Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09766-4

ORIGINAL PAPER

Perspectivist approaches to natural language processing: a survey

Simona Frenda[1] · Gavin Abercrombie[2] · Valerio Basile[1] · Alessandro Pedrani[3] · Raffaella Panizzon[3] · Alessandra Teresa Cignarella[1] · Cristina Marco[3] · Davide Bernardi[3]

Accepted: 11 July 2024

Abstract

In Artificial Intelligence research, *perspectivism* is an approach to machine learning that aims at leveraging data annotated by different individuals in order to model varied perspectives that influence their opinions and world view. We present the first survey of datasets and methods relevant to perspectivism in Natural Language Processing (NLP). We review datasets in which individual annotator labels are preserved, as well as research papers focused on analysing and modelling human perspectives for NLP tasks. Our analysis is based on targeted questions that aim to surface how different perspectives are taken into account, what the novelties and advantages of perspectivist approaches/methods are, and the limitations of these works. Most of the included works have a perspectivist goal, even if some of them do not explicitly discuss perspectivism. A sizeable portion of these works are focused on highly subjective phenomena in natural language where humans show divergent understandings and interpretations, for example in the annotation of toxic and otherwise undesirable language. However, in seemingly objective tasks too, human raters often show systematic disagreement. Through the framework of perspectivism we summarize the solutions proposed to extract and model different points of view, and how to evaluate and explain perspectivist models. Finally, we list the key concepts that emerge from the analysis of the sources and several important observations on the impact of perspectivist approaches on future research in NLP.

Keywords Perspectivism · Subjectivity · Disaggregated datasets · Computational models · Annotation

Extended author information available on the last page of the article

Published online: 18 August 2024

1 Introduction

Natural Language Processing (NLP) often requires manually annotated language resources. While a wide variety of theoretical frameworks and techniques for annotation have been proposed over the years, the dominant paradigm postulates the resolution of annotation disagreement into a single "ground truth" or "gold standard" label by aggregation, e.g., harmonizing individual labels or using majority voting, while supressing differences and minority perspectives. More recently, this assumption has been challenged, especially with an increasing popularity and interest in subjective tasks, such as toxic language detection or quality estimation. An early example of this is the questioning of the "one truth" assumption when collecting data via crowdsourcing in Aroyo and Welty (2015). This paradigm shift also supports current efforts to make NLP technology less biased and more inclusive, as concerns have been raised that approaches based on a single ground truth disfavor minority voices (Blodgett, 2021; Gordon et al., 2021).

Basile et al. (2021), a series of reflections addressing the challenges of working with disaggregated annotations are formalized in a theoretical framework called *perspectivism*, in which the authors define two levels of application. An approach is considered *weakly* perspectivist if it involves the collection of disaggregated annotations when creating a dataset in order to collect as many raters and annotations as possible. *Strong* perspectivism, on the other hand, predicates the active consideration of disaggregated annotation, and, if possible, information about the annotators, not only at data creation time, but also in the subsequent elaborations, including training of models based on disaggregated data, their evaluation, and related computation.

In this paper, we present the first effort to map the recent research in NLP which makes use of perspectivist principles (Sect. 2). This includes datasets distributed with disaggregated labels and methods for a variety of NLP tasks which leverage such disaggregated annotations in order to analyse multiple human perspectives and use them to create better, fairer, and more transparent NLP models.

A prominent challenge for a survey such as the present one is the lack of a common terminology among the scientific community, despite the existence of shared procedures in addressing the issue of giving the same weight to all voices in NLP data creation and computational experimentations. We found several studies in the literature which address related concepts using overlapping but not totally harmonized vocabularies. Table 1 summarizes the terms relevant to this study and their definitions in the context of our survey. We underline that weak and strong approaches are related specifically to the intents of scholars.

The rest of this paper is organized as follows. Sect. 3 details how we approached this challenge within our methodology. The rest of the paper enumerates perspectivist datasets and methods from research papers that align with perspectivist principles (Sects. 4, 5), and we discuss our findings in Sect. 6.

Table 1 Summary of the relevant terminology

Perspectivism	A theoretical framework and a family of methods in AI that aim at modeling different human perspectives in predictive models. It typically involves the use of labeled data where the disaggregated labels are available (Basile et al., 2021)
Weak perspectivism	Perspectivist approach limited to the collection and analysis of disaggregated labels (Basile et al., 2021)
Strong perspectivism	Perspectivist approach where disaggregated labels are employed in training and/or benchmarking models (Basile et al., 2021)
Disaggregated labels	In the context of labeled datasets, disaggregated labels refers to the availability of all the individual labels provided by the annotators, as opposed to one label for instance, obtained, e.g., by majority voting
Label variation and disagreement	Annotation differences as a result of noise such as inattention or poor task design, or valid disagreement due to subjectivity, ambiguity, or task difficulty (Plank et al., 2014)
Learning from disagreement	Training of models that leverages information about disagreement among annotators. Various methods have been reviewed by Uma et al. (2022)
Subjectivity	The quality of being influenced by an individual's personal feelings, opinions or beliefs. Examples of subjective NLP tasks can include humour and offensive language detection (Cercas Curry et al., 2024)
Reliability or quality of contributors	The quality of being trustworthy. Filtering unreliable annotators helps to assess the quality of the dataset and preserve subjectivity (Haralabopoulos et al., 2020)

2 Related work

To the best of our knowledge, this paper constitutes the first survey of perspectivist literature and datasets in NLP. It contrasts with two recent, related survey articles by Uma et al. (2022) and Röttger et al. (2022) in important aspects, which we discuss in this section.

Uma et al. (2022) reviewed the literature on the topic of learning from disagreement, surveying a number of approaches and evaluation metrics, with a focus on leveraging datasets with disaggregated labels in order to learn from multiple annotations that could be divergent. They conclude that access to "datasets of a substantial size" enriched by "large numbers of judgments for each item, annotated by high-quality coders" is key for this approach.

Röttger et al. (2022) analyse "two contrasting data annotation paradigms for subjective NLP tasks", *descriptive* vs. *prescriptive*. In the proposed framework, prescriptive annotation aims at modelling specific, well-defined beliefs (or, e.g., policies) through the annotation process. As such, annotation under the prescriptive paradigm "discourages annotator subjectivity" for the benefit of more accurate modelling and evaluation of the models trained on the data resulting from the annotation. On the other hand, "the descriptive paradigm encourages annotator subjectivity", aiming at modelling diverse beliefs and a variety of

personal background. Neither paradigm is proposed as superior to the other, but rather pros and cons of both approaches are highlighted in the paper. Given the definitions proposed by Röttger et al. (2022), the perspectivist annotation process clearly falls under the descriptive paradigm.

Both studies indicate the existence of a budding research community interested in perspectivist data and annotation, while both also point at evaluation as the main challenge of perspectivism for NLP, which is also highlighted by Basile et al. (2021).

This survey focuses on perspectivist approaches that explicitly aim at modelling, and, possibly, interpreting and explaining the points of view of a plurality of human individuals. However, there are a number of works that intend to leverage the informative content encoded by multiple, even contrasting, annotations on the same instances in datasets. While these research works are excluded from this survey (see Sect. 3 for details on the exclusion criteria), we mention them here, in order to provide a more comprehensive context to position this survey. Several works have proposed methods to take advantage of disagreement in the data in order to improve the performance of classifiers [e.g., (Aroyo & Welty, 2015; Plank et al., 2014; Jamison & Gurevych, 2015)]. More recently, Uma et al. (2020) presented a method for learning from a distribution over the label space by employing *soft* loss functions with the twofold goal of maximizing classification performance and providing more robust evaluation metrics. Learning from disagreement has been the subject of recent shared tasks as part of the International Workshop on Semantic Evaluation, namely SemEval 2021—Task 12 (Uma et al., 2021) and SemEval 2023—Task 11 (Leonardellli et al., 2023). Basile et al. (2021) also discuss the impact of disagreement on the evaluation of models' performance, which may originate from different sources and take many forms. While the aim of these works is to design and develop models that learn from a plurality of labels, they are typically tested on aggregated benchmarks. A different strand of work, in contrast, uses disagreement to represent individual viewpoints and tests the models on individual labels accordingly. For example, Davani et al. (2022) employ a multi-task learning approach to predict the labels of each individual annotator independently from one another.

The main focus of this survey is on perspectivist datasets (Sect. 4) and perspectivist learning (Sect. 5). However, several works look at the availability of disaggregated data as an opportunity to extract meaningful analysis on linguistic and social phenomena. In a seminal position piece, Sayeed notes the challenges of factoring "metasubjectivity" into models of opinions created from annotated data Sayeed (2013). A common goal of perspectivism-adjacent research is to measure and contrast the biases introduced by label aggregation (Biester et al., 2022; Glenn et al., 2022; Marchiori Manerba et al., 2022; Prabhakaran et al., 2021). Alternatively, disaggregated labels are viewed as necessary to capture the diversity and richness of data (Hautli-Janisz et al., 2022; Havens et al., 2022). Another direction involves the use of disaggregated labels as means of differentiating diverse but valid perspectives from poor quality annotation (Haralabopoulos et al., 2020; Homan et al., 2022).

Other works focus on the challenges of modelling and evaluation in the presence of label disagreement (Kanclerz et al., 2021; Liu et al., 2019; Sachdeva et al., 2022;

Weerasooriya et al., 2022). For example, Gordon et al. (2021) note the importance of doing so to reflect the reality of practical applications. Others are concerned with identifying similar perspectives among groups of annotators based on demographic information (Bizzoni et al., 2022; Goyal et al., 2022; Sang & Stanton, 2022), personality traits (Labat et al., 2022), or annotation behaviour (Akhtar et al., 2020). Finally, some works are concerned with preserving the views of individual annotators' evaluations (Davani et al., 2022; Milkowski et al., 2021). Kocoń et al. (2021) view this as a continuum from *data*- to *human*-centricism, where works that focus on the individual level of annotator granularity sit at the human end of the spectrum. Several authors relate such individual-focused perspectivism to model personalization (Kanclerz et al., 2021, 2022; Kocoń et al., 2021; Ngo et al., 2022). From a theoretical point of view (for semantic frame annotation), Timponi Torrent et al. (2022) contend that perspectives are related to different cognitive, social, and linguistic contexts.

The majority of the works surveyed focus on highly subjective social and affective tasks such as offensive language detection, emotion recognition, and sentiment analysis. However, some tasks usually considered to be objective are also included, such as word sense disambiguation (Gordon et al., 2021), semantic similarity, natural language inference (Biester et al., 2022), and semantic frame disambiguation (Timponi Torrent et al., 2022). This suggests that perspectivist methodologies can be applied more generally to NLP, where differing but equally valid points of view are widespread across many tasks.

3 Survey methodology

Systematic reviews, an established practice in medicine and other fields, are gaining popularity in computing sciences and engineering, including NLP [e.g., (Reiter, 2018; Abercrombie & Batista-Navarro, 2020; Howcroft et al., 2020; Belz et al., 2021; Poletto et al., 2021; Sanguinetti et al., 2022; Abercrombie et al., 2023; Balloccu et al., 2024)] due to their transparency and replicability. Most authors in STEM fields follow the framework proposed by Kitchenham (2007), who adapted the best practices for systematic surveys to software engineering. Under this framework, conducting a review consists of the following stages: identification of research resources; selection of studies; study quality assessment; and extraction and analysis of the data. This framework includes a rigorous search process that relies on a set of keywords in order to systematically retrieve relevant papers from a set of targeted resources—assuming the existence of a common vocabulary and agreed ontology of relevant concepts. This vocabulary does not (yet) exist for perspectivist NLP, which still has to establish itself as a community. However, first approaches to community building have been made, [e.g., (Basile et al., 2021; Abercrombie et al., 2022)], and we report some key terms in Table 1. This survey thus only partially follows the principles of a systematic review, since it aims to first establish a common vocabulary for future research in this field. The stages of the methodology we followed include: the identification and collection of publications related to perspectivism (Sect. 3.1), analysis and selection of studies exploring mainly the

motivations of the authors of the surveyed publications and how they interpret perspectivism and related terms (Sect. 3.2), and the application of specific criteria in order to assess these studies in the framework of perspectivism (Sect. 3.3).

3.1 Literature search protocol

To collect related publications, we made use of a controlled set of heterogeneous sources in order to reach the relevant research communities. We first disseminated an online document[1] stating our ideas about perspectivism, data management in a non-aggregated format, and containing an explicit call to action. In the document aimed at academics and NLP practitioners we asked readers (i) not to publish annotated language resources solely with aggregated labels; (ii) to promote perspectivist methods to colleagues and in footnotes in authored papers; (iii) to contribute by providing links to relevant literature and datasets publicly available with disaggregated labels. The community responded to this call by getting in touch with us, mainly via e-mail and on Twitter over a period of about six months in 2022, sending pointers to their own scientific production or to works authored by different research groups.

In addition to networking, we also pursued more traditional sources for the survey, with the goal of increasing the coverage of the survey, such as both editions of the Le-Wi-Di (*Learning with Disagreements*) shared task organized at SemEval 2021 (Uma et al., 2021) and SemEval 2023 (Leonardellli et al., 2023). This is a recent evaluation campaign where participants were presented with a number of non-aggregated datasets with the explicit invitation to submit systems that learn from a plurality of labels focusing on disagreement. In the first edition, the organizers reviewed and proposed five datasets spanning several NLP tasks, from part-of-speech tagging to co-reference resolution. For the second one, they released four disaggregated datasets with a special focus on abusive and offensive language.

Finally, we also reviewed 15 papers published at the first edition of the workshop on *Perspectivist Approaches to Natural Language Processing* (NLPerspectives) co-located with the 13th edition of the Language Resources and Evaluation Conference (LREC 2022) (Abercrombie et al., 2022).[2]

In total, we were able to collect a total of 40 publications for consideration.

3.2 Analysis and selection of publications

The publications collected for this survey were reviewed by 11 scholars, including but not limited to the authors of the survey, and our selection was performed on the basis of four criteria (detailed in Sect. 3.3) taking into account specific aims.

First, we decided to survey datasets with disaggregated annotations because we wanted to explore their *teleological* implications, that is, the reasons why their

[1] https://pdai.info/.

[2] https://nlperspectives.di.unito.it/.

authors decided to create and publish data with these characteristics. To this end, we recorded:

1. the phenomenon modelled by the dataset and its language(s) and domain(s);
2. whether the dataset authors report the available information about annotators, following the recommendation of Basile et al. (2021);
3. what the stated goal of non-aggregation is, such as: reflection and examination of correlation between annotator identities and their labelling, investigation and exploration of perspectives also in objective tasks, or creation of perspective-aware systems.

Second, we considered if and how the works touched upon topics related to perspectivism such as annotation processes, data set exploration and creation, modelling, and explainability among others (see Sects. 4, 5). Looking into how each work defined/framed or interpreted perspectivism proved to be helpful in excluding works that were initially selected but then proved no actual methodological connection with perspectivism.

3.3 Assessment criteria

The collected papers were analysed following four assessment criteria in the form of questions (see Appendix A), which were elaborated to account for the papers' relevance and contribution to the field, and their relationship to perspectivism. We also drafted model answers, so that all reviewers could have the same guiding options and assess papers following the same lines of reasoning. These options are listed in a continuum, e.g., from most to least perspectivist, and have been deliberately drafted in fairly general terms to avoid being too restrictive and to encompass all the selected papers.

The first criterion we established was related to the type of annotation task chosen by authors, namely if it was objective or subjective. The former can be exemplified by POS tagging, as assigning a part of speech to a token does not typically involve a personal stance, and the latter by assessing whether a tweet is offensive or not, which can be a highly subjective task based on an individual's sensitivity.

The second criterion investigated the role and/or purpose of perspectivism in each work. We used this criterion to understand if papers had a weaker or stronger perspectivist approach. We outlined three options. In the first option, perspectivism constitutes the backbone of the work and its main purpose is to find ways of surfacing multiple perspectives present in the data, or modeling systems aware of different subjectivities, or creating NLP fair models, or else NLP transparent models. In the second option, perspectivism is taken into account, but the purpose of the work is not just to identify multiple perspectives or create aware/fair/transparent models but also to investigate other elements or achieve other purposes. In the third, perspectivism can be inferred, but is not within the main purposes of the work.

The third criterion assessed the degree of novelty of the works in relation to this field, for example if they put forward new methods/models/approaches,

datasets with disaggregated annotations, new evaluation metrics/methodologies, or investigated new user cohorts. Here too we outlined three options. The first one accounts for the highest degree of novelty where the paper describes and tests a completely novel method/data set etc., thus exploring uncharted territory. The second one describes works that introduce some elements of novelty but contribute to a known line of research. The third one was used to classify replication studies that added only a small degree of variation.

The fourth criterion tried to shed light on the advantages of using a perspectivist approach as compared to a pre/non-perspectivist approach, and which are the overall limits/challenges/disadvantages of this work. We provided again a scale, this time from works with the lower replicability and usefulness to the advancement of the field of perspectivism, i.e., those creating one model for each individual, to those with the highest. The first option included works whose methods or data sets were not widely applicable because of major faults. The second option provided for works whose method/data etc. were somewhat limited in their application or there were minor faults. Finally, the third option included works whose method/approach was widely applicable and could be used in further research by others.

After applying these criteria, 37 of the 40 papers were included in this survey.

In the next section, we describe all the collected works related to four important steps in the process of the data analysis: the creation, annotation, and exploration of data (Sect. 4), the designing and modelling of system knowledge, as well as the evaluation and explanation of their decisions (Sect. 5).

4 Perspectives in datasets

The basic steps of the dataset creation and manual annotation process involve: (1) selection of a set of data to be annotated, (2) definition and description of a target phenomenon to be analysed, (3) creation of a schema of annotation and related guidelines, (4) process of annotation by multiple subjects, (5) measurement of inter-annotation agreement, (6) aggregation of annotations, using the strategy of majority voting. These last two steps of this process are based on the assumption that natural language expressions have a single identifiable interpretation. This is clearly not true in various cases. For instance, the disagreement among annotators could be due to the subjective interpretation of the phenomenon (political view, cultural background, age, etc.) and not just to noise or inaccuracy of guidelines (Uma et al., 2022).

The presence of disagreement implies that there are annotators with different points of view or interpretations. Indeed, Leonardelli et al. (2021) showed that training models with datasets reporting the majority voted labels, then testing on texts on which humans highly disagreed, resulted in very poor performance. This suggests that the different *perspectives* of annotators are evidently not represented in the training set.

4.1 Annotation process

A first step towards transparency about the perspectives involved in the datasets is the reflection provided by Prabhakaran et al. (2021), who propose a set of recommendations to increase the transparency of datasets, following up on the existing data statement presented in Bender and Friedman (2018). In particular, the authors underline the importance of release the annotator-level labels, the sociodemographic information of annotators to show the equitable representation of various social groups, and information on recruitment, selection, and assignment of annotators, in order to ensure representational diversity. Indeed, in their work, they have shown how the aggregation step may lead certain groups' perspectives to be under-represented, especially in contexts of annotation of subjective phenomena such as offensive language or emotion detection, and sentiment analysis. Basile et al. (2021) also present a list of important recommendations to ensure the transparency of the created dataset, including the collection of raters' confidence in order to identify the most difficult cases.

The importance of the role played by individual perspectives in the annotation process is supported by different studies in various NLP tasks. Almanea and Poesio (2022), for instance, observed that the disagreements in misogyny and sexism annotation of the ArMis corpus (Table 2) are due to the different degree of religious beliefs of annotators. Further evidence of the influence of individual traits, and specifically of the personalities of annotators, is provided by Labat et al. (2022) who investigate variation in the expression and annotation of emotions in human-computer conversations in the domain of customer service. They discovered in particular that extroverted personality is positively correlated with the dimension of dominance.

In the same vein, Viridiano et al. (2022) present the results of a preliminary study on frame annotation in a multimodal setting, where annotators were asked to indicate the frames and frame elements elicited by images. The experiments show how the first language of the annotators influences perception of the frames. While this work is focused purely on the dimension of language, in another work Timponi Torrent et al. (2022), they present a tool for manipulating semantic frames taking into account both language and cultural background.

Sang and Stanton (2022), instead, propose a methodology based on interviews, concept mapping exercises, and self-reporting items, for evaluating the annotation of a corpus of hate speech. On the one hand, this method helps comprehension of how people with different characteristics annotate hate speech and thus the origin of disagreement. On the other, their statistical analysis highlights that annotators respond differently based on their age and personality, and obtaining consistency among annotators with different backgrounds and personalities may never be fully successful.

In dealing with different annotations, the question about the assessment of the quality of corpus remains open. To this end, Haralabopoulos et al. (2020) study the problem of *subjectivity* in crowdsourced annotation of tasks such as sentiment analysis. The authors propose a method for injecting objectivity into the annotation tasks, thus identifying the source of disagreement. They do this by inserting

Table 2 Surveyed perspectivist datasets

Dataset	Language	Source	Access	Phenomenon	Annotators	Size	Info
IMSyPP_EN (Ljubešić et al., 2021)	EN	YouTube comments	Open	Hate speech	10	73,173	Not available
IMSyPP_IT (Cinelli et al., 2021)	IT	YouTube comments	Open	Hate speech	8	70,406	Not available
IMSyPP_SL (Kralj Novak et al., 2021)	SL	Tweets	Open	Hate speech	10	60,000	Not available
MHS (Kennedy et al., 2020)	EN	YouTube, Twitter, Reddit Comments	Open	Hate speech	11,143	50,070	Various
Brexit-HS (Akhtar et al., 2021)	EN	Tweets	Open	Hate speech, aggressiveness, offensiveness	6	1,120	Origin
ArMis (Almanea & Poesio, 2022)	AR	Tweets	Open	Misogyny	3	943	Religious belief
MultiDomain-Agreement (Leonardelli et al., 2021)	EN	Tweets	Open	Offensiveness	> 800	10,000	Not-available
ConvAbuse (Cercas Curry et al., 2021)	EN	Conversations Humans-AI	Open	Abusive language	8	6,837	Not-available
ToxCR (Kumar et al., 2021)	EN	Twitter, Reddit, 4chan comments	On request	Toxic language	17,280	107,620	Various
JSRPData (Goyal et al., 2022)	EN	Newscomments	Open	Toxic language	15	25,500	Social identities
PEJOR (Dinu et al., 2021)	EN, ES	Tweets	Open	Pejorative language	3	2,104	Not-available
Humour (Simpson et al., 2019)	EN	Short texts	On-request	Humour detection	272	18,002	Not-available
Concr-EMWE (Muraki et al., 2022)	EN	Multi-word expressions	Open	Concreteness	2,825	66,458	Not-available
MegaVeridicality (White et al., 2018)	EN	Clause-embedding verbs	Open	Inference	10	3,938	English native
PDIS (Poesio et al., 2019)	EN	Phrases	On-request	Phrase detection	1,741	96,305	Not-available
Gimpel-corpus (Plank et al., 2014)	EN	Words/tokens	On-request	POS tagging	177	14,000	Not-available
DBooks (Bizzoni et al., 2022)	DA	Book Reviews	Not-available	Book quality assessment	57,369	14,647	Gender

Table 2 (continued)

Dataset	Language	Source	Access	Phenomenon	Annotators	Size	Info
QT30NONAGGR (Hautli-Janisz et al., 2022)	EN	Dialogues	Open	Dialogical argumentation	38	10,818	Not-available
VQA_2.0+VIZWIZ (Bhattacharya et al., 2019)	EN	Visual question-answers	Open	VQA	10	45,000	Not-available
IC-LABELME (Rodrigues & Pereira, 2018)	EN	Images	On-request	Outdoor image classification	59	10,000	Not-available
CIFAR-10H (Krizhevsky & Hinton, 2009)	EN	Images	On-request	Subject identification	2,457	10,000	Not-available

In this table, the label `various` in the column *info* includes different sociodemographic information, such as education, gender, origin, religion, and so on

objective terms related to the target class and evaluating the annotators' labels solely on these instances. In this way, their method aims to determine whether the disagreement is caused by the subjectivity of the annotation task, or if it is caused by the unreliability of the annotators. This proposal for assessing the quality of a dataset is analagous to the traditional use of *majority voting strategies* that result in the filtering out of outliers in a non-perspectivist manner. Bizzoni et al. (2022) propose a mixed approach aimed at bridging the gap between subjectivity and the ground truth in the task of quality assessment of literary work. They build a new dataset (reported in the Table 2 below as DBooks) containing Danish literary reviews from 2010 and 2021 and clustered readers in multiple classes (i.e., gender or media type).

Other studies focus mainly on the exploration and management of disagreement. To explore reasons for disagreement, Hautli-Janisz et al. (2022) investigate annotation judgements in the QT30nonaggr corpus[3] (in Table 2) of disaggregated argument annotation, compiling a taxonomy of the types of label disagreements in argument annotation, and identifying the categories of 'annotation errors' (low interpretation of guidelines), 'fuzziness', and 'ambiguity'.

Regarding management of disagreement, Sachdeva et al. (2022) use Rasch Measurement Theory to aggregate labels on a continuous score, allowing annotator disagreement to be statistically summarized. The created dataset is the Measuring Hate Speech corpus (reported as MHS in Table 2) Kennedy et al. (2020),[4] which contains identity group targets and the annotators' sociodemographic information, as well as their estimated survey interpretation bias, difficulty, and rarity of decision.

The process of annotation and consequent creation of datasets is the basis of data-driven approaches to the design and building of models to solve NLP tasks. In the perspectivist framework, it is also one of the most important steps, as the perspectives of different groups of people are encoded in the datasets. As we have seen, different works address the theoretical problems of this step, proposing: a set of guidelines that encourages other researchers to dive deep into perspectivism (Prabhakaran et al., 2021); investigations that confirm the need to consider individual perspectives even in traditionally considered objective tasks, such as semantic disambiguation (Viridiano et al., 2022); and methods for assessing the quality of disaggregated datasets (Haralabopoulos et al., 2020) or for managing the disagreement (Sachdeva et al., 2022). Meanwhile, a greater quantity of literature is dedicated to the creation of disaggregated datasets on various topics.

4.2 Disaggregated datasets

From the sources selected in Sect. 3, the identified perspectivist datasets span a range of different NLP application and languages, such as the detection of toxic language, understanding differences in human behaviour in visual question answering, or event factuality predictions through linguistic inferences. Not all datasets are used to train

[3] http://corpora.aifdb.org/qt30nonaggr.

[4] https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech.

supervised ML models; some are compiled so that they can be made available to the research community for further analysis or to carry out specific language-oriented investigations.

The corpora outlined in Table 2 contain disaggregated annotations of textual and multimodal data, and some of them also report sociodemographic information about annotators (in the column *Info*). The majority of these datasets focus on the mapping and automatic detection of various types of toxic language, either on the internet or in spoken interactions; and only a few of them respect the recommendation of reporting information about annotators (Prabhakaran et al., 2021; Basile et al., 2021).

Some of the datasets in Table 2 have been released as part of the two editions of Le-Wi-Di at SemEval 2021 (Uma et al., 2021) and 2023 (Leonardellli et al., 2023): Gimpel-corpus, PDIS, Humour, IC-labelme and Cifar-10 h in 2021; and Brexit-HS, ArMis, ConvAbuse and MultiDomain-Agreement in 2023.

Only a few datasets report the sociodemographic information of annotators and thus allow perspective-based analysis (Sachdeva et al., 2022); Almanea & Poesio, 2022; Goyal et al., 2022; Bizzoni et al., 2022) or the creation of perspective-aware models (Akhtar et al., 2021; Kumar et al., 2021).

Regarding perspective-based analysis, in addition to the works already described above, Goyal et al. (2022) examine how annotator identities affect their labelling of toxicity when they are members of the target group. The authors explicitly target under-represented and minority groups that have been shown to be underserved by non-perspectivist approaches. In particular, they contracted crowd workers who self-identify as African American, LGBTQ, or neither for annotating examples from the Civil Comments dataset[5] with the presence or absence of toxic language (insult, obscene, threat, etc.). The created dataset, listed in Table 2 as JSRPData (Jigsaw Specialized Rater Pools Dataset), reports the annotator-level label and their identities. On the basis of this information, they found statistically significant variability, and showed through regression experiments the informative value of having annotators' explicit demographic metadata.

Others focused on building new disaggregated corpora to create perspective-aware systems in particular for abusive language detection. Akhtar et al. (2021), for instance, constructed Brexit-HS, a dataset of English tweets regarding Brexit annotated for hate speech, aggressiveness, offensiveness, stereotype, and irony. The dataset is annotated by six people divided into two groups: a target group composed of three Muslim immigrants in the United Kingdom, and a control group of three young researchers with western backgrounds. Brexit-HS was created with the explicit intent of representing the point of view of the victims of hate speech. As such, it served as a basis for studies on training perspective-aware models encoding different perspectives based on automatic clustering of the annotators.

Observing that current systems of toxic comment detection tend to fail in the consideration of various concerns of users, Kumar et al. (2021) created ToxCR to tune existing models in a more personalized direction. ToxCR is an English dataset,

[5] Available on https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification.

which includes the results of a survey of 17,280 residents in the United States aiming to understand how toxic content perception differs across demographics, beliefs, and personal experiences with harassment. Exploiting this new resource, the authors showed how current toxicity classification algorithms, such as Perspective API from Jigsaw,[6] can improve in accuracy by 86 percent on average by tuning the model while taking individual annotator differences into account.

The other datasets reported in Table 2 have been created and released for addressing various NLP tasks, ranging from the highly subjective to others traditionally considered objective. The majority of the considered datasets report annotation of abusive language in different genres of texts. The datasets IMSyPP_EN,[7] IMSyPP_IT,[8] and IMSyPP_SL[9] were created within the IMSyPP project (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech) with the aim of identifying terms, entities and concepts to understand the triggers and facilitators of hate speech, develop models for hate speech detection, and monitor it in English, Italian, and Slovenian respectively. Specifically, Ljubešić et al. (2021) built the English dataset containing YouTube comments annotated with various hate speech types (appropriate, inappropriate, offensive, and violent) and targets (racism, migrants, Islamophobia, antisemitism, religion, sexism, homophobia, ideology, media, politics, individual, and other). Cinelli et al. (2021) and Kralj Novak et al. (2021) collected similar datasets for Italian and Slovenian YouTube comments, respectively, following the same schema and method of annotation used for the English part.

On a similar topic, Cercas Curry et al. (2021) investigated the lesser researched field of abusive language directed towards AI conversational agents They present ConvAbuse,[10] a corpus of conversations between humans and three conversational AI systems which are labelled by multiple domain experts selected on the basis of being members of the groups typically targeted by such abuse, or being experts in such issues.

Dinu et al. (2021) investigated pejorative language in social media, understood as a mainly lexical phenomenon which occurs when a word is used with a negative connotation and/or within a negative context. The authors propose a multilingual lexicon and dataset of English, Spanish, Italian, and Romanian tweets[11] selected from existing datasets.

The rest of the datasets address tasks related to meaning, veridicality and question-answering. Muraki et al. (2022) provided the first large concreteness rating dataset for multi-word expressions (Concr-EMWE).[12] Although the annotators were instructed to give a concreteness rating ranging from 1 (very abstract) to 5 (very

[6] https://www.perspectiveapi.com/.
[7] https://www.clarin.si/repository/xmlui/handle/11356/1454.
[8] https://www.clarin.si/repository/xmlui/handle/11356/1450.
[9] https://www.clarin.si/repository/xmlui/handle/11356/1398.
[10] https://github.com/amandacurry/convabuse.
[11] http://nlp.unibuc.ro/resources.html#pejor.
[12] https://osf.io/ksypa/.

concrete), in this work the possibility of having multiple meanings for an expression was not taken into account.

White et al. (2018) studied linguistic inferences that are triggered by the interaction of lexical and syntactic information for studying precisely how to predict the factuality of an event. To this purpose, they created the MegaVeridicality dataset,[13] that contains veridicality judgments by English native and not-native speakers for verbs that embed declarative clauses.

Finally, Bhattacharya et al. (2019) labelled two datasets (VQA_2.0 and VizWiz) for visual question answering (VQA), i.e., the task of returning an answer to an image-based question. For each image-question pair, 10 crowdsourced answers were collected, with metadata indicating the reasons why the 10 responses to each question are different.[14] Indeed, this is an intrinsically complex and ambiguous task, where different annotators often provide different answers. The authors performed a preliminary analysis, building the basis for a new algorithm that allows to predict directly from a visual question which reasons will cause answer differences.

4.3 Exploring perspectives

Jointly with the creation of datasets that encode the perspectives defined and identified *a priori* by scholars, an interesting and novel process introduced by perspectivism is the exploration of perspectives and variance of annotation of different phenomena.

To this purpose, Havens et al. (2022) present exploratory text visualization techniques as a complement to data perspectivism, investigating patterns and outliers in annotations of the text in a perspectivist context. Authors identify limits in the current visualization tools for perspectivist research and encourage the development of new platforms with interactive, exploratory text visualizations, in which data analysis becomes an intuitive process relying on human vision and cognition.

Another contribution is a framework for analysing the variation of per-item annotator response distributions to data for humans-in-the-loop machine learning (Homan et al., 2022). The authors employ a crowdsourced dataset of hard-to-classify examples of the Open Images Dataset archive, providing visualizations and analysing the variance of the annotation. The experiments show that hypothesis testing can be used to identify particularly anomalous distributional patterns due to individual annotators' perspectives.

Along with that, Biester et al. (2022) describe a new robust framework to test for differences in annotation across various demographic groups. The case study presented in their work focuses on annotators' gender in four benchmark datasets (Affective Text, Word Similarity, Sentiment Analysis and Natural Language Inference).

[13] http://megaattitude.io/projects/mega-veridicality/.
[14] https://vizwiz.org/.

Although more weakly *perspectivist* than some of these approaches, Marchiori Manerba et al. (2022) propose a preliminary method of bias discovery in human annotators on tasks involving subjective judgements. In this work, bias is manifested especially in the rating of highly sensitive topics such as toxicity, and has negative repercussions on the reliability of supervised machine learning by reinforcing and propagating such bias. Disaggregated data are used in this work to both preserve the individual perspective of raters even in the face of disagreement and to inform the models' predictions.

5 Perspectives in models

In this section, we showcase papers that present innovative methods for designing models capable of leveraging disaggregated datasets to model perspectives (Sect. 5.1), the evaluation of designed models (Sect. 5.2) and the explainability of their decisions (Sect. 5.3).

5.1 Explicitly modelling perspectives

The modelling can be done at different levels of depth and can take into account various aspects such as biases, personalization, or disagreement. We have identified several macro categories of relevant papers.

Modelling label distribution The first category involves modelling the distribution of labels in a disaggregated dataset without intending to model underlying perspectives. The only such work is Liu et al. (2019), who present a novel approach to learn distributions of disaggregated labels, which reduces the number of required labels for classification tasks. The authors achieve this by clustering semantically similar instances and aggregating the labels within each cluster into a single larger sample.

Modelling disagreement The two papers in the second category instead aim at modelling the disagreement among groups of annotators and thus modelling multiple perspectives. Akhtar et al. (2020) do so by proposing an innovative method for dividing annotators into (two) groups. Their approach seeks to maximize intragroup agreement and minimize intergroup agreement. While their experimental results on hate speech detection demonstrate the potential for improved classification performance, the authors' aim extends beyond this objective. They seek to identify annotator clusters and model the perspectives of each group. Davani et al. (2022) instead, evaluate three novel strategies for training on to shift the majority vote-based aggregation step from the dataset creation stage to the final prediction stage after model training and are based on multi-task approaches. Instead of attempting to mitigate or eliminate disagreement among annotators, the effort of the authors to develop models with different internal tasks, reveals their aim to model such disagreement and in turn individual perspectives in hate speech and emotion detection. Notably, their results demonstrate that a multi-task approach performs better and is more effective at estimating uncertainty in predictions.

Modelling individual perspectives The third category includes three works leveraging personalized approaches to modelling perspectives. Kanclerz et al. (2022) propose three user-centred methods for detecting hate speech on three different datasets. They argue that aggregating labels into a gold standard, particularly in subjective NLP tasks, leads to overlooking individual perspectives during the training of predictive models. The authors demonstrate this by showing that personalized methods are more effective in allowing individual perspectives to emerge, compared to generalized approaches. The study's findings indicate that personalized approaches have significant potential for improving the accuracy of hate speech detection in NLP tasks. Kanclerz et al. (2021) used a personalized approach for discriminating perceptions of offensive, toxic, and aggressive documents based on an individual's previous annotations. To achieve this, the authors utilize conformity-based personalization, class-based embeddings, and annotation-based embeddings which model personalized context variables. The results indicate that the approach is valid, especially for controversial documents, and effective for subjective NLP tasks such as recognizing emotions. Overall, the personalized approach demonstrates great potential for improving the accuracy of perception discrimination in NLP. Finally, Ngo et al. (2022) use personalized models for emotion prediction. Their models are based on three sources of information: the annotators themselves, their previous annotations, and an encoding of the text that they annotated. The authors conducted experiments on a new disaggregated dataset, StudEmo, and found that each of the features contributed to improved performance, albeit to different extents, when compared to a non-personalized baseline. Of particular note is the finding that personalized knowledge about user beliefs was especially helpful in making decisions about controversial emotions.

Modelling human beliefs The fourth category is about modelling human beliefs from the annotations. For instance, Kocoń et al. (2021) address the problem of modelling the human perspective of a user towards a specific text presenting an approach to enable adaptive models to predict personal and group beliefs, rather than agreed-upon aggressiveness labels. The authors identify three levels of content analysis: macroscopic (perspective of the whole society), mesoscopic (group perspective), and microscopic (individual perspective). Kocoń et al. (2021) then extend the approach to other subjective phenomena, such as emotions and toxicity. In particular, inspired by the Neural Collaborative Filtering model used in recommender systems, they represent human beliefs as latent vectors. Disaggregated data, mainly from the Wikipedia Detox project,[15] are used by the authors (i) to quantify the *human bias*, intended as the distance between the known annotations of a given user and the average annotations provided by other annotators, and (ii) to create a model of human beliefs. The results are promising and confirm the significant impact of the personalized representations on the performance of the models in subjective tasks.

Modelling human bias The fifth (and last) category is focused on human biases. Milkowski et al. (2021) introduce a novel measure estimating Personal Emotional

[15] https://meta.wikimedia.org/wiki/Research:Detox.

Bias (PEB) in evaluating opinions. PEB measures the extent to which previously known annotations of a given user differ from the average annotations provided by all others for a given emotional category, aggregated over all documents. Authors have experimentally validated that models incorporating knowledge provided by PEB obtain a gain in predicting emotions from an individual human perspective, improving the quality of personalized reasoning.

5.2 Perspectivist evaluation

While the majority of works in the field of data evaluation relies on a single aggregated ground truth, there have been a few notable contributions toward evaluations that take into account multiple perspectives. One such work is Gordon et al. (2021), which addresses the difference between evaluating machine learning systems and human-computer interaction. While the former typically uses measures based on generalization error to evaluate technical performance, the latter focuses on collecting user opinions to report on user-facing experience.

To bridge this gap, the authors proposed the *disagreement deconvolution method*, which considers the multiple ground truths provided by annotators and the probability that a given annotation may differ from the most common annotator's response. By adjusting the label distribution of responses for each item using this probability, they built a new test set containing many annotations per item from the disaggregated corpus. This represents a step forward in establishing useful performance evaluations for social computing classifiers, such as toxicity detectors. When applied to the Jigsaw dataset, this method demonstrated how current metrics may overstate the performance of ML-based systems.

The limited number of works we discovered addressing perspectivist evaluation highlights the need and opportunities for further research in this area.

5.3 Explainability with perspectives

The last step in the pipeline is explainability of model decisions. An innovative work on in this direction within the perspectivist realm is presented by Mastromattei et al. (2022). The authors designed a method to enhance the explainability of large language models applied to text classification. The method is based on analysing the activation of syntax subtrees and contrasting the predictions made by different perspective-aware classifiers trained on disaggregated data, in order to highlight the syntactic structures that influence a certain stance. The syntax-based models used in the study are KERMIT (Zanzotto et al., 2020) and its specialization aimed at the analysis of hate speech KERM-HATE (Mastromattei et al., 2022). A pilot experiment is presented on Brexit-HS (Table 2), a perspectivist dataset annotated with hate speech (Akhtar et al., 2020). While KERM-HATE obtains the best classification performance, the most relevant contribution is the proposed perspectivist approach to provide visual syntax-based representations of the syntactic components that are responsible for the predictions of the models along different perspectives.

The scarcity of works exploring explainability with perspectives presents an opportunity for the research community to explore this area further.

6 Discussion and conclusions

In this survey, for the first time, we have selected and analysed a number of datasets and research papers proposing resources and methods intended to leverage the information encoded in disaggregated annotations in order to represent and account for individual and group perspectives in NLP data and tasks, i.e., perspectivist approaches in NLP.

One of the goals of this survey was to identify and understand the key concepts around perspectivism in NLP and their mutual relations. As a result of the analysis presented here, we find that three factors play intertwined roles in this area.

- *Disagreement* is the observed, quantifiable phenomenon that arise, to a different extent, whenever manual annotation is conducted.
- *Subjectivity* is an intrinsic characteristic of a phenomenon in natural language, intended as the extent to which the individual perception of the annotator influences their annotation. Subjectivity is one of the possible causes of disagreement.
- *Reliability* is a characteristic of annotators (and the labels they produce). Unreliable annotators induce noise, which may increase the disagreement even when the subjectivity of the task is low. As suggested by Haralabopoulos et al. (2020), it is important to discriminate between (un)reliability and subjectivity as possible causes of the observed disagreement.[16]

This survey represents a first step towards disentangling these concepts and their relationships.

Another important goal of this survey is the exploration of existing literature that make use of perspectivist principles. In particular, the research papers we surveyed (Sects. 4, 5) were selected mainly on the basis of their teleological dimension, i.e., the focus of this survey is on research work whose main objective is of a perspectivist nature. The datasets (Table 2) were instead included on the more relaxed basis of being published with disaggregated annotation, but we still considered the teleological aspects in their analysis.

More than half of the *perspectivist datasets* we analysed revolve around hate speech and other undesirable language phenomena. This is not surprising, considering that the added value of using disaggregated data is more evident the more *subjective* is the observed linguistic phenomenon (in the sense of subjective

[16] Reliability has traditionally been assessed with Intra-Annotator Agreement measures. However, in the perspectivist framework, disagreement in not necessarily treated as a sign of poor reliability. For a better understanding of the underlying reasons for disagreement, Abercrombie et al. (2023) propose combining these measures with analysis of annotator *stability*.

perception having a stronger role during annotation). The annotation of hate speech and toxic language, and related phenomena, is indeed a prime example of a situation where traditional, pre-perspectivist approaches to annotation fall short (Fortuna et al., 2020). However, we found, several datasets on a variety of domains and tasks which are distributed with disaggregated annotations, indicating how a perspectivist stance may provide a new angle to interpret language phenomena such as argumentation and inference. Moreover, bearing in mind the recommendations for approaching tasks in a perspectivist manner (Basile et al., 2021; Prabhakaran et al., 2021), we noticed that only a few datasets report sociodemographic information about annotators, and only Sachdeva et al. (2022) also include the difficulty and rarity of the annotators' decision. The access to this information allows scholars and researchers to create and design more accurate perspective-aware models.

The *perspectivist methods* analysed in the surveyed papers may be summarized and categorized in different ways. Firstly, they propose solutions focused on different steps of a hypothetical "perspectivist pipeline": (i) *training* of models capable of encoding different perspectives and using them for their predictions (Sect. 5.1); (ii) *evaluation* of the models against disaggregated benchmarks, in order to test their performance in encoding multiple perspectives (Sect. 5.2); (iii) leveraging perspectivist approaches to provide an *explanation* of the models that account for multiple perspectives (Sect. 5.3). Furthermore, some works pertain to what could be considered a preliminary step of said pipeline, that is, supporting the creation of datasets on which perspectivist approaches will rely (Sect. 4).

On the other hand, the surveyed research works can be analysed along more traditional dimensions, such as the task or domain on which the approach is focused. Most methods to improve data curation, while in principle domain-agnostic, are motivated by the creation of datasets for subjective phenomena such as hate speech, toxicity, sentiment, and emotions. Notable exceptions are represented by the multimodal contribution of Homan et al. (2022) and the works by researchers of the FrameNet Brazil project,[17] focusing on the semantic annotation of frames.

The computational methods for modelling and analysing perspectives are often designed with specific tasks as their goal, in particular emotion and toxic language detection. Some of them also attempt to build strongly personified models, like Kocoń et al. (2021) which could be applied in more general contexts. However, a number of proposals are relatively task-agnostic and can be considered as general methods to leverage disaggregated data to train more informed models.

Perspectivist evaluation and explainability are the two least represented areas, according to our survey. This may be due to the inherent difficulty of performing a perspectivist evaluation, which requires a significant overhaul of the structure of benchmarks and evaluation metrics. Indeed, the importance (and difficulty) of evaluation has been highlighted since the first position papers proposing the paradigm shift towards perspectivism in NLP (Basile, 2021; Basile et al., 2021). Furthermore, as "late" steps in the perspectivist pipeline mentioned earlier, effective

[17] https://www2.ufjf.br/framenetbr-en/.

evaluation explainability methodologies will require more robust foundations in the earlier steps to build upon.

The results of the analyses carried out on the resources and papers presented in this survey prompted a few observations. First of all, we notice how almost all the surveyed works are from 2020 onward, therefore they are quite recent. At the same time, while the survey is focused on the area of NLP, several authors are active in other areas of Computer Science, such as Privacy (Kumar et al., 2021), Computer Vision (Bhattacharya et al., 2019), and Healthcare (Kennedy et al., 2020). Moreover, various surveyed works borrow and integrate techniques originally developed from recommender systems, user modelling, or personalization. These two observations jointly paint the picture of a new research community in-the-making, centred on NLP yet interdisciplinary.

Looking at the datasets in Table 2, it is evident how the issues relevant to perspectivism are raised more often in conjunction with the study of highly subjective linguistic phenomena. In particular, offensive and other undesirable language often triggers reflections and proposals for modelling and leveraging different opinions and perspectives. Nevertheless, the number of datasets and methods presented in this survey that do not pertain to subjective phenomena (e.g., frame annotation in Timponi Torrent et al. (2022) and Viridiano et al. (2022)), suggests that perspectivist approaches may benefit NLP tasks regardless of their subjectivity. This is in line with the prescriptions of Basile et al. (2021) of adopting a perspectivist stance even for tasks in "unsuspected domains, like medical diagnosis".

Finally, in analysing the sources, we realized that a paradigm shift is currently being proposed, ignited by the studies on the informativeness of annotation disagreement (Sect. 2) and moving forward to cover all aspects of NLP research based on manually annotated data. At the technical level, the main challenges towards the adoption of the new perspectivist paradigm are represented by the need to develop new methodologies for training and evaluating models that leverage disaggregated datasets for perspectivist purposes. At the same time, preliminary studies appear promising towards the development of a new generation of fair, explainable NLP models under the perspectivist paradigm.

Appendix A

Questions and sample answers used for the analysis of papers.

1. *How is perspectivism defined/framed or interpreted?*
2. *Are the authors proposing a perspectivist approach for a subjective or an objective task?*
3. *What's the role/purpose of perspectivism in this work?*

    a. Perspectivism constitutes the backbone of this work and its main purpose is to find ways of surfacing multiple perspectives present in the data, or modeling systems aware of the subjectivities or creating NLP fair models or NLP transparent models.

   b. Perspectivism is taken into account but the purpose of this work is not just to identify multiple perspectives or create aware/fair/transparent models but also [something else].
   c. Perspectivism can be inferred but it's not within the main purposes of this work.

4. *What are the novelties (i.e., new methods/models/approaches, datasets with disaggregated annotations, new evaluation metrics/methodologies, new user cohorts) of this work?*

   a. This work describes and tests a completely novel method/data set etc. thus exploring uncharted territory;
   b. This work introduces some elements of novelty and contributes to a known line of research;
   c. This work is a replication study with some variations (or something along these lines)

5. *What are the advantages of their perspectivist approach in respect to a pre/not-perspectivist approach?* This is an open-ended question since each work may present a different set of advantages rather than options on a gradient.
6. *What are the limits/challenges/disadvantages of this work?*

   a. The work proposed is not widely applicable because of major faults.
   b. The method/data etc. are somewhat limited in their application/there are minor faults/some factors were not taken into consideration.
   c. The method/approach proposed is widely applicable and can be used in further research by others.

Acknowledgements The work of S. Frenda and V. Basile was funded by the Multilingual Perspective-Aware NLU Project in partnership with Amazon Alexa. G. Abercrombie was supported by the EPSRC projects 'Gender Bias in Conversational AI' (EP/T023767/1) and 'Equally Safe Online' (EP/W025493/1). The work of A.T. Cignarella was funded by the International project 'STERHEOTYPES - Studying European Racial Hoaxes and sterEOTYPES' funded by the Compagnia di San Paolo and VolksWagen Stiftung under the 'Challenges for Europe' call for Project (CUP: B99C20000640007).

Author Contributions We can resume the contribution of the authors as follows. Conceptualization: V. Basile, G. Abercrombie; Methodology: V. Basile, S. Frenda; G. Abercrombie, R. Panizzon; Revision of sources: V. Basile, S. Frenda; G. Abercrombie, R. Panizzon, A. Pedrani, D. Bernardi, A. T. Cignarella, C. Marco; Writing—original draft preparation: S. Frenda, V. Basile, G. Abercrombie, R. Panizzon, A. Pedrani; Writing—review and editing: S. Frenda, D. Bernardi, G. Abercrombie; Funding acquisition: V. Basile; Resources: V. Basile, S. Frenda; Supervision: V. Basile, S. Frenda.

Funding Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

Abercrombie, G., Basile, V., Tonelli, S., Rieser, V., & Uma, A. (Eds.) (2022). *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022*. European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1

Abercrombie, G., Hovy, D., & Prabhakaran, V. (2023). Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In Prange, J., Friedrich, A. (Eds.), *Proceedings of the 17th linguistic annotation workshop (LAW-XVII)* (pp. 96–103). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.law-1.10. https://aclanthology.org/2023.law-1.10

Abercrombie, G., Jiang, A., Gerrard-abbott, P., Konstas, I., & Rieser, V. (2023). Resources for automated identification of online gender-based violence: A systematic review. In: Y.-l. Chung, P. Röttger, D. Nozza, Z. Talat, & A. Mostafazadeh Davani (Eds.), *The 7th workshop on online abuse and harms (WOAH)* (pp. 170–186). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.woah-1.17. https://aclanthology.org/2023.woah-1.17

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science, 3*, 245–270. https://doi.org/10.1007/s42001-019-00060-w

Akhtar, S., Basile, V., & Patti, V. (2021). Whose opinions matter? Perspective-aware models to identify opinions of Hate Speech victims in Abusive Language detection. arXiv:2106.15896.

Akhtar, S., Basile, V., & Patti, V. (2020). Modeling annotator perspective and polarized opinions to improve Hate Speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 8*(1), 151–154.

Almanea, D., & Poesio, M. (2022). ArMIS—The Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the 13th language resources and evaluation conference* (pp. 2282–2291). European Language Resources Association. https://aclanthology.org/2022.lrec-1.244

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine, 36*(1), 15–24. https://doi.org/10.1609/aimag.v36i1.2564

Balloccu, S., Schmidtová, P., Lango, M., & Dušek, O. (2024). Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of EACL 2024*. Association for Computational Linguistics.

Basile, V., Cabitza, F., Campagner, A., & Fell, M. (2021). Toward a perspectivist turn in ground truthing for predictive computing. arXiv:2109.04270.

Basile, V., Caselli, T., Guerini, M., Cignarella, A. T., Poesio, M., Stranisci, M .A., Sanguinetti, M., Cabitza, F., Patti, V., Rieser, V., Derczynski, L., Ravelli, A. A., Abercrombie, G., Tonelli, S., Miltenburg, E., Rosso, P., Camacho-Collados, J., Dudy, S., Dinu, L. P., Manerba, M. M., Homan, C. M., Havens, L., Frenda, S., Ciucci, D., & Markantonatou, S. (2021). The Perspectivist Data Manifesto. Retrieved July 29, 2022, from https://pdai.info/

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., & Uma, A. (2021). We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future* (pp. 15–21). Association for Computational Linguistics (Online). https://doi.org/10.18653/v1/2021.bppf-1.3. https://aclanthology.org/2021.bppf-1.3

Basile, V. (2021). It's the end of the gold standard as we know it. In M. Baldoni & S. Bandini (Eds.), *AIxIA 2020—Advances in Artificial Intelligence* (pp. 441–453). Springer.

Belz, A., Agarwal, S., Shimorina, A., & Reiter, E. (2021). A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume* (pp. 381–393). Association for Computational Linguistics (Online). https://doi.org/10.18653/v1/2021.eacl-main.29. https://aclanthology.org/2021.eacl-main.29

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Bhattacharya, N., Li, Q., & Gurari, D. (2019). Why does a visual question have different answers? In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4271–4280). https://openaccess.thecvf.com/content_ICCV_2019/papers/Bhattacharya_Why_Does_a_Visual_Question_Have_Different_Answers_ICCV_2019_paper.pdf

Biester, L., Sharma, V., Kazemi, A., Deng, N., Wilson, S., & Mihalcea, R. (2022). Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 10–19). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.2

Bizzoni, Y., Lassen, I.M., Peura, T., Thomsen, M.R., & Nielbo, K. (2022). Predicting literary quality how perspectivist should we be? In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 20–25). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.3

Blodgett, S. L. (2021). *Sociolinguistically driven approaches for just Natural Language Processing* (Ph.D thesis, University of Massachusetts Amherst). https://doi.org/10.7275/20410631. https://scholarworks.umass.edu/dissertations_2/2092

Cercas Curry, A., Abercrombie, G., & Rieser, V. (2021). ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 7388–7403). Association for Computational Linguistics (Online)https://doi.org/10.18653/v1/2021.emnlp-main.587. https://aclanthology.org/2021.emnlp-main.587

Cercas Curry, A., Abercrombie, G., & Talat, Z. (2024). Subjective *Isms*? On the sanger of conflating hate and offense in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)* (pp. 275–282), Mexico City, Mexico: Association for Computational Linguistics. https://aclanthology.org/2024.woah-1.22.pdf

Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Kralj Novak, P., & Zollo, F. (2021). Italian YouTube Hate Speech Corpus. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1450

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, **10**, 92–110. https://doi.org/10.1162/tacl_a_00449 https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00449/1986597/tacl_a_00449.pdf.

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics, 10*, 92–110.

Dinu, L. P., Iordache, I.-B., Uban, A. S., & Zampieri, M. (2021). A computational exploration of pejorative language in social media. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 3493–3498). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.296. https://aclanthology.org/2021.findings-emnlp.296

Fortuna, P., Soler, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of Hate Speech datasets. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6786–6794). European Language Resources Association. https://aclanthology.org/2020.lrec-1.838

Glenn, P., Jacobs, C. L., Thielk, M., & Chu, Y. (2022). The viability of best-worst scaling and categorical data label annotation tasks in detecting implicit bias. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 32–36). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.5

Gordon, M.L., Zhou, K., Patel, K., Hashimoto, T., & Bernstein, M. S. (2021). The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21*. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445423

Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T., & Bernstein, M. S. (2021). The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In: Proceedings of the 2021 CHI conference on human factors in computing systems. CHI '21. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445423

Goyal, N., Kivlichan, I., Rosen, R., & Vasserman, L. (2022). Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. In *Proceedings of ACM in human computer interaction in ACM conference on computer-supported cooperative work and social computing CSCW 2022*.

Haralabopoulos, G., Tsikandilakis, M., Torres Torres, M., & McAuley, D. (2020). Objective assessment of subjective tasks in crowdsourcing applications. In *Proceedings of the LREC 2020 workshop on "Citizen Linguistics in Language Resource Development"* (pp. 15–25). European Language Resources Association. https://aclanthology.org/2020.cllrd-1.3

Hautli-Janisz, A., Schad, E., & Reed, C. (2022). Disagreement space in argument analysis. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 1–9). European Language Resources Association (ELRA). https://aclanthology.org/2022.nlperspectives-1.1.pdf

Havens, L., Bach, B., Terras, M., & Alex, B. (2022). Beyond explanation: A case for exploratory text visualizations of non-aggregated, annotated datasets. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 73–82). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.10

Homan, C., Weerasooriya, T. C., Aroyo, L., & Welty, C. (2022). Annotator response distributions as a sampling frame. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @ LREC2022* (pp. 56–65). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.8

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., Miltenburg, E., Santhanam, S., & Rieser, V. (2020). Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th international conference on natural language generation* (pp. 169–182). Association for Computational Linguistics. https://aclanthology.org/2020.inlg-1.23

Jamison, E., & Gurevych, I. (2015). Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 291–297). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1035. https://www.aclweb.org/anthology/D15-1035

Kanclerz, K., Figas, A., Gruza, M., Kajdanowicz, T., Kocon, J., Puchalska, D., & Kazienko, P. (2021). Controversy and conformity: From generalized to personalized aggressiveness detection. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers)* (pp. 5915–5926). Association for Computational Linguistics (Online). https://doi.org/10.18653/v1/2021.acl-long.460. https://aclanthology.org/2021.acl-long.460

Kanclerz, K., Gruza, M., Karanowski, K., Bielaniewicz, J., Milkowski, P., Kocon, J., & Kazienko, P. (2022). What if ground truth is subjective? Personalized deep neural Hate Speech detection. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 37–45). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.6

Kennedy, C. J., Bacon, G., Sahn, A., & Vacano, C. (2020). Constructing interval variables via faceted Rasch measurement and multitask deep learning: A Hate Speech application. arXiv preprint arXiv: 2009.10277.

Kitchenham, B. (2007). *Guidelines for performing systematic literature reviews in software engineering* (pp. 1–65). EBSE Technical Report EBSE-2007-01.

Kocoń, J., Gruza, M., Bielaniewicz, J., Grimling, D., Kanclerz, K., Miłkowski, P., & Kazienko, P. (2021). Learning personal human biases and representations for subjective tasks in Natural Language Processing. In *2021 IEEE international conference on data mining (ICDM)* (pp. 1168–1173). https://doi.org/10.1109/ICDM51629.2021.00140

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management, 58*(5), 102643. https://doi.org/10.1016/j.ipm.2021.102643

Kralj Novak, P., Mozetič, I., & Ljubešić, N. (2021). Slovenian Twitter Hate Speech dataset IMSyPP-sl. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1398

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. MSc thesis. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Kumar, D., Kelley, P.G., Consolvo, S., Mason, J., Bursztein, E., Durumeric, Z., Thomas, K., & Bailey, M. (2021). Designing toxic content classification for a diversity of perspectives. In *17th symposium on usable privacy and security (SOUPS 2021)* (pp. 299–318). arXiv:2106.04511

Labat, S., Ackaert, N., Demeester, T., & Hoste, V. (2022). Variation in the expression and annotation of emotions: A wizard of oz pilot study. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 66–72). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.9

Leonardelli, E., Menini, S., Aprosio, A. P., Guerini, M., Tonelli, S. (2021). Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10528–10539). Association for Computational Linguistics (Online). https://aclanthology.org/2021.emnlp-main.822

Leonardellli, E., Abercrombie, G., Almanea, D., Basile, V., Fornaciari, T., Plank, B., Poesio, M., Rieser, V., & Uma, A. (2023). SemEval-2023 Task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th international workshop on semantic evaluation*. Association for Computational Linguistics.

Liu, T., Venkatachalam, A., Sanjay Bongale, P., & Homan, C. (2019). Learning to predict population-level label distributions. In *Companion proceedings of The 2019 World Wide Web Conference. WWW '19* (pp. 1111–1120). Association for Computing Machinery. https://doi.org/10.1145/3308560.3317082

Ljubešić, N., Mozetič, I., Cinelli, M., & Kralj Novak, P. (2021). English YouTube Hate Speech Corpus. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1454

Marchiori Manerba, M., Guidotti, R., Passaro, L., & Ruggieri, S. (2022). Bias discovery within human raters: A case study of the jigsaw dataset. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022* (pp. 26–31). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.4

Mastromattei, M., Basile, V., & Zanzotto, F. M. (2022). Change my mind: How syntax-based Hate Speech recognizer can uncover hidden motivations based on different viewpoints. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 117–125). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.15

Mastromattei, M., Ranaldi, L., Fallucchi, F., & Zanzotto, F. M. (2022). Syntax and prejudice: Ethically-charged biases of a syntax-based Hate Speech recognizer unveiled. *PeerJ Computer Science, 8*, 859. https://doi.org/10.7717/peerj-cs.859

Milkowski, P., Gruza, M., Kanclerz, K., Kazienko, P., Grimling, D., & Kocon, J. (2021). Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: student research workshop* (pp. 248–259). Association for Computational Linguistics (Online). https://doi.org/10.18653/v1/2021.acl-srw.26. https://aclanthology.org/2021.acl-srw.26

Muraki, E. J., Abdalla, S., Brysbaert, M., & Pexman, P. M. (2023). Concreteness ratings for 62,000 English multiword expressions. *Behavior research methods, 55*(5), 2522–2531. https://doi.org/10.3758/s13428-022-01912-6

Ngo, A., Candri, A., Ferdinan, T., Kocon, J., & Korczynski, W. (2022). StudEmo: A non-aggregated review dataset for personalized emotion recognition. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 46–55). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.7

Plank, B., Hovy, D., & Søgaard, A. (2014). Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 507–511). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2083. https://www.aclweb.org/anthology/P14-2083

Poesio, M., Chamberlain, J., Paun, S., Yu, J., Uma, A., & Kruschwitz, U. (2019). A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. 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)* (pp. 1778–1789). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1176. https://aclanthology.org/N19-1176

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for Hate Speech detection: A systematic review. *Language Resources and Evaluation, 55*(2), 477–523.

Prabhakaran, V., Mostafazadeh Davani, A., & Diaz, M. (2021). On releasing annotator-level labels and information in datasets. In *Proceedings of the joint 15th linguistic annotation workshop (LAW) and 3rd designing meaning representations (DMR) workshop* (pp. 133–138). Association for Computational Linguisticshttps://doi.org/10.18653/v1/2021.law-1.14. https://aclanthology.org/2021.law-1.14

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics, 44*(3), 393–401. https://doi.org/10.1162/coli_a_00322

Rodrigues, F., & Pereira, F. (2018). Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v32i1.11506

Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. B. (2022). Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*. Association for Computational Linguistics. https://aclanthology.org/2022.naacl-main.13.pdf

Sachdeva, P., Barreto, R., Bacon, G., Sahn, A., Vacano, C., & Kennedy, C. (2022). The measuring Hate Speech corpus: Leveraging Rasch measurement theory for data perspectivism. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 83–94). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.11

Sang, Y., & Stanton, J. (2022). The origin and value of disagreement among data labelers: A case study of individual differences in Hate Speech annotation. In M. Smits (Ed.), *Information for a better world: Shaping the global future* (pp. 425–444). Springer.

Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., & Zeldes, A. (2023). Treebanking user-generated content: A UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation* 57, 493–544. https://doi.org/10.1007/s10579-022-09581-9

Sayeed, A. (2013). An opinion about opinions about opinions: Subjectivity and the aggregate reader. In L.. Vanderwende, H. Daumé III, & K. Kirchhoff (Eds.), *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 691–696). Association for Computational Linguistics. https://aclanthology.org/N13-1081.

Simpson, E., Do Dinh, E.-L., Miller, T., & Gurevych, I. (2019). Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5716–5728). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1572. https://aclanthology.org/P19-1572

Timponi Torrent, T., Lorenzi, A., Matos, E. E., Belcavello, F., Viridiano, M., & Andrade Gamonal, M. (2022). Lutma: A frame-making tool for collaborative FrameNet development. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 100–107). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.13

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., & Poesio, M. (2021). SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)* (pp. 338–347). Association for Computational Linguistics (Online). https://doi.org/10.18653/v1/2021.semeval-1.41. https://aclanthology.org/2021.semeval-1.41

Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2020). A case for soft-loss functions. In *Proceedings of the 8th AAAI conference on human computation and crowdsourcing* (pp. 173–177). https://ojs.aaai.org/index.php/HCOMP/article/view/7478

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2022). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research, 72*, 1385–1470. https://doi.org/10.1613/jair.1.12752

Viridiano, M., Timponi Torrent, T., Czulo, O., Lorenzi, A., Matos, E., & Belcavello, F. (2022). The case for perspective in multimodal datasets. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 108–116). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.14

Weerasooriya, T. C., Ororbia, A., & Homan, C. (2022). Improving label quality by jointly modeling items and annotators. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022* (pp. 95–99). European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.12

White, A. S., Rudinger, R., Rawlins, K., & Van Durme, B. (2018). Lexicosyntactic inference in neural models. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4717–4724). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1501. https://aclanthology.org/D18-1501

Zanzotto, F. M., Santilli, A., Ranaldi, L., Onorati, D., Tommasino, P., & Fallucchi, F. (2020). KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp.

256–267). Association for Computational Linguistics (Online). https://doi.org/10.18653/v1/2020.emnlp-main.18. https://aclanthology.org/2020.emnlp-main.18

Authors and Affiliations

Simona Frenda[1] · Gavin Abercrombie[2] · Valerio Basile[1] · Alessandro Pedrani[3] · Raffaella Panizzon[3] · Alessandra Teresa Cignarella[1] · Cristina Marco[3] · Davide Bernardi[3]

✉ Simona Frenda
simona.frenda@unito.it

Gavin Abercrombie
g.abercrombie@hw.ac.uk

Valerio Basile
valerio.basile@unito.it

Alessandro Pedrani
pedrana@amazon.it

Raffaella Panizzon
panizzor@amazon.it

Alessandra Teresa Cignarella
alessandrateresa.cignarella@unito.it

Cristina Marco
marcocri@amazon.it

Davide Bernardi
dvdbe@amazon.com

[1] Department of Computer Science, University of Turin, Corso Svizzera, 185, 10149 Torino, Piemonte, Italy

[2] Interaction Lab, Heriot-Watt University, The Avenue, Edinburgh EH14 4AS, Scotland

[3] Alexa AI, Amazon, Amazon Development Centre Italy, Via Lugaro 15, 10126 Torino, Piemonte, Italy