Linguistic Frontiers • 8(1) • 2025
DOI: 10.2478/lf-2025-0008

Linguistic **Frontiers**

Constructing a Gold Corpus of Annotated Youtube Comments for Discursive Strategies Span Classification

Original Study

Linda Feld, Friedrich Schiller University Jena, E-mail: lindafeld22@web.de
Dr. Lidiia Wegert-Melnyk, Friedrich Schiller University Jena, E-mail: lidiia.melnyk@gmail.com

Received: 11. 10. 2024; Accepted: 8. 5. 2024

Abstract. Our research project is focused on the creation of a corpus of German YouTube comments relating to the topic of gender diversity. It aims at the identification of dominant arguments for and against gender diversity and the multiple discursive strategies, such as victim-blaming, denial, proposal, etc., used to convey them. This is crucial for understanding the pro and contra communication as well as developing linguistic strategies to refute misconceptions and fake information around the topic. This paper discusses the theoretical background, data collection, corpus creation, and the annotation process as a way of reconstructing discursive strategies identified by works in critical discourse analysis and bringing to light new ones.

Keywords: Critical Discourse Analysis, discursive strategy, Krippendorff's Alpha, Cohen's Kappa

INTRODUCTION

Our project explores the discourse surrounding human identity, particularly concerning sex and gender, which has garnered significant attention in recent years. While traditionally viewed as binary categories, discussions have evolved to acknowledge their fluidity along a spectrum. However, conflicting viewpoints persist, leading to heated debates both online and offline. YouTube serves as a platform where individuals contribute to this discourse through video uploads and comments, offering valuable insights into diverse opinions and arguments.

Our primary goal is to uncover diverse opinions, moods, and arguments on gender diversity topics and examine the linguistic strategies employed. We aim to identify discursive tactics such as proposal, legitimation, and anecdote within a large dataset of 350.000 German YouTube comments on gender diversity. Following the Critical Discourse Analysis (CDA) paradigm, we analyze language patterns to reveal how power dynamics are both perpetuated and challenged in discourse (Fairclough 2003).

In our project, sentiment and stance analysis of the comments revealed that the discourse on gender diversity on YouTube tends to carry a negative sentiment, with most commenters adopting a neutral stance. Surprisingly, comments expressing both positive sentiment and a supportive stance were extremely rare, suggesting dominance by those skeptical or opposed to gender diversity (Melnyk, Feld 2022).

We also focused on identifying, classifying, and analyzing the discursive strategies utilized by commenters. After data cleaning and standardization, we trained different language models using a gold corpus formed from annotated and corrected output generated by LanguageTool. The most effective results were obtained using a German GPT-2 model (Melnyk, Feld 2023).

This paper presents the discourse analytical aspect of our project, starting with a brief introduction to the theoretical framework, followed by detailed presentations of the data and corpus. We then elaborate on the deductively and inductively arrived at discursive strategies and provide an overview of the annotation process and results. The paper concludes with an analysis of

data challenges, proposed solutions, and an outlook on the final project.

CRITICAL DISCOURSE ANALYSIS

Our investigations are rooted in the broader framework of Critical Discourse Analysis, informed by the works of Fairclough (1989, 1992), Reisigl and Wodak (2009), and van Dijk (1992, 2012). From a linguistic perspective, we view discourse as semiotic practices situated within specific contexts and related to particular topics (Reisigl, Wodak 2009). From a sociological standpoint, language use constitutes discourse, functioning as a social practice shaped and influenced by social factors (Fairclough 1992). Discourse serves as both a reflection and a constitutive element of social structures, encompassing norms, conventions, and societal relations (Fairclough 1992).

Understanding or producing discourse requires knowledge on the part of both the speaker and the listener. Discourse serves as a medium for the communication and acquisition of knowledge (van Dijk, 2011, 2012). Following van Dijk's socio-cognitive approach (1993a, 1993b), discourse is underpinned by and contributes to social cognitions, encompassing personal mental representations, shared world knowledge, opinions, attitudes, and ideological systems. Discourse involves mental operations such as interpretation, thinking, argumentation, inferencing, and learning (van Dijk 1993b).

Discourse is intricately intertwined with the dynamics of power in society. Those who have privileged access to discourse wield significant influence over its content and structure, thereby shaping societal norms and values (van Dijk 1993a, 1993b, 2001; Fairclough 1992; Reisigl, Wodak 2009). This control over discourse is instrumental in reinforcing existing power structures, as it enables dominant groups to perpetuate their social and political advantages.

However, the relationship between discourse and power is not static. Resistance to domination frequently emerges within society, leading to ongoing struggles over hegemony within discourses (van Dijk 1992; Fairclough 1992; Macgilchrist 2007). Marginalized and oppressed groups often utilize discourse as a tool for challenging prevailing power dynamics, contesting dominant narratives, and advocating for social change.

While some discourses serve to maintain the status quo and uphold existing power relations, others offer spaces for dissent and alternative perspectives (van Dijk 1992; Fairclough 1992; Macgilchrist 2007). By engaging in discourse analysis, scholars can uncover these power struggles embedded within language, shedding light on the complex interplay between discourse, power, and social change.

Critical Discourse Analysis (CDA) is a research approach that primarily examines how social power dynamics, abuse, dominance, and inequality are manifested, perpetuated, and challenged through language use (van Dijk 2001; Wodak 2001). It focuses on various discourses, including institutional, political, gender, and media discourses, which reveal underlying struggles and conflicts (Wodak 2001). Specifically, CDA involves critically analyzing the content, topics, discursive strategies, and linguistic expressions used in discourse to either reinforce or challenge social power relations (Reisigl, Wodak 2009). In our project, this entails studying the strategies employed by individuals from dominant groups (e.g., those denying gender diversity) and examining how others (e.g., supporters of gender diversity) resist and challenge existing social norms, advocating for progressive change.

While CDA traditionally requires a thorough qualitative analysis of the data, there have been attempts to combine such qualitative work with quantitative methods, allowing the researcher to investigate larger amounts of data (cf. Herring 2004; Herring, Androutsopoulos 2015). However, a computer-mediated analysis requires the researcher to restrict their focus to concepts "operationalizable […] in empirically measurable terms" (Herring 2004, 351), which can then be coded and counted. Since natural language is complex and its comprehension relies on more than what is superficially detectable, it appears natural that certain aspects remain out of the scope of CMD analysis and require qualitative and interpretative work. Moreover, in the rapidly growing field of Natural Language Processing (NLP)*, there* have been several advances in 'qualitatively' analyzing large amounts of textual data, e.g., in terms of the data's sentiment, stance, discourse structure, or arguments. However, for valid results to be produced by language models, labeled data is needed as input, which again requires human annotation, i.e., qualitative work.

DATA

Topic: Gender Diversity

Gender and sex are fundamental aspects of identity, with the male/female binary deeply entrenched in Western society (Schilt, Westbrook 2009). This binary stems from the belief that sex, determined by biology and genital shape, equates to gender. However, this simplistic view overlooks the diversity of human anatomy and experience (Rushton et al. 2019; Matsuno, Budge 2017). While sex is typically understood as biologically ascribed, and gender as socially constructed, recent research suggests that both are influenced by social and cultural factors (Schilt, Westbrook 2009; Rushton et al. 2019).

Contrary to the prevailing binary view, gender and sex are fluid and independent categories. This means that a person's gender may not align with their sex assigned at birth, and there exists a diversity of genders and sexes beyond the traditional binary (Schilt, Westbrook 2009; Rushton et al. 2019). Examples include inter*, non-binary, and trans* individuals, whose identities may not conform to societal expectations of male or female.

In contrast, cisgender individuals identify with the gender assigned to them at birth (Taylor 2010). Cisnormativity places them at the top of the gender hierarchy, granting them societal acceptance without

question (Worthen 2021). However, this normativity marginalizes and stigmatizes those whose genders do not adhere to binary standards, leading to discrimination and violence against non-cisgender individuals (Worthen 2016, 2021).

Numerous surveys conducted by national and international institutions (ADSB 2010; EK 2011; FRA 2015; DIM 2017; FRA 2020) illustrate that individuals whose gender, sex, and sexuality diverge from societal norms (cis and hetero) face systematic discrimination across various domains of public life. They encounter rejection, hatred, and violence both publicly and privately, significantly impacting their health and well-being, often elevating the risk of suicide. While recognized by activists, scholars, and policymakers, recent global developments, particularly the ‚antifeminist backlash', have intensified opposition to efforts aimed at addressing these inequalities. Despite scientific recognition of the fluidity of sex and gender constructs, societal debates persist. Gender and sex are no longer viewed as purely private matters but are central to redefining fundamental values and rights within liberal democracies.

Medium: Youtube

Over the past two decades, YouTube has emerged as the largest and most influential video platform globally (Ceci 2021b), boasting over two billion active monthly users who engage in activities such as rating, sharing, and commenting on videos. The platform witnesses an astonishing influx of content, with more than 500 hours of video uploaded every minute (YouTube). Moreover, beyond YouTube itself, videos are extensively shared, watched, and commented on various social media platforms like Facebook and Twitter. According to Schultes et al. (2013), popular videos can accumulate over 500 comments daily and garner approximately 10.000 ratings throughout their lifespan on YouTube. Notably, videos related to news, politics, science, and technology tend to attract the highest number of comments (Thelwall et al. 2012). This underscores the conversational nature of YouTube, as users utilize the platform not only for video sharing but also for engaging in discussions and interactions (Dynel 2014).

YouTube channels have the autonomy to establish their own commenting policies, allowing uploaders to determine whether comments are permitted, who can comment (e.g., specific users), and whether comments undergo moderation before appearing publicly (Madden et al. 2013). They also have the ability to remove comments and utilize automated flagging systems to filter out inappropriate or spam content. Comments that violate YouTube's community guidelines, such as spam, threats to child safety, harassment, and hateful or abusive content, are subject to automatic deletion by the platform (Ceci 2021a). The platform employs both automated (99%) and manual human flagging procedures (1%) to ensure that comments do not promote hatred, violence, or discrimination. Despite these measures, YouTube comments frequently contain emotionally charged opinions and arguments. Additionally, YouTube maintains anonymity for both creators and commenters, allowing individuals to express opinions without fear of identity exposure or repercussions.

YouTube comments, while sometimes detached from the video's content and lacking coherence, serve various purposes such as self-expression, emotional support, reminiscence, grieving, and advice (Madden et al. 2013). Despite this, they can evolve into topical discussions and facilitate the sharing, negotiation, agreement, and challenging of opinions (Dynel 2014; Bou-Franch et al. 2012). The worldwide accessibility, anonymity, and unrestricted access of YouTube enable commenters from diverse backgrounds to participate in discussions and exchange knowledge, making it a significant venue for public commentary and debate on important issues (Thelwall et al. 2012).

However, YouTube comments have limitations. The visibility of comments depends on factors such as recency and interaction with the video, potentially privileging certain comments over others (Dynel 2014). Moreover, the scope of discussion is constrained by the requirement for users to be registered on YouTube, and comments may lack constructive attributes, fostering overestimation and polarization of opinions (Ziegele 2016). While online discourse has the potential to promote inclusivity and democratization, it is not always realized on platforms like YouTube (Herring, Stoerger 2014).

Furthermore, the unmonitored nature of comments can facilitate the spread of misinformation, posing a significant threat to societal well-being (Del Vicario et al. 2016). Nevertheless, YouTube comments remain a widely used means of communication, with each comment potentially reaching thousands of viewers, making them socially significant spaces for genuine debates and opinion exchanges (Schultes et al. 2013; Thelwall et al. 2012). Despite their drawbacks, YouTube comments offer valuable insights for argument mining and statistical opinion analysis (Li et al. 2018), providing researchers with a rich source of public opinion on contentious social and political issues.

Data Collection and Corpus Creation

Data collection relies on an automated scraping procedure. To avoid recall failure while selecting the videos, we created a list of keywords with neutral, positive, and negative sentiment regarding the topic and generated a list of unique video IDs, where the videos have at least one of the keywords in their title or description. The total number of unique keywords is 49, expanded to 183 by incorporating different spelling variations in line with classic German orthography (see Appendix). While links are case-insensitive, making capitalization unimpactful, the variations included adjustments like "+" and "*", commonly used in user queries and URLs, ensuring compatibility with YouTube's search functionality and greater accessibility for diverse typing styles. The links were then obtained with the help of a Python script.

The keywords included (among others):

Sample of keywords

- transgender
- transsexuell
- genderidentität
- transsexuellengesetz
- selbstbestimmungsgesetz
- detransition
- geschlechtsangleichung
- transition
- nicht-binär
- genderfluid
- genderqueer
- non-binär
- geschlechtsumwandlung
- drittes Geschlecht
- intergeschlechtlichkeit
- intersexuell

A total of 450 links were obtained. Only links to videos with an active comments function were included in the project. All the comments were scrapped with JavaScript code making use of the YouTube API on Google Apps Script. The script targeted every single activity under the video. All the comments to the video as well as replies to the comments were retrieved. Replies to the comments are later on considered as independent comment units. Information obtained through the YouTube API includes the commenter's nickname, the date of comment publication, the text of the comment, the number of likes and replies it gained, and the same set of information for the replies. Out of privacy and data protection reasons, the nicknames of the users will not be disclosed and only the comments' texts, creation dates, and links to the videos will be provided in open access.

We did not impose any time limit on the comments, because we assumed that the topic is rather narrow, and, therefore, it would be possible to scrape comments from as many videos as possible in order to collect a corpus representative of general opinions expressed and discursive strategies applied in YouTube comments regarding gender diversity. However, the commenting activity on YouTube is very dynamic with rapid upward trends. Thus, while the actual date range of the corpus lies between April 2015 and January 2022, statistically significant commenting activity can only be traced back to the second half of 2017. Therefore, the final corpus only includes comments from September 2017 until the beginning of 2022.[1] After removing duplicates and non-German comments, the total number of comments in the final corpus is 35.000.[2]

Relevance and Argumentativeness Detection

The data derived from YouTube comments presents challenges for academic research due to its multifaceted nature. Typically textual, these comments may include emoticons, symbols, or nonstandard spellings to convey emotions or emphasize sentiments (Thelwall et al. 2012). Additionally, users often employ symbols like "@" to direct their comments to specific individuals or "#" to signify meta-tags. The structure and content of YouTube comments are complex. Commenters have the freedom to discuss various topics, sometimes unrelated to the video content or previous comments. Some comments stand alone, conveying self-contained messages, while others respond directly to the video or other comments. In the latter case, relevant information may be omitted, assuming a level of shared understanding within the community. Understanding these nuances is crucial for effective analysis of YouTube comment data in academic research. Researchers must employ nuanced methodologies to capture the diverse nature of communication within the platform.

The subsequent phase in corpus construction entailed the meticulous filtration of extraneous comments. This critical process entails a dual-pronged approach, commencing with the initial determination of comment relevance vis-à-vis the discourse's thematic purview. Comments harboring discernible keywords characteristic of the discourse lexicon, such as ‚Geschlecht', ‚trans,' ‚LGBT,' ‚inter', or ‚non-binary,' or those manifesting implicit or explicit references to the discourse, are subjected to further scrutiny. Subsequently, a qualitative assessment is employed to ascertain whether the comment contributes substantively, albeit minimally, to the discourse topic. Comments germane to the video's intrinsic content, including aspects such as the video creator's identity, production nuances, or contextual background, as well as those expressing evaluative sentiments towards the video or its producer, are deemed irrelevant. Thus, while the first exemplar fails to satisfy the initial criterion for relevance and the second passes the preliminary screening but falters in the subsequent stage, the third example successfully navigates both tiers of scrutiny, thereby warranting consideration for subsequent analytical exploration.[3]

1) Stefanie ist eine tolle Frau. Dazu sehr lieb und möchte nix anderes als geliebt und akzeptiert werden. Warum sind die Menschen so gemein zu ihr? 😔 was soll das?! Zum Glück hat sie eine tolle Mutter!
2) "Milan will ein Mann sein, weil er ein Mann ist" - was für ein schöner Abschlusssatz, was gibts da noch zu sagen :)!
3) Es gibt eine blaue Flagge für homosexuallität Hi ich bin w.13 und bin bisexuell ich identifiziere mich so weil ich Frauen attraktiv finde aber war bis jetzt nur mit Jungs zusammen.

In the subsequent phase, the critical criterion involves determining whether a comment presents an argument. This task, termed "argumentativeness detection" by Peldszus (2017), aligns with his definition of argumentative text as those seeking to persuade the

1 The scraping procedure was carried out in 2022. The corpus will be extended with newer comments for the final article of the project.
2 For more information on the creation of the corpus and for the list of keywords used to obtain the links to the YouTube videos, please see Melnyk & Feld (2022).
3 All of the numbered examples are taken from our dataset, existing of 350.000 comments scraped from YouTube between 2017 and 2022. For the full corpus, visit https://www.kaggle.com/datasets/lidiiamelnyk/youtube-comments-on-gender-diversity.

reader to embrace a particular proposition, characterization, evaluation, or course of action. Thus, a comment qualifies as argumentative if it articulates a discernible position and employs discursive strategies to convey its argument – in essence, if it contains any Argumentative Discourse Units (ADUs) as outlined in section 3.4. While a comment may be relevant in terms of its topic, such as exemplified in (3), it cannot be deemed argumentative if it fails to articulate a position or employ persuasive discursive techniques. The following comment, in contrast, passes the first two stages of relevancy as well as the third stage of being argumentative:

4) Man sollte dankbar dafür sein das man leben darf und dankbar das man ein man oder eine Frau ist. Wenn man sich als mann als Frau identifiziert ist das ja ok aber wieso sollte man dann zwanzigtausend neue Geschlechter erfinden man kann sich doch nicht als etwas fühlen was es nicht gibt. Ich stelle mich doch auch nicht vor meine Klasse und sage ich fühle mich als kampfhelikopter wer mich als man sieht ist transfeindlich.

In our study, we conducted an annotation process on a corpus consisting of 2,193 comments to ascertain the presence of argumentative statements. As anticipated, the dataset exhibited an inherent class imbalance, with only 37 % of comments (819 out of 2,193) containing substantive statements related to the research topic, while the remaining 63 % comprised non-argumentative content or primarily discussed meta-features of the video under consideration. To address this binary classification task of discerning between argumentative and non-argumentative comments, we opted to fine-tune the German DistilBERT model, a state-of-the-art transformer-based language model, on our specific transfer task.

Utilizing the TFAutoModelForSequenceClassification API, we carefully selected hyperparameters to achieve optimal performance. These included a learning rate of 2e-5, a batch size of 64, and training for 5 epochs. We employed the Adam Optimizer to optimize the binary cross-entropy loss function, a standard choice for binary classification tasks. Fine-tuning the model yielded promising results, with training and validation accuracies reaching 85% and 81%, respectively. However, given the resulting accuracy scores, we sought to enhance the model's flexibility by introducing a confidence threshold of 0.7 during inference.

Subsequently, we applied the fine-tuned model to our dataset to filter out non-argumentative comments, resulting in the identification of 1,000 annotated comments deemed argumentative from a total of 4,643 comments in the dataset. This post-processing step was crucial in refining the dataset and improving the reliability of subsequent analyses by mitigating the risk of low recall and ensuring a more focused examination of substantive discourse pertaining to the research topic.

DISCURSIVE STRATEGIES

Related Work

Our approach aligns with Reisigl and Wodak's definition of strategy as intentional practices, including discourse, aimed at specific social, political, psychological, or linguistic goals (2009). In political discourse, Reisigl and Wodak delineate five prevalent discursive strategies: nomination, predication, argumentation, perspectivization, and intensification or mitigation. Similarly, van Dijk identifies six persuasive techniques, including argumentation, rhetorical figures, lexical style, storytelling, structural emphasis, and quoting credible sources (1993b, 1992).

Additionally, van Dijk explores discourse's role in perpetuating dominance through justification, positive self-presentation, negative other-presentation, and denial strategies. These denial strategies encompass disclaimers, reverse charges, transfers, defenses, mitigations, reversals, justification, excuse, and victim-blaming (1992). Strategies aim to legitimize or delegitimize discourses and propositions, categorized by van Leeuwen as authorization, moral evaluation, rationalization, and mythopoesis (2008).

In computational research, Persing and Ng (2016), Zhang et al. (2017), Eger et al. (2017), Peldszus (2017), Stab and Gurevych (2017), and Stab et al. (2018) focus on discourse and argumentation structures and relations. They delve into categories such as question-answer dynamics, (dis-)agreement, support and attack, (counter-) claim, and premise (Persing, Ng 2016; Zhang et al. 2017; Eger et al. 2017; Peldszus 2017; Stab, Gurevych 2017; Stab et al. 2018). Some investigations adopt a semantically oriented approach to discourse acts, as seen in Park et al.'s annotation of various support types for propositions in online comments (2015).

Card et al. (2015), Al-Khatib et al. (2016), and Da San Martino et al.'s (2020) contributions provide valuable insights into classifying discourse units. Card et al. constructed a corpus of news articles to identify media frames, assessing agreement using Krippendorff's α and αU (2015). Al-Khatib et al. categorized Argumentative Discourse Units in news editorials, measuring inter-annotator agreement with Fleiss' κ (2016). Da San Martino et al. examined propaganda techniques in news articles, assessing agreement with Mathet et al.'s γ (2019, 2020). SemEval-2020 Task 11, organized by Da San Martino and colleagues, focused on detecting propaganda techniques in news articles (2020), achieving notable results in span identification and technique classification (Morio et al. 2020; Jurkiewicz et al. 2020).

Discursive Strategies in Youtube Comments on Gender Diversity

The formulation of discursive strategies for annotation involved a hybrid approach combining deductive and inductive elements. Initially, a list of potential strategies was compiled based on existing literature and related studies. Subsequently, manual skimming of comments

was conducted to identify these strategies within the data. This process led to the refinement of the strategy list, resulting in a reduced number of candidates and a more nuanced description of their implementation in the comments. Additionally, novel strategies emerged from this manual inspection that had not been previously defined in the literature. The final set encompasses ten discursive strategies, comprising both general discursive moves independent of the discourse's topic and those specifically associated with gender diversity, yet adaptable to other topics.[4]

1) Evaluation

The author evaluates or judges a claim, a situation, or an issue based on rationality (reasonable vs. pointless), emotions (happy vs. sad), or morals (good vs. bad). Such units often contain the first-person singular pronoun 'ich' and evaluative adjectives or emotional language, often in combination with thinking or feeling verbs expressing the commenter's stance (e.g., finden). In this research, we define emotional language as phrasing that either conveys the emotional state of the author or is intentionally designed to evoke specific emotions in the audience (Wilce 2009).

5) Das ist nicht nur ein Seelen Striptease das ist nich menschenwürdig
6) Geschlechterdifferenzierung macht doch an vielen Stellen auch Sinn.
7) Deswegen finde ich es auch richtig einige Steine in denn weg zu legen in diesem Fall […].

2) Denial

The author expresses doubt, mitigates or denies the existence of diverse genders and/or sexes, the issues related to it, or the seriousness of the topic. Such units often contain negations, verbs or adjectives expressing doubt, disparaging or ironic language.

8) Menschen machen Probleme wo keine sind…
9) Ich spekuliere bei manchen auch auf Wichtigtuerei. […] Da kann man doch nicht ernsthaft glauben daß das bei allen echt ist.
10) das sich plötzlich alle angegriffen und diskriminiert fühlen wollen. Ja Wollen. Das scheint ebenfalls eine neue Mode Erscheinung zu sein.

3) Proposal

The author proposes an action or idea, wishes or demands something. Such units often contain modal verbs, imperatives, the adverb 'bitte', or expressions positively or negatively evaluating the respective proposition (e.g., 'am besten').

11) Dann sollten einfach zwei Geschlechter drinstehn.
12) Lass doch jeden Menschen so sein wie er sich wohl fühlt.
13) Man muss nicht verstehen können wie diese Menschen sich fühlen, sondern einfach nur respektieren, dass sie sich selber am besten kennen.

4) Defense/Blaming

The author reverses the roles of victim and perpetrator, claiming victimhood by inferring negative consequences for themselves resulting from gender diversity and the issues related to it (perceived threat). The author often defends themselves, and/or charges others with guilt. Such units often contain self-references, emotional language, demands, questions, or exaggerations.

14) Man erwartet von mir, dass ich ihn mit ihn*sie anspreche? […] Lasst meine Freiheit in Ruhe.
15) Eine Person mit einem augenscheinlichen Minderwertigkeitskomplex oder Aufmerksamkeitsdefiziet verschwendet die Zeit von Richtern, Psychologen und Ämtern […].
16) Und nicht aller Welt seine Meinung aufdrücken, mit dem Argument, ach es tut doch keinem weh. Nur weil man nicht alles mit macht ist man nicht Respektvoll.

5) Authority

The author presents their proposition as a 'given fact' and verifies it by including a reference to some personal (e.g., experts, witnesses, groups and organizations) or impersonal authority (e.g., laws, state institutions), or by referring to science (biology in particular) and "results or conclusions of quantitative research, studies, empirical data analyses, or similar" (Al-Khatib et al. 2016, 3436). Such segments often contain exact numbers and percentages as well as formal language and scientific terms. Authority might also be established through reference to tradition (e.g., through adverbials like 'traditionell', 'gewöhnlich', 'schon immer'), the general public, or the majority.

17) Das ist vom Gesetzgeber mit der dritten Option auch so anerkannt.
18) Geschlechtsdysphorie ist keine Krankheit und nicht im ICD-10 aufgezählt.
19) Welchem [Geschlecht] wir angehören, bestimmt die Natur.

6) Comparison

The author compares the topic of gender diversity – or any issue related to it – to how it has been dealt with in the past or it is dealt with in other countries (in the present, past, or future) or by drawing on another 'generally accepted' fact. Next to drawing on a comparison for the purpose of legitimating one's proposition, this might serve the goal of diverting attention away from the issue by introducing another topic or of downplaying the issue by contrasting it with other serious topics/problems.

4 The following definitions and selected examples are taken from the final guidelines the annotators were provided with. (see Sections 6.2 and 6.3)

20) In anderen Ländern wird man dafür strafrechtlich leider immer noch verfolgt und teilweise gesellschaftlich geächtet […]
21) es gibt auch in der Geschichte viele Völker, die kein binäres System haben.
22) Frauen laufen doch auch im Anzug mit Fliege oder Krawatte rum […]

7) Anecdote

"The [segment] gives evidence by stating personal experience of the author, an anecdote, a concrete example, an instance, a specific event, or similar." (Al-Khatib et al. 2016, 3436, emphasis added) The commenter may be personally affected by the issue being discussed, or they may reference some close friend or family member as a source of verification. The proposition may be expressed in the form of a question.

23) Ich selbst habe auch 2 Transgender Freunde, deren Weg ich teilweise miterlebt habe […]. Diese beiden Jungs sind einfach so glücklich endlich im "richtigen" Körper zu sein.
24) ich hab es gemacht übers TSG mit 2 Gutachten etc hat 1 Jahr gedauert […]
25) Wenn man non binarys fragt, warum sie sich so fühlen wie sie fühlen begründen sie es damit, Interessen zu haben die nicht ihrem Geschlecht entsprechen.

8) Simplification/Hypothesizing

The author reinforces their proposition by positing simplified causes, reasons, or consequences likely to ensue, often in a binary manner. Frequently, hypothetical scenarios are presented or future outcomes anticipated. This is typically achieved through comparative clauses, modal verbs, and futurate forms, or by framing the assertion as a question.

26) In 30 Jahren wollen Menschen Hunde werden.... Der Weltuntergang naht aber sowas von…
27) [könnte sein], dass Kriminelle darin einen Schlupfwinkel sehen könnten, z.B. ihre Identität zu verändern aus verbrecherischen Gründen […]
28) während man aber ausgelacht werden würde, wenn man behauptet "Ich bin ein Asiate im Körper eines Afrikaners" oder "Ich bin ein 45jähriger im Körper eines 24jährigen" […]
29) Man kann ein Mann sein und sich schminken genauso eine Frau, die dies nicht tut.

9) Relativization/Disclaimer

The author either denies or mitigates the potential offensiveness of the proposition or their own remarks by shifting the blame or responsibility to another entity, such as other individuals or institutions. These segments frequently feature negations or modal adverbs and may be followed by a proposition introduced by a contrasting conjunction like 'aber'.

30) (Das soll nicht beleidigend gegen Transgender sein)
31) diese Sache mit dem dritten Geschlecht macht manche Menschen sogar so wütend, dass Hass auf dieses dritte Geschlecht geschürt wird.
32) Von mir aus sollen sich die Leute fühlen als was sie wollen, aber ernsthaft […].
33) Also ich würde mich schon als tolerant bezeichnen, aber […]

10) Assumption

This subtype is reserved for any "assumption, conclusion, judgment, or opinion of the author, a general observation, possibly false fact, or similar" (ibid.) which cannot be classified as either of the above categories.

34) nicht-binär zu sein, ist keine Krankheit.
35) Es gibt mehr als zwei Geschlechtsidentitäten.
36) non-binary geht es so bei beiden Zuordnungen (auch bei Pronomen, "Mann/ Frau" usw.), weil sie damit etwas verbinden, was sie selbst nicht sind.

Annotation of Discursive Strategies

The annotation of discursive strategies involves two separate steps, the first of which is the identification of a relevant, or argumentative piece of text. The approach of identifying elements in a text without predefined boundaries and locations is called unitizing. "Unitizing is identifying within a medium – within an initially undifferentiated continuum – contiguous sections containing information relevant to a research question." (Krippendorff 2004, 219) This process does not follow grammatical or other formal criteria but requires an in-depth analysis of the semantics and pragmatics of a given text. As Krippendorff puts it,

> […] units should not be considered givens. They emerge in processes of reading and thus implicate the experiences of the analyst as a competent reader. Units are often regarded as a function of the empirical tenacity of what is observed, but it is the act of unitizing that creates them and recognizes them as such. This act crucially depends on the analyst's ability to see meaningful conceptual breaks in the continuity of his or her reading experiences, on the purposes of the chosen research project, and on the demands made by the analytical techniques available to date. (Krippendorff 2004, 98)

The second step in annotating discursive strategies involves assigning specific labels to the identified spans of text, a process commonly referred to as coding. According to Krippendorff (2004, 220), coding entails transcribing, recording, categorizing, or interpreting units of analysis into a data language for comparison and analysis. In our study, annotators assign nominal categories to the identified units, which necessitates familiarity with the discourse topic as well as understanding the meanings of character strings, references, connotations, and contents of expressions (ibid., 110).

Annotation Software

The annotations were conducted using Hexatomic, a versatile and OS-independent platform tailored for deep multi-layer linguistic corpus annotation. Hexatomic aims to enhance research efficiency by providing a unified tool capable of managing various annotation types and corpus formats. This is facilitated through its generic graph-based data model and converter framework for importing and exporting corpus data. Additionally, Hexatomic boasts robust search capabilities, including linguistic structure-based queries. It offers extensibility through plugins, catering to specific annotation needs. As desktop software, it can be downloaded and used offline, catering to researchers working in diverse environments. Implemented in Java as an Eclipse e4 application, Hexatomic offers an intuitive interface, simplifying the annotation process.

In Hexatomic, the annotation process for user corpora involves importing data from PAULA format and opening documents in the Graph Editor. Researchers then filter annotations to focus on constituent trees, select segments of interest, and add new annotations, such as root nodes, using the console. Once annotations are added or modified, projects are saved to preserve changes, ensuring continuity of work. The intuitive interface of Hexatomic streamlines these steps, facilitating an efficient and effective annotation process for linguistic corpora.

Annotators using Hexatomic received comprehensive training on navigating and utilizing the tool for annotation tasks. Detailed instructions and guidelines on Hexatomic usage were provided to ensure consistency and accuracy in the annotation process (see Evaluation metrics). This included guidance on filtering annotations, selecting relevant segments, and adding annotations using the console. Annotators were also equipped to address any challenges or issues encountered during the annotation process.

Annotation Process

For the annotations, guidelines were compiled that contain a brief introduction to the project, detailed definitions of the ten discursive strategies along with multiple examples and notes regarding ambiguous cases, clarifications of competing categories, information on the desired length of the annotated spans, and step-by-step instructions for using the annotation software (Hexatomic). Further, a decision tree was included to simplify the annotation process for each comment. As noted by Krippendorff (2004), such schemes can minimize confusion on the part of the annotator and are especially helpful when annotators have to decide between a large number of alternatives.

The annotation process unfolded in three primary phases, each designed to refine the annotation scheme and ensure consistency among annotators. In the initial phase, annotators familiarized themselves with the annotation tool and tested the scheme on a sample of 100 comments. Subsequent meetings addressed annotations, questions, and issues, leading to adjustments in the guidelines to enhance clarity and exhaustiveness. While some disagreements arose regarding the relevance of material tangential to the discourse, guidelines were reinforced to exclude such ambiguous content from annotation.

The second phase involved annotating seven additional sets of comments, resulting in a total of 1.000 annotated comments over a four-month period. This phase aimed to further solidify the annotation scheme's precision and consistency.

In the third phase, the authors reconciled annotations to create the gold corpus. Spans were aligned based on overlap and category matching among annotators, prioritizing larger overlapping areas or spans with matching categories. Tool-related errors were addressed, and spans solely annotated by one annotator were removed to eliminate disagreements. The resulting dataset was refined to 536 labeled spans, ensuring a higher level of consensus and reliability for subsequent analysis.

Overall, these phases meticulously shaped the annotated dataset, providing a robust foundation for analysis while maintaining consistency and reliability.

Evaluation Metrics

Evaluating interannotator agreement (IAA) involves two key steps: unitizing agreement, which concerns identifying spans, and coding agreement, which pertains to the labels assigned to those spans. Various metrics have been developed to measure each type of agreement. Common techniques for measuring coding agreement include Cohen's kappa and Krippendorff's alpha.

Cohen's kappa is a coefficient used to gauge coding agreement for nominal categories. It measures the proportion of joint judgments in which there is agreement after chance agreement is excluded. Cohen's kappa ranges from +1.00 to -1.0, where a value of 1 represents perfect agreement, 0 equals chance agreement, and values less than 0 indicate that observed agreement is less than expected by chance. Landis and Koch proposed labels and ranges to describe the relative strength of agreement based on kappa scores (Landis, Koch 1977). However, one limitation of Cohen's kappa is its restriction to measuring agreement between only two coders. To address this limitation, Fleiss proposed an adaptation of Cohen's kappa that can measure agreement among more than two coders (Fleiss 1971).

Krippendorff's alpha (α) is a widely-used coefficient for assessing agreement among observers in coding tasks, regardless of the number of observers or the type of data being analyzed (Krippendorff 2004). It accounts for discrepancies arising from unequal utilization of values and is robust even in situations with missing data or small sample sizes. The calculation of alpha involves comparing the observed disagreement (Do) with the expected disagreement (De) when chance alone dictates the coding decisions. By subtracting the expected disagreement from the observed disagreement and dividing by the maximum possible disagreement,

Krippendorff's alpha produces a value that ranges from 0 to 1. A value of 1 indicates perfect agreement, while a value of 0 suggests that the observed agreement is no better than what would be expected by chance.

In addition to alpha, Krippendorff proposed a family of uα-coefficients, which offer a nuanced examination of the reliability of unitized continua (Krippendorff et al. 2016). These coefficients provide insights into various aspects of reliability and can pinpoint sources of inconsistency associated with specific categories or values assigned to units. For instance, cuα focuses on the agreement between values assigned to intersecting units, while (k)uα allows researchers to identify sources of unreliability linked to specific categories or values. This comprehensive approach enables researchers to understand the reliability of their coding schemes in greater detail, facilitating more informed interpretations of their data.Krippendorff et al. provide a Java-based software to compute the different uα-coefficients, allowing researchers to choose the preferred metric, select the level of statistical significance for the confidence limits, and specify the minimum reliability required for their data to be taken seriously (Krippendorff et al. 2016).

In conclusion, evaluating interannotator agreement is crucial for ensuring the reliability and validity of annotated datasets. Various metrics, including Cohen's kappa and Krippendorff's alpha, have been proposed to measure coding agreement between annotators while considering chance agreement. The uα-coefficients provide nuanced insights into the reliability of unitized continua, particularly valuable for datasets with undefined spans of text. Utilizing these metrics empowers researchers to understand annotator agreement comprehensively, ensuring confident interpretation and utilization of annotated data.

Annotation Results

Cohen's Kappa values per category were computed for both the complete dataset and the dataset with spans having less than two annotators removed. The results revealed differences in average Kappa values between the two scenarios. When considering all spans, lower average Kappa values were observed, indicating the potential influence of disagreements on inter-annotator agreement metrics. However, after the removal of spans with insufficient annotator agreement, the Kappa values improved, presenting a more accurate reflection of agreement levels across different categories.

This meticulous data preprocessing approach highlights the significance of refining datasets to ensure a more nuanced and accurate representation of inter-annotator agreement. The subsequent analysis benefits from the removal of instances with minimal agreement, contributing to a more robust evaluation and enhancing the overall reliability of the research findings.

Overall, the trend indicates a positive impact on agreement levels following the data preprocessing. Notably, categories with initially moderate agreement, such as Assumption (ASS) and Denial (DEN), exhibited substantial improvements after removing ambiguous spans,

Category	Cohen's Kappa before spans removal	Cohen's Kappa after spans removal
ASS	0.56	0.68
ANEC	0.79	0.79
AUT	0.73	0.79
COM	0.69	0.76
DEF	0.86	0.89
DEN	0.73	0.83
EVA	0.81	0.87
PROP	0.89	0.93
REL	0.91	0.95
SIM	0.79	0.85

Table 1 Cohen's Kappa scores per category

highlighting the critical role of careful curation in refining the dataset. Conversely, categories with high initial agreement, like Defense/Blaming (DEF) and Relativization/Disclaimer (REL), maintained strong consistency even after the removal, suggesting robust annotator consensus. The findings emphasize the importance of addressing data quality issues to enhance the reliability of inter-annotator agreement assessments in natural language processing tasks.

While Cohen's Kappa scores indicate substantial agreement among annotators, this agreement does not extend to the lengths of the annotated spans. The identified spans exhibit considerable variation in length, ranging from 11 to 35 characters on average. The calculated average overlap between the spans is 74 %. As we proceed with further analysis, considering both span length and the assigned category, it is noteworthy that the task of span identification proved to be challenging, with lower levels of agreement adversely impacting Krippendorff's alpha scores. Out of the 1481 identified spans, 75 % were found to be overlapping, suggesting a relatively high recall rate between annotators. This observation underscores the complexity of the task and highlights the need for a nuanced assessment that incorporates both span length and category assignment for a comprehensive understanding of inter-annotator agreement in natural language processing tasks.

The image reveals varying span lengths across different discourse strategy categories. DEF and DEN show minimal outliers, indicating consistency in span lengths. In contrast, categories like ASS, PROP, SIM, and AUT exhibit greater diversity in span lengths, with numerous outliers suggesting a broader range of textual expressions. REL and EVA have the shortest spans, denoting concise relational and evaluative discourse. DEN follows suit, likely due to the need for swift refutation. Conversely, AUT stands out with longer spans, suggesting more elaborate discussions on autonomy-related topics.

Overall, the analysis highlights how discourse strategies manifest through different span lengths, reflecting the complexity and diversity of communication styles within each category. The diversity of span

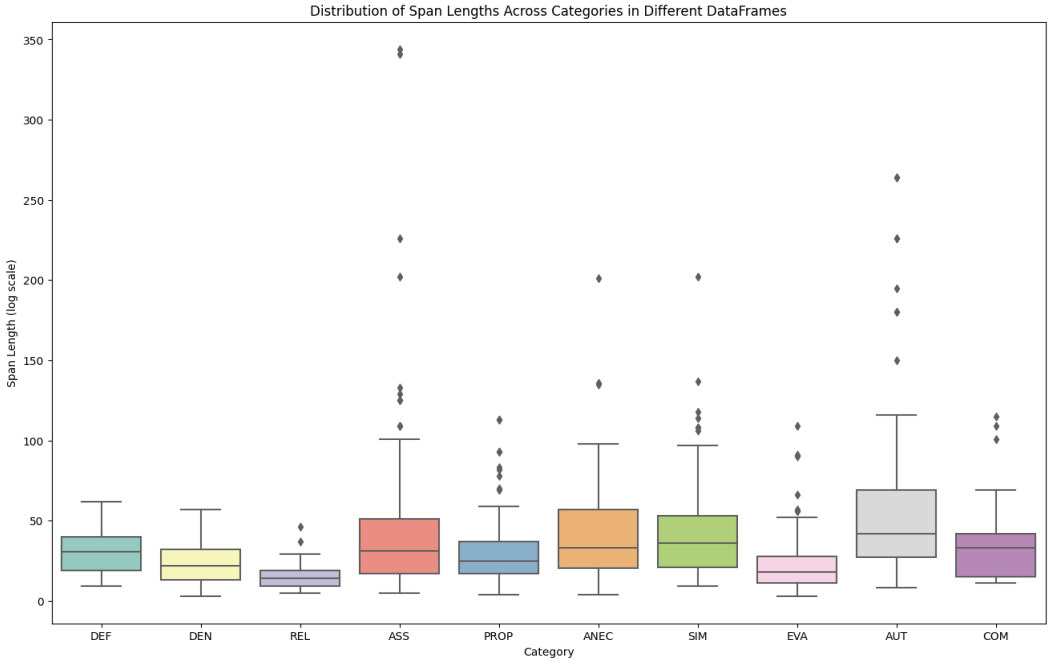

Fig 1 Average span length across categories

Round	ASS	PROP	SIM	EVA	DEN	REL	ANEC	AUT	COM	DEF
1	0.56	0.54	0.99	0.36	0.58	0.50	0.33	0.19		0.04
2	0.66	0.30	0.57	0.14	0.53	0.69	0.64	0.65	0.24	
3	0.56	0.52	0.64	0.53	0.98	0.44	0.36	0.29	0.70	0.75
4	0.63	0.62	0.79	0.52	0.58		0.48	0.68	0.33	0.43
5	0.56	0.42	0.61	0.48	0.44	0.96	0.44	0.67	0.59	0.87
6	0.59	0.51	0.48	0.43	0.56	0.95	0.53	0.28	0.50	
7	0.56	0.41	0.65	0.44	0.77	0.50	0.92	0.82		
8	0.32	0.62	0.44	0.29		0.61		0.19		
9	0.63	0.77	0.68	0.45	0.99		0.99	0.48	0.54	0.19
10		0.6								
Total	0.56	0.53	0.65	0.41	0.68	0.67	0.59	0.47	0.48	0.46

Table 2 Krippendorff's alpha per annotation category and annotation round

length might also affect the interannotator agreement scores, where the categories with shorter spans and less variation might display better annotation results. We can test the hypothesis with Krippensorff's alpha defined above. We can group the categories per Krippendorff's score into:

- categories with $\alpha > 0.6$: SIM, REL, DEN
- categories with $0.5 < \alpha < 0.6$: ASS, PROP, ANEC
- categories with $0.4 < \alpha < 0.5$: EVA, AUT, COM, DEF

The strategies categorized under SIM, REL, and DEN demonstrated the highest level of agreement among annotators, with an alpha coefficient exceeding 0.6, indicating substantial consensus. DEN and REL are also among the categories with shorter and less diverse spans. This suggests a robust understanding and identification of these strategies within the annotated data.

In contrast, the categories of ASS, PROP, and ANEC exhibited moderate agreement, with alpha coefficients falling within the range of 0.5 to 0.6. While not as strong as the aforementioned categories, this level of agreement still indicates a noteworthy degree of consensus among annotators regarding the identification and labeling of these strategies.

The remaining categories, namely EVA, AUT, COM, and DEF, were characterized by fair agreement, with alpha coefficients ranging from 0.4 to 0.5. Although less

consistent than the categories with higher agreement scores, these strategies still garnered a level of consensus among annotators, albeit to a lesser extent. While AUT category has the most diverse and also longest spans, EVA spans are rather short and similar in length, which, therefore, shows that there is no significant connection between the length of spans and the overlapping level of them in the annotations.

Given the intricacy of the task and the comprehensive measures taken to ensure the inclusion of overlapping spans, thereby enhancing recall rates, we have made the decision to include categories with moderate agreement (0.5 < alpha < 0.6) into our gold corpus. This adjustment allows for a more comprehensive representation of the annotated strategies, taking into account the nuanced interpretations and diverse perspectives of the annotators.

In summary, while some categories exhibited stronger agreement than others, the incorporation of categories with moderate agreement enriches the depth and inclusivity of our gold corpus, contributing to a more thorough understanding and analysis of discourse strategies in the given context.

Challenges Of The Dataset

In our strategic approach to address the discrepancies in span definitions within the corpus, we have undertaken a meticulous process to enhance the quality of our labeled data. While we exhibit a high level of confidence in the label annotations, disparities in span definitions have prompted us to implement a series of corrective measures. Having excluded the four categories with the lowest interannotator agreement, we ended up with the dataset comprising of the following examples:

The first step involves the removal of spans lacking overlap with at least one additional span, as defined by the other annotator. Subsequently, when confronted with overlapping spans, we opt for the larger span, establishing it as the definitive representation in our gold corpus. Additionally, we address spans within close proximity (up to 5 characters) that share the same label, amalgamating them to account for potential mechanical errors during the annotation process.

Category	Value counts
ASS	193
PROP	137
SIM	92
DEN	45
ANEC	38
REL	31

Table 3 Value counts per category

However, the resulting corpus poses challenges that could potentially impact the efficacy of machine learning in subsequent research stages. The dataset, comprising 536 labeled spans, may be deemed insufficient for the effective fine-tuning of a transformer model. Furthermore, the imbalance in both span sizes (ranging from 11 to 35 characters on average) and label distribution raises concerns about overfitting for well-represented labels and potential recall issues for underrepresented ones.

To enhance the dataset quality, we propose the generation of synthetic data for the less represented labels. Leveraging the open-source Llama-2-13b-chat-german model5 —a variant of Meta's Llama 2 13b Chat model finetuned on an additional dataset in the German language—we prompt the model to rephrase examples while ensuring the preservation of the original comment's meaning. Notably, Llama-2-13b-chat-german is optimized for German text, exhibiting proficiency in understanding, generating, and interacting with German language content. However, it is essential to acknowledge that the model is not yet fully optimized for the German language due to its training on a small, experimental dataset and limited parameter count.

The Llama-2-13b-chat-german model has been fine-tuned to excel in tasks like factual retrieval, ensuring accurate responses based on contextual information without hallucination. By generating synthetic data from this model, we aim to create a more balanced dataset, particularly for labels with fewer examples. This augmentation will result in each label having 193 examples, contributing to a more robust dataset.

It is important to note that further discussions on this topic, including strategies for span detection and classification, as well as additional details on data augmentation, will be covered in our upcoming paper. This next paper will delve deeper into these methodologies, providing a comprehensive analysis of our approach and its implications for dataset quality and model performance.

OUTLOOK

Using the final annotated dataset, Da San Martino et al. (2019) conducted experiments with various BERT-based models to address two key objectives: predicting whether a sentence contains propaganda techniques and identifying both the spans and the type of propaganda technique present. Their research achieved promising results, particularly with a multi-granularity network on top of BERT using ReLU, which yielded an F1 score of 60.98. Additionally, other studies have explored similar tasks, such as span identification and classification of discursive strategies, employing a combination of models like BERT, GPT-2, RoBERTa, XLNet, and XLM (Morio et al. 2020)

In our paper, we aim to develop a comprehensive methodology to narrow down and qualitatively assess computationally derived opinions on gender diversity. This involves leveraging automated argument mining techniques, which encompass automatic argument detection, argument span segmentation, and identification of relations between arguments. Our specific focus lies in

5 https://huggingface.co/jphme/Llama-2-13b-chat-german

identifying the discursive strategies employed to express various viewpoints on gender diversity in Germany, utilizing argument unit segmentation as a guiding framework.

To achieve our objectives, we propose breaking down our task into two main parts. Firstly, we will explore traditional approaches such as the Spacy-based method for span detection, adapting named entity recognition techniques to identify spans of discourse (Maurya et al. 2022). Additionally, we will fine-tune BERT-based models (Naim et al. 2022) for sequence labeling, treating span detection as a binary classification task where each token is classified as belonging to a span or not.

The second part of our research will involve multi-label classification of the identified spans. To expand our dataset and improve model performance, we plan to utilize generative models to generate additional data. Subsequently, we aim to fine-tune open-source language models such as LLama2 (Zhu et al. 2023) or RoBERTa for text classification tasks, enabling accurate classification of the identified discourse strategies (Jurkiewicz et al. 2020).

Overall, our approach integrates both traditional and state-of-the-art techniques in natural language processing to comprehensively analyze and classify discursive strategies related to gender diversity discourse in Germany. Through this methodology, we aim to contribute to a deeper understanding of the rhetorical patterns and discourse dynamics surrounding this critical social issue.

REFERENCES

Al-Khatib, K., Wachsmuth, H., Kiesel, J. et al., 2016. A News Editorial Corpus for Mining Argumentation Strategies. In Y. Matsumoto & R. Prasad (Eds.), Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. The COLING 2016 Organizing Committee, pp. 3433–3443.

Bou-Franch, P., Lorenzo-Dus, N., & Blitvich, P. G., 2012. Social Interaction in YouTube Text-Based Polylogues: A Study of Coherence. Journal of Computer-Mediated Communication, 17, 501–521.

Card, D., Boydstun, A. E., Gross, J. H. et al., 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In: 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing Conference Proceeding (Short Papers), China: Beijing, pp. 4171–4186.

Ceci, L., 2021a. Distribution of video comments removed from YouTube worldwide Q3 2021, by reason. Edited by Google. Statista, available at: <https://www.statista.com/statistics/1133165/share-removed-youtube-video-comments-worldwide-by-reason/>.

Ceci, L., 2021b. YouTube - Statistics & Facts. Statista, available at: <https://www.statista.com/topics/2019/youtube/#dossierKeyfigures>.

Da San Martino, G., Yu, S., Barrón-Cedeño, A. et al., 2019. Fine-Grained Analysis of Propaganda in News Articles. In: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing Conference Proceeding, China: Hon Kong, pp. 5636–5646.

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H. et al., 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In: 14th International Workshop on Semantic Evaluation Proceeding conference, Barcelona, Spain, pp. 1377–1414.

Del Vicario, M., Bessi, A., Zollo, F. et al., 2016. The spreading of misinformation online. Proceedings of the National Academy of Sciences of the United States of America, 113(3), 554–559.

Demata, M., Heaney, D., & Herring, S. C. (Eds.)., 2018. Language and Discourse in Social Media: New Challenges, New Approaches. Altre Modernità: Università degli Studi di Milano.

Deutsches Institut für Menschenrechte (DIM)., 2017. Gutachten: Geschlechtervielfalt im Recht. Status quo und Entwicklung von Regelungsmodellen zur Anerkennung und zum Schutz von Geschlechtervielfalt. In Bundesministerium für Familie, Frauen, Senioren und Jugend (Eds.), Begleitmaterial zur Interministeriellen Arbeitsgruppe Inter- und Transsexualität – Band 8, available at: <https://www.bmfsfj.de/resource/blob/114066/8a02a557eab695bf7179ff2e92d0ab28/imag-band-8-geschlechtervielfalt-im-rechtdata.pdf>.

Devlin, J., Chang, M.-W., Lee, K. et al., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019, Minneapolis, Minnesota, USA, pp. 4171–4186.

Dynel, M., 2014. Participation framework underlying YouTube interaction. Journal of Pragmatics, 73, 37–52.

Eger, S., Daxenberger, J., & Gurevych, I., 2017. Neural End-to-End Learning for Computational Argumentation Mining. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, pp. 11–22.

European Union Agency for Fundamental Rights (FRA)., 2015. Being Trans in the EU. Comparative analysis of the EU LGBT survey data. Summary. European Union Agency for Fundamental Rights (FRA)., 2020. EU-LGBTI II: A long way to go for LGBTI equality. Luxembourg.

Europäische Kommission, Generaldirektion Justiz (EK)., 2011. Trans- und intersexuelle Menschen. Diskriminierung von trans- und intersexuellen Menschen aufgrund des Geschlechts, der Geschlechtsidentität und des Geschlechtsausdrucks. Amt für amtliche Veröffentlichungen der Europäischen Gemeinschaften, available at: <https://op.europa.eu/o/opportal-service/download-handler?identifier=9b338479-c1b5-4d88-a1f8a248a19466f1&format=pdf&language=de&productionSystem=cellar&part=>

Fairclough, N., 1989. Language and power. Longman Group.

Fairclough, N., 1992. Discourse and Social Change. Polity Press.

Fleiss, J. L., 1971. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

Franzen, J., & Sauer, A., 2010. Benachteiligung von Trans*Personen, insbesondere im Arbeitsleben. Antidiskriminierungsstelle des Bundes, available at: <https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/expertise_benachteiligung_von_trans_personen.pdf?__blob=publicationFile&v=3>.

Herring, S. C., 2004. Computer-Mediated Discourse Analysis: An Approach to Researching Online Behavior. In S. A. Barab, R. Kling, & J. H. Gray (Eds.), Designing for Virtual Communities in the Service of Learning, Cambridge University Press, pp. 338–376.

Herring, S. C., & Androutsopoulos, J., 2015. Computer-mediated discourse 2.0. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), The handbook of discourse analysis (2nd ed.), John Wiley & Sons, pp. 127–151.

Herring, S. C., & Stoerger, S., 2014. Gender and (A) nonymity in Computer-Mediated Communication. In J. Holmes, M. Meyerhoff, & S. Ehrlich (Eds.), Handbook of Language, Gender, and Sexuality (2nd ed.), John Wiley & Sons, pp. 567–586.

Jeong, A. C., 2003. The Sequential Analysis of Group Interaction and Critical Thinking in Online Threaded Discussions. The American Journal of Distance Education, 17(1), 25–43.

Jurkiewicz, D., Borchmann, Ł., Kosmala, I. et al., 2020. ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them. In Proceedings of the 14th International Workshop on Semantic Evaluation, Barcelona, Spain (Online), pp. 1415–1424.

Krippendorff, K., 2004. Content Analysis. An Introduction to Its Methodology (2nd ed.). Sage.

Li, T., Lin, L., Choi, M. et al., 2018. YouTube AV 50K: An Annotated Corpus for Comments in Autonomous Vehicles. In iSAI-NLP 2018 Proceedings, IEEE, pp. 1–5.

Liu, Y., Ott, M., Goyal, N. et al., 2019. RoBERTa: A robustly optimized bert pretraining approach, available at: <https://arxiv.org/pdf/1907.11692.pdf>.

Macgilchrist, F., 2007. Positive Discourse Analysis: Contesting Dominant Discourses by Reframing the Issues. Critical Approaches to Discourse Analysis Across Disciplines, 1(1), 74–94.

Madden, A., Ruthven, I., & McMenemy, D., 2013. A classification scheme for content analyses of YouTube video comments. Journal of Documentation, 69(5), 693–714.

Mathet, Y., Widlöcher, A., & Métivier, J.-P., 2015. The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. Computational Linguistics, 41(3), 437–479.

Maurya, P., Jafari, O., Thatte et al., 2022. Building a comprehensive NER model for Satellite Domain. SN Computer Science, 3(3), 199.

Mochales, R., & Moens, M.-F., 2011. Argumentation mining. Artificial Intelligence and Law,19, 1–22.

Morio, G., Morishita, T., Ozaki, H. et al., 2020. Hitachi at SemEval-2020 Task 11: An Empirical Study of Pre-Trained Transformer Family for Propaganda Detection. In Proceedings of the 14th International Workshop on Semantic Evaluation, Barcelona, Spain (Online), pp. 1739–1748.

Naim, J., Hossain, T., Tasneem, F. et al., 2022. Leveraging fusion of sequence tagging models for toxic spans detection. Neurocomputing, 50, 688–702.

Park, J., Katiyar, A., & Yang, B., 2015. Conditional Random Fields for Identifying Appropriate Types of Support for Propositions in Online User Comments. In C. Cardie (Ed.), Proceedings of the 2nd Workshop on Argumentation Mining. 2nd Workshop on Argumentation Mining, Denver, Association for Computational Linguistics, pp. 39–44.

Peldszus, A., 2017. Automatic recognition of argumentation structure in short monological texts PhD thesis. Institutional Repository of the University of Potsdam, Potsdam, Germany.

Persing, I., & Ng, V., 2016. End-to-End Argumentation Mining in Student Essays. In Proceedings of NAACL-HLT 2016, San Diego, California, pp. 1348–1394.

Reisigl, M., & Wodak, R., 2009. The Discourse-Historical Approach (DHA). In R. Wodak & M. Meyer (Eds.), Methods of Critical Discourse Analysis, Sage, pp. 87–121.

Rushton, A., Gray, L., Canty, J. et al., 2019. Review. Beyond Binary: (Re)Defining "Gender" for 21st Century Disaster Risk Reduction Research, Policy, and Practice. International Journal of Environmental Research and Public Health, 16(3984), 1–14.

Schilt, K., & Westbrook, L., 2009. Doing Gender, Doing Heteronormativity. "Gender Normals," Transgender People, and the Social Maintenance of Heterosexuality. Gender & Society, 23(4), 440–464.

Schultes, P., Dorner, V., & Lehner, F., 2013. Leave a Comment! An In-Depth Analysis of User Comments on YouTube. In R. Alt & B. Franczyk (Eds.), Proceedings of the 11th International Conference on Wirtschaftsinformatik (WI2013), pp. 659–674.

Stab, C., & Gurevych, I., 2017. Parsing Argumentation Structures in Persuasive Essays. Computational Linguistics, 43(3), 619–660.

Stab, C., Miller, T., Schiller, B. et al., 2018. Cross-topic Argument Mining from Heterogeneous Sources. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium.

Taylor, E., 2010. Cisgender Privilege: On the Privileges of Performing Normative Gender. In K. Bornstein & S. B. Bergmann (Eds.), Gender outlaws: The next generation, Seal Press, pp. 268–272.

Thelwall, M., Sud, P., & Vis, F., 2012. Commenting on YouTube Videos: From Guatemalan Rock to El Big Bang. Journal of the American Society for Information

Science and Technology, 63(3), 616–629.
van Dijk, T. A., 1992. Discourse and the denial of racism. Discourse & Society, 3(1), 87–118.
van Dijk, T. A., 1993a. Analyzing Racism Through Discourse Analysis. Some Methodological Reflections. In J. H. Stanfield & R. M. Dennis (Eds.), Race and Ethnicity in Research Methods, Sage, pp. 92–134.
van Dijk, T. A., 1993b. Principles of critical discourse analysis. Discourse & Society, 4(2), 249–283.
van Dijk, T. A., 1995. Discourse, power and access. In C. R. Caldas-Coulthard & M. Coulthard (Eds.), Texts and Practices. Readings in Critical Discourse Analysis, Routledge, pp. 84–104.
van Dijk, T. A., 2001. Critical Discourse Analysis. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), The Handbook of Discourse Analysis, Blackwell, pp. 352–371.
van Dijk, T. A., 2011. Discourse, knowledge, power and politics. Towards critical epistemic discourse analysis. In C. Hart (Ed.), Critical Discourse Studies in Context and Cognition, John Benjamins, pp. 27–63.
van Dijk, T. A., 2012. A note on epistemics and discourse analysis. British Journal of Social Psychology, 51, 478–485.
van Leeuwen, T., 2008. Discourse and Practice. New Tools for Critical Discourse Analysis. Oxford University Press.
Wilce, J. M., 2009. Language and emotion. Cambridge University Press.
Wodak, R., 2001. What CDA is about ± a summary of its history, important concepts and its developments. In R. Wodak & M. Meyer (Eds.), Methods of Critical Discourse Analysis, Sage, pp. 1–13.
Worthen, M. G. F., 2016. Hetero-cis-normativity and the gendering of transphobia. International Journal of Transgenderism, 17(1), 31–57.
Worthen, M. G. F., 2021. Why Can't You Just Pick One? The Stigmatization of Non-binary/ Genderqueer People by Cis and Trans Men and Women: An Empirical Test of Norm-Centered Stigma Theory. Sex Roles, 85, 343–356.
YouTube. (n.d.). YouTube by the Numbers, available at: <https://blog.youtube/press/>.
Zhang, A. X., Culbertson, B., & Paritosh, P., 2017. Characterizing Online Discussion Using Coarse Discourse Sequences. In Proceedings of the International AAAI Conference on Web and Social Media. International AAAI Conference on Web and Social Media, Montréal, pp. 357–366.
Ziegele, M., 2016. Nutzerkommentare als Anschlusskommunikation. Theorie und qualitative Analyse des Diskussionswertes von Online-Nachrichten. Springer VS.
Zhu, X., Cao, J., Tang, D. et al., 2023. Text as Image: Learning Transferable Adapter for Multi-Label Classification. arXiv preprint arXiv:2312.04160.

APPENDIX

Keywords used for finding suitable YouTube videos:
Transgender (transgender)
transgender
Trans-Gender
trans-gender
Transsexuell (transsexual)
transsexuell
Transsexueller
transsexueller
Trans-Sexuell
trans-sexuell
Trans-Sexueller
trans-sexueller
Genderidentität (gender identity)
genderidentität
Gender-Identität
gender-identität
Genderidentitaet
genderidentitaet
Gender-Identitaet
gender-identitaet
Transgender-Identität (transgender identity)
transgender-Identität
Transgender+Identität
transgender+Identität
Transgender-Identitaet
transgender-identitaet
Transgender-Rechte (transgender rights)
transgender-Rechte
Transgender+Rechte
transgender+rechte
Gleichberechtigung+Transgender (equal rights for transgender people)
gleichberechtigung+transgender
Gleichberechtigung-Transgender
gleichberechtigung-transgender
Transsexuellengesetz (law for transsexual people)
transsexuellengesetz
Transsexuellen-Gesetz
transsexuellen-gesetz
Trans-Gesetz
trans-gesetz
TSG
TSG-Gesetz
TSG+Gesetz
TSG Gesetz
tsg-gesetz
Selbstbestimmungsgesetz (self-identification law)
selbstbestimmungsgesetz
Selbstbestimmungs-Gesetz
selbstbestimmungs-gesetz
Transsexuellenrecht (transsexual rights)
transsexuellenrecht
Transsexuellen-Recht
transsexuellen-recht
Reform+Transsexuellengesetz (reform of the transsexual law)

reform+transsexuellengesetz
Reform-Transsexuellengesetz
reform-transsexuellengesetz
Neuregelung+des+Transsexuellengesetzes (new regulation of the transsexual law)
neuregelung+des+transsexuellengesetzes
Neuregelung-des-Transsexuellengesetzes
neuregelung-des-transsexuellengesetzes
Transgesetz (trans law)
transgesetz
TransLaw
translaw
Trans-Gesetz
trans-gesetz
Reform+des+Transgesetzes (reform of the trans law)
reform+des+transgesetzes
Reform-des-Transgesetzes
reform-des-transgesetzes Neuer+Transgesetz (new trans law)
neuer+transgesetz
Neuer-Transgesetz
neuer-transgesetz
Diskussion+zum+Transsexuellengesetz (discussion about the transsexual law)
Debatte+Transsexuellengesetz
debatte+transsexuellengesetz
Debatte-Transsexuellengesetz
debatte-transsexuellengesetz
Transrechte-Gesetzesentwurf (trans rights law draft)
Trans-Recht+Gesetz-Entwurf
trans-recht+gesetz-entwurf
Trans-Recht-Gesetz-Entwurf
trans-recht-gesetz-entwurf
Entwurf+Transgesetz
entwurf+transgesetz
Entwurf-Transgesetz
entwurf-transgesetz
Transgeschlechtlichkeit (transgenderism)
transgeschlechtlichkeit
Trans-Geschlechtlichkeit
trans-geschlechtlichkeit
Transmann (transman)
transmann
Trans*Frau (transwoman)
TransFrau
transFrau
Transfrau
transfrau
Detransition (detransition)
detransition
De-Transition
de-transition
Retransition (retransition)
retransition
Re-Transition
re-transition
Geschlechtsangleichung (gender alignment)
geschlechtsangleichung
Geschlechts-Angleichung
geschlechts-angleichung
Nicht-binär (non-binary)
Nicht+binär
nicht+binär
Nicht-binär
nicht-binär
nicht-binaer
Nicht-binaer
Diversgeschlechtlich (diverse gender)
diversgeschlechtlich
Agender (agender)
agender
Bigender (bigender)
bigender
Genderfluid (genderfluid)
genderfluid
Genderqueer (genderqueer)
genderqueer
Geschlechtsumwandlung (gender transformation)
geschlechtsumwandlung
Geschlechtsumwandlungen (gender transformations)
geschlechtsumwandlungen
GA-OP (gender-affirming surgery)
ga-op
Metaidoioplastik (metaidioplasty)
metaidoioplastik
Transition (transition)
transitionTransMenschen (transpeople)
transmenschen
Trans-Menschen
trans-menschen
Trans-Solidarität (trans* solidarity)
trans-solidarität Trans-Solidaritaet
trans-solidaritaet
Intergeschlechtlichkeit (intersexuality)
intergeschlechtlichkeit
Intergeschlecht (intersex)
intergeschlecht
Geschlechtsvielfalt (gender diversity)
geschlechtsvielfalt
Geschlechts-Vielfalt
geschlechts-vielfalt
Transsexualismus (transsexuality)
transsexualismus
Transgenderformen (transgender forms)
transgenderformen
Transgender-Formen
transgender-formen
Mann-zu-Frau-Transsexuelle (Man-to-Woman transsexual)
mann-zu-frau-transsexuelle
Frau-zu-Mann-Transsexuelle (Woman-to-Man transsexual)
frau-zu-mann-transsexuelle
Heteronormativität (heteronormativity)
heteronormativität
Heteronormativitaet
heteronormativitaet
Geschlechtsinkongruenz (gender incongruity)
geschlechtsinkongruenz
Geschlechts-Inkongruenz

geschlechts-inkongruenz
X-gender (X-gender)
x-gender
Drittes+Geschlecht (third gender)
drittes+geschlecht
Drittes-Geschlecht
drittes-geschlecht
Transgression (transgression)
transgression
Phalloplastik (phalloplasty)
phalloplastik Geschlechtsangleichende+Operation (sex reassignment surgery)
geschlechtsangleichende+operation
Geschlechtsangleichende-Operation
geschlechtsangleichende-operationon