Linking a Hypothesis Network From the Domain of Invasion Biology to a Corpus of Scientific Abstracts: The INAS Dataset

Marc Brinner
Bielefeld University
marc.brinner@uni-bielefeld.de

Sina Zarrieß
Bielefeld University
sina.zarriess@uni-bielefeld.de

Tina Heger
Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin
t.heger@tum.de

Abstract

We investigate the problem of identifying the major hypothesis that is addressed in a scientific paper. To this end, we present a dataset from the domain of invasion biology that organizes a set of 954 papers into a network of fine-grained domain-specific categories of hypotheses. We carry out experiments on classifying abstracts according to these categories and present a pilot study on annotating hypothesis statements within the text. We find that hypothesis statements in our dataset are complex, varied and more or less explicit, and, importantly, spread over the whole abstract. Experiments with BERT-based classifiers show that these models are able to classify complex hypothesis statements to some extent, without being trained on sentence-level text span annotations.

1 Introduction

In many disciplines of science, researchers need to develop specific hypotheses that make it possible to confront general scientific claims with empirical evidence (Lloyd, 1987). For instance, studies in invasion biology, a sub-discipline of biodiversity research, investigate why certain species can establish in new ecosystems and typically formulate hypotheses specific to the species or the forms of invasion success they address (see Figure 2). It is essential for a researcher to be aware of the existing hypotheses in these fields, but, to date, structured information on claims and hypotheses investigated in a field is often hardly available. In some cases, though, valuable resources and overviews are compiled manually by domain experts as, for instance, Jeschke and Heger (2018)'s hierarchical network of hypotheses synthesizing research in the field of invasion biology. In this paper, we propose to leverage this resource as a new dataset for domain-specific information extraction from scientific publications and explore the potential of state-of-the-art off-the-shelf NLP models for automatic hypothesis identification.

Extracting domain-specific information on hypotheses from scientific publications is still a considerable challenge for state-of-the-art approaches in NLP and IE. Research on IE for the biodiversity domain provides many annotated datasets and models with domain-specific labeling schemes for named entities and relations – e.g., species, locations and habitats (Nguyen et al., 2019) – but does not account for more complex entities like claims, research questions or hypotheses. Work on argumentation mining for scientific texts (Fisas et al., 2016; Lauscher et al., 2018) annotate argumentative spans of texts, including claims, but do not link them to domain-specific knowledge. However, the lack of domain-specific categories is a major gap in existing search repositories for biodiversity researchers, as shown by (Löffler et al., 2021).

In this work, we perform initial studies on the automatic extraction of information on hypotheses investigated in scientific publications. We compile a corpus of scientific abstracts, based on metadata in Jeschke and Heger (2018)'s hypothesis network for invasion biology. We release the resulting INAS dataset that links 954 scientific papers (with abstracts and titles) to nodes in a hypothesis network. Similar to datasets in relation extraction (Mintz et al., 2009), the INAS dataset is weakly labeled, as the hypotheses are linked to the abstract as a whole, and not annotated in terms of text spans. We present a pilot analysis on hypothesis statements within the texts and find that they are complex, varied and spread over the whole abstract, challenging existing labeling schemes in IE. We carry out experiments on labeling abstracts with BERT-based classifiers and show that these models are able to detect fine-grained hypothesis categories to some extent, without being trained on text span annotations. This shows that domain-specific resources on hypotheses provide a valuable starting point for this complex IE task, and points to some challenges for future research on automatic hypothesis extraction.

2 Related Work

Our work combines ideas from named entity recognition (NER) and relation extraction (RE), which typically targets domain-specific tagging schemes, with ideas of domain-general mining of claims, which aims at discovering complex statements of claims in text. We will briefly discuss related work from these areas in the following.

2.1 Entity and Relation Extraction in Scientific Texts

Extracting information on scientific studies from publications is a well-known problem in IE (Augenstein et al., 2017; Gábor et al., 2018). Within this area, biomedical text is one of the most widely and deeply explored domains, cf. (Demner-Fushman et al., 2022), with many datasets and tools that tag, e.g., diseases (Doğan et al., 2014), drugs and chemicals (Li et al., 2016), or drug-protein relations (Miranda et al., 2021) (among many others). In the domain of biodiversity, NER datasets focus on tagging species (Gerner et al., 2010; Pafilis et al., 2013), specific concepts like bacteria and their locations (Deléger et al., 2016), or combinations of species, habitats, locations (Nguyen et al., 2019). Löffler et al. (2020) present the QEMP benchmark, which further extends the types of entities and links them to existing ontologies in biovdiversity research. The INAS dataset follows a similar direction, as our hypothesis tags are taken from an existing network of hypotheses.

2.2 Mining Claims in Scientific Texts

In argument mining, different annotation schemes for aspects of scientific arguments have been proposed, such as argumentative zones (Teufel et al., 1999, 2009), argumentation schemes (Green, 2015), or argumentative components (Lauscher et al., 2018). Due to the importance of claims in argumentative structures, several studies focus specifically on the detection of claims in a variety of domains (Aharoni et al., 2014; Lippi and Torroni, 2015; Daxenberger et al., 2017; Habernal and Gurevych, 2017), using binary schemes that mark individual sentences or spans of texts as being claims or not. Blake (2010) present a more detailed annotation study for claims in scientific texts, distinguishing between different types of claim formulations (e.g., explicit claim vs. implicit claim) and roles that different parts of the claim fulfill. Accuosto et al. (2021) annotate scientific abstracts from computational linguistics and biomedicine with a variety of tags and relations related to argumentative structure, and Fergadis et al. (2021) annotate claims and topics in scientific abstracts on sustainable development, with both studies performing experiments on automatic prediction of these annotations. None of theses datasets, though, links annotations of claims to domain-specific concepts.

3 The INAS Dataset

We now introduce the INAS dataset[1], which is based on an existing resource that organizes papers from the field of invasion biology into a network of hypotheses. In the following, we will describe this network (Section 3.1), provide an overview of the dataset we created from this resource (Section 3.2, 3.3), present a qualitative and preliminary quantitative analysis of hypothesis statements (Section 3.4) and discuss its intended use (Section 3.5).

3.1 Hi-Knowledge Network of Hypotheses

Invasion biology is concerned with researching the human-induced spread of species outside of their native ranges, caused by factors like global transport and trade. For example, plants are imported as exotic garden plants, and small insects, plant seeds, and even reptiles and mammals are regularly transported as hitchhikers with traded goods around the globe, sometimes leading to an establishment of viable populations in the wild and spread to new locations within the new range (Elton, 1958; Davis, 2009). One aim of invasion biology is to explain why it is possible for these species to establish and often even flourish in areas in which they did not evolve. Over time, many major hypotheses have been developed as potential explanations for this phenomenon. For example, the "enemy release hypothesis" states that the absence of a species' natural enemies in the exotic range can be a cause of invasion success. Other major hypotheses are more concerned with the conditions under which

[1] https://github.com/inas-argumentation/inas-abstracts

an introduced, non-native species will be able to establish amongst the native species as, e.g., the "biotic resistance hypothesis" stating that an ecosystem with high biodiversity is more resistant against non-native species than an ecosystem with lower biodiversity.

Many empirical studies in invasion biology aim to test such major hypotheses. In order to do this, researchers have to decide on a specific study system (i.e., focal organisms and habitat) and a research method (e.g., observational survey, lab experiment), and they often also have to choose which specific aspect of the hypothesis they address. In the case of the enemy release hypothesis, one group of empirical studies tests whether invasive species actually are released from their enemies and a second group studies whether invaders show enhanced performance if they are released from enemies. Each of these groups can be further subdivided into studies focusing on specialist enemies (i.e., species only preying on specific other species) or generalist enemies (i.e., enemies without specific preferences, e.g., slugs). All these decisions progressively instantiate more general concepts from the main hypothesis until a concrete, testable sub-hypothesis is reached.

Jeschke and Heger (2018) identified these specific instantiations of the main hypotheses as well as the underlying decision process and organized them in a hierarchical hypothesis network based on the Hierarchy-of-Hypotheses (HoH) approach (Jeschke et al., 2012; Heger et al., 2013, 2021). Therefore, each node in the hierarchy represents a hypothesis at a certain level of abstraction while links to nodes on higher/lower levels connect each hypothesis to its more abstract or more specific versions, respectively. The underlying decision process of replacing abstract components of hypotheses by more specific instantiations thereby induces a tree structure, meaning that each node can have several child nodes but at most one parent node.

In (Jeschke and Heger, 2018), ten out of the 12 main hypotheses were depicted as such hierarchies of hypotheses, and a large literature survey was conducted to quantify the level of empirical support for each of them. In this process, a list of papers for each main hypothesis was collected, with each paper being annotated with the necessary information to correctly place it in the hierarchy, so that a group of empirical studies that address the specific hypothesis can be linked to each node in the hierarchy. A visualization of the hierarchical hypothesis network, as well as the underlying data, are available[2] (see Figure 1).

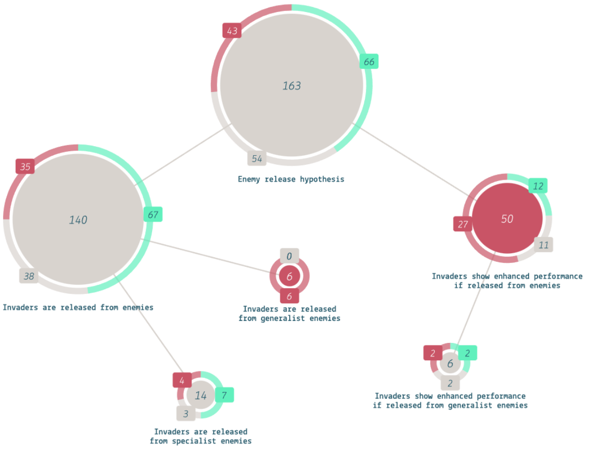


Figure 1: The sub-hypothesis structure for the enemy release hypothesis, one of the ten main hypotheses.

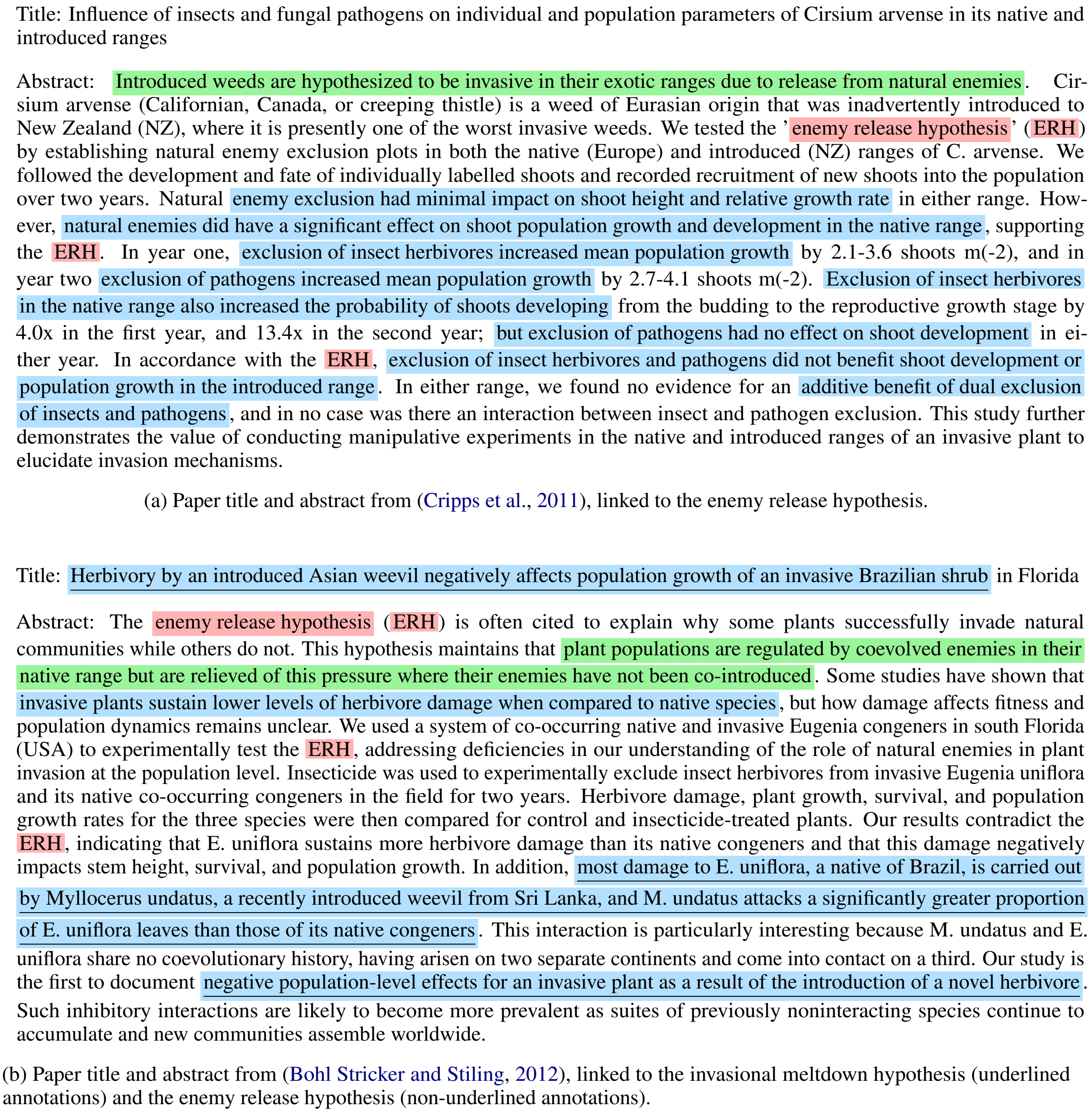
3.2 Dataset for Hypothesis Detection

The basis for the INAS dataset is a collection of Excel files (one for each main hypothesis) containing paper titles from the field of invasion biology in combination with further information about each paper. Since this data is not easily accessible for automatic processing, we extracted the paper titles as well as the information needed to determine the placement of the papers in the hierarchical hypothesis network from the Excel files and subsequently used a web scraper to obtain the corresponding abstracts. This was possible for 954 samples, leading to the final dataset of 954 paper titles, abstracts, and hierarchical hypothesis labels. The dataset also includes written statements of all hypotheses from the hypothesis network to provide the option of introducing general information about the hypotheses in different prediction settings.

Since the basis for this dataset are scientific paper titles and abstracts it is not possible to publish all texts from this dataset due to copyright. Instead, we release the paper titles with corresponding DOIs and links to the websites the papers are published on to allow for easy automated scraping of the necessary data.

Figure 2 shows two example abstracts, one of which is linked to the enemy release hypothesis (Figure 2a) while the second abstract is linked to two hypotheses in the network (Figure 2b).

[2] https://hi-knowledge.org/invasion-biology/



Title: Influence of insects and fungal pathogens on individual and population parameters of Cirsium arvense in its native and introduced ranges

Abstract: Introduced weeds are hypothesized to be invasive in their exotic ranges due to release from natural enemies. Cirsium arvense (Californian, Canada, or creeping thistle) is a weed of Eurasian origin that was inadvertently introduced to New Zealand (NZ), where it is presently one of the worst invasive weeds. We tested the ’enemy release hypothesis’ (ERH) by establishing natural enemy exclusion plots in both the native (Europe) and introduced (NZ) ranges of C. arvense. We followed the development and fate of individually labelled shoots and recorded recruitment of new shoots into the population over two years. Natural enemy exclusion had minimal impact on shoot height and relative growth rate in either range. However, natural enemies did have a significant effect on shoot population growth and development in the native range, supporting the ERH. In year one, exclusion of insect herbivores increased mean population growth by 2.1-3.6 shoots m(-2), and in year two exclusion of pathogens increased mean population growth by 2.7-4.1 shoots m(-2). Exclusion of insect herbivores in the native range also increased the probability of shoots developing from the budding to the reproductive growth stage by 4.0x in the first year, and 13.4x in the second year; but exclusion of pathogens had no effect on shoot development in either year. In accordance with the ERH, exclusion of insect herbivores and pathogens did not benefit shoot development or population growth in the introduced range. In either range, we found no evidence for an additive benefit of dual exclusion of insects and pathogens, and in no case was there an interaction between insect and pathogen exclusion. This study further demonstrates the value of conducting manipulative experiments in the native and introduced ranges of an invasive plant to elucidate invasion mechanisms.

(a) Paper title and abstract from (Cripps et al., 2011), linked to the enemy release hypothesis.

Title: Herbivory by an introduced Asian weevil negatively affects population growth of an invasive Brazilian shrub in Florida

Abstract: The enemy release hypothesis (ERH) is often cited to explain why some plants successfully invade natural communities while others do not. This hypothesis maintains that plant populations are regulated by coevolved enemies in their native range but are relieved of this pressure where their enemies have not been co-introduced. Some studies have shown that invasive plants sustain lower levels of herbivore damage when compared to native species, but how damage affects fitness and population dynamics remains unclear. We used a system of co-occurring native and invasive Eugenia congeners in south Florida (USA) to experimentally test the ERH, addressing deficiencies in our understanding of the role of natural enemies in plant invasion at the population level. Insecticide was used to experimentally exclude insect herbivores from invasive Eugenia uniflora and its native co-occurring congeners in the field for two years. Herbivore damage, plant growth, survival, and population growth rates for the three species were then compared for control and insecticide-treated plants. Our results contradict the ERH, indicating that E. uniflora sustains more herbivore damage than its native congeners and that this damage negatively impacts stem height, survival, and population growth. In addition, most damage to E. uniflora, a native of Brazil, is carried out by Myllocerus undatus, a recently introduced weevil from Sri Lanka, and M. undatus attacks a significantly greater proportion of E. uniflora leaves than those of its native congeners. This interaction is particularly interesting because M. undatus and E. uniflora share no coevolutionary history, having arisen on two separate continents and come into contact on a third. Our study is the first to document negative population-level effects for an invasive plant as a result of the introduction of a novel herbivore. Such inhibitory interactions are likely to become more prevalent as suites of previously noninteracting species continue to accumulate and new communities assemble worldwide.

(b) Paper title and abstract from (Bohl Stricker and Stiling, 2012), linked to the invasional meltdown hypothesis (underlined annotations) and the enemy release hypothesis (non-underlined annotations).

Figure 2: Two abstracts from the INAS dataset, annotated with explicit (green) and implicit (blue) hypothesis statements, and hypothesis names (red). The first example is classified correctly by all trained classifiers (Section 4). In the second example, the enemy release hypothesis is always classified correctly again, while the invasional meltdown hypothesis is only recognized by one out of ten trained classifiers (BioBERT base).

3.3 Dataset Analysis

Scientific abstracts are usually short and concise, which is also the case in the INAS dataset: On average, an abstract from the dataset consists of 10.26 sentences, with only 3.1% of samples surpassing the usual limit of 510 tokens for BERT models if the concatenation of paper title and abstract are tokenized using a standard BERT tokenizer. The class distribution among the ten main hypotheses is uneven, mirroring the true distribution of papers addressing the different hypotheses in the literature: The most dominant class contains about 21.8% of the samples (Invasional meltdown hypothesis) while about 1.8% of samples are assigned the most infrequent class (Island susceptibility hypothesis). This uneven distribution is even more pronounced among the sub-hypotheses, with some being assigned only a single sample while the most frequent hypothesis on the lowest level is addressed by 6.8% of papers. Importantly, every paper can address multiple (sub-)hypotheses (5.5% of samples address two main hypotheses) and can also be only assigned to hypotheses that are not on the lowest level in the hierarchy if non of the hypotheses on the next lower level matches the research conducted in it.

3.4 Hypothesis Statements

Since the hypothesis labels for the INAS dataset were created based on the full-text papers, it is unclear whether the titles and abstracts contain enough information to correctly identify every hypothesis that the corresponding papers address. Additionally, different ways of conveying hypothesis information can be more challenging to recognize, with domain knowledge being required regularly. Both these factors potentially affect the performance of automatic hypothesis identification models (compare Section 4), so that gaining insight into the typical ways that hypothesis information is stated in these abstracts is a mandatory basis for many analyses. To this end, together with a domain expert from invasion biology, we carried out a qualitative analysis of hypothesis statements and formulations within abstracts in the INAS dataset. We observe that hypothesis statements are extremely varied, ranging from explicit statements of hypothesis names in the case of some of the most well-known hypotheses to implicit hypothesis statements through, e.g., descriptions of experiments. In this initial analysis, we identified the following types of hypothesis statements:

Hypothesis name Explicit mentions of the hypotheses by their name (see text spans marked in red in Figure 2a). Some hypotheses are named after the main concepts they represent (e.g., *biotic resistance hypothesis*), a mention of these concepts provides almost the same information as an explicit hypothesis name and is therefore also annotated.

Explicit hypothesis statement Sentences stating the general hypothesis addressed in the paper, but without naming it (see green text span in Figure 2b).

Hypothesis fragment Spans of text that contain important parts of the hypothesis that is addressed in the paper but that do not belong to a complete hypothesis statement.

Implicit hypothesis statement Spans of text that reveal the hypothesis that is addressed in the paper without actually formulating it (e.g., descriptions of experiments, see blue text spans in Figure 2a and 2b).

Tag Type	Title	Abstract	Both
Name	.10/0.10	.30/0.64	.34/0.74
Statement	0/0	.42/0.58	.42/0.58
Fragment	.24/0.30	.56/1.08	.60/1.38
Implicit	.28/0.28	.80/1.86	.80/2.14
All	.62/0.68	.96/4.16	.96/4.84

Table 1: Distribution of the different tags in our subset of 50 annotated samples, broken down into presence of the tags in the titles, abstracts, or both (titles and abstracts combined). The statistics provided are the fraction of texts containing the specific tag at least once as well as the average number of annotated spans of the tag per text.

These different types of hypothesis statements we observed correspond to different types of tasks addressed in existing work on IE. While hypothesis names would be covered by NER schemes and systems (though existing NER schemes in biodiversity do not include them), explicit hypothesis statements are more similar to claims annotated in argument mining (Fergadis et al., 2021). Implicit claims are not well covered by both approaches, except in Blake (2010)'s study on claim formulations. Interestingly, the qualitative examples in Figure 2 suggest that implicit hypothesis statements are the most frequent, an observation that will be supported by data analyses that follow.

We conduct a pilot study to evaluate the presence of different types of hypothesis statements in scientific titles and abstracts from the field of invasion biology. To do this, we asked an expert annotator who was familiar with (Jeschke and Heger, 2018)'s hypothesis network to annotate a set of 50 titles and abstracts from the test set of the INAS dataset on span-level with the statement types introduced in Section 3.4. The set of annotated samples allows us insight into several interesting properties of the distribution of information about hypotheses in the dataset: Even though every paper addresses at least one hypothesis from the network, only 42% of titles and abstracts contain an actual hypothesis statement, while 34% state the name of the hypothesis. Accounting for the overlap in these groups, only 56% of samples provide concrete information about the hypothesis that the paper addresses in the title or abstract. Instead, authors often rely on hypothesis fragments (60% of samples) or implicit hypothesis statements (80% of samples) to make clear which hypothesis is addressed in their work. A detailed breakdown of the distribution of hypoth-

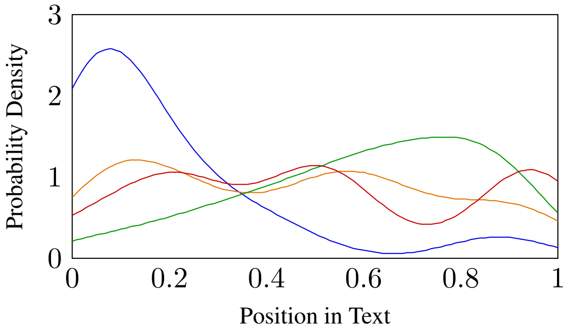


Figure 3: Empirical probability density function of the positionings of hypothesis statements (blue), hypothesis names (red), hypothesis fragments (orange), and implicit hypothesis statements (green) in the abstracts, created using a kernel density estimator using a Gaussian kernel (bandwidth=0.1).

esis information in the titles and abstracts is given by Table 1, also including the average number of annotated spans of a certain tag in the dataset. The averages of 1.38 hypothesis fragments and 2.14 implicit hypothesis statements per text as well as the average of 4.84 annotated spans of all classes per text here clearly indicate that the information about the hypothesis addressed in a paper can be seldom found in a single sentence: Instead, information from different parts of the text needs to be used for correct identification of the hypothesis.

Additional interesting patterns arise if we analyze the likelihood of specific tags being located at different positions in the abstract. To do this, we define the position of an annotated span as the average token index of all tokens in the span divided by the total number of tokens in the text, resulting in positions in the range $[0, 1]$. We can then plot the empirical probability density function (created using a kernel density estimator) for each tag, as is done in Figure 3. While hypothesis names and hypothesis fragments have a rather uniform probability of appearing at any position in the text, hypothesis statements are made mainly at the beginning, while implicit hypothesis statements are more likely to be made later in the abstract. The reasons for this are that abstracts regularly begin with an explicit description of the hypothesis while ending with details about experiments and observations, which often fall in the category of implicit hypothesis statements.

3.5 Discussion

Current datasets labelling claims in scientific texts mostly focus on a precise span-level annotation instead of providing detailed semantic labels (see Section 2). While studies addressing also the semantic content of claims exist, the claims often address a variety of very distinct topics that can often be easily differentiated by non-experts as, e.g., claims addressing residency vs. claims addressing foreign policy in DebateNet-mig15 (Lapesa et al., 2020). This stands in stark contrast to the INAS dataset, where all of the hypotheses in the hierarchy address the same phenomenon of invasive species being successful in a new domain, which already is a rather narrow subfield of general biology. Therefore, even with respect to the highest level of the hierarchy, the correct identification of the hypothesis addressed in an abstract is a very challenging problem that requires expert knowledge, with many lower levels in the hierarchy representing even more subtle differences that are harder to distinguish. We argue that researchers in the scientific domain will benefit most from tools differentiating on such a precise level because subtle semantic information about the hypothesis addressed in a paper can be of high importance in judging the relation between scientific studies or the relevance with respect to a search query. Therefore, the INAS dataset adds a new and important facet to the general landscape of datasets on IE for scientific text. At the same time, the fine-graininess of the hypothesis network combined with the varied nature of hypothesis formulations in abstracts (Section 3.4) creates challenges for fully supervised labeling of the dataset. A complete annotation of hypothesis statements linked to the network would require experts familiar with the domain as well as linguistic aspects of annotation (which is very unrealistic). For this reason, we now explore whether "weak" abstract-level hypothesis labels in the current dataset provide useful information for state-of-the-art NLP models.

Model	**F1(S)**	**F1(M)**
Naive Bayes	.702	-
BERT base	.665 (±.047)	.659 (±.051)
BERT large	.670 (±.045)	.674 (±.032)
BioBERT base	**.758** (±.025)	.751 (±.033)
BioBERT large	.734 (±.020)	.731 (±.065)
PubMedBERT base	**.758** (±.027)	.757 (±.026)

Table 2: Classification F1 scores for all models tested in our study. F1(S) denotes the F1 scores in the single label classification setting while F1(M) refers to the multi-label classification setting.

4 Hypothesis Identification as Abstract Classification

In this Section, we report baseline experiments on modeling the automatic identification of hypotheses in the INAS dataset.

4.1 Experimental Set-up

We frame hypothesis detection as a classification problem where the input is the concatenation of title and abstract of a paper and the output is a label of the major hypotheses that are addressed in the corresponding paper, with major hypotheses meaning the ten hypotheses on the highest level of the hierarchical hypothesis network.

We test different models that allow us to gain insight into different properties of the dataset: On the one hand, we test the performance of a naive Bayes classifier working on unigrams after removing stop words and highly frequent/infrequent words, allowing us to explore how much simple word frequency statistics already reveal about the hypothesis that is addressed in a paper. On the other hand, we test more complex neural classifiers in the form of standard BERT classifiers (Devlin et al., 2019) (base and large) as well as BERT classifiers trained on texts from a domain that presumably more closely resembles the domain of invasion biology abstracts: BioBERT models (Lee et al., 2019) (base and large) and the PubMedBERT model (Gu et al., 2022) (base), all trained on scientific abstracts and full-text papers from the biomedical domain. The training is done on a training set comprising 75% of the samples from the dataset, evaluation and testing are done on subsets containing 10% and 15% of the samples, respectively.

Due to the fact that a single paper can address multiple hypotheses, the classification is a multi-label classification problem. The naive Bayes classifier is only applicable to single-label classification, though, so we train it by inserting the samples with multiple labels repeatedly into the training set, once with each label. We proceed in the same way for the test and validation splits, meaning that the classifier will not be able to achieve perfect accuracy. To be able to compare the results, we test the BERT models in the same single-label setting (using a softmax classification layer). Additionally, we test the BERT classifiers in the multi-label setting by predicting an individual probability for each class. In this case, we still force the classifier to predict at least one positive label for each sample since this lead to increased performance. For all BERT classifiers, we reduce the effect of variance during training on our results by training ten classifiers for each model type and classification setting and report the average macro F1 score as well as the standard deviation.

4.2 Results

Table 2 displays the classification results in terms of the macro F1 score for both the single-label and the multi-label classification setting.

Notably, the naive Bayes classifier performs reasonably well and even outperforms the standard BERT classifiers, indicating that simple word frequency statistics provide significant information about the correct label. An analysis of the naive Bayes classifier weights revealed that hypothesis-specific concepts, as well as parts of the hypothesis names, were strong indicators for the specific classes, but also some species and country names that mostly appear in the context of specific hypotheses were used as a basis for the classification. The advantage of the naive Bayes classifier compared to the BERT classifiers might originate in the fact that many domain-specific terms might be unknown to the BERT models and the small training set might not be enough to fully learn these new concepts.

The classifiers based on variants of BERT that are adapted to texts from the biomedical domain consistently outperformed the naive Bayes classifier, which is consistent with earlier results that show that in-domain fine-tuning generally leads to improved performance (Gururangan et al., 2020). Notably, especially the smaller BERT$_{base}$ models show better performance as well as reduced variance, making them the best performing models in our study. We also observe that the ability to do multi-label predictions generally does not yield an improvement, which can be explained by the small number of cases where multi-label prediction is necessary.

Even though the BioBERT and PubMedBERT models show increased performance compared to the naive Bayes classifier, the difference appears to be moderate considering the large difference in complexity. All BERT models should be able to process the same word frequency information as the naive Bayes classifier, meaning that their ability to combine the information from different words and sentences is only responsible for a 7%

performance increase. We believe that this indicates that the BERT classifiers are not able to understand the full semantic content of hypothesis statements, especially if they are only made implicitly. Instead, the increase in performance might simply be caused by the classifier's ability to detect slightly more complex patterns than unigrams (e.g., n-grams) and by its ability to nonlinearly combine the information about the presence of these still simple patterns.

4.3 Ablation Study

We use the domain expert annotations from 3.3 to evaluate which kinds of information are most important for the neural network classifiers. To test this, we perform an ablation study in which we train a classifier (BioBERT base) on: (i) only the title, (ii) the first two sentences from the abstract, or (iii) the last two sentences from the abstract.

The evaluation of the ablated classifiers on the test set yielded an F1 score of 0.61 for the titles and an equal score of 0.53 for the first two and for the last two sentences. Therefore, the title contains on average more information that is useful for the classification, which is to be expected since a good title should clearly indicate the key aspects of the underlying study while it is not necessary that every sentence in the abstract has the same density of information. The equal performance on the first and the last sentences from the abstract is more surprising since it implies that the different types of information that are commonly found at these positions (hypothesis statement vs. implicit hypothesis statement) seem to be equally useful for the classification.

An alternative explanation for this result is that the human annotations do not generally correspond to information that is used by the neural network classifier. To explore this hypothesis, we divide the 50 annotated samples into 10 folds, in a way that, beginning from fold one, each fold progressively contains samples that contain more annotated spans and thus contain more information about the hypotheses according to our annotation. We then measure the performance of BioBERT base on each of these folds and plot the average number of annotations in each fold against the micro F1 score the model achieved on that fold (see Figure 4). To better see the correlation, we also fit a kernel regression model (Nadaraya, 1964; Watson, 1964) to the data, resulting in a clearly visible positive correlation between the number of annotated spans in a sample and the classification performance of the neural classifier. This correlation is mainly caused by two low-scoring batches that both have a low number of annotated spans, which means that samples with few annotated spans have an increased probability of being misclassified while the probability stays relatively constant for samples that have at least four annotated spans. This indicates that our annotations correspond to useful information for the classification and therefore indicates that the general annotation scheme that allows for a distributed annotation of hypotheses is reasonable.

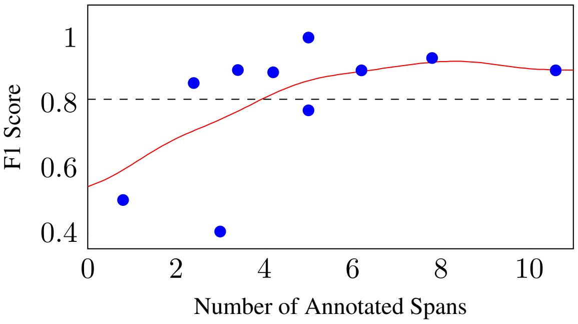


Figure 4: Classification micro F1 score vs. number of annotated spans for ten folds from the test set. The data was split into ten folds so that, beginning from fold one, each fold progressively contains samples with more annotated spans. The dashed line indicates the F1 score on all 50 samples, the red line is fitted to the data via a kernel regression model (Gaussian kernel with bandwidth=2.5).

In combination with the fact that the distributed annotations also correspond to the intuition of the domain expert, our study shows that the annotation of hypotheses and claims as single spans of text is limited and can be insufficient for certain domains like scientific texts. For this reason, our study shifts the focus from the simple, binary classification of sentences as claims to more fine-grained semantic categories, and at the same time, shifts the focus from detailed annotations of text spans to more general abstract- or paragraph-level annotation of hypotheses. We also note that the latter type of annotation may be more intuitive and faster for domain experts, which may not be trained linguistic annotators familiar with the complexities in text annotation.

5 Conclusion

In this work, we proposed and published the INAS dataset and conducted initial analyses and experiments on it. Our studies revealed interesting insights into the availability and distribution of information about the hypotheses in scientific paper titles and abstracts from the field of invasion biology. We believe that there is great potential for a variety of different studies to be performed using this dataset, some of which we plan on conducting in future work. These include further classification experiments like exploring the full hierarchical classification problem, trying to improve classification performance by conducting pretraining on full-texts from the field of invasion biology, or testing one-shot classification leveraging the written hypothesis descriptions. Further, our annotation experiment could enable studies on span-level hypothesis detection, e.g. in a weakly-supervised manner or in a one-shot classification setting. Finally, we also hypothesize that the introduction of human-engineered knowledge (e.g., in the form of ontologies) into, for example, the classification process can help overcome the problem of a lack of domain-knowledge of current language models.

References

Pablo Accuosto, Mariana Neves, and Horacio Saggion. 2021. Argumentation mining in scientific literature: from computational linguistics to biomedicine. Accepted: 2021-05-19T07:47:36Z Publisher: CEUR Workshop Proceedings.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2):173–189.

Kerry Bohl Stricker and Peter Stiling. 2012. Herbivory by an introduced asian weevil negatively affects population growth of an invasive brazilian shrub in florida. *Ecology*, 93:1902–11.

Michael Cripps, Graeme Bourdôt, David Saville, Hariet Hinz, Simon Fowler, and Grant Edwards. 2011. Influence of insects and fungal pathogens on individual and population parameters of cirsium arvense in its native and introduced ranges. *Biological Invasions - BIOL INVASIONS*, 13.

Mark A. Davis. 2009. *Invasion Biology*. Oxford University Press.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. *arXiv preprint arXiv:1704.07203*.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessieres, and Claire Nédellec. 2016. Overview of the bacteria biotope task at bionlp shared task 2016. In *Proceedings of the 4th BioNLP shared task workshop*, pages 12–22.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2022. *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Dublin, Ireland.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Number: arXiv:1810.04805 arXiv:1810.04805 [cs].

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Charles S. Elton. 1958. *The Ecology of Invasions by Animals and Plants*. Methuen & Co. Ltd.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A Multi-Layered Annotated Corpus of Scientific Papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3081–3088, Portorož, Slovenia. European Language Resources Association (ELRA).

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1–17.

Nancy Green. 2015. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23. ArXiv:2007.15779 [cs].

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Tina Heger, Carlos A Aguilar-Trigueros, Isabelle Bartram, Raul Rennó Braga, Gregory P Dietl, Martin Enders, David J Gibson, Lorena Gómez-Aparicio, Pierre Gras, Kurt Jax, Sophie Lokatis, Christopher J Lortie, Anne-Christine Mupepele, Stefan Schindler, Jostein Starrfelt, Alexis D Synodinos, and Jonathan M Jeschke. 2021. The hierarchy-of-hypotheses approach: A synthesis method for enhancing theory development in ecology and evolution. 71(4):337–349.

Tina Heger, Anna T Pahl, Zoltan Botta-Dukát, Francesca Gherardi, Christina Hoppe, Ivan Hoste, Kurt Jax, Leena Lindström, Pieter Boets, Sylvia Haider, et al. 2013. Conceptual frameworks and methods for advancing invasion ecology. *Ambio*, 42(5):527–540.

J. Jeschke, Lorena Gómez Aparicio, S. Haider, Tina Heger, C. Lortie, P. Pyšek, and D. Strayer. 2012. Support for major hypotheses in invasion biology is uneven and declining.

J. M. Jeschke and T. Heger. 2018. Invasion biology: hypotheses and evidence.

Gabriella Lapesa, Andre Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, Jonas Kuhn, and Sebastian Padó. 2020. DEbateNet-mig15:Tracing the 2015 Immigration Debate in Germany Over Time. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 919–927, Marseille, France. European Language Resources Association.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An Argument-Annotated Corpus of Scientific Publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proc. of IJCAI*.

Elisabeth A Lloyd. 1987. Confirmation of ecological and evolutionary models. *Biology and Philosophy*, 2(3):277–293.

Felicitas Löffler, Nora Abdelmageed, Samira Babalou, Pawandeep Kaur, and Birgitta König-Ries. 2020. Tag me if you can! semantic annotation of biodiversity metadata with the QEMP corpus and the BiodivTagger. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4557–4564, Marseille, France. European Language Resources Association.

Felicitas Löffler, Valentin Wesp, Birgitta König-Ries, and Friederike Klan. 2021. Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs? *PloS one*, 16(3):e0246099.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2021. Overview of drugprot biocreative vii track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

E. A. Nadaraya. 1964. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Nhung TH Nguyen, Roselyn S Gabud, and Sophia Ananiadou. 2019. Copious: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity data journal*, (7).

Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Geoffrey S. Watson. 1964. Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372. Publisher: Springer.