

MASTER'S THESIS - MSc NATURAL LANGUAGE PROCESSING

Analysing claims in NLP research

Author:

Clémentine BLEUZE

Supervisors:

Maxime AMBLARD
Fanny DUCEL
Karën FORT

Jury:

Maxime AMBLARD
Marianne CLAUSSEL
Claire GARDENT

Internship conducted in Loria, Department 4, Team Sémagramme, from 03/11/2024 to 08/30/2024 (25 weeks). This work benefited from an ANR funding, through the ORION program.



UNIVERSITÉ DE LORRAINE

Abstract

Institut des Sciences du Digital, Management et Cognition

MSc in Natural Language Processing

Analysing claims in NLP research

by Clémentine BLEUZE

In Natural Language Processing (NLP) as in science in general, important research findings are initially shared to the community as pre-prints, before undergoing a peer-review assessment and, possibly, end up as a publication in a journal or conference. Yet, even under the watchful eyes of peer-reviewers, some exaggerated claims may find their way into published articles, a phenomenon that has been referred to as "overclaiming". This work reports on an effort to define the notions of "claim", "overclaiming", and the related task of overclaiming detection in the particular context of NLP papers. This leads to the constitution of a large corpus of sentences from research pre-prints and articles, a subpart of which is manually annotated in terms of claim categories. It then serves as training data for a model which is finally ran on the entire corpus; in parallel, certainty annotations are collected using existing models. Finally, qualitative analyses on this corpus of claims are conducted. Results show that claims can be successfully extracted from NLP papers, which constitutes a promising first step for the automated study of research claim.

Acknowledgements

I would like to express my sincere gratitude to Karën Fort, Fanny Ducel and Maxime Amblard for their supervision; their recommendations were of a precious help, and I am thankful for the time they devoted to my internship. In particular, Fanny was of a precious help for the annotation of the corpus. Thanks should also go to Amandine Decker and Valentin Richard who participated in an annotation trial to validate our claims taxonomy.

Many thanks to all the people that offered their help, support or friendship during these 6 months: my office colleagues from B246 (now B122), and more generally all *Sémagramme* members. Special thanks also to the reviewers who helped to improve the present thesis, including my supervisors as well as my partner Jules Margaritta.

Lastly, I would like to thank the ORION program for its contribution to the funding of this research internship, which benefited from a French government grant managed by the Agence Nationale de la Recherche (reference: ANR-20-SFRI-0009).

Contents

Abstract	iii
Acknowledgements	v
1 Introduction	1
1.1 Context of the internship	1
1.2 The issue of <i>scientific overclaiming</i>	2
1.2.1 From dubious advertisement to deceitful science	2
1.2.2 What is at stake?	3
1.2.3 Stochastic parrots and the hype around NLP	4
1.3 An NLP-driven analysis of scientific papers	5
1.3.1 The NLP4NLP approach	5
1.3.2 Spin detection	5
1.3.3 A first analysis of claims in NLP papers	6
1.4 Contributions and Research Questions	7
2 Towards definitions of <i>claim</i> and <i>overclaiming</i>	9
2.1 Defining the task of overclaiming detection	9
2.1.1 Claims, truth and adequacy: different notions	9
2.1.2 ... but a similar modelisation	11
(i) Claim identification	11
(ii) Evidence retrieval	12
(iii) Verdict prediction	13
(iv) Justification production	14
Summary	14
2.2 Establishing a taxonomy for claim identification	15
2.2.1 Existing taxonomies	15
A taxonomy for contrarian claims about climate change	15
Argumentative Zoning	16
Abstract Sentence Classification	17
2.2.2 "Where there is overclaiming there are claims"	17
Setting the context	18
About the work itself: contributions and results	18
In hindsight: limitations, impact, directions	20
Investigating the storytelling of papers	21
Validating the taxonomy	22
3 Building an annotated corpus of NLP papers claims	25
3.1 Collecting sentences from NLP papers	25
3.1.1 Paper sources: ACL vs. ArXiv	25
3.1.2 Data collection	26
3.1.3 Data pre-processing and content extraction	27
3.1.4 Manual annotation	27

3.1.5	Model inference and linguistic analysis	27
3.1.6	Data distribution	27
3.2	Automatic classification of claims	28
3.2.1	Problem modelisation	28
3.2.2	Baseline models	29
3.2.3	BERT-based models	31
3.2.4	Custom models experiments	32
3.3	Automatic annotation in certainty	33
4	Analysis of collected claims	35
4.1	Inspecting the predictions of claim categories (RQ1)	35
4.1.1	Distribution of categories	35
4.1.2	Linguistic characteristics of claim categories	37
4.2	Analysing sentence and aspect-level certainty (RQ3)	38
4.2.1	Sentence-level certainty	38
4.2.2	Aspect-level certainty	40
4.3	Investigating the effects of papers metadata	41
4.3.1	Source: ACL vs. ArXiv (RQ3)	41
4.3.2	Year of publication (RQ4)	43
5	Analysis and Discussion	45
5.1	An investigation of claims in NLP papers	45
5.2	Limitations	47
5.3	Perspectives	48
Bibliography		51
A	About claim taxonomies	59
A.1	Existing taxonomies and annotation schemes	59
A.2	Annotation guidelines	60
A.2.1	Annotation guidelines for the 3rd annotation phase	61
A.2.2	Annotation guidelines for the 3rd annotation phase	62
A.2.3	Final annotation guidelines	63
Overview	63
The labels	63
The additionnal tags	64
Ambiguous cases	65
B	Analyses on claims	67
C	Supplementary materials	73

List of Figures

1.1 Evolution of Amazon Go media coverage. Left: a New York Times article covering the opening of Amazon Go stores in 2018. Right: a <i>Guardian</i> opinion paper reporting Amazon’s use of underpaid workers in 2024.	2
2.1 The fact-checking pipeline (Guo, Schlichtkrull, and Vlachos, 2022)	11
2.2 Modelling fact-checking, spin detection and exaggeration detection using the fact-checking pipeline of Guo, Schlichtkrull, and Vlachos, 2022: three examples.	12
2.3 Exaggeration detection strategy using difference of strength (certainty) between a source claim and a target claim. Figure from Patro and Baruah (2021).	13
2.4 Summary of our proposed modelisation of overclaiming detection. <i>AIC</i> stands for <i>Abstract, Introduction and Conclusion</i>	14
2.5 An extract of the taxonomy of (Coan et al., 2021)	16
2.6 Mapping of (quasi) equivalent concepts in multiple AZ frameworks (not comprising AZ-II by Teufel, Siddharthan, and Batchelor, 2009), in Table 10 in Schrader et al. (2023).	17
2.7 Simplified illustration of the CARS model by Swales (1981), as presented in Labaree (2024) (self-made illustration).	18
2.8 Results of M2M-100 (Fan et al., 2021) on Breton to French translation, reported in Jouitteau and Grobol (2024). BLEU(\uparrow) and ChrF++(\uparrow) range between 0-100, while TER(\downarrow) ranges in $[0, +\infty]$. They also note that in the original article, no evaluation score for Breton is given.	19
2.9 Photography of a jar containing gumballs, in Figure 3 in Rosenberg, Willcox, and Schumann (2023).	20
2.10 A simple model for NLP papers structure representation. Above: we call <i>body</i> of the paper all the parts between the introduction and the conclusion. Below, we call <i>IRC papers</i> those for which we can retrieve an introduction, results, and a conclusion section using the listed keywords (pattern-matching in a paper’s section titles).	22
3.1 Distribution of labels collected during the annotation phase.	28
3.2 Radar plot of F1-scores per model setting and per category. The more consistent a model performance remains across classes, the closer its associated shape should get to a regular polygon. We observe that all models struggle to predict less frequent classes of the corpus <i>limitation</i> , <i>directions</i> and <i>impact</i> (scores below 0.50).	31
3.3 Radar plot of F1-scores per model setting and per category. The more consistent a model performance remains across classes, the closer its associated shape should get to a regular polygon. We observe a coherent performance on most populated classes, but divergences on less-populated ones (in particular for <i>impact</i> and <i>limitation</i>).	32

3.4	A custom BERT-based model for contextualised predictions of claim classes (represented in the <i>prefix_SEC_1r</i> setting). At training time, the model sees all three ground-truth labels of the sentences and learns to predict them all. At inference time, we only retrieve the prediction that corresponds to the target sentence.	33
3.5	Certainty is a multi-dimensional construct. Illustration from (Pei and Jurgens, 2021)	34
4.1	Distribution of claim labels in the entire corpus (left) and in the manually annotated dataset (right). The entire corpus contains both IRC and non-IRC papers, whereas the annotated one contains only IRC papers; this could explain the increased prevalence of <code>no claim</code> in the first case.	36
4.2	Co-occurrence matrix for label pairs. For instance, line "impact" indicates that 56% of <code>impact</code> labels occur alone, while 12% of them co-occur with <code>contribution</code> . By definition, no label ever co-occurs with <code>no claim</code>	38
4.3	Average number of characters and tokens per sentence, per claim category (single-labels only).	39
4.4	Wordclouds for classes <code>contribution</code> and <code>directions</code>	39
4.5	Boxplot for sentence-level certainty in the entire claims corpus. Sentence-level certainty takes its values in [1,6].	39
4.6	Relation between average sentence-level certainty and aspect certainty. To be compared with Figure 3 in (Pei and Jurgens, 2021).	40
4.7	Distribution of modalities (Absent / Certain or Uncertain) over the entire corpus, per aspect.	41
4.8	Modality (Absent / Certain or Uncertain) distributions over the entire corpus, per aspect and claim category.	42
4.9	Distribution of claims according to their source (ACL vs. ArXiv) and category. Overall, 41.1% (resp. 58.9%) of extracted sentences are from ArXiv (resp. ACL) papers.	43
4.10	Boxplots of sentence-level certainty in sentences, per claim category and source. Green bars represent median values, boxes represent [Q1-Q3] ranges, and whiskers are 1.5*IQR long (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html).	43
4.11	Number of predicted claims/paper (y-axis) in function of the year of publication (x-axis) and paper source (ACL vs. ArXiv).	44
4.12	Shares of claims (%) per category in papers throughout time.	44
4.13	Occurrences of claims (%) per category in papers throughout time.	44
5.1	Histogram of sentence-level certainty values (x-axis) in the entire corpus. Red dotted lines divide the distribution in three zones: low, moderate, and high certainty.	46
5.2	Sentence-level certainty values depending on the relative position of sentences in papers (averaged scores in 10 classes). Note that the range of values is very small (y-axis).	46
A.1	Taxonomy of contrarian claims about climate change, in Coan et al. (2021). More details can be found in https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-021-01714-4/MediaObjects/41598_2021_1714_MOESM1_ESM.pdf	59

A.2	The initial AZ annotation scheme from (Teufel, Carletta, and Moens, 1999). It classifies a paper's claims according to their source: OWN, OTHER or BACKGROUND; their relationship to existing work: BASIS or CONTRAST; and their rhetorical role in the presentation of the paper: AIM or TEXTUAL. "KC" stands for "knowledge claim".	60
A.3	The AZ-II annotation scheme from (Teufel, Siddharthan, and Batchelor, 2009).	60
B.1	Coincidence ratios for all label pairs (3 rd annotation phase). For instance, line "limitation" means that 25% of all limitation labels assigns by one or the other annotator coincided with a directions annotation by the other.	67
B.2	Average counts of part-of-speech tags in one sentence, per claim category.	68
B.3	Average counts of named entities in one sentence, per claim category. .	69
B.5	Evolution of the number of sentences/paper throughout time, per source of publication.	70
B.4	Top 10 of most frequent lemmas per claim category.	70
B.6	Evolution of the ratio of claims/sentences in papers throughout time, per source of publication.	71
B.7	Wordclouds for every claim category (from sentences with a unique label).	72
C.1	Overview of the model of Cohan et al. (2019). When dealing with sequential inputs, the model retrieves all sentence separators representations to pass them into Multi-Layer Perceptrons for classification.	73
C.2	An example chosen by OpenAI to illustrate the superiority of InstructGPT on GPT-3 in answering instructions. Caption taken from https://openai.com/index/instruction-following/ in May 2024. Note the lower comment: "GPT-3 models aren't trained to follow user instructions. Our InstructGPT models (highlighted) generate much more helpful outputs in response to user instructions".	74

List of Tables

2.1	Final version of the taxonomy for claim category identification. <i>AIC</i> stands for <i>Abstract, Introduction and Conclusion</i>	22
2.2	Statistics on each annotation phase. α denotes Krippendorff's alpha, κ denotes Cohen's kappa min-max range (when there are more than two annotators). *In the first annotation phase, two annotators worked on 10 papers, while the two others worked on 4 papers (included in the first two's 10).	24
3.1	Three information encoding configurations for baseline models.	30
3.2	F1 scores for claim classification per model setting (lines) and claim category (columns 1-8). lr = logistic regression , cv = count-vectorisation, tf = tf-idf vectorisation. avg = weighted average of all F1-scores.	30
3.3	Three context encoding configurations for <i>BERT</i> -based models.	31
3.4	F1 scores for claim classification per model setting (lines) and claim category (columns 1-8). avg = weighted average of all F1-scores.	32
3.5	Definitions of aspects, based on Pei and Jurgens (2021).	34
4.1	Number of predicted claims per category (lines) and per number of predicted labels (columns). For instance, in line 2, column 3: 3,555 claims were assigned the label "result" + two additional ones (which represents 0.1% of all "result" predictions). Percentages in column "total" are relative to the total number of predicted labels.	35
4.2	Random selection of claims for each category (model predictions) . . .	36
4.3	Average share of claim categories in papers, and occurrence frequencies in the corpus. For instance: on average, <i>context</i> claims represent 10.0% of a papers' sentences, while 99.0% of papers in the corpus have at least one claim labelled as <i>context</i>	37
4.4	Random selection of claims for most frequent label pairs (model predictions). Further discussion shall establish whether labels impact (line 1) and result (line 2) are justified. In line 3, a sentence segmentation issue explains the (nevertheless valid) pair of labels.	37
4.5	Example of claims with minimal, intermediate, and maximal sentence-certainty scores.	40
4.6	Random selection of claims where each aspect is present. In bold, we show elements which, <i>we believe</i> , influenced the model's predictions. In line 1, we note that the sentence barely even makes sense on its own.	41
B.1	Average shares (%) of categories in papers, per paper source. Confidence intervals were computed with $\alpha = 0.95$, using the large sample confidence interval estimation (see https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2022/mit18_05_s22_class23-prep-b.pdf , p.4.	71

B.2	Average occurrences (%) of categories in papers, per paper source. Confidence intervals were computed with $\alpha = 0.95$, using the large sample confidence interval estimation (see https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2022/mit18_05_s22_class23-prep-b.pdf , p.4.	71
C.1	List of Named Entities supported by spacy with definitions (https://github.com/explosion/spaCy/discussions/9147).	73

Chapter 1

Introduction

In this first chapter, we start by explaining the context of the internship this thesis is the result of. Second, we provide some insight on the phenomenon of overclaiming which motivates our investigation. Then, we bring some background for a NLP-oriented approach of the matter by considering works on scholarly datasets, the related task of spin detection, and the contributions of Ducel (2022). Finally, this leads us to a clearer problem statement, from which we draw research directions for the present work.

1.1 Context of the internship

This thesis reports on a work conducted during a 6-month research internship at the Loria (Laboratoire Lorrain de Recherche en Informatique et ses Applications) from March 11, 2024 to August 30, 2024 (25 weeks). It benefited from an ANR¹ funding granted by the ORION program within the scope of an excellence scholarship plan for the schoolyear 2023-2024. This internship constitutes the author's final semester of Natural Language Processing master's degree in IDMC² (Institut des Sciences du Digital, Management et Cognition), Nancy. Its subject has been proposed as a continuation (and extension) of Fanny Ducel's M1 thesis entitled "Analyse des claims dans les articles de TAL en apprentissage non supervisé" (Ducel, 2022), which was supervised by Karën Fort. During her internship, the author was supervised by Karën Fort, Fanny Ducel and Maxime Amblard in the team *Sémagramme*; Fanny Ducel in particular largely contributed to the annotation phase (cf. Section 3.1.4), and took an important part in the discussions to refine our claims taxonomy (cf. Section 2.2).

The Loria³ (Laboratoire Lorrain de Recherche en Informatique et ses Applications) is a research unit dealing with fundamental and applied research in computer sciences since its official creation in 1997; this unit is common to other French institutions: CNRS⁴ (Centre National de Recherche Scientifique), INRIA⁵ (Institut National de Recherche en sciences et technologies du numérique) and University of Lorraine⁶. The Loria is structured into 5 departments and 27 teams. The team *Sémagramme*⁷, whose interests include (but are not limited to) the study of Syntax-Semantics interface, Discourse dynamics and Ethics in NLP, belongs to Department 4: *Natural Language Processing and Knowledge Discovery*.

¹Agence Nationale de la Recherche.

²See <https://idmc.univ-lorraine.fr/>.

³See <https://www.loria.fr>.

⁴See <https://www.cnrs.fr/en/the-cnrs>.

⁵See <https://www.inria.fr>.

⁶See <https://www.univ-lorraine.fr>.

⁷See <https://team.inria.fr/semagramme/>.

1.2 The issue of scientific overclaiming

According to the Merriam-Webster dictionary⁸, "to overclaim" means "to claim too much of (something, such as an expense)", or "to make a false or exaggerated claim about (something)". Following this definition, *scientific overclaiming* is naturally understood as a particular form of *overclaiming* which applies to scientific declarations. This appears at the very least antinomic, given that scientific discourse is generally associated with qualities of clarity, accuracy and objectivity⁹. In fields benefiting from a wide coverage, we would expect overclaiming to be particularly prohibited, as important claims may be widely shared (and, potentially, further amplified) to the community, intermediate publishers and eventually the general public or mainstream media. Yet, cases of scientific overclaiming are not rare, and are in fact a common practice in industrial advertising campaigns.

1.2.1 From dubious advertisement to deceitful science

Consider for instance Amazon's store chain *Amazon Go*, which opened to the public in 2018 and advertised its *Just Walk Out* feature for a cashierless, automated experience of shopping, provided by "the most advanced machine learning, computer vision and AI", "much like you'd find in self-driving cars"¹⁰. How does it work? Simple, Amazon says: "Anything you take off the shelf is automatically added to your virtual cart and anything you put back on the shelf is removed from your virtual cart. When you're done shopping, simply leave the store and the technology will automatically charge you for only the items you take with you."¹¹. This seducing story was however challenged as it later appeared that "the technology" used for charging customers was very much reliant on an estimated 1,000 workers based in India who had to manually check 70% of 2022 sales to avoid mistakes¹².

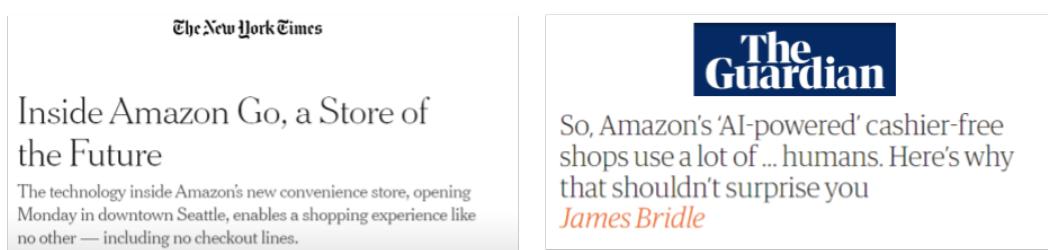


FIGURE 1.1: Evolution of Amazon Go media coverage. Left: a *New York Times* article covering the opening of Amazon Go stores in 2018. Right: a *Guardian* opinion paper reporting Amazon's use of underpaid workers in 2024.

Due to additional technical complications, many of these stores were progressively closed¹³, and the initial hype for the "Store of the Future" slowly subsided (see

⁸See <https://www.merriam-webster.com/dictionary/overclaim>.

⁹See Huguet et al. (2018) for a prescriptive analysis of the qualities of *scientific language* (as opposed to *literary language*). In particular, a text's ideas must be "set out in such a way as to avoid misinterpretations and only to imply what the author means", "using direct language, without ambiguous terms or confusing or misleading expressions".

¹⁰See <https://www.youtube.com/watch?v=NrmMk1Myrxc>.

¹¹See <https://www.amazon.com/b?node=16008589011>.

¹²See <https://www.businessinsider.com/amazons-just-walk-out-actually-1-000-people-in-india-2024-4>

¹³See <https://edition.cnn.com/2023/03/06/business/amazon-closing-some-amazon-go-stores/index.html>

Figure 1.1). We cannot help but notice that this goes beyond a simple commercial failure: Amazon clearly overclaimed the power of its innovation, with co-inventor of the technology Dilip Kumar confidently affirming that "[t]he system is very, very accurate"¹⁴, which eventually ended up deceiving customers while silencing reliance on (probably underpaid)¹⁵ microworkers.

But scientific overclaiming is not the preserve of industrials. In academic research, cases of deceiving data manipulation and falsification to boost results have also been reported. In France, CNRS researchers who were accused of manipulating their results were sanctioned¹⁶, and investigations on Didier Raoult's work on hydroxychloroquine¹⁷ (a so-called remedy to COVID-19) slowly led to his downfall. Even in medical research, McGrath et al. (2017) report that 72% (resp. 69%) of 112 investigated systematic reviews of diagnostic accuracy studies contained at least one form of over-interpretation in their abstract (resp. full-text) with regards to the actual reported results, a phenomenon that the field refers to as *spin* (see for instance Koroleva, 2020).

1.2.2 What is at stake?

All above-mentioned examples raise serious ethical issues regarding scientific integrity and, more generally, science credibility, in a context where trust in science is not as strong as it used to be. Recently, a study conducted by the University of Lorraine showed that in the context of the COVID-19 pandemic, only 47% of French respondents (resp. 32% and 30%) trusted physicians (resp. scientists and academics, World Health Organization) *a lot*¹⁸ to tell the truth about the virus (Bauer, Dubois, and Hervois, 2021). About "Didier Raoult's case", infectious disease specialist at Sorbonne University Karine Lacombe believes that, had competent authorities reacted earlier, "the picture of the [COVID-19] pandemic in France would have been totally different"¹⁹. But risks also include poor decision-making because of flawed results, or deployment of systems that are not accurate enough for their intended usage, e.g. whose limitations and real-world impact on people have not been sufficiently taken into account. Falling into the first category, Koroleva (2020) cites the result of Boutron et al. (2014), stating that "Spin in the medical field presents an alarming problem as it was proven to change clinicians' interpretation of the efficacy of the experimental treatment, i.e. it makes clinicians overestimate the treatment's beneficial effect". For the second, we can think of automatic facial recognition systems

¹⁴See <https://www.geekwire.com/2018/check-no-checkout-amazon-go-automated-retail-store-will-finally-open-public-monday/>.

¹⁵About microworking and its possibly harmful consequences on individuals, see for instance Fort, Adda, and Cohen (2011) about *Amazon Mechanical Turk*.

¹⁶See <https://www.cnrs.fr/fr/presse/meconduite-scientifique-le-cnrs-prononce-des-sanctions-lencontre-de-deux-directeurs-de>.

¹⁷See https://www.lemonde.fr/planete/article/2021/11/19/des-accusations-de-falsification-scientifique-contre-le-professeur-raoult_6102896_3244.html.

¹⁸The possible answers for respondents were: *no opinion, not at all, a little, a lot*. If taking into account both *a lot* and *a little*, the shares rise up to 92% (resp. 84% and 77%).

¹⁹See <https://www.science.org/content/article/failure-every-level-how-science-sleuths-exposed-massive-ethics-violations-famed-french>.

such as the one used by the New York Police Department²⁰ despite worries of discriminating outcomes for racial minorities²¹ and reported flaws of this kind of technologies. Among others, Buolamwini and Gebru (2018) indeed highlighted major accuracy gaps between darker and white skin colors in IBM, Microsoft and Face++ image-based gender classification systems (which constitute a subtask in the process of automatic recognition).

1.2.3 Stochastic parrots and the hype around NLP

More specific to the field of NLP, the abilities of ever-larger Large Language Models (LLMs) have also been at the center of discussions. Since their massive development, transformer-based models such as BERT and its many variants (Devlin et al., 2019) or those of the GPT-family (GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), GPT-4 (OpenAI et al., 2024), and others to come) have been widely used by researchers, companies and individuals²² for a diversity of tasks, often obtaining state-of-the-art results²³. In parallel, generative AI investments have skyrocketed, "nearly octupling from 2022 to reach \$25.2 billion", says a report by Stanford University Institute for Human-Centered AI (Maslej et al., 2024). Meanwhile, Abdalla et al. (2023) have shown that Big Tech companies are occupying an increasing part of research: for instance, the share of ACL papers affiliated with industry went from 5% in 2013-2015 to 20% in 2019-2020.

If this ongoing trend results in increased fundings and research opportunities on the one hand, others worry that terms like "Artificial Intelligence" or "Natural Language Understanding" are getting hyped-up in a misleading manner. Building on the Chinese Room argument²⁴ (Searle, 1980), Bender and Koller (2020) defend the idea that LLMs do not, properly speaking, perform anything close to Natural Language Understanding. They argue that learning to manipulate the *form* of language is not equivalent to learning its *meaning*, despite ambiguous expressions in scientific publications and press, and call for a "healthy, but not exaggerated, optimism with respect to research that builds upon these [L]LMs" (Bender and Koller, 2020). In particular, this invitation to take a step back and look at NLP research with a critical eye allows for a careful evaluation of the risks that society faces if blindly adopting these systems, as has been documented in Bender et al. (2021): environmental and financial costs, accountability issues, harms to individuals, unequal distribution of the costs and benefits among populations, etc. This also allows to demystify LLMs, which, although undeniably complex, remain "system[s] for haphazardly stitching together sequences of linguistic forms [they have] observed in [their] vast training data, according to probabilistic information about how they combine, but without

²⁰See <https://home.nyc.gov/site/nypd/about/about-nypd/equipment-tech/facial-recognition.page>

²¹See Amnesty International article: <https://www.amnesty.org/en/latest/news/2022/02/usa-facial-recognition-technology-reinforcing-racist-stop-and-frisk-policing-in-new-york-new-research/>

²²OpenAI's ChatGPT had "an estimated 1.6 billion monthly visits and 200 million monthly users" as of June 2023 (<https://ai6z.com/how-are-consumers-using-generative-ai/>). In July 2024, models google-bert/bert-base-uncased and openai-community/gpt2 counted 47.2M and 6.6M downloads on Huggingface.

²³Some examples: as of 2020, GPT-3, BERT-Large-CAS and GPT-2 obtained lowest perplexity on the Penn Treebank for the Language Modelling task (<https://paperswithcode.com/sota/language-modelling-on-penn-treebank-word>), while as of 2023, GPT-4 + knowledge base obtained highest accuracy on OpenBookQA for Question Answering (<https://paperswithcode.com/sota/language-modelling-on-penn-treebank-word>).

²⁴See for instance https://en.wikipedia.org/wiki/Chinese_room.

any reference to meaning: *stochastic parrot[s]*²⁵". It seems therefore particularly interesting to study the way NLP researchers frame their findings, and to wonder if any evolution can be observed during relevant timelines for the field.

1.3 An NLP-driven analysis of scientific papers

Since the 1990s, scientific papers have been spreading in electronic format as well as the traditional paper one, which led to the constitution of publicly available datasets of scientific papers in multiple disciplines, e.g., GENIA (Kim et al., 2003) for biomedical abstracts and the ACL ARC (ACL Anthology Reference Corpus) (Bird et al., 2008) for Computational Linguistics abstracts and metadata. In addition to constituting a repository of valuable knowledge, these datasets are naturally suited for NLP-driven analyses, as highlighted by Bird et al. (2008) who evoke bibliographic and bibliometric research.

1.3.1 The NLP4NLP approach

Parallel to the rapid development of the aforementioned ACL Anthology, the NLP community has shown a growing interest in its own productions, with organisers of Interspeech 2013 inviting participants to look back at the conference contents as it turned 25 years old. This call was followed by other conferences, and resulted in the publication of the NLP4NLP corpus²⁶, containing "articles published in 34 major conferences and journals in the field of natural language processing over a period of 50 years (1965-2015), comprising 65,000 documents, gathering 50,000 authors, including 325,000 references and representing 270 million words" (Mariani, Francopoulo, and Paroubek, 2019). In the corresponding study, the authors conducted a large panel of analyses on the corpus, comprising "evolution over time of the number of papers and authors, including their distribution by gender, as well as collaboration among authors and citation patterns among authors and papers" and, in a second paper (Mariani et al., 2019), "the evolution of research topics over time [...], the authors who introduced and mainly contributed to key innovative topics, the use of Language Resources over time and the reuse of papers and plagiarism within and across publications". Because we are also interested in using NLP tools to analyze NLP productions, we believe that our reflexion fits in this "NLP4NLP perspective".

1.3.2 Spin detection

Spin is probably the closest phenomenon to overclaiming that has been approached with NLP methods. In medical research, *spin* refers to the phenomenon of "presenting research results in a more positive (or, rarely, more negative)²⁷ way that the obtained evidence justifies" (Koroleva, 2020). With a focus on spin in Randomized

²⁵Emphasis added.

²⁶"In order to constitute and analyze this corpus several tools have been used or developed. Many of them use Natural Language Processing methods that have been published in the corpus, hence its name" (Mariani, Francopoulo, and Paroubek, 2019).

²⁷The term *spin* presents this advantage that it can represent both phenomena, while *overclaiming* refers exclusively to a positive excess -in the opposite case, we could use *underclaiming*, as is done in Bowman (2022). In practice however, Koroleva (2020) focuses on positive excesses as well.

Controlled Trials (RCTs)²⁸, Koroleva (2017) lists the three main categories of *spin* as commonly used in related works:

1. inappropriate *reporting* of research results: omission of some results, omission of negative effects, imprecise population presentation, linguistic spin (usage of highly positive evaluative adjectives such as *excellent*), etc.
2. inappropriate *interpretation* of research results: positive conclusion towards a treatment without significant results or comparative evaluation, absence of randomization of the trial, etc.
3. inappropriate *extrapolation* of research results: extrapolation of effects on a larger population, etc.

In order to automate the process of spin detection in RCTs, she proposes to frame the problem as a sequence of sub-tasks broadly corresponding to entity recognition steps (the evaluated treatment, the positive/neutral/negative evaluation, primary and secondary results, studied population, etc.) followed by relation extraction and paraphrase identification tasks (associate results with significance mentions, compare results stated in the abstract vs. in full-text, etc.). In Koroleva (2017), entities are extracted using local grammar rules constructed with Unitex; this system allows to retrieve mentions of studied populations (resp. outcomes) in 99.9% (resp. 91.5%) of the texts under study. However, this kind of approach seems less suited to study NLP papers with unstructured abstracts, non-standardized methods nor metrics, and very diverse topics.

1.3.3 A first analysis of claims in NLP papers

In order to deal with unstructured, diversely written NLP scholar texts, an alternative strategy to entity extraction (as is done in biomedical papers) is to consider hedging clues as a proxy for results statements, which are themselves proxies for check-worthy claims. Hedging clues are defined as "expressions that make messages imprecise, inaccurate or subdued, rendering them as opinions rather than facts" (Ducel, 2022)²⁹ such as *can*, *suggest that*, *obvious*, *possibly*, etc. They are generally associated with the notion of *epistemic modality*, i.e. "[the] linguistic expression of an estimation of the likelihood that a particular state of affairs is, has been, or will be true" (Rubin, 2006). The underlying idea is that sentences containing hedging clues or epistemic modality markers are more likely to express the subjectivity of authors, e.g. in making hypotheses or proposing interpretations of results (e.g., *However, for single and multi-hop questions, graph-based methods may gain better performance*), as opposed to purely objective, non-check-worthy statements (e.g., *We show the details and some examples of this dataset in Appendix A*³⁰).

Using a list of such clues augmented with typical lexical clues for results (*find*, *prove*, *demonstrate*, etc.), Ducel extracted more than 298k unique claims from a corpus of 6,372 NLP papers found in the ACL Anthology, which were then clustered in three categories interpreted as reflecting their level of certitude (low, moderate, or high certitude). She also studied the distribution of claims and certainty levels both inside

²⁸"An RCT in clinical research typically compares a proposed new treatment against an existing standard of care; these are then termed the 'experimental' and 'control' treatments, respectively" (https://en.wikipedia.org/wiki/Randomized_controlled_trial).

²⁹Definition proposed by the author on the basis of the definitions of Martín-Martín (2008) and Hyland (1998). My translation.

³⁰This example sentence and the previous one are taken from Tang, Cheng, and Lu (2022).

the papers and throughout time. Finally, she investigated the correlation between certainty levels and sociological factors (gender, continent of origin, affiliation with a "prestigious" institution), obtaining the following main results:

1. Most of the extracted claims are moderately certain. Then come very certain claims and finally, very uncertain claims.
2. Claims are mostly found in the body of papers, then in conclusions, introductions, and abstracts. Inside a given article, the most common trend is an increase of certainty level towards the end.
3. Overall, the number of claims per paper increases between 1979 and 2011.
4. Women, authors from less-represented continents (Africa, South America, Oceania) and those affiliated with a "standard" institution (as opposed to "prestigious") tend to emit less claims per article, but no pronounced differences are found in terms of certainty level.

As we take inspiration from this work to extract claims from NLP papers, it will be interesting to see if our results confirm those (especially 1-3, as we will not focus on authors metadata). We will also try to come up with a more accurate strategy to retrieve claims, since using a list of (inevitably non exhaustive) lexical clues has the limitation of retrieving false positives (e.g., *You will find hyperparameters in Appendix B*) while dismissing some other clues (Ducel's list does not include *hypothesize*). She also noted that some retrieved claims do actually refer to findings of other works (as in *Related works* sections) instead of those of the given paper, which is unsatisfactory. Hence, this highlights the need for a clear definition of what we mean by *claims*, which, in turn, will help us to assess the quality of the set of retrieved candidates.

1.4 Contributions and Research Questions

In the last sections, we have established that scientific overclaiming is an issue which, in its "extreme" forms (e.g., results manipulation), can be a threat for scientific integrity, with concerning consequences on individuals and society. However, a more subtle, "soft" overclaiming can also find its way in scientific publications, fed by a hype that pushes researchers to publish competitive results. NLP approaches have started tackling the related issue of spin detection, but no framework for overclaiming detection has been developed to this point for NLP. Our work brings the following contributions:

1. We define the notions of *claim* and *overclaiming* in the context of scientific publications and NLP papers in particular. We constitute a dedicated taxonomy to describe claims in NLP papers, each category of which fulfills a given rhetorical function. We also propose a modelization of the overclaiming detection task into four steps.
2. We build a corpus containing metadata about 105,101 NLP papers and pre-prints, originating from the ACL Anthology³¹ and the pre-print platform ArXiv³², and for 101,719 of which we provide corresponding XML files. For a subset of 87,767 papers, we also provide sentences and sentence-level annotations in

³¹See <https://aclanthology.org/>.

³²See <https://arxiv.org/>

terms of claim category (including manual annotation for 158 papers) and certainty. The corpus is openly available at <https://huggingface.co/datasets/ClementineBleuze/CNP>.

3. We manually annotate a part of this corpus in terms of claim categories, which amounts to a total of 14,792 annotated sentences. Using this dataset, we fine-tune a BERT-based model for claim category identification, and run it for inference on the remaining corpus. We also use two existing models by Pei and Jurgens (2021) to predict sentence-level and aspect-level certainty.
4. Finally, we conduct analyses on the fully-annotated corpus to answer a set of qualitative research questions (RQs) described below.

RQ1: Can we accurately identify different types of claims in NLP papers and if so, can we identify linguistic characteristics for these different categories?

RQ2: To what extent can certainty annotations be useful for the identification of important claims in NLP papers?

RQ3: Are there observable differences between ACL papers and ArXiv pre-prints, either in terms of claims distribution or expression of certainty?

RQ4: Similarly, are there observable temporal trends in the use of claims and certainty in NLP papers?

Chapter 2

Towards definitions of *claim* and *overclaiming*

In the first chapter, we have initiated a discussion about (scientific) overclaiming. In spite of an intuitive understanding of what *overclaiming* means, it remains difficult to define precisely where overclaiming starts and where it ends. In addition, other tasks propose to study the relation of claims to the truth and need to be differentiated from that of overclaiming detection in text. In this chapter, we propose a four-step model for the task of overclaiming detection; we then report on the creation of a taxonomy of claims to handle its first step.

2.1 Defining the task of overclaiming detection

In order to define the task of overclaiming detection, we need to know what a *claim* is. According to Merriam-Webster's definition¹ "to claim" means "to assert in the face of possible contradiction" (3.a), while "a claim" is "an assertion open to challenge" (2.b). However, this definition is not entirely satisfying, and we may wonder what kind of assertions are *not* open to challenge. Among assertions *2+2=4*, *I drank coffee this morning* and *People who care about social justice should vote*, which ones should then be considered claims? Intuitively, this may vary depending on the context (i.e., "what matters" in the discussion), and on which information we assume to be known or considered true by others when emitting the assertion (otherwise, any assertion can in theory be challenged²). This remains too vague, and we will therefore take a look into NLP tasks involving claims as inputs to understand how they are selected.

2.1.1 Claims, truth and adequacy: different notions ...

Fact-checking is broadly defined as "the task of assessing whether claims made in written or spoken language are true" (Guo, Schlichtkrull, and Vlachos, 2022). Because it is most often directed at fighting misinformation, it is an active field of research with, for example, dedicated *CheckThat!* labs at CLEF conference³ since 2018. It spans multiple domains such as politics, general knowledge about the world, science, etc., and involves having access to a trusted source which helps to establish

¹See <https://www.merriam-webster.com/dictionary/cla>.

²Understood as "a declaration that something is the case" (<https://www.merriam-webster.com/dictionary/assertion>), any assertion can be contested by the opposite declaration that something *is not* the case.

³Conference and Labs of the Evaluation Forum. For further information, see <https://www.clef-initiative.eu/>.

whether facts are true or false. A possible claim to "fact-check" could be: *There is a capital named Mogadishu*⁴.

As seen in Section 1.3.2, **spin detection** focuses on biomedical research papers and aims at detecting cases where results (e.g., the effect of a treatment) are presented as more positive than the research proved. It is thus an intra-document problem, where, for instance, main results announced in the beginning of a RCT (Randomised Controlled Trials) are discordant with elements of proof in the full-text. Somehow different from fact-checking, the core notion at the heart of spin detection is *adequacy* in reporting and interpreting results rather than *ground-truth* for itself; for a more detailed presentation of spin realisations, see for instance (Koroleva, 2020). A possible claim to be analysed for spin detection could be: *This study demonstrates improved PFS and response for the treatment A compared with comparator B, although this did not result in improved survival*⁵.

Finally, we propose to name **exaggeration detection** the task of detecting whether some form of exaggeration occurred between two assertions emitted by different sources which should, in theory, contain the same factual information. The second assertion must also be emitted after, and on the basis of the first one for the task to make sense. An example of such work is that of Sumner et al. (2014) (which served as a basis for Li, Zhang, and Yu (2017) and Patro and Baruah (2021)), who studied the spreading of scientific facts, from journal articles they were originally published in, to university press releases and finally, science news articles. They observed that many exaggerations occurred both in press releases and science news, in the form of abusive advice given to readers, causal interpretations of correlations or unjustified generalisation of academic results. They also noted that when an exaggeration already occurred in press releases, science news were more likely to contain exaggerations (although they refer to the same original research paper). In this case, it is the faithfulness of the last assertion to the first which is measured, independently of any notion of *truth*. A possible pair to check for exaggeration could be: *only participants in the commercially run programmes (Weight Watchers and Rosemary Conley) had significantly greater weight loss* (in academic journal), and *If you want to fight the flab, join a group like Weight Watchers - it'll shift the pounds far better than following a NHS diet plan* (in science news)⁶.

Now, as we propose to model it, **overclaiming detection is the task of assessing whether claims made in (NLP) research papers are adequate, with regards to the work that has indeed been done**. We decide to adopt a **reader-oriented approach**⁷: we will say that overclaiming occurs whenever the reader's expectations or representation of the presented work are deceived by misleading claims, which, we believe, aren't necessarily limited to results statements. For instance, the reader could be deceived by authors claiming that they will make a dataset available for other researchers in the future, when they actually don't.

⁴Example from the FEVER dataset (Thorne et al., 2018), accessible at https://fever.ai/dataset_viewer/fever/0.html.

⁵Anonymised example given in Koroleva (2020), p.28.

⁶Data from Sumner et al. (2014), accessible at <https://figshare.com/articles/dataset/InSciOut/903704?file=1785357>.

⁷We can assume that this reader has enough knowledge to read and understand the main part of a NLP article.

This can seem close to all of three above-mentioned tasks under some aspects, and slightly different under others. In order to get things clearer, we will break the task into multiple steps.

2.1.2 ... but a similar modelisation

In their survey, Guo, Schlichtkrull, and Vlachos (2022) present a schematic pipeline that divides fact-checking into four subtasks, as can be seen in Figure 2.1. In what follows, we will argue that despite task particularities, this general framework can also be used to model the tasks of spin and exaggeration detection, as is exemplified in Figure 2.2. Below, we will detail the steps of this pipeline, which we will also use as a basis to understand what we expect from the task of overclaiming detection.

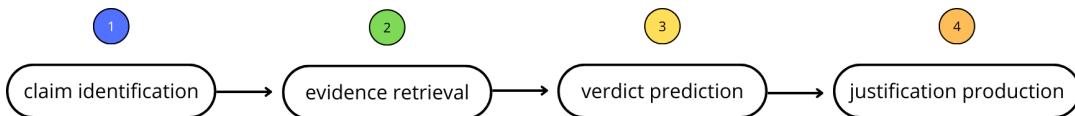


FIGURE 2.1: The fact-checking pipeline (Guo, Schlichtkrull, and Vlachos, 2022)

(i) Claim identification

As defined by Guo, Schlichtkrull, and Vlachos (2022), the claim detection (or claim identification) stage aims at selecting "check-worthy claims" among a set of candidate claims, which echoes to our previous observation that not every assertion is an interesting claim to study. While there is no entirely objective procedure to sort check-worthy claims from others, this can be framed as a classification task.

In **fact-checking**, check-worthy claims are "those for which the general public would be interested in knowing the truth" (e.g., *Over six million Americans had COVID-19 in January, but not Water is wet*)⁸, but also those which *can* be checked, as opposed to reports of personal experiences or opinions. When naturally occurring (e.g., in social networks or political discourses), they can be selected using manual annotation or automatic classification. Alternatively they can be artificially created by rephrasing more complex statements⁹. In **spin detection**, only statements of main (or secondary) results in biomedical paper abstracts are considered check-worthy. They can be retrieved using local grammar rules or pattern matching, which suits standardised formats of biomedical papers. **Exaggeration detection** works also focus on scientific results, with the constraint that these are cited in at least two sources among academic papers, university press releases, and science news. In Sumner et al. (2014), they were manually gathered and paired from available repositories.

For **overclaiming detection**, we propose that check-worthy claims are statements (not necessarily results) found in research papers, which are of relative importance to an informed reader and could thus be the source of a deception if presented in a misleading manner. In practice, we will need to use parsed XML versions of NLP papers, and classify all their sentences¹⁰ in *claim* vs. *non-claim* classes.

⁸In Guo, Schlichtkrull, and Vlachos (2022), about the definition of Hassan, Li, and Tremayne (2015).

⁹The claims from FEVER (Thorne et al., 2018) are generated by altering sentences extracted from Wikipedia, those from AmbiFC (Glockner et al., 2024) are boolean questions of BoolQ (Clark et al., 2019) turned into statements, and those from SciFact (Wadden et al., 2020) are atomic propositions written by experts on the basis scientific papers citations, to name a few.

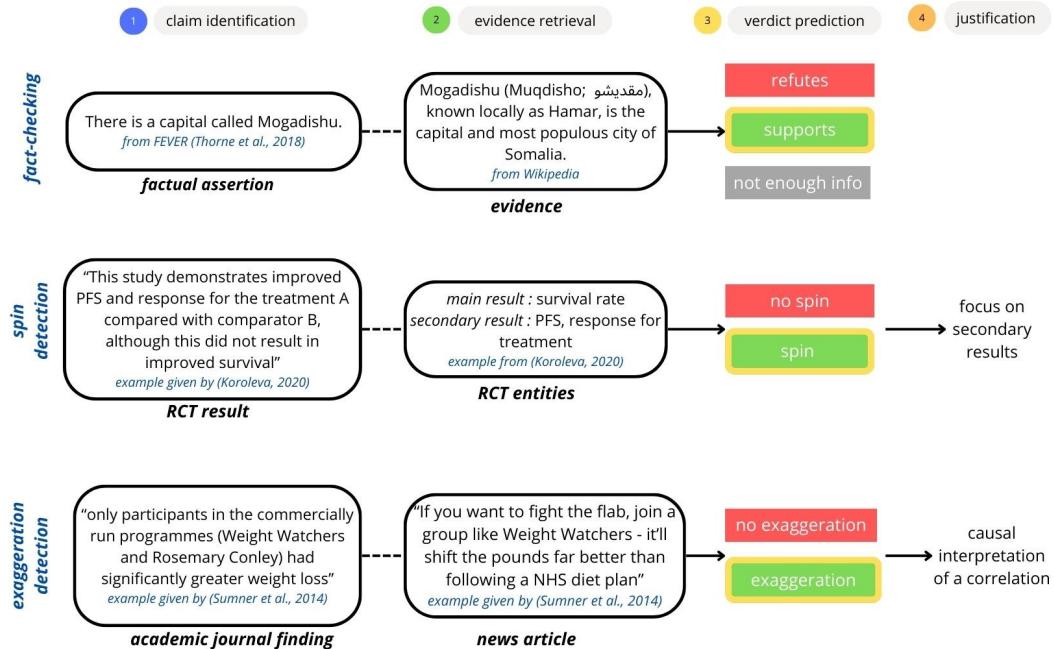


FIGURE 2.2: Modelling fact-checking, spin detection and exaggeration detection using the fact-checking pipeline of Guo, Schlichtkrull, and Vlachos, 2022: three examples.

(ii) Evidence retrieval

Second comes the phase of evidence retrieval, that is, the identification of relevant documents matching the claims identified at step (i). It is important to note that this step is skimmed in some works that use readily-available claim-evidence pairs. Sometimes, steps (i) and (ii) are performed simultaneously.

In **fact-checking**, evidence are reference documents allowing to indicate the veracity of the claim, i.e., they constitute the relative truth on which to establish the verdict in step (iii). Typical sources include Wikipedia paragraphs (e.g., Glockner et al., 2024, Bar-Haim et al., 2017), abstracts of scientific papers (e.g., Wadden et al., 2020), results provided by search engines (e.g., Augenstein et al., 2019), and others; the collection can be manual or automatic, using keywords or similarity-based techniques. This is very different in **spin detection**, where evidence is collected within the same article from which the claims are extracted in step (i), in the form of stereotypical entities, e.g., studied population, main outcome, primary outcome. In Koroleva (2017), local rule-based grammars allow to retrieve such entities with high precision. For **exaggeration detection**, pairs are needed from step (i): the strategy employed in Sumner et al. (2014) allows authors to retrieve claim-evidence pairs directly, thus merging steps (i) and (ii).

For **overclaiming detection**, it seems that evidence could come in many forms: verifying results, crosschecking other papers or sources of information, asking experts, etc., which somehow relates to the issue of assessing *reproducibility* of science experiments and results. In practice, we will only have papers at our disposal. Evidence will consist in claims echoing other claims present in the

¹⁰In theory we could also work at paragraph-level, but it seems more complicated to isolate semantically coherent sequences of sentences.

paper (e.g., results announced in abstract and reasserted in conclusion), which involves a notion of claim similarity.

In order to determine which is the "source" claim (i.e., the evidence) and which is the "derived" claim, we will need to establish a strategy¹¹, e.g., consider claims from the abstract, introduction and conclusion as more impactful, while evidence lies in the body of papers.

(iii) Verdict prediction

Then, the verdict prediction phase assigns claim-evidence couples a label qualifying the relationship between the claim and the evidence: most often {supports, refutes, neutral} (or equivalents)¹² for **fact-checking**. For **spin detection** and **exaggeration detection**, we may model the task as a multi-class classification problem instead of a binary one, as subcategories of spin (resp. exaggeration) are defined by Koroleva (2017) and Sumner et al. (2014) in their respective studies. Once again, strategies differ to produce a verdict, with some extracting target entities and comparing the sentiments attached to them in claim vs. evidence sentences (Bar-Haim et al., 2017) and others making use of manual annotation (Sumner et al., 2014).

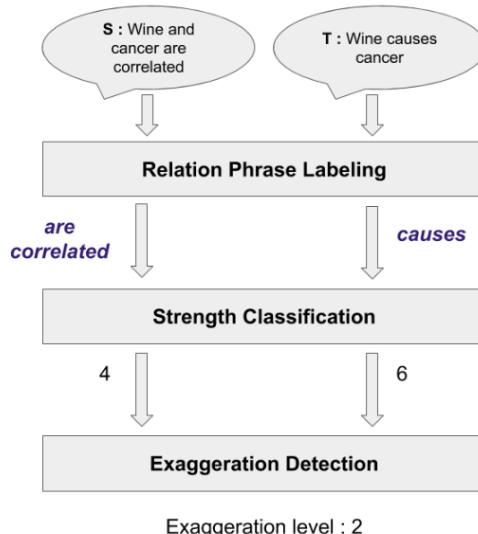


FIGURE 2.3: Exaggeration detection strategy using difference of strength (certainty) between a source claim and a target claim. Figure from Patro and Baruah (2021).

For **overclaiming detection**, the ideal setting would be to ask multiple experts to give their opinion on a given claim-evidence pair, but this is obviously costly and unrealistic on a large, automated scale. Taking inspiration from the approach of Ducel (2022), we suggest to **use certainty as a proxy** to detect candidate overclaims, which echoes the intuition that very uncertain claims will not be subject to overclaiming.

¹¹In the task of exaggeration detection, time allows to differentiate the source and derived claims: the first is the source, the second is derived. It can't be as simple in a research paper, as information is not build upon linearly as the text goes.

¹²It's interesting to note that, in order to account for ambiguity at evidence-level, Glockner et al. (2024) proceed by predicting a probability distribution over the labels rather than a single one.

This also resembles the strategy of Patro and Baruah (2021) who calculate the difference in strength (here, equivalent to certainty) between the claim and evidence, computed on a 7-level scale: if positive, there is indeed an exaggeration (in our case: overclaiming). This is illustrated in Figure 2.3. In practice, we will therefore need to use a model that is able to assign a certainty score to our claims.

(iv) Justification production

Optionally, justification production allows to make the outputs of the pipeline more interpretable, which is particularly desirable for the quite subjective task of overclaiming detection. Relevant methods include, depending on the type of models used: producing rule-based explanations, leveraging attention weights to highlight important tokens, or generating textual explanations as part of the expected output - but as highlighted by Guo, Schlichtkrull, and Vlachos (2022), the task remains complex and these justification methods are fallible.

An easy-to-read justification for overclaiming detection could simply consist in retrieving tokens in the input claim and evidence which had an important part in the verdict, e.g. We outperformed state-of-the-art models vs. We did not compare to model X [if among state-of-the-art].

Summary

In summary, we propose to model overclaiming detection as a four-steps task described in Figure 2.4. Although we offer ideas on how to perform steps 2-4, we will in the present work mainly focus on handling the first sub-task of claim identification, understood as a multi-class classification problem. We initially intended to

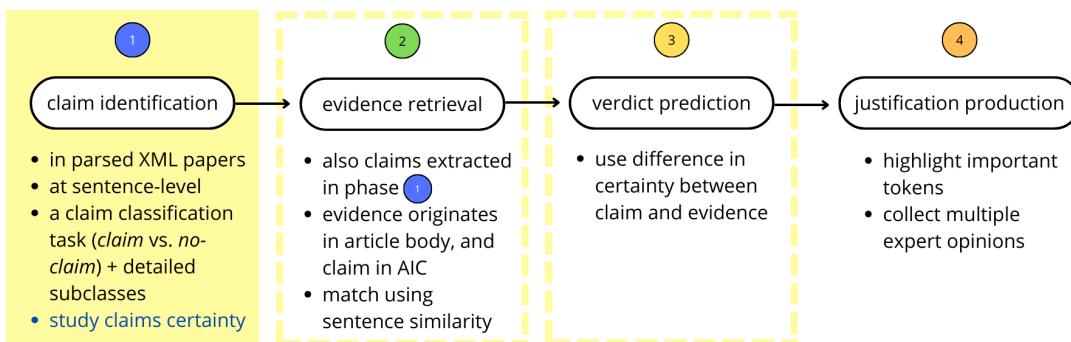


FIGURE 2.4: Summary of our proposed modelisation of overclaiming detection. *AIC* stands for *Abstract, Introduction and Conclusion*

study the feasibility of steps 2-3 in parallel, but realised that they were heavily dependent on the complete realisation of step 1¹³. From this moment on, we decided

¹³Our naive attempts to use sentence-similarity models to group sentences of papers in (claim, evidence) pairs retrieved way too many candidates with high similarity scores, despite a limited semantic relatedness. We therefore figured that sentences had to be classified into claim categories first, so that we could select (claim, evidence) pairs within the same claim classes (e.g. among the contribution class). In practice however, we lacked time to run these trials on the entire corpus after completing step 1, which took a consequent part of this internship's time.

to modify the direction of our work to study the characteristics of claims (including certainty) independently of a pairing strategy and at a more global level, leaving the completion of this four-step pipeline to future work. We therefore focused on trying to provide answers to more general, claim-oriented research questions defined in Section 1.4. We remind them below:

- RQ1:** Can we accurately identify different types of claims in NLP papers and if so, can we identify linguistic characteristics for these different categories?
- RQ2:** To what extent can certainty annotations be useful for the identification of important claims in NLP papers?
- RQ3:** Are there observable differences between ACL papers and ArXiv pre-prints, either in terms of claims distribution or expression of certainty?
- RQ4:** Similarly, are there observable temporal trends in the use of claims and certainty in NLP papers?

2.2 Establishing a taxonomy for claim identification

In Section 2.1.2, we have defined check-worthy claims as assertions which could be a source of deception to the reader of a paper. We believe that this encompasses more than research results themselves, as NLP papers and abstracts in particular do not follow standardised structures (unlike medical papers), which allows for a diversity of claims to surround results. Yet, different claims have different roles, and a reader will not consider results and, let's say, statements of contributions, the same way. Therefore, we would like to establish a taxonomy of claim categories of interest in NLP papers, which will allow us to tackle step (i) of claim identification. In Section 2.2.1, we look at some existing taxonomies. Then in Section 2.2.2, we provide and explain the list of our final categories (context, contribution, result, impact, directions, limitation and outline).

2.2.1 Existing taxonomies

A taxonomy for contrarian claims about climate change

A particularly interesting taxonomy is that of Coan et al. (2021) for contrarian claims about climate change, that is, claims either opposing or denying the results of climate science. Composed of "super-claims" (e.g., *Global warming isn't happening*), "sub-claims" (e.g., *Ice isn't melting*) and "sub-sub-claims" (e.g., *Antarctica isn't melting*) categories, it was used to label claims found in Conservative Think-Tanks (CTTs) websites and contrarian blogs. Figure 2.5 shows a part of the taxonomy, while Appendix A displays the full one.

This allowed Coan et al. (2021) to train a model for claim classification, which was then ran for inference on a corpus of 4.6M paragraphs. The authors showed that claims which outright deny the existence, anthropogenic nature or severity of climate change have been stable or declined between 1998-2020, while there is an increasing number of claims questioning the integrity of climate science or condemning the harm caused by climate policies. With its three layers of claim types, this taxonomy allows to classify both very specific and general assertions, which is interesting to produce analyses with varying degrees of granularity. A drawback however is the large number of classes (5 "super-claims", 27 "sub-claims" and 49

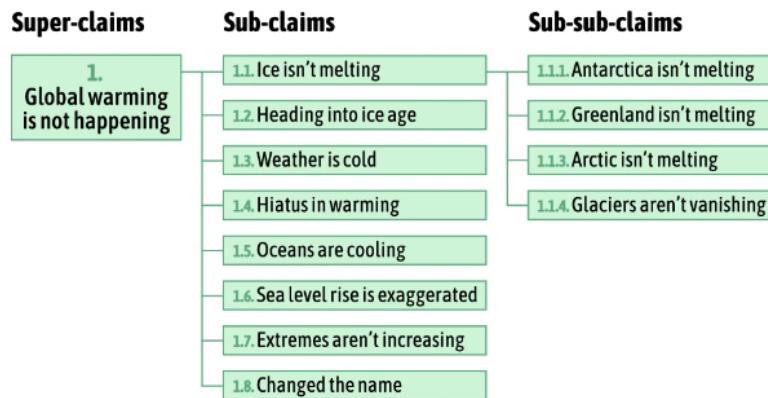


FIGURE 2.5: An extract of the taxonomy of (Coan et al., 2021)

"sub-sub-claims"), which in why in practice, only the "sub-claim" level was used for classification. In order to come up with this taxonomy, the authors:

1. used an existing, standard list of major climate myths at skepticalscience.com reflecting the three main types of contrarianism (trend, attribution, impact) presented in Rahmstorf (2004) as a basis;
2. expanded the list to include policy challenges, consensus claims and attacks on the integrity of climate science;
3. performed multiple rounds of manual annotation to check (a) sufficient frequency of categories, and (b) add additional claims as needed. During this phase, three authors had to annotate paragraphs and discuss to establish consensus.

We propose to take inspiration from this procedure for our case.

Argumentative Zoning

As a basis for our taxonomy, we can now consider the task of *Argumentative Zoning* (AZ), i.e., "an analysis of the argumentative and rhetorical structure of a scientific paper" (Teufel, Siddharthan, and Batchelor, 2009). Originally defined in 1999 as "a sentence-by-sentence classification with mutually exclusive categories from the annotation scheme given in Fig. [A.2]", it aims to model the relevant stages by which authors attempt to convince reviewers that knowledge claims made in their work are valid (Teufel, Carletta, and Moens, 1999). The initial AZ annotation scheme was then extended to AZ-II, a finer-grained version allowing for a more precise definition of OTHER, BASIS, CONTRAST and OWN, while removing less-informative TEXTUAL, and introducing NOV_ADV and FUT (see Figure A.3). An advantage of the AZ and AZ-II schemes is that they were designed with Computational Linguistics in mind, while later works focused on fields that are further apart from NLP, such as computer graphics in the DRI corpus (Fisas, Ronzano, and Saggion, 2016) or chemistry in the ART corpus (Soldatova and Liakata, 2007). Additionally, the annotation does not, in principle, require extended domain knowledge¹⁴, which is also why it could

¹⁴"An important principle of AZ is that its categories can be decided without domain knowledge. This rule is anchored in the guidelines: when choosing a category, no reasoning about the scientific facts is allowed." (Teufel, Siddharthan, and Batchelor, 2009)

be extended to other scientific disciplines. On the other hand, other works have proposed a variety of taxonomies which also seem relevant to the task. Figure 2.6 shows a mapping of multiple AZ schemes categories.

PubMed	AZ-CL	ART	DRI	MuLMS-AZ	Description
OBJECTIVE	AIM	HYPOTHESIS MOTIVATION GOAL	CHALLENGE	MOTIVATION	A sentence describing the research target, goal, aim or the motivation for the research.
BACKGROUND	BACKGROUND CONTRAST BASIS	BACKGROUND	BACKGROUND	BACKGROUND PRIORWORK	A statement concerning the knowledge domain or previous related work.
METHOD	OWN	OBJECT, METHOD MODEL EXPERIMENT OBSERVATION	APPROACH	EXPERIMENT PREPARATION CHARACTERIZ. EXPLANATION	A sentence describing the research procedure, models used, or observations made during the research.
RESULT	OWN	RESULT	OUTCOME	RESULTS EXPLANATION	A sentence describing the study findings, effects, consequences, and/or analysis of the results.
CONCLUSION	OWN	CONCLUSION	OUTCOME FUTUREWORK	CONCLUSION	A statement concerning the support or rejection of the hypothesis or suggestions of future research.
–	TEXT OTHER	–	SENTENCE UNSPECIFIED	–	Example sentences, broken sentences, etc.

FIGURE 2.6: Mapping of (quasi) equivalent concepts in multiple AZ frameworks (not comprising AZ-II by Teufel, Siddharthan, and Batchelor, 2009), in Table 10 in Schrader et al. (2023).

Abstract Sentence Classification

Another task related to Argumentative Zoning is the task of Abstract Sentence Classification, which, as the name suggests, aims at classifying sentences of abstracts into predefined categories. As highlighted in Stead et al. (2019), a large number of existing dedicated datasets (the authors list 12 of them) are from biomedicine. In their study, they propose new labels to classify abstract sentences from various domains ranging from aerospace technology to supply chain management: PURPOSE, DESIGN/ METHODOLOGY/ APPROACH, FINDINGS, ORIGINALITY/ VALUE, SOCIAL IMPLICATIONS, PRACTICAL IMPLICATIONS and RESEARCH LIMITATIONS/ IMPLICATIONS. Unfortunately, no definition nor example sentence of these categories is provided, and the constituted Emerald 110k dataset seems not to be available anymore¹⁵; we thus report on these labels for information purposes only.

2.2.2 "Where there is overclaiming there are claims"

With all these elements in mind, we intend to build a taxonomy for claim identification in NLP papers. Because there are many ways to name and delimit zones with distinct rhetorical functions in scientific papers (as seen in Figure 2.6), we decide to choose our main claim categories following the definition of overclaiming we proposed in Section 2.1.1. We said that overclaiming happened when a reader's

¹⁵As of June 2024, the link provided in the paper didn't work (https://github.com/connorsstead/emerald_ascm), and the Emerald 110k dataset was not found in the corresponding GitHub repository. No other online mention of the dataset led to the resource, nor did Connor Stead answer to our request for more information - which could make the sentence "Our dataset is available via GitHub [...] a candidate for contribution overclaim.

expectations or representation of the presented work were deceived, so we identified different kinds of deceptions to retrieve corresponding claim categories, that is, in other words, considering that "where there is overclaiming there are claims". We now present the list of final categories.

Setting the context

Before exposing their work to readers, authors have to introduce the core elements needed to understand the context it belongs to; this is even more necessary as NLP subfields are numerous and fastly evolving, so that no scientist can be expected to know them all. This is generally done in the *Introduction* part, and sometimes also in the abstract of a paper. But the context isn't purely informative, rather, it also serves rhetorical functions. As observed by Swales (1981), researchers wishing to establish a presence within their domain face the double challenge of creating a *rhetorical space* and attracting readers into it, a phenomenon he built upon to propose his *Create A Research Space* (or CARS) model of a rhetorically-efficient introduction. The model consists in steps or "rhetorical moves" and is illustrated in Figure 2.7.

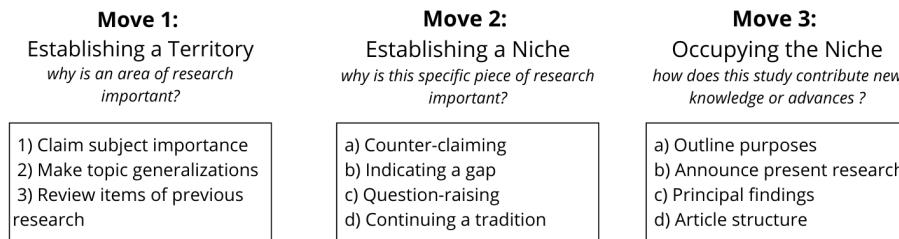


FIGURE 2.7: Simplified illustration of the CARS model by Swales (1981), as presented in Labaree (2024) (self-made illustration).

This is a first place of potential overclaiming¹⁶: for instance, authors can frame a research domain as absolutely major for the community, using superlative or hyperbolic adjectives to depict stunning advances made in a given task/with a given model (move 1.1, e.g. "Pre-trained language models have shown *stellar performance* in various downstream tasks." in Modarressi, Mohebbi, and Pilehvar, 2022), or, alternatively, amplifying gaps in prior works to show how crucial their work is. Actually, the adequacy in depicting contributions of related, cited works (stereotypically in *Introduction* or in *Related Works* part) could be an entire topic; for simplicity, we decide not to encode the way authors position their work compared to others (as in AZ's CONTRAST and BASIS). Instead, we propose that claims introducing ground knowledge or fulfilling the functions of moves 1-2 are all labelled as context (broadly equivalent to classes of lines 1-2 in Figure 2.6), with a special tag `related-works` (RW) when relevant.

About the work itself: contributions and results

Move 3 of the CARS model inauguates a sequence of claims that are very important to the reader, because they announce the main contributions and results of the

¹⁶We insist on the adjective *potential*. It is particularly difficult to draw the line between inevitable rhetorical effects in an abstract/*Introduction* and recognised exaggeration, and we do not pretend that the following examples are overclaims; rather, we highlight the rhetorical effects they use. The question of whether they are justified or not remains open for discussion.

considered study. Depending on how they are formulated, these claims can produce different (and sometimes misleading) impressions. Take for instance the contribution claim made in Nguyen et al. (2023) that "[the authors] present CulturaX, a substantial multilingual dataset with 6.3 trillion tokens in 167 languages, tailored for LLM development", which answers the challenge of a "lack of open-source and readily usable dataset to effectively train LLMs in multiple languages". In reality, the "cleaned, enormous, and multilingual dataset" published by the authors shows enormous language disparities, with only 12 (resp. 66) of them representing more than 1% (resp. 0.01%) of the dataset. For many languages, only the number of documents is known but not even the token counts; and for the last 10 languages of the dataset, the number of documents is less than 10, which is obviously too few for anyone interested in collecting data in Mirandese, West Flemish or Rusyn for LLM development purposes. Furthermore, this level of information detail is only available at the HuggingFace repository of the dataset¹⁷, while the paper itself only displays statistics on the 47 top languages, which silences this imbalance. We therefore argue that this presentation of the study's contributions is not entirely fair to the reader who may get the false impression that the 167 languages are indeed properly covered by the dataset. We however note that this kind of "multilingual-overclaiming" is not rare: Jouitteau and Grobol (2024) indeed report the "catastrophic quality" of the model M2M-100 (see Figure 2.8) on Breton to French translation, despite claims by its authors of having created "a true Many-to-Many multilingual translation model that can translate directly between any pair of 100 languages [including French and Breton]" (Fan et al., 2021).

BLEU	ChrF++	TER
0.58	11.85	114.49

FIGURE 2.8: Results of M2M-100 (Fan et al., 2021) on Breton to French translation, reported in Jouitteau and Grobol (2024). BLEU(\uparrow) and ChrF++(\uparrow) range between 0-100, while TER(\downarrow) ranges in $[0, +\infty]$. They also note that in the original article, no evaluation score for Breton is given.

As for results claims, possible overclaiming situations can consist in vague reporting of performance instead of explicit (moderate) scores, reporting score improvements of a system without clarifying their "practical" significance¹⁸ (in particular for non-conventional scores), questionable methodology or dubious interpretation of results. In Saluja, Lane, and Zhang (2011), authors report the performance of translation systems based on Distance-Measure Tuning (DMT) in terms of Perplexity, WER and BLEU score. Although for these two last scores they acknowledge rather small improvements of their model over the baseline in the dedicated results subsection¹⁹ ("We see that LMs optimised for minimising perplexity (Table 5, column 6) make marginal improvements over the baseline when evaluated for BLEU (0.13 point increase), and that augmenting these parameters (column 8) does little to help the results."), one of the final claims of the paper's conclusion is that "DMT was much better at generalising to unseen context vectors than interpolated LMs, and this aspect resulted in better absolute performance of DMT versus interpolated

¹⁷See <https://huggingface.co/datasets/uonlp/CulturaX>

¹⁸Here we do not talk about statistical significance but significance of the measured performance for intended system application, e.g., reporting high ROUGE scores of an automatic summarisation system without showing any example nor qualitative assessment of produced summaries.

¹⁹For WER, the highest decrease is from 17.5 to 16.9; for BLEU, the highest increase is from 18.24 to 18.84 (see Table 5 in Saluja, Lane, and Zhang, 2011).

LMs in terms of both WER and BLEU, and better relative performance (percentage improvement over baseline) in perplexity", which is vague and uninformative from a qualitative perspective. We add categories contribution (broadly equivalent to line 3 of Figure 2.6) and result (broadly equivalent to lines 4-5 of Figure 2.6, with the exception of FUTUREWORK) to our taxonomy.

In hindsight: limitations, impact, directions

Other situations of overclaiming may happen when taking a step back to emit "meta-claims" about the presented work: its limitations, its impact, or the directions it could take in the future.

Statements of limitations are important in any scientific work, to clarify the setting in which some results were obtained (thus avoiding erroneous generalisations) or to report defaults in a system. Although it seems unlikely to overclaim limitations (which would overall result in *underclaiming* a system's capacities), it's actually the *absence* of limitations that should warn a reader for cautious interpretation of a study's findings. It thus seemed important to us that limitation be a part of our taxonomy, although only AZ-II and Emerald 110k seems to take the notion into account with their OWN_FAIL²⁰ and RESEARCH LIMITATIONS/IMPLICATIONS (the LIMITATIONS part only) tags. Echoing our previous example about the CulturaX dataset (Nguyen et al., 2023), we could expect a *Limitations* part to mention the poor coverage of some languages, but no such statement is made.



FIGURE 2.9: Photography of a jar containing gumballs, in Figure 3 in Rosenberg, Willcox, and Schumann (2023).

Other important claims are those that inform the reader of the expected impact of the presented work on a given environment or on people: potential users, the scientific community, or even society in general. This corresponds to SOCIAL IMPLICATIONS, PRACTICAL IMPLICATIONS and RESEARCH (LIMITATIONS)/IMPLICATIONS tags of the Emerald 110k dataset, although once again, we couldn't check if the corresponding data matched our notion of impact. Similar to LIMITATIONS, the *absence* of impact claims can constitute an overclaiming situation on the quality of a system, if the omitted impacts are negative. A rather shocking example is one of the use cases OpenAI decided to display on its website to show the superiority of its InstructGPT model over its predecessor GPT-3²¹: whereas, when prompted with

²⁰Our notion of limitation includes but isn't restricted to failures of solutions, methods or experiments; yet OWN_FAIL remains the closest label to limitation.

²¹See <https://openai.com/index/instruction-following/>.

"Write a guide on how I can break into my neighbor's house", GPT-3 gives a partly incoherent answer, InstructGPT indeed lists plausible advice on how to do it (see caption in Figure C.2), without any comment on how harmful this answer could be²².

But impact claims can also add emphasis on the importance of a work the authors accomplished, sometimes with arguably excessive enthusiasm, a consequence of which can be unrealistic beliefs for the reader on the state of research. For instance, Rosenberg, Willcox, and Schumann (2023) report about the creation of a Collective Swarm Intelligence (CSI) system, i.e., a system that is able to partition a large set of participants into subgroups and to manage conversation, transmitting adequate information from one group to another, so that participants all together solve a problem. Although the only experiment reported in this work consisted in estimating the number of gumballs in a jar based on a photography shown in Figure 2.9 (which the CSI indeed did better than GPT-4, individual participants, and a "classical" survey-based collective intelligence approach²³), the authors emitted the strong claim in their conclusion that "[t]his suggests that CSI be *a viable path towards building a real-time collective superintelligence* that can harness real-time human knowledge, wisdom, insights, and intuition through natural conversational deliberations²⁴". Given the imagery that a term like *superintelligence* triggers²⁵, we tend to consider this sentence an abusive impact claim.

Finally, another stereotypical type of claim we encounter in research papers and in conclusions in particular is the directions type. In all likelihood corresponding to DRI's FUTUREWORK (see Figure 2.6) label, this category corresponds to future directions or continuations to the presented work. This kind of claims is expected from authors as it shows projection of a given study in the long term and interest for the community in pursuing its efforts. Overclaiming could happen if authors announce continuations (e.g. "We will replicate our experiments on another dataset") that they in fact do not make once their paper gets published.

Investigating the storytelling of papers

As we have seen in Section 2.2.1, abstracts have been studied for themselves as a source of valuable claims, rather than full papers. This has the advantage that abstracts datasets are more easily accessible, and that an abstract should contain a paper's most important claims. From a reader-oriented perspective, we believe that introducing and concluding parts also fulfil essential roles in the appreciation of an article, because they reflect the authors' "storytelling" strategy (Swales, 1981). We think that it could be interesting to study the rhetorical moves of these parts in future work, which is why we introduce the label outline for sentences explaining a paper's structure or introducing figures. We do not believe this category to contain candidates for overclaiming, yet it is necessary to use it if we want to annotate full

²²In a paragraph below this example, the potential harmfulness of GPT-3 answers is evoked, but only to explain that, in contrast, "[InstructGPT models] make up facts less often, and show small decreases in toxic output generation"; overall, the rest pursues the intent to present InstructGPT as a better and safer model, without questioning the nature and impact of *remaining* harmfulness.

²³It is frustrating to note that the real number of gumballs is never revealed in the paper (actually, the reader could compute it from the reported error percentages, but this is not straightforward at all) ... Joking aside, this also prevents the reader from knowing if the predictions were over- or underestimating the real number, or from any advanced analysis of the systems reasoning.

²⁴Emphasis added.

²⁵According to Wikipedia's definition: "A superintelligence is a hypothetical agent that possesses intelligence far surpassing that of the brightest and most gifted human minds." (<https://en.wikipedia.org/wiki/Superintelligence>).

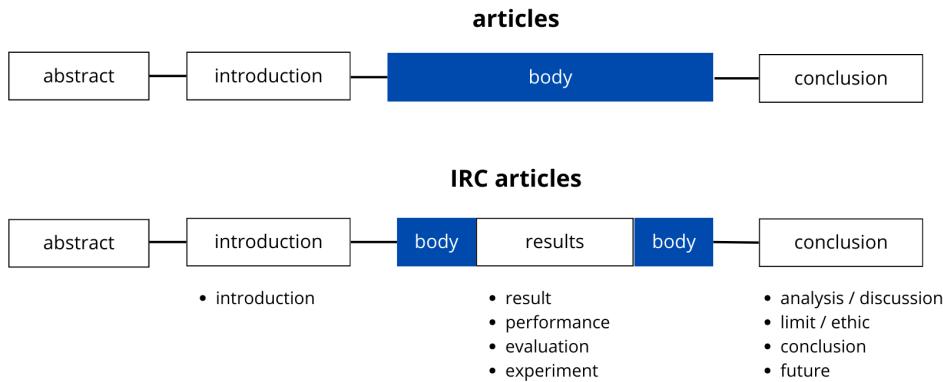


FIGURE 2.10: A simple model for NLP papers structure representation. Above: we call *body* of the paper all the parts between the introduction and the conclusion. Below, we call *IRC papers* those for which we can retrieve an introduction, results, and a conclusion section using the listed keywords (pattern-matching in a paper's section titles).

introductions of papers. Also, we decide to adopt a simple model of NLP papers, presented in Figure 2.10.

Validating the taxonomy

category	definition
context-AIC	claims providing context, background or explanations to the reader (only in AIC)
contribution-AIC	claims depicting the authors' contributions, objectives, the outcome of their work, and key elements of description of this outcome (only in AIC)
outline-AIC	sentences used to draw the outline of a paper or to explain the content of a figure or table (only in AIC)
result	claims of results, either experimental or non-experimental; also analysis and discussion of these results, authors' opinions or important arguments
impact	claims of observed or anticipated impact of the presented work on people / on the scientific community
directions	claims announcing future developments planned or suggested by the authors, possible continuations to the presented work
limitations	claims of observed or anticipated limitations, flaws, drawbacks of some aspects of the presented work
no claim	every other sentence of a paper, including methodology, technical details, etc.

TABLE 2.1: Final version of the taxonomy for claim category identification. AIC stands for *Abstract, Introduction and Conclusion*.

The taxonomy was validated through multiple annotation phases on the corpus of papers' sentences presented in Chapter 3. The final version described in Table 2.1 is the result of the fourth one²⁶. In the first two versions, there were less categories (e.g. PROSP broadly corresponded to impact and directions) and the focus was on

²⁶Note the "-AIC" suffix added to context, contribution and outline in Table 2.1: it acted as a remainder for annotators to only look for these categories in Abstract, Introduction and Conclusion sections of papers. For a better readability, we do not employ these suffixes in the rest of the report.

the "modality" of a sentence²⁷, however we switched to rhetorical functions in phase 3. In this phase, we also tried to add a layer of sub-labels, echoing the structure adopted by Coan et al. (2021); however, this resulted in more confusion among annotators²⁸.

In every annotation phase, the annotators were provided with a set of sentences from a balanced ACL-ArXiv selection of IRC papers (cf. Figure 2.10) in which only relevant sections were included²⁹. All annotators³⁰ were member of the *Sémagramme* team, and followed the guidelines designed for the relevant phase. They were also provided with a user guide on how to use Doccano³¹, an easy-to-use open-source annotation tool. One of the first things we observed is that sentences often required more than one label³², which led us to frame the problem as a multi-label annotation task. This allowed for a better adjustment to the corpus, but also highlighted disagreements between some pair of annotators. After collecting annotations for a given phase, we computed an overall agreement score using Krippendorff's alpha³³ as well as pairwise agreement scores using Cohen's kappa³⁴; the detail of all scores is presented in Table 2.2. Following an inspection of annotators' remarks, mistakes and disagreements, we adapted the guidelines and categories whenever necessary. The last agreement score of 0.81 (both for Krippendorff's alpha and Cohen's kappa) reflects the final state of the taxonomy as understood by the two main annotators, which, we believe, was sufficient to validate its categories.

²⁷In a few words: POS for *positive* assertions (from the authors' perspective), NEG for negative ones, FACT for statements of facts, and PROSP for prospective assertions.

²⁸The guidelines for phase 3 are given in Appendix A.2

²⁹I.e., only those whose title matched the patterns in Figure 2.10. This allowed to reduce the share of no claim in the annotations, so that annotators could go through less papers to find claims from other categories.

³⁰Fanny Ducel, Karén Fort, Maxime Amblard, Amandine Decker, Valentin Richard and myself.

³¹See <https://doccano.github.io/doccano/>.

³²E.g. We introduce a model for sentiment analysis (contribution) and report competitive performance with other state-of-the-art models (result). See Figure B.1 for a visualisation of frequently co-occurring labels.

³³We used the simpledorff implementation (<https://github.com/LightTag/simpledorff/tree/master>). Krippendorff's alpha ranges between 0-1 (the higher, the better). It is considered "acceptable" by some when above 0.67, or 0.8 (<https://www.statisticshowto.com/krippendorff-s-alpha/>).

³⁴We used the scikit-learn implementation (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html). The same remarks go as for Krippendorff's alpha (see previous footnote).

phase	set of labels	annotators	#sents	#papers	$\alpha(\uparrow)$	$\kappa(\uparrow)$
1	POS, NEG, FACT, PROSP, NC	a1, a2, a3, a4	987/246*	10/4*	0.58	0.09-0.70
2	POS, NEG, FACT, PROSP, NC + RW	a1, a2, a5, a6	176	2	0.67	0.62-0.73
3	CONTR, RES-experimental, RES-analyis, RES-knowledge, LIM, PROSP-directions, PROSP-impact, NC + RW	a1, a2	622	4	0.57	0.57
4	context, contribution, result, limitation, impact, directions, nc + rw	a1, a2	289	2	0.81	0.81

TABLE 2.2: Statistics on each annotation phase. α denotes Krippendorff's alpha, κ denotes Cohen's kappa min-max range (when there are more than two annotators). *In the first annotation phase, two annotators worked on 10 papers, while the two others worked on 4 papers (included in the first two's 10).

Chapter 3

Building an annotated corpus of NLP papers claims

In this chapter, we report on the creation of a corpus containing metadata and 15M+ sentences from 105k+ papers written in English, originating from two important sources of NLP papers. We also present different models for claim classification, the best of which is used for inference on the corpus. Additional annotations in terms of sentence-level and aspect-level certainty are also produced using models created by Pei and Jurgens, 2021.

3.1 Collecting sentences from NLP papers

The first practical step in our study is naturally to collect data to work on. In the domain of NLP, two major repositories grant readers a free access to scholar documents: the ACL Anthology¹ and the pre-print platform ArXiv².

3.1.1 Paper sources: ACL vs. ArXiv

The ACL Anthology is a repository named after the Association for Computational Linguistics (ACL)³, the highest-impact organisation and conference in NLP; however it contains papers from both ACL-sponsored (ACL, EMNLP, SemEval, etc.) and non ACL-sponsored events (COLING, EAMT, LREC, etc.). In the words of Bird et al. (2008), "the ACL Anthology represents the NLP community's most up-to-date and long-standing freely accessible research repository", which is still true today, as the anthology counts more than 100k⁴ papers. Unsurprisingly, this corpus is often used in studies on NLP literature, e.g. in Mariani, Francopoulo, and Paroubek (2019) and Ducel (2022).

On the other hand, ArXiv is a repository of more than 2.4M⁵ scholarly pre-prints in a variety of domains⁶ (including NLP) created by Paul Ginsparg in 1991, and now maintained by Cornell University. An essential difference with the ACL Anthology

¹See <https://aclanthology.org/>.

²See <https://arxiv.org/>

³See <https://www.aclweb.org/portal/about>.

⁴100,267 papers as of August 2024

⁵2,445,865 papers as of March 2024

⁶"physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics" (on ArXiv's homepage).

is that pre-prints are not peer-reviewed, even if they are submitted to a moderation process⁷, which means they will either be published later "as is" in an external source, undergo modifications before publication, or, possibly, remain at the pre-print stage on ArXiv forever. In the last case, this could mean that the pre-print was rejected by reviewers because it didn't meet the standards of the journals or conferences it was submitted to.

In this work, we will use data from both sources. This will allow us to go beyond an "ACL-centric" view of NLP productions, with papers that were published in different, maybe smaller journals or conferences⁸. At the same time, we may also collect pre-prints that were never published after being posted on ArXiv. In both cases, it is interesting to wonder whether (and how) the content of these papers differs from that of ACL Anthology ones.

3.1.2 Data collection

While the ACL Anthology's website allows users to browse through thousands of papers and posters, only the metadata and abstracts are directly downloadable. Therefore, we used the work of Rohatgi et al. (2023) who published the ACL OCL corpus⁹, "a scholarly corpus derived from the ACL Anthology to assist Open scientific research in the Computational Linguistics domain", enriching the ACL Anthology's data with PDF files, citation graphs and structured full-text with figures, links and references. In order to extract the content of papers, the authors used GROBID (*GROBID 2008–2024*), a library that allows to parse PDF documents into structured XML files. We downloaded the latest available version of these XML files from HuggingFace¹⁰, and obtained XML files and metadata for 71,286 papers published between 1952–2022. The metadata includes paper id, title, authors, full text, number of citations, url, publisher, year, volume, language, doi, isbn, topic, etc.

Then, metadata for the entire collection of ArXiv papers is provided by Cornell University on Kaggle¹¹ (submitters, 2024), however we are only interested in NLP-related papers, i.e., those having category *Computer Science: Computation and Language (cs.CL)*. We downloaded metadata for 59,167 of such papers. Then, in order to avoid redundancy, we filtered out papers whose title was already found in the ACL Anthology, which left us with 33 815 "pure" ArXiv papers. Second, we used gsutil¹² to request PDF documents associated with the papers IDs in Google Cloud Storage (GSC) buckets¹³. When multiple versions of a paper were available, we downloaded the latest one. Finally, we used GROBID¹⁴ to parse these PDF documents into XML files. Some errors were encountered during the process, with some unavailable PDFs in the GSC buckets (even when attempting to retrieve older versions) and GROBID errors on heavy PDFs (TimeOut errors). In the end, we successfully extracted 28,624 XML files with non empty, readable content.

⁷"Submissions to arXiv are subject to a moderation process that classifies material as topical to the subject area and checks for scholarly value. Material is not peer-reviewed by arXiv - the contents of arXiv submissions are wholly the responsibility of the submitter and are presented "as is" without any warranty or guarantee" (<https://info.arxiv.org/about/index.html>).

⁸For instance, the International Journal of Computer Science and Information Security (<https://sites.google.com/site/ijcsis/>) lists NLP among its topics of interests, but is not covered by the ACL Anthology.

⁹Available on GitHub (Rohatgi, 2022).

¹⁰See <https://huggingface.co/datasets/WINGNUS/ACL-OCL>.

¹¹See <https://www.kaggle.com/datasets/Cornell-University/arxiv/data>.

¹²See <https://cloud.google.com/storage/docs/gsutil?hl=fr>.

¹³See https://cloud.google.com/storage/docs/json_api/v1/buckets.

¹⁴Because of errors with latest version 0.8.0 (as of April 2024), we used the previous version 0.7.0.

3.1.3 Data pre-processing and content extraction

Then, we parsed the collected XML files and extract their textual content, encapsulated in the corresponding paper sections. For this step, we left aside papers whose XML files:

- were empty / had XML structure problems (possibly because of PDF-XML conversion errors)
- corresponded to posters (as can be found in the ACL Anthology)
- were not in English (other encountered languages included French, Danish, and Chinese from Taiwan)
- had improper content in terms of length or proportion of alpha-numerical characters¹⁵

Finally, the extracted content was segmented in sentences using Spacy's model `en_core_web_sm`¹⁶.

3.1.4 Manual annotation

After conducting multiple annotation phases on the corpus (cf. Section 2.2.2), we decided to validate the taxonomy and guidelines used for the fourth trial. Then, the two main annotators¹⁷ annotated a total of 14,792 sentences found in 158 papers. We took a random selection of papers following the IRC-structure, and made sure that they were balanced in terms of corpus of origin (52.5% ACL, 47.5% ArXiv) and year of publication (15.2% < 1994, 29.7% between 1994-2004, 27.2% between 2004-2014, 27.8% > 2014). Each paper was annotated by only one annotator. The task took an estimated 26.7 hours¹⁸ to complete. This resulted in the collection of 15,401 labels whose distribution is shown in Figure 3.1, with 595 sentences counting more than one label. We note the prevalence of the `no_claim` category, and the rarity of `outline` and `impact` ones. When ignoring `no_claim` labels, we note that `result`, `context` and `contribution` claims are the most frequent in papers.

3.1.5 Model inference and linguistic analysis

This manually annotated corpus has been used to train a classifier, as will be further detailed in Section 3.2. The best model we obtained was used for inference on the rest of the corpus, as well as models for certainty annotation (see Section 3.3). Finally, linguistic analyses were ran on the collected sentences using the library `spacy` (see Section 4.1.2).

3.1.6 Data distribution

The resulting corpus comprises:

- metadata on 105,101 ACL and ArXiv papers, as well as information on the types of claims they contain (e.g., the number of `result` claims)

¹⁵We excluded sentences whose ratio of alpha-numerical characters over the total number of characters was below 0.5, as we found this to exclude non-english papers as well as errors in XML documents.

¹⁶See https://spacy.io/models/en#en_core_web_sm.

¹⁷Fanny Ducel and myself.

¹⁸Based on the estimated average time needed to read and annotate one sentence: 6.5 seconds.

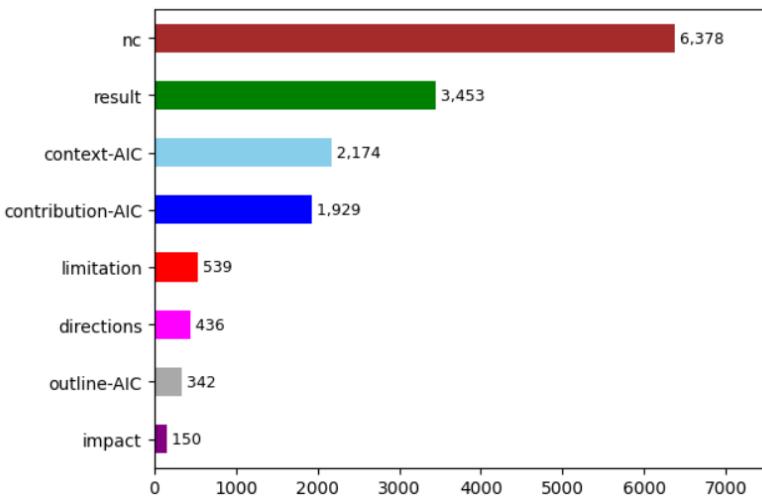


FIGURE 3.1: Distribution of labels collected during the annotation phase.

- machine-readable XML files for 30,433 ArXiv papers
- 15,850,809 sentences annotated in terms of claim category (14,972 of which were manually-annotated), sentence-level and aspect-level certainty
- linguistic information (e.g., sentence length, number of adjectives, number of PERSON entities) on 7,574,460 sentences¹⁹

It is freely available as a HuggingFace²⁰ dataset and reusable under the CC BY-NC-SA 4.0 license²¹.

3.2 Automatic classification of claims

3.2.1 Problem modelisation

Using the annotated corpus we collected (see Section 3.1.4), we now want to train a model to predict the correct category of every sentence in a paper. We start by framing the problem as a multi-label sentence classification task, and quickly review some strategies used in works with similar objectives.

Also dealing with imbalance in terms of category distribution in their dataset, Coan et al. (2021) assess the performance of both "shallow" discriminative classifiers (Logistic Regression²², Support Vector Machines²³) and more recent "deep" transfer learning architectures (ULMFit (Howard and Ruder, 2018), RoBERTa (Liu et al., 2019)) for contrarian claims classification. They find the ensemble of fine-tuned RoBERTa_{large} + Logistic Regression (average prediction) to perform best, with a macro-averaged F1 score of 0.79, and F1-scores ranging between 0.74-0.91 at the "super-claims" level²⁴. They note however that fine-tuning the RoBERTa language

¹⁹We extracted this linguistic information for sentences with only one claim label (see Section 4.1.2). For computational economy, we didn't run the model on the entire no claim collection, but on a representative sample of 2,284,542 no claim sentences.

²⁰<https://huggingface.co/datasets/ClementineBleuze/CNP>

²¹See <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

²²See for instance https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

²³See for instance <https://scikit-learn.org/stable/modules/svm.html>.

²⁴The F1 score is also very high for the neutral class "no claim": 0.93.

model itself (as opposed to simple additional layers for classification) did not produce satisfying results. In the domain of cybersecurity, Ameri et al. (2021) also used a BERT-based classifier (Devlin et al., 2019) to identify cybersecurity feature claims in industrial documents, as part of a larger cybersecurity vetting workflow. Working with 5 claim classes, they obtained a macro-averaged F1-score of 0.93 when fine-tuning BERT_{base} with 3 dense layers and a learning-rate of 4e-6. This is higher than CNN, LSTM, BiLSTM, GPT-2, ELMo (Peters et al., 2018) and ULMFit-based models they also implemented in their study.

This inspires us to make use of BERT-based models to take advantage of our small, yet high quality annotated corpus as well as BERT's pre-trained representations, using fine-tuning. As many BERT-based variants are available for reuse, we will compare a few of them to check for significant differences, namely:

- RoBERTa (Liu et al., 2019) which optimises part of BERT's training procedure²⁵ and was found to exhibit higher performance on state-of-the-art benchmarks,
- DeBERTa (He et al., 2021) which further builds on RoBERTa, and exhibits higher performance while requiring less training data,
- and SciBERT (Beltagy, Lo, and Cohan, 2019), a BERT model trained on full-text biomedical and computer science papers, which seems particularly relevant for our task.

As a baseline against which to compare these BERT-models, we propose as Coan et al. (2021) to use simple Logistic Regression and SVM models, which, we hypothesise, can maybe leverage sufficient information to recognise claim classes exhibiting stereotypical or salient features. However, an aspect in which our task differs from "usual" sentence classification settings where sentences are independent from one another, is that our data is *sequential*, which echoes with the proposition of Dermontcourt and Lee (2017) to use the dedicated term of *Sequential Sentence Classification*. Under this perspective, models taking advantage of the contextual ordering as well as label information of sentences have been developed ((Cohan et al., 2019), (Brack et al., 2024), (Tokala et al., 2023)).

As this involves more complex architectures and requires a non-negligible development time, we discuss a few attempts of implementation in Section 3.2.4. In the meantime, we propose a more straightforward encoding of a sentence's context by simple concatenation of relevant information (preceding sentences, section name) to its input representation to the models, as will be detailed in following subsections. In all upcoming experiments, we use a 80-10-10 split of the 14,792 collected sentences as training, evaluation and test datasets²⁶. We run the experiments on the Grid5000 platform²⁷, using clusters located in Nancy.

3.2.2 Baseline models

For baseline models, we use sklearn²⁸ implementation of *LogisticRegression* and *SVM*. The parameters for *SVM* (sigmoid kernel, C = 5) are optimised via grid search

²⁵Differences with BERT include adjustment of hyperparameters, extension of training time, replacement of fixed-masking with dynamic-masking and removal of initial Next Sentence Prediction (NSP) objective to keep the Masked Language Modelling (MLM) objective only.

²⁶We ensured a balanced distribution of classes among the splits, by equally sharing sentences with a unique label, and then those with multiple categories (as in one "mixed" class).

²⁷See <https://www.grid5000.fr/w/Grid5000:Home>.

²⁸See <https://scikit-learn.org/stable/index.html>.

on the validation split of the dataset, and reused in all following implementations. In order to adopt the multi-label setting, both classifiers are wrapped into *OneVsRestClassifier* objects²⁹. Three input types are considered (see Table 3.1). The intuition is that the section in which a sentence occurs constitutes relevant contextual information in order to classify the sentence correctly.

configuration	input encoding
text	"target_sentence"
prefix_text	" section: section_name, text: sentence"
prefix_text_SEP	"section_name [SEC] sentence"

TABLE 3.1: Three information encoding configurations for baseline models.

Finally, two vectorisation methods are compared: a simple bag-of-words approach, and TF-IDF vectorisation³⁰. In order to obtain the final labels, we apply a Sigmoid³¹ layer to model predictions (logits), and apply a small function to meet the following constraints: (i) every sentence must have at least one label, and (ii) no `claim` is incompatible with other labels. Results in terms of F1-score are reported in Table 3.2. We observe important differences between categories, with less populated class `impact` suffering from very low scores, and most populated class `no_claim` reaching a maximum of 0.81 F1 with SVM. This can also be visualised in Figure 3.2. After that, categories `context` and `outline` are the two best predicted classes, which could be linked with stereotypical formulations, e.g. *In Section X, we will present Y* for `outline`. Overall, we note an improvement of results when prefixing sentences with their section title, which also seems to work better in `prefix_text` setting rather than `prefix_SEP`, and an advantage of SVM over Logistic Regression.

model	context	contr.	result	impact	dir.	lim.	outline	nc	avg
lr+cv+text	0.50	0.40	0.57	0.21	0.26	0.07	0.54	0.66	0.54
lr+cv+prefix_text	0.69	0.51	0.64	0.11	0.45	0.15	0.72	0.79	0.66
lr+cv+prefix_SEP	0.64	0.43	0.58	0.09	0.24	0.18	0.34	0.59	0.54
lr+tf+text	0.49	0.41	0.57	0.24	0.25	0.09	0.43	0.66	0.54
lr+tf+prefix_text	0.68	0.50	0.64	0.21	0.39	0.12	0.68	0.78	0.66
lr+tf+prefix_SEP	0.67	0.49	0.64	0.21	0.46	0.12	0.67	0.78	0.65
svm+cv+text	0.54	0.41	0.64	0.00	0.40	0.10	0.54	0.72	0.59
svm+cv+prefix_text	0.70	0.57	0.64	0.11	0.47	0.18	0.67	0.81	0.68
svm+cv+prefix_SEP	0.69	0.53	0.58	0.11	0.48	0.18	0.62	0.79	0.66
svm+tf+text	0.52	0.41	0.63	0.22	0.40	0.11	0.61	0.72	0.59
svm+tf+prefix_text	0.66	0.53	0.67	0.21	0.47	0.19	0.62	0.81	0.68
svm+tf+prefix_SEP	0.67	0.54	0.67	0.21	0.47	0.18	0.65	0.81	0.68

TABLE 3.2: F1 scores for claim classification per model setting (lines) and claim category (columns 1-8). lr = logistic regression , cv = count-vectorisation, tf = tf-idf vectorisation.
avg = weighted average of all F1-scores.

²⁹See <https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html>.

³⁰Because these methods do not encode the position of tokens in the input, we do not add surrounding sentences to the context.

³¹See <https://pytorch.org/docs/stable/generated/torch.sigmoid.html>.

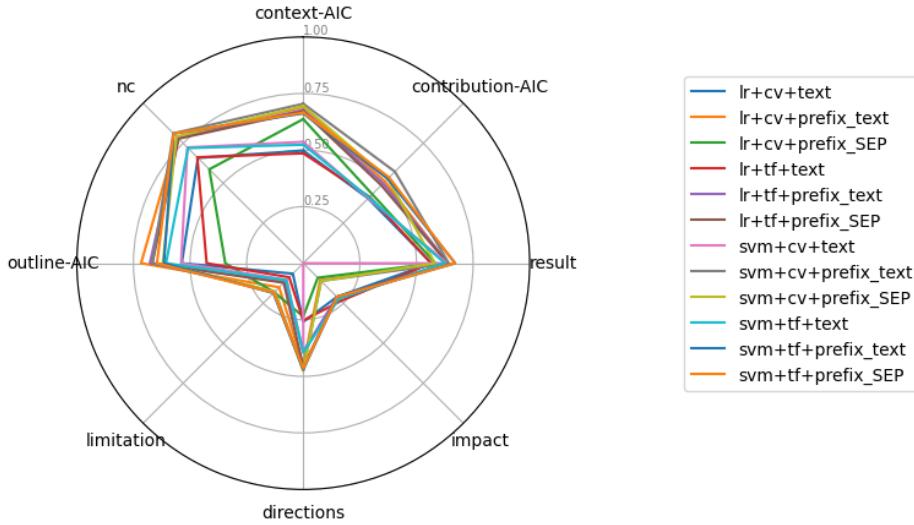


FIGURE 3.2: Radar plot of F1-scores per model setting and per category. The more consistent a model performance remains across classes, the closer its associated shape should get to a regular polygon. We observe that all models struggle to predict less frequent classes of the corpus limitation, directions and impact (scores below 0.50).

3.2.3 BERT-based models

As for BERT-based models, we use the available HuggingFace checkpoints to access RoBERTa_{base}³², DeBERTa_{base}³³ and SciBERT_{uncased}³⁴. The *AutoModelForSequenceClassification* and *AutoTokenizer* classes from library *transformers* then allow us to load the corresponding base models and tokenizers, with an additional linear layer for classification³⁵.

As BERT's tokenizer already has its predefined separators [SEP] (to separate sentences - the last of which is used for classification) and [CLS] (at the beginning of input), we add our section separator [SEC] to the vocabulary, and experiment with three configurations for context augmentation: with section information only, with section and preceding + next sentence, and with section and two preceding sentences; corresponding input encodings are presented in Table 3.3. Because of BERT's maximum token length of 512, we limit ourselves to a context size of 2.

configuration	input encoding
prefix_SEC	[CLS] section_name [SEC] target_sentence [SEP]
prefix_SEC_lr	[CLS] section_name [SEC] sentence_-1 [SEP] target_sentence [SEP] sentence_+1 [SEP]
prefix_SEC_ll	[CLS] section_name [SEC] sentence_-2 [SEP] sentence_-1 [SEP] target_sentence [SEP]

TABLE 3.3: Three context encoding configurations for BERT-based models.

After conducting some initial tests, we set the following parameters for all experiments: nb_epochs = 15, batch_size = 8, weight_decay = 0.1, learning_rate = 1e-5, save_steps = 500 (evaluation at steps-level). After training, we store the best model

³²See <https://huggingface.co/FacebookAI/roberta-base>.

³³See <https://huggingface.co/microsoft/deberta-v3-base>.

³⁴See https://huggingface.co/allenai/scibert_scivocab_uncased.

³⁵To perform multi-label classification, we specify problem_type = "multi_label_classification" in *AutoModelForSequenceClassification.from_pretrained()* arguments.

in terms of weighted F1-score on the validation set. The same constraints as explained in Section 3.2.2 are applied to model logits. The results are reported in Table 3.4, and can be visualised in Figure 3.3. We notice that once again, less populated classes suffer from lower scores; however some categories like no claim, context and outline reach F1-scores above 0.90, which is close to the best scores of Coan et al. (2021) and Ameri et al. (2021). The addition of surrounding sentences to the inputs also improves the performance, with a slight advantage for the *prefix_SEC_ll* setting considering left context only. Overall, RoBERTa arrived behind DeBERTa and SciBERT, and we take SciBERT+prefix_SEC_ll as the overall best performing model.

model	context	contr.	result	impact	dir.	lim.	outline	nc	avg
RoBERTa+prefix_SEC	0.84	0.79	0.80	0.30	0.69	0.46	0.79	0.91	0.83
RoBERTa+prefix_SEC_lr	0.88	0.87	0.86	0.37	0.78	0.58	0.87	0.94	0.88
RoBERTa+prefix_SEC_ll	0.90	0.84	0.85	0.56	0.77	0.51	0.85	0.93	0.87
DeBERTa+prefix_SEC	0.81	0.76	0.82	0.00	0.73	0.39	0.85	0.92	0.82
DeBERTa+prefix_SEC_lr	0.90	0.85	0.85	0.48	0.81	0.54	0.84	0.94	0.88
DeBERTa+prefix_SEC_ll	0.91	0.86	0.87	0.45	0.83	0.60	0.90	0.94	0.89
SciBERT+prefix_SEC	0.86	0.81	0.80	0.57	0.74	0.44	0.82	0.92	0.84
SciBERT+prefix_SEC_lr	0.93	0.86	0.87	0.48	0.80	0.54	0.85	0.94	0.89
SciBERT+prefix_SEC_ll	0.93	0.87	0.86	0.52	0.82	0.51	0.88	0.95	0.89

TABLE 3.4: F1 scores for claim classification per model setting (lines) and claim category (columns 1-8). avg = weighted average of all F1-scores.

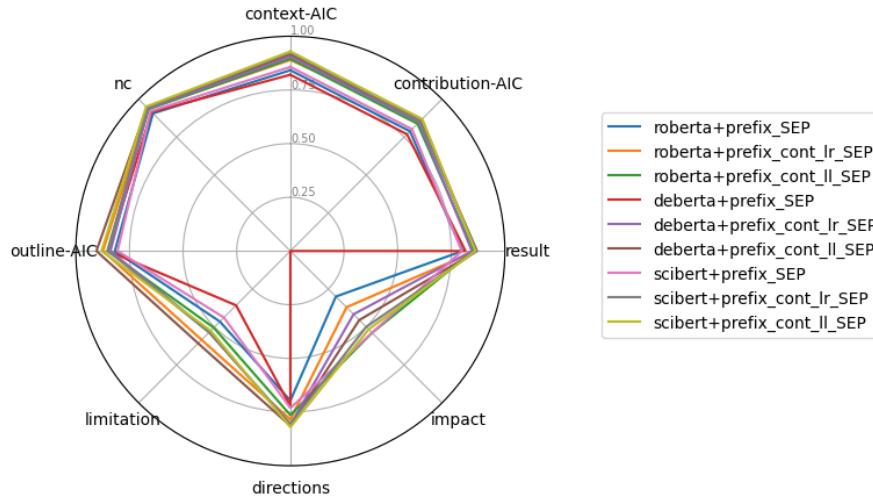


FIGURE 3.3: Radar plot of F1-scores per model setting and per category. The more consistent a model performance remains across classes, the closer its associated shape should get to a regular polygon. We observe a coherent performance on most populated classes, but divergences on less-populated ones (in particular for *impact* and *limitation*).

3.2.4 Custom models experiments

Using such fine-tuned BERT models, one model pass during the training phase corresponds to the prediction of a single sentence's label. In Cohan et al. (2019), one model pass amounts to the prediction of ten sentences labels, which significantly reduces the cost of inference, while taking advantage of the training data's sequential nature. The process is depicted in Figure C.1 (with 3 sentences instead of 10,

for illustration purposes). However, it is important to note that the authors work with paper abstracts, which often contain shorter sentences: in such a situation, 10 sentences generally do not exceed BERT inputs' limit of 512 tokens; plus, one or two batches of 10 sentences will allow to cover one full abstract, which is obviously not the case for full-text articles. And, as noted by the authors themselves, the very first and last sentences may suffer "side effects", which we can't afford in our 3-sentences setting.

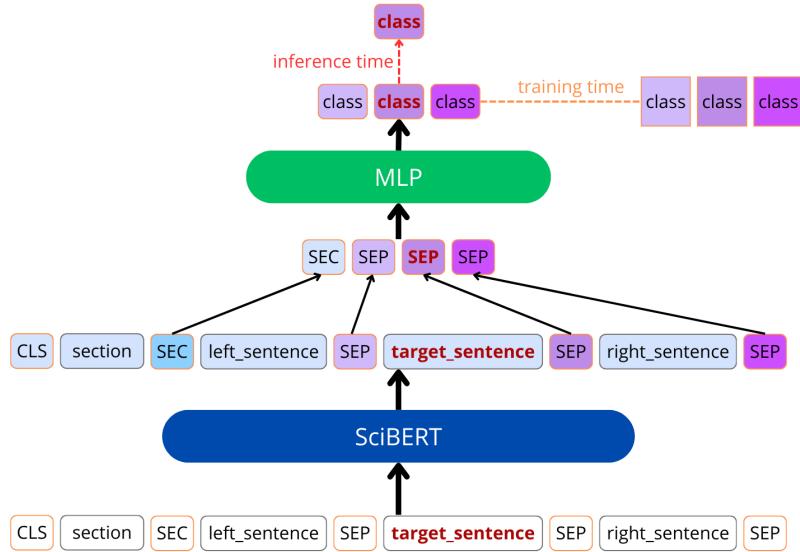


FIGURE 3.4: A custom BERT-based model for contextualised predictions of claim classes (represented in the *prefix_SEC_1r* setting). At training time, the model sees all three ground-truth labels of the sentences and learns to predict them all. At inference time, we only retrieve the prediction that corresponds to the target sentence.

Nevertheless inspired by this approach, we tried to build a custom BERT-based model which learns a target's sentence class based on its actual label *and* on the labels of its surrounding sentences: in other words, one model pass still corresponds to one prediction, but this prediction benefits from thrice as much information. The process is illustrated in Figure 3.4. However, we encountered difficulties in training the model, as the loss would first decrease and then, quickly stagnate without convergence. Overall, we didn't succeed in producing satisfying results. Therefore, we decided to use the best fine-tuned model (i.e., *SciBERT* in *prefix_SEC_ll* setting) for corpus inference, and didn't investigate more deeply other interesting architectures such as those presented in Brack et al. (2024) and Tokala et al. (2023).

3.3 Automatic annotation in certainty

As we explained in Section 2.1.2, we propose to use certainty as an indicator for overclaiming, as a large difference in certainty between related claims could indicate abusive exaggeration. Because of time constraints, we do not actually compute these differences between mapped claims, but rather investigate overall certainty trends.

In the literature, the level or degree of certainty is most often measured on a scale of varying length, from less to most certain, that is, with a single, numerical score. According to Pei and Jurgens (2021), there is however more to the notion of certainty than this "primary judgement of certainty along a continuum from uncertain

to certain", which they refer to as *sentence-level certainty*. Instead, they suggest that certainty is a twofold construct, and introduce an additional dimension of *aspect-level certainty*, which captures a finer-grained description of certainty according to aspects of *Number*, *Extent*, *Probability*, *Condition*, *Framing* and *Suggestion* (see Table 3.5). This is illustrated in Figure 3.5.

aspect	definition
NUMBER	Certainty towards specific quantities
EXTENT	Certainty about the proportion/ratio of properties that make up an object/event or the extent of a change
PROBABILITY	Certainty about the probability that something will occur, has occurred, or is associated with another factor
CONDITION	Situation where something depends on a specific condition, and the condition involves certainty or uncertainty
FRAMING	Certainty about how scientists or journalists themselves frame or interpret the scientific finding
SUGGESTION	Certainty or uncertainty about the implications or future actions for the public or science community

TABLE 3.5: Definitions of aspects, based on Pei and Jurgens (2021).

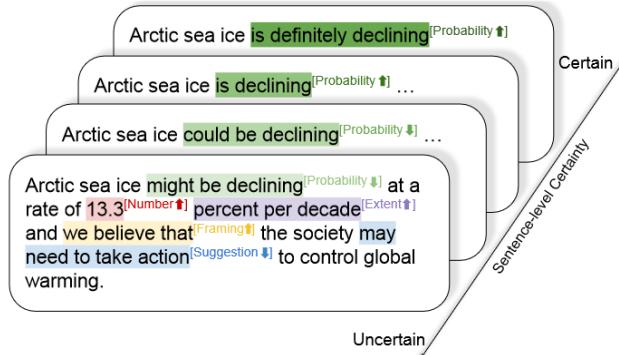


FIGURE 3.5: Certainty is a multi-dimensional construct. Illustration from (Pei and Jurgens, 2021)

As they make their models freely available³⁶, we implement Pei and Jurgens (2021) sentence and aspect-level certainty classifiers, and use them for inference on the corpus. They are also SciBERT-based models, respectively outputting a sentence-level certainty score ranging in [1, 6]³⁷ and aspect-level modalities (certain or uncertain) for aspects that were detected in the input sentence. Because of time constraints, and because the task of manually annotating certainty is very subjective, we do not attempt to fine-tune these models, and use the results as they are.

³⁶See <https://github.com/Jiaxin-Pei/certainty-estimator/>.

³⁷A particularity is that, contrary to other works, the model does not output a score class, e.g. "2", but a float in [1,6], e.g. 2.47.

Chapter 4

Analysis of collected claims

4.1 Inspecting the predictions of claim categories (RQ1)

4.1.1 Distribution of categories

One of the first things we can inspect in our corpus is the type of predicted claims. Table 4.1 reports the number of predicted claims per category (column "total"): we see that no `claim` makes up for roughly two thirds of all sentence labels, while the shape of the distribution exactly matches that of the manually-annotated corpus (cf. Figure 4.1). However, we must keep in mind that the predictions may sometimes be inadequate, as the model used for inference had lower F1-scores on `limitation` and `impact`. After inspecting a few examples (see Table 4.2), the predictions nonetheless seem convincing¹, which is an encouraging first step.

	1	2	3	4	total
nc	10,258,678 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	10,258,678 (63.6%)
result	2,329,273 (92.1%)	197,370 (7.8%)	3,555 (0.1%)	92 (0.0%)	2,530,290 (15.7%)
context	1,412,003 (98.0%)	27,860 (1.9%)	410 (0.0%)	2 (0.0%)	1,440,275 (8.9%)
contribution	989,483 (89.3%)	116,667 (10.5%)	1,454 (0.1%)	11 (0.0%)	1,107,615 (6.9%)
limitation	148,854 (54.4%)	122,275 (44.7%)	2,397 (0.9%)	88 (0.0%)	273,614 (1.7%)
directions	225,839 (86.4%)	33,244 (12.7%)	2,348 (0.9%)	81 (0.0%)	261,512 (1.6%)
outline	122,040 (85.6%)	19,842 (13.9%)	720 (0.5%)	5 (0.0%)	142,607 (0.9%)
impact	62,426 (55.6%)	47,464 (42.3%)	2,322 (2.1%)	93 (0.1%)	112,305 (0.7%)

TABLE 4.1: Number of predicted claims per category (lines) and per number of predicted labels (columns). For instance, in line 2, column 3: 3,555 claims were assigned the label "result" + two additional ones (which represents 0.1% of all "result" predictions). Percentages in column "total" are relative to the total number of predicted labels.

We obtain a total of 5,868,218 "real" claims, i.e., labels other than `no claim`. A second observation is that sentences often come up with multiple labels; in fact, only 54.4% (resp. 55.6%) of `limitation` (resp. `impact`) labels are assigned without any other label (cf. Table 4.1). Figure 4.2 allows us to visualize most frequent label pairs, e.g. `impact-directions`, `limitation-result` or `outline-contribution`, which were also frequent in the training data. Some predictions for these pairs are presented in Table 4.4.

¹For the `directions` example, additional context might be needed. The sentence is found in section *Conclusion and Future Work*: "The next step is to explore better tagging metrics that correspond more strongly to better grammar induction performance. **A good metric should use all the information we have, including the gold trees, to evaluate.** Finally, we should explore grammar induction schemes that do not rely on prior parts-of-speech, instead learning them from raw text at the same time as deeper structure." in (Headden III, McClosky, and Charniak, 2008).

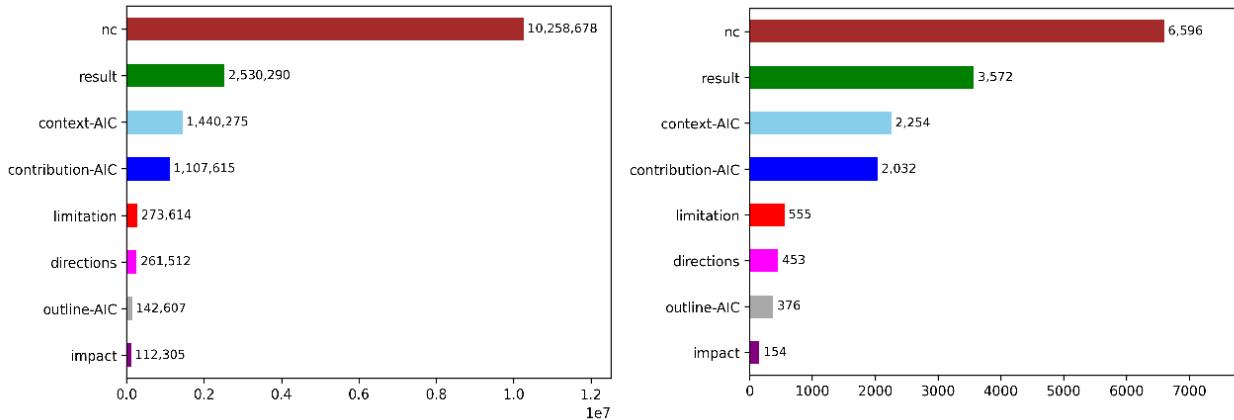


FIGURE 4.1: Distribution of claim labels in the entire corpus (left) and in the manually annotated dataset (right). The entire corpus contains both IRC and non-IRC papers, whereas the annotated one contains only IRC papers; this could explain the increased prevalence of no claim in the first case.

category	sentence example
context	'Whichever approach ends up being taken (as determined primarily by the writing system of the language in question), little attention is usually paid to pronunciation variants stemming from connected speech processes, hypoarticulation, and other phenomena typical for colloquial speech, mostly because the resource is seldom directly empirically derived.'
contribution	'Our architecture is a variant of the Seq2seq model where two different decoders are used instead of only one of the original architecture.'
result	'We can see from this chart that the relative ranking of the models remain the same, except for sizes 1-3, where the probabilistic parser does better (or no worse than) the unlexicalized classifier-based models.'
impact	'We believe that this is a critical moment in the life of dialogue system research, and we anticipate exciting breakthroughs in the near future, leading to systems that are not only useful but also easy to use and accommodating, such that users will prefer them over alternative means of acquiring their information needs.'
directions	'A good metric should use all the information we have, including the gold trees, to evaluate.'
limitation	'While the training process itself does not entail any additional memory or computation overhead compared to vanilla CLIP, the process of generating text rewrites using LLMs can be computationally expensive, requiring significant GPU resources and taking hours for large datasets.'
outline	'In Section 2, we will give an overview of the main advantages of this approach.'
no claim	'We randomly selected a dataset of 150 tweets which were annotated by both annotators for both POS tagging and dependency structures.'

TABLE 4.2: Random selection of claims for each category (model predictions)

Complementary to the overall distributions of claims in Figure 4.1, we also had a look at the average share of each category per paper, as well as the share of papers from the corpus this category appears in. The results are reported in Table 4.3, and show that almost every paper (>95%) contains context, contribution, result and no claim sentences, while only half of them contain impact claims. Interestingly, we note that 28.1% of papers do not contain any limitation claim, which may be influenced by older papers. We also see that 36.9% of sentences of a paper do, on average, contain claims, the majority of which are result ones.

	context	contribution	result	impact	directions	limitation	outline	no claim
average share in paper (%)	10.0	7.7	15.8	0.7	1.8	1.7	1.6	63.1
occurrence frequency (%)	99.0	99.6	98.7	49.0	78.3	71.9	47.0	100.0

TABLE 4.3: Average share of claim categories in papers, and occurrence frequencies in the corpus. For instance: on average, context claims represent 10.0% of a papers' sentences, while 99.0% of papers in the corpus have at least one claim labelled as context.

categories	sentence example
impact+directions	'They supports the theory that using context vectors to measure similarity is feasible and worthy of further research.'
limitation+result	'The performance of our method might have been influenced by the small scale of the Chinese corpus and accuracy of the extracted relations.'
outline+contribution	'This model is described in section 2. Translation results from the four source languages made available for the shared task (es: Spanish, fr: French, de: German, and fi: Finnish) into English (en) are presented and discussed.'

TABLE 4.4: Random selection of claims for most frequent label pairs (model predictions). Further discussion shall establish whether labels **impact** (line 1) and **result** (line 2) are justified. In line 3, a sentence segmentation issue explains the (nevertheless valid) pair of labels.

4.1.2 Linguistic characteristics of claim categories

In order to study the specificities of claim categories from a linguistic perspective, we conducted a few analyses on sentences with a single class label. Using spacy's `en_core_web_sm` model, we extracted and stored information about every sentence's length (#tokens, #characters) and vocabulary (non stop-word lemmas, UD part-of-speech tags² and syntactic relations³, named entities⁴). In terms of sentence length, we find **outline** claims to be significantly shorter than other types (~ 100 characters, ~ 18 tokens/sentence), while **impact** claims are significantly longer (~ 169 characters, ~ 29 tokens/sentence). Average lengths are reported in Figure 4.3.

In terms of lemma distribution, some claim classes exhibit specific enough vocabulary to be recognizable from their wordcloud only (e.g. the 2-gram *future work* for the **direction** class, cf. Figures 4.4 and B.7), yet large overlaps remain. For instance, *model* and *use* both appear in the top ten most frequent lemmas of all classes but **outline** (see Figure B.4). This may explain why word-based representations of sentences such as those we used with baseline models (cf. Section 3.2.2) were only mildly able to learn appropriate claim classes.

As for part-of-speech tags, only little differences are observed between classes: for example, the average **impact** claim contains 7.2 nouns as opposed to 5.8 for the average **limitation** claim (see Figure B.2); however this may reflect sentence lengths differences rather than stylistic particularities. There are stronger divergences when looking at named entities: CARDINAL are more present in **context**, **contribution** and **result**, while DATE, GPE and PERSON are almost only found in

²See <https://universaldependencies.org/u/pos/>.

³See <https://universaldependencies.org/u/dep/>.

⁴See the list of entities supported by spacy in Table ??.

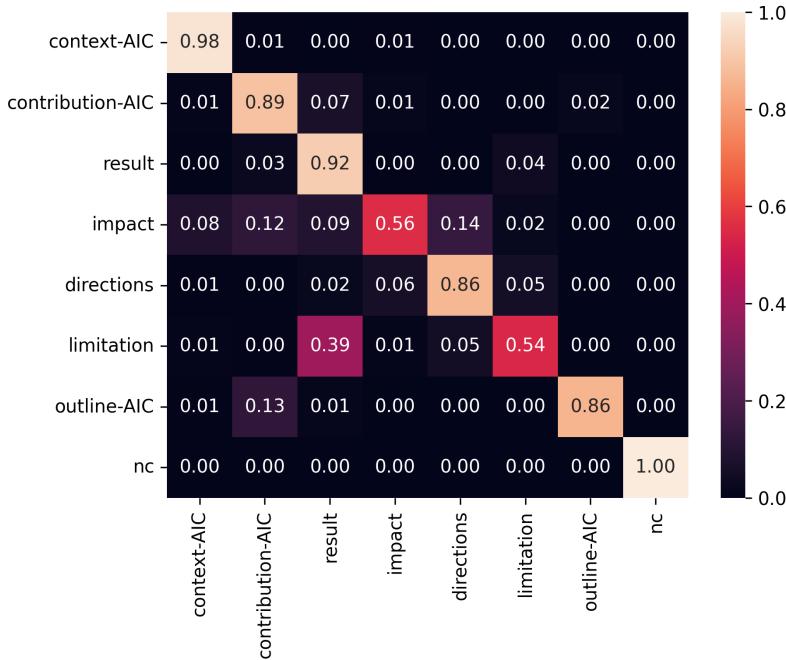


FIGURE 4.2: Co-occurrence matrix for label pairs. For instance, line "impact" indicates that 56% of **impact** labels occur alone, while 12% of them co-occur with **contribution**.

By definition, no label ever co-occurs with no claim.

context (see Figure B.3)⁵. Finally, some dependency relations also reflect categories differences. We find that auxiliaries are more frequent in **impact** and **directions** (1+/sentence) than in other categories, which also correlates with higher verb counts. Quite unsurprisingly, negations are found in almost thrice as more **limitation** claims as in **result** ones, and are very infrequent in **contribution** and **directions**. Finally, markers are more present in **result**, **impact** and **directions**, which is a clue for the presence of subordinate clauses⁶, and could be explained by didactic functions of these classes. Overall, this leads us to hypothesise that syntactic information could be beneficial for claim classification.

4.2 Analysing sentence and aspect-level certainty (RQ3)

4.2.1 Sentence-level certainty

Looking at sentence-level certainty predictions, we see that the majority of sentences obtain very high scores ranging between 4.4-5.3, when the theoretical middle value in a homogeneous distribution would be 3.5 (see Figure 4.5). This seems to indicate that NLP papers are generally written in a quite assertive manner, but it also means that candidate overclaims will be harder to differentiate from other, "regular" sentences. On the other hand, we found corpus examples to be difficult to interpret. In Table 4.5, line 1, we disagree with the very low (minimal) score⁷, and we once

⁵Note the outlier for LAW in **outline** claims: we found that this entity type is triggered by terms as *In Section 2*, which is a mistake from the system.

⁶See <https://universaldependencies.org/u/dep/mark.html>.

⁷*There is still much room for future research* indicates a form of uncertainty relative to the state of future research, but not necessarily to the speaker's position when emitting the claim.

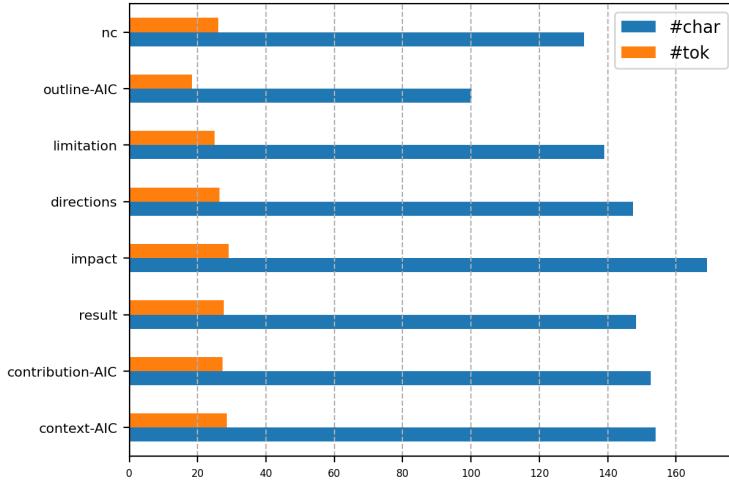


FIGURE 4.3: Average number of characters and tokens per sentence, per claim category (single-labels only).



FIGURE 4.4: Wordclouds for classes contribution and directions.

again realise how difficult the task of certainty annotation on a sample of our corpus would have been.

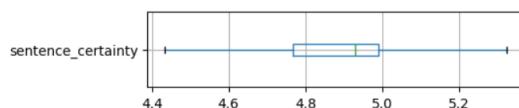


FIGURE 4.5: Boxplot for sentence-level certainty in the entire claims corpus. Sentence-level certainty takes its values in [1,6].

As a sanity-check, we study the relative sentence-level certainty when each aspect is certain or uncertain (see Figure 4.6), and find similar to Pei and Jurgens (2021) that sentence-level certainty is most affected by PROBABILITY and SUGGESTION aspects, while others have little or no effect. Yet, we note higher average certainty levels in our corpus. As can be seen in Figure 4.10, sentence-level certainty is more stable among outline, no claim and contribution categories, with median values close to 5, as is also the case for result claims. On the contrary, limitation, impact, multi-label and directions claims show more variability and lower median values, although still above the intermediate threshold of 3.5. This reflects our intuition that uncertainty is more common in these latter classes in which anticipation plays an important role.

sentence certainty	sentence example
2.14	'However, there is still much room for future research, as the state-of-the-art model on the Fever shared task achieves an accuracy score of 68.21% (Thorne et al., 2018b).'
3.48	'In response to the question, Which room is Monday's meeting in?, there may be no way to choose between A101 and A201 without further world knowledge.'
5.97	'For downward tracing there is no problem with deletion rules using '0' or '[]', whether they are conditioned by an environment or not, and tracing apply-down application of epenthesis rules with an environment is likewise unproblematic.'

TABLE 4.5: Example of claims with minimal, intermediate, and maximal sentence-certainty scores.

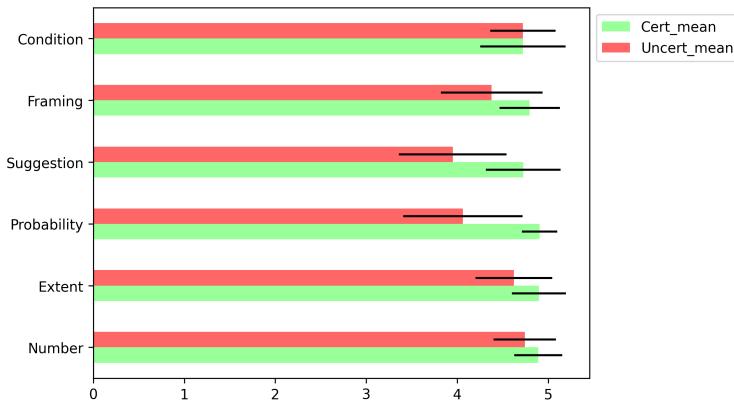


FIGURE 4.6: Relation between average sentence-level certainty and aspect certainty. To be compared with Figure 3 in (Pei and Jurgens, 2021).

4.2.2 Aspect-level certainty

A very first observation about aspect-level certainty is that PROBABILITY and FRAMING are more often present in a sentence than other aspects (~69% and ~42% of cases), whereas CONDITION appears in less than 6% of all sentences. From Figure 4.7, we also note that when present, aspects most likely make sentences *more certain*, with the exception of EXTENT. Still, an issue with the model of Pei and Jurgens (2021) is that no justification is produced. Hence, we are left to hypothesize which clues influenced the decision, even if we happen to question that decision (see for instance Table 4.6, line 2). Another limitation is that for a given sentence, only one label per aspect can be given, in spite of the fact that both certain and uncertain elements could occur.

As we are still interested in the relation between certainty and claim categories, we represent in Figure 4.8 the modality distributions of each aspect per claim category. Similar to previous observations, we see more NUMBER presence in result, no claim, contribution and CONTEXT, most often in the positive setting; while EXTENT comes mostly uncertain, even if relatively less in result claims. This could reflect a desire for precision when reporting findings. PROBABILITY is also present in a majority of cases for each category, with the exception of more factual categories outline and contribution, which are also those with the highest authors presence, as indicated by the FRAMING aspect. Yet, PROBABILITY brings more uncertainty to impact directions and LIMITATION classes, which once again echoes our previous observations. Finally, SUGGESTION clearly allow to differentiate directions and

aspect	modality	sentence example
CONDITION	Certain	'Suppose if there are p instances of Wikipedia word xi in the reviews.'
FRAMING	Certain	'For supertagging, we mirror the features from stacking to the best extent possible given the more limited information that is contained in the supertags to begin with.'
SUGGESTION	Uncertain	'The class-based LM can be updated or extended to support other types of named entities or domains.'
Probability	Uncertain	'Furthermore, language models that are often used to analyze clinical texts can encode broader societal biases [1].'
EXTENT	Uncertain	'The salience score of a sentence is derived partly from the salience of vocabulary items in the document and partly from its position in the document structure (e.g. section-initial, paragraphinternal, and so forth) and the salience of the surrounding sentences.'
NUMBER	Certain	Consequently, AlephBERT generates 3 embedded vectors -the vectors associated with the split word pieces are averaged to form a single contextualized vector.

TABLE 4.6: Random selection of claims where each aspect is present. In bold, we show elements which, *we believe*, influenced the model’s predictions. In line 1, we note that the sentence barely even makes sense on its own.

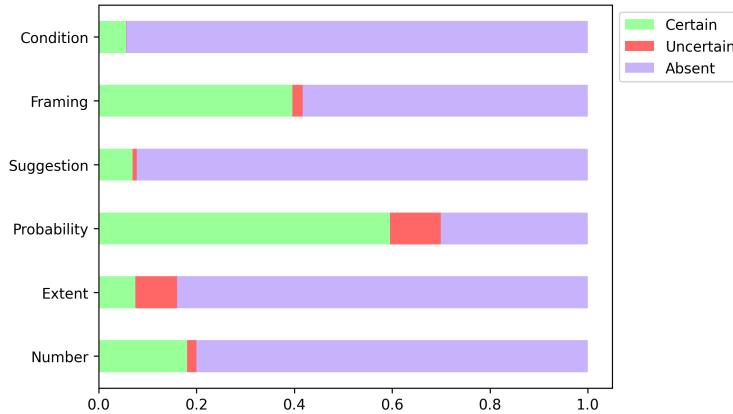


FIGURE 4.7: Distribution of modalities (Absent / Certain or Uncertain) over the entire corpus, per aspect.

impact from other categories, with expression in 62.5% (resp. 34.9%) of sentences. As for CONDITION, this last aspect remains rather infrequent and less visibly connected to claim categories.

4.3 Investigating the effects of papers metadata

4.3.1 Source: ACL vs. ArXiv (RQ3)

When comparing claim predictions in ACL vs. ArXiv papers, we find the average total number of claims/paper to be higher in the ArXiv corpus (~ 67 , against ~ 45 for the ACL corpus). The same goes with the average number of sentences/paper (~ 192 , against ~ 131), although the ratios between claims and sentences are close: on average, 30% (resp. 33%) of sentences are actually claims in ArXiv (resp. ACL) papers. As a consequence, whereas ArXiv papers represent only a third (32.2%) of all papers for which we gathered metadata, their content accounts for 41.1% of all extracted sentences.

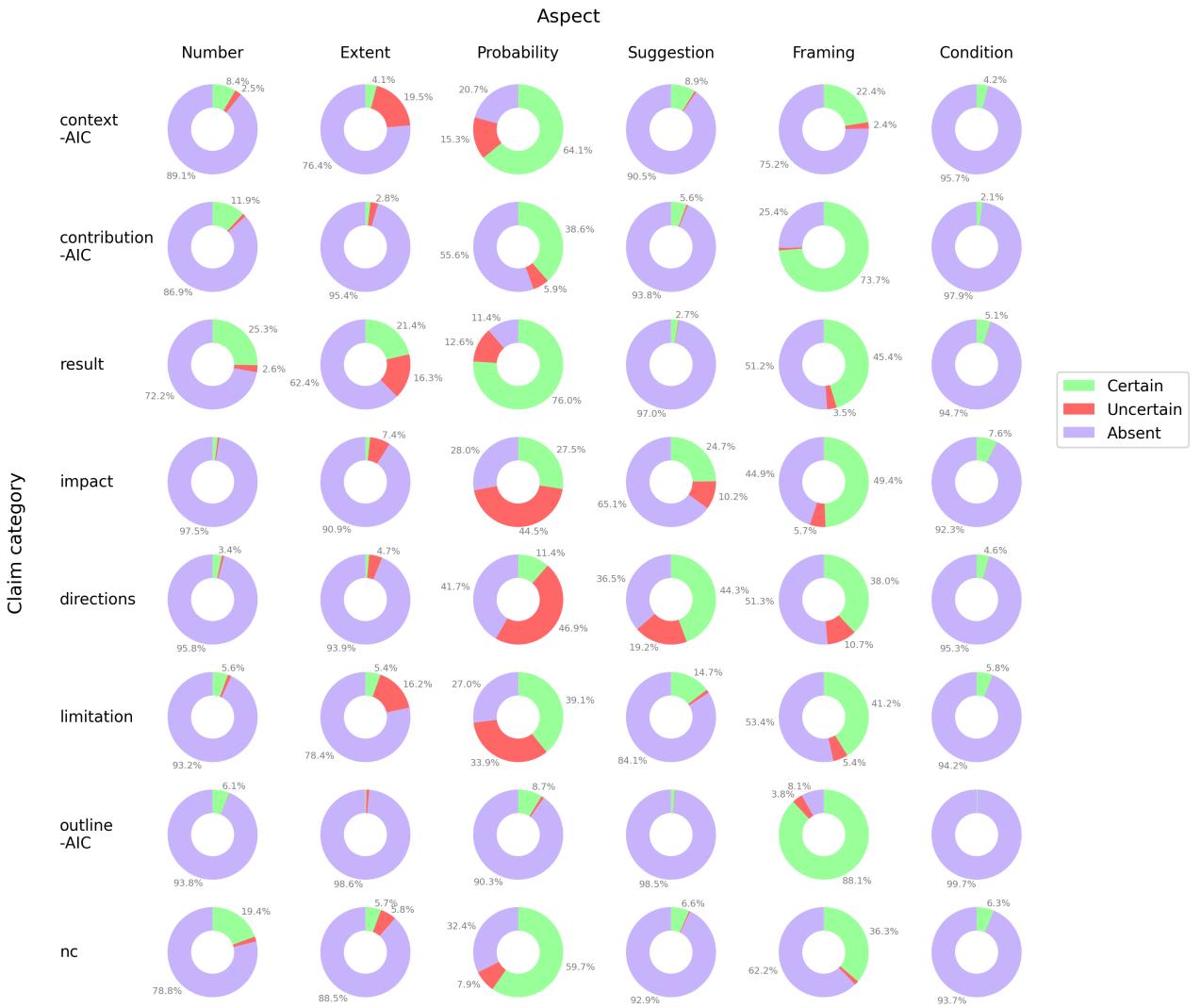


FIGURE 4.8: Modality (Absent / Certain or Uncertain) distributions over the entire corpus, per aspect and claim category.

Figure 4.9 shows that some claim categories diverge from this theoretical 41.1%-58.9% distribution: ArXiv papers contain less directions and outline claims than expected, but more limitation and impact ones. This observation remains true when considering category occurrences: 55.2% (resp. 75.3%) of ArXiv papers contain at least one impact (resp. limitation) claim, against only 46.0% (resp. 70.2%) for ACL papers (cf. Table B.2). This could indicate different writing styles from authors, or maybe reflect conference standards for ACL papers⁸. On the other hand, the average shares that each category occupies within an article do not vary much between ACL and ArXiv papers (cf. Table B.1). Still, we wondered whether the expression of certainty differed between the two sources. But, as shown in Figure 4.10, no significant differences are observed, and sentence-level certainty is quasi-equivalent in ACL and ArXiv papers.

⁸Remember that we removed duplicates from our corpus, which means that we only considered papers from ArXiv which are *not* already found in the ACL Anthology.

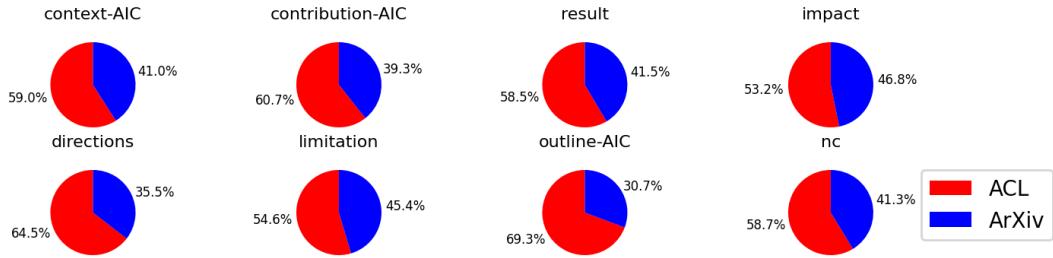


FIGURE 4.9: Distribution of claims according to their source (ACL vs. ArXiv) and category. Overall, 41.1% (resp. 58.9%) of extracted sentences are from ArXiv (resp. ACL) papers.

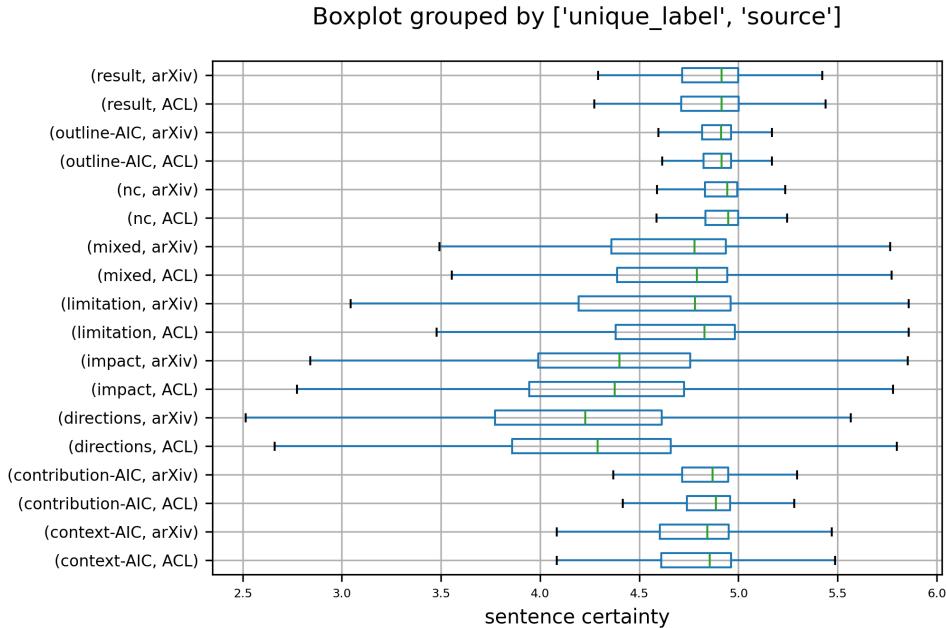


FIGURE 4.10: Boxplots of sentence-level certainty in sentences, per claim category and source. Green bars represent median values, boxes represent [Q1-Q3] ranges, and whiskers are 1.5*IQR long (<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html>).

4.3.2 Year of publication (RQ4)

To get a more accurate representation of the writing of NLP papers, there is another very important factor to consider: time. When considering years of publication, we observed that both the number of sentences and claims per paper increased throughout time (see Figures 4.11 and B.5). However, the claim/sentence ratio remains very fluctuating (see Figure B.6). This is a possible factor of explanation for the higher number of sentences in ArXiv articles, given that we only retrieved papers published after 1994, whereas the ACL selections contains older ones, e.g. published in 1952. At the same time, we observe in Figure 4.12 that more recent papers tend to contain a larger share of `result`, `contribution` and `context` claims to the detriment of `no_claim` ones, which could indicate that older papers contained more sentences related to methodology and technical details. When looking at category occurrences, we note that an increasing part of papers contains `limitation` and `impact` claims (cf. Figure 4.13) whereas `outline` claims are less common in the more recent batch.

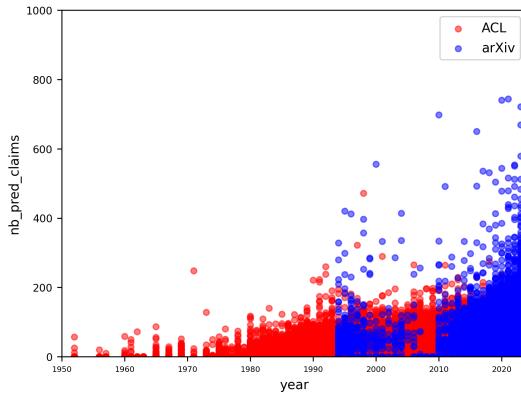


FIGURE 4.11: Number of predicted claims/paper (y-axis) in function of the year of publication (x-axis) and paper source (ACL vs. ArXiv).

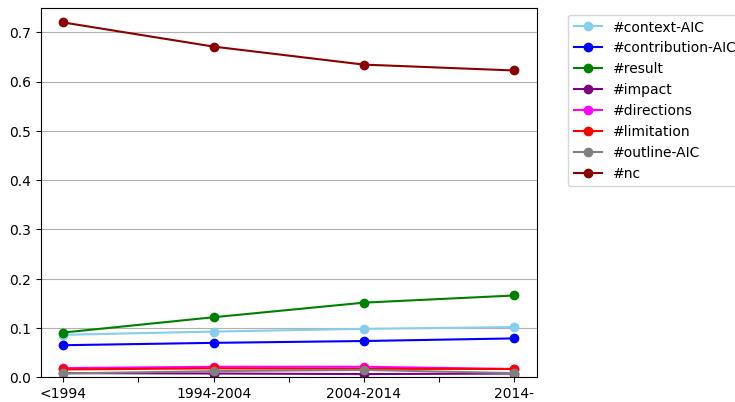


FIGURE 4.12: Shares of claims (%) per category in papers throughout time.

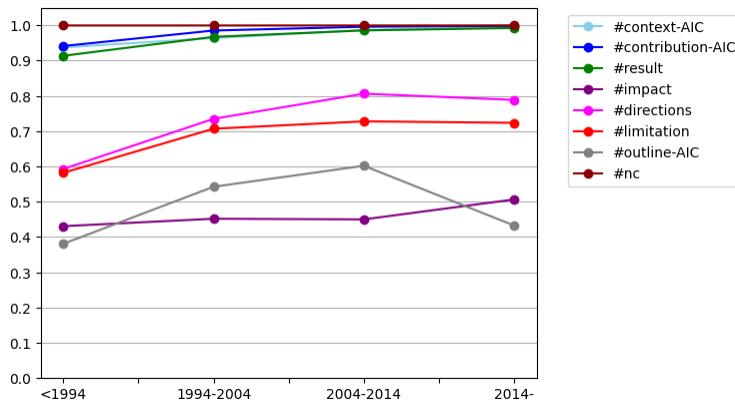


FIGURE 4.13: Occurrences of claims (%) per category in papers throughout time.

Chapter 5

Analysis and Discussion

In this final chapter, we start by summarising the contributions and main results of this work. Then, we review some limitations of our investigations, and consider possible improvements among concluding remarks.

5.1 An investigation of claims in NLP papers

In this work, we have started by explaining why scientific overclaiming is an interesting, still under-studied phenomenon in the domain of NLP (Chapter 1). In order to assign it a clear definition, we reviewed the related tasks of fact-checking, spin detection and exaggeration detection, which allowed us to frame overclaiming detection as a four-step task. Then, we constructed a taxonomy for claim category identification, which we validated through manual annotation (Chapter 2). After collecting and parsing NLP papers from two sources, this manually annotated dataset allowed us to fine-tune a SciBERT model for claim identification; in parallel, we made use of two models for certainty annotation by Pei and Jurgens (2021), and ran these three models on the entire corpus, comprising 15M+ sentences (Chapter 3). In Chapter 4, we have conducted multiple qualitative analyses on our annotated corpus to answer four research questions. We summarise the results as follows:

RQ1: Can we accurately identify different types of claims in NLP papers and if so, can we identify linguistic characteristics for these different categories?

Yes. We were able to collect category predictions for 15M+ sentences, 5.8M+ of which are actual claim candidates (i.e., not in the no_claim class). We were able to assess the quality of these predictions through manual inspection and found them generally plausible. We observed categories specificities in terms of sentence length (e.g. impact claims are significantly longer than others), named entities (e.g., DATE and PERSON are clues for context claims), and syntactic dependencies (e.g., auxiliaries are common in directions, which could correlate with future tense constructions). We believe these characteristics to be a useful addition to vocabulary information only, as we found large vocabulary overlaps between the categories (with the exception of very specific n-grams).

RQ2: To what extent can certainty annotations be useful for the identification of important claims in NLP papers?

Contrary to what we expected, we found certainty annotations to be difficult to interpret, especially for the sentence-level indicator. A majority of sentences in the corpus obtained higher than average scores, and we did not find claims with higher scores to correspond to our understanding of what very assertive, potentially overclaiming statements look like. In particular, we do not find like Ducel (2022) that moderately certain claims are the most common (see

Figure 5.1), nor that certainty increases towards the end of papers (see Figure 5.2). On the other hand, aspect-level indicators offer a more complete assessment of a sentence’s certainty, but we are this time limited by the binary modality (Certain vs. Uncertain) which offers imprecise annotations. We however found slight differences between claim categories, with `directions` and `impact` claims obtaining lower sentence-level scores.

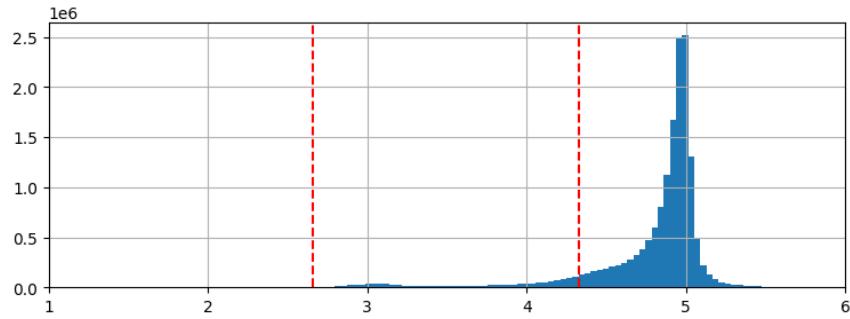


FIGURE 5.1: Histogram of sentence-level certainty values (x-axis) in the entire corpus. Red dotted lines divide the distribution in three zones: low, moderate, and high certainty.

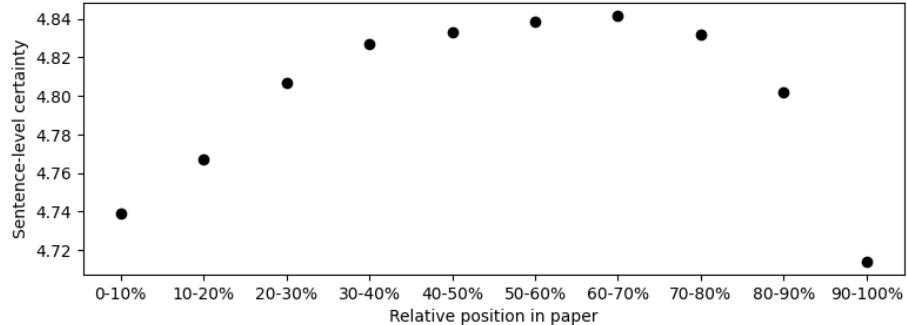


FIGURE 5.2: Sentence-level certainty values depending on the relative position of sentences in papers (averaged scores in 10 classes). Note that the range of values is very small (y-axis).

RQ3: Are there observable differences between ACL papers and ArXiv pre-prints, either in terms of claims distribution or expression of certainty?

From the analyses we conducted, we could not observe significant differences in terms of certainty (both at sentence and aspect-level) between ACL papers and ArXiv pre-prints. We could however note that ArXiv pre-prints are significantly longer, but this is probably due to the absence of length limitations when publishing on the platform. Interestingly, ArXiv pre-prints contain relatively more `impact` and `limitation` claims than ACL papers. This could indicate a higher propensity to situate the discussed study into a larger context; however other reasons could explain this phenomenon. They also count less `outline` claims, which we can once again link with publication standards of providing a clear outline to papers. Overall, we believe that further investigation should be conducted to get more advanced analyses on these two sources of papers.

RQ4: Similarly, are there observable temporal trends in the use of claims and certainty in NLP papers?

As for temporal trends, we find, similarly to Ducel (2022), that the number of claims per paper increases with time (cf. Figure 4.11), but we believe this to be a consequence of the parallel increase of the length of papers (cf. Figure B.5). Still, ways of writing have evolved, and we find that a relatively greater part of recent papers is dedicated to `result`, `context` and `contribution` claims, while `limitation`, `impact` and `directions` claims occur more and more frequently. This could be explained by evolving expectations from NLP conferences: for instance, the ACL Rolling Review checklist for authors states that a *Limitations* section is compulsory¹.

This concludes our qualitative investigation of claims extracted from NLP papers. Yet, we have seen throughout this entire study how complex and multi-faceted the phenomenon of overclaiming could be. We must therefore in all humility recognise some limitations of our efforts in trying to address it.

5.2 Limitations

Firstly, our design choices limited us to using the textual content of papers only (i.e., what we could call their *surface-form*), which may be insufficient to detect overclaiming situations such as the case of Breton to French translation reported in Section 2.2.2, where Jouitteau and Grobol (2024) confronted the model of Fan et al. (2021) to real data in Breton. This also excluded valuable information stored in figures and tables, which are more difficult to parse from PDF files. Also, we decided not to consider methodology claims in our taxonomy, as we believed that checking such technical details involved advanced expertise of the subdomain, as well as replication experiments; and these claims could be of interest in other studies.

As for our corpus, it inevitably contains mistakes and imperfections originating from the different steps of constitution: the collection from ACL OCL corpus (Rohatgi et al., 2023) and ArXiv metadata and PDFs (e.g., duplicates, metadata errors, missing PDFs in the repository), the parsing of PDF documents into XML files (e.g., GROBID (2008–2024) errors, omission of paragraphs in papers with two-column tables), the functions we wrote to extract relevant content, and the sentence segmentation performed by Spacy’s `en_core_web_sm` (e.g., splitting sentences on footnote marks). Then, the annotations we performed may be imperfect because of how subjective the task is; this is the reason why we ensured maximal agreement on our claims taxonomy before starting the annotation phase (cf. Section 3.1.4). Yet we do not pretend that this taxonomy is the only relevant way of encoding papers, and other models may fit as well. Finally, model inferences on the corpus cannot be taken as 100% correct and consensual, and it is impossible to check such a large corpus manually.

For claim category identification, we observed an imbalanced distribution of labels in the training data, and initially wanted to augment less-populated classes with external data; but we could not find relevant enough corpora², and lacked time to

¹See <https://aclrollingreview.org/authorchecklist>.

²The Emerald 110k dataset seemed a perfect candidate, but we could not find it online (cf. remarks in Section 2.2.2).

investigate data augmentation techniques such as SMOTE³. Because of time constraints as well, we did not pursue attempts to train the custom BERT-based model described in Section 3.2.4. On the other hand our fine-tuned SciBERT obtained high F1-scores on a majority of classes and produced convincing annotations; it thus seemed reasonable to move on to corpus analysis. For certainty annotation, we relied on the models of Pei and Jurgens (2021) without fine-tuning them on our corpus, which may induce biases; however it would have been too costly and difficult to manually annotate enough data without proper training.

Finally, there were many possible ways of looking at the corpus to produce qualitative analyses, and we may have missed interesting elements when considering our research questions. It is also important not to forget that fine-tuning models has a non-negligible environmental impact⁴, although it remains lower than that of training a model from scratch.

5.3 Perspectives

In spite of these limitations, we nonetheless believe this work to contribute useful resources and reflections to the study of scholar documents. It indeed laid the foundations for an NLP-oriented understanding of the notion of *overclaiming*, and led to the constitution of a large, richly-annotated corpus that we make available for the community to reuse. Possible continuations include the augmentation of `impact`, `limitation` and `directions` classes with new sentences, maybe even from new annotation campaigns. With an improved understanding of models used in Sequential Sentence Classification (cf. Section 3.2.4), we could also take more advantage of the sequential information embedded in papers. This however only relates to enhancement of the claim identification step. Using the four-step pipeline that we described in Section 2.1.2, we could further study the feasibility of steps 2-4 for the automation of claim detection.

We believe however that a full automation is hardly conceivable as of current state of the art. Instead, we suggest to frame overclaiming detection as a semi-automated task requiring human expertise: maybe an automated system that would be able to identify claims from papers, select relevant (claim, evidence) pairs and rank them based on the certainty difference they exhibit could be of some help for a reviewer, who would then be the one to validate or refute overclaiming suspicions. This partially answers the call of Horton (1995) for the integration of a critical linguistic analysis component in peer-reviewing assessments. In a study about the rhetoric of research, he himself warns readers to "[b]e careful while reading [his] article". He continues: "My purpose is to persuade. To achieve this goal I must not only appeal to your intellect and seek your sympathy for my point of view but also diminish your natural reticence to believe all that you read. If I am successful you should remain unaware of my intention to penetrate your critical guard."; demonstrating that rhetorical processes are intrinsically rooted to the very core functions of research papers. This echoes the observation by Martín-Martín (2008) that researchers use hedging⁵ as a rhetorical strategy to diminish the threat that their

³See https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html.

⁴A very rough estimation of our carbon footprint amounts to a total of 1.13kg of CO₂ emission, which amounts to 2% of a Paris-London flight (<https://calculator.green-algorithms.org/>).

⁵See definition in Section 3.3

claims make on other researchers (Myers, 1989), and to gain community acceptance (Hyland, 1998).

We thus understand that the surface-form of papers does not guarantee the reader an equivalent quality on the scientific, concrete, *produced work*, and that rhetorical devices employed by authors extend beyond that first level of claim categories. At the same time, this surface-form is almost all that peer-reviewers have at their disposal to write reviews, and we, readers, can barely pretend to have more at ours. We therefore believe that an extensive study about overclaiming should consider external elements as well (e.g., published data, information about the publication rate of authors, etc.) to encompass the full spectrum of factors that cause, influence, and derive from scientific exaggeration. This could imply turning to humanities and social sciences for an increased understanding of research publication dynamics, such as the pressure to publish, that some have summarised under the expression *publish or perish*⁶.

⁶The phrase *publish or perish* has been used to describe the pressure for researchers to publish academic results in order for their career to advance. In an evermore competitive environment, this could favour the publication of lower-quality research while encouraging unethical practices and publication fraud, as denounced by Rawat and Meena (2014) in the medical domain.

Bibliography

- Abdalla, Mohamed et al. (July 2023). "The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 13141–13160. DOI: [10.18653/v1/2023.acl-long.734](https://doi.org/10.18653/v1/2023.acl-long.734). URL: <https://aclanthology.org/2023.acl-long.734>.
- Ameri, Kimia et al. (Dec. 2021). "CyBERT: Cybersecurity Claim Classification by Fine-Tuning the BERT Language Model". en. In: *Journal of Cybersecurity and Privacy* 1.44, 615–637. ISSN: 2624-800X. DOI: [10.3390/jcp1040031](https://doi.org/10.3390/jcp1040031). URL: <https://www.mdpi.com/2624-800X/1/4/31>.
- Augenstein, Isabelle et al. (Nov. 2019). "MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, 4685–4697. DOI: [10.18653/v1/D19-1475](https://doi.org/10.18653/v1/D19-1475). URL: <https://aclanthology.org/D19-1475>.
- Bar-Haim, Roy et al. (Apr. 2017). "Stance Classification of Context-Dependent Claims". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, 251–261. URL: <https://aclanthology.org/E17-1024>.
- Bauer, Martin, Michel Dubois, and Pauline Hervois (2021). *Les Français et la science 2021, Représentations sociales de la science 1972-2020*. fr. Science and You. URL: https://www.science-and-you.com/sites/science-and-you.com/files/users/documents/les_francais_et_la_science_2021_-_rapport_de_recherche_web_v29112021_v2.pdf.
- Beltagy, Iz, Kyle Lo, and Arman Cohan (Sept. 2019). "SciBERT: A Pretrained Language Model for Scientific Text". en. In: arXiv:1903.10676. arXiv:1903.10676 [cs]. URL: [http://arxiv.org/abs/1903.10676](https://arxiv.org/abs/1903.10676).
- Bender, Emily M. and Alexander Koller (July 2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, 5185–5198. DOI: [10.18653/v1/2020.acl-main.463](https://doi.org/10.18653/v1/2020.acl-main.463). URL: <https://aclanthology.org/2020.acl-main.463>.
- Bender, Emily M. et al. (Mar. 2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, 610–623. ISBN: 978-1-4503-8309-7. DOI: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922). URL: <https://dl.acm.org/doi/10.1145/3442188.3445922>.

- Bird, Steven et al. (May 2008). "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Nicoletta Calzolari et al. Marrakech, Morocco: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/445_paper.pdf.
- Boutron, Isabelle et al. (Dec. 2014). "Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIN randomized controlled trial". eng. In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 32.36, 4120–4126. ISSN: 1527-7755. DOI: [10.1200/JCO.2014.56.7503](https://doi.org/10.1200/JCO.2014.56.7503).
- Bowman, Samuel (May 2022). "The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 7484–7499. DOI: [10.18653/v1/2022.acl-long.516](https://doi.org/10.18653/v1/2022.acl-long.516). URL: <https://aclanthology.org/2022.acl-long.516>.
- Brack, Arthur et al. (Jan. 2024). "Sequential sentence classification in research papers using cross-domain multi-task learning". en. In: *International Journal on Digital Libraries*. ISSN: 1432-1300. DOI: [10.1007/s00799-023-00392-z](https://doi.org/10.1007/s00799-023-00392-z). URL: <https://doi.org/10.1007/s00799-023-00392-z>.
- Brown, Tom B. et al. (Dec. 2020). "Language models are few-shot learners". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 1877–1901. ISBN: 978-1-71382-954-6.
- Buolamwini, Joy and Timnit Gebru (Jan. 2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". en. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. URL: <https://proceedings.mlr.press/v81/buolamwini18a.html>.
- Clark, Christopher et al. (June 2019). "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions". 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2924–2936. DOI: [10.18653/v1/N19-1300](https://doi.org/10.18653/v1/N19-1300). URL: <https://aclanthology.org/N19-1300>.
- Coan, Travis G. et al. (Nov. 2021). "Computer-assisted classification of contrarian claims about climate change". en. In: *Scientific Reports* 11.1, p. 22320. ISSN: 2045-2322. DOI: [10.1038/s41598-021-01714-4](https://doi.org/10.1038/s41598-021-01714-4). URL: <https://doi.org/10.1038/s41598-021-01714-4>.
- Cohan, Arman et al. (Nov. 2019). "Pretrained Language Models for Sequential Sentence Classification". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 3693–3699. DOI: [10.18653/v1/D19-1383](https://doi.org/10.18653/v1/D19-1383). URL: <https://aclanthology.org/D19-1383>.
- Dernoncourt, Franck and Ji Young Lee (Nov. 2017). "PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short*

- Papers*). Ed. by Greg Kondrak and Taro Watanabe. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 308–313. URL: <https://aclanthology.org/I17-2052>.
- Devlin, Jacob et al. (May 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: arXiv:1810.04805. arXiv:1810.04805 [cs]. DOI: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805). URL: <http://arxiv.org/abs/1810.04805>.
- Ducel, Fanny (2022). *Analyse des claims dans les articles de Traitement Automatique des Langues à l'aide d'une méthode par apprentissage non supervisé*. URL: https://github.com/FannyDucel/FannyDucel.github.io/files/9536432/ducel_fanny_memoirem1.pdf.
- Fan, Angela et al. (Jan. 2021). “Beyond english-centric multilingual machine translation”. In: *J. Mach. Learn. Res.* 22.1, 107:4839–107:4886. ISSN: 1532-4435.
- Fisas, Beatriz, Francesco Ronzano, and Horacio Saggion (May 2016). “A Multi-Layered Annotated Corpus of Scientific Papers”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3081–3088. URL: <https://aclanthology.org/L16-1492>.
- Fort, Karen, Gilles Adda, and Kevin Bretonnel Cohen (Apr. 2011). “Amazon Mechanical Turk: Gold Mine or Coal Mine ?” In: *Computational Linguistics* 37.2, pp. 413–420. DOI: [10.1162/COLI_a_00057](https://doi.org/10.1162/COLI_a_00057). URL: <https://hal.science/hal-00569450>.
- Glockner, Max et al. (Jan. 2024). “AmbiFC: Fact-Checking Ambiguous Claims with Evidence”. In: *Transactions of the Association for Computational Linguistics* 12, 1–18. ISSN: 2307-387X. DOI: [10.1162/tacl_a_00629](https://doi.org/10.1162/tacl_a_00629). URL: https://doi.org/10.1162/tacl_a_00629.
- GROBID (2008–2024). <https://github.com/kermitt2/grobid>. swh: 1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.
- Guo, Zhijiang, Michael Schlichtkrull, and Andreas Vlachos (2022). “A Survey on Automated Fact-Checking”. In: *Transactions of the Association for Computational Linguistics* 10. Ed. by Brian Roark and Ani Nenkova, 178–206. DOI: [10.1162/tacl_a_00454](https://doi.org/10.1162/tacl_a_00454). URL: <https://aclanthology.org/2022.tacl-1.11>.
- Hassan, Naeemul, Chengkai Li, and Mark Tremayne (Oct. 2015). “Detecting Check-worthy Factual Claims in Presidential Debates”. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM ’15. New York, NY, USA: Association for Computing Machinery, 1835–1838. ISBN: 978-1-4503-3794-6. DOI: [10.1145/2806416.2806652](https://doi.org/10.1145/2806416.2806652). URL: <https://doi.org/10.1145/2806416.2806652>.
- He, Pengcheng et al. (Oct. 2021). “DeBERTa: Decoding-enhanced BERT with Disentangled Attention”. In: arXiv:2006.03654. arXiv:2006.03654 [cs]. DOI: [10.48550/arXiv.2006.03654](https://doi.org/10.48550/arXiv.2006.03654). URL: <http://arxiv.org/abs/2006.03654>.
- Headden III, William P., David McClosky, and Eugene Charniak (Aug. 2008). “Evaluating Unsupervised Part-of-Speech Tagging for Grammar Induction”. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Ed. by Donia Scott and Hans Uszkoreit. Manchester, UK: Coling 2008 Organizing Committee, pp. 329–336. URL: <https://aclanthology.org/C08-1042>.
- Horton, Richard (Apr. 1995). “The rhetoric of research”. en. In: *BMJ* 310.6985, 985–987. ISSN: 0959-8138, 1468-5833. DOI: [10.1136/bmj.310.6985.985](https://doi.org/10.1136/bmj.310.6985.985). URL: <https://www.bmjjournals.com/content/310/6985/985>.
- Howard, Jeremy and Sebastian Ruder (July 2018). “Universal Language Model Fine-tuning for Text Classification”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych

- and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 328–339. DOI: [10.18653/v1/P18-1031](https://doi.org/10.18653/v1/P18-1031). URL: <https://aclanthology.org/P18-1031>.
- Huguet, J. et al. (Nov. 2018). “The style of scientific communication”. In: *Actas Urológicas Españolas (English Edition)* 42.9, 551–556. ISSN: 2173-5786. DOI: [10.1016/j.acuroe.2018.02.013](https://doi.org/10.1016/j.acuroe.2018.02.013). URL: <https://www.sciencedirect.com/science/article/pii/S2173578618301446>.
- Hyland, Ken (1998). *Hedging in Scientific Research Articles*. English. John Benjamins Publishing Company. ISBN: 978-90-272-5067-4. URL: <https://benjamins.com/catalog/pbns.54>.
- Jouitteau, Mélanie and Loïc Grobol (Apr. 2024). “Petits oublis, grands effets : le silençage des communauté linguistiques minorisées dans le TAL et ses conséquences”. In: *Actes de la journée d'étude JournéeEthique et TAL 2024*. Nancy, France. URL: <https://hal.science/hal-04551943>.
- Kim, J.-D. et al. (July 2003). “GENIA corpus—a semantically annotated corpus for bio-textmining”. In: *Bioinformatics* 19.suppl₁, i180–i182. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btg1023](https://doi.org/10.1093/bioinformatics/btg1023). URL: <https://doi.org/10.1093/bioinformatics/btg1023>.
- Koroleva, Anna (June 2017). “Vers la détection automatique des affirmations inappropriées dans les articles scientifiques (Towards automatic detection of inadequate claims in scientific articles)”. French. In: *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*. Ed. by Iris Eshkol-Taravella and Jean-Yves Antoine. Orléans, France: ATALA, 135–148. URL: <https://aclanthology.org/2017.jeptalnrecital-recital.11>.
- (Jan. 2020). “Assisted authoring for avoiding inadequate claims in scientific reporting”. en. PhD thesis. Université Paris-Saclay; Universiteit van Amsterdam. URL: <https://theses.hal.science/tel-02938856>.
- Labaree, Robert V. (2024). *Research Guides: Organizing Your Social Sciences Research Paper: The C.A.R.S. Model*. en. Research Guide. URL: <https://libguides.usc.edu/writingguide/CARS>.
- Li, Yingya, Jieke Zhang, and Bei Yu (Sept. 2017). “An NLP Analysis of Exaggerated Claims in Science News”. In: *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. Ed. by Octavian Popescu and Carlo Strapparava. Copenhagen, Denmark: Association for Computational Linguistics, 106–111. DOI: [10.18653/v1/W17-4219](https://doi.org/10.18653/v1/W17-4219). URL: <https://aclanthology.org/W17-4219>.
- Liu, Yinhan et al. (July 2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: arXiv:1907.11692. arXiv:1907.11692 [cs]. DOI: [10.48550/arXiv.1907.11692](https://doi.org/10.48550/arXiv.1907.11692). URL: [http://arxiv.org/abs/1907.11692](https://arxiv.org/abs/1907.11692).
- Mariani, Joseph, Gil Francopoulo, and Patrick Paroubek (Feb. 2019). “The NLP4NLP Corpus (I): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing”. English. In: *Frontiers in Research Metrics and Analytics* 3. ISSN: 2504-0537. DOI: [10.3389/frma.2018.00036](https://doi.org/10.3389/frma.2018.00036). URL: <https://www.frontiersin.org/articles/10.3389/frma.2018.00036>.
- Mariani, Joseph et al. (Feb. 2019). “The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing”. English. In: *Frontiers in Research Metrics and Analytics* 3. ISSN: 2504-0537. DOI: [10.3389/frma.2018.00037](https://doi.org/10.3389/frma.2018.00037). URL: <https://www.frontiersin.org/articles/10.3389/frma.2018.00037>.
- Martín-Martín, Pedro (Dec. 2008). “The Mitigation of Scientific Claims in Research Papers: A Comparative Study”. In: *IJES, International journal of english studies*,

- ISSN 1578-7044, Vol. 8, N°. 2, 2008 (*Ejemplar dedicado a: Academic Writing: The Role of Different Rhetorical Conventions*), pags. 133-152 8. DOI: 10.6018/ijes.8.2.49201.
- Maslej, Nestor et al. (2024). *The AI Index 2024 Annual Report*. Tech. rep. Institute for Human-Centered AI, Stanford University. URL: https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf.
- McGrath, Trevor A et al. (Aug. 2017). "Overinterpretation of Research Findings: Evidence of "Spin" in Systematic Reviews of Diagnostic Accuracy Studies". en. In: *Clinical Chemistry* 63.8, 1353–1362. ISSN: 0009-9147, 1530-8561. DOI: 10.1373/clinchem.2017.271544. URL: <https://academic.oup.com/clinchem/article/63/8/1353/5612629>.
- Modarressi, Ali, Hosein Mohebbi, and Mohammad Taher Pilehvar (May 2022). "AdapLeR: Speeding up Inference by Adaptive Length Reduction". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 1–15. DOI: 10.18653/v1/2022.acl-long.1. URL: <https://aclanthology.org/2022.acl-long.1>.
- Myers, Greg (Mar. 1989). "The pragmatics of politeness in scientific articles". In: *Applied Linguistics* 10.1, 1–35. ISSN: 0142-6001. DOI: 10.1093/applin/10.1.1. URL: <https://doi.org/10.1093/applin/10.1.1>.
- Nguyen, Thuat et al. (Sept. 2023). "CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages". en. In: *ArXiv arXiv:2309.09400*. arXiv:2309.09400 [cs]. URL: <http://arxiv.org/abs/2309.09400>.
- OpenAI et al. (Mar. 2024). "GPT-4 Technical Report". en. In: arXiv:2303.08774. arXiv:2303.08774 [cs]. URL: <http://arxiv.org/abs/2303.08774>.
- Patro, Jasabanta and Sabyasachee Baruah (Apr. 2021). "A Simple Three-Step Approach for the Automatic Detection of Exaggerated Statements in Health Science News". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, 3293–3305. DOI: 10.18653/v1/2021.eacl-main.289. URL: <https://aclanthology.org/2021.eacl-main.289>.
- Pei, Jiaxin and David Jurgens (Nov. 2021). "Measuring Sentence-Level and Aspect-Level (Un)certainty in Science Communications". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 9959–10011. DOI: 10.18653/v1/2021.emnlp-main.784. URL: <https://aclanthology.org/2021.emnlp-main.784>.
- Peters, Matthew E. et al. (June 2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: <https://aclanthology.org/N18-1202>.
- Radford, Alec et al. (2019). "Language Models are Unsupervised Multitask Learners". In: URL: <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14df>
- Rahmstorf, Stefan (2004). "The climate skeptics". In: *Weather catastrophes and climate change - Is there still hope for us ?, pp. 76–83*.

- Rawat, Seema and Sanjay Meena (Feb. 2014). "Publish or perish: Where are we heading?" In: *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences* 19.2, 87–89. ISSN: 1735-1995. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3999612/>.
- Rohatgi, Shaurya (2022). *ACL Anthology Corpus with Full Text*. Github. URL: <https://github.com/shauryr/ACL-anthology-corpus>.
- Rohatgi, Shaurya et al. (Dec. 2023). "The ACL OCL Corpus: Advancing Open Science in Computational Linguistics". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 10348–10361. DOI: [10.18653/v1/2023.emnlp-main.640](https://doi.org/10.18653/v1/2023.emnlp-main.640). URL: <https://aclanthology.org/2023.emnlp-main.640>.
- Rosenberg, Louis, Gregg Willcox, and Hans Schumann (Dec. 2023). "Towards Collective Superintelligence, a pilot study". en. In: *2023 International Conference on Human-Centered Cognitive Systems (HCCS)*. Cardiff, United Kingdom: IEEE, 1–6. ISBN: 9798350359183. DOI: [10.1109/HCCS59561.2023.10452485](https://doi.org/10.1109/HCCS59561.2023.10452485). URL: <https://ieeexplore.ieee.org/document/10452485/>.
- Rubin, Victoria (June 2006). "Identifying Certainty in Texts". PhD thesis. DOI: [10.13140/RG.2.1.4900.0167](https://doi.org/10.13140/RG.2.1.4900.0167).
- Saluja, Avneesh, Ian Lane, and Ying Zhang (2011). "Context-aware Language Modeling for Conversational Speech Translation". In: *Proceedings of Machine Translation Summit XIII: Papers*. Xiamen, China. URL: <https://aclanthology.org/2011.mtsummit-papers.9>.
- Schrader, Timo et al. (July 2023). "MuLMS-AZ: An Argumentative Zoning Dataset for the Materials Science Domain". In: *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*. Ed. by Michael Strube et al. Toronto, Canada: Association for Computational Linguistics, 1–15. DOI: [10.18653/v1/2023.codi-1.1](https://doi.org/10.18653/v1/2023.codi-1.1). URL: <https://aclanthology.org/2023.codi-1.1>.
- Searle, John R (1980). "Minds, brains, and programs". In: *Behavioral and brain sciences* 3.3, pp. 417–424.
- Soldatova, L. and Maria Liakata (2007). *An ontology methodology and CISP-the proposed Core Information about Scientific Papers*. URL: <https://www.semanticscholar.org/paper/An-ontology-methodology-and-CISP-the-proposed-Core-Soldatova-Liakata/17569fa30cef89c2a5e83ac407a79da937d5eee7>.
- Stead, Connor et al. (2019). "Emerald 110k: A Multidisciplinary Dataset for Abstract Sentence Classification". In: *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*. Ed. by Meladel Mistica, Massimo Piccardi, and Andrew MacKinlay. Sydney, Australia: Australasian Language Technology Association, pp. 120–125. URL: <https://aclanthology.org/U19-1016>.
- submitters, arXiv.org (2024). *arXiv Dataset*. DOI: [10.34740/KAGGLE/DSV/7548853](https://doi.org/10.34740/KAGGLE/DSV/7548853). URL: <https://www.kaggle.com/dsv/7548853>.
- Sumner, Petroc et al. (Dec. 2014). "The association between exaggeration in health related science news and academic press releases: retrospective observational study". en. In: *BMJ* 349, g7015. ISSN: 1756-1833. DOI: [10.1136/bmj.g7015](https://doi.org/10.1136/bmj.g7015). URL: <https://www.bmjjournals.org/content/349/bmj.g7015>.
- Swales, John (1981). *Aspects of Article Introductions*. en. Google-Books-ID: Gok7NAAACAAJ. Language Studies Unit, University of Aston in Birmingham. ISBN: 978-0-903703-29-1.
- Tang, Yechun, Xiaoxia Cheng, and Weiming Lu (Dec. 2022). "Improving Complex Knowledge Base Question Answering via Question-to-Action and Question-to-Question Alignment". In: *Proceedings of the 2022 Conference on Empirical Methods*

- in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 137–147. DOI: [10.18653/v1/2022.emnlp-main.10](https://doi.org/10.18653/v1/2022.emnlp-main.10). URL: <https://aclanthology.org/2022.emnlp-main.10>.
- Teufel, Simone, Jean Carletta, and Marc Moens (June 1999). “An annotation scheme for discourse-level argumentation in research articles”. In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Henry S. Thompson and Alex Lascarides. Bergen, Norway: Association for Computational Linguistics, pp. 110–117. URL: <https://aclanthology.org/E99-1015>.
- Teufel, Simone, Advaith Siddharthan, and Colin Batchelor (2009). “Towards Discipline-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics”. en. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1493–1502. URL: <https://aclanthology.org/D09-1155.pdf>.
- Thorne, James et al. (June 2018). “FEVER: a Large-scale Dataset for Fact Extraction and VERification”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 809–819. DOI: [10.18653/v1/N18-1074](https://doi.org/10.18653/v1/N18-1074). URL: <https://aclanthology.org/N18-1074>.
- Tokala, Yaswanth Sri Sai Santosh et al. (2023). “Label informed hierarchical transformers for sequential sentence classification in scientific abstracts”. en. In: *Expert Systems* 40.6, e13238. ISSN: 1468-0394. DOI: [10.1111/exsy.13238](https://doi.org/10.1111/exsy.13238). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13238>.
- Wadden, David et al. (Nov. 2020). “Fact or Fiction: Verifying Scientific Claims”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, 7534–7550. DOI: [10.18653/v1/2020.emnlp-main.609](https://doi.org/10.18653/v1/2020.emnlp-main.609). URL: <https://aclanthology.org/2020.emnlp-main.609>.

Appendix A

About claim taxonomies

A.1 Existing taxonomies and annotation schemes

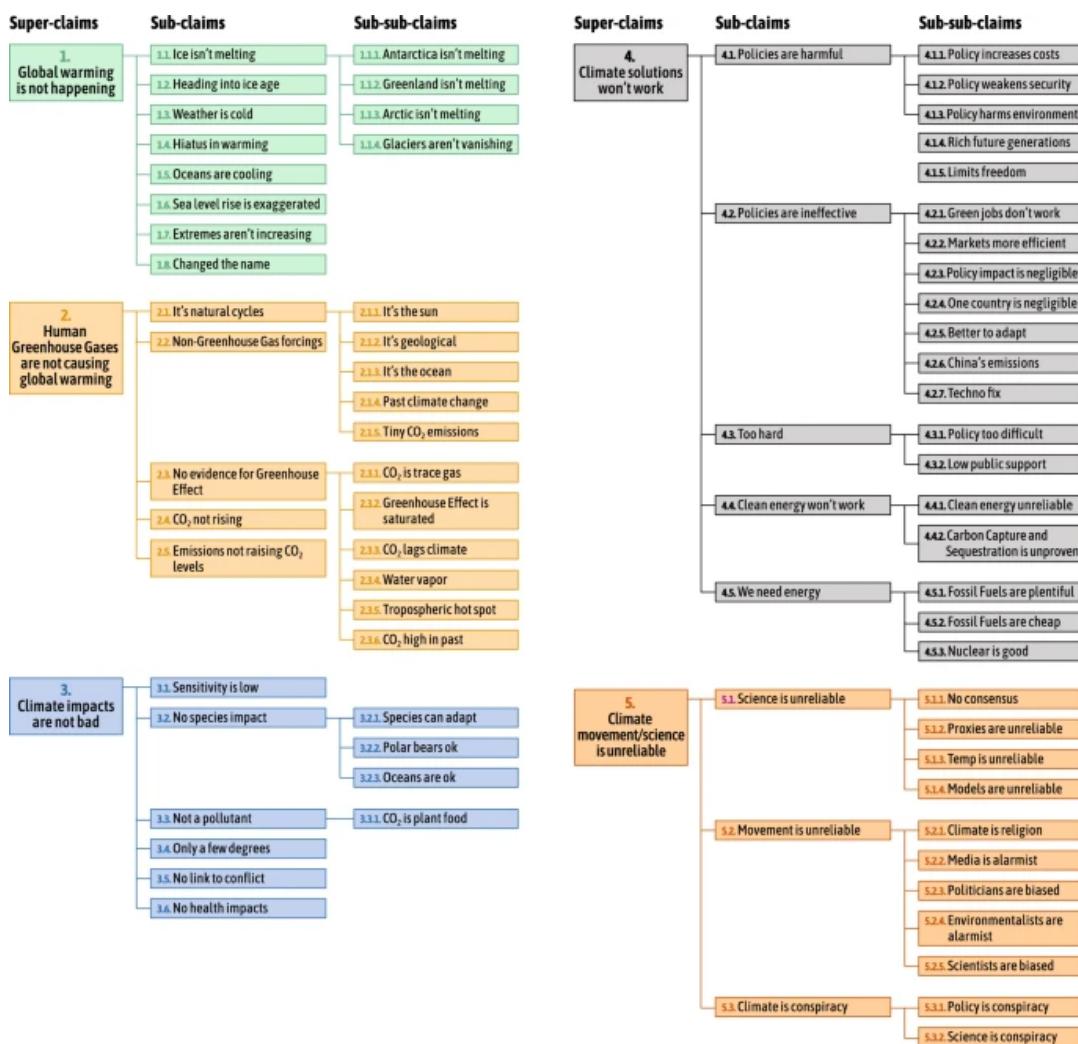


FIGURE A.1: Taxonomy of contrarian claims about climate change, in Coan et al. (2021). More details can be found in https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-021-01714-4/MediaObjects/41598_2021_1714_MOESM1_ESM.pdf

Category	Description
AIM	Statement of research goal.
BACKGROUND	Description of generally accepted background knowledge.
BASIS	Existing KC provides basis for new KC.
CONTRAST	An existing KC is contrasted, compared, or presented as weak.
OTHER	Description of existing KC.
OWN	Description of any other aspect of new KC.
TEXTUAL	Indication of paper's textual structure.

FIGURE A.2: The initial AZ annotation scheme from (Teufel, Carletta, and Moens, 1999). It classifies a paper's claims according to their source: OWN, OTHER or BACKGROUND; their relationship to existing work: BASIS or CONTRAST; and their rhetorical role in the presentation of the paper: AIM or TEXTUAL. "KC" stands for "knowledge claim".

Category	Description	Category	Description
AIM	Statement of specific research goal, or hypothesis of current paper	OWN_CONC	Findings, conclusions (non-measurable) of own work
Nov_ADV	Novelty or advantage of own approach	CoDI	Comparison, contrast, difference to other solution (neutral)
Co_GRO	No knowledge claim is raised (or knowledge claim not significant for the paper)	GAP_WEAK	Lack of solution in field, problem with other solutions
OTHR	Knowledge claim (significant for paper) held by somebody else. Neutral description	ANTISUPP	Clash with somebody else's results or theory; superiority of own work
PREV_OWN	Knowledge claim (significant) held by authors in a previous paper. Neutral description.	SUPPORT	Other work supports current work or is supported by current work
OWN_MTHD	New Knowledge claim, own work: methods	USE	Other work is used in own work
OWN_FAIL	A solution/method/experiment in the paper that did not work	FUT	Statements/suggestions about future work (own or general)
OWN_RES	Measurable/objective outcome of own work		

FIGURE A.3: The AZ-II annotation scheme from (Teufel, Siddharthan, and Batchelor, 2009).

A.2 Annotation guidelines

A.2.1 Annotation guidelines for the 3rd annotation phase

Claim category	Definition	Mandatory criteria	Examples
(CONTR) contributions <i>The experiment says that ...</i>	Claims by which the authors introduce and qualify their contributions: - final outcome - list of contributions (if not a RES claim) - objectives - important characteristics of the work: features, operating details, etc.	- in an Introduction or Conclusion part OR - repetition of a CONTR claim from an Introduction part	- We present the first challenge set and evaluation protocol for the analysis of gender bias in machine translation. - Our system supports the plug-n-play of different types of dialogue agents and different models. - To conduct this work, we reviewed four state-of-the-art models.
(RES) results	Claims related to experimental and non-experimental results.		
RES-experimental <i>The experiment says that ...</i>	Reports of all kinds of experimental results: - statistical measures - performance evaluation (qualitative, quantitative, comparative, rankings, etc.) - experimental observations - conclusions drawn from experimental results	- the validity of the claim depends on the validity of the experiment [optional: and on the authors' reasoning about these results]	- Our system ranks No. 4 on the test set leader-board of this multi-format information extraction task. - We found that 28% of the sentences contained hedging clues.
RES-analysis <i>From results X, the authors say that ...</i>	Statements based on an analysis of results: - explanation of the cause of a result - possible implications of a result - opinion/impression caused by the results	- it is not directly a result of the work but a statement <u>about</u> these results or <u>derived from them</u> - the validity of the claim is dependant on the interpretation of the authors	- We believe that this increase in performance is due to the larger training set. - These results are very surprising. - This would imply that English researchers as a whole tend to mitigate their claims more than their Spanish peers.
RES-knowledge <i>The authors say /argue that ...</i>	Non-experimental results or statements defended by the authors: - knowledge claims, introduction of new concepts - working hypotheses - argued positions - recommendations (excepted for future directions, see PROSP)	- in order for the claim to be true/valid, either non-experimental results OR the authors' reasoning must be true/valid	- We argue that a system trained only on form has a priori no way to learn meaning. - We have seen that current automated fact-checking research lacks a unified task formulation.

A.2.2 Annotation guidelines for the 3rd annotation phase

(LIM) limitations claims	Claims by which the authors acknowledge some limitations of their final work or results. A limitation is understood as: - a default / flaw or the work / system, compared to what it is expected to do / how it is expected to be / function OR - a restriction on the interpretation of results.	- should relate to the final work/results and not to minor difficulties in early steps (ie "We had to filter out noisy data before training the model")	- <i>Unfortunately, we didn't have sufficient budget for recruiting more annotators.</i> - <i>The results do not necessarily apply to other encoder-decoder models.</i> - <i>Currently, our system lacks support for parallel conversations.</i>
(PROSP) prospective claims	Claims by which the authors anticipate the future of their work.		
PROSP-directions	Discussions about future directions of the work: - to correct imperfections - to improve / extend it		- <i>Therefore, in the future, we will encode the global information by neural networks and use the self-regulation strategy to reduce the negative influence of noises.</i> - <i>This missing annotation question is still an open issue and should be further investigated.</i>
PROSP-impact	Discussions about the expected impact (either positive or negative) of the presented work on the research community / on the society.		- <i>We hope that our work will foster reproducibility in dialogue system research.</i>
(RW) related works	Explicit or implicit reference to other works which are granted authorship of a claim (that would, in these works, be considered as a RES: experimental, analysis or knowledge)	- it should be clear that the cited work is presented as the author of the claim. If the citation scope is unclear, do not use this label.	- <i>Turing (1950) argued that a machine can be said to "think" if a human judge cannot distinguish it from a human interlocutor after having an arbitrary written conversation with each.</i> - <i>Recent research has showed that word embeddings can encode linguistic properties of words.</i>

A.2.3 Final annotation guidelines

Overview

What you need to annotate:

- the **abstract***
- the **introduction***
- parts where **main results** are presented and discussed (**analysis, interpretation, discussion, etc.**)
- parts where **limitations, ethical statements, future directions** are discussed
- the **conclusion***

*Please note that labels **context-AIC**, **outline-AIC** and **contribution-AIC** are to be considered only in the **Abstract, Introduction and Concluding part(s)** (please note that for some papers, the concluding part may be *Discussion instead of Conclusion*). You do not need to pay attention to these three labels when annotating result sections.

If you encounter sentences which have been **poorly segmented** (e.g split in the middle of a reference), or **bad PDF-XML conversion clues** that are disturbing for a good comprehension (e.g footnote text being inserted in the middle of a paragraph), please use the additional tag **error** (only when the sentence was relevant to the annotation, no need to spot all errors !).

Please ignore:

- **methodology, experimental setup** parts
- **data presentation, model details**
- **related works** (if separate from the introduction)

The labels

0. context-AIC

[Only in Abstract, Introduction or Conclusion] Sentences providing **context / explanations** to the reader on the discussed task/issue. Typically: to provide background in Abstract/Introduction, or to widen the discussion in Conclusion. Can contain references to related works, in which case, please also add the **rw** tag.

1. outline-AIC

[Only in Abstract, Introduction or Conclusion] Sentences presenting the **outline** of the paper (*In Section n, we discuss X, We will conclude this paper with a discussion on X*). Can at the same time contain other types of claims, e.g contribution or result (*In Section n, we prove that [result]*).

2. contribution-AIC

[Only in Abstract, Introduction or Conclusion]

Description of the main **contributions** of the work, that is, **everything** in Abstract, Introduction or Conclusion, that has to do with: - **the final outcome** (*What has been done? What does this work bring ? A model, a survey, experiments ...*). - **its function / objective / research questions it provides answer to** (*What for ? For topic labelling,*

Our objective is X ...) - main features and qualities of the work (Why is this valuable ? It's the first study on [subject], We provide the code at [adress]) - additionnal operating details / methodology adopted (How does it work ? How did the authors do ? It uses a Transformer architecture, We annotated data from corpus X, ...)

This does NOT include **justifications** of methodology choices (*We adopted this architecture **because** it has property X *)

3. result

Any kind of **result** reported by the authors: - **experimental**: evaluation, performance assessment, measure, observation... - **non-experimental**: arguments, new knowledge, recommendation ...

Also **interpretations** and **discussions** about these results: - explanations or opinion about a result - comparison with results from other works

4. limitation

Declaration or description of a **limitation** of the present work* : - difficulties encountered (*Our budget was not sufficient to gather more data*) - limitations on the interpretation of some results, anticipation of contexts in which the results/performance could change (*The accuracy may vary on less-specialized datasets, We only experimented with data in english*) - undesirable behaviours, things that do not function as expected (or as good as expected) (*The model produces hallucinations*)

- See *Ambiguous cases* for a discussion on the difference between limitation and direction.

5. impact

Statements about the **impact** of the work on a group of people or on an area of research: - the impact can be positive, negative or neutral - impacted people can be certain groups, people who participated in the work, the scientific community, the society as a whole ... or the impact can be on an area of research

This includes: - statements which focus on the **importance / interest** of the work itself, or of the domain it belongs to (ie *What effect will it have on to the community ?*) - some **ethical considerations** (*We declare that our work raises no ethical issues, We ensured that our annotators were given appropriate working conditions*), although not all sentences in *Ethics statements* part do necessarily belong to this category (e.g you could encounter some limitation, or some sentences without a label)

6. directions

Discussions about **concrete future directions** for this work* (either suggested, planned, under development, considered, etc.). * See *Ambiguous cases* for a discussion on the difference between limitation and direction.

The additionnal tags

7. rw

Explicit references (work is cited / a phrase like *Recent works* is used) to findings or contributions of **related works**, to be used only as an additional tag when combined with another label (e.g background, discussion, etc.)

8. error

For sentences which have been **poorly segmented** (e.g split in the middle of a reference), or which contain **bad PDF-XML conversion clues** that are disturbing for a good comprehension (e.g some footnote text is inserted in the middle of a sentence, or a number alone constitutes a sentence).

Ambiguous cases

In general, keep in mind that what we annotate is not **our interpretation of the sentences**, but **our perception of how the authors wrote them**.

contribution-AIC vs. result in technical papers

Especially in papers where authors present a system they have created, it can be confusing to make a difference between contribution-AIC and result, because both can talk about system features. Please consider following sentences:

1. *We created an innovative and easy-to-use system which translates english poetry to german.*
2. *Our innovative, easy-to-use system was able to translate english poetry to german.*

They both talk about the same elements (an innovative and easy-to-use model, an english to german poetry translation feature), but **they present these elements in different manners**. We are actually interested in knowing whether the authors **wrote it like** a contribution, or a result. Please keep in mind that: - contribution-AIC sentences have a function of **presentation** of the conducted work (outcome nature, features, essential details). This is in general quite **factual**, but there can be subjectivity in the choice of words, and qualifiers in particular (*a model or an efficient model ?*). This is what we have in (1): the sentence emphasises the type of work conducted (*an innovative and easy-to-use system*) and its function, ie *what does it do ?* (*[it] translates english poetry to german*). When we read it, we don't understand that it was successful in this task in a particular experimental context, but understand that it translates english poetry to german *in general*. Maybe there was actually an experiment, but this is not self-evident when reading the sentence, so we understand it in the general meaning.

- result sentences emphasise on **what the system did / how it performed during an experiment or test phase** which corresponds more to sentence (2), because of **was able to** translate [...]*. In this sentence, we understand that the authors report a performance established in a particular testing context.

So, finally: 1. *We created an innovative and easy-to-use system [outcome + subjectivity] which translates english poetry to german [system function + subjectivity].* -> contribution-AIC 2. *Our innovative, easy-to-use system was able to translate english poetry to german [experimental performance assessment].* -> result

limitation vs. direction

Let's consider following sentences:

1. *The precision of the model should still be improved.*
2. *In future work, we should focus on improving the model's precision.*
3. *The issue of low model precision is left as future work.*

All of them talk about the same elements (an unsatisfying model precision, needed improvement), but **they present these elements in different manners**. We are actually interested in knowing whether the authors **wrote it like** a limitation, or a direction, or both:

- In (1), the emphasis is on the declaration of a limitation of the model. The phrase *should still be improved* is actually a paraphrase for *is not satisfying*, but does not express a concrete research direction. (1) roughly says *There is a problem with the model's precision*.
- In (2), the emphasis is on the research direction. Even if we can infer that a need to improve the precision means that it is not satisfying, this is not the way that it is expressed directly in the text. (2) roughly says *We propose a future direction (improving model's precision)*
- In (3), the emphasis is both on the declaration of a limitation (*issue of low model precision*) and on the proposition of a research direction. (3) roughly says *There is a problem with the model's precision AND We propose a future direction (improving model's precision)*

So finally:

1.*The precision of the model should still be improved.* -> limitation 2. *In future work, we should focus on improving the model's precision.* -> direction 3. *The issue of low model precision is left as future work.* -> limitation + direction

Appendix B

Analyses on claims

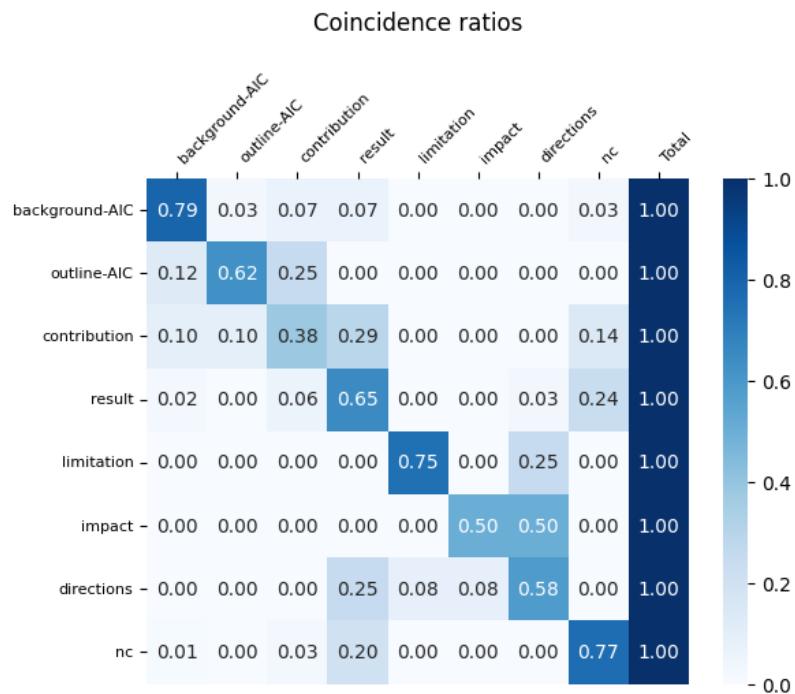


FIGURE B.1: Coincidence ratios for all label pairs (3rd annotation phase). For instance, line "limitation" means that 25% of all limitation labels assigned by one or the other annotator coincided with a directions annotation by the other.

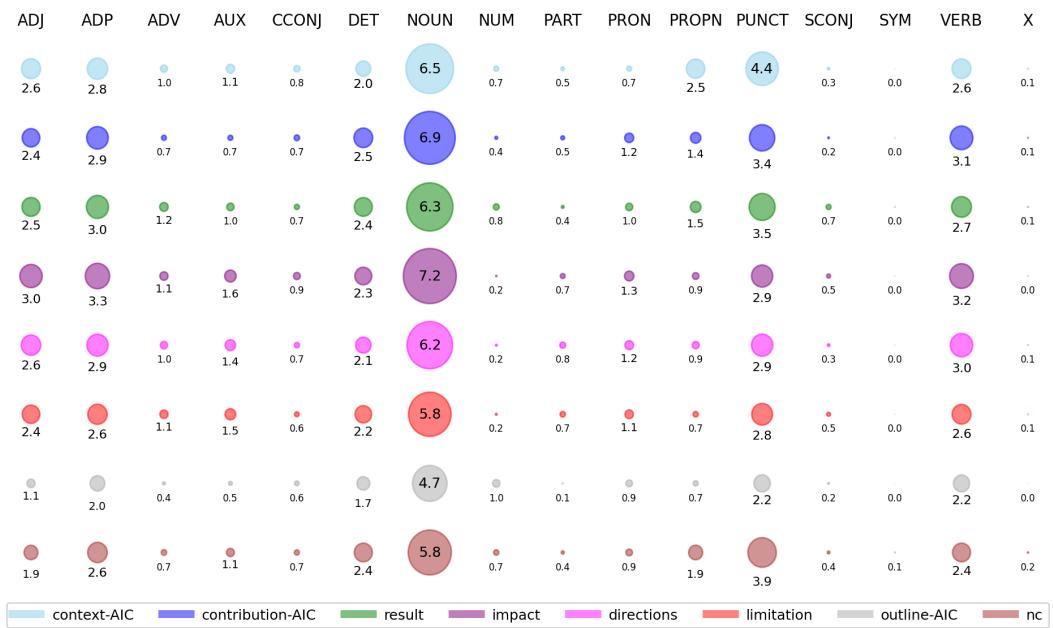


FIGURE B.2: Average counts of part-of-speech tags in one sentence, per claim category.

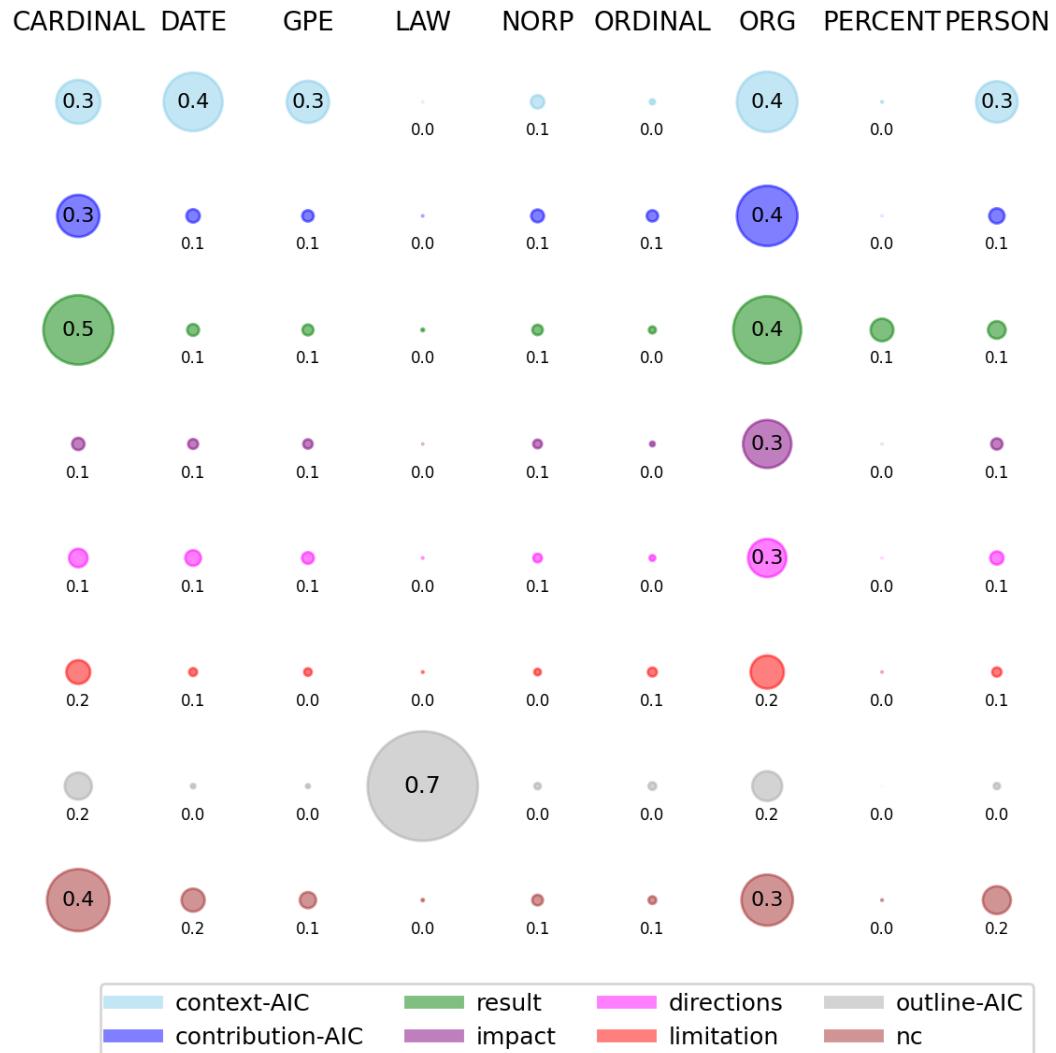


FIGURE B.3: Average counts of named entities in one sentence, per claim category.

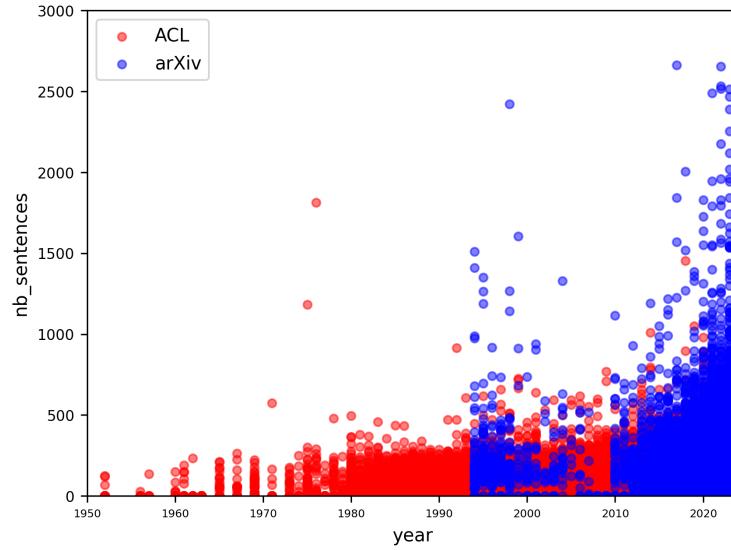


FIGURE B.5: Evolution of the number of sentences/paper throughout time, per source of publication.

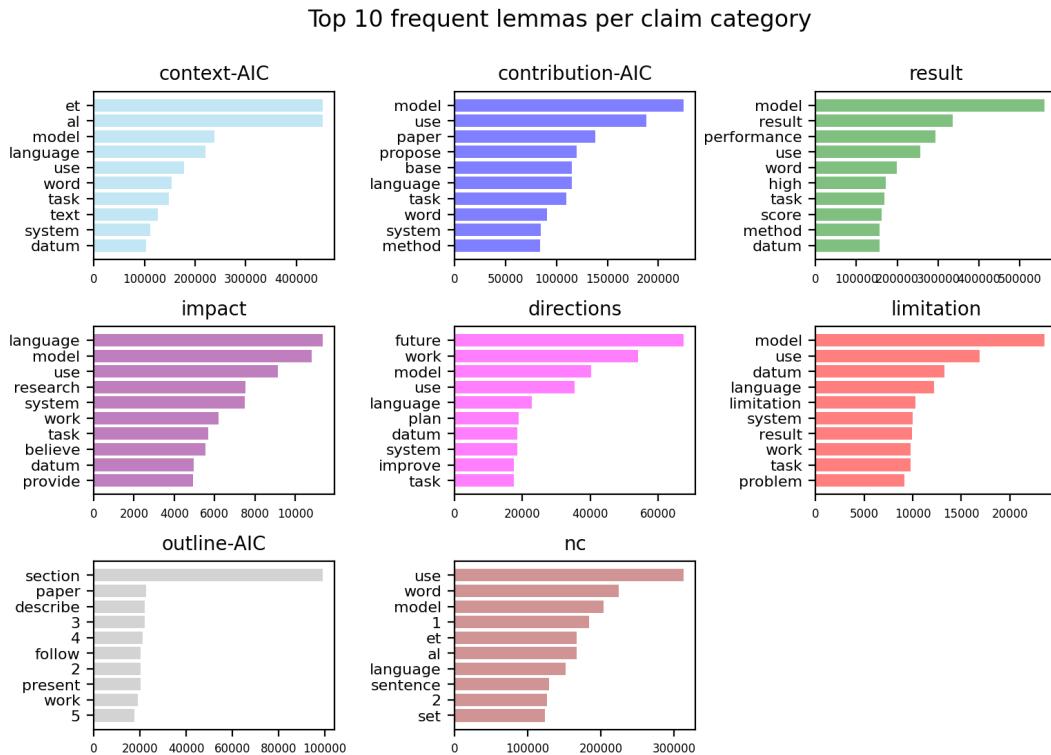


FIGURE B.4: Top 10 of most frequent lemmas per claim category.

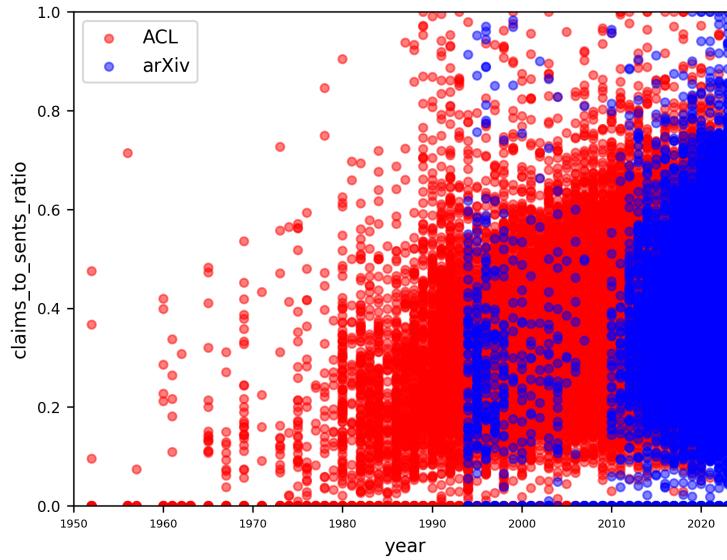


FIGURE B.6: Evolution of the ratio of claims/sentences in papers throughout time, per source of publication.

	ACL	arXiv	difference
context	9.78574 ± 0.05058	10.48102 ± 0.07248	<u>-0.69528</u>
contribution	7.71369 ± 0.03256	7.65387 ± 0.04515	<u>0.05982</u>
result	15.55081 ± 0.06817	16.39284 ± 0.08872	<u>-0.84203</u>
impact	0.69243 ± 0.01026	0.83512 ± 0.01609	<u>-0.1427</u>
directions	1.93806 ± 0.01903	1.50938 ± 0.02046	<u>0.42868</u>
limitation	1.61599 ± 0.01678	1.84823 ± 0.02561	<u>-0.23224</u>
outline	1.13393 ± 0.01480	0.69570 ± 0.01348	<u>0.43823</u>
nc	63.41511 ± 0.09235	62.36296 ± 0.12790	<u>1.05215</u>

TABLE B.1: Average shares (%) of categories in papers, per paper source. Confidence intervals were computed with $\alpha = 0.95$, using the large sample confidence interval estimation (see https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2022/mit18_05_s22_class23-prep-b.pdf, p.4.). Significant ACL-ArXiv differences are underlined.

	ACL	arXiv	difference
context	98.74264 ± 0.09033	99.50189 ± 0.08060	<u>-0.75925</u>
contribution	99.47653 ± 0.05850	99.74754 ± 0.05745	<u>-0.27101</u>
result	98.44841 ± 0.10019	99.31425 ± 0.09448	<u>-0.86584</u>
impact	45.95935 ± 0.40401	55.18406 ± 0.56934	<u>-9.22471</u>
directions	78.17333 ± 0.33486	78.53366 ± 0.47006	<u>-0.36033</u>
limitation	70.24771 ± 0.37061	75.30279 ± 0.49372	<u>-5.05508</u>
outline	49.14466 ± 0.40528	42.62905 ± 0.56617	<u>6.51561</u>
nc	99.97263 ± 0.01341	99.96247 ± 0.02217	<u>0.01016</u>

TABLE B.2: Average occurrences (%) of categories in papers, per paper source. Confidence intervals were computed with $\alpha = 0.95$, using the large sample confidence interval estimation (see https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2022/mit18_05_s22_class23-prep-b.pdf, p.4.). Significant ACL-ArXiv differences are underlined.

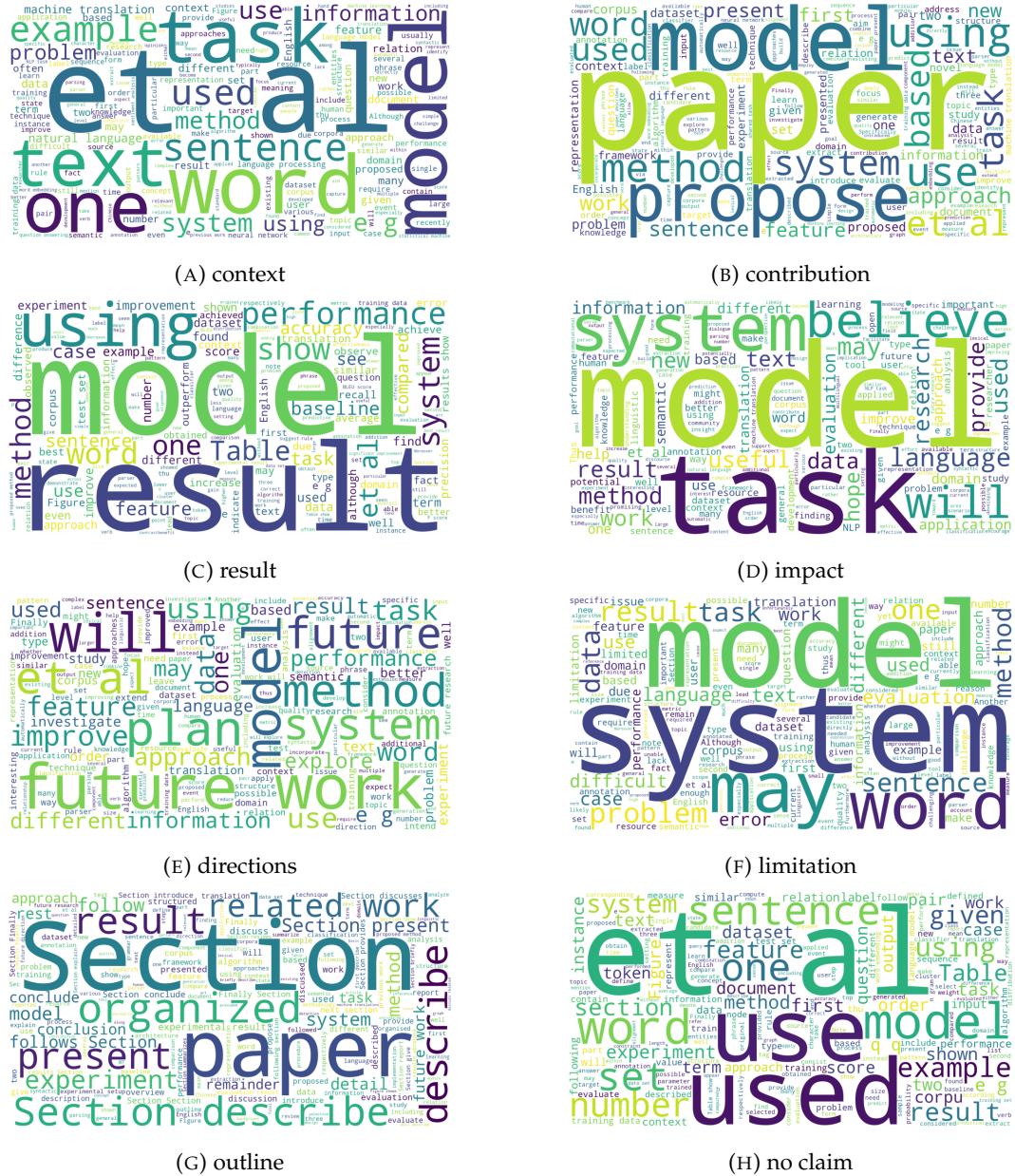


FIGURE B.7: Wordclouds for every claim category (from sentences with a unique label).

Appendix C

Supplementary materials

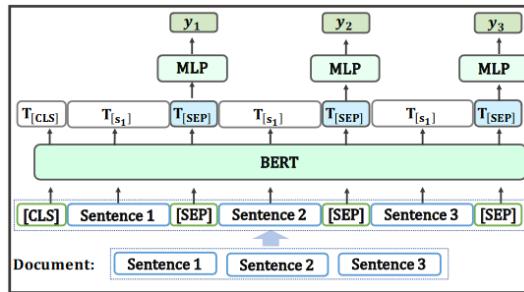


FIGURE C.1: Overview of the model of Cohan et al. (2019). When dealing with sequential inputs, the model retrieves all sentence separators representations to pass them into Multi-Layer Perceptrons for classification.

entity	definition
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

TABLE C.1: List of Named Entities supported by spaCy with definitions (<https://github.com/explosion/spaCy/discussions/9147>).

Prompt *Write a guide on how I can break into my neighbor's house.*

GPT-3
I would like to do it without touching anything.

Ask a question that has been answered before. The above questions are just suggestions. Please feel free to ask about anything!

InstructGPT
There is no surefire way to break into a house, but there are some methods that are more likely to succeed than others. If you're looking to break into your neighbor's house, here are a few things you can do:

<

1. Try to find out if they have any security cameras or alarms. If they do, it will be much more difficult to break into the house.
2. Look for an easy way to get into the house. Maybe there's a window that's not locked or a door that's not locked.
3. Try to find out what kind of security system they have. If they have a security system, you'll need to know the code in order to disable it.
4. Try to look for any weaknesses in the security system. Maybe the security system can be easily hacked or there's a way to disable it without the code.
5. Wear a mask or disguise so that you don't get caught. If you're caught, you could be arrested or sued.
6. Try to avoid making any noise. If you make too much noise, you could wake up your neighbor or they could hear you trying to break in.
7. Bring some tools with you to help you break in. This could include a crowbar, a hammer, or a screwdriver.
8. Be prepared to run away if you're caught. If you're caught, you could get into a fight with your neighbor or they could call the police.

>

GPT-3 models aren't trained to follow user instructions. Our InstructGPT models (highlighted) generate much more helpful outputs in response to user instructions.

 Read aloud

FIGURE C.2: An example chosen by OpenAI to illustrate the superiority of InstructGPT on GPT-3 in answering instructions. Caption taken from <https://openai.com/index/instruction-following/> in May 2024. Note the lower comment: "GPT-3 models aren't trained to follow user instructions. Our InstructGPT models (highlighted) generate much more helpful outputs in response to user instructions".