ModaFact: Multi-paradigm Evaluation for Joint Event Modality and Factuality Detection

Marco Rovera, Serena Cristoforetti, Sara Tonelli
Fondazione Bruno Kessler, Trento, Italy
{m.rovera,satonelli}@fbk.eu

Abstract

Factuality and modality are two crucial aspects concerning events, since they convey the speaker's commitment to a situation in discourse as well as *how* this event is supposed to occur in terms of norms, wishes, necessity, duty and so on. Capturing them both is necessary to truly understand an utterance meaning and the speaker's perspective with respect to a mentioned event. Yet, NLP studies have mostly dealt with these two aspects separately, mainly devoting past efforts to the development of English datasets. In this work, we propose ModaFact, a novel resource with joint factuality and modality information for event-denoting expressions in Italian. We propose a novel annotation scheme, which however is consistent with existing ones, and compare different classification systems trained on ModaFact, as a preliminary step to the use of factuality and modality information in downstream tasks. The dataset and the best-performing model are publicly released and available under an open license.

1 Introduction

Event *factuality* is the level of commitment a speaker assigns to a situation in discourse. On the other hand, linguistic *modality* conveys the relationship a situation is supposed to have with respect to wishes, norms, goals, authority, etc. Together, factuality and modality play a substantial role in determining the actual meaning of an utterance, thereby establishing different views and degrees of commitment about the occurrence of events in the world. Consider for example the following sentences:

> (a) La nave, con 65 migranti a bordo, è approdata in porto. (*The ship, with 65 migrants on board, landed in harbor).*
>
> (b) La nave, con 65 migranti a bordo, è dovuta approdare in porto. (*The ship, with 65 migrants on board, had to land in harbor).*
>
> (c) Il comandante della nave, con 65 migranti a bordo, ha dichiarato l'intenzione di approdare in porto.
> (*The captain of the ship, with 65 migrants on board, declared his intention to land in harbor).*

While sentence (a) merely presents the occurrence of an event, sentence (b) expresses the event as originating from a *necessity* and sentence (c) as originating from a *will*. This is what we call modality. Also, while the speaker in (a) and (b) shows to be sure that the event happened, this does not hold in (c). This defines the factuality of the event.

In public discourse, especially in social media, the arising and consolidation of certain beliefs or stances with respect to events is crucial, both for the emergence of opinions and for the effects they might produce on the behaviour of individuals. This aspect could be accurately captured by jointly modelling factuality and modality, yet little attention has been devoted to analysing their role in public discourse and most existing NLP works have focused on modeling and detecting them separately. Furthermore, several recent works in the disinformation area have used the term *factuality* as a synonym of "factual correctness with respect to world knowledge" (Devaraj et al., 2022; Chen et al., 2023; Feng et al., 2023; Augenstein et al., 2024), overshadowing the long-established concept of factuality in linguistic terms, which regulates the position conveyed by a speaker w.r.t. the occurrence of a certain event without making any assumption on world knowledge. While it is important to detect the veracity of an utterance to ensure the circulation of high-quality information, we argue that studying factuality and modality as we do in this work is equally important also in terms of impact, as it allows capturing *how* a speaker describes a certain

Proceedings of the 31st International Conference on Computational Linguistics, pages 6378–6396
January 19–24, 2025.

event. For example, in tasks like check-worthy claim detection (Nakov et al., 2021), the presence of certainty cues strongly affects whether a claim presenting an event is likely to be considered true by the readers, thus needing to be prioritized for fact-checking. redFurthermore, recent works have shown that LLMs are sensitive to to changes in the modality of the prompts (Leidinger et al., 2023) as well as to shifts in factuality and modality in QA tasks (Zhou et al., 2023), with more confident prompts leading to worse generation.

In this work, we therefore focus on the development of a novel dataset, ModaFact, where factuality and modality are jointly annotated. The resource is in Italian, a language for which only one dataset for *factuality* detection was created before (Minard et al., 2014). We then use it to evaluate different approaches for factuality and modality detection. The main contributions of this work can thus be outlined as follows:

(i) we develop and release the novel ModaFact dataset, covering jointly for the first time factuality and modality; the resource contains over 10,000 event mentions, manually annotated according to a carefully designed scheme, while providing back-compatibility with established schemes;

(ii) we use ModaFact to evaluate different language models and three learning paradigms (Masked Language Models, Sequence-to-Sequence and Causal Language Models) to perform joint factuality and modality detection.

We argue that factuality and modality are not only interesting phenomena from a linguistic point of view, but may play a relevant role in downstream tasks like fact-checking (Yao et al., 2021) and language data analysis (Prieto et al., 2020).

The dataset[1] and the best-performing fine-tuned model[2] for joint modality and factuality detection (see Section 5) are available for download from our Huggingface repository. Also, instructions on how to use the model for inference can be found on GitHub[3].

[1] https://huggingface.co/datasets/dhfbk/modafact-ita

[2] https://huggingface.co/dhfbk/modafact-ita

[3] https://github.com/dhfbk/ModaFact

2 Related Work

2.1 Terminology and Positioning

Traditionally, factuality has been a subject of study in linguistics and refers to the degree of commitment a speaker makes with respect to an event or, in other words, to the level of (un)certainty *in a linguistic utterance*. Recently, however, the term entered the scientific debate with regard to disinformation, in relation to the ability of LLM to generate more or less *true* contents (Devaraj et al., 2022; Chen et al., 2023; Feng et al., 2023), thus indicating the situation when "the generated text is not factually correct with respect to world knowledge" (Augenstein et al., 2024). In this work, we will refer to the former interpretation of the term.

2.2 Factuality and Modality Annotation

Factuality. In the literature, a certain level of agreement has been reached in defining the analysis of factuality as being based on three axes: *certainty*, *polarity* and *time*. The first systematic effort to create an annotated dataset for factuality in English can be referred to FactBank (Saurí and Pustejovsky, 2009), where the authors proposed to evaluate factuality over two dimensions of *certainty* and *polarity*. The *time* dimension was implicit in this case, as FactBank builds upon TimeBank (Pustejovsky et al., 2003). Later on, Van Son et al. (2014) introduced the *time* dimension in order to distinguish between future and non-future events and proposed to combine factuality detection and sentiment analysis in order to support perspective analysis. The three-dimensional scheme has also been employed by Tonelli et al. (2014) in the Newsreader corpus, as well as by Minard et al. (2014) for the creation of the ITA-Timebank, so far the only existing Italian corpus annotated with factuality. Based on previous analysis provided in Diab et al. (2009), García and Montraveta (2020) slightly diverged from this model by employing the *commitment* dimension in place of the *certainty* dimension, when they created TAGFACT, an annotated corpus for factuality in Spanish. Yao et al. (2021) crowdsourced a large dataset of Covid-19-related news, annotated with events, sources and modal dependencies and cast the factuality detection as a modal dependency parsing problem.

Modality. Unlike factuality, for which there is a preferred scheme, modality is a more multifaceted phenomenon (Morante and Sporleder, 2012; Ghia et al., 2016; Pyatkin et al., 2021) and also for this

reason, several frameworks have been proposed (Palmer, 1986; Fintel, 2005). However, an established subdivision often adopted in NLP works distinguishes between *epistemic*, *deontic* and *dynamic* modality (Ruppenhofer and Rehbein, 2012; Marasović and Frank, 2016). In Ruppenhofer and Rehbein (2012) the authors proposed an annotation scheme for modal verbs in English with sense-annotations on the MPQA news corpus (Wiebe et al., 2005). In a more articulated setting, Ghia et al. (2016) proposed to consider modality as a function of the trigger-target-relation triad, called construction.

2.3 Factuality and Modality detection

Marasović and Frank (2016) framed modality detection as a Word-sense Disambiguation (WSD) task, focusing only on verbs and using a Convolutional Neural Network (CNN) architecture. Also, the authors adopt the traditional epistemic/deontic/dynamic modal scheme, experimenting on three different datasets in English and German. Rudinger et al. (2018) tested two different bidirectional LSTMs models for event factuality detection, on four different English datasets, finding that lexical features have minimal impact on performance, while across-dataset multi-task setting results in increased performance. Conversely, Pouran Ben Veyseh et al. (2019) used a Graph Convolutional Network (GCN) architecture for factuality detection, aiming at directly integrating syntactic and semantic information, using the same datasets as Rudinger et al. (2018), while Liu et al. (2022) experimented on the same datasets reporting performance improvements with Direct Labeled Graph Recurrent Networks (DLGRN). Yao et al. (2021) used a multitask approach, with an attention-based feed-forward network to label spans and reconstruct the modal dependency tree for each text, showing the benefits of using a joint over a pipelined approach. More recently, Murzaku et al. (2023) proposed to cast factuality detection as a text-to-text problem, focusing on the FactBank corpus. This was the first - and, to the best of our knowledge, the only so far - attempt to employ generative language models for the task. Beside focusing on factuality only, however, the authors evaluate a single sequence-to-sequence model without any comparison with other architectures, which is one of the primary contributions of this work.

3 ModaFact dataset

3.1 Data selection

In order to ensure a wide coverage of different event types, the sentences to be included in ModaFact have been *uniformly* sampled from EventNet-ITA (Rovera, 2024), a large dataset for Event Frame Parsing in Italian. Since this resource has been created starting from Wikipedia texts, also ModaFact can be published and reshared without restrictions, while ensuring its full compatibility with event extraction systems possibly built upon EventNet-ITA. Nevertheless, pre-existing event information has not been made available to ModaFact annotators.

3.2 Selection of event-denoting expressions

ModaFact's goal is to encode in a joint way the factuality and modality values of *event-denoting expressions* in text. Moreover, we aim at keeping the annotation framework as light as possible, without compromising the expressiveness of the scheme.

Event trigger identification. The first step in the annotation process is therefore the identification of event-evoking textual spans. Following previous work (Tonelli et al., 2014; Yao et al., 2021) and consistently with the data source from Rovera (2024), target spans are *any event-denoting noun, verb or multi-word expression*.[4] Event factuality and modality in ModaFact are annotated at token level, using the IOB-2 format, which allows discontinuous mentions.

After selecting an event trigger, annotators are asked to specify both a (mandatory) factuality label and an (optional, if applicable) modality label following the scheme described below (see Section 3.3). Note that given a sentence, all possible targets are annotated.

3.3 Annotation scheme

To annotate factuality given an event trigger, we devise two representations, a *fine-grained* and a *coarse-grained* one, so to cover all linguistic dimensions of factuality while providing a lighter scheme that may be useful in downstream tasks.

In line with Minard et al. (2014), Van Son et al. (2014) and Tonelli et al. (2014), all based on Saurí

[4] Considering the very low frequency of adjectives and adverbs found in Minard et al. (2014) for Italian, these parts-of-speech were not considered as targets for event-denoting triggers.

Author	Dataset	Lang	Factuality components
Saurí and Pustejovsky (2009)	FactBank	Eng	Certainty, Polarity
Van Son et al. (2014)	MPQA (*)	Eng	Certainty, Polarity, Time
Tonelli et al. (2014)	Newsreader (*)	Eng	Certainty, Polarity, Time
Minard et al. (2014)	Ita-TimeBank	Ita	Certainty, Polarity, Time
García and Montraveta (2020)	TAGFACT	Spa	Event type, Commitment, Polarity, Time
Yao et al. (2021)	Modal Dependency	Eng	Certainty, Polarity

Table 1: Existing factuality annotated datasets and corresponding modeled features. Datasets marked with (*) have not been created by the authors for the purpose and factuality annotation has been added at a later stage.

and Pustejovsky (2009), we annotate factuality on three dimensions, namely CERTAINTY, POLARITY and TIME. As for modality, a scheme has been devised that brings together the most referred classes from different schemes in the literature, aiming at maintaining both consistency and coverage.

Fine-grained Factuality The annotation scheme for factuality follows the one established in the literature already discussed above and structured along three dimensions: CERTAINTY, POLARITY and TIME.

The CERTAINTY dimension expresses the epistemic commitment of the source w.r.t. the occurrence of a given event. It has four possible values: CERTAIN, POSSIBLE, PROBABLE and UNDERSPECIFIED. An event is CERTAIN if it is assumed to belong to the world of facts (Saurí and Pustejovsky, 2009). In order to distinguish between POSSIBLE and PROBABLE events then, we rely on a test proposed by Saurí (2008): if an event can be copredicated with its PROBABLE opposite polarity counterpart, its certainty value is POSSIBLE. If the copredication with the PROBABLE event of opposite polarity leads to a contradiction, its certainty value is PROBABLE. Finally, the UNDERSPECIFIED value is assigned to events whose certainty status can not be assessed, due to the lack of necessary evidence in the utterance.

The POLARITY dimension refers to whether the event is in the scope of a negation and possible values are POSITIVE, NEGATIVE and UNDERSPECIFIED. The latter is an uncommon label, used almost exclusively for events introduced by the conjunction *se* (*if/whether*), in events embedded in indirect interrogative clauses and in clauses introduced by verbs of knowledge and awareness (see examples in Appendix B).

The TIME dimension accounts for the temporal placement of an event with respect to the moment the utterance has been produced. This dimension can assume three distinct values: FUTURE, PRESENT/PAST, UNDERSPECIFIED. Non-future events are not further analyzed since present and past time values have the same impact on the final factuality value (see Section 2). Conversely, the FUTURE value is one of the two causes of uncertainty that can render an event NON-FACTUAL.

Coarse-grained Factuality While in the fine-grained scheme each event trigger is annotated with three attributes, each representing one factuality dimension, we devise also a coarse-grained scheme, in which the three values are conflated into a single factuality value. Thus, after manually annotating the dataset with the fine-grained scheme, we automatically map the labels to such coarse-grained values. The conversion scheme is reported in Table 2. At this level, an event mention can be labeled alternatively as FACTUAL, NON-FACTUAL, COUNTERFACTUAL or UNDERSPECIFIED. In general, we rely on the assumptions by Saurí (2008) and Van Son et al. (2014) about the inconsistency of classifying future events as FACTUAL. We assume that uncertainty can originate both from a future event occurrence and from low commitment of the speaker. Once one of these two factors is present, the factuality value of the event is NON-FACTUAL, regardless of the labels assigned to the other dimensions. Therefore, an event is FACTUAL only if it is annotated as CERTAIN-POSITIVE-PRESENT/PAST. Finally, if the event mention is annotated as CERTAIN-NEGATIVE-PRESENT/PAST, the event is COUNTERFACTUAL. Cases in which a CERTAIN label is combined with UNDERSPECIFIED polarity are in theory possible, however we do not report any occurrence in our corpus.

Modality values We consider modality as a separate dimension with respect to factuality. In fact, while any mention of an event is by definition bearer of a factuality value, not all mentions of events are necessarily marked by a modality value. In ModaFact's scheme we selected 10 modality values (summarised in Table 3) which map the tra-

Fine-Grained			Coarse-grained
Certainty	Polarity	Time	Factuality value
CERTAIN	POSITIVE	PRESENT/PAST	FACTUAL
CERTAIN	NEGATIVE	PRESENT/PAST	COUNTERFACTUAL
CERTAIN	*any value*	UNDERSPECIFIED	UNDERSPECIFIED
POSSIBLE/PROBABLE	*any value*	*any value*	NON_FACTUAL
any value	*any value*	FUTURE	NON_FACTUAL
UNDERSPECIFIED	*any value*	PRESENT/PAST/TUND	UNDERSPECIFIED

Table 2: Conversion table from fine-grained to coarse-grained factuality in ModaFact.

Feature	Values
Certainty	CERTAIN, POSSIBLE, PROBABLE, UNDERSPECIFIED
Polarity	POSITIVE, NEGATIVE, UNDERSPECIFIED
Time	PRESENT/PAST, FUTURE, UNDERSPECIFIED
Modality	WILL, FINAL, CONCESSIVE, POSSIBILITY, CAPABILITY, DUTY, COERCION, EXHORTATIVE, COMMITMENT, DECISION

Table 3: Fine-grained factuality and modality values in ModaFact.

ditional four modal categories in Italian[5], as well as extend to the categories of *purpose*s, *commitment*s and *decision*s. The different values are briefly described below, while the full annotation guidelines, along with detailed examples from the dataset, are reported in Appendix B.

WILL. This modality value applies to events that are in the scope of the intention of an animated entity.

FINAL. When an aim or purpose is attributed to the agent of an event, the event is annotated as FINAL. This modal sense shares some features with WILL, but marks such cases where the will to achieve some goals (the event) is implicitly expressed in the utterance.

POSSIBILITY. This category is used to mark an *epistemic* application of the modal verb *potere* (can) for events that are introduced as one of the possible scenarios.

CONCESSIVE. It is used to label events that are subject to the authority of an entity other than their own agent. This maps the *deontic* use of the verb *potere* (can/may).

CAPABILITY. The capability mode applies to all events that represent an agent's ability, which can be originated by the availability of adequate means, mapping the *dynamic* use of the verb *potere* (can) (Ruppenhofer and Rehbein, 2012).

DUTY. The DUTY modal sense applies to events that are not directly prompted by the intention of an animated entity, but that occur as result of external circumstances or states of affairs.

COERCION. This label applies to events that occur as a result of a force exerted by an individual who is not the agent of the action itself. This class marks the *deontic* use of the verb "dovere", in that it marks an obligation towards the occurrence of an action, imposed by an external entity holding a certain degree of authority.

EXHORTATIVE. This modal sense shares some features with COERCION, such as the intention of an external entity on the agent's actions. However, we model it as a separate modal sense in that the event in focus is proposed as a mere suggestion or recommendation, an indication that implies no forceful imposition.

COMMITMENT. Every time an agent expresses self-commitment towards the occurrence of an event, the event is attributed a COMMITMENT modality. The agent takes a certain degree of responsibility for the future event to take place.

DECISION. Deciding is a cognitive action that one can undertake individually, and in this case it means forming a resolution towards a future action.

To summarise, in the fine-grained annotation scheme each event-denoting expression is labeled with three factuality values and one (optional) modality value, while in the coarse-grained one it is assigned only one factuality and one (optional) modality label.

3.4 Inter-annotator Agreement

The corpus was annotated by an Italian native speaker with a background in NLP and lexical semantics. In parallel, a portion of the data (302 sentences, approximately 10% of the whole corpus) has been annotated by a second native speaker with the same background.

We compute inter-annotator agreement (IAA) at

[5] *Volere* (want), *potere* (can/could/may/might), *dovere* (must/shall/should/have to)

three levels: event-triggers, trigger+labels spans, and agreement by single label. At each stage, we consider both Jaccard Index, a simple measure of difference between the two annotated sets, and Cohen's Kappa (Cohen, 1960). IAA scores, reported in Table 4, have been computed on the fine-grained, 4-dimensional version of the annotated corpus (Certainty + Polarity + Time + Modality).

	Jaccard	Cohen's K
Event triggers	0.903	0.903
Triggers + labels	0.744	0.740
Labels only	0.935	0.956

Table 4: IAA scores on 302 sentences from ModaFact. All scores are intended as exact match.

Event triggers are the minimal unique sequences of tokens in a sentence that denote an event occurrence and are annotated with factuality and modality values (see Section 3.2). At this stage we evaluate, via exact match comparison, only whether the correct tokens have been marked as events. Consider for instance the following sentence:

> Inoltre, è in grado di esercitare la già sperimentata influenza illecita su persone e strutture, con cui potrebbe fuggire all'estero dove ha proprietà e conoscenze. (*Moreover, he is able to exert the already proven illicit influence over people and structures, with which he could flee abroad where he has property and connections.*)

A1: `esercitare`, `fuggire`, `conoscenze`
A2: `esercitare`, `fuggire`, `ha conoscenze`

In the above example, the two annotators marked the same event tokens `esercitare` (*exert*) and `fuggire` (*flee*), while for the third one disagreement arises about whether the auxiliary `ha` should be part of the event extent or not. In this case, despite the core event being the same, a mismatch is computed.

Event triggers + labels. In this setting we compute the agreement, between each event trigger-labels pair, for example:

A1: *utilizzare*=POSSIBLE-POS-FUTURE-FINAL
A2: *utilizzare*=CERTAIN-POS-UNDERSPECIFIED-FINAL

Note that if the event spans do not fully match or some of the four annotated attributes are not the same in the two annotations, the agreement score is equal to zero. No partial score is assigned.

Labels only. Taking as input only the subset of trigger-labels pairs where the two annotator agree on the trigger, we evaluate the agreement on the label assignment, again as exact match.

Overall, we observe that agreement is quite high despite the conservative approach adopted to compute the scores, showing that the task and the annotation guidelines are well-defined.

Sentences	3,039
Words	73,784
Annotations	10,445
Unique label assignments	33,029
Words per sentence (avg.)	24.28
Annotations per sentence (avg.)	3.44
Unique label assignments per sentence	10.87

Table 5: Dataset statistics for ModaFact.

3.5 Dataset statistics

Overall, ModaFact consists of 3,039 sentences containing 10,445 event mentions annotated with factuality and modality labels. The complete statistics for the dataset are shown in Table 5, while further details on the label distribution are reported in Appendix A. Concerning factuality, the class distribution is rather skewed, with the CERTAIN-POSITIVE-PRESENT/PAST pattern covering the large majority of events. Modality is labeled in around 16% of the cases, and for this task the ten classes are more evenly distributed, with FINAL being the most frequent modality value.

4 Experimental Setting

Starting from the ModaFact dataset, we aim at implementing a system able to replicate factuality and modality annotation. We therefore compare different approaches for fine-grained and coarse-grained classification. For each version of ModaFact (fine- or coarse-grained), we create 5 train/dev/test splits (60-20-20) via *stratified* sampling, in order to keep the same label distribution over sets in each fold. Detailed hyper-parameter settings for each model are provided in Appendix C.

4.1 Models

We experiment by *fine-tuning* five different monolingual and multilingual supervised learning models, this way covering all the main deep learning

paradigms currently in use: Masked Language Models, Sequence-to-Sequence models and Causal (autoregressive) Lannguage Models. Beside the quantitative evidence, we aim at understanding what tool is most suitable for the automatic labelling task, as well as comparing the behavior of generative models on extractive tasks like ours. We employ the exact same data over all models, with different formatting that we document in this Section, thus enabling full comparability of the results.

Masked Language Models (MLM). Despite recent advances brought by generative large language models, transformer-based encoder-only models like BERT (Devlin et al., 2019) achieved the state-of-the-art in classification and extraction tasks for a long time, so we adopted it as a baseline. BERT-like architectures are bidirectional, which makes them particularly suitable for token-level tasks, in that they have access both to previous and following context. We used the Italian BERT-base model[6] along with the MaChAmp toolkit (van der Goot et al., 2021) and experiment with two different learning settings, in particular:

(a) token-level sequence labeling with a conditional random fields (CRF) decoding layer, enforcing the BIO scheme (seqBIO). In this setting, *for each token*, all assigned labels are conflated to form a unique label (Table 6a).

(b) multitask sequence labeling, where factuality and modality are learned in parallel as two separate seqBIO tasks (Table 6b).

	T1 - Factuality+modality
token_{n-1}	O
token_{n}	B-CERTAIN-POS-FUTURE-WILL
token_{n+1}	I-CERTAIN-POS-FUTURE-WILL
token_{n+2}	O

(a) seqBIO setting.

	T1 - Factuality	T2 - Modality
token_{n-1}	O	O
token_{n}	B-CERTAIN-POS-FUTURE	B-WILL
token_{n+1}	I-CERTAIN-POS-FUTURE	I-WILL
token_{n+2}	O	O

(b) Multitask setting.

Table 6: Label formatting in (a) seqBIO and (b) multitask settings respectively.

Sequence-to-Sequence (Seq2Seq. Sequence-to-sequence models (Sutskever et al., 2014), based on an encoder-decoder architecture, have commonly been used for tasks like machine translation and question answering, and are trained to transform (transduct) an input sequence into an output sequence. Despite the unidirectionality of their decoder, Seq2Seq models can still access the context by attending to the latent representation of the full input via the attention layer. For our task, we fine-tuned a pre-trained instance of mT5[7] (Xue et al., 2021), a multilingual model supporting 101 languages, including Italian.

Causal Language Models (CLM). Unlike MLM and Seq2Seq, causal models are typically based on autoregressive, decoder-only architecture and are trained to predict the next token based on previous tokens, by sampling it from a probability distribution. We fine-tuned two models: Minerva,[8] a recent mono(/bi)lingual model specifically trained on Italian data, and Aya[9] (Üstün et al., 2024), a fully multilingual model trained on specifically curated data. In the case of Seq2Seq and CLM, the task is cast as a text-to-text problem (Paolini et al., 2021; Murzaku et al., 2023). However, as inline formatting tends to produce longer output sequences, thus increasing computational and time requirements, we opt for an attribute-value format, producing as output a sequence of span-label for each sentence, for example:

> L'alleanza con Silla diede inoltre prova di essere molto utile, grazie all'abilità di attaccare Goguryeo da opposte direzioni. (*The alliance with Silla also proved to be very useful, thanks to the ability to attack Goguryeo from opposite directions.*)
>
> alleanza = CERTAIN - POS - PRESENT/PAST
>
> attaccare = UNDERSPECIFIED - POS - UNDERSPECIFIED - CAPABILITY

5 Evaluation and Results

Consistently with the methodology used for the IAA, the performance of each model has been evaluated on three levels: event-trigger detection, trigger+labels, labels only. Although the task was tackled end-to-end, this three-layered evaluation approach highlights the capabilities of the models at

[6] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[7] https://huggingface.co/google/mt5-xxl

[8] https://huggingface.co/sapienzanlp/Minerva-3B-base-v1.0

[9] https://huggingface.co/CohereForAI/aya-23-8B

different stages of the task. Furthermore, the model performance is evaluated both for fine-grained and for coarse-grained classification.

Event triggers. We first assess the model's ability to correctly detect event-denoting text spans. Results are shown in Table 7.

Model	FG-F1	CG-F1
BERT (seqBIO)	**0.860**	**0.865**
BERT (multitask)	0.856	0.862
mT5-xxl-13B	**0.873**	**0.871**
Aya23-8B	0.825	0.831
Minerva-3B	0.520	0.657

Table 7: F1-scores for **event-trigger** prediction on fine-grained (FG) and coarse-grained (CG) data, exact match, 5-fold average. **Baseline**, **best performance**.

Triggers + labels. At this stage we evaluate the ability of each model to correctly detect trigger-labels pairs. Results are reported in Table 8.

Labels. To provide a more detailed analysis, at this stage we perform per-label evaluation (averaged results are reported in Table 9, detailed per-class results are presented in Appendix D). We apply the following rationale: we first evaluate *gold* event triggers, *i)* if a gold trigger is matched by a prediction, we move on to evaluate each associated label individually, *ii)* else, if the gold trigger is not matched, each associated label is considered a false negative. Finally, *iii)* for each unmatched *predicted* trigger+labels pair, all labels are computed as false positives.

6 Discussion

Results across different models, in terms of macro F1 average and weighted F1 average over 5 folds, are reported in Table 9, while per-label results are presented in Appendix D. Given the severe class imbalance in ModaFact, we also computed a weighted average, in order to provide more realistic insights. We chose the BERT seqBIO setting as a baseline as this is the most efficient architecture in terms of required preprocessing, computation, training and evaluation. Overall, we observe that mT5 (Seq2Seq) performs best across both tasks and both data configurations, followed by encoder-only models. On the other hand, decoder-only generative models struggle to perform competitively, failing to beat the baseline in any task. All models considered in this work are very clearly precision oriented, with a precision/recall delta ranging from 0.08 in mT5 up to 0.3 in Minerva. In particular, two trends can be observed: firstly, compared to MLMs and Seq2Seqs, if on the one hand CLMs yield reasonably high precision, they suffer from low recall, proving unable to generate all the token+labels sequences. Also, this low performance seems to originate in two different ways in Aya and Minerva: while Aya shows a good ability to detect event triggers and its loss of performance mainly arises in the labeling part, Minerva shows issues right from the trigger detection phase, causing a cascading drop in performance (see Tables 7 and 8). Overall, fine and coarse-grained versions of the data perform comparably, with differences that are mostly local, due to single labels. Concerning the different classes, for fine-grained factuality UNDERSPECIFIED values prove to be most difficult to detect (this is true also at coarse grained level), along with the PROBABLE certainty label. With regard to modality values, the most difficult labels to predict are PROBABLE, COERCION and CAPABILITY, while FINAL, WILL and DECISION score the best performance (Appendix D). Finally, we compared the per-label inter-annotator agreement scores with the performance of the model. Results, reported in Table 10, reveal a very strong positive linear correlation between human (dis)agreement and per-class prediction performance, showing that the models tend to faithfully approximate human judgement and agreement degree over the labelset.

Model	FG-F1	CG-F1
BERT (seqBIO)	**0.715**	**0.749**
BERT (multitask)	0.709	0.745
mT5-xxl-13B	**0.748**	**0.766**
Aya23-8B	0.687	0.714
Minerva-3B	0.405	0.534

Table 8: F1-scores for **trigger+labels pairs** prediction on fine-grained (FG) and coarse-grained (CG) data, exact match, 5-fold average. **Baseline**, **best performance**.

7 Conclusions and Future Work

In this work we presented ModaFact, a novel resource for Italian containing over 10,000 event mentions annotated with factuality and modality values. The dataset has been made freely available to the research community to foster future works on event processing taking into account speaker's perspective and beliefs. We also presented an evaluation of three different classification paradigms,

		Fine-grained				Coarse-grained			
		Factuality		Modality		Factuality		Modality	
	Lang	F1	w-F1	F1	w-F1	F1	w-F1	F1	w-F1
BERT (seqBIO)	ITA	*0.67*	*0.837*	*0.712*	*0.733*	*0.719*	*0.836*	*0.726*	*0.744*
BERT (multitask)	ITA	0.67	0.837	**0.733**	**0.751**	0.7	0.83	**0.737**	**0.755**
mT5-xxl-13B	MULTI (101)	**0.716**	**0.862**	**0.761**	**0.768**	**0.732**	**0.849**	**0.763**	**0.778**
Aya23-8B	MULTI (101)	0.645	0.820	0.695	0.705	0.680	0.817	0.698	0.709
Minerva-3B	ITA, ENG	0.466	0.591	0.527	0.531	0.563	0.700	0.544	0.547

Table 9: Macro F1 average and F1 weighted average scores, 5 folds. *Baseline*, **results above baseline**, best results.

Model	Pearson r	p-value
BERT (seqBIO)	0.926	4.78^{-9}
BERT (multitask)	0.918	1.12^{-8}
mT5-xxl-13B	0.897	7.82^{-8}
Aya23-8B	0.914	1.64^{-8}
Minerva-3B	0.931	2.50^{-9}

Table 10: Pearson Correlation Coefficient between Cohen's K score on the inter-annotator agreement set and the overall models' performance.

for a total of five different classification models, showing that a Sequence-to-Sequence model, mT5, consistently outperforms the others both for fine- and coarse-grained classification, as well as for subtasks such as event trigger prediction.

Future works will follow two directions: on the one hand, we plan to go beyond model fine-tuning, by experimenting with more diverse learning techniques like in-context learning an zero-shot, in order to assess the performance of LLMs on the joint task of factuality and modality detection. On the other hand, ModaFact will be tested in downstream tasks like disinformation detection, where factuality and modality patterns will be analysed in relation to how different news sources present true or false information.

Limitations

As for generative models, we did not perform hyperparameter optimization, which could result in suboptimality of the presented results. This was due to the high computational cost of fine-tuning LLMs (even in their quantized versions). Secondly, we observe that modality is a strongly language-dependent phenomenon. Therefore, the scheme we propose may not be *directly* portable to other languages. Also, we acknowledge that the aforementioned scheme might be open to extensions, which have not been examined in this work.

Ethics Statement

ModaFact has been built starting from Wikipedia data, therefore it will be released under Creative Commons Attribution-ShareAlike 4.0 license following copyleft principles. No sensitive information is present in the data.

Concerning annotation effort, both annotators carried out data annotation as part of their work as employees at the host institution.

Acknowledgments

We acknowledge the support of the PNRR project FAIR- Future AI Research (PE00000013), under the NRRP MUR program funded by NextGeneration EU. This work was also funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101135437 (AI-CODE).

References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, pages 1–12.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44502–44523. Curran Associates, Inc.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345,

Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. pages Proceedings of the Third Linguistic Annotation Workshop, 68–73.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952, Singapore. Association for Computational Linguistics.

Kai Von Fintel. 2005. Modality and language. In Donald M. Borchert, editor, *Encyclopedia of Philosophy*, pages 20–27. macmillan reference.

Glòria Vázquez García and Ana María Fernández Montraveta. 2020. Annotating factuality in the TAGFACT corpus. page Multiperspectives in Analysis and corpus design.

E Ghia, L Kloppenburg, M Nissim, P Pietrandrea, V Cervoni, et al. 2016. A construction-centered approach to the annotation of modality. In *Proceedings of the 12thISO Workshop on Interoperable Semantic Annotation*.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Xiao Liu, Heyan Huang, and Yue Zhang. 2022. End-to-end event factuality prediction using directional labeled graph recurrent network. *Information Processing & Management*, 59(2):102836.

Ana Marasović and Anette Frank. 2016. Multilingual modal sense classification using a convolutional neural network. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 111–120, Berlin, Germany. Association for Computational Linguistics.

Anne-Lyse Minard, Alessandro Marchetti, and Manuela Speranza. 2014. Event factuality in Italian: Annotation of news stories from the Ita-TimeBank. page Fourth International Workshop EVALITA 2014.

Anne-Lyse Minard, Manuela Speranza, and Tommaso Caselli. 2016. The EVALITA 2016 event factuality annotation task (FactA). pages EVALITA. Evaluation of NLP and Speech Tools for Italian, 32–39.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260.

John Murzaku, Tyler Osborne, Amittai Aviram, and Owen Rambow. 2023. Towards generative event factuality prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 701–715, Toronto, Canada. Association for Computational Linguistics.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval*, pages 639–649, Cham. Springer International Publishing.

Frank Robert Palmer. 1986. Mood and modality.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Ma Jie, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, Stefano Soatto, et al. 2021. Structured prediction as translation between augmented natural languages. In *ICLR 2021-9th International Conference on Learning Representations*, pages 1–26. International Conference on Learning Representations, ICLR.

Paola Pietrandrea. 2005. Epistemic modality: functional properties and the Italian system.

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4393–4399, Florence, Italy. Association for Computational Linguistics.

Mario Prieto, Helena Deus, Anita De Waard, Erik Schultes, Beatriz García-Jiménez, and Mark D Wilkinson. 2020. Data-driven classification of the certainty of scholarly assertions. *PeerJ*, 8:e8871.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, David Day, Lisa Ferro, Robert Gaizauskas, Marcia Lazo, Andrea Setzer, and Beth Sundheim. 2003. The TimeBank corpus. pages Corpus Linguistics, 647–656.

Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. The possible, the plausible, and the desirable: Event-based modality detection for language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 953–965, Online. Association for Computational Linguistics.

Marco Rovera. 2024. EventNet-ITA: Italian frame parsing for events. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 77–90, St. Julians, Malta. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Josef Ruppenhofer and Ines Rehbein. 2012. Yes we can!? annotating English modal verbs. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1538–1545, Istanbul, Turkey. European Language Resources Association (ELRA).

Roser Saurí. 2008. A Factuality Profiler for Eventualities in Text. pages P.h.D. Thesis, Brandeis University.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. pages Language Resources and Evaluation, 43(3)–227.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Sara Tonelli, Rachele Sprugnoli, and Manuela Speranza. 2014. NewsReader Guidelines for Annotation at Document Level. pages Extension of Deliverable D3. Technical Report NWR–2014–2.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Chantal Van Son, Marieke Van Erp, Antske Fokkens, and Piek Vossen. 2014. Hope and fear: Interpreting perspectives by integrating sentiment and event factuality. pages Proceedings of the 9th Language Resources and Evaluation Conference, 26–31.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jiarui Yao, Haoling Qiu, Jin Zhao, Bonan Min, and Nianwen Xue. 2021. Factuality assessment as modal dependency parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1540–1550, Online. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

A Dataset Statistics

This section provides a detailed overview on the dataset, in particular about label distribution in different settings. This appendix is referenced in Section 3.5.

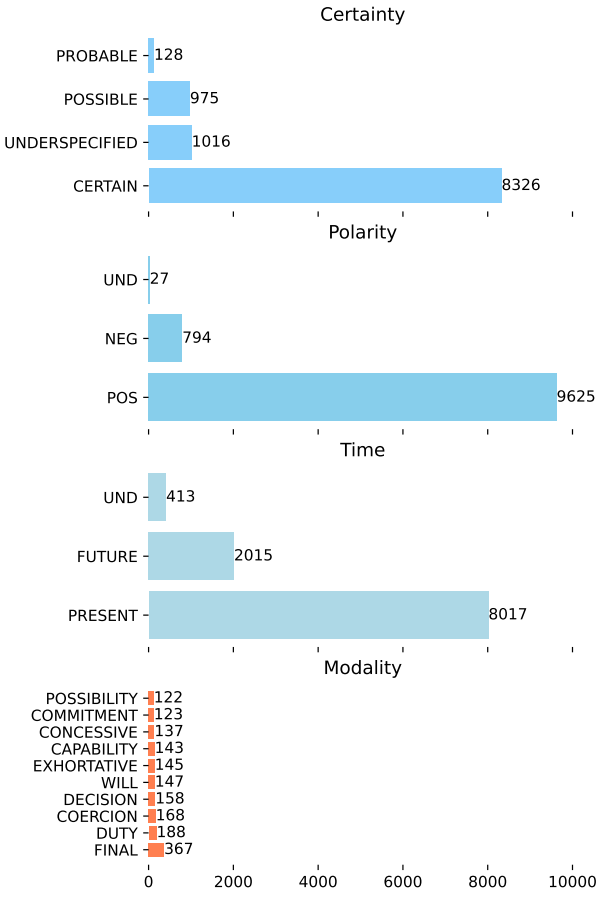

Figure 1: Label distribution in the fine-grained version of the dataset.

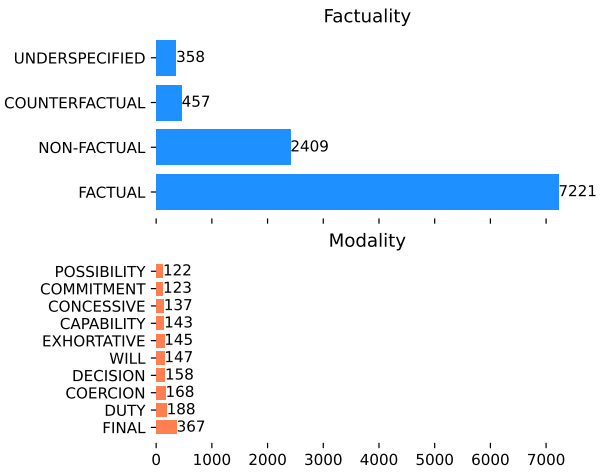

Figure 2: Label distribution in the coarse-grained version of the dataset.

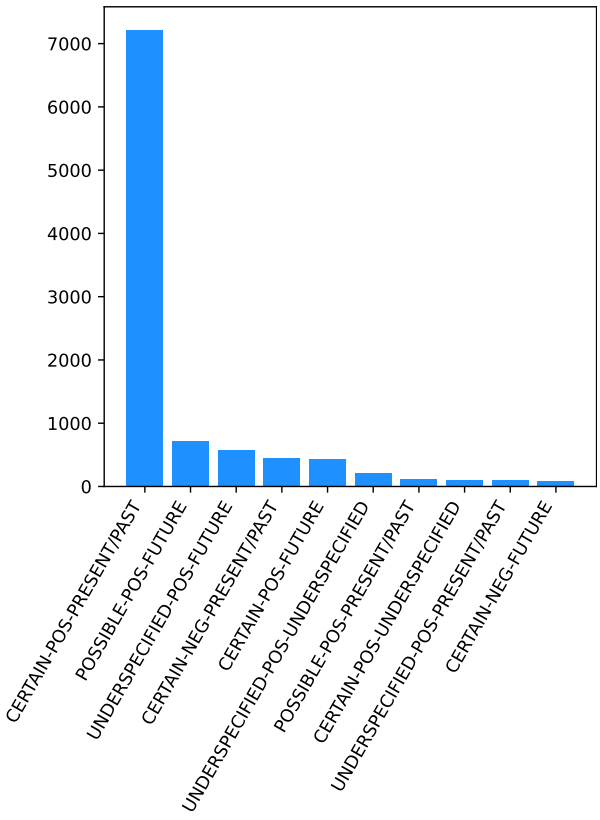

Figure 3: 10 most frequent fine-grained combinations of factuality labels in the dataset.

B Annotation Guidelines

In this Section we provide further details and examples related to Section 3.3 about ModaFact's annotation guidelines for factuality and modality. Each example sentence is prepended with its unique identifier in the dataset.

B.1 Factuality

CERTAINTY. Information about the certainty value of an event is most commonly conveyed by the verb tense, but can also be provided by adjectival, adverbial and prepositional phrases. Moreover, separate clauses and distant components of the period concur in the definition of the certainty value of an event. In the following example, the absolute certainty of the event *risparmiare* (to save) is mitigated by the presence of a conditional subordinate clause. Taking this condition into account, the certainty value of the event is POSSIBLE:

> w90711_66: se in tutta Sodoma e Gomorra avesse trovato solo "dieci giusti", a motivo di quei dieci, avrebbe sicuramente [risparmiato POSSIBLE] le città dalla distruzione.
> (*if he was to find ten righteous people in the cities of Sodom and Gomorrah, for the sake of those ten people, he would certainly spare the cities from destruction.*)

POLARITY. Several lexical features contribute to the assignment of the label, in addition to the standard negation adverb *non* (not). The assessment of the polarity value needs to take into consideration the inferences deriving from the interaction of conjunctions, embedding predicates and verb tenses with negation:

> w2907293_279: I 66 italiani rimanenti non poterono che [arrendersi POSITIVE].
> (*The 66 remaining Italians couldn't do anything but surrender.*)
>
> I 66 italiani rimasti non poterono [arrendersi NEGATIVE].
> (*The 66 remaining Italians could not surrender.*)

On the polarity dimension, UNDERSPECIFIED is an uncommon label, used almost exclusively for events introduced by the conjunction *se*. The Italian word *se* translates both the English conjunctions *if* and *whether*. The latter is the usage that indicates uncertainty or multiple possibilities. We consider it to be neutral with respect to the polarity of the embedded events. A typical case in which polarity remains underspecified is that of events embedded in indirect interrogative clauses, such as

> w115269_73: Nel film documentario [...] gli venne chiesto perché fosse entrato nelle SS e se [provasse rimorso UNDERSPECIFIED] nell'aver fatto tale scelta.
> (*In the documentary film [...] he was asked why he had joined the SS and whether he felt remorse about that choice.*)

and events in clauses introduced by verbs of knowledge and awareness:

> w5032529_106: Il titolo è stata un' aggiunta postuma, originariamente Marco intitolò l' opera "A se stesso", ma non si sa se [avesse intenzione UNDERSPECIFIED] di renderla pubblica.
> (*Marco titled the work "A se stesso", but it is not known whether he intended to make it public.*)

The expression *o meno* (or not) is an indicator that the event and its polar opposite are equally valid options in the utterance:

> w3191251_54: Dopo l' annuncio dell' armistizio trascorse tre giorni in mare, nel dubbio se [obbedire UNDERSPECIFIED] o meno all' ordine di consegnarsi agli Alleati a Bona.
> (*After the communication of the armistice he spent three days at sea, doubting whether he should obey or not the order to surrender to the allies in Bona.*)

The polarity of an event is not driven exclusively by the presence of shallow features such as a lexicalized negation. In the following example, indicating the ending of a temporal span during which the event is not *yet* taking place signals that it did in fact take place consequently:

> w2570514_115: Anche se l'accordo venne firmato ufficialmente il 1° ottobre 1800, esso non [entrò in vigore POSITIVE] fino al 1802.
> (*Although the agreement was officially signed on October 1, 1800, it did not go into effect until 1802.*)

TIME. The annotation of the TIME dimension must take into account the relationship between the moment of utterance and the moment of occurrence of the event reported in the utterance:

> w3723080_53: In realtà poco dopo Massimino, Costantino e Licinio si [coalizzarono PRESENT/PAST] per [eliminare FUTURE] il primo dei quattro "augusti".
> (*In fact, shortly after, Massimino, Costantino and Licinio joined forces to kill the first of the four "augusti".*)

In this example with a final clause, the event *eliminare* (to kill) is secondary to the main event *si coalizzarono* (joined forces) both logically and temporally.

Unless further specified, we assume the verb tense to carry the most reliable information about the temporal placement of an event. However, the time value of an event might or might not correspond to its grammatical tense. There might be elements in the clause that suggest different classifications, such as specific dates:

> w245577_69: Il 3 febbraio dello stesso anno la società comunica tuttavia che verrà [reintegrato FUTURE] nella rosa a partire dal giorno seguente.
> (*However, on February 3rd of the same year, the club breaks the news that he would be reintegrated in the roster from the following day.*)
>
> w2604095_83: Tentando di raggiungere l'Etiopia [morirà PRESENT/PAST] a Suakin (Sudan) nel 1641; è sepolto a Goa in India.
> (*In an attempt to reach Ethiopia, he would die in Suakin (Sudan) in 1641; he is buried in Goa (India).*)

On the other hand, an event expressed in future tense needs to be evaluated in its context for an accurate assessment of its real time value. Other elements in the sentence can help decide whether the grammatical future reflects a reference to an actual future point in time, or it is used as a rhetorical device.

B.2 Modality

WILL
The label WILL characterises events that are the object of the intention of an animated entity. The phrase embedding the event usually contains the modal verb *volere* (want) or can be rephrased with it without losing its core meaning. Contrary to COERCION, the agent of the prospective events may or may not be the same entity that expresses the desire that they occur:

> w546153_78: Kopaszewski inizialmente rifiutò l'offerta dell'Angers, volendosi [trasferire WILL] al Reims, e i dirigenti del Nœux-les-Mines desideravano che il giocatore [restasse WILL] una stagione in più per poi partire per Reims.
> (*Kopaszewski initially rejected the proposal by Angers, because he wanted to transfer to Reims, and the directors of Nœux-les-Mines wanted the player to stay one more season and then leave for Reims.*)
> w4576624_53: [...] sia Giorgio II che diversi ministri dell'Hannover erano intenzionati a [riprendere WILL] la guerra.
> (*[...] both George II and many Hanoverian ministers intended to resume the war.*)

FINAL
When an aim or purpose is expressed through a final construction, the event is annotated as FINAL. The usual lexical triggers of this label are the preposition *per* (for, to) and constructions such as *allo scopo di* (with the aim to), *al fine di* (in order to), *in modo da* (so that), etc. They are often followed by a verb in its infinitive form.

> w2110656: dei nobili militari avevano preso le armi per [rovesciare FINAL] lo zar, non però per [sostituirlo FINAL] con uno altro più gradito, ma per [porre fine FINAL] all'autocrazia.
> (*some military nobles had taken up arms to overthrow the czar, however, not to replace him with a more welcome one, but to end autocracy.*)

It is worth mentioning that not every occurrence of the preposition *per* before an event implies that it should be marked with the label FINAL:

> w100937_3: [...] comparve nello show con una certa frequenza, per poi [abbandonarlo FINAL] dopo la prima stagione.

(*[...] she appeared on the show with some frequency, only to drop out after the first season.*)

POSSIBILITY

The label POSSIBILITY is used to mark an epistemic application of the modal verb *potere* (can), in other words, for events that are introduced as one of the possible scenarios. The distinction between deontic and epistemic use of the modal verb *potere* or semantically similar constructions is not always straightforward. As noted in (Pietrandrea, 2005), the deontic and the epistemic meaning of verbs such as *potere* can be expressed by the same lexical forms, and the receiver of the message relies on the context to disambiguate it. An event is marked as POSSIBILITY when it is merely presented as an opportunity, not related to one's own capabilities or restricted by somebody else's authorization.

Besides the modal verb, other representative lexical realizations that trigger this kind of events are: *avere la possibilità di* (have the possibility to), *avere modo di* (have a chance to), *avere l'opportunità/ occasione di* (have the opportunity/occasion to).

w642306_257: Operata alcuni giorni dopo, comincia la fase di riabilitazione in vista di un possibile [ritorno POSSIBILITY] alle competizioni.
(*Having undergone surgery a few days later, she starts the rehabilitation phase ahead of a possible return to competition.*)

CONCESSIVE

In contrast with POSSIBILITY, the label CONCESSIVE marks the *deontic* modality of an event, in that it indicates the regulatory influence of an entity holding a certain degree of authority over its occurrence. The concept of concession refers to the modal verb *potere*, but also shows analogies with the label COERCION, its counterpart within deontic modality. They are in fact the two aspects of exertion of power: a prohibition is closely akin to an obligation *not to do* something.

w2916753_181: [...] all' azienda guidata un tempo da Maybach e suo figlio venne proibito di [costruire UNDERSPECIFIED-NEGATIVE-FUTURE-CONCESSIVE] dirigibili.
(*the company once led by Maybach and his son was prohibited from building airships.*)

w1241546_40: In pratica gli allora feudatari del luogo, i Savelli, nel 1321 diedero l' autorizzazione a [smantellare UNDERSPECIFIED-POSITIVE-FUTURE-CONCESSIVE] le strutture della villa.
(*Basically, the then feudal lords of the place, the Savelli family, in 1321 gave permission to dismantle the structures of the villa.*)

w604247_167: Nell' ottobre 1778 chiese il permesso a Washington e al Congresso di [tornare UNDERSPECIFIED-POSITIVE-FUTURE-CONCESSIVE] a casa in licenza.
(*In October 1778 he asked permission from Washington and Congress to return home on furlough.*)

CAPABILITY

The CAPABILITY label applies to all the events that represent an ability of an (animated) entity. The ability to carry out an action can be originated by the presence of adequate means:

w2836434_42: La nave necessitava di operazioni di revisione, che però i turchi non erano in grado di [effettuare CAPABILITY] a causa della mancanza di bacini di carenaggio adeguati.
(*The ship needed overhaul operations, but the Turks were unable to carry them out due to the lack of adequate dry docks.*)

Common lexical units triggering a CAPABILITY event are, among others: *essere in grado* (to be able to), *sapere* (to know), *essere capace* (to be capable of), *abilità* (ability), *capacità* (capability), *potere* (can).

The verb *riuscire* (be able to, manage to) can assume an analogous role as well, usually when introducing durative or recurrent events. Nonetheless, it is not to be confused with an indication of the successful outcome of an action. The following examples show the verb *riuscire* in its two meanings of 1) inherent ability and 2) result of an action, respectively:

w2296151_1: Nel libro viene narrata la vita di una giovane coppia, Frank e April Wheeler, che non riesce a [comunicare

CERTAIN-NEGATIVE-PRESENT/PAST-CAPABILITY], ingarbugliati tra nevrosi, bugie e furibonde liti.
(*The book chronicles the lives of a young couple, Frank and April Wheeler, who fail to communicate, entangled in neuroses, lies, and furious arguments.*)

w318413_132: Un migliaio di manifestanti riescono ad [arrivare CAPABILITY] in città ed effettuano un sit-in.
(*A thousand protesters manage to arrive to the city and carry out a sit-in.*)

The certainty value of a CAPABILITY event depends largely on its polarity value. Whereas a negative polarity indicates the impossibility of said eventuality, like in sentence w2296151_1, the certainty value of a positive event remains open to assessment. If more specific hints are not provided by the context, the value remains UNDERSPECIFIED. The same is true for the TIME axis, since the action of which the subject is capable does not refer to a specific occurrence, but rather to a general activity that does not require an anchoring in time. In this aspect, this case is similar to a generic statement.

w4948532_193: Dopo una rischiosa operazione chirurgica, è infatti capace di respirare UNDERSPECIFIED-POSITIVE-UNDERSPECIFIED-CAPABILITY] sott'acqua [...].
(*After a risky surgical operation, he is able to breathe underwater [...].*)

DUTY

The expressions triggering DUTY are often the same as for COERCION. An event is considered to fall in the DUTY rather than in the COERCION category if the external force is not an animated entity, that is, if it arises from the lack of alternatives rather than from an external will:

w1827439_25: [...] la scarsa autonomia costrinse l'incrociatore ad [interrompere DUTY] anticipatamente la missione.
(*The low autonomy forced the cruiser to interrupt the mission ahead of time.*)

COERCION

The label COERCION is used to annotate events triggered by orders, requests, assignments and obligations. These triggers can in turn be lexically expressed by verbs (*ordinare*, (to order) *forzare* (to force), *incaricare* (to charge), etc.) and substantives (*comando* (command), *richiesta* (request), etc.).

In its most basic form, that is, if no other clue is given about the outcome of the request, an event introduced by an order or a request does not provide information about its own certainty. The event is still only speculated and no concrete anchoring to the real world is evident. As suggested in (Minard et al., 2016), such an event is annotated with underspecified certainty. There are, however, cases in which the introducing phrase arguably conveys clues about the certainty of the embedded event, such as

w1239385_33: Rembrandt la fece [rinchiudere COERCION] in un manicomio di Gouda [...]
(*Rembrandt had her locked up in a prison in Gouda [...].*)

The verb *fare* in a preterite tense, suggests that the events in its span have indeed taken place.

Lexical units such as *dare/ricevere l'incarico* might indicate a variety of situations ranging from an order to the assignment of a role, so each case needs to be evaluated based on its pragmatic context.

w1436073_42: [...] tra i quali primeggia quello di Cagliari, ove più volte gli fu dato l' alto incarico di [reggere COERCION] l'ufficio del Provveditore agli studi.
(*[...] among which that of Cagliari stands out, where multiple times he was assigned the high office of School Superintendent.*)

The attribution of an institutional role is considered to be the assignment of a job position, rather than an imposition, therefore lacking the forceful component inherent of the COERCION modal label.

A blurry line divides circumstances caused by willing entities from impersonal phenomena, so the relevance of agentive intervention should be considered when evaluating a potential case of COERCION. The following example shows a borderline case, that is labeled as NECESSITY although *arrivo* (arrival) is not a strictly natural cause:

w3843686_117: Alle elezioni di gennaio ad Hanoi vi fu il trionfo dei Viet

Minh, ma l'imminente arrivo a nord di truppe francesi, previsto per marzo, costrinse Ho a [negoziare CERTAIN POSITIVE PRESENT/PAST NECESSITY].
(*In the January elections in Hanoi there was a Viet Minh triumph, but the imminent arrival of French troops in the north, scheduled for March, forced Ho to negotiate.*)

EXHORTATIVE

The label EXHORTATIVE shares some features with COERCION, such as the intention of an external agent on one's actions. However, it represents a different modality in that the embedded event is proposed as a mere suggestion or recommendation: an indication that implies no forceful imposition. The embedded event is expected to take place at a later point in time than the act of advising. As for events marked as COERCION, a simple exhortation does not offer enough elements to assess the certainty of an event, so its value remains underspecified:

w735294_35: L'unica cosa che Gandalf può consigliare al povero Frodo è di [lasciare UNDERSPECIFIED-POS-FUTURE-EXHORTATIVE] furtivamente la Contea.
(*The only thing Gandalf can advise poor Frodo to do is to stealthily leave the Shire.*)

On the other hand, if the event is reported as a fact, the annotation of the axis is to be done accordingly:

w488708_165: Ludendorff, su suggerimento di Hoffmann, [vergò CERTAIN-POS-PRESENT/PAST-EXHORTATIVE] il dispaccio al Kaiser [...]
(*Ludendorff, at Hoffmann's suggestion, penned the dispatch to the Kaiser [...].*)

COMMITMENT

An event labeled with COMMITMENT modality takes place at a later point in a time with respect to the statement itself. Nonetheless, its time value has to be evaluated by the annotator in the light of the information given by the whole sentence. The same holds for the certainty value. In case no further information is given about the outcome, the COMMITMENT event introduced by verbs such as *promettere* (to promise), *giurare* (to swear), *assicurare* (to assure) and nouns such as *promessa* (promise), *giuramento* (swearing), impegno (commitment) is annotated as CERTAIN. This is due to the fact that the source of the information is the entity who makes the promise and, from their point of view, there is no doubt that the object of commitment is going to come into being.

w1665507_32: Di conseguenza, i Francesi [...] promisero un [supporto CERTAIN-POS-FUTURE-COMMITMENT] nella causa d'indipendenza ungherese.
(*Accordingly, the French [...] promised support in the cause of Hungarian independence.*)

Nonetheless, if the sentence reveals that the commitment turns out to be not certainly fulfilled, or there are conditions set for its fulfilment, the factuality values of the event should be corrected accordingly:

w5435565_205: Egli promise di [tornare POSSIBLE-POS-FUTURE-COMMITMENT] ad Atene nel caso che il colpo di stato avesse avuto successo.
(*He promised to return to Athens in case the coup was succesful.*)

DECISION

Deciding is a cognitive action that one can undertake individually, and in this case it means forming an opinion or a resolution about an event. Conversely, a decision can also be made by a group of individuals, and in this case its meaning is closer to the concept of agreement, or formal deliberation to regulate a future behaviour.

Words that cause an event in their object span to be marked with the label DECIDE are those belonging to the semantic class of the verbs *decidere* (to decide), *stabilire* (to establish), *determinare* (to determine), *deliberare* (to sanction) and the nouns *decisione* (decision), *risoluzione* (resolution), etc. Also verbs and substantives with a weaker relation to this semantic family can be considered to trigger a DECIDE event when they can be rephrased with, for example, the verb *decidere* (to decide) and keep roughly the same meaning:

w6881230_177: [...] scelse di non [intraprendere DECIDE] alcuna azione sostanziale in tal senso.
(*He decided non to take any substantial action in that sense.*)

w11693_65: La Luxemburg, non sostenuta da tutti i suoi redattori, preferì [dimettersi DECIDE] il 2 novembre.
(*Luxemburg, not supported by all her redactors, preferred resigning on November 2nd.*)
w4450359_74: Jaime rifiutò di [crederci DECIDE] e fuggì.
(*Jaime refused to believe it and fled.*)

As shown in the last example, the introducing predicate can determine the negative polarity of an event. The examples suggest another distinction to be made in the annotation: as far as the TIME axis is concerned, the value can change between FUTURE and PRESENT/PAST depending on the predominant perspective of the sentence. If the event is only presented as the mere object of a resolution, it is projected to the future and is annotated consequently. If the event is instead clearly reported as an effective occurrence, its primary reading is that of a factual event, so its time value is PRESENT/PAST. The following sentence offers an example of complements signaling unequivocally that the event has in fact taken place:

w103179_89: Fu deciso di [avviarli CERTAIN-POSITIVE-PRESENT/PAST-DECIDE] disarmati e appiedati verso Bir Hacheim dove giunsero il 29 mattina.
(*It was decided to set them off unarmed and on foot toward Bir Hacheim, where they arrived in the morning of the 29th.*)

w1610853_77: Durante il dibattimento Friedjung non poté esibire gli originali dei documenti fotografici in questione perché Aehrenthal si era rifiutato di [fornirglieli CERTAIN-NEGATIVE-PRESENT/PAST-DECIDE].
(*During the hearing Friedjung could not produce the originals of the photographic documents in question because Aehrenthal had refused to provide them to him.*)

As far as the certainty dimension is concerned, the annotation guidelines in (Minard et al., 2016) suggest the annotation of this semantic class of events with the value CERTAIN. We maintain this as a rule of thumb, since the source expresses full conviction that the event will take place.

C Hyperparameters

BERT models have been trained on an Nvidia RTX 5000 with 16GB RAM using the MaChAmp toolkit (van der Goot et al., 2021). Both BERT (seqBIO) and BERT (multitask) have been trained with

- `batch size = 4`
- `learning rate = 1e-4`
- `weight decay = 0.01`

for 40 epochs, by picking the best model according to MaChAmp defined `span-F1` and `multi-accuracy` metrics[10], respectively.

mT5-xxl, Aya-23-8B and Minerva-3b-base-v0.1 models have been trained on an Nvidia A40, equipped with 48GB RAM, using

- `batch size = 4`
- `learning rate = 1e-4`

For these models we used 8-bit quantization and Parameter-Efficient Fine-Tuning (PEFT), with the following LORA parameters:

- `r = 32`
- `alpha = 64`
- `dropout = 0.05`

D Detailed Performance Results

This appendix presents per-label F1 scores for the five models, averaged over 5 folds, in both fine- and coarse-grained configurations (Table 11). This appendix is referenced in Sections 5 and 6.

[10] https://github.com/machamp-nlp/machamp/blob/master/docs/metrics.md

	seqBIO	multitask	mT5	Aya	Minerva
CERTAIN	0.878	0.878	**0.899**	0.865	0.615
PROBABLE	0.415	0.42	**0.499**	0.395	0.223
POSSIBLE	0.703	0.702	**0.741**	0.665	0.511
UNDERSPECIFIED	0.686	0.682	**0.727**	0.656	0.515
POSITIVE	0.914	0.911	**0.923**	0.896	0.656
NEGATIVE	0.810	0.808	**0.861**	0.822	0.615
UNDERSPECIFIED	0.670	0.564	**0.676**	0.257	0.171
PRESENT/PAST	0.871	0.868	**0.894**	0.864	0.604
FUTURE	0.793	0.791	**0.841**	0.800	0.657
UNDERSPECIFIED	0.382	0.375	**0.448**	0.359	0.236
WILL	0.809	0.814	**0.829**	0.744	0.483
FINAL	0.848	0.853	**0.870**	0.767	0.583
CONCESSIVE	0.675	0.717	0.771	**0.792**	0.624
POSSIBILITY	0.486	**0.586**	0.538	0.455	0.316
CAPABILITY	0.64	0.721	**0.738**	0.635	0.411
DUTY	0.778	0.803	**0.835**	0.709	0.389
COERCION	0.636	0.623	**0.699**	0.632	0.493
EXHORTATIVE	0.739	0.69	**0.763**	0.714	0.652
COMMITMENT	0.711	0.75	**0.772**	0.725	0.625
DECISION	**0.797**	0.777	**0.797**	0.776	0.692
FACTUAL	0.876	0.872	**0.888**	0.863	0.733
NON-FACTUAL	0.791	0.782	**0.804**	0.764	0.693
COUNTERFACTUAL	0.720	0.734	0.777	0.725	0.549
UNDERSPECIFIED	0.490	0.412	**0.460**	0.369	0.277
WILL	0.829	**0.836**	0.807	0.775	0.596
FINAL	0.854	**0.859**	0.858	0.772	0.775
CONCESSIVE	0.705	0.736	**0.794**	0.775	0.698
POSSIBILITY	0.525	**0.523**	**0.523**	0.384	0.214
CAPABILITY	0.712	0.688	**0.753**	0.609	0.454
DUTY	0.778	0.779	**0.819**	0.702	0.469
COERCION	0.610	0.639	**0.699**	0.658	0.517
EXHORTATIVE	0.738	0.723	**0.758**	0.736	0.605
COMMITMENT	0.758	0.770	**0.802**	0.757	0.636
DECISION	0.750	0.819	**0.821**	0.812	0.684

Table 11: Per-class F1 average across 5 folds. In boldface, best performance score for the class.