% of texts. This means that the addition of the feature combination aspectual classes, named entities and sentiment mitigates the error.

For the event type LIFE, two error sources are reduced when using the modified model. First, the confusion with ACCIDENT which amounted to 5% for the baseline model predictions, only occurs with 2%. Second, the faulty prediction of the event type CRIME is decreased by 3% (8% vs. 5%). The partial avoidance of both errors by the modified model then results in a higher (recall) score for the class.

Errors related to WIN LOSE and their differences between the two models spread over a number of classes and are therefore more complex. Wrong predictions of ACCIDENT, CRIME, DISCOVERY, GOVERNMENT and LIFE can be observed with regard to the baseline model. For the modified model, an additional, albeit small, error source is the wrong prediction of DIPLOMACY. In contrast to FINANCE and LIFE described above where some errors remained the same between models, changes in error rates are present for all event types. In the case of ACCIDENT, CRIME, DISCOVERY and LIFE, the modified model reaches lower error rates. The improvement ranges from 1% (ACCIDENT, CRIME) to 8% (DISCOVERY). This means that the greatest decrease in errors connected to the event type WIN LOSE is found for the confusion with the semantically similar event type DISCOVERY. Still, an increase of errors can be seen for DIPLOMACY (1%) and GOVERNMENT (1%). These increases are however minor. In general, it can be stated that while there is more complexity in the errors connected to the event type WIN LOSE, the most distinct advantage of the modified model can be found for the differentiation of WIN LOSE and DISCOVERY texts.

For DIPLOMACY, a rise in errors is detected for the modified model's predictions. The confusion matrix reveals that this rise is connected to three event types, namely DISCOVERY, DISORDER and GOVERNMENT. DISCOVERY and DISORDER show an increase of errors of 1% and 2% respectively. A rise of 3% is observed for GOVERNMENT.

The error analysis revealed that the modified model works similar to the baseline model. Systematic error that were observed for the baseline model are mitigated by the modified model. However, error sources are not dissolved entirely. Still, the modified model shows improvements. The combination of all three features supports event type classification in this case.

8 Discussion

The previous chapter presented the results of the main experiment of this work. The modified models were contrasted with the baseline models. For a selection of models, detailed error analyses were conducted. To reiterate the key figures again, namely, mean macro F1 scores, standard deviations and the differences in scores between test sets, table 41 gives an overview of all models. Deep green cells represent the best model, medium green cells indicate the second best, light green shows the third best.

The results of the main experiment can be discussed in the context of the research questions that were posed after the preliminary experiment (cf. chapter 5.3). The first research question was:

1. Do transformer models like RoBERTa benefit from the explicit addition of manually chosen features for the task of event type classification?

All in all, 35 models were trained with additional features. Each model was tested on two separate test sets. This results in 70 points of comparison in total between the modified and the baseline models. 57 times, the modified models reached higher macro F1 scores than the corresponding baseline models. This equals 81.43%. Differences between the small and the big test set results can be observed. Out of the 35 test runs made on the small test sets, 27 (77.14%) resulted in higher scores for the modified models. For the big test set, the modified models had an advantage over the baseline for 30 of the 35 test set predictions (85.71%). This means that the use of the additional features seems to contribute slightly more to the big test set predictions which focused on testing the models' ability to generalize.

The difference between the small and big test set classification results can be interpreted as follows: The additional features that were chosen for the task represent information that is not dependent on individual tokens that were seen during training. Aspectual classes, for example, focus on a tempo-

	Mean macro F1 small	SD small	Mean macro F1 big	SD big	Diff between test sets
Baseline	0.8967	0.0239	0.84	0.0247	0.0567
Aspectual classes	0.9094	0.0176	0.8644	0.0166	0.045
Named entities	0.9203	0.0096	0.8607	0.0105	0.0596
Sentiment	0.9175	0.0192	0.8665	0.0177	0.051
Named entities + sentiment	0.9003	0.035	0.8623	0.0096	0.038
Named entities + aspectual classes	0.9047	0.0122	0.8565	0.0139	0.0482
Aspectual classes + sentiment	0.9196	0.0214	0.8595	0.0181	0.0601
Aspectual classes + named entities + sentiment	0.9128	0.0313	0.8539	0.0192	0.0589

Table 41: Overview all model results

ral dimension of the main verb in its context defined by notions of dynamicity, telicity and duration. For event type classification, the main verb oftentimes represents the event trigger (cf. Linguistic Data Consortium 2005), making it important for classification. This temporal dimension can stay the same for texts belonging to the same class without including (nearly) identical tokens. Consequently, this means that the pattern created by using the additional linguistic feature is transferable. The resulting advantage can be seen especially for texts that include textual patterns that diverge more strongly from the database used for training (i.e., for the big test set).

Differences between models that resulted in statistical significance with a preset level of 0.05 were less frequent. Significance was reached 15 times in total. Only once it was found for the small test set, the remaining 14 times were found for the big test set. This again points to the fact that the addition of features is more helpful for the classification of the big test set and therefore seems to support the prediction of more diverse textual instances. Considering a less strict preset level of significance of 0.1, eight additional significant differences are detected, four for the classification of the small test set, another four for the big test set.

It has to be taken into account that p-values are dependent on the sample size, meaning that "[t]he larger the data size, the larger the accuracy of the statistical test, and therefore, the stronger the evidence against or in favor of the null hypothesis" (Gómez-de-Mariscal et al. 2021: 1). As the name suggests, the big test set, holding 2000 entries, is larger than the small test sets which hold approximately 200 instances each. It can be assumed that this has an influence on statistical significance testing. The results observed for the big test set are therefore more reliable than those seen for the small test set.

In summary, the results that were obtained during the main experiment indicate that the addition of explicit features can help the task of event type classification. The analysis of table 41 supports this. It shows that the baseline

models reached the lowest mean macro F1 scores for the classification of both test sets, a comparatively high SD with regard to the small test set and the highest SD regarding the big test set.

The second research question:

2. Which features are sensible for the task of event type classification?

aimed at the evaluation of the chosen features in comparison to each other. Features that were used are aspectual classes, named entities and sentiment.

A number of insights can be gained when comparing the results of the models using the individual features and feature combinations to each other (cf. table 41). First, it is seen that models trained with three specific data modifications appear to be most promising. Those modifications are: aspectual classes, named entities as well as the combination of named entities and sentiment.

The models incorporating only named entity information have their main advantage when it comes to the classification of the small test set. They reach the highest mean macro F1 score of all models (0.9203) and display the smallest SD (0.0096). However, the models also display a relatively high difference of 0.0596 between mean test set scores which makes them slightly less stable than the baseline which displayed a difference of 0.0567. This means that the models using the feature encounter difficulties for the classification of the big test set.

The error analysis of model 5 in chapter 7.3 already showed this effect. It was seen that systematic errors, namely the confusion of DIPLOMACY and GOVERNMENT as well as WIN LOSE and DISCOVERY, increased for the big test set in contrast to the baseline. These shortcomings are not only observed for model 5 but for all models that use the named entity feature. Table 24 shows this with regard to the mean F1 scores achieved for DIPLOMACY and DISCOVERY. Both event types are on average classified with

lower scores by the modified models for the big test set.

However, as the models that use named entities reach on average at least the same and at times notably higher scores (cf. DISORDER and FINANCE) for the remaining event types, the mean macro F1 score is still higher than the one reached by the baseline models (0.8607 vs. 0.84) for the big test set.

The small and the big test set differ with regard to their textual content. The small test sets include texts that consist of event argument sentences. It can be argued that the named entity feature works especially well for this database as event arguments, and therefore the named entities that represent them, stay the same for events of the same type. The big test set includes entire newspaper articles. In these articles the broader context of the event is described. This might lead to diverging amounts of the named entity classes. The reduction of texts to event arguments is not always feasible for vast amounts of data. Therefore, the advantage of the named entity feature is not as pronounced for the classification of more diverse data.

The usage of aspectual classes as well as the combination of named entities and sentiment are advantageous for the big test set classification. The two models reach the second and third highest mean macro F1 scores of 0.8644 and 0.8623 respectively. Additionally, the combination of named entities and sentiment results in the smallest SD for the big test set (0.0096). Aspectual classes and named entities and sentiment also display the two smallest differences between the test set predictions. The average difference between the mean macro F1 scores of the small and the big test set of all models is 0.0522. For the aspectual class feature, a difference of 0.045 is observed, for named entities and sentiment the difference is 0.038. This means that both features not only support the classification of the big test set but also exhibit stable classification performances.

As mentioned in chapter 7.3.4, the feature incorporating named entities and sentiment has a higher discriminatory power than the two individual features. The named entity feature as well as the sentiment feature already reach

comparatively high scores. For the named entity feature the main advantage is seen for the small test set. For the sentiment feature it is seen for the big test set (cf. table 41). Their combination leads to robust predictions by the models.

A distinctiveness that is observed for the aspectual class feature is that the resulting models reach higher scores than the baseline for all ten test set predictions. It can be argued that the aspectual class feature is most closely associated with the texts. It is focused on temporal characteristics of verbs and their context. The pre-training of transformer models is implemented by processing large amounts of data. Therefore, it can be assumed that the features that define aspectual classes, i.e., dynamicity, duration and telicity, are already encoded in the resulting contextual embeddings. The usage of aspectual classes as an overt feature for classification of event types reinforces the encoded patterns but does not introduce entirely new information. Therefore, the models using the aspectual class feature follow the baseline models closely.

Additionally it was observed that p-values improve for all big test set predictions made by the models using the aspectual class feature in contrast to the small test set predictions and generally converge towards significance. This trend is not observed this clearly for any of the other features. Table 42 shows how the other features performed in contrast to the baseline.

The comparison of the performance of the modified models reveals patterns. Five of the seven modified models reach lower scores than the baseline for the small test set classification with model 2. The analysis of baseline model 2 in chapter 5.2.1 hinted strongly at the fact that model 2 is overfitting which leads to a high macro F1 score for the small test set. Consequently, the modified models which do not appear to be overfitting reach mostly lower scores. The prediction of more diverse data as included in the big test, however, results in a stark decrease in macro F1 score for the baseline model. The advantage that is provided by the additional linguistic features can then be observed: Six of

Feature	Number of test set scores higher than the baseline (out of 10)	Models with a lower score; which test set	Models with an increase in p-value for the big test set
Aspectual classes	10	-	-
Named entities	9	Model 5; big test set	Model 5
Sentiment	9	Model 2; small test set	Model 3, model 5
Named entities + sentiment	7	Model 2; small test set Model 5; both test sets	Model 3, model 5
Named entities + aspectual classes	7	Model 2; small test set Model 4; small test set Model 3; big test set	Model 3, model 5
Sentiment + aspectual classes	7	Model 2; small test set Model 3; small test set Model 4; big test set	Model 4, model 5
Aspectual classes + named entities + sentiment	8	Model 2; small test set Model 5; big test set	Model 3, model 5

Table 42: Feature comparison

the modified models trained with split 2 reach significantly higher scores for the big test set.

Another pattern that is observed is that three of the seven modified models achieve lower scores than the baseline model for the classification of the big test set with model 5. P-values increase for the classification of the big test set for six modified models. Baseline model 5 exhibits a small difference between test sets as described in chapter 5.2.1. This was interpreted as a good capability of the model to classify more diverse data. Accordingly, the increase in p-values can be explained: The small test set classification of baseline model 5 results in a relatively low score, the modified models reach higher scores and p-values are converging towards significance. The classification of the big test set by baseline model 5, however, leads to a comparatively high macro F1 score. The difference between the baseline and the modified models becomes smaller and p-values increase again. The only modified model that still shows a decrease in p-value in this case is the one using aspectual classes.

In summary, it can be stated that all features that were chosen for this work improve the mean macro F1 scores reached for the classification of the small and the big test set as well as the SDs for the big test set. There are however differences between the features as such.

Research question 3 focuses on the comparison of errors made by the baseline and modified models.

3. Systematic errors represent shortcomings of the chosen model with regard to solving the task with the given data, i.e., how capable is the model as such to work with the database/task. Do the added features target the identified main error sources or are other effects observed?

Systematic errors that were identified for the baseline models are connected to

- the confusion of semantically similar event types (specifically, DIPLO-

MACY texts classified as GOVERNMENT, DISORDER texts classified as CRIME, ORGANISATION texts classified as either DIPLOMACY or GOVERNMENT, WIN LOSE texts classified as DISCOVERY),

- the incorrect prediction of event types for which only small amounts of data are available (FINANCE) and
- the models' restricted ability to generalize to more diverse data as shown by the difference between the results of the small and the big test set classification.

To answer research question 3, the mean F1 scores that were reached for the individual event types are consulted. Overall, there are 154 points of comparison (11 event types for each of the seven features for two test sets respectively). In 106 instances, the modified models reach higher scores than the baseline models. This equals 68.83%. 35 times the event types are on average predicted with the same score (22.73%) by the baseline and the modified models. This leaves 13 instances (8.44%) that display higher scores for the baseline models. Of these 13 instances, ten are found for the small test set, three for the big test set.

The differences in event type scores between the baseline and the modified models vary in size. An overview of the mean differences between the baseline models on the one hand and all modified models on the other hand is shown in table 43. Cells of event types that are on average predicted with higher scores by the modified models are colored green. No difference between models is indicated by white cells. Red cells show event types that were predicted with higher scores by the baseline models.

The event types that benefit the least from the addition of the linguistic features are: ACCIDENT, CRIME, DIPLOMACY and NATURAL DISASTER. ACCIDENT, CRIME and NATURAL DISASTER generally exhibit the highest F1 scores with regard to the baseline models. ACCIDENT is classified

	Small test	Big test
Accident	0.01	0.01
Crime	0	0.01
Diplomacy	-0.01	0
Discovery	0.03	0
Disorder	0	0.06
Finance	0.06	0.09
Government	0.01	0.02
Life	0.04	0.01
Natural Disaster	0	0.01
Organisation	0.02	0.02
Win Lose	0.02	0.03

Table 43: Mean differences of event type scores between modified models and baseline models

with mean scores of 0.97 (small test set) and 0.95 (big test set), CRIME with 0.96 and 0.91 and NATURAL DISASTER with 0.94 and 0.95. This means that these three event types do not pose a problem for classification. The error analyses of the models conducted in this work supports this assumption. No obvious systematic problems with regard to the classification of these event types were identified.

The distributions of event types in the training, development and test sets shown in figures 11 - 14 reveal that ACCIDENT, CRIME and NATURAL DISASTER are represented with the most samples in the datasets. The availability of a high number of training samples contributes to classification success. Even though the added features do not improve the scores for these specific event types noticeably, they also do not hinder classification. This means that no diverging patterns are introduced by the features that could influence the linguistic representation of the event types negatively. Therefore, the addition of the features is unproblematic.

The fourth event type that does not improve with the addition of the features, DIPLOMACY, is classified with lower scores by the baseline models than ACCIDENT, CRIME and NATURAL DISASTER. It reaches a mean F1

score of 0.85 for the small test set and a mean F1 score of 0.77 for the big test set. Consequently, there is a general need and room for improvement. A more detailed consultation of the scores reached by the different modified models with regard to DIPLOMACY shows that the models that use named entities and sentiment as a feature (cf. table 30) contribute the most to the negative difference of -0.01 observed for the small test set. The feature in this case appears to actually interfere with the correct classification of the event type and leads to a mean F1 score of 0.77 for the small test set. A reoccurring error that is detected for the classification of DIPLOMACY is the confusion with GOVERNMENT. DIPLOMACY and GOVERNMENT have very similar distributions of sentiment classes (cf. figure 27) and similar amounts of texts that are annotated with the named entity person (cf. figure 26). The combination of the two features seems to strengthen the error for the classification of the small test set.

Even though the mean F1 score reached for DIPLOMACY does not increase for the big test set, errors connected to it can be consulted shortly. Table 44 includes the event type confusions that were identified for the baseline models in chapter 5.2.1 again, along with the corresponding error rates of the modified models. The main errors are written in bold.

In chapter 7.2 it was assumed that the main error connected to DIPLOMACY, i.e., the confusion with GOVERNMENT, might be improved by the aspectual class feature. Table 44 does show a small decrease of the mean error rate when the feature is used.

The table additionally shows an increase for the main error of DIPLOMACY with regard to the named entity feature. As mentioned before, named entities, which oftentimes represent event arguments in the data, are similar for DIPLOMACY and GOVERNMENT. The named entity feature introduces a disadvantage for the classification of DIPLOMACY for the big test set.

On average, the modified models display the greatest improvements for the F1 scores of the event types DISORDER, FINANCE, GOVERNMENT,

ORGANISATION and WIN LOSE. The wrong prediction of the event types DISORDER, FINANCE, ORGANISATION and WIN LOSE was identified as a common error source for the task of event type classification in this work (cf. chapter 5.2.1). The classification of texts belonging to the event type GOVERNMENT was not as error-prone. However, GOVERNMENT oftentimes constituted the class that was predicted instead of the correct class for a number of event types (cf. table 14). The way in which the classification improves differs between event types and features.

For DISORDER, no increase in mean F1 score is observed for the small test set. The classification of DISORDER by the baseline models already results in a very high mean F1 score of 0.99. The modified models match this performance. The addition of linguistic features improves the classification of DISORDER when it comes to the big test set. The error analyses in chapter 7.3 already showed that for the selection of modified models that was analyzed in more detail, the confusion of DISORDER and CRIME is less frequent for the big test set predictions. Table 44 gives further insights into changes connected to the confusion of the event types. It can be seen that all features mitigate the error. This improvement was expected (cf. chapter 7.2).

The performance of the sentiment feature is surprising to an extent. Only small improvements were expected due to the similarities in sentiment distributions for the event types DISORDER and CRIME in the data. The consultation of the error rates reached by the individual models offers an explanation: In the case of model 3 the error increases by 2%. For models 1, 2 and 5, improvements of 4% are observed. For model 4, however, an improvement of 10% is seen. Model 4 using the sentiment feature is the only model trained with data split 4 that reached significance in contrast to the baseline (cf. chapter 7.3.3). This suggests that the event types' representations learned by the model are generally more robust than the ones learned by other models. The great improvement for the confusion of DISORDER and CRIME for model 4 might therefore not only be due to the addition of the feature but also to the

Event type	Classified as	Baseline	Aspectual classes	Named entities	Sentiment	Named entities + sentiment	Named entities + aspectual classes	Aspectual classes + sentiment	Aspectual classes + named entities + sentiment
ACCIDENT	NATURAL DISASTER	2	2	2	1	2	1	1	1
CRIME	ACCIDENT	2	2	2	2	2	2	3	2
CRIME	GOVERNMENT	3	4	3	4	3	3	3	4
DIPLOMACY	ACCIDENT	2	3	2	3	2	3	3	2
DIPLOMACY	**GOVERNMENT**	**11**	**10**	**13**	**7**	**10**	**10**	**11**	**8**
DIPLOMACY	ORGANISATION	4	5	6	6	5	6	3	3
DISCOVERY	LIFE	2	2	1	2	2	2	2	2
DISORDER	ACCIDENT	5	5	5	6	5	5	5	5
DISORDER	**CRIME**	**13**	**10**	**11**	**9**	**11**	**10**	**10**	**10**
DISORDER	GOVERNMENT	9	6	3	4	6	3	4	6
FINANCE	DIPLOMACY	3	3	4	3	3	7	3	7
FINANCE	**GOVERNMENT**	**26**	**14**	**11**	**14**	**7**	**11**	**17**	**11**
FINANCE	LIFE	3	0	4	0	3	1	0	4
FINANCE	NATURAL DISASTER	8	8	4	7	6	6	6	8
FINANCE	ORGANISATION	7	1	4	4	11	13	7	10
GOVERNMENT	CRIME	4	3	4	3	4	4	4	3
GOVERNMENT	DIPLOMACY	3	3	3	3	3	3	3	4
LIFE	ACCIDENT	2	2	2	2	2	3	5	3
LIFE	CRIME	7	5	6	4	6	6	6	5
LIFE	GOVERNMENT	4	3	2	2	3	3	3	5
ORGANISATION	CRIME	2	2	1	1	2	2	3	4
ORGANISATION	**DIPLOMACY**	**11**	**10**	**9**	**10**	**8**	**9**	**12**	**10**
ORGANISATION	**GOVERNMENT**	**14**	**14**	**11**	**11**	**11**	**11**	**12**	**10**
WIN LOSE	**DISCOVERY**	**11**	**11**	**13**	**11**	**11**	**14**	**10**	**8**
WIN LOSE	GOVERNMENT	5	4	3	4	3	4	3	5
WIN LOSE	LIFE	3	3	3	2	3	3	2	2

Table 44: Mean error rates: all features (%)

model's overall capability of differentiating the event types.

The average increase in mean macro F1 score of 0.06 emphasizes that the classification of DISORDER is in general aided by the added features. This means that smaller errors that are connected to the event type might also be targeted. The analysis of the features' mean error rates reveals a reduction of the confusion of DISORDER with GOVERNMENT. With regard to the features named entities and sentiment, the improvements achieved are on average even greater than those achieved for the confusion of DISORDER and CRIME. As DISORDER and GOVERNMENT are inherently less similar to each other than DISORDER and CRIME, the addition of the features contributes more strongly to their discrimination.

The usage of models trained with additional features leads to distinctly higher mean F1 scores for both the small and the big test set for the event type FINANCE. The difference to the baseline is on average 0.06 for the small test set and 0.09 for the big test set. As was noted before, only a small amount of instances of the event type is included in the dataset. This makes its classification error-prone. The average increase achieved by the modified models illustrates that the FINANCE classification is supported by the added features. It can be argued that small classes benefit especially from explicitly added information as new and statistically relevant patterns are created. These patterns make them better separable from other classes in the corpus. This was already mentioned in chapter 7.2. The assumptions made regarding the improvements facilitated by the features therefore hold true.

The evaluation of the mean errors made by the modified models shows that especially the main error of confusing FINANCE texts with the event type GOVERNMENT is lessened by all features. Smaller errors connected to the classification of FINANCE are also reduced in parts. However, for some features, the assumption arises that errors might be redistributed, not avoided. This can be seen for the features named entities and sentiment, named entities and aspectual classes as well as aspectual classes, named entities and

sentiment. While the confusion with GOVERNMENT is less problematic, the confusion with ORGANISATION is strengthened in these cases. GOVERNMENT and ORGANISATION are semantically similar. Therefore, it is possible that texts that are no longer incorrectly classified as GOVERNMENT are incorrectly classified as ORGANISATION. Still, the general amount of errors connected to the event type is notably reduced by means of the added features.

The classification of the event type ORGANISATION is improved by 0.02 for both test sets. This makes its difference to the baseline smaller than the one seen for DISORDER and FINANCE. For the baseline models, the main errors connected to ORGANISATION were the confusions with the event types DIPLOMACY and GOVERNMENT. It was assumed that the named entity feature will be especially helpful for the discrimination of the event types. The consultation of the mean errors support this assumption. Models using features that incorporate information about named entities, i.e., named entities as a single feature as well as the combinations named entities and sentiment, named entities and aspectual classes and aspectual classes, named entities and sentiment, improve the errors connected to the wrong classification of DIPLOMACY the most.

Named entities also aid the distinction of ORGANISATION and GOVERNMENT. With regard to this error, improvements are similar for all features, except aspectual classes. For the aspectual class feature, the error rate does not change. ORGANISATION and GOVERNMENT show a similar distribution of aspectual classes in the data (cf. figure 25). The feature therefore does not contribute to more distinct representations of ORGANISATION and GOVERNMENT.

The improvement seen for WIN LOSE is also not as pronounced as the one observed for the event types DISORDER and FINANCE. The average difference between the baseline and modified models for the small test set is 0.02. This difference increases to 0.03 for the big test set. The main problem

with regard to the classification of WIN LOSE texts is the confusion with the event type DISCOVERY. Both event types include texts that deal with sports. This makes their textual descriptions similar. In this case, the overlap between event types seems to be more problematic than for other event types. This becomes obvious when considering the errors made by the modified models listed in table 44. For three features, the error rate does not change, two features increase the error rate and only two features show an improvement with regard to the error. Those are aspectual classes and sentiment as well as aspectual classes, named entities and sentiment. More distinct improvements were expected for this error, especially for the sentiment feature (cf. chapter 7.2). It can be stated that class discrimination remains difficult for the models. This is possibly due to the mentioned reports on sports related events which are included for both event types. The features are not suitable for introducing a means of discriminating these reports according to the correct event type. The higher mean macro F1 score that is observed for WIN LOSE is thus due to improvements of smaller error sources, i.e., the confusion with GOVERNMENT as well as LIFE.

As mentioned previously, GOVERNMENT constitutes the event type that was predicted instead of the correct event type for a number of classes. The improvement that is seen for GOVERNMENT is therefore due to improvements with regard to these confusions. Table 44 shows that a number of smaller errors involving GOVERNMENT are mitigated, e.g., the wrong prediction of GOVERNMENT for DISORDER and WIN LOSE texts. Additionally, the mitigation of two main errors, i.e. the wrong prediction of GOVERNMENT for FINANCE and ORGANISATION texts, also improves the overall score of the event type.

In summary, it was seen that the usage of additional linguistic features targets almost all of the main problems that were identified for the baseline models. Systematic confusions between semantically similar event types are mitigated by the given linguistic features. Additionally, the ability of the mod-

els to generalize to more diverse data is supported.

Specific assumptions were made regarding the individual features' performances (cf. chapter 7.2). It was found that most of these assumptions hold true. An exception was however found with regard to the main error connected to the classification of WIN LOSE texts. In this case, the features did not support classification in the way it was expected.

For smaller errors made by the baseline models, i.e., error rates between 2 and 3%, no notable improvements were observed. Examples of smaller errors are ACCIDENT classified as NATURAL DISASTER, CRIME classified as ACCIDENT, WIN LOSE classified as LIFE. It can be assumed that these errors are due to data points that diverge strongly from the norm. This means that individual texts are notably different to the majority of texts belonging to a specific event type. The linguistic features are supposed to strengthen the concise representation of the event types based on actual similarities that are present in the data but not exploited by the model. Therefore, they do not aid the classification of texts that include diverging patterns.

9 Conclusion

The general goal of the work at hand was the classification of texts according to their event types. A combination of transformer models and manually constructed linguistic features was proposed to reach this goal.

Chapter 2 described the evolution of text classification with a focus on available language representations. A trade-off between the interpretability of a model and manual effort that needs to be invested was observed: Classification decisions of models that use automatically learned features, such as transformer models, are opaque, however, no manual effort is required to enable classification when these kinds of features are used. The fine-tuning of transformer models has thus been established as the preferred approach for numerous NLP tasks.

Chapter 3 showed that this preference for transformer models also holds true for the task of event type classification: Event types are dominantly classified with the help of transformer models. Classification is based on the entire input text without supplying additional features to support the correct classification. Features are learned automatically.

Chapter 4 introduced the experimental set up. Next to implementation libraries and evaluation methods, the data that was used for the experiments was described. The annotation scheme of the corpus was simplified and texts were reduced to their essence, i.e., to sentences that include an event argument, for model training.

The preprocessing of the data resulted in two distinct ways of testing the trained models. On the one hand, test sets that also include texts summarized according to event arguments were drawn upon. On the other hand, a separate test set including texts that were not summarized was created. The usage of this test set had the aim of evaluating the models' ability to generalize to more complex data. The classification of this second test set was assumed to be more challenging as the essential event information was spread over longer

texts and other events than the annotated main event were recorded.

In chapter 5, two preliminary experiments were conducted. These experiments aimed at furthering the understanding of the probable textual features that are automatically determined and focused on by transformer models in order to solve the task of event type classification and the observable consequences for classification that arise from this automatic selection. The first experiment concentrated on analyzing the importance of individual word classes for classification. The results of this experiment indicated that when using transformer models and therefore automatically learned features, the most relevant word class for classification is nouns. This stands in contrast to existing definitions of events which state that event types are most distinctly determined by verbs. The second experiment evaluated the influence that the automatically learned features have with regard to model performance. To this end, five baseline models were trained and evaluated. The results revealed systematic confusions in the prediction of certain event types. The classification of texts belonging to semantically similar event types and event types for which only a few texts were available for training was particularly error-prone. It was hypothesized that the addition of explicit linguistic features to the database can improve the classification of event types. The considerations resulting from the preliminary experiments formed the basis for the rest of this work.

Chapter 6 gave an overview of the features that were chosen: aspectual classes, named entities and sentiment. Linguistic features were added to the training data automatically. Accordingly, models that were capable of annotating the texts were needed. Open-source models were used for named entity recognition and sentiment classification. Custom models were trained and evaluated for the classification of aspectual classes. Similar to the classification of event types, it was found that the focus of the classification model seems to diverge from linguistic definitions. In the case of aspectual class classification this means that sentences were dominantly categorized based

on only the main verb. Other linguistic indicators that change the aspectual class of a sentence, as for example temporal modifiers, were not always considered.

The main experiment was detailed in chapter 7. In total, 35 models were trained with the modified database that included the additional features. Specific assumptions were made about expected improvements connected to the individual features. The models using the linguistic features were compared to the baseline models with the help of classification reports, paired bootstrap testing and confusion matrices. For a selection of models a detailed error analysis was conducted. These analyses revealed differences regarding the suitability of the individual features for the task.

The detailed discussion of the results in chapter 8 showed that there is a general improvement in classification when linguistic information is added explicitly to the texts that are used for model training and testing. The greatest improvements were observed for texts belonging to event types that are represented by a small number of samples in the data and event types that are semantically similar to others. This means that the addition of features targeted the main problems of classification that were identified in the preliminary experiment. Additionally, the added information seemed to especially support the classification of more diverse data, i.e., the big test set. However, it was noted that sample size influenced the results of significance testing which means that the results observed for the big test set are in general more reliable than those seen for the small test set. It was also seen that the expected improvements stated in chapter 7.2 held true for almost all event types and features. An exception was seen for WIN LOSE. The improvements for this event type were due to the mitigation of smaller errors. The main error source was not improved reliably.

It was also found that the added features do not hinder the classification of event types that already display high scores for the baseline models. This suggested that the features were well chosen. To summarize, it was found

that constructing a small number of features manually and adding those to the database which is processed with the help of state of the art transformer models can improve classification still. Criticism that feature engineering for event type classification is outdated and time-consuming can therefore not be accepted fully.

It can be assumed that future research focused on event type classification will undergo another change. Generative approaches will gain traction. It remains to be seen if and how these approaches will refocus the task of event type classification. Determining event types with the help of prompts like: What happened? might for example force LLMs to consider verbs more strongly again.

References

Ahn, David. 2006. The stages of event extraction. In: *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. 1–8. Sydney, Australia. https://aclanthology.org/W06-0901/.

Alikhani, Malihe & Stone, Matthew. 2018. Arrows are the Verbs of Diagrams. In: *Proceedings of the 27th International Conference on Computational Linguistics*. 3552–3563. Santa Fe, New Mexico, USA. Association for Computational Linguistics. https://aclanthology.org/C18-1301/.

Bakker, Femke & Van Heusden, Ruben & Marx, Maarten. 2024. Timeline Extraction from Decision Letters Using ChatGPT. In: *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*. 24–31. St. Julians, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.case-1.3/.

Becker, Nils & Krumbiegel, Theresa. 2021. FKIE_itf_2021 at CASE 2021 Task 1: Using Small Densely Fully Connected Neural Nets for Event Detection and Clustering. In: *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. 113–119. Online. Association for Computational Linguistics. https://aclanthology.org/2021.case-1.15/.

Berg-Kirkpatrick, Taylor & Burkett, David & Klein, Dan. 2012. An Empirical Investigation of Statistical Significance in NLP. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 995–1005. Jeju Island, Korea. Association for Computational Linguistics. https://aclanthology.org/D12-1091/.

Boutaleb, Allaa & Picault, Jerome & Grosjean, Guillaume. 2024. BERTrend: Neural Topic Modeling for Emerging Trends Detection. In: *Proceedings of the Workshop on the Future of Event Detection (FuturED)*. 1–17. Miami, Florida, USA. Association for Computational Linguistics. https://aclanthology.org/2024.futured-1.1/.

Brent, Michael R.. 1991. Automatic Semantic Classification of Verbs From Their Syntactic Contexts: An Implemented Classifier for Stativity. In: *Fifth Conference of the European Chapter of the Association for Computational Linguistics*. Berlin, Germany. Association for Computational Linguistics. https://aclanthology.org/E91-1039/.

Carlson, Greg. 1977a. *Reference to Kinds in English*. PhD thesis. University of Massachusetts, Amherst.

Carlson, Greg. 1977b. A unified analysis of the English bare plural. In: *Linguistics and Philosophy* 1. 413–456.

Chen, Yubo & Xu, Liheng & Liu, Kang & Zeng, Daojian & Zhao, Jun. 2015. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 167–176. Beijing, China. Association for Computational Linguistics. https://aclanthology.org/P15-1017/.

Comrie, Bernard. 1976. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: Cambridge University Press.

Davani, Aida Mostafazadeh & Yeh, Leigh & Atari, Mohammad & Kennedy, Brendan & Portillo Wightman, Gwenyth & Gonzalez, Elaine & Delong, Natalie & Bhatia, Rhea & Mirinjian, Arineh & Ren, Xiang & Dehghani, Morteza. 2019. Reporting the Unreported: Event Extraction for Analyzing the Local Representation of Hate Crimes. 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)*. 5753–5757. Hong Kong, China. Association for Computational Linguistics. https://aclanthology.org/D19-1580/.

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *CoRR* abs/1810.04805. http://arxiv.org/abs/1810.04805.

Dowty, David R. 1977. Toward a Semantic Analysis of Verb Aspect and the English 'Imperfective' Progressive. In: *Linguistics and Philosophy* 1. 45–77.

Dowty, David R. 1979. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Dordrecht: Springer Netherlands.

Efron, Bradley. 1979. Bootstrap Methods: Another Look at the Jackknife. In: *The Annals of Statistics* 7 (1). 1–26.

Efron, Bradley & Tibshirani, Robert J. 1993. *An introduction to the bootstrap*. London: Chapman and Hall.

Farsi, Salman & Eusha, Asrarul Hoque & Arefin, Mohammad Shamsul. 2024. CUET_Binary_Hackers at ClimateActivism 2024: A Comprehensive Evaluation and Superior Performance of Transformer-Based Models in Hate Speech Event Detection and Stance Classification for Climate Activism. In: *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*. 145–155. St. Julians, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.case-1.20/.

Fellman, Evan & Tyo, Jacob & Lipton, Zachary. 2024. The Future of Web Data Mining: Insights from Multimodal and Code-based Extraction Methods. In: *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE*

2024). 1–5. St. Julians, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.case-1.1/.

Feng, Xiaocheng & Huang, Lifu & Tang, Duyu & Ji, Heng & Qin, Bing & Liu, Ting. 2016. A Language-Independent Neural Network for Event Detection. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 66–71. Berlin, Germany. Association for Computational Linguistics. https://aclanthology.org/P16-2011/.

Filip, Hana. 2011. Aspectual class and Aktionsart. In: Heusinger, Klaus von & Maienborn, Claudia & Portner, Paul (Ed.), *Semantics (HSK 33.1)*. 1186–1217. De Gruyter Mouton. Berlin, Boston. https://doi.org/10.1515/9783110255072.1186.

Firth, John R.. 1962. A synopsis of linguistic theory, 1930–1955. In: *Studies in Linguistic Analysis*. Blackwell. Oxford.

Friedrich, Annemarie & Gateva, Damyana. 2017. Classification of telicity using cross-linguistic annotation projection. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2559–2565. Copenhagen, Denmark. Association for Computational Linguistics. https://aclanthology.org/D17-1271/.

Friedrich, Annemarie & Palmer, Alexis & Pinkal, Manfred. 2016. Situation entity types: automatic classification of clause-level aspect. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1757–1768. Berlin, Germany. Association for Computational Linguistics. https://aclanthology.org/P16-1166/.

Goldberg, Yoav. 2017. *Neural Network Methods for Natural Language Processing*. San Rafael: Morgan & Claypool. https://32931414.s21i.faiusr.com/61/ABUIABA9GAAgxLqCuwYojOPTXQ.pdf.

Gómez-de-Mariscal, Estibaliz & Guerrero, Vanesa & Sneider, Alexandra & Jayatilaka, Hasini & Phillip, Jude M. & Wirtz, Denis & Muñoz-Barrutia, Arrate. 2021. Use of the p-values as a size-dependent function to address practical differences when analyzing large datasets. In: *Scientific Reports* 11. https://doi.org/10.1038/s41598-021-00199-5.

Hamborg, Felix & Donnay, Karsten. 2021. NewsMTSC: A Dataset for (Multi -)Target-dependent Sentiment Classification in Political News Articles. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1663–1675. Online. Association for Computational Linguistics. https://aclanthology.org/2021.eacl-main.142/.

Hermes, Jürgen & Richter, Michael & Neuefeind, Claes. 2015. Automatic Induction of German Aspectual Verb Classes in a Distributional Framework. In: *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*. 122–129.

Duisburg-Essen, Germany. German Society for Computational Linguistics and Language Technology.

Honnibal, Matthew & Montani, Ines & Van Landeghem, Sofie & Boyd, Adriane. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Huang, Ting-Hao Kenneth & Ferraro, Francis & Mostafazadeh, Nasrin & Misra, Ishan & Agrawal, Aishwarya & Devlin, Jacob & Girshick, Ross & He, Xiaodong & Kohli, Pushmeet & Batra, Dhruv & Zitnick, C. Lawrence & Parikh, Devi & Vanderwende, Lucy & Galley, Michel & Mitchell, Margaret. 2016. Visual Storytelling. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1233–1239. San Diego, California. Association for Computational Linguistics. https://aclanthology.org/N16-1147/.

Hürriyetoğlu, Ali & Tanev, Hristo & Thapa, Surendrabikram & Uludoğan, Gökçe (Ed.). 2024a. *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*. St. Julians, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.case-1.0/.

Hürriyetoğlu, Ali & Thapa, Surendrabikram & Uludoğan, Gökçe & Dehghan, Somaiyeh & Tanev, Hristo. 2024b. A Concise Report of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text. In: *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*. 248–255. St. Julians, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.case-1.34/.

Hutto, C.. 2014. vaderSentiment. Github. https://github.com/cjhutto/vaderSentiment.

Hutto, C. & Gilbert, Eric. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1). 216–225. 10.1609/icwsm.v8i1.14550. https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Ide, Nancy & Suderman, Keith. 2004. The American National Corpus First Release. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal. European Language Resources Association (ELRA). https://aclanthology.org/L04-1313/.

Jurafsky, Daniel & Martin, James H. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition draft. https://web.stanford.edu/~jurafsky/slp3/.

Kenny, Anthony. 1963. *Action, Emotion and Will*. London: Routledge.

Kent, Samantha & Krumbiegel, Theresa. 2021. CASE 2021 Task 2 Socio-political Fine-grained Event Classification using Fine-tuned RoBERTa Document Embeddings. In: *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. 208–217. Online. Association for Computational Linguistics. https://aclanthology.org/2021.case-1.26/.

Klavans, Judith L. & Chodorow, Martin. 1992. Degrees of Stativity: The Lexical Representation of Verb Aspect. In: *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*. Nantes France. https://aclanthology.org/C92-4177/.

Kranich, Svenja. 2010. *The Progressive in Modern English: A Corpus-Based Study of Grammaticalization and Related Changes*. Amsterdam: Rodopi.

Krumbiegel, Theresa & Decher, Sophie. 2022. NLP4ITF @ Causal News Corpus 2022: Leveraging Linguistic Information for Event Causality Classification. In: *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*. 16–20. Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://aclanthology.org/2022.case-1.3/.

Krumbiegel, Theresa & Pritzkau, Albert & Schmitz, Hans-Christian. 2021. Event Extraction and Discourse Monitoring. In: *Proceedings of the Workshop on Humanities-Centred Artificial Intelligence co-located with 44th German Conference on Artificial Intelligence*. Berlin, Germany. CEUR. https://ceur-ws.org/Vol-3093/paper3.pdf.

Lakoff, George. 1966. *On the nature of syntactic irregularity*. PhD thesis. Indiana University.

Lehman-Wilzig, Sam & Seletzky, Michal. 2010. Hard news, soft news, 'general' news: The necessity and utility of an intermediate classification. In: *Journalism* 11. 37–56. 10.1177/1464884909350642.

Li, Qian & Li, Jianxin & Sheng, Jiawei & Cui, Shiyao & Wu, Jia & Hei, Yiming & Peng, Hao & Guo, Shu & Wang, Lihong & Beheshti, Amin & Yu, Philip S. 2021a. A Survey on Deep Learning Event Extraction: Approaches and Applications. In: *IEEE Transactions on Neural Networks and Learning Systems* 35. 6301–6321. https://api.semanticscholar.org/CorpusID:253063434.

Li, Sha & Ji, Heng & Han, Jiawei. 2021b. Document-Level Event Argument Extraction by Conditional Generation. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 894–908. Online. Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.69/.

Liebman, Sam. 2018. Mapping Word Embeddings with Word2vec. https://medium.com/data-science/mapping-word-embeddings-with-word2vec-99a799dc9695.

Lin, Tsung-Yi & Maire, Michael & Belongie, Serge & Bourdev, Lubomir & Girshick, Ross & Hays, James & Perona, Pietro & Ramanan, Deva & Zitnick, C. Lawrence & Dollár, Piotr. 2015. Microsoft COCO: Common Objects in Context. https://arxiv.org/abs/1405.0312.

Linguistic Data Consortium. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*. https://www.ldc.upenn.edu/Projects/ACE/.

Liu, Shulin & Chen, Yubo & He, Shizhu & Liu, Kang & Zhao, Jun. 2016. Leveraging FrameNet to Improve Automatic Event Detection. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2134–2143. Berlin, Germany. Association for Computational Linguistics. https://aclanthology.org/P16-1201/.

Liu, Yinhan & Ott, Myle & Goyal, Naman & Du, Jingfei & Joshi, Mandar & Chen, Danqi & Levy, Omer & Lewis, Mike & Zettlemoyer, Luke & Stoyanov, Veselin. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: *CoRR* abs/1907.11692. http://arxiv.org/abs/1907.11692.

Loerakker, Meagan & Müter, Laurens & Schraagen, Marijn. 2024. Fine-Tuning Language Models on Dutch Protest Event Tweets. In: *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*. 6–23. St. Julians, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.case-1.2/.

Marneffe, Marie-Catherine de & Manning, Christopher D. & Nivre, Joakim & Zeman, Daniel. 2021. Universal Dependencies. In: *Computational Linguistics* 47 (2). 255–308. 10.1162/coli_a_00402. https://aclanthology.org/2021.cl-2.11/.

Maron, M. E. 1961. Automatic Indexing: An Experimental Inquiry. In: *J. ACM* 8 (3). 404–417. ISSN 0004-5411. https://doi.org/10.1145/321075.321084.

McKinley, David. 2021. All the News Dataset. https://www.kaggle.com/datasets/davidmckinley/all-the-news-dataset.

Metheniti, Eleni. 2023. *What do you know, BERT? Exploring the linguistic competencies of Transformer- based contextual word embeddings.* PhD thesis. Université Toulouse le Mirail - Toulouse II.

Metheniti, Eleni & Van De Cruys, Tim & Hathout, Nabil. 2022. About Time: Do Transformers Learn Temporal Verbal Aspect? In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. 88–101. Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/2022.cmcl-1.10/.

Mikolov, Tomas & Chen, Kai & Corrado, Greg & Dean, Jeffrey. 2013a. Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781.

Mikolov, Tomas & Sutskever, Ilya & Chen, Kai & Corrado, Greg & Dean, Jeffrey. 2013b. Distributed Representations of Words and Phrases and their Compositionality. https://arxiv.org/abs/1310.4546.

Mittwoch, Anita. 2019. Aspectual Classes. In: Truswell, Robert (Ed.), *The Oxford Handbook of Event Structure*. Oxford University Press. Oxford.

Moens, Marc & Steedman, Mark. 1988. Temporal Ontology and Temporal Reference. In: *Computational Linguistics* 14 (2). 15–28. https://aclanthology.org/J88-2003/.

Nandwani, Pansy & Verma, Rupali. 2021. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In: *Social network analysis and mining* 11. 10.1007/s13278-021-00776-6.

Nguyen, Thien Huu & Grishman, Ralph. 2016. Modeling Skip-Grams for Event Detection with Convolutional Neural Networks. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 886–891. Austin, Texas. Association for Computational Linguistics. https://aclanthology.org/D16-1085/.

Pedregosa, Fabian & Varoquaux, Gaël & Gramfort, Alexandre & Michel, Vincent & Thirion, Bertrand & Grisel, Olivier & Blondel, Mathieu & Müller, Andreas & Nothman, Joel & Louppe, Gilles & Prettenhofer, Peter & Weiss, Ron & Dubourg, Vincent & Vanderplas, Jake & Passos, Alexandre & Cournapeau, David & Brucher, Matthieu & Perrot, Matthieu & Duchesnay, Édouard. 2011. Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12. 2825–2830.

Petrovic, Sasa & Osborne, Miles & McCreadie, Richard & Macdonald, Craig & Ounis, Iadh & Shrimpton, Luke. 2021. Can Twitter Replace Newswire for Breaking News? In: *Proceedings of the International AAAI Conference on Web and Social Media* 7 (1). 713–716. https://ojs.aaai.org/index.php/ICWSM/article/view/14450.

Pritzkau, Albert. 2023. Investigating Propaganda Considering the Discursive Context of Utterances. In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*. Jaén, Spain. CEUR. https://ceur-ws.org/Vol-3496/.

Pritzkau, Albert. 2024. *Interaktive (visuelle) Supervision für Modelle des Maschinellen Lernens zur Analyse und Bewertung von Informationskampagnen*. PhD thesis. Universität Leipzig.

Qiu, Jinhui. 2024. An Analysis of Model Evaluation with Cross-Validation: Techniques, Applications, and Recent Advances. In: *Advances in Economics, Management and Political Sciences* 99. 69–72. 10.54254/2754-1169/99/2024OX0213.

Radford, Alec & Narasimhan, Karthik & Salimans, Tim & Sutskever, Ilya. 2018. Improving Language Understanding by Generative Pre-Training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Rajapakse, Thilina C. & Yates, Andrew & Rijke, Maarten de. 2024. Simple Transformers: Open-source for All. In: *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. SIGIR-AP 2024. 209–215. New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3673791.3698412.

Raleigh, Clionadh & Linke, Andrew & Hegre, Håvard & Karlsen, Joakim. 2010. Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature. In: *Journal of Peace Research* 47 (5). 651–660. 10.1177/0022343310378914. https://doi.org/10.1177/0022343310378914.

Ray, Marjolaine & Wang, Qi & Mélanie-Becquet, Frédérique & Poibeau, Thierry & Mazoyer, Béatrice. 2024. An Incremental Clustering Baseline for Event Detection on Twitter. In: *Proceedings of the Workshop on the Future of Event Detection (FuturED)*. 18–24. Miami, Florida, USA. Association for Computational Linguistics. https://aclanthology.org/2024.futured-1.2/.

Řehůřek, Radim & Sojka, Petr. 2010. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45–50. Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Ribeiro, Marco Tulio & Singh, Sameer & Guestrin, Carlos. 2016a. Local Interpretable Model-Agnostic Explanations (LIME): An Introduction. A technique to explain the predictions of any machine learning classifier. https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/.

Ribeiro, Marco Tulio & Singh, Sameer & Guestrin, Carlos. 2016b. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. 1135–1144. New York, NY, USA. Association for Computing Machinery. ISBN 9781450342322. https://doi.org/10.1145/2939672.2939778.

Ruppenhofer, Josef & Ellsworth, Michael & Petruck, Miriam R. L. & Johnson, Christopher R. & Scheffczyk, Jan. 2010. *FrameNet II: Extended Theory and Practice*.

Sha, Lei & Qian, Feng & Chang, Baobao & Sui, Zhifang. 2018. Jointly Extracting Event Triggers and Arguments by Dependency-Bridge RNN and Tensor-Based Argument Interaction. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1). 10.1609/aaai.v32i1.12034. https://ojs.aaai.org/index.php/AAAI/article/view/12034.

Sharma, Piyush & Ding, Nan & Goodman, Sebastian & Soricut, Radu. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2556–2565. Melbourne, Australia. Association for Computational Linguistics. https://aclanthology.org/P18-1238.pdf.

Siegel, Eric V. & McKeown, Kathleen R.. 1996. Gathering statistics to aspectually classify sentences with a genetic algorithm. In: *Proceedings of the Second International Conference on New Methods in Language Processing*.

Smith, Carlota S. 1997. *The Parameter of Aspect*. 2nd edition. Dordrecht: Kluwer Academic Publishers.

Stone, M. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 36. 111–147. http://www.jstor.org/stable/2984809.

Sun, Chi & Qiu, Xipeng & Xu, Yige & Huang, Xuanjing. 2019. How to Fine-Tune BERT for Text Classification? In: *CoRR* abs/1905.05583. http://arxiv.org/abs/1905.05583.

Supriyono & Wibawa, Aji Prasetya & Suyono & Kurniawan, Fachrul. 2024. Advancements in natural language processing: Implications, challenges, and future directions. In: *Telematics and Informatics Reports* 16. ISSN 2772-5030. https://doi.org/10.1016/j.teler.2024.100173. https://www.sciencedirect.com/science/article/pii/S2772503024000598.

Tanev, Hristo. 2024. Leveraging Approximate Pattern Matching with BERT for Event Detection. In: *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*. 32–39. St. Julians, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.case-1.4/.

Tetreault, Joel & Nguyen, Thien Huu & Lamba, Hemank & Hughes, Amanda (Ed.). 2024. *Proceedings of the Workshop on the Future of Event Detection (FuturED)*. Miami, Florida, USA. Association for Computational Linguistics. https://aclanthology.org/2024.futured-1.0/.

Tong, MeiHan & Xu, Bin & Wang, Shuai & Han, Meihuan & Cao, Yixin & Zhu, Jiangqi & Chen, Siyu & Hou, Lei & Li, Juanzi. 2022. DocEE: A Large-Scale and Fine-grained Benchmark for Document-level Event Extraction. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3970–3982. Seattle, United States. Association for Computational Linguistics. https://aclanthology.org/2022.naacl-main.291/.

Tunstall, Lewis & Werra, Leandro von & Wolf, Thomas. 2022. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. Sebastopol: O'Reilly Media, Incorporated.

Uludoğan, Gökçe & Yüksel, Atıf Emre & Tunçer, Ümit & Işık, Burak & Korkmaz, Yasemin & Akar, Didar & Özgür, Arzucan. 2024. Detecting Hate Speech in Turkish Print Media: A Corpus and A Hybrid Approach with Target-oriented Linguistic Knowledge. In: *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*. 205–214. St. Julians, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.case-1.29/.

Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan N. & Kaiser, Łukasz & Polosukhin, Illia. 2017. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010. Red Hook, NY, USA. Curran Associates Inc.

Vendler, Zeno. 1957. Verbs and times. In: *Philosophical Review* 66. 143–160.

Wadden, David & Wennberg, Ulme & Luan, Yi & Hajishirzi, Hannaneh. 2019. Entity, Relation, and Event Extraction with Contextualized Span Representations. 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)*. 5784–5789. Hong Kong, China. Association for Computational Linguistics. https://aclanthology.org/D19-1585/.

Wang, Tianlu & Sridhar, Rohit & Yang, Diyi & Wang, Xuezhi. 2022. Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models. In: *Findings of the Association for Computational Linguistics: NAACL 2022*. 1719–1729. Seattle, United States. Association for Computational Linguistics. https://aclanthology.org/2022.findings-naacl.130/.

Wang, Yixin & Jordan, Michael I. 2021. Desiderata for Representation Learning: A Causal Perspective. In: *ArXiv* abs/2109.03795. https://api.semanticscholar.org/CorpusID:237440146.

Yagcioglu, Semih & Erdem, Aykut & Erdem, Erkut & Ikizler-Cinbis, Nazli. 2018. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1358–1368. Brussels, Belgium. Association for Computational Linguistics. https://aclanthology.org/D18-1166/.

Yang, Sen & Feng, Dawei & Qiao, Linbo & Kan, Zhigang & Li, Dongsheng. 2019. Exploring Pre-trained Language Models for Event Extraction and Generation. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5284–5294. Florence, Italy. Association for Computational Linguistics. https://aclanthology.org/P19-1522/.

Young, Peter & Lai, Alice & Hodosh, Micah & Hockenmaier, Julia. 2014. From image descriptions to visual denotations: New similarity metrics for

semantic inference over event descriptions. In: *Transactions of the Association for Computational Linguistics* 2. 67–78. 10.1162/tacl_a_00166. https://aclanthology.org/Q14-1006/.

A Appendix – Event Arguments

Event type	Event Arguments
Air Crash	Crew Service Years Flight No. Aircraft Agency Cause Alternate Landing Place Casualties and Losses Passengers Date Survivors Taking-off Place Location Accident Investigator Scheduled Landing Place
Appoint_Inauguration	Appointee Appointment Time Employment Agency Last Job of the Appointee Position Inauguration Time Predecessor Appointer Age of the Appointee Term of Office
Armed Conflict	End Time Conflict Duration Weapon and Equipment Damaged Facility Casualties and Losses Attacker Date Military Strength Target Location
Awards Ceremony	Winner Employed Institution Award Reason Host Award Date Location Award Field

Bank Robbery	Weapon Used Arrested Hostage Investigating agency Transportation Perpetrators Bank Name Date Recovered Amount Investigating Spokesperson Location Stolen Amount
Break Historical Records	The Grades of the Previous Record Holder Last Time the Record was Broken Grades Record-breaking Project Record Breaker Date Location Previous Record Holder
CommitCrime – Accuse	Evidence Prosecution Lawyer Defense Lawyer Time of the Case Witness Prosecutor Charged Crime Accused People
CommitCrime – Arrest	Criminal Evidence Police Arrest Time Suspect The Charged Crime Arrest Location
CommitCrime – Investigate	Person under Investigation Investigative Agency Date Cause Head of Investigation Team
CommitCrime – Release	Judge Release Time Prison Term Defense Lawyer Jail Time Released People Release Reason Sentencing Location Charged Crime

CommitCrime – Sentence	Judge Detention Start Time Judgment Result/Prison Term Prosecution Lawyer Release Time Defense Lawyer Victim Court The Sentence Claimed by the Defense Lawyer Suspect Court Time Prison Accusation The Sentence Claimed by the Prosecutor's Lawyer
Diplomatic Talks	Participants Summit Name Summit Theme Date Achievement Location
Diplomatic Visit	Visitor Host Date Country Visited
Disease Outbreaks	Number of People Hospitalized Outbreak Location Way for Spreading Disease Number of Vaccinated People Cured Rate Special Medicine Areas affected Cause Symptoms Cured Cases Suspected Cases Precautionary Measure Date Epidemic Data Issuing Agency Death Rate Complications Susceptible Population Vaccine Research and Development Organisation Death Cases Confirmed/Infected Cases

Droughts	Economic Loss Damaged Crops & Livestock Related Rivers or Lakes Production Cuts Areas Affected Date The Worst-Hit Area Cause Influenced People
Earthquakes	Number of Rebuilding House Number of Destroyed Building Economic Loss Epicenter Number of Trapped People Number of Aftershocks Aid Agency Number of Evacuated People Casualties and Losses Affected Area Temporary Settlement Date Number of Rescued People Magnitude Aid Supplies/Amount
Election	Voting method Election Goal The Final Seats Result Turnout People Casting Key Votes Candidate Scandal (Allegations of fraud, etc.) Candidates and their Political Parties Date The Final Votes and Percentages Electoral System Location Election Name
Environment Pollution	Solution Cause Number of Victims Pollution Source Party Responsible for Pollution Date Anti-pollution People/Organisations Location

Famine	Solution Economic Loss Affected Areas Aid Agency Date Casualties and Losses Number of Influenced People Cause Aid Supplies/Amount
Famous Person – Death	State Before Death Location/Hospital Age Deceased Date Doctor Profession Death Reason Perpetrator
Famous Person – Divorce	Child Husband Marriage Duration Child Custody Property Division Date Cause Wife Announce Platform
Famous Person – Give a Speech	Speaker News Release Agency Date Speaker Status Location Ways to Watch the Speech
Famous Person – Marriage	Cost Husband Invited Person Wedding Dress Designer How Many Times Get Married Date Wedding Venue Witness Wife
Famous Person – Recovered	Disease Cost Location/Hospital Sequelae Doctor and Medical Team Date People Treatment Method

Famous Person – Sick	Symptom Location/Hospital Illness Doctor and Medical Team Date People Cause
Financial Aid	Aid Reason Beneficiary Date Sponsor Location Funding
Financial Crisis	Policy Proposals Economic Loss Affected Industries Economists who Predicted the Crisis Unemployed Rate Affected Area Bankrupt Business Start Date End Date Cause
Fire	Economic Loss Aid Agency Number of Damaged Houses Casualties and Losses Temporary Settlement Date Location Number of Rescued People Magnitude Cause Aid Supplies/Amount
Floods	Economic Loss Disaster-Stricken Farmland Water Level Aid Agency Affected Areas Casualties and Losses Related Rivers or Lakes Temporary Settlement Missing Date Number of Damaged Houses Number of Rescued People Maximum Rainfall Number of Evacuated People Cause Aid Supplies/Amount

Gas Explosion	Compensation Economic loss Casualties and Losses Date Location Attending Hospital Survivors Cause Accident Investigator
Government Policy Changes	Announcement Date Cause Effective Date Policy Content Invalid Date Deliberating Agency Policy Name & Abbreviation Bill Drafting Agency Location
Insect Disaster	Influenced Crops and Livelihood Economic Loss Response Measures Aid Agency Affected Areas Date Pests Cause Aid Supplies/Amount
Join in an Organisation	Organisation Leader Announcement Date Cause Effective Date Join Conditions Organisation Members Organisation Name Declarer Countries Joined the Organisation
Mass Poisoning	Compensation Economic Loss Casualties and Losses Date Poisoning Type Location Attending Hospital Cause Accident Investigator

Military Exercise	Participating Countries Goal Commanders and their Position Weapon and Equipment Military Exercise Scale Army Date Location Projects
Mine Collapses	Economic Loss Compensation Number of Trapped People Rescue Start Time Trapped Days Cause Casualties and Losses Trapped Depth Date Attending Hospital Survivors Location Accident Investigator
Mudslides	Number of Destroyed Building Economic Loss Aid Agency Affected Areas Casualties and Losses Influenced People Missing Temporary Settlement Date Number of Rescued People Number of Evacuated People Cause Aid Supplies/Amount
New Achievements in Aerospace	Astronauts Carrier Rocket Launch Date Launch Country Spokeswoman/Spokesman Spacecraft Mission Research Agency Mission Duration Launch Result Cooperative Agency Launch Site Spacecraft

New Archaeological Discoveries	Reasons for the Formation of the Historical sites Archaeologist Discover Location Discover Time Artifacts and Their Chronology Archaeologist Organisation Historical Sites
New Wonders in Nature	Spectacle Start Time Spectacle Location Types of the Spectacle Best Way to Shoot Live Broadcast Platform Forecasting Agency Spectacle Duration Spectacle End Time Cause
Organisation Closed	Head of the Institution Organisation Date Location Cause
Organisation Established	Organisation Registered Capital Date Head of Institution Spokesmen Location Cause
Organisation Fine	Fine Reason Lawyer Fined Agency Penalty Amount Date Location Regulatory Authority
Organisation Merge	Merger Terms Head of the Merged Organisation Organisation Industry Date Acquirer Acquisition amount Acquiree

Protest	Government Reaction Method Arrested Protest Reason Damaged Property Casualties and Losses Date Location Protest Slogan Protesters
Regime Change	Time for Dignitaries to Resign Commanders of the Army Country Army Date Head of the Government Provisional Government
Resignation_Dismissal	Age of the Person that Resigned Employment Agency Resign Reason Successor Approver Position Date Person that Resigned Term of Office
Riot	Economic Loss Riot Reason Casualties and Losses Weapon Belligerents Date Location
Road Crash	Compensation Number of Vehicles Involved in the Crash Economic loss Casualties and Losses Date Location Attending Hospital Survivors Cause Accident Investigator

Shipwreck	Shipwreck Reason State of the Hull Hull Location Rescue Organiser Rescue Start Time Hull Discovery Time Casualties and Losses Rescue Tool or Method Missing Date Ship Agency Lost Contact Time Survivors Ship No. Location Accident Investigator
Sign Agreement	Contracting Parties Agreement Validity Period Agreement Name Date Agreement Content Location
Sports Competition	Score End Time Lasting Time Host Country Contest Participant MVP Start Time Competition Items Game Name Champions Postpone Time Location Postpone Reason

Storm	Storm Formation Location People/Organisation who predicted the disaster Amount of Precipitation Storm Hit Location Influence People Storm Direction Storm Hit Time Storm Formation Time Storm Center Location Storm Warning Level Storm Movement Speed Storm Name Maximum Wind Speed
Strike	End Date Economy Loss Start Date Strike Industry Duration Strike Agency Strikers Strike Reason Strikers Status Strike Outcome Boycotted Institutions
Tear Up Agreement	Agreement Members Agreement Name Date Tear Up Reason The Agency Who Broke the Agreement
Train Collisions	Economic Loss Cause Train No. Casualties and Losses Missing Date Attending Hospital Survivors Location Accident Investigator Train Agency

Tsunamis	Tsunami Warning Level Number of Destroyed Building Economic Loss Warning Device Magnitude (Tsunami heights) People/Organisation who predicted the disaster Aid Agency Casualties and Losses Date Aid Supplies/Amount Number of Rescued People Tsunamis Cause Area Affected
Volcano Eruption	Outbreak Date Economic Loss Refuge Fire Warning Level Number of Damaged House People/Organisation who predicted the disaster Affected Areas Number of Evacuated People Casualties and Losses Volcano Name Last Outbreak Time Cause The State of the Volcano (Dormant or Active)
Withdraw from an Organisation	Organisation Leader Countries Withdrawing from the Organisation Announcement Date Exit Conditions Effective Date Organisation Members Withdraw Reason Organisation Name Declarer