Kerkhof, Anna; Reich, Valentin

Working Paper

Gender Stereotypes in User-Generated Content

CESifo Working Paper, No. 10578

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Kerkhof, Anna; Reich, Valentin (2023) : Gender Stereotypes in User-Generated Content, CESifo Working Paper, No. 10578, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/279329

CESifo WORKING PAPERS

10578
2023

July 2023

Gender Stereotypes in User-Generated Content

Anna Kerkhof, Valentin Reich

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10578*

Gender Stereotypes in User-Generated Content

Abstract

Gender stereotypes pose an important hurdle on the way to gender equality. It is difficult to quantify the problem, though, as stereotypical beliefs are often subconscious or not openly expressed. User-generated content (UGC) opens up novel opportunities to overcome such challenges, as the anonymity of users may eliminate social pressures. This paper leverages over a million anonymous comments from a major German online discussion forum to study the prevalence and development of gender stereotypes over almost a decade. To that end, we develop an innovative and widely applicable text analysis procedure that overcomes conceptual challenges that arise whenever two variables in the training data are correlated, and changes in that correlation in the prediction sample are subject of examination themselves. Here, we apply the procedure to study the correlation between gender (i.e., does a comment discuss women or men) and gender stereotypical topics (e.g., work or family) in our comments, where we interpret a strong correlation as the presence of gender stereotypes. We find that men are indeed discussed relatively more often in the context of stereotypical male topics such as work and money, and that women are discussed relatively more often in the context of stereotypical female topics such as family, home, and physical appearance. While the prevalence of gender stereotypes related to stereotypical male topics diminishes over time, gender stereotypes related to female topics mostly persist.

JEL-Codes: C550, J160.

Keywords: gender bias, gender stereotypes, natural language processing, machine learning, user-generated content, word embeddings.

*Anna Kerkhof**
ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Munich / Germany
a.kerkhof@lmu.de

Valentin Reich
ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Munich / Germany
reichv@ifo.de

*corresponding author

July 18, 2023
We thank Maja Adena, Jean-Victor Alipour, Elliot Ash, Milena Djourelova, Ruben Durante, Florian Englmaier, Thomas Fackler, Oliver Falck, Moritz Goldbeck, Felix Hagemeister, Marina Happ, Ines Helm, Valentin Lindlacher, Johannes Loh, Johannes Münster, Andreas Peichl, Claudia Steinwender, Katharina Werner, Sebastian Wichert, and participants of numerous workshops and seminars for helpful comments and suggestions. Maica Pham, Emil Philipp, Clara Strasser, and Aleksandar Vasic provided excellent research assistance. This project is funded by the Bavarian State Ministry of Science and the Arts in the framework of the bidt Graduate Center for Postdocs. Anna Kerkhof further acknowledges funding from the Joachim Herz Stiftung. All errors are our own.

1. Introduction

Despite advances during the past decades, important hurdles remain on the path to gender equality. In particular, gender stereotypes persist (Bertrand, 2020). Gender stereotypes reflect general expectations about attributes, characteristics, and roles of women and men. E.g., assertiveness and performance are often ascribed to men, while warmth and care for others are attributed to women (e.g., Cuddy et al., 2008; Kite et al., 2008; Fiske, 2010). Recent empirical evidence demonstrates that gender stereotypes affect how we perceive others and how we perceive ourselves (Ellemers, 2018), confining both personal choices and professional careers (Jensen and Oster, 2009; La Ferrara et al., 2012; Kearney and Levine, 2015).

How prevalent are gender stereotypes? It is difficult to address this question, as stereotypical beliefs are not always conscious, and even if they are, they may not be openly expressed (Blackburn, 2017).[1] The growing importance of user-generated content (UGC) opens up novel opportunities to overcome such biases, though. In particular, the anonymity of users in online discussion fora may eliminate social pressures and allow individuals to voice what they think but would otherwise not say (Hsueh et al., 2015; Wu, 2018). At the same time, recent developments in automated text analysis (Gentzkow et al., 2019; Ash and Hansen, 2022) provide the necessary tools to assess gender stereotypes in UGC at large-scale.

This paper leverages a unique dataset of more than a million anonymous comments from a major German online discussion forum to examine the prevalence and development of gender stereotypes over time. To that end, we develop a novel text analysis procedure that overcomes conceptual challenges that arise whenever two variables in the training data are correlated, and changes in that correlation in the prediction sample are subject of examination themselves. Specifically, we assess (i) whether a comment discusses men or women (or no person at all), and (ii) whether a comment covers topics that the literature unanimously classifies as stereotypical male (related to work and money) or stereotypical female (related to family, home, and physical appearance) (Fiske, 2010; Ellemers, 2018; Marjanovic et al., 2022). This allows us to examine if men are mentioned more often than women in the context of male, and women more often than men in the context of female topics at a given point in time. Based on that, we can document whether, where, and to what extent gender stereotypes exist in our data, and how they develop over time. The text analysis procedure is innovative, widely applicable, and we perceive it as a major contribution of our paper.

The topic classification of comments is conceptually challenging. In particular, we wish to assess which topics are being discussed such that the inference is not driven by gender itself. E.g., a classic supervised machine learning (ML) algorithm could learn patterns like "Comment talks about women, thus higher likelihood of topic *family*" from the training data and transfer them to the sample of interest. As a result, we would not be able to detect differences in gender stereotypes between the training and the prediction sample and, crucially, we would not be able to detect changes in gender stereotypes over time. Dictionary methods that use curated lists of words related to specific topics could address this issue. However, classic dictionary methods are prone to yield both false positives and false negatives, and they are sensitive to prefixes,

[1] E.g., social desirability bias – the tendency to provide answers that adhere to social norms – is likely to confound self-reported measures (Podsakoff, 2003; Fisher, 1993).

suffixes, synonyms, and typographical errors.

We propose an innovative solution to these challenges by enriching unbiased dictionaries with the flexibility and "understanding" of *word embeddings* (Mikolov et al., 2013). Word embeddings represent the semantic meaning of words in an n-dimensional space, where the embedding vectors of words with similar meaning are close to each other. We exploit this feature by transforming words associated to specific (gender stereotypical) topics – e.g., work or family – from a dictionary into their word embedding representation.[2] Then, we generate a large number of linear combinations of the word embeddings associated to one specific topic, where the resulting vectors lie somewhere in between the original embeddings. Under the key assumption that word embeddings associated to specific topics are clustered in the vector space, we can use these linear combinations as unbiased training data for a supervised ML algorithm (Support Vector Machine) that is ultimately able to predict if a particular comment covers a specific topic or not.

To apply the trained model to our sample of interest, we must make multi-word comments comparable to word-level embeddings. To this end, we determine each comment's most important words through a *clustered tf-idf* approach. Then, we transform these words into their word embedding representation and compute their linear combination, using their normalized *tf-idf*-values as weights. Each comment is thus ultimately represented by a linear combination of word embeddings that is projected onto the same vector space as our training data, whereby we can apply the trained model for topic classification.

In contrast to the more ambiguous (gender stereotypical) topics, the occurrence of men and women as part of the discussion in our comments is relatively explicit. As a result, we can base our gender classification on a composite of classic dictionary approaches. To minimize the number of false positives, we restrict the procedure to carefully selected gender specific names and terms. To minimize the number of false negatives, we combine three different dictionary approaches that complement each other.

Based on our topic and gender classification, we present strong evidence for the prevalence and persistence of gender stereotypes in our data. In particular, we show that men are relatively more often discussed in the context of stereotypical male topics like work and money, and women are relatively more often discussed in the context of stereotypical female topics like family, home, and physical appearance. Moreover, while gender stereotypes related to work, money, and physical appearance slightly diminish over time, we find no such pattern for domestic issues like family and home. These findings are further supported by regression analyses that control for comment characteristics as well as user and news section fixed effects. The results are robust to excluding offensive language from our data, and they are not driven by potential stereotypes in the news articles that the comments were originally attached to.

Researchers have recently started to distinguish between hostile and benevolent sexism (e.g., Glick and Fiske, 2001, 2018). While both are based on gender stereotypes, hostile sexism conveys a clear antipathy, whereas benevolent sexism is positive in tone but imparts patronizing beliefs about women.[3] To examine whether the gender stereotypes in our data are driven by hostile

[2]Specifically, we use the *Linguistic Inquiry and Word Count Dictionaries* ("LIWC"); see Section 3.2 for details.

[3]E.g., a man's comment to a female colleague on how "cute" she looks, however well-intentioned, may undermine her feelings of being taken seriously as a professional (see Glick and Fiske, 2018).

or benevolent sexism, we first determine their sentiment, and then use standardized sentiment scores as weights for our comments. In line with our analysis of offensive language, we find just small evidence for the existence of benevolent sexism in the context of work, money, and physical appearance, and no evidence for either benevolent or hostile sexism in the context of domestic issues.

Our paper makes two major contributions to the literature. First, we advance the broad and timely research on gender inequality and gender discrimination (e.g., Bertrand and Duflo, 2017). As far as we know, we are the first who leverage the anonymity of UGC to provide a clean and extensive analysis of the prevalence and development of gender stereotypes over almost a decade.

Second, we develop a novel ML-based procedure to classify UGC, where we enrich classic dictionaries with the flexibility and understanding of word embeddings. This procedure can be used more generally for document-level topic classification; potential applications include all types of novel and unconventional text as data such as social media and other online platforms. In terms of application, the procedure is especially useful if relationships between variables in the training data are subject of examination themselves and should thus not be transferred to the sample of interest. E.g., one could further apply the procedure to study changes in racial biases of online users, the association with certain topics such as climate change or gun control with political parties, or the political leaning of news media; see Section 6 for further discussion. To further support research in that direction, our method is available as a Python package on `https://github.com/VFMR/WEELex`.

The remainder of the paper is organized as follows. Section 2 reviews the related literature on gender discrimination and stereotypes, UGC, as well as on recent advances in automated text analysis. Section 3 describes our data and illustrates both the topic and the gender classification in detail. In Section 4, we apply these classifications to our data and illustrate the prevalence and development of gender stereotypes over time. Section 5 provides a battery of robustness checks on our main results. Section 6 concludes.

2. Related literature

Our paper is related to three strands of literature. First, we add to the vast research on gender inequality, in particular to studies on gender discrimination (e.g., Altonji and Blank, 1999; Blau and Kahn, 2017; Charles and Guryan, 2011; Bohnet, 2016; Bertrand and Duflo, 2017) and gender norms and stereotypes (e.g., Akerlof and Kranton, 2000; Bordalo et al., 2016, 2019; Ellemers, 2018; Bertrand, 2020; Ash et al., 2021a,b). Most of this literature considers gender discrimination in specific contexts (e.g., in the workplace) or discusses the prevalence of gender stereotypes at a given point in time. We contribute by examining the prevalence and development of gender stereotypes in UGC over the course of almost a decade, where the anonymity of users allows us to overcome unconscious and social desirability biases that often confound self-reported measures. Moreover, despite the growing importance of online discussions, gender stereotypes in UGC have hardly been studied before.

Most closely related are the papers by Wu (2018) and Marjanovic et al. (2022). Wu (2018) studies the prevalence of gender stereotypes in the "Econ Job Market Rumors" forum and finds that the discourse becomes significantly less academic oriented, and more about personal

information and physical appearance, when users talk about female researchers. Relatedly, Marjanovic et al. (2022) examine gender stereotypes in about ten million comments on male and female politicians from Reddit and show that female politicians are more often described in relation to their body, clothing, and family than males. We extend these analyses in three important ways. First, we analyze an extensive amount of comments on a broad range of topics from a general interest discussion forum, which enhances the external validity of our results compared to the existing studies. In particular, our findings are not limited to gender stereotypes held by a subset of economists, or gender stereotypes related to politicians.[4] Second, both Wu (2018) and Marjanovic et al. (2022) provide static analyses, while our paper examines the prevalence and development of gender stereotypes over time. Third, in contrast to our study, neither of them addresses potential gender biases in the topic classification of UGC.

The second strand of related literature examines UGC (see Luca, 2016, for a survey). The lion's share of this research focuses on the analysis of consumer reviews (e.g., Chevalier and Mayzlin, 2006; Mayzlin et al., 2014; Anderson and Magruder, 2012) or incentives to contribute UGC (e.g., Wang, 2010; Anderson et al., 2013; Easley and Ghosh, 2013; Zhang and Zhu, 2011). While text analysis – especially sentiment analysis – is not new to this literature, UGC has thus far not been tapped to examine the prevalence and, in particular, the development of large societal phenomena such as gender stereotypes. Moreover, the anonymity of users has rarely been considered as a feature, but rather as a problem, e.g. in the context of hate speech (Gagliardone et al., 2015).

Third, we propose a new procedure to classify UGC and thereby add to the growing research on text as data (Grimmer and Stewart, 2013; Gentzkow et al., 2019; Ash and Hansen, 2022). The novelty of our approach is to enrich classic dictionary methods with the flexibility and understanding of word embeddings as developed by Mikolov et al. (2013) and Bojanowski et al. (2017). We thereby contribute to a vibrant literature that incorporates NLP and ML methods to answer economic questions that could not be addressed before (Athey, 2019; Athey and Imbens, 2019). Our paper is especially close to Garg et al. (2018), who use word embeddings to quantify historical trends and social change in gender and ethnic stereotypes. However, while Garg et al. (2018) explicitly allow their word embeddings to capture gender stereotypes, our approach is especially designed to prevent this. In addition, most of the literature on text as data studies English corpora, while analyses involving other languages are rare. We contribute to closing this gap by developing a classification procedure that we apply to German data, but which could principally be used for all languages that feature appropriate (unbiased) dictionaries and pre-trained word embeddings.[5]

Our classification procedure as such is furthermore related to two recent sub-strands of research in text analysis. First, it links to *Correlation Explanation (CorEx) Topic Modelling* (Gallagher et al., 2017), an anchored topic modelling approach using seed words – i.e., a dictionary – to assign documents to topics. This method uses the entire corpus to determine the best fitting topics, though, whereby it is susceptible to issues of gender bias as described above.

[4] The readership of *Spiegel Online* is predominantly male, middle-aged, well educated, and well earning; see `https://app.powerbi.com/` for the most recent readership data collected by the Working Group on Media Analysis (*Arbeitsgemeinschaft Media-Analyse e.V.*, homepage: `https://www.agma-mmc.de/`).

[5] Alternatively, if pre-trained word embeddings do not exist, a sufficient requirement is to leverage a corpus large enough to train one's own embedding vectors.

Second, our approach is similar to *Latent Semantic Scaling* (Watanabe, 2021), which combines a dictionary with word embeddings, too, but is limited to predictions along a single axis (e.g., a sentiment score or political polarity). Likewise, the *Word Embedding Association Test* by Caliskan et al. (2017) employs word embeddings to measure the similarity of words to predefined topics, but also operates on just one dimension. We add to this literature by developing a topic classification procedure that avoids gender bias and is furthermore able to predict multiple topics that are not mutually exclusive.

3. Data and classification

Our analysis of gender stereotypes in UGC features a unique sample of about 7.5 million comments that we classify through a novel text analysis procedure that combines classic dictionary methods, word embeddings, and supervised ML algorithms. This section illustrates the raw data and describes our topic, gender, and sentiment classification procedures in detail.

3.1. Data

Our data comprises 7,345,166 comments that we retrieved from the public *Spiegel Online* (*"SPON"*) discussion forum by the end of 2019.[6] *SPON* attracts around 19 million users per month[7] and ranks among Germany's top five online news websites.[8]

SPON allows its users to comment and discuss its news content. The comments are organized in threads that are attached to *Spiegel Online*'s news articles, but the discussion could also be accessed through a central interface that aggregates all threads. Around 70% of all news articles allow for comments; the remaining 30% typically involve sensitive issues such as migration, terror attacks, and sexual harassment (Dachwitz, 2016).

For each comment, we retrieve information on the user alias (i.e., the nickname of the user who has written the comment), the time and date of upload, position in the thread, and the content of the comment itself. Note that we cannot infer the users' gender from their aliases, and that individual comments usually do not explicitly refer or respond to previous comments from the same thread. Appendix A.1 displays some exemplary comments, Appendix D illustrates one exemplary discussion thread in detail.

Figures 1a to 1f describe our raw data in more detail. Figure 1a depicts the absolute number of comments posted within each of *SPON*'s news sections from Jan 2010 to Dec 2018. Plausibly, the majority of comments is attached to articles on politics or economics, which are *SPON*'s most important news sections. While the absolute number of comments per month is impressive (e.g., 126,990 comments were posted just in March 2011), Figure 1a also reveals that it has been shrinking over time. However, Figure 1b shows that part of the effect can be explained by a diminishing number of articles that allow for discussion on behalf of the users, especially after the 2015 refugee crisis (we observe a total of 782,431 articles/threads). There is ample

[6] Since Jan 2020, users must log in to the forum to read and write comments, which eliminates the anonymity that we wish to exploit for our analysis.

[7] See `https://meedia.de/2017/04/13/agof-welt-rueckt-dank-n24-traffic-an-spon-heran-focus-dank-rekordzahlen-fast-gleichauf-mit-bild/` (Dec 2022).

[8] See IVW, `https://ausweisung.ivw-online.de/index.php?it=1\&setc=1` (Dec 2022).

heterogeneity in the number of comments per thread: while the median (mean) thread features 10 (9.43) comments, the minimum number is equal to 1 and the maximum number equal to 80. Similarly, the comments' average length varies a lot, with a median (mean) length of 336 (467.81), and a maximum length of 23,239 characters (Figure 1c).

Considering the users ($n = 272,023$), we find that the majority of comments is written by a minority of users. E.g., the median user posts just two comments, the mean user 27, and the most active users several thousands (Figure 1d). While some users just post one comment and never come back again, others remain active for considerable time periods. In particular, we find that the minimum amount of time between the first and the last comment is equal to zero for the majority of users, but that there is a long tail of users who remain active for several years (Figure 1e). The users are not too specialized in terms of topics that they contribute to. Specifically, Figure 1f shows that, conditional on writing at least two comments, many of them contribute to discussions related to two or more news sections.

To examine the content of the comments beyond the classification of gender stereotypical topics in further detail, we use BERTopic (Grootendorst, 2022), a state-of-the-art NLP topic modeling technique to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions. To focus on the most important aspects, we restrict the analysis to comments that we eventually classify as discussing men or women (i.e., that we classify as *male* or *female*, respectively).

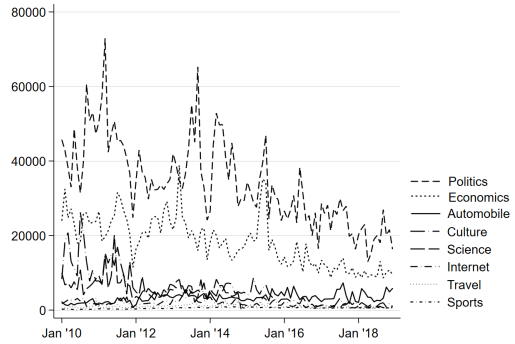
Figure 2a displays the most important terms for the most important topics in comments classified as *male* or *female*, respectively. We find that political issues prevail; in particular, an immense proportion of comments classified as *female* seems to be about Angela Merkel. Exluding such comments from the analysis (Figure 2b) reveals that many comments about women discuss gender related issues such as sexism, feminism, and leadership quotas. We also find that the relevance of the topics varies over time; e.g., Figure 3 shows that the financial crisis in Greece was a major topic in 2015 and that debates on muslims and kurds domineered in 2016, shortly after the infamous terror attacks in Paris. Similarly, we find that debates on sexism gained importance with the #MeToo movement in 2018.

3.2. Topic classification

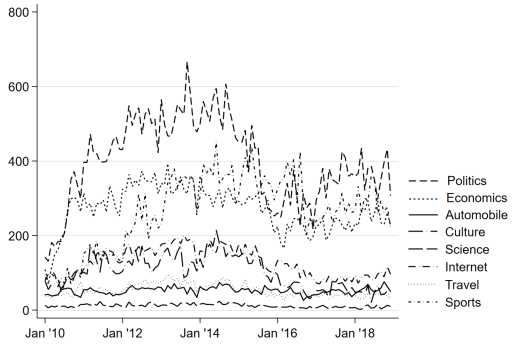
To examine the prevalence and development of gender stereotypes in our data, we must assess whether and which (gender stereotypical) topics are being discussed in the comments, and whether the discussions center on women or men. This section illustrates our automated topic classification procedure, where we propose an innovative combination of dictionary and word embedding approaches to resolve crucial conceptual challenges such as gender bias, false positives, and false negatives. The gender (assessing whether comments discuss women or men), sentiment, and offensive language classifications of comments are specified in Sections 3.3 to 3.5 below.

3.2.1. Conceptual challenges

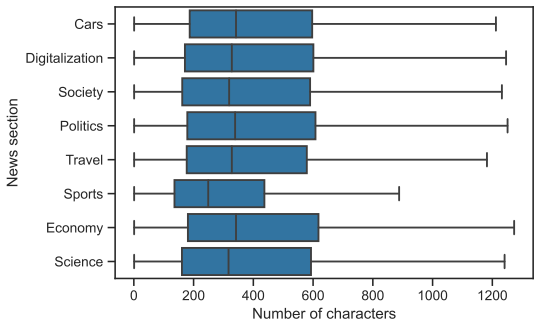
As argued above, the automated topic classification of comments could suffer from three pitfalls: gender bias, false positives, and false negatives. Gender bias is likely to arise in a naive super-



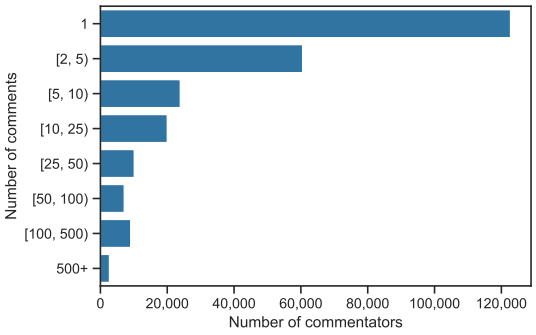
(a) Absolute number of comments per news section



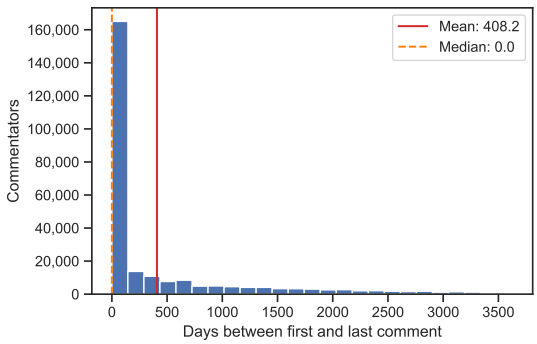
(b) Absolute number of threads per news section



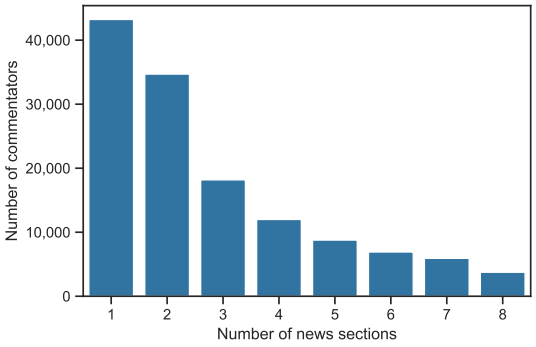
(c) Boxplot number of characters per comment by news section (no outliers).



(d) Distribution of the number of comments per user.

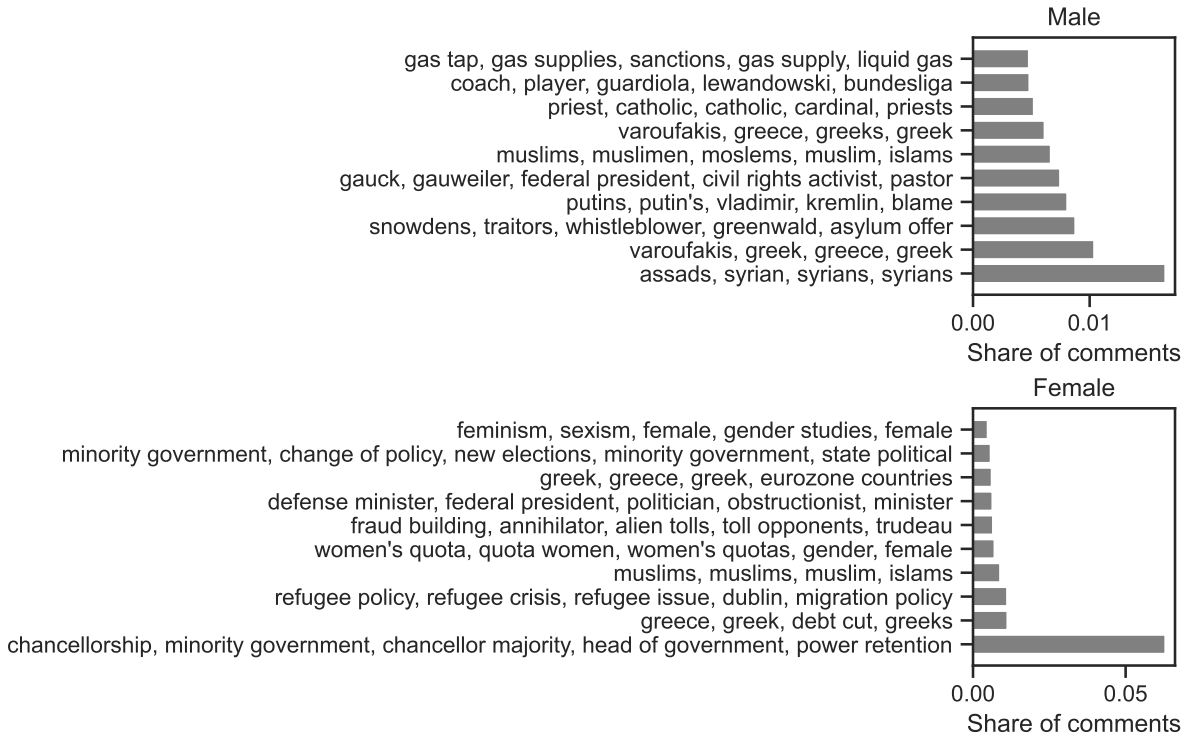


(e) Activity of users.

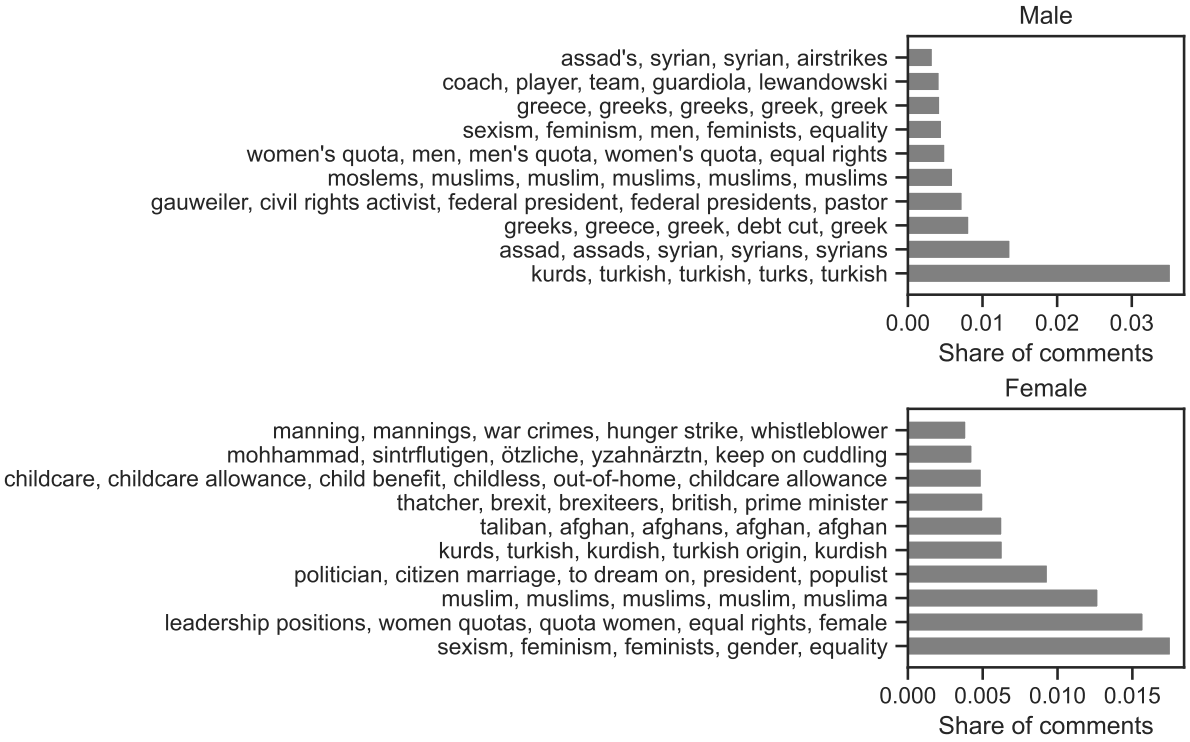


(f) Number of different news sections that a user contributes to, for users with at least two contributions.

Figure 1: Descriptives of the raw data.



(a) Most important terms of the most frequently occurring topics.



(b) Most important terms of the most frequently occurring topics, without comments on Angela Merkel.

Figure 2: BERTopic output.

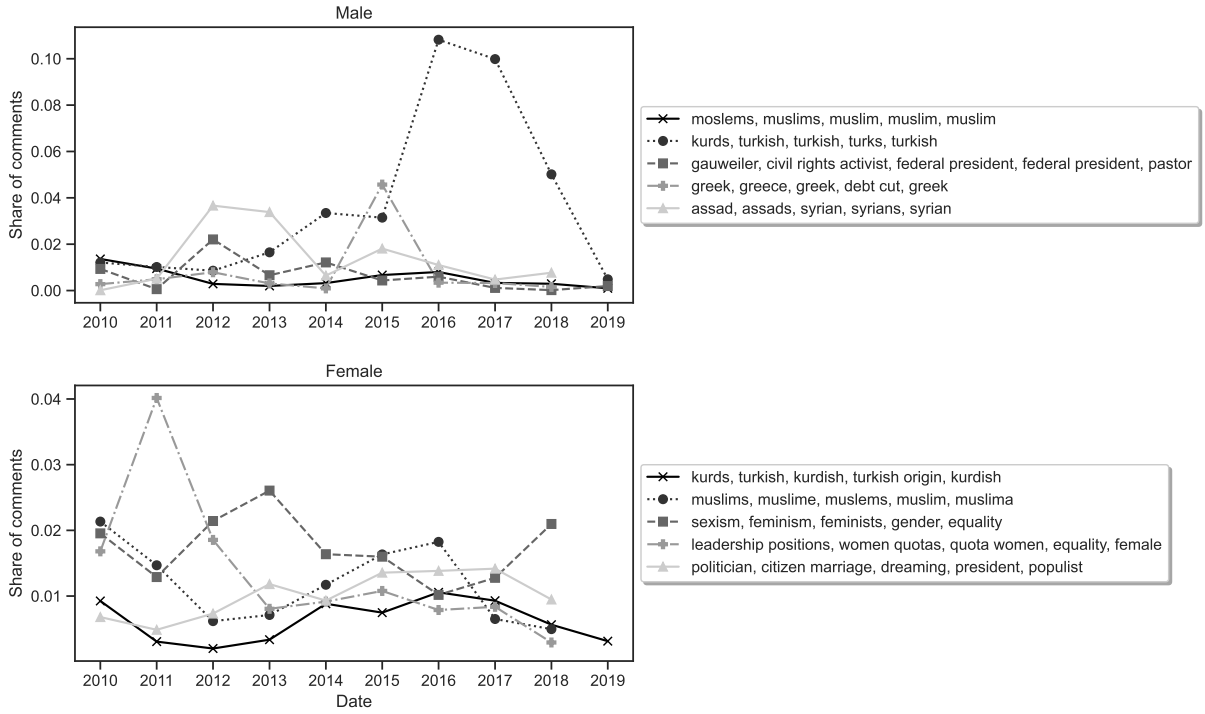


Figure 3: Development of the importance of the most frequently occurring topics over time, without comments on Angela Merkel.

vised ML approach, where human coders manually classify whether comments cover a certain stereotypical topic or not. If this information was used to train a supervised ML algorithm, the algorithm would pick up existing gender stereotypes from the training data and transfer them to the sample of interest. E.g., suppose that comments in the training data discuss women more often than men in the context of *family*. A supervised ML algorithm would pick up this joint pattern and classify comments on women in the prediction sample accordingly. As a result, we would not be able to catch differences in gender stereotypes between comments in the training and in the prediction sample and, crucially, we would not be able to detect changes in gender stereotypes over time. In other words, any topic classification procedure that is driven by the occurrence of gender in a specific comment is likely to yield biased results.

Dictionary methods that use curated lists of words or expressions related to a specific topic – typically put together by linguistic researchers – could solve this issue. Practitioners usually apply such methods by counting the occurrences of dictionary terms in a corpus (e.g., Tetlock, 2007). However, dictionary methods come with two disadvantages. First, they could yield false positives, as it is not trivial how to select or aggregate words in a corpus or a document to capture just the relevant and unambiguous ones. Second, they could yield false negatives, because the selection of words in a dictionary is naturally limited. In addition, dictionary methods are sensitive to prefixes, suffixes, typographical errors, or synonyms, especially when considering morphologically rich languages like German and error-prone online discussions.

In this paper, we propose a solution to these challenges by enriching unbiased dictionary methods with the flexibility and understanding of *word embeddings* (Mikolov et al., 2013). Word embeddings represent the semantic meaning of words by vectors in an n-dimensional space, where

words with a similar meaning are represented by vectors that are close to each other.[9] Under the key assumption that words related to a specific (gender stereotypical) topic are clustered in the embedding vector space, this feature allows us to predict the topic(s) of a comment based on words that are semantically *similar* to those in an unbiased dictionary.[10]

3.2.2. Procedure

Our topic classification procedure comprises two main parts – *training* and *prediction* – which consist of several smaller steps, respectively. Figure 4 provides an overview of the procedure, further details are discussed below.

Part 1: Training

Step 1: Dictionary pre-processing Part 1 of our topic classification procedure is based on *Linguistic Inquiry and Word Count Dictionaries* ("LIWC" henceforth), which provide extensive human-validated lists of words that correspond to certain topics.[11] E.g., the topic *work* includes words like *labor*, *office*, and *politician*, while the topic *family* includes words like *mother*, *brother*, and *childcare*. Following the recent literature from social psychology (e.g., Fiske, 2010; Ellemers, 2018; Marjanovic et al., 2022), we identify six of the topics in LIWC as gender stereotypical: *work* and *money* for men, and *family*, *home*, *body* and *sexual* for women. These topics are also in line with and can be integrated into the more general Stereotype Content Model (SCM, Cuddy et al., 2008), whereby gender stereotypes can be understood along two primary dimensions: warmth and competence. According to the SCM, women are associated with high warmth and low competence, while men are associated with high competence and low warmth. In our application, the topics *work* and *money* correspond to high competence, the topics *body* and *sexual* to low competence, and the topics *family* and *home* to high warmth. Let T denote the set of all, and $T^{gender} \subset T$ the set of gender stereotypical topics in LIWC.

We start by removing all ambiguous words from all topics $t \in T^{gender}$. E.g., the topic *work* features words like *negotiate* and *request*, which could be related to workplace activities but also to other contexts. Thus, we let two Research Assistants independently decide which words unambiguously describe a certain topic and proceed only with those that both of them agreed upon. Further pre-processing steps include capitalization of nouns and replacing words that are designed to find match patterns with words that actually exist.[12]

Step 2: Word embeddings Next, we transform each of the remaining words from each topic $t \in T^{gender}$ into its real-valued 300-dimensional *FastText* word embedding (Bojanowski et al., 2017). Word embeddings are computed via neural network architectures that require huge amounts of data. Therefore, we do not compute the word embeddings ourselves, but rely on externally pre-trained data, as is standard in applied research (e.g., Garg et al., 2018; Kozlowski

[9] See Gentzkow et al. (2019) and Ash and Hansen (2022) for intuitive discussions of word embeddings.
[10] We discuss all assumptions in Appendix C.
[11] Specifically, we use the German adaption *DE-LIWC2015* (Meier et al., 2019) of the English original developed by Pennebaker et al. (2015). For some supportive tasks, we also consider terms from the 2001 version of the LIWC (Wolf et al., 2008), which is based on the English original by Pennebaker et al. (2001).
[12] E.g., we replace *administrati** with *administration* and *analyse** with *analyse*.

Training

Body	Work	Family
breasts	workplace	aunt
body	employment	brother
face	colleague	children
pretty	wage	married

↓

FastText vectorization

↓

Prediction

"I was sitting at my desk when my boss showed up to talk to me about my paycheck."

↓

$\boldsymbol{x} = \text{ctfidf}_1 * \boldsymbol{v}_1 + \text{ctfidf}_2 * \boldsymbol{v}_2 + \text{ctfidf}_3 * \boldsymbol{v}_3$

↓

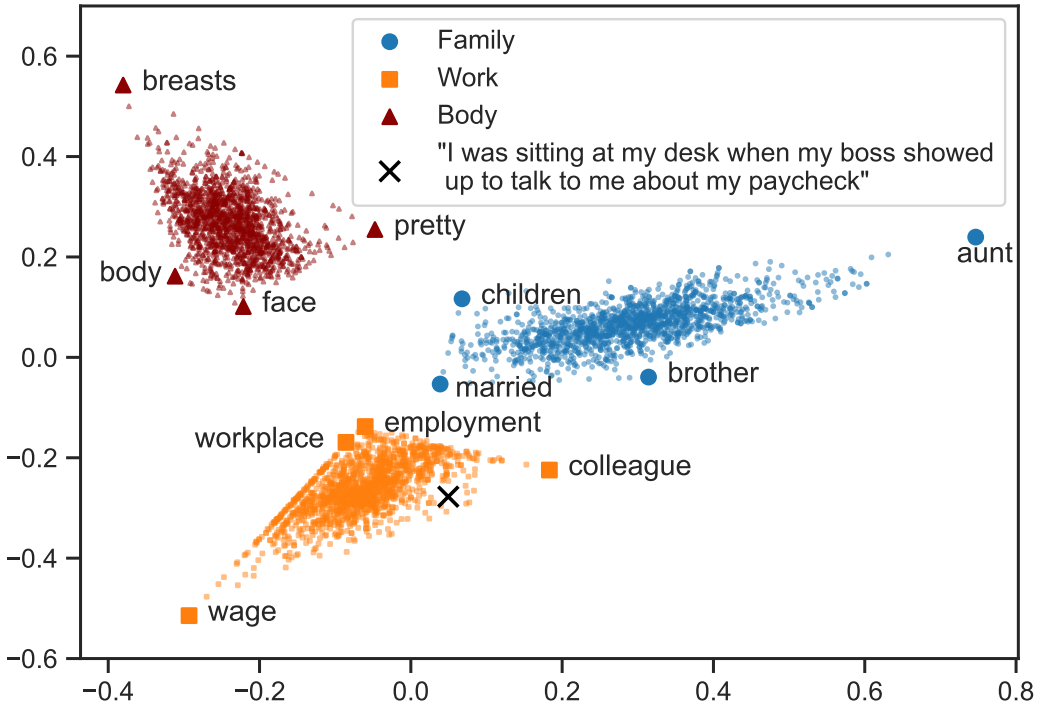


Figure 4: Stylized example of our topic classification procedure

Notes: In this example, we transform words from a dictionary featuring the topics *Body*, *Work*, and *Family* into their *FastText* word embedding representation. (In contrast to our actual model, we apply a *principal component analysis* to reduce the 300-dimensional vector space to just two dimensions; this enables us to visualize the data.) Large circles, squares, and triangles represent the word embeddings. Small circles, squares, and triangles represent linear combinations of the word embeddings, which we use as training data for a Support Vector Machine. Note that the word embeddings in this figure are based on our actual data.
The upper RHS displays a hypothetical comment with work-related content. Note, however, that none of its words appear in our stylized dictionary. We retrieve the comment's most important words with our *clustered tf-ifd* and compute their linear combination, using the normalized *clustered tf-ifd*-scores as weights. The corresponding vector is represented by the dark cross. Under the key assumption that words related to a specific topic are clustered in the embedding vector space, our trained model will predict the topic *work* with high, and the remaining topics with low probability.

et al., 2019).[13] The *FastText* word embeddings are pre-trained by Mikolov et al. (2018), based on Common Crawl's web archive and the entire Wikipedia. This vast amount of training data ensures high quality results; moreover, since the ultimate goal of our procedure is to classify UGC, we perceive this kind of training data as particularly adequate.[14] The main difference between *FastText* and more traditional embedding methods like *Word2Vec* is that *FastText* is trained on n-grams rather than full words. This leads to practically no out-of-vocabulary words during prediction and performs well for morphologically rich languages like German.[15] Hence, *FastText* word embeddings are robust towards common dictionary concerns such as synonyms, compound words, prefixes, suffixes, and typographical errors.[16]

Step 3: Generate training data Based on the *FastText* word embeddings, we generate our training data. To this end, we split the word embeddings into three groups – training, test, and validation – where the training vectors are used as input for a supervised ML model. Crucially, we train a *separate* model for each topic $t \in T^{gender}$. Thus, each model will ultimately be able to predict whether a particular comment covers a particular topic t or not, but the individual topics are not mutually exclusive (e.g., a comment could be classified as being related to *work and* being related to *money*). Specifically, we conduct the following procedure *for each* of the six topics $t \in T^{gender}$:

Denote the focal topic as t^f (e.g., *work*). To generate *one* training observation i:

1. Randomly select one further topic $t^i \in T$, where t^i can be equal to the focal topic t^f or any other topic $t \neq t^f$ in T.[17]

2. Pick three random word embeddings $\boldsymbol{v}^i$ from t^i and three random scalars w^i that add up to one. The linear combination of $\boldsymbol{v}^i$, using w^i as weights is given by

$$\boldsymbol{x}^i = w_1^i \boldsymbol{v}_1^i + w_2^i \boldsymbol{v}_2^i + w_3^i \boldsymbol{v}_3^i, \tag{1}$$

   where $\boldsymbol{x}^i$ is a new vector in the same vector space and with the same dimensionality as the original word embeddings $\boldsymbol{v}^i$. Crucially, any $\boldsymbol{x}^i$ lies somewhere in between $\boldsymbol{v}^i$.

3. Finally, let y^i be a binary target variable, where $y^i = 1$ if $t^i = t^f$ (here: if t^i is equal to *work*), and $y^i = 0$ otherwise.

[13] We use the *gensim* software library (Řehůřek and Sojka, 2010) to load and apply the pre-trained vectors.

[14] Note that the *FastText* word embeddings are likely to outperform any word embeddings that we train ourselves, simply because the training data used by Mikolov et al. (2018) is many times larger than our sample of comments.

[15] E.g., even if the term *Wahlumfrage* (election survey) does not occur in the training data, *FastText* is able to compute the embedding vector as a combination of its vectors for *Wahl* (election) and *Umfrage* (survey) and can thus capture similarities to both of its components.

[16] It has recently been argued that pre-trained word embeddings may be gender biased themselves (e.g., Gonen and Goldberg, 2019). While discarding this is beyond the scope of our paper, we believe that any potential gender bias in the word embeddings is smaller than the bias we would generate if we used a supervised ML approach on our data.

[17] Using the entire set of topics T instead of just T^{gender} enriches our collection of words from different contexts, whereby our algorithm will ultimately be better able to disambiguate them. We further support this approach with a short self-compiled list of words relating to *cars* and *politics*, since these topics play a dominant role in the UGC that we wish to classify.

Steps [1] to [3] are repeated n times such that the y^i are roughly balanced with respect to being equal to 0 or 1.[18] Thus, we ultimately generate n training observations for each category $t \in T^{gender}$, where each training observation i consists of a 300-dimensional vector $\boldsymbol{x}^i$ (which, in turn, is a linear combination of the word embeddings $\boldsymbol{v}^i$) and a binary target variable y^i that indicates if $\boldsymbol{x}^i$ is a linear combination of word embeddings from the focal topic t^f or not.

Step 4: Training of the Support Vector Machine Next, we use the training observations from Step 3 as input for a supervised ML model (one model per topic $t \in T^{gender}$).[19] To this end, we let an ensemble of *Support Vector Machine* algorithms (SVM) use the $(n \times 300)$ input matrix $\boldsymbol{X}_t = (\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, ..., \boldsymbol{x}_t^n)^T$ to predict the vector of binary target variables $\boldsymbol{y}_t = (y_t^1, y_t^2, ..., y_t^n)^T$ for each $t \in T^{gender}$. Specifically, we consider an ensemble of three models for each t, where each of these is a sub-ensemble of SVM algorithms. Each algorithm in each sub-ensemble is trained on a different random draw of input data. We use 5-fold cross validation to tune the hyperparameters of all algorithms such that each algorithm within the sub-ensemble features an identical set of hyperparameters and differs only in the input data drawn at random. Then, we aggregate the three ensembles with the best performing sets of hyperparameters into a final ensemble. The SVM algorithms essentially search for borders that optimally separate observations belonging to t from observations that do not, slicing the vector space into areas that correspond to the individual topics $t \in T^{gender}$. Our key assumption here is that word embeddings from the same topic are clustered within the vector space.

Step 5: Intermediate Evaluation As an intermediate evaluation of our trained model, we come back to the yet unused validation word embeddings (see Step 3). In particular, we use our model to determine the probability with which each of these word embeddings corresponds to each topic $t \in T^{gender}$. Then, we compare our prediction with the actual topics that the validation word embeddings correspond to.[20]

Table 1 displays four of the most frequently used evaluation metrics for binary classification. All of these metrics for all topics $t \in T^{gender}$ are close to 1, thus demonstrating that our trained model performs extremely well. Note, however, that the results in Table 1 are not (yet) informative about the topic classification of UGC, which we conduct in Part 2 of our procedure.

Part 2: Prediction

Step 1: Collapse comments by clustered tf-idf Before we can apply the trained model to our sample of interest, we must make the multi-word comments comparable to word-level embedding vectors. To this end, we use *tf-idf* (term frequency / inverse document frequency) to identify the most relevant words per comment. Specifically, since regular *tf-idf* ignores semantic similarity of words (which would reduce the flexibility and understanding that we gained through the word embeddings), we develop a *clustered tf-idf* approach, where words of similar meaning are considered together.

[18] n is some multiple of the number of word embeddings in t^i. This topic specific multiplier value is found via hyperparameter tuning with 5-fold cross validation.
[19] Model training and prediction were executed with the Python library *scikit-learn* (Pedregosa et al., 2011).
[20] Recall that the actual topics of the validation word embeddings are known.

Table 1: Validation of the SVM ensemble

	Accuracy	Precision	Recall	F1
Work	0.947	0.815	0.791	0.803
Money	0.935	0.897	0.821	0.857
Family	0.988	0.885	0.885	0.885
Home	0.988	0.810	0.895	0.850
Body	0.946	0.899	0.860	0.879
Sexual	0.968	0.830	0.830	0.830

Notes: Prediction metrics for the validation word embeddings. *Accuracy* is the proportion of correct predictions. *Precision* is the proportion of correct positives. *Recall* measures the proportion of positives captured by the positive predictions. The f_1-*score* is the harmonic mean of precision and recall.

The *clustered tf-idf* comprises three steps. We start by computing regular *tf-idf* weights for all words in our corpus. Then, we use an unsupervised ML algorithm to cluster the words' *FastText* embedding vectors.[21] The algorithm is tuned to identify many clusters with few word embeddings, respectively, which assures that the embeddings within a cluster are semantically close to each other. We then aggregate the regular *tf-idf* weights of all words that correspond to the embedding vectors within one cluster to a *clustered tf-idf* weight, which, in turn, is assigned to all words within that cluster. If a cluster comprises just one word embedding, the *clustered tf-idf* corresponds to the regular *tf-idf* weight of the corresponding word.[22]

Based on the *clustered tf-idf* weights, we identify the three most relevant clustered word embeddings per comment.[23] Analogous to the words from LIWC, we transfer these nouns into their 300-dimensional *FastText* word embeddings. Then, we compute their linear combination, using their normalized *clustered tf-idf* weights such that they add up to one. Thus, each comment is ultimately represented by a linear combination of word embeddings that is projected onto the same vector space as the training data, whereby we can apply the trained model from Part 1.

Step 2: Predict topics Finally, we use our trained model to predict the probability with which each of the collapsed comments discusses a specific gender stereotypical topic $t \in T^{gender}$. In particular, we classify a comment as discussing topic t if $\Pr(t) > 0.5$.[24] Note that the topics are *not* mutually exclusive; e.g., a comment could be classified as discussing both *work* and *money*. See Appendix A.1 for three examples of our topic classification.

3.2.3. Validation

We pursue two approaches to validate our automated topic classification. First, we consider the pairwise correlation between the predicted topics and the news outlet section that the comments were originally attached to (Figure 5). Plausibly, comments that we classify as covering the topics *work* and *money* are most strongly correlated to the news outlet's economy section, whereas

[21] More specifically, we use agglomerative hierarchical clustering (Murtagh and Contreras, 2012).
[22] See Appendix C for further details.
[23] We identify nouns with the part-of-speech tagging capabilities of the *spacy* software library in Python. We focus on nouns, because they are less ambiguous in terms of their topic correspondence than adjectives or verbs.
[24] Section 5 shows that our results do not hinge on this binary classification.

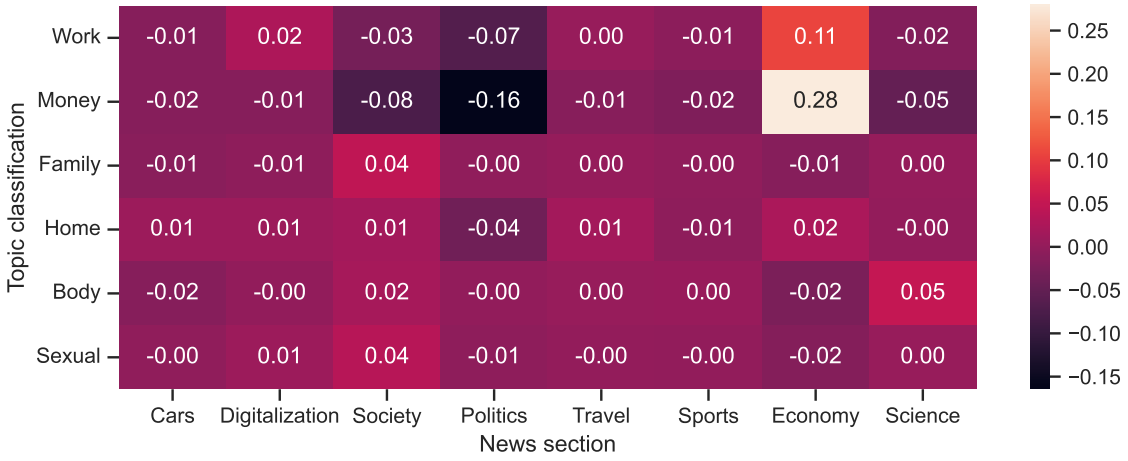


Figure 5: Correlation between news sections and topic classification

Notes: Values in cells are Pearson correlation coefficients between the binary classification predictions of comments and the binary indicator for the newspaper section the comment is located in. Brighter shadings indicate a larger positive correlation.

comments that cover *family* and *sexual* appear most frequently in the news outlet's society, and comments on *body* in the science section.

Second, we apply our automated topic classification to chunks of text whose content is known. In particular, we screen the Wikipedia category tree[25] for the categories that best match the six gender stereotypical topics that we consider (Table A.1 provides an overview).[26] E.g., Wikipedia articles from the category "working environment" are very likely to cover the topic *work*; hence, our topic classification procedure should classify those articles accordingly.

Figure 6 shows that our procedure performs extremely well. In particular, we find that Wikipedia articles from categories that correspond to a certain topic $t \in T^{gender}$ are classified accordingly, while the average prediction probabilities for unrelated topics are low. E.g., our model predicts that Wikipedia articles from the categories "finance", "means of payment", and "money transfers" cover the topic *money* with a probability of up to 74%, and the remaining topics with a probability close to 0%.

3.3. Gender classification

In contrast to the more ambiguous (gender stereotypical) topics, the occurrence of men and women as part of the discussion in our comments is relatively explicit. E.g., if a comment mentions "Harry" or "Mr. Smith", it is clear that a man is being discussed. As a result, we can base the gender classification of our comments on a composite of simple dictionary approaches. To minimize the number of false positives, we restrict the procedure to few gender specific names and terms that are unambiguous in this context as well as to celebrities whose gender is publicly known. To minimize the number of false negatives, we combine three different dictionary approaches whose results complement each other.

[25] See `https://en.wikipedia.org/wiki/Special:CategoryTree` (May 2022).

[26] We use the Wikipedia API to retrieve the first paragraph of all articles that belong to the selected categories. From this list, we remove articles about individuals, interest groups, and redirects. Then, we apply our algorithm to classify each of the collected paragraphs.

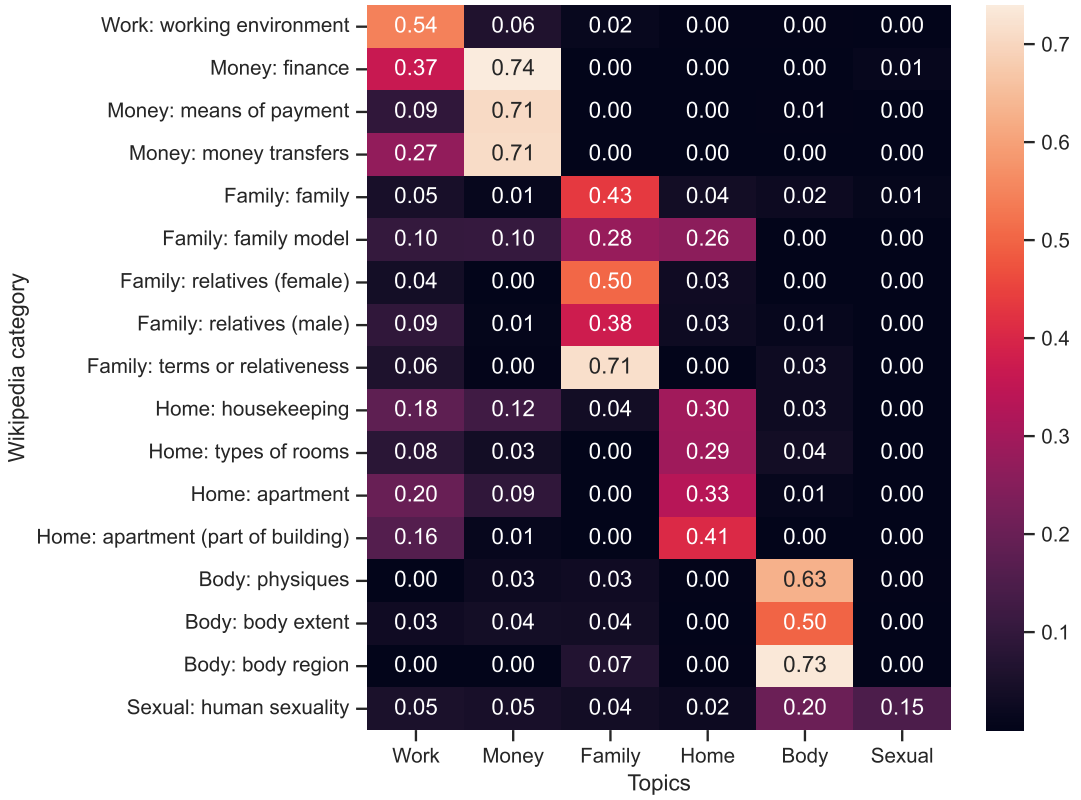


Figure 6: Topic classification of Wikipedia articles from known categories

Notes: The figure depicts the average predicted probabilities that articles from a specific Wikipedia category cover each of our gender stereotypical topics $t \in T^{gender}$. Brighter shadings indicate larger probabilities.

3.3.1. Procedure

The gender classification of our comments is based on three dictionary approaches:

First names We start by retrieving a list of the 100 most popular German male and female first names from *The Society for the German Language*'s website and remove ambiguous names such as *Ernst* ("serious").[27] Then, we search each comment for the occurrence of one or several male or female first names. If a comment features at least one female first name, it is classified as *female*, if it features at least one male first name, it is classified as *male*, and if it features no popular first name at all, it is classified as *none*. Note that a comment could be classified as both *male and female* at this stage; ties are resolved when we compile the results from all three approaches.

Gender specific terms Second, we search the comments for unambiguous gender specific terms like *lady* or *gentleman*.[28] Analogous to the above procedure, we classify a comment as *female* (*male*) if it contains at least one of these terms, and as *none* otherwise.

Celebrities Third, we let a Research Assistant read several thousand comments and compile a list of all celebrities that she came across (e.g., Donald Trump, Angela Merkel, Beyoncé). Based on this list, we searched all comments for the occurrence of celebrities whose gender is publicly known.[29] As above, we classify a comment as *female* (*male*) if it features at least one female (male) celebrity, and as *none* otherwise.

Composition We compile the results from our dictionary approaches in three steps. We first consider consonant classifications. In particular, we ultimately classify a comment as *female* (*male*) if at least one of the dictionary approaches classifies the comment accordingly, and the other approaches either agree or classify the comment as *none*.

In a second step, we resolve conflicting classifications *across* our dictionary approaches (i.e., if one approach classifies a comment as *male* and another approach as *female*). Since gender specific terms are less ambiguous than first names, and celebrities are less ambiguous than gender specific terms, our third approach overrules the second one, and the second approach overrules the first. Section 5 demonstrates that our results are robust to alternative composition rules, such as the first names or the gender specific terms overruling the other approaches.

Finally, we resolve ties *within* our composite classification (i.e., if a comment is classified as both *male and female* after resolving conflicting classifications across the approaches). Specifically, if we find that a comment discusses both men *and* women, we set its classification to *none*. Section 5 shows that our results are robust to classifying such cases according to the majority of male/female first names, gender specific terms, and celebrity occurrences (i.e., classify a com-

[27] *Gesellschaft für deutsche Sprache e.V.*, see `https://gfds.de/vornamen/beliebteste-vornamen/#` (Nov 2022) for further details.

[28] In particular, we search for the gender specific male terms *herr, mann, männ* and the gender specific female terms *frau, dame, weib, mädchen, fräulein*.

[29] The procedure yields a total of 1,491 male and of 511 female celebrities. When we search the comments , we take different spellings and spelling mistakes of the celebrities into account.

Table 2: Most predictive words for *female* and *male* comments

(1) *female*	(2) *male*
wife	money
family	chancellor
society	war
child	party
life	politics
mother	politician
victim	president
party	law
quota	government
law	people

Notes: This table displays the ten most predictive words for comments classified as *male* or *female*. The words are obtained via two separate Lasso-Logistic propensity score models based on a random sample of 9,000 comments, where 3,000 are classified as *male*, 3,000 are classified as *female*, and 3,000 are classified as *none*.

ment as *female* if there are more female than male classifiers and vice versa). See Appendix A.1 for three examples of our gender classification.

3.3.2. Validation

As argued above, the explicit discussion of men and women in our comments joint with the careful selection of terms for our dictionary approaches curtails the risk of generating false positives and negatives. To validate the performance of our gender classification procedure nonetheless, we use a Lasso-Logistic propensity score model and examine the words that are most predictive for comments classified as *male* or *female*. Specifically, we draw a random sample of 3,000 *male, female,* and *none* comments, respectively, and use the trained *tf-ifd* from Section 3.2 to vectorize the comments.[30] Then, we run two separate Lasso-Logistic regressions, where we use the *male* classifier as dependent variable in the first, and the *female* classifier as dependent variable in the second regression.

Table 2 displays the ten most predictive terms for comments classified as *female* (column 1) and *male* (column 2), respectively. The results are compelling: while words such as *wife*, *mother*, *family*, and *child* are most predictive for comments classified as *female*, words like *money*, *war*, and *president* are most predictive for comments classified as *male*. This does not only validate our gender classification, but also prefigures our main results on gender stereotypes that we present in Section 4.

3.4. Sentiment

To examine if gender stereotypes are driven by hostile or benevolent sexism (Glick and Fiske, 2001, 2018), we also determine the sentiment of our comments. To this end, we apply Latent Semantic Scaling (LSX, Watanabe, 2021) to compute a sentiment score for each comment. Similar to our topic classification, LSX adopts a polarity lexicon, where seed words are assigned

[30] As in our main specification (see Section 4), we exclude all comments about Angela Merkel from the analysis.

Table 3: Evaluation metrics offensive comments

Accuracy	Precision	Recall	F1
0.80	0.74	0.60	0.66

to a positive or negative class. These seed words are then transferred to their word embedding representation, and the polarity of other words can be inferred from the similarity of their word embeddings to the embeddings from the dictionary.

To apply LSX to our analysis, we use the SentiWS sentiment dictionary (Remus et al., 2010), which provides an extensive list of German words along with a polarity score ranging from -1 to 1. We restrict the analysis to words with an absolute score above 0.5, which gives us about 120 words, and use *FastText* to transfer these words into their word embeddings. Then, we compute the similarity of all nouns, verbs, adjectives, and adverbs in each of our comments with the word embeddings from the dictionary, weight the words with the corresponding polarity scores, and use the *clustered tf-idf* method from above to compute an aggregate sentiment score for each of our comments. Finally, we standardize the comment-level sentiment scores such that comments that are more negative than the average feature a negative, and comments that are more positive than the average feature a positive score.

3.5. Offensive language

Since we are mainly interested in subtle forms of gender stereotypes, we identify all comments that use offensive language and separate them from more common speech in our subsequent analyses. To this end, we employ a multilingual *BERT* model (Devlin et al., 2018), i.e., a large pre-trained language model that we fine-tune for the supervised prediction of offensive language in our comments. To this end, we use German Tweets from Wiegand et al. (2018) and Struß et al. (2019), which come with a crowdsourced indicator for offensive language as training data.[31] Then, we apply the trained model to our sample of comments to predict the probability with which each comment features offensive language. Analogous to our topic classification, we classify a comment as offensive if $\Pr(offensive) > 0.5$.

While we cannot validate the performance this model on our comments, we can validate how well it performs on a sample of held out validation Tweets and assume that the Tweets and the offensive language in them are sufficiently similar to our comments. Table 3 shows that the model produces relatively few false positives but does miss out on some offensive posts.

4. Results

This section presents the results from applying the topic, gender, sentiment, and offensive language classification to our sample of interest. We start by providing (static) descriptive evidence, then we present the results on the prevalence and development of gender stereotypes over time.

[31] The website for this labelling task defines offensive language as "hurtful, derogatory or obscene comments made by one person to another person" (https://fz.h-da.de/iggsa).

4.1. Descriptives

Gender classification Since our main analysis is based on comments that are classified as either *male* or *female*, we start by considering the results of our gender classification. From our initial sample of 7,345,166 comments, 1,375,252 are classified as discussing either women or men. From these, we exclude all comments that mention Angela Merkel, as she is likely to be an outlier in terms of the subtle and unconscious gender stereotypes that we wish to examine (see also Figure 2).[32] This reduces the number of comments for our main analysis to 1,162,735, where 200,261 comments (17.22%) are classified as *female*, and 962,474 (82.78%) are classified as *male*.

Topic classification Based on the 1,162,735 comments from above, Table 4 summarizes the results of our topic classification. The topics that appear most frequently in our main sample are *work* and *money*, i.e., those that we perceive as stereotypical male. In contrast to that, topics that we identify as stereotypical female – *family*, *home*, *body*, and *sexual* – appear relatively seldom. To further check the validity of our main results, we also introduce two placebo topics – *time* and *space* – which are arguably unrelated to gender. Hence, when we compare how often men and women are mentioned in the context of *time* and *space*, we should not be able to observe any differences between these groups.

Some of our gender stereotypical topics are conceptually similar (e.g., *family* and *home*). To take this into account – and to present the prevalence and development of gender stereotypes as concisely as possible – we pool comments that are classified as *work* or *money* (or both) as *professional*. Analogously, we pool *family* and *home* as *domestic*, *body* and *sexual* as *physical*, and *time* and *space* as *placebo*. Section 5 shows that our results are qualitatively similar when we consider each of those topics individually.

Figure 7 displays the proportion of comments classified as *professional*, *domestic*, *physical*, and *placebo* for *male* and *female* comments, respectively. While the proportion of *female* comments classified as *professional* is smaller than for *male* comments, it is considerably larger for comments classified as *domestic* and *physical*, which strongly suggests that gender stereotypes exist in our data. The difference between *male* and *female* comments for *placebo*, in contrast, is negligible.

Sentiment Figure 8 shows the results of our sentiment classification. The left panel depicts the average sentiment score for comments classified as *male* or *female* in each month of our observation period. We find that both types of comments have negative sentiment scores on average, where comments classified as *male* are usually more negative than comments classified as *female*. The development of sentiment is mostly parallel for *male* and *female* comments: the average sentiment scores increase until about Jan 2014, then decline steadily with a particularly sharp drop for *female* comments by the end of 2017.

To facilitate the interpretation of our sentiment score, we standardize the values to have a mean of zero and a standard deviation of one. The right panel in Figure 8 shows the results:

[32] We use a simple dictionary approach to identify referrals to Angela Merkel. Section 5 provides a robustness check, where we keep comments on her in our sample.

Table 4: Topic classification

Topic	No. comments	Share
Original		
work	113,334	9.75%
money	157,618	13.56%
family	19,145	1.65%
home	15,141	1.30%
body	67,965	5.85%
sexual	5,096	0.44%
time	2,916	0.25 %
space	18,180	1.56%
Pooled		
professional	255,690	21.99%
domestic	33,932	2.92%
physical	72,714	6.25%
placebo	21,087	1.81%

Notes: Results of our topic classification. Note that the topic classification is not mutually exclusive, i.e., a comment could be classified as covering zero, one, or several topics.

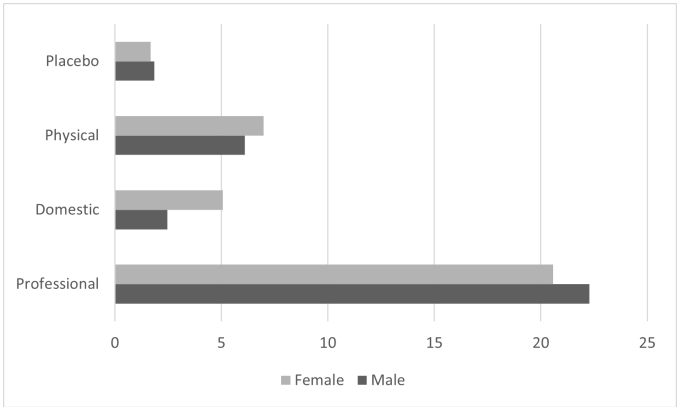


Figure 7: Pooled topic classification by gender (in %).

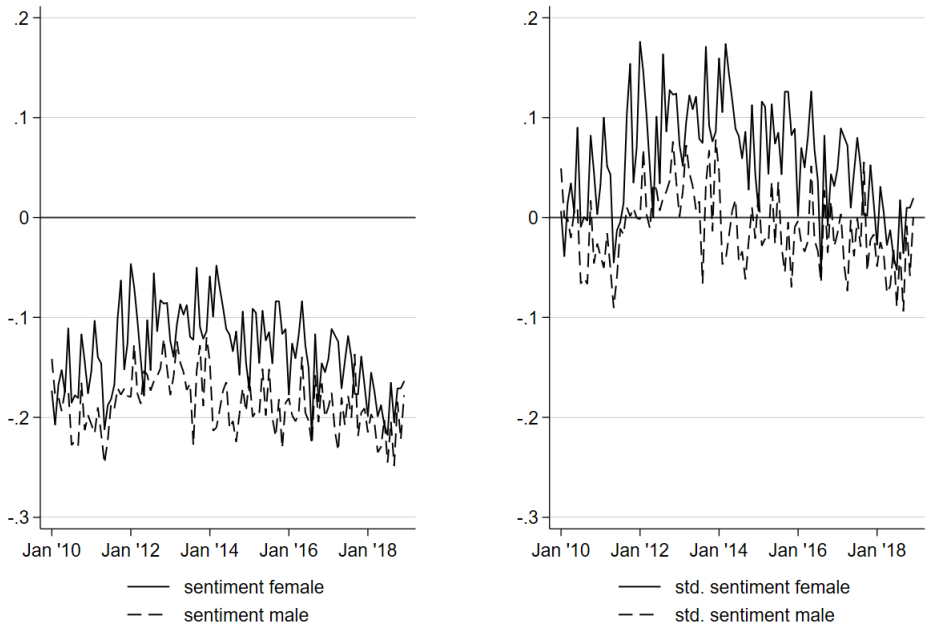


Figure 8: Average sentiment scores per month and gender. Left panel: raw sentiment scores. Right panel: standardized sentiment scores.

after standardization, the average sentiment score for comments classified as *male* is close to and fluctuates around zero, whereas the average sentiment score for comments classified as *female* is largely positive. As illustrated above, we use these standardized sentiment scores as weights for our comments. In particular, we multiply each comment i that is classified as covering a gender stereotypical topic $t \in T^{Gender}$ with $(1 + \text{std\_sentiment\_score}_i)$. Thus, comments with average sentiment are given the same weight as in our main analysis, whereas more benevolent comments are given larger, and more hostile comments are given lower weight than before.

Offensive language We find that $133,266$ or (11.46%) of our comments are classified as *offensive*. In particular, $13,67\%$ of comments that we classify as *female*, and 11% of comments that we classify as *male* feature offensive language. Note that this result does not conflict with our findings on sentiment: in particular, although there are relatively more offensive comments on women, their sentiment is on average more positive than the average sentiment of offensive comments about men.

4.2. Prevalence and development of gender stereotypes

4.2.1. Index

We use our text analysis procedure to document the prevalence and development of gender stereotypes over the course of almost a decade. To this end, we compute an index that captures the degree to which gender stereotypes exist in our data at a given point in time. More specifically, we consider each of our pooled (gender stereotypical) topics t in a particular month τ. For that topic and month, we count how many comments i are classified as *female*, and

how many are classified as *male*. To take into account that there are generally fewer comments about women than about men, we normalize these counts with the absolute number of *female* and *male* comments in month τ, respectively. Finally, we compute the difference between these normalized counts for each topic t and month τ:

$$index_{t,\tau} = \frac{\sum_i (female_{i,\tau} \cap t_{i,\tau})}{\sum_i female_{i,\tau}} - \frac{\sum_i (male_{i,\tau} \cap t_{i,\tau})}{\sum_i male_{i,\tau}}. \quad (2)$$

If women are mentioned relatively more often than men in the context of a specific topic t in month τ, the index in Eq. (2) is positive. If, in contrast, men are mentioned relatively more often than women, the index in Eq. (2) is negative. In other words, a negative index for the topic *professional*, as well as positive indices for the topics *domestic* and *physical*, would be in line with the existence of gender stereotypes in our data.

The main advantage of our index is that it is very easy to interpret. However, its absolute magnitude depends on how frequently a specific topic t appears in general. E.g., if t appears relatively seldom both in the context of male and female comments, our index will be relatively small. As an alternative, one could set the first and second term in Eq. (2) in relation to each other, whereby the index would no longer depend on the overall frequency of t. However, the index could then grow to infinity or even become undefined if the denominator approaches zero. We therefore stick to the specification given by Eq. (2) when we present our main results and provide alternative specifications as robustness checks in Section 5.

Baseline Figure 9 shows our main results, which document the prevalence and persistence of gender stereotypes in UGC. We find that men are discussed more often in the context of *professional* topics than women (index predominantly negative), and that women are discussed more often in the context of *domestic* and *physical* topics than men (index consistently positive). This prevalence of gender stereotypes is relatively stable over time. In particular, our indices remain roughly within the same range over the entire observation period of nine years. However, while we observe no time trend for our index on *domestic*, gender stereotypes in the context of *professional* and *physical* diminish slightly. Specifically, the index for *professional* moves closer towards zero and is even temporarily positive after Jan '13. The index for *physical* approaches zero by the end of 2017. Reassuringly, the index for our placebo topics is close to zero over the entire time period.

The magnitude of our indices can best be interpreted in relation to the overall occurrence of the gender stereotypical topics. E.g., 21.99% of comments in our main sample are classified as *professional*, whereby the corresponding index in Figure 9 is relatively small. In contrast to that, only only 2.92% of comments are classified as *domestic*, which means that the difference between women and men that our index documents is relatively large. 6.25% and 1.81% of our comments are furthermore classified as *physical* and as *placebo*, respectively, whereby the corresponding indices are of intermediate magnitude. We further discuss the issue in Section 5, where we present results that are based on relative indices.

The short-term development of our indices can in parts be linked to eminent national and international events. E.g., the more gender balanced discussion on *professional* topics in 2013/14 coincides with the famous National Socialist Undergrounds (NSU) Trial that centered on the

alleged (female) terrorist Beate Zschäpe and gained huge media attention in Germany. Similarly, the downward movement for our index on *physical* by the end of 2017 coincides with the global #MeToo-movement, and the 2018 instances where it becomes negative could also be explained with the football world cup. In sum, however, our indices remain relatively stable over time, suggesting that gender stereotypes prevail irrespective of what is happening around the world.

Note that our indices capture the ultimate prevalence of gender stereotypes at any given point in time, but they remain agnostic about whether the gender stereotypes are statistically accurate or not (Fraser et al., 2023). Moreover, the indices cannot tell what drives the differences between women and men. E.g., we show that women are discussed relatively less often in the context of professional issues than men, and that this difference diminishes over time, but the index as such is not informative about whether this trend is caused by specific events, more prominent female figures in the public debate, a change in users' perception of gender roles over time, or the exit/entry of users with more or less gender stereotypical perceptions, to name just a few potential explanations. We consider this as a feature, rather than a short-coming, of our analysis. In particular, our main objective is to provide a clean documentation of the prevalence and development of gender stereotypes over time, which is just equivalent to studying the aggregate effect of all potential mechanisms mentioned above. In other words, if our main interest is to measure the absolute prevalence of gender stereotypes in UGC – which is arguably what matters most for public policy – rather than studying selected aspects of it, potential mechanisms that could drive the index play an interesting but secondary role. We further discuss this issue in Section 4.2.2 below, where we conduct regression analyses that control for comment and user characteristics.

Sentiment Figure A.3 presents our sentiment-weighted indices, which are very much comparable to our main results. While the indices for *professional* and *physical* are slightly more positive than before, the index for *domestic* is largely unaffected. This is in line with what we report in Figure 8 and the results on offensive language that we discuss below. Hence, there is just small (if any) evidence for the presence of benevolent, and no evidence for hostile sexism in our data.

Offensive language Figure A.4 shows that our indices are as good as unaffected when we exclude comments classified as *offensive* from our data, emphasizing again that our approach captures subtle and unconscious gender stereotypes that are not expressed in terms of explicit harassment. In other words, the gender stereotypes that we document in our paper are ingrained into users' common speech and do not come along with verbal offenses. Figure A.4 also illustrates that our main results are robust to potential time variation in forum moderation policies: even when we remove every comment that features offensive language, our main results prevail.[33]

News articles Our main argument for studying the prevalence and development of gender stereotypes in UGC is that users' anonymity allows them to voice what they think but would otherwise not say. However, it could be that the comments merely take up gender stereotypes from the news articles that they were originally attached to. In this case, the above results would not be informative about subtle and unconscious stereotypes on behalf of the users.

[33]Our main results are also robust to applying even stricter classifications of offensive language.

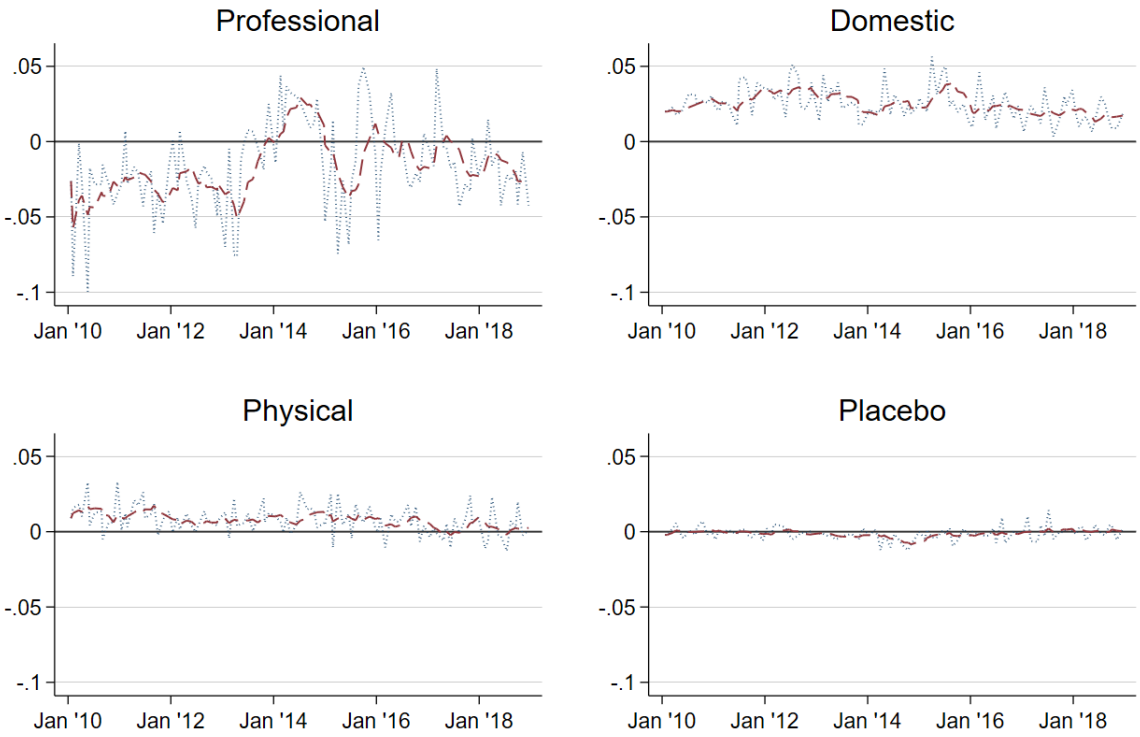


Figure 9: Main results

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The blue dotted line corresponds to the index as illustrated in Section 4.2.1. The red dashed line corresponds to a moving average based on the current and the five previous months.

To demonstrate that the prevalence and development of gender stereotypes in our comments is independent of the news articles, we retrieved the text body of all articles that the comments are attached to.[34] Then, we classify the articles analogous to the procedures that we describe in Section 3.2.2 and compute indices as specified in Section 4.2.1.

Figures A.5 shows that although the indices for UGC and news articles move to some extent in parallel, the latter fluctuate more around zero, indicating that news coverage is gender balanced. Hence, while it is plausible that the two types of indices have a similar shape – since both are likely to be affected by the same eminent events around the world – we find no evidence for gender stereotypes in the news articles, and hence conclude that the users' discussion reflects their own inherent, and not just potential gender stereotypes from the news articles.

4.2.2. Regression analyses

To further explore the prevalence and development of gender stereotypes over time, this section provides the results from two types of regression analyses, where we control for comment and user characteristics. This allows us to study if and to what extent our results are driven by observable features such as length of the comment, news outlet section, or user fixed effects.

We start by estimating the regression equation

$$Topic_{i,\tau} = \beta_0 + \beta_1 female_i + \beta_2 month_\tau + \beta_3 female_i * month_\tau + \theta X_i + \lambda_s + \lambda_u + \varepsilon_{i,\tau} \qquad (3)$$

by OLS, where $Topic_{i,\tau}$ indicates whether comment i is classified as covering one of our pooled gender stereotypical topics, respectively, $female_i$ is a dummy equal to one if comment i is classified as *female*, and $month_\tau$ is a continuous variable capturing a linear time trend. The vector X_i comprises length of a comment, sentiment, and an indicator for offensive language. Finally, λ_s and λ_u capture news section and user fixed effects. Standard errors are clustered on the thread level. The parameters of interest are β_1 and β_3. Specifically, β_1 measures the average difference in the propensity to be mentioned in the context of a particular gender stereotypical topic between women and men for the entire observation period of almost a decade, conditional on our controls. β_3, on the other hand, measures if this difference has risen or fallen over time.

Table 5 shows the OLS estimates using each of our pooled gender stereotypical topics as dependent variable. Consistent with the main results in Section 4.2.1, the estimate for $female_i$ is negative for *professional*, positive for *domestic* and *physical*, and close to zero for the placebo topics, irrespective of the empirical specification. Similarly, we find evidence that the prevalence of gender stereotypes in the context of professional topics declines over time: the corresponding estimate is positive and statistically significant. In contrast to that, the evidence for a decline in gender stereotypes in the context of domestic or physical issues is less clear. Although the corresponding estimates are negative and statistically significant (also owing to our large sample size), they are extremely small both in absolute terms and also relative to our estimates for $female_i$. Interestingly, the estimates hardly change when we include our fixed effects, indicating that the results are driven by variation within users and news sections. In other words, it is not the case that users or news sections with smaller gender bias become more important over time,

[34] $N = 70,235$, equivalent to the number of threads that we consider in our main analysis.

but that the same users and news sections undergo (small) changes.

Table 5: Regression results

	Professional			Domestic		
	(1)	(2)	(3)	(4)	(5)	(6)
$female_i$	-0.0301***	-0.0243***	-0.0228***	0.0291***	0.0262***	0.0234***
	(0.00186)	(0.00178)	(0.00189)	(0.00097)	(0.00097)	(0.00101)
$female_i * month_\tau$	0.00030***	0.00031***	0.00025***	-0.00008***	-0.00007***	-0.00006***
	(0.00003)	(0.00003)	(0.00003)	(0.00002)	(0.00002)	(0.00002)
X_i		Yes	Yes		Yes	Yes
λ_s		Yes	Yes		Yes	Yes
λ_r			Yes			Yes
N	1,148,313	1,148,313	1,094,588	1,148,313	1,148,313	1,094,588
		Physical			Placebo	
	(7)	(8)	(9)	(10)	(11)	(12)
$female_i$	0.0129***	0.00989***	0.00908***	-0.00090	-0.00050	-0.00049
	(0.00117)	(0.00116)	(0.00123)	(0.00057)	(0.00057)	(0.00061)
$female_i * month_\tau$	-0.00010***	-0.00009***	-0.00009***	-0.00001	-0.00002	-0.00001
	(0.00002)	(0.00002)	(0.00002)	(0.00001)	(0.00001)	(0.00001)
X_i		Yes	Yes		Yes	Yes
λ_s		Yes	Yes		Yes	Yes
λ_r			Yes			Yes
N	1,148,313	1,148,313	1,094,588	1,148,313	1,148,313	1,094,588

Notes: Robust standard errors in parentheses. Standard errors are clustered on the thread level. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$

In a second regression analysis, we regress $Topic_i$ on X_i, λ_s, and λ_u alone and replace the topic indicator $t_{i,\tau}$ in the computation of our gender stereotype index from equation (2) with the residuals $\hat{\varepsilon}_{i,\tau}$ from that regression. The idea is that these residuals represent the probability of a specific gender stereotypical topic conditional on observed comment characteristics and user and news section fixed effects.[35] In line with the results from Table 5, Figure A.6 shows that our indices are closer to zero but qualitatively similar to those that we present in Section 4.2.1, suggesting that the prevalence and development of gender stereotypes in UGC is not predominantly driven by any of our controls.

5. Robustness checks

Angela Merkel Our main analysis excludes all comments on Angela Merkel as outliers. Figure A.7 shows that our results are qualitatively comparable when we keep those observations in our sample. In particular, our index for *professional* remains predominantly below, and our index for *domestic* predominantly above zero. In contrast to our main results, the index for *physical* is *negative*; moreover, the index for *domestic* and is closer to zero than above. This finding is intuitive: Angela Merkel does not correspond to classic female gender stereotypes and is seldom discussed in the context of family, home, and physical appearance. Thus, considering comments on her in the analysis shifts these indices downwards.

Interpreting the index for *professional* requires closer examination. From Jan '10 to about Jan '15, the index is on average closer to zero than in Figure 9, i.e., the discussion is more gender balanced. Afterwards, the index is on average further away from zero than in Figure 9,

[35] Section 5 demonstrates that it hardly makes a difference whether we base our indices on topic indicators or (continuous) predicted probabilities.

i.e., the discussion becomes less gender balanced. A plausible explanation is Merkel's prominent role in the refugee crisis starting in Spring 2015. In particular, Merkel pursued a very warm and welcoming policy towards Syrian refugees and thereby triggered plentiful debates among politicians and the public, including users from our discussion forum. As a result, Merkel appeared in many comments that are not related to work or money, thus shifting the index further away from gender balance.

Alternative topic classifications Next, we explore the robustness of our results to alternative topic classifications. In particular, we show that we obtain similar results when we consider each gender stereotypical topic separately (i.e., when we do not pool related topics), and when we use a non-binary topic classification.

Figure A.8 displays our index from Section 4.2.1 for each gender stereotypical topic $t \in T^{gender}$ as well as for our two placebo topics *time* and *space*. With the exception of *work*, all indices are similar to those that we present in Section 4.2.1. Specifically, the index for *money* is predominantly negative, the indices for *home*, *family*, *body*, and *sexual* are predominantly positive, and the indices for *time* and *space* are close to zero. In contrast to our main results, the index for *work* fluctuates around zero, indicating that gender stereotypes in the context of *professional* are mainly driven by gender stereotypes in discussions about money-related issues.

The indices in Figure A.9 are based on a non-binary topic classification. Specifically, we do not assign a dummy equal to one if our algorithm predicts that comment i covers topic t with $Pr(t) > 0.5$, but use the predicted probabilities $Pr(t)$ themselves to compute the index from Section 4.2.1. This makes the indices harder to interpret, but also preserves information that would otherwise get lost (e.g., if the predicted probabilities for a certain topic are often positive, but smaller than 0.5).

Figure A.9 shows that, with the exception of *work*, our indices are nearly unaffected. In particular, the predicted probabilities $Pr(t)$ are either close to zero or close to one, whereby using them instead of dummies does not make much of a difference. In contrast to that, comments classified as *female* often feature a small but positive probability to cover work-related issues. In consequence, the index for *work* is consistently above zero, suggesting that women are *more* likely to be discussed in the context of work than men. We perceive this result as slightly misleading, though. In particular, small but positive predicted probabilities to cover a specific topic are more indicative of a comment *not* covering than actually covering that topic and should be interpreted accordingly (which is, e.g., facilitated by a binary classification as in our main specification).

Alternative gender classification We present the results from three alternative gender classification procedures. First, we re-consider ties *within* our composite gender classification. Specifically, we do not exclude observations that are classified as both *male* and *female* from the analysis, but resolve the ties with respect to the number of male and female instances within one comment. In particular, we count the absolute number of male and female first names, gender specific terms, and celebrities, and classify a comment as *male* if the former outweighs the latter and vice versa. Only if the absolute number of male and female instances is exactly equal to each other, the comment is classified as *none* and dropped from the sample. Figure A.10

shows that our results are as good as unchanged when we base our indices on this alternative gender classification.

Second and third, we re-consider ties *across* our composite gender classification. In particular, we let (i) gender specific terms and (ii) first names overrule the results from the other approaches. As above, our main indices remain nearly unchanged with this new specification, and are thus omitted.

Alternative computation of the index As argued in Section 4.2.1, the magnitude of the index given by Eq. (2) depends on the overall frequency of a specific topic t. To erase this feature, we compute

$$index'_{t,\tau} = \frac{\sum_i (female_{i,\tau} \cap t_{i,\tau})}{\sum_i female_{i,\tau}} / \frac{\sum_i (male_{i,\tau} \cap t_{i,\tau})}{\sum_i male_{i,\tau}}. \tag{4}$$

as an alternative. If $index'_{t,\tau} > 1$ ($index'_{t,\tau} < 1$), women are discussed relatively more (less) often than men in the context of topic t. Note that $index'_{t,\tau} \in [0, \infty]$, i.e., it approaches zero if the nominator becomes very small, and it grows to infinity and eventually becomes undefined if the denominator shrinks to zero.

Figure A.11 shows the results. In line with what we discuss in Section 4.2.1, the index for *professional* is very close to but predominantly smaller than one, i.e., compared to the overall occurrence of the topic *professional*, the difference between women and men is relatively small. Similarly, the index for *physical* is close to but predominantly larger than one. In contrast to that, the index for *domestic* fluctuates around 2, hence, the difference between women and men is relatively large.

6. Conclusion

Gender stereotypes – i.e., general expectations about attributes, characteristics, and roles of women and men – pose an important hurdle on the way to gender equality. It is difficult to quantify the problem, though, since gender stereotypes are not always conscious, and even if they are, they may not be openly expressed. This paper exploits the anonymity of UGC to overcome such challenges. In particular, we develop a novel text analysis procedure that enriches unbiased dictionaries with the flexibility and understanding of word embeddings to classify more than a million user-written comments from a major German discussion forum in terms of (stereotypical) topics, gender, and sentiment. Based on that, we can document the prevalence and development of gender stereotypes over time.

We find strong evidence for the existence and persistence of gender stereotypes in our data. Specifically, we show that men are discussed relatively more often in the context of work and money than women, while women are discussed relatively more often in the context of family, home, and physical appearance than men. While the prevalence of gender stereotypes associated to male topics like work and money diminish slightly, gender stereotypes associated to female topics such as family and home mostly persist over time. This result is supported by regression analyses that control for comment characteristics as well as for user and news section fixed effects. The results are also robust to excluding offensive language from our data, and they are not driven by potential stereotypes in the news articles that the comments were originally

attached to. Moreover, we find just small evidence for benevolent, and no evidence for hostile sexism as drivers of gender stereotypes.

Assessing the prevalence and development of gender stereotypes in our society is a necessary requirement to take further actions towards gender equality. In particular, it is important to understand more subtle and unconscious stereotypes, as these are harder to address than explicit discrimination and harassment. At the same time, however, subtle gender stereotypes are way more difficult to measure. As far as we know, our paper is the first that leverages the anonymity of UCG for a clean and extensive analysis of the prevalence and development of (subtle and potentially unconscious) gender stereotypes over time. We thus advance a paramount societal debate concerning academics, policy makers, and the general public. Our paper presents sharp evidence for the existence of gender stereotypes in UGC. Above all, however, our findings indicate that gender stereotypes prevail despite all measures that have been taken so far and despite global social media movements like #MeToo, thus calling for intensified efforts or alternative remedies.

We develop a novel procedure for the automated topic classification of UGC that can be applied far beyond this paper. Two features are especially advantageous. First, our procedure allows for topic classification in the absence of labeled training data and for flexible dictionary classification even with small dictionaries. These features are particularly beneficial in the context of novel and unconventional data such as text from social media and other online platforms, languages where extensive dictionaries do not exist, and all types of text as data that have rarely been studied before and thus do not exhibit large training data. Second, in terms of application, the procedure is especially useful if relationships between variables in the training data are subject of examination themselves and should thus not be transferred to the sample of interest. E.g., one could apply the procedure to study changes in racial biases of online users, the association with certain topics such as climate change or gun control with political parties, or the political leaning of news media. To further support research in that direction, our method is available as a Python package on `https://github.com/VFMR/WEELex`.

Our paper has several limitations that open up avenues for further research. First, while we document the prevalence and development of gender stereotypes in UGC, we stay agnostic about their relation to actual attributes of women and men. In other words, assessing whether and to what extent gender stereotypes in UGC are a precise or biased reflection of real world circumstances is beyond the scope of our paper. However, gender stereotypes in terms of people's *expectations* about characteristics and roles of women and men pose a substantial problem by themselves – irrespective of the actual status quo – and thus require close examination. In addition, associating men with work and money, and women with family, home, and physical appearance is gender stereotypical by definition, no matter whether the gender stereotype is statistically accurate or not (see, e.g., Fraser et al., 2023, for a taxonomy of gender stereotypes).

Second, users of online discussion fora represent a certain selection of users, whereby the external validity of our findings is limited to that circle. However, given the global reach and growing importance of UGC as well as the public attention that vociferous actors from the online world receive, the population of users that we study is highly influential and thus of inherent relevance.

Finally, as argued above, we do not consider hate speech or open sexual harassment in our

analysis but focus on more subtle forms of gender stereotypes. Although this limits the scope of our findings, we perceive it as a feature of our study: while it is relatively easy to detect gender discrimination in terms of open assaults and offenses, assessing subtle and subconscious gender stereotypes is way more difficult.[36] We provide an important contribution to addressing this challenge by proposing a novel classification procedure that allows us to document the prevalence and development of (subtle) gender stereotypes over time.

[36] In addition, focusing on subtle and subconscious gender stereotypes eliminates potential confounds regarding the supervision of online discussion forums. In particular, hate speech and open sexual harassment are often deleted by moderators. Since we discard such comments from our analysis, our results are unaffected by any potential moderation policies of the forum.

A. Omitted figures

A.1. Exemplary comments

The purpose of a publicly traded company is to generate a sufficient rate of return. And it's every investor's right to exert pressure on the company management. Lamentations out of place.

→ *work* → *money*

That's right. I forgot about family reunion. Wouldn't have thought that young men leave their women and kids to come to Germany.

→ *family*

It's not about proving something. Noone should be forced to join a demonstration. I guess I wouldn't have gone myself, because I'm a lazy bastard. But it's a scandal that official associations distance themselves from demos.

Figure A.1: Topic classification: three examples

Notes: Comment 1 is classified as *work* and *money*, comment 2 is classified as *family*, and comment 3 is classified as not covering a gender stereotypical topic.

I have been saying for long what Mr. Steinbrück said. It's a pity that he was so abandoned by his party allies during the election campaign, especially by Mr. Gabriel. A true Social Democrat who has been impressed and fostered by Helmut Schmidt. Back then, he was the only MP who would disclose his income. The others were too craven and mocked him. He is the most sincere politician whom I know. I can only beg him to return to policy and to show and teach his party allies true Social Democratic policy.

→ *male*

Ms. Kässmann is and will be an idol to me. Smart, good-looking, courageous, warm, coherent, good mother, faithful Christian. The Protestant Church did not suffer, of course, to the contrary. People set standards for dealing with fault.

→ *female*

The author did not care about whether the small sales are truly just due to the design or due to the price as well. I personally prefer to drive a rare car on German streets, a Daihatsu-Copen... Even after 6 years people keep asking me what kind of car that is. However, the Copen is too small for many people, because it just has 2 sears and a small trunk, where in the summer the roof is stored to drive overtly. I also have to say that the Copen was only available as right-hand drive car in its first years. Tuning pieces are only available in Japan for high prices, plus German customs with more than 20%.

→ *none*

Figure A.2: Gender classification: three examples

Notes: Comment 1 is classified as *male*, comment 2 is classified as *female*, and comment 3 is classified as neither *male* nor *female*.

A.2. Further indices

Figure A.3: Sentiment-weighted indices

Notes: The figure displays our sentiment-weighted indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The blue dotted line corresponds to the index as illustrated in Section 4.2.1. The red dashed line corresponds to a moving average based on the current and the five previous months.

Figure A.4: Exclude comments with offensive language

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The indices are based on a subsample that excludes all comments classified as *offensive*. The blue dotted line corresponds to the index as illustrated in Section 4.2.1. The red dashed line corresponds to a moving average based on the current and the five previous months.

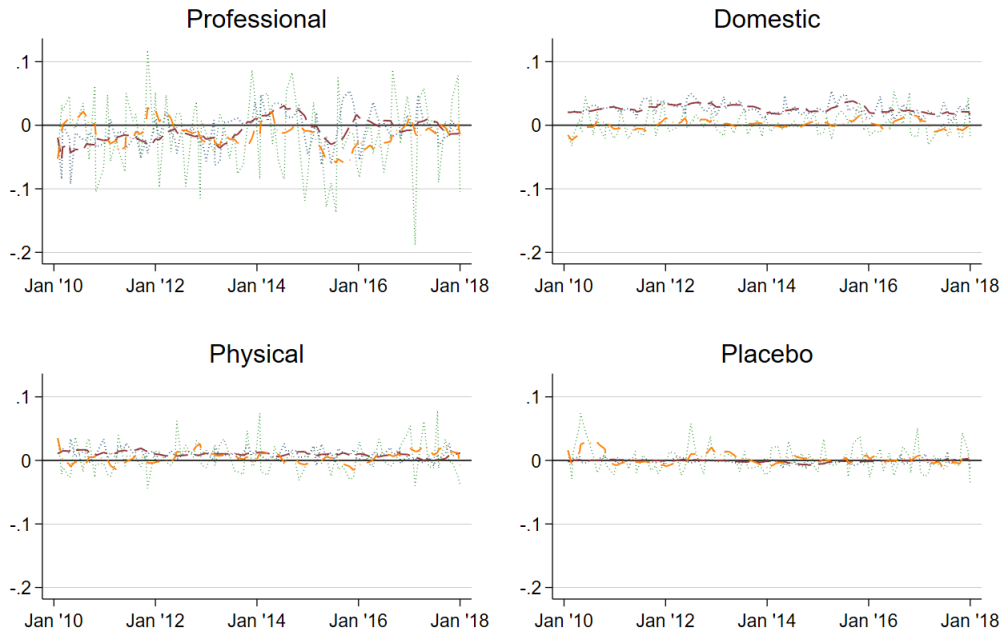


Figure A.5: Gender stereotypes in news articles

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The blue and the green dotted lines correspond to the index as illustrated in Section 4.2.1, where the blue line is based on comments, and the green line is based on news articles. The red and orange dashed lines correspond to a moving average based on the current and the five previous months, where the red line is based on the indices for comments, and the red line based on the indices for news articles.

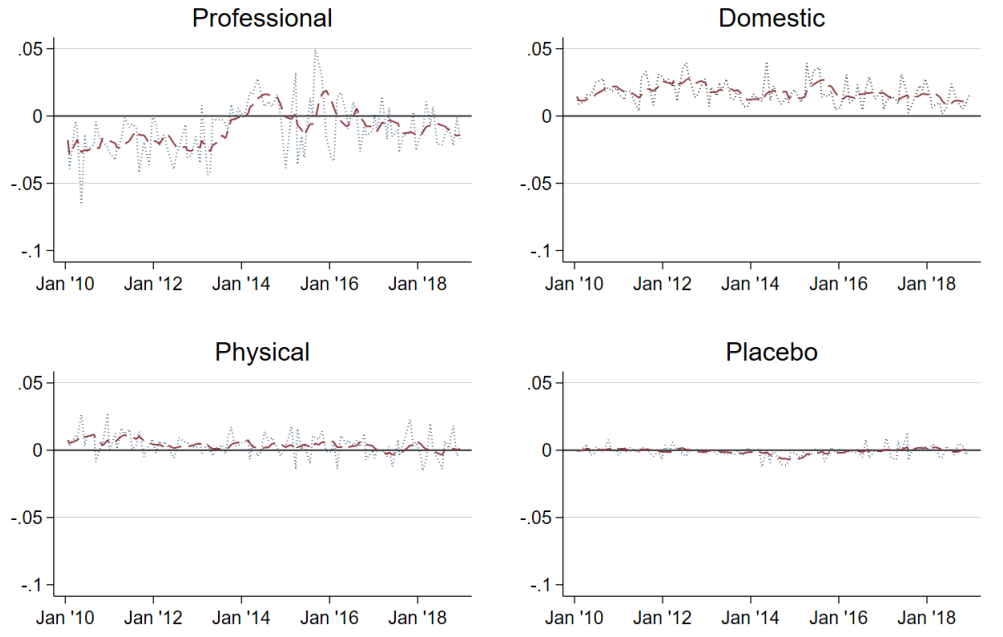


Figure A.6: Indices based on residuals

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*, based on the residuals from an OLS regression of each topic indicator on observable comment and user characteristics. The blue dotted line corresponds to the index as illustrated in Section 4.2.1. The red dashed line corresponds to a moving average based on the current and the five previous months.

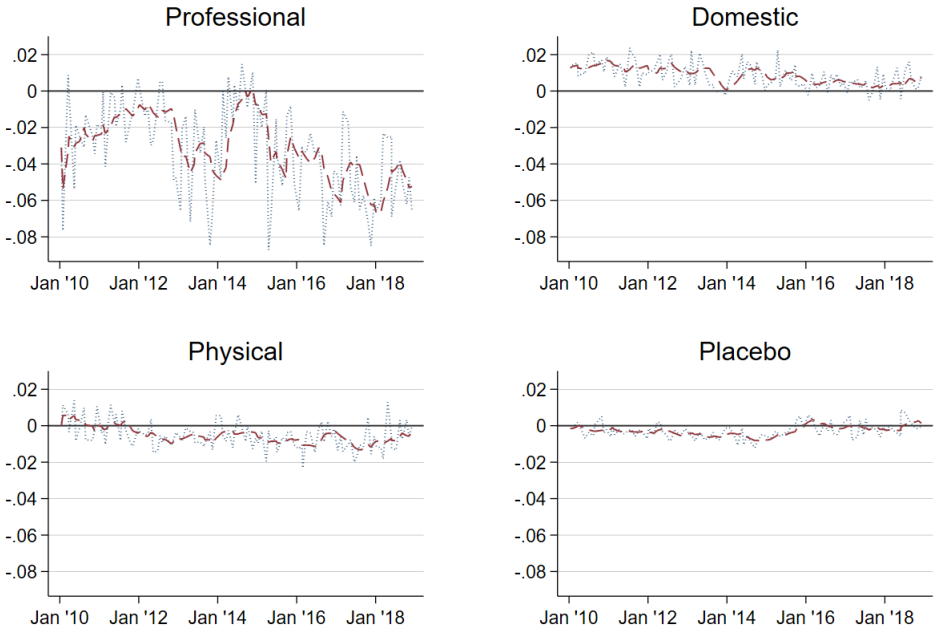


Figure A.7: Include comments on Angela Merkel

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The indices are based on a sample that includes all comments on Angela Merkel. The blue dotted line corresponds to the index as illustrated in Section 4.2.1. The red dashed line corresponds to a moving average based on the current and the five previous months.

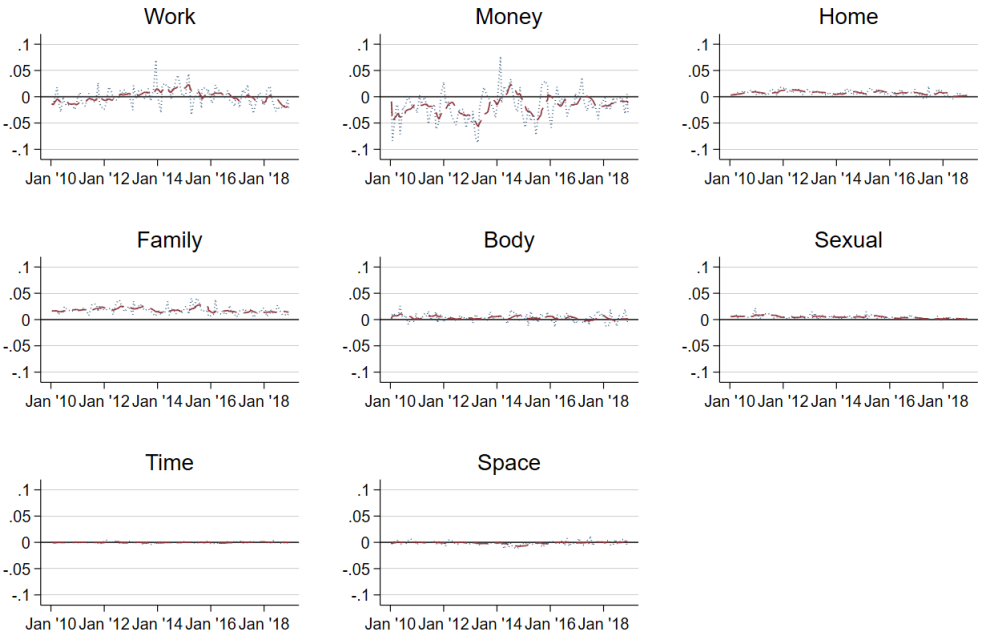


Figure A.8: Non-pooled topics

Notes: The figure displays our indices for the topics *work*, *money*, *home*, *family*, *body*, *sexual*, *time*, and *space*. In contrast to our main analysis, related topics are not pooled together. The blue dotted line corresponds to the index as illustrated in Section 4.2.1. The red dashed line corresponds to a moving average based on the current and the five previous months.

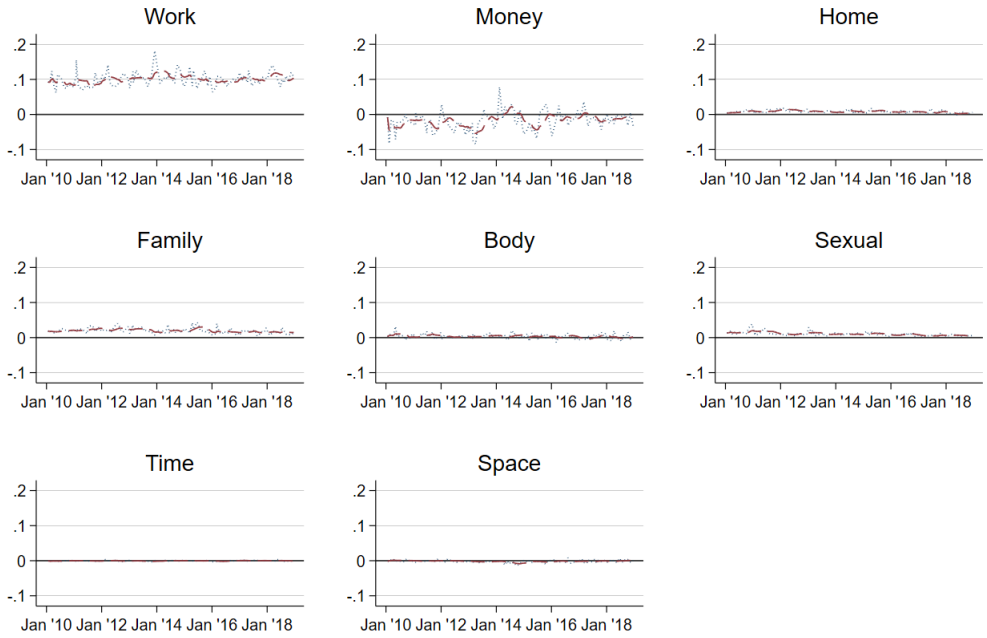


Figure A.9: Non-binary classification

Notes: The figure displays our indices for the topics *work*, *money*, *home*, *family*, *body*, *sexual*, *time*, and *space*. In contrast to our main analysis, related topics are not pooled together. Moreover, the index is based on raw continuous probabilities for topics instead of topic indicators. The blue dotted line corresponds to the index as illustrated in Section 4.2.1. The red dashed line corresponds to a moving average based on the current and the five previous months.

Figure A.10: Alternative gender classification

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The indices are based on an alternative gender classification as illustrated in Section 5. Otherwise, the blue dotted line corresponds to the index as illustrated in Section 4.2.1. The red dashed line corresponds to a moving average based on the current and the five previous months.

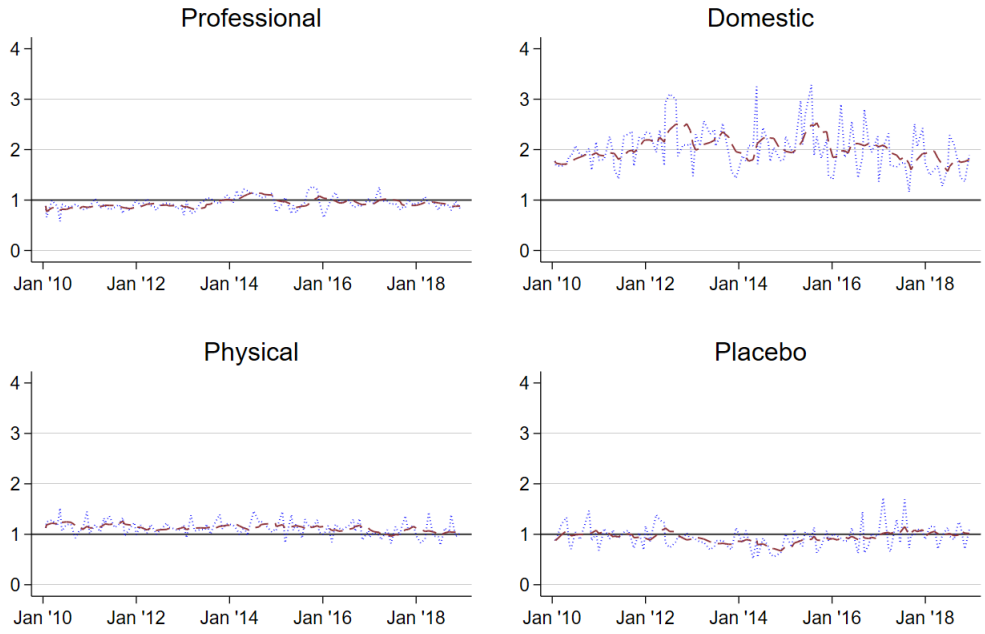


Figure A.11: Alternative gender classification

Notes: The figure displays our indices for the pooled topics *professional*, *domestic*, *physical*, and *placebo*. The indices are based on relative indices as illustrated in Section 5. The red dashed line corresponds to a moving average based on the current and the five previous months.

B. Omitted tables

Table A.1: Wikipedia categories

LIWC topic	Wikipedia category	no. articles
work	working environment	62
money	finance	48
	means of payment	28
	money transfers	143
family	family	72
	family model	8
	relatives (male)	16
	relatives (female)	4
	terms of relativeness	14
home	housekeeping	50
	types of rooms	151
	apartment	25
	apartment (part of building)	16
body	physiques	27
	body extent	67
	body region	12
sexual	human sexuality	100

Notes: Table A.1 shows the Wikipedia categories, the corresponding gender stereotypical topics, and the associated number of Wikipedia articles that we retrieve to validate our topic classification procedure.

C. Technical details

C.1. Clustered tf-ifd

This section provides some technical details on our *clustered tf-idf* approach.

Notation Suppose that there D documents (here: comments) and J clusters of words with similar meaning. Each cluster j comprises W_j words, where W_j is small. Denote these words by $w_{1,j}, ..., w_{k,j}, ..., w_{W_j,j}$.

Clustered term frequency (ctf) The clustered term frequency for cluster $j \in \{1, ..., J\}$ and document $d \in \{1, ..., D\}$ is given by

$$\text{ctf}_{j,d} = \frac{\sum_{k=1}^{W_j} \#(w_{k,j} \text{ in } d)}{\max\left(\sum_{w=1}^{W_1} \#(w_{k,1} \text{ in } d), \sum_{w=1}^{W_2} \#(w_{k,2} \text{ in } d), ..., \sum_{w=1}^{W_J} \#(w_{k,J} \text{ in } d)\right)}, \tag{5}$$

where $\#(w_{k,j} \text{ in } d)$ is the number of occurrences of word $w_{k,j}$ in document d.

Clustered inverse document frequency (cidf) The clustered inverse document frequency for cluster $j \in \{1, ..., J\}$ and document $d \in \{1, ..., D\}$ is given by

$$\text{cidf}_{j,d} = \log\left(\frac{N}{\max\left(\text{df}(w_{1,j}), \text{df}(w_{2,j}), ..., \text{df}(w_{W_j,j})\right) + 1}\right), \tag{6}$$

where $\text{df}(w_{k,j})$ is the document frequency of word $w_{k,j}$.

Clustered tf-idf Given the clustered term frequency and the clustered inverse document frequency, the *clustered tf-idf* is given by

$$\text{ctfidf}_{j,d} = \text{ctf}_{j,d} \times \text{cidf}_{j,d}. \tag{7}$$

C.2. Assumptions

Our method requires different assumptions for both the training and the prediction stage. For model training, the core assumption is that words within a topic are clustered in the embedding vectors space. There are two parts to this. First, the clusters need to capture a topical similarity. Here, we assume that these clusters are convex sets in the embedding vector space. A common way to relate embedding vectors to each other is via the cosine similarity, i.e., the angle between vectors. Thus, embedding vectors are usually considered similar when they are close in terms of the cosine distance. Our approach considers vectors similar if they are close in terms of their euclidean distance. As long as the origin is not contained within the cluster, a short euclidean distance implies a short cosine distance but not vice versa. This makes our clustering a bit more conservative than if it were based on cosine distances, trading off recall in favor of precision.[37]

Second, the actual words in the dictionary must be clustered. One way to verify this is via the cross validation metrics of the training stage. If both precision and recall are sufficiently high, the topics are well separated. We support this in our application by removing ambiguous terms from the dictionary.

For model prediction, the core assumption is that the topics represent the topics in the corpus well. For example, this would be violated with a dictionary about nutritional sciences and a text corpus about planes.

[37] Additionally, cosine distance based clusters would require extrapolation since one needs to assume a hypothetical vector belongs to the same topic if it has a short angle to the observed dictionary words but is otherwise far away from these vectors in terms of their euclidean distance.

D. Qualitative description of an exemplary thread

To better illustrate our data, this section provides an in-depth qualitative description of one exemplary thread in the *SPON* discussion forum. Specifically, we consider a thread from the *Society* section that was originally attached to an article entitled *Why are people prone to believe in higher beings?*, published on January 1st, 2013. We first provide a translation of all comments in the thread, then we discuss structure and content in detail.

Table A.2: Exemplary discussion thread

No.	Time	User ID	Comment
1	11:05am	User_1	1 + 1 = 2, 1 + 0 = 1, 0.75 + 0.25 = 1. How do you think the second equation should be interpreted? Who is Jesus, who is god? How the third one? How the countless others that are still imaginable? How would you face the existence of evil, knowing about the omniscience of God (The omniscent Creator forms the imperfect world? Why? So that it suffers? So that it can be screwed, what can hardly be denied (selling of indulgences, Luther and his wizards, Moses, Abraham, etc.)?) By the way, you wanted to have my opinion. Here you have what I think about the assumption of oneness.
2	12:23am	User_2	Both emanation models, the theistic and the scientific, are incomplete. If you leave out all the historic nonsense, then the only difference is that the theistic model presumes that the creation of World is based on a will.
3	12:38am	User_1	The existence of laws of nature, love, evil, belongs to Creation. Men as part of Creation are no puppets of God. They were provided with reason and conscience. But they often think that they are the actual Lords of Creation. They switch off their conscience and hold God responsible for the consequences. It's just as in the economy: privatize revenues, socialize losses.
4	1:12pm	User_2	Ah, again the question of all questions. Well, reality is really very sad sometimes. But you could shoot a 24/7-soap opera: playing his (eternal) life: Sumerer. Plot: Eating the best food, then sex, then a bit of sleep, and then sending love comments with the computer (appeared in ep. 8 by flipping fingers) into the world. At the latest in ep. 4389 you want to step in and let Sumerer digress from the plot and ask what the shit is all about. Then the stage director smiles and says that he forgot to say that every actor has a free will, of course. In ep. 4390 then, Sumerer nibbles of the Tree of Knowledge, and later even more evil things come to his mind. But somehow I know this film – this story – already (at the very beginning of bible).

Continued on next page

Table A.2: Exemplary discussion thread (Continued)

No.	Time	User ID	Comment
5	1:21pm	User_1	I guess this is why they threw Ashera, Jahwe's intimate partner, together with further heavenly legions, out of the temple and later palmed the sculpture of Virgin Mary off on him? How disappointed must God have been?
6	2:14pm	User_3	You are Gods. Evil things? Nonsense. Jealous gods drive each other to utter fury. Happens that one scratches eyes, breaks noses, bans or hijacks one's lover, blows up figures. Over the course of time, what happens is forgiven and forgotten. And merrily they proceed.
7	2:25pm	User_4	Only in a free economy do culture and civilization blossom (science), because the money is invested lest to lose value (like the flour in the jug in Thomas 97). There needs to be an anticipated liquidity payment if the money is not being invested, monthly, annual, or even daily. The fruit cannot generate further fruit by lending, because there is no interest any more. Yet there is no inflation, if the money is being invested in the medium or long run (in a bank, not in a jug of course), then there are no anticipated liquidity payments. It is such a system that releases the true productive, scientific, and social powers of men and bans sweet idleness. That requires of course, that one cannot sidestep to the monopoly of private property. These two monopolies lending og property and lending of money must be suppressed. In such an economy, Apple would be under control swiftly. Only this way, culture and civilization can develop sustainably.
8	2:35pm	User_5	How is this droll utopia related to the topic? Your evangelical zeal for the prophet society would make you hero of every religious community, though.
9	3:48pm	User_6	Explain to me what "atheism" is supposed to be – I don't know. What substance, which meaning does this verbiage have?
10	16:07pm	User_6	Please compare the secular states of Europe, where every person can believe in whatever he or she wants, with the "theocracies". Then you realize where people are better off.
11	16:11pm	User_7	There is no "atheism". What's that supposed to be anyways?

Continued on next page

Table A.2: Exemplary discussion thread (Continued)

No.	Time	User ID	Comment
12	16:33pm	User_8	If you are talking about the Hitler regime or Communism, you just offended all non-religois people. Non-religious people are neither (Neo-)Nazis nor Communists and therefore not responsible for the crimes of these regimes. Your problem is – in my opinion – that you do not understand the term "non-religious". For you there is only religious people. According to the principle: Everyone believes in something. That is not true. There are people who do not follow a leader. Neither a religious one, nor an Ideologist.
13	16:34pm	User_9	Then you gave whatever humanity does not know yet the name "God". Like the mathematician calls an unknown "x". But you didn't explain anything. Your x just has a different name now: God.
14	17:06pm	User_10	One possible reason for "believe in higher beings" was not mentioned yet: the religious person does not only use gods to explain the world, he also wants to be protected. That's why gods have protective functions in many religions. Man prays to these gods so that they can help him. Sometimes it is ghosts, too: `http://www.spiegel.de/panorama/gesellschaft/aberglaube-in-thailand-wie-geister-das-leben-der-menschen-bestimmen-a-872769.html`. Modern Christians like to call believe in gods or spirits "superstition". And misses that the Holy Ones of Catholic Church are nothing else. Christianity knows evil spirits and demons, too. And the Vatican still offers classes for exorcism. Bottom line: modern Christianity has not developed far away from superstition of "primitive people". Even in modern Europe the world seems to be populated with invisible good and bad beings for religious people.

Continued on next page

Table A.2: Exemplary discussion thread (Continued)

No.	Time	User ID	Comment
15	17:31pm	User_7	S.Freud was dealing with the topic and reaches two special points that he classifies as thought control through religion. He shows understanding for men's search for solace and comfort, and counts in religion, which is especially effective for granting the oldest and strongest wishes for protection and care via a mythologised father figure. Religion as illusion. Freud's main argument against religion is not, however, that it prohibits to enjoy life, but that it overdoes it and punishes resistance with oppression. Who submits to thought control is not able to reach the "psychological ideal, the primate of intelligence". Suppression of base instincts: Freud did not deal with scientific theology, in Roman-Catholic Austria or even Protestant theology, that resisted suppression of thought successfully since the end of the 18th century. He drew his knowledge about religion from his direct experience with Judaism to which he confessed. In his self-portrayal he writes in 1935: "Early absorption in biblical history, just after I mastered the art of reading, has, how I later realized, determined the direction of my thinking." His final work "The Man Moses and the monotheistic religion", published 1939, appreciates Judaism, because its strict rules and the suppression of basic instincts brought about the "triumph of intellectuality over sensuality".
16	17:52pm	User_8	You are definitely mixing up cause and impact, because the "theocracies", especially in the Islamic world, did not cause the political and social crises, but were the consequence. Until a few decades ago, Iran was nearly as secular as the countries from the Arabic Spring were until recently. Whether it is the Mideast conflict, the lopsided support of Israel by the West, trade and oil, or simply the severe inferiority complex against the West that got the radical Islamists to power is a question that should be dealt with in other threads.
17	18:41pm	User_7	Now it's getting ridiculous. Just use Google if you still don't know it even after your umpteenth contribution to the forum, where you rattle off neo-atheistic points of view.

Continued on next page

Table A.2: Exemplary discussion thread (Continued)

No.	Time	User ID	Comment
18	18:55pm	User_7	Slowly read again your quote above, you claim by yourself that religious theocracies are responsible if their states do not fare as well as we do in secular Europe. But whoever – just like you! – blames non-secular countries for the political and social grievances must be consequent and blame the non-religious people there for the existing grievances, or not? And wherever – like in Communist states – atheism becomes doctrine (look up Karl Marx!), then it does not help the atheists to hit and run and have nothing to do with the massacres that were committed in the name of humanism. Because these were nothing else than atheistic theocracies! Horrible crimes were committed in Christianity that Christians are being reminded of all the time. If I as a Christian am held reliable, you as atheist should be too. Any further questions?
19		User_11	One should positively mention the Egyptian Pharaoh Hatschepsut, who introduced the multi-day Opet festival. The beer, oh the beer flew like water. One reckons that the festival had a positive impact on fertility at the Nile, while, what is sad but true, Jahwe's bride was hijacked later on in Israel and he was lonely every since.
20		User_7	I don't know right know, but I think that some psychoanalyst once called the exile from Egypt "birth". Was that Freud or Jung?
21		User_12	Adam and Eve could be cast out of paradise. That means that paradise has boundaries and is not endless. This means, that there is only a certain number of squared meters of paradise available. Who evangelizes is responsible for congestion. If you turn everyone to faithful, it's gonna be like subway in Tokyo up there. – free quote after Marc Uwe Kling, "The kangaroo manifest".

The discussion thread features 21 comments from 12 unique users (we replaced the original user names wit User IDs). All comments were written within a few hours on the date of publication of the underlying article. The comments vary in length: while some comprise just two or three sentences (e.g., comments 2 and 5), others are considerably longer (e.g., comments 7 and 15). All comments are somehow related to religion, which is the primary topic of the underlying article, but starting from comment 7, the discussion digresses towards political and economic issues, too. Some users reply to each other, but this is not always the case. E.g., comments 1 to 3 are seemingly unrelated to each other, but comment 4 is a direct reply to comment 3. Similarly, comments 6 and 7 are related to the general discussion in the thread but do not respond to any previous comments; comment 8, in contrast, is a reaction to comment 7. While many comments contribute to an overall (developing) discussion within the thread,

some comments are just random (e.g., comment 18). We also observe that direct interactions between users are relatively short-lived: e.g., User_1 and User_2 have a brief "conversation" in the beginning of the discussion – although they do not always immediately react to each other – and User_7 and User_8 have a brief conversation towards the end. These two conversations are not related to each other. In sum, the path dependency of the discussion within the thread is rather limited.

References

Akerlof, George A and Rachel E Kranton (2000) "Economics and Identity," *The Quarterly Journal of Economics*, **115** (3), 715–753.

Altonji, Joseph G and Rebecca M Blank (1999) "Race and gender in the labor market," *Handbook of Labor Economics*, **3**, 3143–3259.

Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec (2013) "Steering User Behavior with Badges," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, 95–106: ACM.

Anderson, Michael and Jeremy Magruder (2012) "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database," *The Economic Journal*, **122** (563), 957–989.

Ash, Elliott, Daniel L Chen, and Arianna Ornaghi (2021a) "Gender attitudes in the judiciary: Evidence from US circuit courts," *Center for Law & Economics Working Paper Series*, **2019** (02).

Ash, Elliott, Ruben Durante, Maria Grebenschikova, and Carlo Schwarz (2021b) "Visual Representation and Stereotypes in News Media," *CESifo Working Papers* (9686).

Ash, Elliott and Stephen Hansen (2022) "Text Algorithms in Economics."

Athey, Susan (2019) "The Impact of Machine Learning on Economics," in *The Economics of Artificial Intelligence*: University of Chicago Press, 507–552.

Athey, Susan and Guido W Imbens (2019) "Machine learning methods that economists should know about," *Annual Review of Economics*, **11**, 685–725.

Bertrand, Marianne (2020) "Gender in the twenty-first century," *AEA Papers and Proceedings*, **110**, 1–24.

Bertrand, Marianne and Esther Duflo (2017) "Field experiments on discrimination," *Handbook of Economic Field Experiments*, **1**, 309–393.

Blackburn, Heidi (2017) "The status of women in STEM in higher education: A review of the literature 2007–2017," *Science & Technology Libraries*, **36** (3), 235–273.

Blau, Francine D and Lawrence M Kahn (2017) "The gender wage gap: Extent, trends, and explanations," *Journal of Economic Literature*, **55** (3), 789–865.

Bohnet, Iris (2016) *What works*: Harvard university press.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017) "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer (2016) "Stereotypes," *The Quarterly Journal of Economics*, **131** (4), 1753–1794.

——— (2019) "Beliefs about gender," *American Economic Review*, **109** (3), 739–73.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan (2017) "Semantics derived automatically from language corpora contain human-like biases," *Science*, **356** (6334), 183–186.

Charles, Kerwin Kofi and Jonathan Guryan (2011) "Studying discrimination: Fundamental challenges and recent progress," *Annual Review of Economics.*, **3** (1), 479–511.

Chevalier, Judith and Dina Mayzlin (2006) "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research*, **43** (3), 345–354.

Cuddy, Amy JC, Susan T Fiske, and Peter Glick (2008) "Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map," *Advances in experimental social psychology*, **40**, 61–149.

Dachwitz, Ingo (2016) "Analyse von Spiegel Online: So tickt Deutschlands größte Nachrichtenseite," *Netzpolitik.org*, URL: `https://netzpolitik.org/2016/analyse-von-spiegel-online-so-tickt-deutschlands-groesste-nachrichtenseite/\#netzpolitik-pw`.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018) "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*.

Easley, David and Arpita Ghosh (2013) "Incentives, gamification, and game theory: An economic approach to badge design," *ACM Transactions on Economics and Computation (TEAC)*, **4** (3), 16.

Ellemers, Naomi (2018) "Gender stereotypes," *Annual Review of Psychology*, **69**, 275–298.

Fisher, Robert J (1993) "Social desirability bias and the validity of indirect questioning," *Journal of Consumer Research*, **20** (2), 303–315.

Fiske, Susan T (2010) "Venus and Mars or down to Earth: Stereotypes and realities of gender differences," *Perspectives on Psychological Science*, **5** (6), 688–692.

Fraser, Kathleen C, Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof (2023) "What Makes a Good Counter-Stereotype? Evaluating Strategies for Automated Responses to Stereotypical Text," *Working Paper*.

Gagliardone, Iginio, Danit Gal, Thiago Alves, and Gabriela Martinez (2015) *Countering online hate speech*: UNESCO Publishing.

Gallagher, Ryan J., Kyle Reing, David Kale, and Greg Ver Steeg (2017) "Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge," *Transactions of the Association for Computational Linguistics*, **5**, 529–542.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018) "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences of the United States of America*, **115** (16), E3635–E3644.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019) "Text as data," *Journal of Economic Literature*, **57** (3), 535–574.

Glick, Peter and Susan T Fiske (2001) "An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality.," *American Psychologist*, **56** (2), 109.

——— (2018) "The ambivalent sexism inventory: Differentiating hostile and benevolent sexism," in *Social cognition*: Routledge, 116–160.

Grimmer, Justin and Brandon M Stewart (2013) "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis*, **21** (3), 267–297.

Grootendorst, Maarten (2022) "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*.

Hsueh, Mark, Kumar Yogeeswaran, and Sanna Malinen (2015) ""Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors?" *Human communication research*, **41** (4), 557–576.

Jensen, Robert and Emily Oster (2009) "The power of TV: Cable television and women's status in India," *The Quarterly Journal of Economics*, **124** (3), 1057–1094.

Kearney, Melissa S and Phillip B Levine (2015) "Media influences on social outcomes: The impact of MTV's 16 and pregnant on teen childbearing," *American Economic Review*, **105** (12), 3597–3632.

Kite, Mary E, Kay Deaux, and Elizabeth L Haines (2008) *Gender stereotypes.*: Praeger Publishers/Greenwood Publishing Group.

Kozlowski, Austin C, Matt Taddy, and James A Evans (2019) "The geometry of culture: Analyzing the meanings of class through word embeddings," *American Sociological Review*, **84** (5), 905–949.

La Ferrara, Eliana, Alberto Chong, and Suzanne Duryea (2012) "Soap operas and fertility: Evidence from Brazil," *American Economic Journal: Applied Economics*, **4** (4), 1–31.

Luca, Michael (2016) "User-generated content and social media," in *Handbook of Media Economics*, **1**: Elsevier, 563–592.

Marjanovic, Sara, Karolina Stańczak, and Isabelle Augenstein (2022) "Quantifying gender biases towards politicians on Reddit," *PloS one*, **17** (10), e0274317.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014) "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, **104** (8), 2421–2455.

Meier, Tabea, Ryan Boyd, James Pennebaker, Matthias Mehl, Mike Martin, Markus Wolf, Andrea Horn, T Meier, R Boyd, J Mehl, M Martin, M Wolf, and M Horn (2019) ""LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE-LIWC2015."

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin (2018) "Advances in Pre-Training Distributed Word Representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig (2013) "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746–751.

Murtagh, Fionn and Pedro Contreras (2012) "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2** (1), 86–97.

Pedregosa, F, G Varoquaux, A Gramfort, Michel V., B Thirion, O Grisel, M Blondel, Prettenhofer P., R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay (2011) "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, **12**, 2825–2830.

Pennebaker, James, Ryan Boyd, Kayla Jordan, and Kate Blackburn (2015) "The Development and Psychometric Properties of LIWC2015."

Pennebaker, James W, Martha E Francis, and Roger J Booth (2001) "Linguistic inquiry and word count: LIWC 2001."

Podsakoff, NP (2003) "Common method biases in behavioral research: A critical review of the literature and recommended remedies," *Journal of Applied Psychology*, **885** (879), 10–1037.

Řehůřek, Radim and Petr Sojka (2010) "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50, Valletta, Malta: ELRA.

Remus, Robert, Uwe Quasthoff, and Gerhard Heyer (2010) "SentiWS-A Publicly Available German-language Resource for Sentiment Analysis.," in *LREC*.

Struß, Julia Maria, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner (2019) "Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language," in *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 354–365, Erlangen, Germany: German Society for Computational Linguistics & Language Technology.

Tetlock, Paul C (2007) "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, **62** (3), 1139–1168.

Wang, Zhongmin (2010) "Anonymity, social image, and the competition for volunteers: a case study of the online market for reviews," *The BE Journal of Economic Analysis & Policy*, **10** (1), 1–35.

Watanabe, Kohei (2021) "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages," *Communication Methods and Measures*, **15** (2), 81–102.

Wiegand, Michael, Melanie Siegel, and Josef Ruppenhofer (2018) "Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language," in *Proceedings of the GermEval*, Vienna, Austria.

Wolf, Markus, Andrea Horn, Matthias Mehl, Severin Haug, James Pennebaker, and Hans Kordy (2008) "Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count," *Diagnostica*, **54**, 85–98.

Wu, Alice H (2018) "Gendered language on the economics job market rumors forum," *AEA Papers and Proceedings*, **108**, 175–79.

Zhang, Xiaoquan (Michael) and Feng Zhu (2011) "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia," *American Economic Review*, **101** (4), 1601–1615.