LDK 2023

Sentiment Inference and Gender Classification for Gender Profiling

Manfred Klenner
Department of Computational Linguistics
University of Zurich, Switzerland
klenner@cl.uzh.ch

Abstract

In this paper we describe the further development of an existing rule-based system for sentiment inference. We have created new resources, trained models for the novel language-specific task of gender classification of nouns and applied it to German gender-tailored profiling in newspaper texts. We discovered an imbalance wrt. gender denoting nouns and the role they take as sources or targets of verbs denoting positive or negative relationships. Our goal was to get empirical access to the perception of gender, their roles and their reciprocal relations as portrayed in the news. Our empirical findings are based on statistical hypothesis testing.

1 Introduction

The identification of gender denoting expressions in texts might serve various purposes. For instance, it could be used to identify bias or other forms of imbalance like gender stereotypes as portrayed by the media. We focus on the detection of polar relations (in favor of, against) and polar roles (e.g. positive or negative actor) that gender referring expressions occupy in three Swiss newspaper texts. Given the sentence *Merkel cheats the people*, we are entitled to infer that the writer claims that Merkel acts against her nation and that she should be regarded as a villain. We have compared the distribution of male and female denoting expressions in such contexts on the basis of 380 polar verbs that express a positive (in favor) or negative (against) relation between the actor (the source) and the theme, patient or recipient (the target). We use the term *sentiment inference* for this task, because the identification of relations or roles is not in every case just a simple lexicon lookup. Subordination and negation has to be taken into account. Take the sentence *The land criticizes that Europe (not) supports the Ukraine*: from the unnegated version we can infer that - among others - the mentioned land is against the EU and the Ukraine. The inference pattern here is: if some actor A (land) is against something (support) that is good for another actor B (Ukraine), than A is against B, at least in a situation specific way. The negated version with *not* gives rise to the opposite inference, that A is in favor of B. In a couple of papers e.g. (Klenner and Amsler, 2016), (Klenner et al., 2017a), (Klenner et al., 2017b), (Klenner, 2018) and (Göhring et al., 2021) we have described the resources and principles behind sentiment inference[1].

In this paper, we focus on (the usage of) a new system component that allows us to do gender tailored analysis, namely our gender aware animacy classifer. Moreover, we not only are carrying out an intrinsic evaluation but also an extrinsic end-to-end evaluation. The goal was to find out whether these two components - the rule-based inference system and the gender classifier are suitable means for gender profiling. *Gender Profiling* strives to identify the contexts male and female denoting expressions[2] occupy according to e.g. the media and whether the distribution is uniform are imbalanced[3]. A finding contributing to the female profile could be, for instance, that female nouns are significantly more often the targets of particular verbs than male denoting nouns.

First, we describe our rule-based approach to sentiment inference, then we introduce our new gender classifier and then we discuss the empirical results of applying these two components to newspaper texts from 2004 to 2022. We also try to find out whether the gender profiles have changed, i.e. whether there is a difference between 2004-2014 and 2018-2022.

[1] See https://pub.cl.uzh.ch/demo/stancer/index.py for an online demo.

[2] Certainly, we do not claim that gender is a binary category; but gender-denoting nouns without explicit indications (e.g. '*') do have a binary reference that we cannot overcome.

[3] We avoid the stronger notion of *bias*, since we cannot determine whether the incidents reported by the news are facts or stem from a biased world view.

2 Sentiment Inference

Sentiment (or opinion) implicature (Deng and Wiebe, 2014) aims to predict positive or negative attitudes of opinion holders towards other persons, groups, etc. or towards inanimate entities (targets). We would like to adopt a broader view and call the resulting task *sentiment inference*[4]. If we read that someone has honored, punished or even hurt someone else, then, strictly speaking, we do not know whether there is an *attitude* of the initiator towards the target: we only know that some action was carried out that affected the target in a positive or negative way[5]. Sentiment inference as we put it is the prediction of positive and negative relations holding between a source (an opinion holder or not) and a target.

The central resource of our model is a verb lexicon comprising about 1,000 different verbs. Verbs might have more than a single reading, so in principle, disambiguation was needed. However, there is no verb disambiguator available for German and we do not have the resources to train one. Fortunately, it turned out that disambiguation partly can be done on the basis of dependency parsing, selectional restrictions and animacy detection (see Klenner and Göhring (2022)).

At first glance, animacy detection seems to be related to semantic role labeling. The semantic role *actor* most naturally would be animate. However, we have shown that existing semantic role labeler for German are not reliable in this respect (Klenner and Göhring, 2022). One problem of the task is metonymy, where e.g. a capital city stands for a government (e.g. in *Wien criticizes Brussels*).

Before we have an example, let us first discuss the kind of information a verb in our lexicon carries. Take *to cheat*. As most of the polar verbs, it has two polar roles, a source and a target. It also expresses a directed relation (here: *against*) that holds between the two. Here, the source is (acting) against the target. Moreover, the source of *cheat* might be regarded as negative actor, a villain, the target as the victim (given that the sentence is factual, i.e. not negated or in modal mode).

Table 1 and table 2 illustrate the kind of specifications in our lexicon. Table 1 defines the first frame of *sorgen für* (Eng. care for). The upper part of the table are restrictions that must be fulfilled in order to instantiate the polar frame (below the line). The (dependency) parse must comprise exactly a subject (subj) and prepositional phrase (pp) with the preposition *für* and the subject and the noun of the pp must be animate (+a).

dep. label	subj	pp-obj
lex. restr.	-	prep=für
sel. restr.	+a	+a
polar role	source	target
polar rel.	in favour	-
polar effect	+actor	+effect

Table 1: Frame I of *sorgen für* (Eng. care for). Dependency label (dep. label), lexical restriction (lex. restr.) and selectional restriction (sel. restr.) as well as the polar profile are shown.

If this is given, the filler of the subject is regarded as the source of an in-favor relation towards the target which is the noun of the pp. The source is claimed to be a positive actor (+actor) and the target to receive a positive effect (+effect).

dep. label	subj	pp-obj
lex. restr.		prep=für
sel. restr.	+a	-a
polar restr.		+pos
polar role	source	target
polar rel.	in favour	
polar effect	+actor	

Table 2: Frame II of *sorgen für* (Eng. care for)

Table 2 specifies frame II of the same verb. It is also an example where animacy is a disambiguating factor. The subcategorization frame II is the same (incl. the preposition) as frame I, but the filler of the pp noun is inanimate (-a). A German example sentence would be: *Sie sorgte für gute Stimmung*. The English translation is: she provided a good atmosphere. A different verb is used in English. Please note that frame II has an additional polar restriction, namely that the filler of the pp noun should be positive (+pos). We have implemented a phrase-level polarity composition on the basis of a polarity lexicon[6] (see Clematide and Klenner (2010)) and composition rules (see Moilanen and Pulman (2007) for the principles of sentiment composition). Here *good atmosphere* is recognized as

[4]The notion has been used in the past, see e.g. (Choi et al., 2016), who defined it as *directed opinion*.

[5]In the sentence *The government destroyed all our hopes*, the government is a negative source, but not an opinion holder.

[6]German Polarity Lexicon: download from the IGGSA website under `https://sites.google.com/site/iggsahome/downloads`

a positive phrase. Only if the pp is positive, the actor is a positive actor, if it is negative (frame 3, not shown) like in *bad atmosphere* the actor also is negative. Given a neutral actor like in (*Sie sorgt für Papier*, Eng. She ensures that there is enough paper), no polar relation or polar role at all should be set.

The selectional restrictions are not gender-specific. But the selectional restriction animate (+a) is fulfilled if either a male or female denoting noun is found as a filler.

For the present study, we used those 368 out of the 1,000 verbs that passed a particular frequency threshold (discussed in section 5). We further divided these 380 verbs into 3 subclasses: verbs denoting physical events (119 cases) like *to hit*, verbs denoting emotional events (101 cases) like *to enjoy* and verbs denoting communicative acts (160 cases) like *to blame*. This subdivision allowed us to focus on differences on a more fine-grained level. A couple of verbs cannot be assigned a definite category, e.g. *to hurt* could happen as a physical or an emotional incident. Such verbs are kept in both classes[7]. Table 3 shows some examples.

verb DE	verb EN	p	e	c
töten	kill	+	-	-
zerstören	destroy	+	-	-
quälen	torture	+	-	-
sorgen	care	-	+	-
verabscheuen	detest	-	+	-
ärgern	annoy	-	+	-
beschuldigen	blame	-	-	+
beschimpfen	insult	-	-	+
anprangern	accuse	-	-	+

Table 3: Verbs for 3 subclasses: p (physical), e (emotional), c (communicative)

Although the division into 3 subclasses is a step towards a more fine-grained analysis, there are commonalities across classes in terms of the strength of a verb. In psychology, but also in the Natural Language Processing (NLP) community, words have been characterized not only in terms of polarity (positive, negative) but also in terms of arousal and dominance (see Mohammad (2018)). Arousal quantifies the intensity of emotion provoked by a stimulus, and dominance the degree of control exerted by a stimulus. For instance, *tired* has low arousal and low dominance, *angry* has high arousal and medium dominance and *vanquish and defeat* has high arousal and high dominance. We used the VAD resource of Mohammad (2018)[8] to assign scores for arousal and dominance to our verbs. 80 out of 380 were not found in this resource. We used fastText (Joulin et al., 2017) embeddings to find scores for these out of vocabulary (oov) verbs. We took the most similar verb of an oov verb and transferred its score to the oov verb. Most of the time, synonyms were found, but sometimes also antonyms. Thus, we manually inspected the pairings and approved the transfer or corrected it, if needed (choosing the best fitting similar word). The higher the arousal and dominance values of negative verbs, the clearer is the source of such a verb regarded as a villain and the target as a victim. This as well might reveal some gender-specific differences.

The model architecture up to the point where we started to create a version of the system for the task of gender profiling consisted of a lexicon of verbs, specifying their polar properties and selectional restrictions, a dependency parser and an animacy classifier. The gender classifier is new, also the classification of verbs as belonging to one of three verb classes and the arousal and dominance assignment to these verbs.

3 Grammatical Gender Classification

The grammatical gender of an animacy denoting expression in German can be either male or female[9]. Detecting male or female reference, i.e. reference to men or women, thus boils down to identify the grammatical gender of animacy denoting expressions. Other gender identifies only recently ĥave been included by using the gender star etc. However, in our texts they are not being used. The most indicative part of a gender denoting expression, e.g. a noun phrase is the nominal head. If we had a complete list of gender denoting nouns, grammatical gender classification might be regarded as a simple lexicon look-up. However, such a list would be huge and could not be claimed to be complete, since e.g. new professions might come into existance. We have a list of 30,000 profession denoting nouns, 13,000 of which are female forms. Some of them are rather specific and probably will never be used in newspaper texts. Rather than searching

[7] So the sum of the verbs from all classes is more than 368.

[8] The VAD resource is available under `https://saifmohammad.com/WebPages/nrc-vad.html`

[9] There are only very few cases where a neutral noun can refer to an animate (human) referent, i.e. *Mädchen*, Eng. girl.

through such a over-specific, but inherently incomplete list each time a noun has to be classified, a learned model for gender classification might be more reasonable since it also has some generative capacity, i.e. is able to classify nouns never seen before. Such a model should learn the footprint of gender denoting nouns as apposed to non-animacy denoting nouns. Word embeddings seem to be the perfect basis for such classifiers, since they capture relatedness. Still we cannot expect that pretrained word embeddings already provide the three needed (class) clusters: male, female, inanimate. But a machine learning approach might be able to properly weight embedding dimension in order carve out the class-specific profiles.

In Klenner and Göhring (2022) we have introduced German animacy classification. On the basis of 13,000 German nouns that were manually classified as denoting either animate or inanimate entities[10] we trained a logistic regression classifier using fastText embeddings (see our paper for the various experiments and a full discussion). The overall accuracy was 96.67%.

In order to create a gold standard for grammatical gender classification, we manually selected those nouns that could be used to refer to women or men. Examples of female denoting nouns are *Schwester, Gastgeberin, Schauspielerin* (Eng. sister, hostess, actress, respectively).

It turned out that the class frequencies were imbalanced, more male than female denoting nouns. In German, by adding the suffix *in* to the end of male denoting noun (most of the time) a female denoting noun can be created, e.g. *Helfer* $\rightarrow$ *Helfer**in*** (Eng. helper). If such a derived wordform was found in a corpus at least twice, it was added to the female list.

These lists of (fe)male denoting nouns were further augmented by exploiting a list of first names. Again, we only kept firstnames which also were found in a corpus and were above a threshold (here: 10 occurrences). Table 4 shows the final distribution (frequency counts) of male, female and inanimate denoting nouns[11]. Our gold standard comprises more than 18,000 nouns.

We then had slightly more female than male nouns. However, since female nouns are in our text corpus - as we had found out - less frequent than male nouns, we intentionally kept the resulting (little) bias.

	inanimate	female	male
nouns	5826	5637	5002
first names	-	966	966
$\sum$	5826	6603	6200

Table 4: Frequency counts of the three classes

The accuracy of the classifier on a random 75/25 train/test split is 96.0%, see table 5 for precision, recall and f-measure of that split. The mean accuracy of a ten-fold cross validation was 95.20%. Since the train set and the test set are exclusive, the good performance of the classifier indicates that word embeedings for this kind of nouns seem to be a proper basis for learning.

	inanimate	female	male
precision	96.0%	96.9%	94.9%
recall	95.7%	97.6%	94.5%
f1	95.8%	97.1%	94.7%

Table 5: Performance of the three-way, gender-aware animacy classification model

Not all German female denoting nouns possess the *in* ending. In our list of female denoting nouns, 50 have endings other than *in* (e.g. *Frisöse*, Eng. hairdresser). On the other hand, a word with an *in* ending is not a reliable indicator of a female noun. In a corpus of 25 million nouns, we found 67,823 words (tokens) ending with *in*. For 36,247 cases of these *in*-words our classifier predicted *female*. The remaining 31,576 *in*-nouns correspond to 4,035 types. We manually classified 1,000 and found only 5 female denoting words. Classifying *in*-words immediately as female denoting nouns would produce quite some errors. This is not what our fastText-based classifier does, although it uses sub-word splitting.

The performance of our classifier with respect to the non- *in* female denoting nouns cannot reliably be evaluated at the moment. It is future work to train models able to deal with such rare cases.

4 Corpus, Corpus Split and Gender Reference in German

Gender profiling in our study is restricted to the monitoring of polar roles and polar relations male and female denoting nouns occupy in newspaper texts. Different profiles then can be identified on

[10] Download at: https://zenodo.org/record/7630043#.Y-aCU9LMJH4

[11] The list of male first names was reduced to the size of the female first names.

the basis of different distributions. Especially, uneven distributions are of interest, since they can be interpreted as gender specific. The basic assumption behind our approach is that the overall prior distribution of each gender should also more or less be reflected in the frequency of the polar roles they take and in the polar relations they enter in. We, thus, were interesting in constellations where the genders are involved less or more often than their prior (gender) probability suggests. We interpret these cases as polar imbalance that reveals the gender-specific perception these newspapers cast.

We have data for different periods of the same three Swiss newspapers (2004-2022). Only the last period from 2018 to 2022 was sampled by us for this study, the former data are provided by colleagues. The data points of the 2004-2014 data come without a timestamp, and only the plain sentences are available, not the texts. No coreference resolution was possible, thus. This reduces the number of hits, but should not skew the underlying distribution too much: there is no reason to believe that female denoting nouns are more or less often pronominalized than male ones. Only this would distort the prior gender probabilities we have found on the basis of gender denoting nouns. Since in German, inanimate objects also might have male or female grammatical gender (e.g. *Brücke* is female, Eng. bridge), counting male and female pronouns cannot provide any additional information about the gender distribution. Also, the plural use of *sie* (Eng. she) in German might refer to male, female or neutral animate or inanimate referents. Again, corpus statistics would not help. We thus only looked at cases where the gender classifier triggered and we omit pronoun fillers.

In German, the male word form of e.g. a profession for a long time was used generically to refer to either gender. This was true for singular and for plural. For instance, *Lehrer* (Eng. teacher) was used as a singular and a plural form to refer to all genders. However, for over 20 years now distinct word forms have been used in newspapers. Singular male *Lehrer* and female *Lehrerin*, plural *Lehrer* is now reserved for male reference, while *Lehrerinnen* is used for female reference. In recent years in the course of the discussions of a non-binary gender inclusive language usage, apart from special characters like the gender star ('*') like in *Lehrer*innen* or the colon (':') like in *Lehrer:innen* the nominalized participle present of verbs is meant to refer to all genders[12]. For instance, the participle present of the verb *lehren* (Eng. teach) is *lehrend* (Eng. teaching), the nominalized plural form *Lehrende* (Eng. roughly: *teachings* to represent teachers) is used as an all-inclusive reference. This ongoing language chance does not affect our current study. Special characters are not used in the three newspapers, they consequently used male and female forms and avoid the participle present[13].

Our experiments are carried out over the whole corpus but partly also period-wise. In the period-wise mode we also tried to find out whether there is some change in the perception of gender. The most recent period, 2018 to 2022, was compared with the oldest one, from 2004 to 2014. Period 2015-2017 was viewed as a transition period.

We dependency parsed all sentences, extracted predicate argument structure from the parse trees (incl. passive voice normalization), applied the gender classifier to all nouns and run the sentiment inference system. We further analyzed those verb instantiations where the source was classified as male or female. The target was allowed to be animate or inanimate.

5 Empirical Setup

The maximum likelihood estimation (MLE) of the probability of the female gender wrt. whole corpus is 0.183 (2,671,140 out of 14,577,122 gender nouns). The assumption, the null hypothesis H_0, was that the overall prior gender probability should also be reflected in the distribution of the sources and targets of the polar verbs. For instance, in 18.3% of all instantiations of e.g. the verb *beschuldigen* (Eng. denounce) the source should be a female denoting noun. If this expectation is significantly violated a gender-specific imbalance is found that is, the null hypothesis H_0 is rejected.

As an operationalization of this research question we relied on hypothesis testing on the basis of the binomial distribution. Male and female denoting nouns are binomially distributed per verb frame

[12] This, however, is only possible if a verb form is available for the noun which is not the case for e.g. *Professor, Professorin* (Eng. professor).

[13] The participle present nominalization - according to German grammar books - should be used to indicate that some person is involved only temporarily (or even only at the moment) in the task denoted by the participle. *Singende* (singing people) are different from *Sänger* (singer), they only currently are singing. The new usage is not conform with this view, however if it gains acceptance, the grammar books had to be rewritten.

role. For instance the source role of *betrügen* (Eng. cheat) requires an animate filler which either could be denoted by a male or female noun. If a gender occupies a particular verb position significantly less or more often than the prior probability suggests, than an imbalance is found. Henceforth, we call under represented (less often) genders *scarce* and over represented (more often) genders *abundant*. For instance, if female nouns are significantly less often sources of a verb, we say that the verb is *female scarce* for that role. We omit the reference to the role name if it is clear from the context.

We give a schematic example of the statistical procedure: if a transitive (active voice) verb has $n = 2000$ instantiations (and thus 2000 sources) and $s = 100$ sources are female, then we determine the cumulative probability of up to 100 cases given 2000 trials with $p = 0.183$ as $\sum_{i=0}^{100} binom(i, 2000, 0.183)$. If this value is below $\alpha = 0.025$, then we reject H_0 and adopt H_1, i.e. we can conclude that female nouns occur significantly less often as sources than male nouns, the verb is, thus, female scarce. It might be the case (but not necessarily) that male nouns occur significantly more often as sources of the same verb. To check this, the probability of having 1900 or more occurrences of male sources given that p=0.817 is determined $(1 - \sum_{i}^{1900} binom(i, 2000, p = 0.817)$.

We only kept verbs where a normal distribution could be assumed. This is given if $np \geq 5$. Resolved for n we have $n \geq 5/0.183 \geq 27.3$. Overall 380 verbs out of 1,000 verbs are above this threshold.

Most of the verbs are negative verbs. This is not only due to the imbalance in our verb lexicon (70% negative verbs), but also presumably due to the fact that news more often are negative than positive. In our discussion we thus focus on negative verbs and only refer briefly to positive cases in the last subsection of section 6.

6 Empirical Study

We first identified the gender-specific distribution of source and target roles given the set of polar verbs: for which gender which verbs (verb roles) are scarce and for which abundant. A particular verb role might be scarce for one gender and abundant for the other one (and vice versa)[14]. In these cases the imbalance is complementary. We call these verbs *gender prompted*. Table 6 shows an example of the constellation *gender prompted*.

	female	male
source	scarce	abundant

Table 6: Example of *gender prompted*: source of verb *ermorden* (Engl. to kill)

We not only looked at the distribution of a single role, but also at the combination of roles, the possible source-target pairings: female-female, female-male,male-male and male-female. If for a particular verb the source role is abundant for one gender and at the same time the target role is abundant for the other one, the verb reveals a gender opposition (because the verb expresses a negative relationship). We call these verbs *gender settled*. These cases represent the strongest gender-specific claim we can made. Table 7 gives a example of the constellation *gender settled*.

	female	male
source	-	abundant
target	abundant	-

Table 7: Example of *gender settled*: verb *bedrängen* (Eng. to harass)

6.1 Source Role

In this setting, we determined the gender specific occupation of the source role. 72 out of the 380 verbs (19%) are either scarce or abundant for some gender, the rest of the verbs shows no significant gender-specific instantiation pattern.

Out of the 72 verbs, 8 verbs are male scarce, 61 male abundant; 51 verbs are female scarce and 11 female abundant (72=61+11). The intersection of male abundant and female scarce (and vice versa) gives us those verbs that we called gender prompted, i.e. the role in question (here: source) is preoccupied by one gender and rarely ever filled by the other one. All 51 female scarce verbs are male abundant. Also all 8 male scarce verbs are female abundant. Thus, 59 of the 72 verbs are gender prompted verbs, that is about 15% of the 380 verbs.

In order to find out whether these scarce or abundant verbs might show a verb class specific gender distribution, we assigned each verb its verb class and determined for each gender and prompt type

[14] Please note that if female is scarce for some verb, male must not necessarily be abundant (and vice versa): the cumulative probabilities (even of complementary priors) do not necessarily distribute the mass of 1.

(scarce, abundant) a distribution (see Table 8).

	physical	emotional	communicative
↓ ♀	47.06	17.65	35.29
↓ ♂	12.50	12.50	75.00
↑ ♀	9.09	18.18	72.73
↑ ♂	45.90	16.39	37.71

Table 8: Verb class specific distribution of gender (female ♀ or male ♂) scarce (↓) and abundant (↑): source

Each row shows the gender-specific verb class distribution of a type, e.g. female scarce (♀ ↓): 47.06 physical, 17.65% emotional and 35.29% communicative verbs. To give an example of a gender prompted constellation: the communicative verbs where male are scarce (75%) and female abundant (72.73%) are gender prompted, they are often verbs of accusation (8 out of 11, in italics): *accuse, betray, blame, denounce, discriminate, dismiss, incriminate, hate*, avenge, reject, sue.

We can see symmetrical pattern: female scarce is mainly in class physical (47.06%), which is the major group of male abundant (45.9%). On the other hand: male scarce (72.73%) is in the class communicative, which is the major group of female abundance (75%). Thus, male nouns are more often sources of physical violence, while female nouns are more often sources of verbal oppositions. We call it *opposition* instead of *violence*, since the class *communicative* is more heterogeneous than the class *physical*. A communicative negative verb might be one that hurts the patient verbally like *insult*, but also one that might be regarded a defense like *reproach* or *accuse*.

Male nouns are abundant sources of verbs like: *abuse, assault, attack, beat, coerce, complain, condemn, deny, despise, destroy, distort, harass, harm, hurt, insult, kill, murder, rage, rape, slaughter, terrorize*.

The source role of the gender prompted verbs in some cases can be further qualified with the strong notion of a villain. For instance, the source of *slaughter* is a highly negative actor, a villain. On the other hand, the actor of *reproach* cannot be further classified on a polar dimension. The most negative verbs are those that refer to physical (to kill), emotional (to hate) or verbal (to excoriate) violence. These verbs are modeled in our lexicon as having a negative actor. For each gender we determined the percentage of negative actorship. For male abundant, 43% out of the 51 verbs are of that type, male denoting nouns can be regarded as negative actors in theses cases. For female abundance this is just about the half, 23%.

6.2 Arousal and Dominance

As discussed, words (verbs) also carry arousal and reveal dominance. Can we also find gender-specific differences for these two parameters? For verbs with female and male actors: Is the gender specific arousal (dominance) associated with the prompted verbs in line with the prior probability?

What does arousal mean in the context of a polar verb? A high arousal of a negative verb indicates that the source is regarded as a rather negative actor (a villain) and the target as someone highly negatively affected (a victim). Dominance means that the target is in a clear subordinate position.

The overall prior probability for female was 0.183. Now that we are looking for the gender-specific arousal mass for source (actor) roles, we rather should use the MLE estimation of the gender-specific probability of being the source (and later the target), not the overall prior. For female nouns the probability of filling the source role is 0.164 (78,643 female sources out of 478,165 sources). The arousal (dominance) mass for female should thus be 16.4% of the total arousal (dominance) mass.

The gender-specific arousal (dominance) mass is the product of the arousal (dominance) value of a verb (with a particular gender as source) multiplied by the frequency of that verb. The total mass is the sum of both gender masses.

The total arousal mass of male and female verb tokens is 13,677 (rounded). Female arousal level should correspond to 16.4% of this mass, which is 2,246, but only 260 (1.9%) was found. The same is true for dominance, the overall mass is 20,836 but only 416 actually has been seen for female (2% instead of 16.4%). We can interpret this in the following way: compared to female, male (negative polar) actions are dominating and are much more negative emotion evoking. The only reason for the imbalanced mass distribution can be the magnitude of the arousal (dominance) level per verb. Male denoting nouns must occur (more often) as sources of verbs with high arousal (dominance) scores than female denoting nouns.

If we look at the arousal and dominance levels for the target (i.e. patient) role, we find that this time female nouns are much more affected than their prior probability predicts. The MLE estimation of the female prior of the target role is 0.177 (32,264 female targets out of 182,530 targets). The arousal mass of verbs with female/male as targets is 2632. Female nouns cover 40.3% of it instead of 17.7%. The negative load for female targets is drastically higher than for male targets. Note however that in this setting the actor might be male or female (see section 6.4 for the gender-paired view). For dominance we get 41%.

6.3 Target Role

In this setting, we determined the gender specific occupation of the target role. 43 out of the 380 verbs (11.3%) are either scarce or abundant for some gender, the rest of the verbs shows no significant gender-specific instantiation pattern. 34 out of 43 are gender prompted (i.e. scarce for one gender, abundant for the other one).

	physical	emotional	communicative
↓♀	21.43	7.14	71.43
↓♂	91.00	9.00	0.00
↑♀	95.00	5.00	0.00
↑♂	30.43	4.35	65.22

Table 9: Verb class specific distribution of gender (female ♀ or male ♂) scarce (↓) and abundant (↑): target

All 14 female scarce verbs are male abundant and all 20 male scarce are female abundant. 34 of the 43 verbs are, thus, gender prompted verbs, that is 8.9% of the 380 verbs.

If we look at the verb classes (Table 9), female are scarcely targets of negative communication (71.43%), while male are (65.22%). Male are scarcely targets of (particular) physical violence (91%), while female are (95%). Note that high scarce male and high abundant male wrt. to a verb class are not contradicting, because the gender-wise intersection of scarceness verbs and abundant verbs is empty: some physical verbs are scarce, some abundant.

Almost 24% (source: 15% + target: 8.9%) of the 380 verbs are gender prompted. For these verbs female and male denoting nouns are complementary (scarce, abundant) fillers of the source or target role. This indicates a significant gender imbalance.

6.4 Verbs of Gender Opposition

Now that we have for each gender the information for which verb role it is abundant, we can find cross gender cases of opposition, namely the constellation which we have called settled: verbs for which male abundant holds for one role and female abundant for the other one (and vice versa).

We have found 11 verbs with male sources and female targets that show gender opposition : *harass, molest, murder, shoot, abuse, coerce, terrorize, kill murder, rape, injure, assault* . All verbs are expressing physical violence.

For the inverse setting (with female source and male targets) three verbs are found: *denounce, incriminate, accuse*. All verbs of the class *communicative*.

From a very condensed point of view we might say that male denoting nouns cover villain roles (female being the victim), while female denoting nouns cover accuser roles (male being the accused).

We could also look into the gender internal pairings. Only rare cases were found. In the pairing male-male the following verbs are settled: *arrest, convict*. For female-female only *discriminate* was found. There are more cases of cross-gender that gender internal opposition.

7 Gender Profile Change

So far, we have discussed gender profiles on the basis of all data from the whole period. An interesting question might be whether this has changed over the years or whether it is a constant pattern in newspaper texts. We have compared the period OLD (2004-2014) with the period NEW (2018-2022). Period 2015-2017 was left out as a potential transmission period.

First of all, the prior probabilities of gender have changed. In period OLD the probability of a female denoting noun is 0.169, in period NEW 0.196. We carried out our experiments with these period-specific probabilities.

	physical	emotional	communicative
↓♀	50 (54.8)	16.7 (9.7)	33.3 (35.5)
↓♂	20 (11.1)	0 (0)	80 (88.9)
↓♀	20 (11.1)	0 (0)	80 (88.9)
↓♂	46.9 (50)	16.3 (13.6)	34.7 (36.4)

Table 10: The distribution of verb class instantiations for the source role: format 2004-2014 (2018-2022)

Table 10 shows the results for the source role for period OLD and NEW (with NEW in brackets). Slight tendencies can be noticed. Male are even more abundant in the class physical (50% instead of 46.9%) and female less (11.1% instead of 20%). Also there is an increase in female abundance for communication verbs (from 80% to 88.9%) while the increase for this class for male is less high.

	physical	emotional	communicative
↓ ♀	10 (0)	20 (0)	70 (100)
↓ ♂	100 (100)	0 (0)	0 (0)
↑ ♀	95 (93.3)	0 (0)	5 (6.7)
↑ ♂	22.7 (45.5)	9.09 (0)	68.2 (54.5)

Table 11: The distribution of verb class instantiations for the target role: format 2004-2014 (2018-2022)

Table 11 shows the target role development. The most striking change is in verb class *physical* for male. Whereas in the period OLD 22.7% were male abundant, in NEW we have 45.4%. At the same time male is less abundant in the class *communicative* (a drop from 68.23% to 54.5%).

As we can see from the two tables, the profiles have slightly chanced. What is surprising is the fact that the number of female denoting nouns has increased only by 2.7% (from 16.9% to 19.6%). We would have guessed a higher increase, given that gender awareness seemed to have raised in recent years.

8 The Positive Dimension

As mentioned previously, negative verbs are much more frequent than verbs expressing a positive relationship. We have focused, thus, on the against relation in this study. However to complete the picture we might have a brief look into positive relations and the gender specific patterns in this section. We start with the source role. Female abundant verbs are *honor, celebrate, rejoice, win, help, love, like, fall in love, forgive, appreciate*. All female abundant verbs are gender prompted (are at the same time male scarce). Male abundant verbs are *accept, liberate, insist, affirm, respect, care, concede, reveal*. Also all male abundant are gender prompted.

	physical	emotional	communicative
↓ ♀	12.50	50.00	37.50
↓ ♂	27.27	63.64	9.09
↑ ♀	27.27	63.64	9.09
↑ ♂	11.11	44.44	44.44

Table 12: The distribution of verb class instantiations for the source role of positive verbs

Table 12 shows the verb class distribution. It is interesting to see that emotion verbs are much more prominent for positive verbs than for negative ones. Communicative verbs are least abundant for female (9.09%) which is quite the opposite to negative verbs (where it was 72.73%). Physical verbs are less important for the positive relationships.

The statistics for the target role case are too meager to be of any significance. There are 5 verbs that are gender prompted, namely *honor, encourage, love, care, fall in love*. They are female abundant and male scarce. Male is scarcely patient of these verbs while female are abundantly often. Statistics for positive gender cross abundance cannot be found in our data set.

9 Related Work

Bias detection and debiasing are important research topics (see Stanczak and Augenstein (2021) for a survey). Researchers use e.g. pointwise mutual information (PMI) to measure the association of words with gender (Stanczak et al., 2021). We are rather interested in statistically supported claims about gender-specific instantiation patterns of verbs.

In an approach more closely related to ours, Sun and Peng (2021) observe a gender-specific tendency to combine personal and professional events in the Wikipedia pages of celebrities, an asymmetric association where e.g. women's personal events appear more often in the career section than for men. They also establish higher efficiency when extracting events (verb denotations) over analyzing raw text for detecting this gender bias. To this aim, they use the odds ratio (OR), calibrate over synthetic sentences to estimate real occurrence frequencies, and select the events with the largest gender differences.

We are not aware of other animacy detection approaches for German. Also there is no gender classifier available apart from ours. In Klenner et al. (2023), the initial version of our gender classifier applied to gender-tailored role labeling was introduced.

Gender classification in English is primarily restricted to predicting the gender of text author(s) (e.g. bloggers, see Mukherjee and Liu (2010)). Other researchers analyzed the ACL anthology to find gender specific research topics (Vogel and Jurafsky, 2012). However this is restricted to the recognition of the gender of person names. Campa et al. (2019) aim to identify whether the subject of an article is female or male based on (the content of) the headlines. A gold standard of headlines was created and used where male and female reference could be found. Among others, a CNN approach reached an accuracy of 86.7%. In contrast, we do not identify the gender of the subject of the whole text, but of source and target roles of verbs.

Gender profiling is also a task in the area of computational forensic linguistics (Sousa-Silva, 2018), see e.g. the shared task on Bots and Gender Profiling 2019[15]. The task is to determine whether a tweet is from a human or a bot and if human which gender. Again, the gender of the author is profiled, not as in our case the gender of text referents.

We are not aware of any sentiment inference approach to German others than ours. For English, a couple of approaches exist. A rule-based approach to sentiment inference is Neviarouskaya et al. (2009). Each verb instantiation is described from an *internal* and an *external* perspective. For example, "to admire a mafia leader" is classified as affective positive (the subject's attitude towards the direct object) given the internal perspective while it is (as a whole) a negative judgment, externally (here the concepts introduced by the Appraisal Theory are used, cf. Martin and White (2005)).

Rashkin et al. (2016) introduce connotation frames to represent various types of connotations using typed relations. They consider the writer's perspective, the entity's perspective, effects, values as well as mental states. For each predicate, they infer a connotation frame composed of 9 relationship aspects. In contrast to our setting (real sentences), their experiments are based on crowd sourcing with artificial, rather simple sentences (just subject/object, no subclauses).

Choi and Wiebe (2014) address methods for creating a sense-level lexicon for opinion inference. They consider expressed opinions towards events that have positive or negative effects on entities. As words have mixtures of senses among the three classes (+/-effect and Null), they develop a sense-level rather than word-level lexicon. The resulting resource is based on WordNet senses, annotated with one of the aforementioned classes. In contrast, our annotations consider not only effects on entities but also relations between entities as well as actors.

A more recent approach is described in Park et al. (2021). The authors call the underlying task *direct sentiment extraction to question answering (DSE2QA)* which essentially is what others have called sentiment implicature (cf. Deng et al. (2014)). On the basis of a manually labeled corpus on the 2016 U.S. presidential election and on COVID-19, a method is developed that is utilizing BERT-like pretrained transformers. Questions (Does X has negative sentiment towards Y) on whether a particular relationship exists or not are used, answers are aggregated to make a final guess. This approach actually anticipates recent developments in the context of GPT-like models like ChatGPT. The authors of Zhang et al. (2023) show that ChatGPT outperforms existing approaches in the area of stance detection. Moreover, it is also able to explain its answer. The authors claim that this is a crucial new property of such models. We have carried out a couple of initial experiments with ChatGPT as well. A sentence and prompt like *Mister Tiber refuses to help his sick neighbor. Is he in favor or against her?* is answered with *Mister Tiber's refusal to help his sick neighbor suggests that he is against her*, after removal of *sick* the chatbot now finds the prompt *difficult to determine*. This hesitant reaction was typical in our experiments. To find the right prompt is the task to solve in such contexts. As soon as the concrete training procedure behind it has been published, stance-tailored versions of ChatGPT might finally prove superior to other approaches. The chatbot is also able to do gender identification. The following question-prompt pair was correctly resolved: *Die ZDF-Moderatorin log die Verantwortliche des Aufsichtsrats an. Wer ist weiblich?*[16] (Eng. *The ZDF presenter lied to the person in charge of the supervisory board. Who is female?*[17]. The correct answer is *ZDF-Moderatorin, Verantwortliche*. Since the idea of science is not to just develop prompting skills, we have to wait until we have access to the exact methodological details of such models.

10 Conclusion and Outlook

In this paper, we focused on a gender-tailored analysis of newspaper texts. We searched for the gender profiles in terms of the gender-specific roles newspapers convey. We strived to fix those events (denoted by verbs) that are gender prompted, i.e.

[15] See https://pan.webis.de/clef19/pan19-web/author-profiling.html

[16] It is not the grammatical (female) gender, ChatGPT referred to - we checked this.

[17] When we tried the same question one day later, ChatGPT failed to give an answer.

descriptions where male or female denoting nouns are occurring significantly less or more often than expected. An even stronger, gender opposition indicating case are gender settled verbs, where the source role is abundantly filled by nouns of one gender and the target role by nouns of another one.

The profiles that we have found clearly cast male nouns as filling negative actor roles while female nouns are as targets negatively affected. Moreover female nouns as source role fillers are accusers and male the accused. The primary goal of this work is not a particular statistical screening but the development of a methodology which allows to validate (confirm or reject) claims that otherwise must be regarded as mere long-shot guesses. Our approach may also be used in other genres (e.g. fiction instead of news) in which a particular imbalance (e.g. men committing physical violence) may not (claim to) reflect reality, but rather some potential bias in the data that must be checked.

From a technical perspective, we introduced the first gender-specific classifier (as far as we are aware of). We combined it with a rule-based sentiment inference system for gender profiling. Our empirical study was carried out in the established statistical setting of hypothesis testing.

In future work, we like would to apply our approach to new data where coreference resolution is possible in order to increase the statistical basis of our claims. Also, other expressions like e.g. noun phrases with polar adjectives modifying gender denoting nouns could supplement our verb-specific view. The overall goal is an ever more fine-grained apparatus for gender profiling. At some point, we also will focus on gender inclusive reference and how to combine this with our current approach.

Acknowledgements

This work was supported by the Swiss National Foundation (SNF) under the project number 105215_179302 from 2018 to 2022. I would like to thank Anne Göhring for her collaboration in the project. Thanks also to Alison Jong-Ju Kim and Dylan Massey for for their valuable support.

Discussion of Limitations

Our method detects gender imbalance by using an existing rule-based system and a grammatical gender classifier. Neither performs perfectly, and we do not claim that our sampling methods produce representative data drawn from the whole population. Rather, we work with a subset that can be identified by our tools. Generalizing from the subset to the population is not our intention; our approach is a attempt to carry out gender-tailored sentiment analysis. We do not claim to find biases in the data, we instead speak of imbalance. Whether the cause of imbalance is bias would require an additional qualitative analysis of the results.

References

Stephanie Campa, Maggie Davis, and Daniela Gonzalez. 2019. Deep & machine learning approaches to analyzing gender representations in journalism. In *https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15787612.pdf*.

Eunsol Choi, Hannah Rashkin, Luke Zettlemoyer, and Yejin Choi. 2016. Document-level sentiment inference with social, faction, and discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 333–343, Berlin, Germany. Association for Computational Linguistics.

Yoonjung Choi and Janyce Wiebe. 2014. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*, pages 1181–1191.

Simon Clematide and Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for german. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 7–13, Lisbon, Portugal.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment propagation via implicature constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 377–385, Gothenburg, Sweden. Association for Computational Linguistics.

Lingjia Deng, Janyce Wiebe, and Yoonjung Choi. 2014. Joint inference and disambiguation of implicit sentiments via implicature constraints. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 79–88, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Anne Göhring, Manfred Klenner, and Sophia Conrad. 2021. Deinstance: Creating and evaluating a german corpus for fine-grained inferred stance detection. In *17th Conference on Natural Language Processing (KONVENS 2021)*, pages 213–217. ACL Anthology.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association* *for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Manfred Klenner. 2018. Offensive language without offensive words (olwow). In *KONVENS 2018 (The Conference on Natural Language Processing). Germeval Task 2018 – Shared Task on the Identification of Offensive Language*. GSCL.

Manfred Klenner and Michael Amsler. 2016. Sentiframes: a resource for verb-centered german sentiment inference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1–4, Paris, France. European Language Resources Association (ELRA).

Manfred Klenner, Simon Clematide, and Don Tuggener. 2017a. Verb-mediated composition of attitude relations comprising reader and writer perspective. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Budapest, Hungary, April 17-23, 2017, Revised Selected Papers, Part II*, volume 10762 of *Lecture Notes in Computer Science*, pages 141–155. Springer.

Manfred Klenner, Anne Goehring, Alison Yong-Ju Kim, and Dylan Massey. 2023. Gender-tailored semantic role profiling for german. In *12th International Conference on Computational Semantics (IWCS)*, Nancy, France.

Manfred Klenner and Anne Göhring. 2022. Semantic role labeling for sentiment inference: A case study. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 144–149, Potsdam, Germany. KONVENS 2022 Organizers.

Manfred Klenner and Anne Göhring. 2022. Animacy denoting german nouns: Annotation and classification. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1360–1364, Marseille, France. European Language Resources Association (ELRA).

Manfred Klenner, Don Tuggener, and Simon Clematide. 2017b. Stance detection in Facebook posts of a German right-wing party. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 31–40, Valencia, Spain. Association for Computational Linguistics.

J. R. Martin and P. R. R. White. 2005. *Appraisal in English*. Palgrave, London.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Karo Moilanen and Stephen Pulman. 2007. Sentiment composition. In *Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP-2007)*, pages 378–382, Borovets, Bulgaria.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA. Association for Computational Linguistics.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Semantically distinct verb classes involved in sentiment analysis. In *Proceedings of the IADIS International Conference Applied Computing 2009, 19-21 November, Rome, Italy, 2 Volumes*, pages 27–35. IADIS Press.

Kunwoo Park, Zhufeng Pan, and Jungseock Joo. 2021. Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4091–4102, Online. Association for Computational Linguistics.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Rui Sousa-Silva. 2018. Computational forensic linguistics: An overview of computational applications in forensic contexts. *Language and Law*, 5(2):119–143.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

Karolina Stanczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2021. Quantifying gender bias towards politicians in cross-lingual language models. *CoRR*, abs/2104.07505.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360, Online. Association for Computational Linguistics.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL Anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2023. How would stance detection techniques evolve after the launch of ChatGPT? *CoRR*, abs/1207.0016.