Placing M-Phasis on the Plurality of Hate: A Feature-Based Corpus of Hate Online

Dana Ruiter, Liane Reiners, Ashwin Geet d'Sa, Thomas Kleinbauer, Dominique Fohr, Irina Illina, Dietrich Klakow, Christian Schemer, Angeliki Monnier

Placing M-Phasis on the Plurality of Hate: A Feature-Based Corpus of Hate Online

Dana Ruiter[†*], Liane Reiners[‡*], Ashwin Geet D'Sa[♣], Thomas Kleinbauer[†], Dominique Fohr[♣], Irina Illina[♣], Dietrich Klakow[†], Christian Schemer[‡], Angeliki Monnier[♠]

[†]Spoken Language Systems Group, Saarland University, Germany
{druiter, kleiba, dietrich.klakow}@lsv.uni-saarland.de

[‡]Department of Communication, Johannes Gutenberg University Mainz
{liane.reiners, schemer}@uni-mainz.de

[♣]Université de Lorraine, CNRS, Inria, LORIA
{ashwin-geet.dsa, dominique.fohr, irina.illina}@loria.fr

[♠] Université de Lorraine, CREM
angeliki.monnier@univ-lorraine.fr

Abstract

Even though hate speech (HS) online has been an important object of research in the last decade, most HS-related corpora over-simplify the phenomenon of hate by attempting to label user comments as *hate* or *neutral*. This ignores the complex and subjective nature of HS, which limits the real-life applicability of classifiers trained on these corpora. In this study, we present the M-Phasis corpus, a corpus of $\sim 9k$ German and French user comments collected from migration-related news articles. It goes beyond the *hate-neutral* dichotomy and is instead annotated with 23 features, which in combination become descriptors of various types of speech, ranging from critical comments to implicit and explicit expressions of hate. The annotations are performed by 4 native speakers per language and achieve high ($0.77 \leq \kappa \leq 1$) inter-annotator agreements. Besides describing the corpus creation and presenting insights from a content, error and domain analysis, we explore its data characteristics by training several classification baselines.

Keywords: Hate Speech, Corpus Creation, Feature-Based, Multi-Disciplinary, Bilingual, Migration

1. Introduction

The internet has made the exchange of information and ideas between individuals easier than ever before. But through the provision of anonymity and filter bubbles, it has also contributed to the propagation of hateful contents that are a threat to the open exchange of opinions. To gain insights into the dynamics and characteristics of different types of hateful speech and to develop counter-measures, communication researchers and computer scientists alike require high-quality annotated data.

To date, published datasets related to hate speech (HS) come with several limitations. Most corpora used to train HS classifiers over-simplify the phenomenon of HS by labelling user content with binary classes, e.g., *hate/neutral*, whose underlying definition varies greatly across corpora (Jurgens et al., 2019). Another common limitation in HS corpora is the use of slur (Rost et al., 2016) or emotion word lists (Paltoglou et al., 2013) to identify and sample hateful content. These approaches are insensitive to more subtle forms of HS (i.e., implicit hate) that is conveyed through syntactical or contextual features (Cohen-Almagor, 2018). Further, most hate speech corpora are based on the same outlet, i.e., Twitter (more than 50% of datasets), and language, i.e., English ($\sim 40\%$) (Vidgen and Derczynski, 2021). Focusing on Twitter data is especially limiting, given that Twitter poses a special case of user interactions due to the small maximum length of tweets. The English-centrism ignores the cross-cultural differences of the manifestations of hate. Also, HS corpora often do not provide access to the conversational context in which a comment is embedded, which is problematic since HS highly depends on the context (Kovács et al., 2021). All of these limitations reduce the generalisability of HS analysis and classifiers.

In this paper, we present ***M**igration and **P**atterns of **Ha**te **S**peech **i**n **S**ocial Media* (M-Phasis), a corpus which focuses on the topic of migration and addresses several of the above limitations of current HS corpora. Our contributions are the following:

1. Collection (Section 3) of user comments which: 1) are sampled from news articles that match **migration-related** regular expression keywords to ensure relevance to the topic of migration; 2) are not sampled based on a list of pre-defined slur keywords and thus also capture **implicit** forms of hate; 3) are derived from a **diverse** set of mainstream and fringe media outlets; 4) are in two lan-

[*] Equal contribution.

guages, **French and German**, for which only few HS corpora exist; 5) are collected as a comment thread to allow **context-sensitive** analysis.

2. Comment annotations are based on **features** such as negative and positive evaluations, contrasting of groups, and expressions of emotion, which in combination become descriptors of various types of speech, ranging from critical comments to implicit and explicit hate.

3. To identify difficulties of the M-Phasis corpus and to provide a guide for future research, we train and evaluate **classification baselines** on it (Section 4) and analyse errors (Section 5).

4. **Analysis** of 1) the frequent agent-victim tuples found in the corpus (Section 6) as well as 2) of the domain differences between comments of different media outlets (Section 7).

2. Related Work

HS classifiers that detect abusive content online and flag it for human moderation or automatic deletion are the most common **computational approach** to counter HS online (Jurgens et al., 2019). These classifiers are furthermore important research tools, e.g., to explore the dynamics of specific types of HS online (Johnson et al., 2019; Uyheng and Carley, 2021) or to identify common targets of abuse that require special protection (Silva et al., 2021). The algorithms behind HS classifiers are manifold (Schmidt and Wiegand, 2017), ranging from statistical machine learning methods (Saleem et al., 2016; Waseem and Hovy, 2016) to neural approaches applying representations of language models (Yang et al., 2019) in single or multi-task (Plaza-Del-Arco et al., 2021) settings.

In **social sciences**, the focus of HS research lies on the analysis of the manifestation of hate, its dynamics and role in society. A common approach is quantitative content analysis. It focuses on the investigation of manifest media content in a systematic, objective and quantitative fashion (Berelson, 1952). Therefore, an extensive annotation protocol is developed. These annotations are more extensive than those typically performed in computer science, and often also take into account the context. Social science distinguishes between different forms of impolite, uncivil or intolerant communication (Coe et al., 2014; Su et al., 2018; Rossini, 2020); more fine-grained than the binary distinction commonly used in HS corpora. What distinguishes HS particularly from other concepts is that the hateful expression is group-oriented (Erjavec and Kovačič, 2012). Often content analyses treat HS as a special form of incivility (Ziegele et al., 2018) or harmful speech (Robert et al., 2016) without investigating it further. But there exist also exclusive HS content analyses focusing on e.g., racist speech (Harlow, 2015), gendered HS (Döring and Mohseni, 2020) or HS targeting refugees and immigrants (Paasch-Colberg et al., 2021).

A **hate speech corpus** that satisfies the different needs of computer and communication scientists requires quantity (to be able to learn detection) and granularity (to analyse various facets of HS). Due to the different research questions addressed in communication science, the granularity-focused corpora are usually not published. This reduces the reproducibility and constantly forces researchers to create their own data annotations, which is money and time consuming. The vast majority of published HS corpora thus favour quantity over granularity, which come with various known limitations. Firstly, most HS corpora focus on a binary classification, e.g., *hate* or *non-hate* (Alakrot et al., 2018), whose underlying meaning varies across corpora based on their annotation protocols. Depending on the focus of the HS corpus, the annotated classes vary greatly (Vidgen and Derczynski, 2021), ranging from: person-directed abuse (e.g., cyber bullying) (Wulczyn et al., 2017; Sprugnoli et al., 2018) to group-directed abuse such as sexism (Jha and Mamidi, 2017) or racism (Waseem and Hovy, 2016; Sigurbergsson and Derczynski, 2019). This diversity of class definitions makes it difficult to effectively combine corpora to train classifiers that generalise well across similar HS tasks (Ruiter et al., 2019; Bose et al., 2021). Further, the binarisation (e.g., *sexist/not-sexist*) of HS phenomena often leads to classifiers that are unreliable and/or biased (Wiegand et al., 2019). More recent corpora try to overcome this limitation by creating tasks of higher granularity, focusing on multi-class tasks which may describe the target type (group vs. individual) or intensity of the abuse (Ousidhoum et al., 2019). Basile et al. (2019) also annotate the aggressiveness of the abuse, focusing on migrants and women. The multi-class approach with a focus on migration makes this corpus the closest to our work. Overall there is a trend towards more complex annotations, but most approaches (including Basile et al. (2019)) still attempt to make judgements about what constitutes hate, which stands in contrast to the complex and subjective nature of HS.

We overcome the difficulty of objectively defining HS by moving beyond judgements of whether a statement is hateful or not. With the content-analytical approach in mind, we focus on annotating HS-related features, a procedure similar to the one of Paasch-Colberg et al. (2021). Further, the M-Phasis corpus is based on user content posted on a variety of mainstream and fringe media platforms in two languages: French (FR) (Chung et al., 2019; Ousidhoum et al., 2019) and German (DE) (Bretschneider and Peters, 2017; Struß et al., 2019; Mandl et al., 2019).

3. The M-Phasis Corpus

3.1. Dataset Collection

Choice of Outlets Instead of focusing on a single data source such as Twitter, we keep our sources diverse by focusing on user content posted in comment sections of several popular news outlets in France and

Germany. To cover a broad political spectrum, we focus on four mainstream news outlets and two/three fringe media outlets in France and Germany, respectively. Concretely, the French data is collected from mainstream outlets *France Info* (*fi*), *Le Figaro* (*lf*), *Le Monde* (*lm*), *Valeurs Actuelles* (*va*) and fringe media *AgoraVox* (*av*), *Riposte Laïque* (*rl*), while the German data stems from mainstream *Tagesschau* (*ts*), *Welt* (*we*), *Zeit* (*ze*), *Focus* (*fo*) and fringe *Compact* (*co*), *Epoch Times* (*et*), *Junge Freiheit* (*jf*). We also ensure that different outlet types are covered, i.e., daily (*lf, lm, we*) and weekly newspapers (*fo, va, ze*) and a public television channel (*fi, ts*). Most outlets have an equivalent in Germany/France with regard to e.g., the type of news source or political stance.

Collection Method and Time Frame The M-Phasis corpus consists of articles and their comment threads. To identify articles on our chosen news articles which are relevant to the topic of migration, we create a list of 8 (FR) or 9 (DE) migration-related keywords (see Appendix). Note that these keywords are only related to migration and not related to HS (i.e., no slurs), and thus leave room for the collection of implicitly hateful comments. We implement a crawler that searches through the outlets' web pages and retrieves an article and its complete comment thread if: 1) the article content matches with one of the migration-related keywords; 2) was published between January 2020 and May 2020; 3) contains at least five comments in its comment thread. We limit the collected comments in the comment thread of a single article to the 100 most recent comments at the time of crawling. Apart from the text, we also collect meta data, i.e., the outlet, title, subtitle, date of publication and author of an article as well as the username and date of a user comment. Each article and comment is assigned a unique ID. Since we reconstruct the hierarchical nature of articles and their comment threads, we also save the ID of the direct parent, i.e., article or other user comment, to which a comment is a direct response. The final data sizes per outlet and language are presented in Table 1.

Privacy and Copy Right Laws To conceal the identity of a user abiding to data privacy laws (GDPR), user names are anonymised using internal user IDs and we do not retain a mapping of the user names to the user IDs. To abide copy right laws, we do not publish the textual content of articles, which are replaced by URLs that point to the corresponding web pages.

3.2. Annotation

Corpus Structure To study different types of hate in user comments without an a-priori definition of HS, we develop an annotation protocol that includes various facets of how hate can be communicated in user comments. There are two units of analysis in the corpus: the article and the corresponding comments in its comment thread.

On the **article** level, we capture the type of news pieces (*fact* or *opinion oriented*), the topic as well as the first three mentioned main agents (e.g., *politicians, migrants, organisations* etc.).

For **comments**, the classes differ between moderation or user comments. For moderation comments (i.e., written by moderators), only the type of moderating action (e.g., *deletion*[1], *referral to the netiquette*, etc.) is annotated. Note that we only have access to publicly available data, thus the proportion of hate comments collected in news outlets with strict moderation policies is affected. For all user comments, we annotate the topic, potential (*agreeing* or *disagreeing*) references to its parent and the use of amplifiers (i.e., stylistic reinforcing elements). The centrepiece of the corpus is the annotation of HS-related phenomena. Instead of giving annotators a definition of HS, we focus on HS features which in their combination become descriptors of hateful content, described below.

Hate Speech Features HS features are annotated across five modules (Figure 1), each containing 1–7 categories (i.e., *questions*), which can each have several classes (i.e., *answers*). We present these together with an example instance: *Keine Migranten mehr aufnehmen. Wir haben genügend eigene Sorgen.* (No more migrants. We have enough worries of our own.):

- `c_ne`: **Negative evaluation** of an agent, e.g., *migrant, politician*; generalisation-level of agent, i.e., *individual* vs. *group*; whether the evaluation is *explicit* or *implicit*; reason for the evaluation, e.g., *hypocrisy, ignorance, financial burden*; the *victim(s)* of the behaviour of the agent; use of *irony* or *swearwords*. Here in the example: *no negative evaluation.*
- `c_pe`: **Positive evaluation** of an agent. The categories are analogous to `c_ne`. Here: *none.*
- `c_act`: **Recommendation of an action** or behaviour, e.g., *adaption, elimination*; *explicitness* of the recommendation; the *agent* suggested to perform the action and its level of *generalisation*; the *victim(s)* of the action and their level of *generalisation*. Here: *a recommendation to treat migrants (victims), as a group, negatively but violence-free. The agent is unclear.*.
- `c_contr`: **Contrasting** between an in- and outgroup, e.g., *elite vs. the people*. Here: *migrants vs. German population*.
- `c_emo`: Expression of a *negative* or *positive* **emotion**, the trigger of the emotion, e.g., *migrants, media* and its level of *generalisation*; whether the emotion is expressed via *sarcasm*. Here: *no expression of emotion*.

[1]When a comment has been deleted but still shows on the webpage as a comment with the text removed (e.g., *This comment has been deleted.*), this counts as a moderation comment with moderation action *deletion*.

Type	Outlet (FR)	#Art.	#Com.	Outlet (DE)	#Art.	#Com.
Mainstream	France Info (*fi*)	20	618	Tagesschau (*ts*)	13	1,020
	Figaro (*lf*)	19	1,056	Welt (*we*)	74	736
	Le Monde (*lm*)	16	554	Zeit (*ze*)	15	999
	Valeurs Actuelles (*va*)	30	614	Focus (*fo*)	13	962
Fringe	AgoraVox (*av*)	11	369	Compact (*co*)	12	282
	Riposte Laïque (*rl*)	35	1,435	Epoch Times (*et*)	2	75
	–	–	–	Junge Freiheit (*jf*)	55	747
Total	–	131	4,646	–	187	4,821

Table 1: Overview of number of news articles (#Art.) and comments (#Com.) collected from both mainstream and fringe media outlets in French (FR) and German (DE) for inclusion in the M-Phasis dataset.

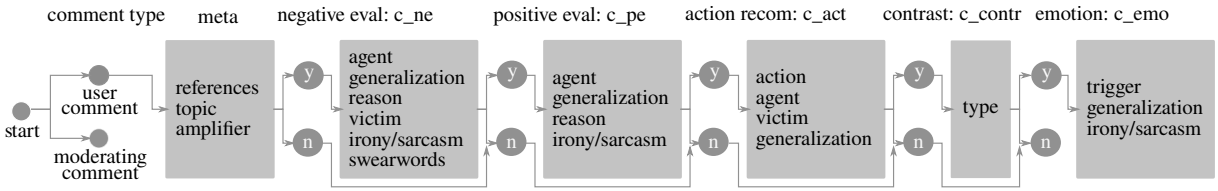

Figure 1: Annotation pipeline for HS features in user comments across five modules (`c_ne`, `c_pe`, `c_act`, `c_contr`, `c_emo`). When a comment fullfills the requirement (*y*) of a module (e.g., contains negative evaluation for `c_ne`), follow-up categories (gray boxes) are annotated, otherwise (*n*) we skip to the next module.

Multiple annotations of a single module are possible e.g., if an annotator wants to annotate several negative evaluations. The HS features above can be combined to create use-case-specific definitions of hate/negativity. For example, the `c_ne` module can be used to focus on explicit negative evaluations of groups to describe *explicit HS*, implicit negative evaluations with a reason for *critical comments* or explicit negative evaluations of individuals to focus on *cyber bullying*. Positive evaluations (`c_pe`) of controversial groups can also be signs of *HS* or *radicalisation*, and the recommendation of actions such as killing individuals or groups (`c_act`) is relevant for identifying HS content illegal in some countries (e.g., according to NetzDG in Germany).
We give more sample annotations and an overview of all annotated modules with their corresponding categories and classes in the Appendix.

Annotators and Annotations We recruited four annotators per country. They are native speakers of French/German interested in studying HS. They were paid the standard monthly salary of research assistants in Germany/France. The annotators went through an extensive training period. Each instance is annotated by a single annotator and the annotations were performed using HUMAN (Wolf et al., 2020).
At the end of the annotation process, we selected 100 user comments in French and German respectively, which were then annotated by two annotators each to calculate **inter-annotator agreement** using Brennan and Prediger's Kappa (κ) (Brennan and Prediger, 1981). κ is calculated for each category, where each individual class per category is treated as a binary *yes/no* decision. This makes it possible to calculate the agreement when classes are not mutually exclusive. Over all categories, we observe high levels of agreement, with all categories being within a reasonable range of $0.77 \leq \kappa \leq 1$. We report the inter-annotator agreement values for all categories in the Appendix.

4. Task-Specific Classification Baselines

4.1. Experimental Setup

To provide first insights into the M-Phasis dataset, we train several baseline models on a number of classification tasks that are based on a subset of classes and categories of the M-Phasis dataset and analyse their performance and limitations. We focus on two classification tasks, namely task E (i.e., *Evaluation* of agents; based on module `c_ne`) and task A (i.e., *Action Recommendation*; module `c_act`), which are sentiment-related tasks. Each task is divided into 5 sub-tasks to replicate the structure of the M-Phasis corpus that is based on gradually more in-depth follow-up questions per module. Task {E|A} have similar structures and are divided into {E|A}-1 (*Does the comment contain a* {*negative or positive evaluation*|*action recommendation*}*?*), {E|A}-2 (*Is the* {*evaluation*|*action recommendation*} *implicit or explicit?*), {E|A}-3a (*Who is the target of the* {*evaluation*|*action recommendation*}*?*), {E|A}-3b (*What is the* {*behaviour of* | *action recommended to*} *the target?*), {E|A}-3c (*Who is the* {|*suggested*} *victim of the behaviour?*). We provide a single train(ing), dev(elopment) and test(ing) split.

Data Taking into account the sparsity of most categories in the original M-Phasis dataset, we create our **task-based dataset** using only those modules which

LA	SP	E-1	E-2	E-3a	E-3b	E-3c	A-1	A-2	A-3a	A-3b	A-3c
DE	Train	2,806	1,931	1,931	1,794	1,078	2,806	624	624	624	624
DE	Dev	500	351	351	320	184	500	114	114	114	114
DE	Test	1,000	681	681	632	365	1,000	225	225	225	225
DE	D_{KL}	0.36_3	0.07_2	0.20_5	0.25_9	0.36_6	0.16_2	0.21_2	0.32_6	0.56_5	0.51_6
FR	Train	2,178	1,741	1,741	1,584	1,323	2,178	680	680	680	680
FR	Dev	500	409	409	382	206	500	327	327	327	327
FR	Test	1,000	795	795	719	607	1,000	327	327	327	327
FR	D_{KL}	0.48_3	0.07_2	0.24_5	0.19_9	0.36_6	0.07_2	0.04_2	0.11_6	0.33_5	0.09_6

Table 2: Number of instances within each sub-task (E, A) in the train, dev and test splits (SP) of the German (DE) and French (FR) language (LA) corpora. The class imbalance per sub-task is given via the Kullback-Leibler divergence (D_{KL}) between the sub-task class distribution of $_c$ classes and a perfectly balanced class distribution.

contain sufficient labelled data. This resulted in using the c_{ne|pe} module for task E and c_act for task A. To avoid strong class imbalances within each sub-task, we clustered classes from the original set of classes together which were similar in their underlying meaning and which were sparse in their respective number of instances. We give a more detailed listing of the mapping between the original classes to the sub-tasks' classes in the Appendix. We remove URLs from the text and replace them with a special token, i.e., [URL]. We randomly sample 1,000 and 500 instances from the corpus as the test and dev splits respectively. The remaining 2,806 (DE) or 2,178 (FR) instances are used for the training data. Sub-tasks {E|A}-3{a|b|c} are lower-resourced than sub-tasks {E|A}-{1|2}, due to their higher number of classes and the smaller amount of available annotations. For each sub-task, the number of instances of the DE/FR train-dev-test splits and the number of classes are given in Table 2. To give insight into the class imbalance per sub-task, we also report the Kullback-Leibler divergence (D_{KL}) between the class distribution of a sub-task and a perfectly balanced class distribution. A rather balanced class distribution would thus lead to a D_{KL} close to 0.

Model Specifications and Evaluation Our baseline models (B) are transformer-based classifiers as implemented in the `transformers` library.[2] Specifically, we use `bert-base-german-cased` (DE) and `camembert-base` (Martin et al., 2020) (FR). To explore whether domain knowledge can be inserted into the models via intermediate masked-language model (MLM) training, we also fine-tune both language models on their respective DE or FR task-based training data for 20 epochs using the MLM objective to obtain task-tuned language models (B+T). We also explore whether the annotations in the German and French data are sufficiently consistent amongst each other to enable a bilingual learning that improves the classification performance in comparison to a monolingual model. Therefore, analogous to B+T, we fine-tune a multilingual model `bert-base-multilingual-cased` on the concatenation of the German and French training data using the MLM objective (M+T) and then learn classification jointly (M+T(J)) or separately (M+T(S)) on the German and French sub-tasks. All classification models are run over 10 seeded runs with early stopping ($\delta = 0.01$, $\text{patience} = 5$) and we report their average Macro F1 on the test set together with standard mean error. For the domain analysis we use the multilingual universal sentence encoder (Yang et al., 2020) to embed user comments, as it works well on semantic similarity tasks (Cer et al., 2018).

[2] https://github.com/huggingface/transformers

4.2. Results

Performing task-based **intermediate MLM fine-tuning** (B+T) leads to limited improvements over the monolingual baselines (B), with improvements up to $+2.9$ (DE, A-3c) on the German data (Table 3). All improvements are seen on the target-victim sub-tasks {E|A}-3{a|b|c}. Task domain knowledge acquired by the intermediate MLM training is thus mostly useful for the lower-resourced sub-tasks. For French, most tasks show no significant difference.

The **multilingual** baselines (M+T) are by far outperformed by their monolingual (B+T) counterparts. The training on both the French and German data jointly (M+T(J)) leads to some significant improvements on the more complex E-3{a|b|c} sub-tasks in comparison to the multilingual model which was trained on French or German separately (M+T(S)), indicating that there is a sufficient overlap in the French and German annotations such that the lower-resourced sub-tasks benefit from the joint learning; the gain of additional samples outweighing the loss obtained by a few noisy samples.

Overall, we observe low F1 scores across all tasks. This underlines the difficulty of the tasks, which is mostly due to the small amount of samples and sparseness of minority classes, especially for the more complex sub-tasks. Methods focusing on low-resource classification (Hedderich et al., 2021) should be explored to overcome the sparsity in the corpus. We give a more detailed account on the error sources in Section 5.

LA	CM	E-1	E-2	E-3a	E-3b	E-3c	A-1	A-2	A-3a	A-3b	A-3c
DE	B	55.6±.5	58.7±.4	49.2±.4	27.8±.9	35.2±.4	**72.3±.4**	56.2±1	**31.0±.8**	30.8±.5	33.1±.4
DE	B+T	55.0±.2	58.6±.4	**51.6±.6**	**29.9±.8**	35.4±.3	71.3±.4	57.8±.9	28.9±.9	30.8±1	**36.0±.5**
DE	M+T(S)	48.3±.5	52.4±2	45.9±.5	23.4±.4	32.1±2	65.1±3	52.3±1	28.9±.9	28.7±2	28.2±.7
DE	M+T(J)	49.0±1	48.1±4	47.5±.4	23.6±.4	34.9±.7	64.1±2	49.5±2	30.7±1	26.8±.7	28.4±.8
FR	B	59.3±.7	63.3±.4	54.1±.5	32.9±.3	**39.0±.3**	66.9±.5	53.7±.8	40.4±.5	42.1±.6	40.8±.6
FR	B+T	59.6±.3	63.4±.3	53.4±.4	33.5±.3	37.1±.6	67.6±.3	53.2±.4	41.1±.5	**43.8±.7**	40.1±.7
FR	M+T(S)	50.3±1	58.8±.5	44.3±.6	23.1±3	32.7±.4	60.2±2	51.2±1	34.5±.8	32.1±.7	34.2±.9
FR	M+T(J)	49.2±.8	49.0±3	45.3±.6	28.0±.4	33.5±.4	51.4±4	52.3±2	36.9±.6	30.6±2	34.4±.6

Table 3: Average Macro F1 of different classification models CM for language LA on the relevant sub-tasks (E,A) test sets. Standard mean errors given as bounds. Top scores outside of the error bounds of other models in **bold**.

5. Qualitative Error Analysis

To further identify shortcomings of the baseline models and difficulties related to the corpus structure, we perform a qualitative error analysis. We focus on the two best models in DE (B) and FR (B+T) on task E-1, as this task focuses on positive/negative evaluations of agents and is thus not far from the popular sentiment analysis task. To this end, we have sampled 100 instances from the DE and FR test set predictions and annotated specific error types (Table 4).

On the German side, the most common error stems from comments without an evaluation but which were classified as containing a negative evaluation (i.e., *over-blacklisting*), which was prevalent in 18% of instances. The most common causes for over-blacklisting are i) naming of countries or places (5%; EX-1), ii) naming of people (especially politicians; 3%) or iii) other trigger words (e.g., *Nazi, Politiker* (politician); 4%; EX-2). This is due to the **topical bias** in the M-Phasis corpus. Its focus is on the topic of migration, which is ensured by selecting news articles based on migration-related keywords. This enables the inclusion of comments containing implicit and explicit forms of hate, as well as positive sentiments. However, due to this topical focus, politicians are frequent recipients of negative evaluations (Section 6), and thus the classifier mistakenly learned to equate the appearance of political actors with a negative sentiment. While topical bias is not uncommon in HS corpora (Wiegand et al., 2019), it should be taken into account when using this data to train models, especially those going into production.

A negative evaluation being ignored by the classifier (i.e., classified as *no evaluation*) is the second most common error (6%). Mistakes in the annotations are one reason, e.g., in cases where a negative action recommendation was mistakenly annotated as a negative evaluation (EX-3). Denoising or similar techniques can be used to mitigate the effects of **noise** in the annotations. Another source of error stems from the models, which only allow to attribute a single label to each instance. However, in some cases several actors are annotated in the original M-Phasis corpus with varying evaluations. A **multi-label** classifier could be used to model this complexity. Lastly, when the negative evaluation is too **implicit or dependent on context**, the classifier was not able to detect it (2%; EX-4). The annotators were always shown the context of a given comment (e.g., the article or comment to which the current instance is referring to), which was ignored by our classifiers. Including this contextual information may improve the classification of implicit evaluations.

On the French side, we observe much less cases of over-blacklisting (2%), while the prevalence of ignoring negative evaluations is the same as for the German model (6%). The reduced prevalence of over-blacklisting might be due to the larger proportion of fringe media content in the French corpus (44.5% vs. 22.8% in Germany), thus reducing the amount of neutral/informative content to be mistakenly black-listed.

6. Target Analysis

One important set of features of the M-Phasis corpus are the target annotations. This includes the annotation of positive and negative evaluations of targets in user comments (i.e., `c_{ne|pe}` or task E in Section 4). Concretely, a target (*agent*) is evaluated by a user based on their actions (*evaluation*) which have caused harm (or benefit) to a third party (*victim*). Analogous to these agent-evaluation-victim triples, we also obtain agent-action-victim triples (i.e., `c_act` or task A), where a user suggests that the agent performs an action (*action*) under which a victim should suffer. We explore some of the main trends found in the triple annotations.

For both the German and the French portions of the M-Phasis corpus, politicians ($2.5k$ (DE) /$4.4k$ (FR) mentions)[3] and migrants ($1k$/$1.7k$) were the most common targets of **negative evaluations**. On the German portion (Figure 2), the most common mentions of political agents with a negative evaluation were *EU* and *Merkel*. Indicating a negative sentiment towards the current government and its handling of the topic of migration. The most frequent negatively evaluated action of these political agents on the German side was *passivity* (712 times). The two major mentioned victims of the behaviour of politicians are, by a large part

[3]Note that these numbers are reduced in Figures 2 and 3, as they only show the most frequent evaluations/actions.

EX	Instance	Type
1	*Es gibt die ersten Verdachtsfälle in Äthiopien. [...]* (There are some first suspected cases in Ehtiopia. [...])	$\emptyset \rightarrow N$
2	*Der Berufswunsch dieses jungen Mannes: Politiker! Mehr ist dazu nicht zu sagen.* (The career aspiration of this young lad: politician! Nothing more to say about this.)	$\emptyset \rightarrow N$
3	*Man muss sie registrieren (eindeutig, Fingerabdrücke etc!), und Versorgung/Sozialleistungen gibt's nur am registrierten Ort. Punkt.* (They need to be registered (unambiguously, finger prints etc!), and aid/social benefits only at the registered location. Done.)	$N \rightarrow \emptyset$
4	*Chouette 2 de moins.* (Cool 2 less.)	$N \rightarrow \emptyset$

Table 4: Example instances (EX) from the DE and FR task E-1 test set with the error type (*reference* $\rightarrow$ *predicted*) of the best performing classification models in DE (B) and FR (B+T). Classes: *none* ($\emptyset$), *negative* (N).

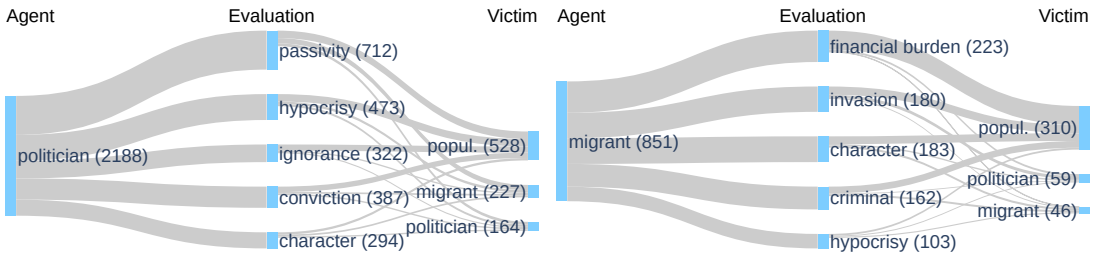

Figure 2: Agent-Evaluation-Victim sankeys for the two most common agents, *politician* (left) and *migrant* (right), in the German portion of the M-Phasis dataset. We show the 5/3 most common evaluations/victims respectively.

(589 times) the German population or, to a smaller extend (299), migrants. When migrants are the recipients of negative evaluations, the action they are most frequently accused of are being a *financial burden* (223), with the population being the most frequently mentioned victim (353). Nevertheless, the fact that politicians are by far more frequently negatively evaluated than migrants on the topic of migration, shows that negative sentiments tend to be directed to the decision makers of migration-related policies. This blaming reflects the notion that migrants also suffer from these policies (as a frequent victim group of politicians).

The two most frequent agents to be addressed in an **action recommendation** by the German users (Figure 3 (left)) are either political actors (792) or the German population (289). German politicians are most frequently called to force foreigners to adapt to German society (240) and when they are called to treat someone negatively, the largest victim group are mostly foreign political entities (*EU*, *Türkei* (Turkey), *Griechenland* (Greece)). When politicians are called to treat someone positively, the suggested beneficiary are most frequently migrants (78). Similar trends are also found in French user comments. While the German calls for action tend to be more moderate, the action recommendations in France (Figure 3 (right)) include more radical actions such as physical violence towards (74) or elimination of (29) foreigners. It is unclear whether this more radical manifestation of hate towards migrants is due to an increased societal radicalisation or a difference in the data, where a) German comments are more likely to be published in mainstream media compared to French comments (Section 7), or b) the German news outlets are more pro-active in their moderation strategies, deleting more radical comments before we could collect them for the M-Phasis corpus.

TS	CM	E-1	E-2	E-3a	E-3b	E-3c
M	B	53.8±.2	56.7±.3	**49.7±.5**	**26.3±.8**	**35.1±.6**
F	B	**57.1±.9**	**61.0±1**	46.6±1	24.7±.1	21.6±2
M	B+T	**53.8±1**	56.8±.2	49.8±.6	**26.6±.7**	**35.2±.5**
F	B+T	51.7±1	**59.3±1**	50.9±.7	22.4±.8	17.4±.8

Table 5: Average Macro F1 of German classification models CM tested on the mainstream (M) or fringe (F) test set (TS) of the relevant sub-tasks (E). Standard mean errors given as bounds. Top scores outside of the error bounds of other models in **bold**.

The sparsity of most triples makes their prediction using classification models especially difficult, which can be observed in the generally low F1 scores of the baselines on sub-tasks {E|A}-3{a|b|c} (Table 3).

7. Domain Analysis

The M-Phasis corpus contains user comments from various news outlets. As these outlets differ in political orientation, their user comments also tend to differ in style. This can lead to domain differences across instances, which can affect the performance of classifiers trained on the M-Phasis dataset. To quantify this domain difference between user comments, we generate embeddings on the concatenation of all user comments belonging to a single news outlet using universal sen-

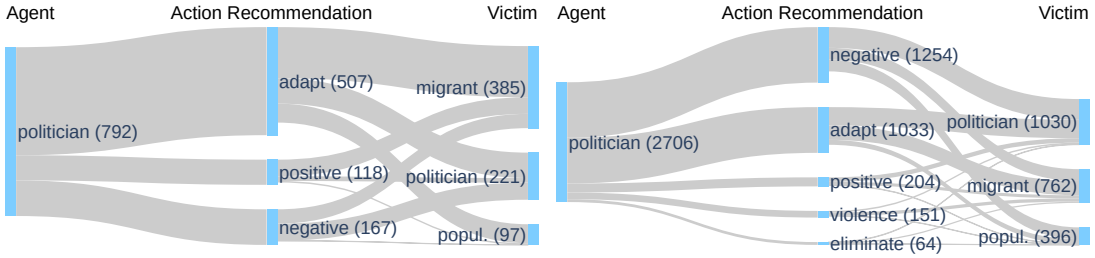

Figure 3: Agent-Action-Victim sankeys for the most common agent (*politician*) in the German (left) and French (right) portions of the M-Phasis dataset. We show up to 5/3 of the most common evaluations/victims respectively.

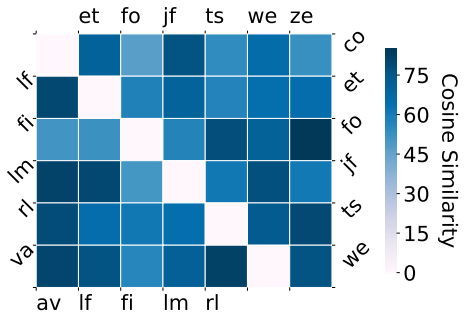

Figure 4: Cosine similarity between universal sentence encoder embeddings of user comments in DE (top-right triangle) or FR (bottom-left triangle) news outlets.

tence encoder. For each language, we calculate the cosine similarity between the embeddings (Figure 4).

There is a strong correlation between the manual categorisation of news outlets as either *mainstream* or *fringe* and the similarity between user comments of these two types of media outlets. Specifically, user comments posted in mainstream media such as *Focus*, *Tagesschau*, *Welt* and *Zeit* are closer to each other than to fringe media such as *Compact*, *Epoch Times* or *Junge Freiheit*. Performing K-Means clustering over the German document embeddings ($k = 2$) yields exactly the same divide between user comments of mainstream and fringe media. For French, there is a three-step divide ($k = 3$), where comments under news outlets are clustered into fringe (*Agoravox*, *Riposte Laïque*, *Valeurs Actuelles*), mainstream (*Le Figaro*, *Le Monde*) and intellectual (*France Info*).

For the German data 77.2% of comments are from mainstream media, while for the French side the domains are more balanced, with 55.5% of comments stemming from either mainstream or intellectual media. To quantify the effect of this domain imbalance on the German data, we evaluate the German B and B+T models on the subset of instances in the test sets that stem from 1) mainstream or 2) fringe media outlets. We focus this analysis on task E, as it contains more samples across its sub-tasks than task A (Table 5).

For the sparse multi-class sub-tasks E-3{a|b|c}, the performance on the more data-rich mainstream comments is comparatively higher (B), underlining the fact that the domain differences in the M-Phasis corpus are especially to be taken into consideration when working with data sparse classification tasks. With intermediate MLM training (B+T) the macro F1 performance on fringe comments drops across most tasks in comparison to the performance without intermediate MLM training (B), indicating that the domain-imbalance on the German data of the M-Phasis corpus was transmitted into the representations of the underlying BERT model, leading to a lower classification performance on the under-represented fringe domain.

8. Conclusion and Future Work

We present M-Phasis, a corpus of $\sim$ $9k$ German and French comments collected from migration-related news articles. While most existing HS corpora rely on an ad-hoc definition of HS, which ignores the complex nature of hate online, the M-Phasis corpus does not attempt to judge whether a user comment is hateful or not. Instead, it focuses on a total of 23 HS-related features, which in their combination become descriptors of various types of hateful content. We discuss baseline results on several sub-tasks created from the M-Phasis corpus together with a qualitative error analysis. We analyse evaluations and action recommendations and quantify the domain differences between comments of different sources included in the M-Phasis corpus.

The M-Phasis corpus leaves room for various types of analysis. Comments are collected with their context, thus the relation and flow of information between instances can be analysed. The feature-based approach allows for analysis on correlations between different HS features. Comments stem from various news outlets and make a cross-media analysis possible. The bilingual nature of the corpus also allows for a cross-cultural study of HS phenomena in France and Germany.

The M-Phasis corpus, the train-dev-test splits, model outputs and annotation protocol are made public under `https://github.com/uds-lsv/mphasis`.

Acknowledgments

We thank our annotators for their keen work. The project on which this paper is based is funded by the DFG (WI 4204/3-1) and ANR (ANR-18-FRAL-0005).

9. Bibliographical References

Alakrot, A., Murray, L., and Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181. Arabic Computational Linguistics.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Berelson, B. (1952). *Content Analysis in Communication Research*. Foundations of communication research. Free Press.

Bose, T., Illina, I., and Fohr, D. (2021). Unsupervised domain adaptation in cross-corpora abusive language detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 113–122, Online, June. Association for Computational Linguistics.

Brennan, R. L. and Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(3):687–699.

Bretschneider, U. and Peters, R. (2017). Detecting offensive statements towards foreigners in social media. In *HICSS*.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *CoRR*, abs/1803.11175.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, July. Association for Computational Linguistics.

Coe, K., Kenski, K., and Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679.

Cohen-Almagor, R. (2018). Taking north american white supremacist groups seriously: The scope and the challenge of hate speech on the internet. *International Journal for Crime, Justice and Social Democracy*, 7(2):38–57, Jun.

Döring, N. and Mohseni, M. R. (2020). Gendered hate speech in youtube and younow comments: Results of two content analyses. *SCM Studies in Communication and Media*, 9(1):62–88.

Erjavec, K. and Kovačič, M. P. (2012). "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6):899–920.

Harlow, S. (2015). Story-chatterers stirring up hate: Racist discourse in reader comments on u.s. newspaper websites. *Howard Journal Of Communications*, 26:21–42, 01.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online, June. Association for Computational Linguistics.

Jha, A. and Mamidi, R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada, August. Association for Computational Linguistics.

Johnson, N., Leahy, R., Restrepo, N., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., and Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573:1–5, 09.

Jurgens, D., Hemphill, L., and Chandrasekharan, E. (2019). A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy, July. Association for Computational Linguistics.

Kovács, G., Alonso, P., and Saini, R. (2021). Challenges of hate speech detection in social media. *SN Computer Science*, 2, 04.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July. Association for Computational Linguistics.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong,

China, November. Association for Computational Linguistics.

Paasch-Colberg, S., Strippel, C., Trebbe, J., and Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, 9(1):171–180.

Paltoglou, G., Theunis, M., Kappas, A., and Thelwall, M. (2013). Predicting emotional responses to long informal text. *Affective Computing, IEEE Transactions on*, 4:106–115, 01.

Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., and Martín-Valdivia, M. T. (2021). A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Robert, F., Ashar, A., Gasser, U., and Joo, D. (2016). Understanding harmful speech online. *Berkman Klein Center for Internet & Society Research Publication*.

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, page 0093650220921314.

Rost, K., Stahel, L., and Frey, B. S. (2016). Digital social norm enforcement: Online firestorms in social media. *PLOS ONE*, 11(6):1–26, 06.

Ruiter, D., Rahman, M. A., and Klakow, D. (2019). Lsv-uds at HASOC 2019: The problem of defining hate. In Parth Mehta, et al., editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 263–270. CEUR-WS.org.

Saleem, H. M., Dillon, K. P., Benesch, S., and Ruths, D. (2016). A web of hate: tackling hateful speech in online social spaces. In *First Workshop on text Analytics for Cybersecurity and Online Safety at LREC 2016*.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics.

Sigurbergsson, G. I. and Derczynski, L. (2019). Offensive language and hate speech detection for danish. *CoRR*, abs/1908.04531.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2021). Analyzing the targets of hate in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):687–690, Aug.

Sprugnoli, R., Menini, S., Tonelli, S., Oncini, F., and Piras, E. (2018). Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium, October. Association for Computational Linguistics.

Struß J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., and Klenner, M. (2019). Overview of germeval task 2, 2019 shared task on the identification of offensive language. Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg, pages 352 – 363, München [u.a.]. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg.

Su, L. Y.-F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., and Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the facebook pages of news outlets. *New Media & Society*, 20(10):3678–3699.

Uyheng, J. and Carley, K. (2021). Characterizing network dynamics of online hate communities around the covid-19 pandemic. *Applied Network Science*, 6, 03.

Vidgen, B. and Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32, 12.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Wolf, M., Ruiter, D., D'Sa, A. G., Reiners, L., Alexandersson, J., and Klakow, D. (2020). HUMAN: Hierarchical universal modular ANnotator. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–61, Online, October. Association for Computational Linguistics.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C.,

Sung, Y.-h., Strope, B., and Kurzweil, R. (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online, July. Association for Computational Linguistics.

Ziegele, M., Koehler, C., and Weber, M. (2018). Socially destructive? Effects of negative and hateful user comments on readers' donation behavior toward refugees and homeless persons. *Journal of Broadcasting & Electronic Media*, 62(4):636–653.

Appendix A: List of Keywords

In order to identify articles related to the topic of migration, we compose a list of regular expression keywords related to this topic. The keywords for France and Germany are equivalent in meaning, however, for Germany there exists one additional keyword, since we include both the more modern and politically correct term *geflüchtete** (refugee) and its older counterpart *flüchtling**, which are both the equivalent of the French keyword *réfugié(s)*.

Concretely, the French keywords are: *étrangers (foreigners)*, *immigré(s) (immigrant(s))*, *migrant(s)*, *réfugié(s)*, *demandeur(s) d'asile (seeker(s) of asylum)*, *asile*, *immigration*, *migration*.

The German keywords are: *zuwander** *(immigrant(s))*, *einwander** *(immigrant(s))*, *migrant**, *flüchtling**, *geflüchtete**, *ausländ** *(foreigner(s))*, *asyl** *(asylum or seeker(s) of asylum)*, *immigra** *(immigration or immigrant(s))*, *migration**.

Appendix B: Annotation Overview

We give an overview over the different modules, categories and classes that annotators annotate for each article or comment instance.

Articles are presented to the annotators in the annotation tool without context. Article annotations only have a single module `article`. The first question (*category*) shown to the annotators is `n_2` (Table 8), which is always followed by `n_3`. If the annotator annotates `n_3` $= 0$ (i.e., *migration not a topic*), then the annotation of the article instance is over and the next instance is shown. If any other class is chosen, then the annotator is asked to also annotate `n_4`, where they are asked to choose the first three mentioned agents in the article.

Comments are presented to the annotators together with their direct parent as context, e.g., the news article or another comment to which the current comment is a reply. As we discern between user comments and moderation comments, the first category shown to annotators is `c_usmod` (Table 9). If the annotator chooses `c_usmod` $= 0$ (i.e., *moderating comment*), then the annotation of this comment instance is over and the annotation tool proceeds to the next instance. Otherwise, if the comment is annotated as a *user comment*, the annotation tool continues with all follow-up questions in the `meta` module, i.e., `c_refn` to `c_amp`. Then, the annotation tool enters the `c_ne` module, where negative evaluations are annotated. If the annotator chooses `c_ne_1` $= 0$, it will skip all following categories in the `c_ne` module and jump to the next module `c_pe`. Otherwise, it will proceed to ask all dependent follow-up categories `c_ne_2` to `c_ne_7`. At the end of the `c_ne` module, the annotation tool asks the annotator whether they want to annotate any further negative evaluations. If this is the case, the tool loops back to the beginning of the `c_ne` module. If not, the tool continues to the next module `c_pe`, where positive evaluations are annotated. This module functions analogous to the `c_ne` module, such that it is only traversed if the annotator states that there is a positive evaluation. Again, multiple traversals are also possible. After the `c_pe` module, the annotation tool goes to the `c_act` module. If the first category `c_act`$= 0$, then the tool skips all follow-up categories, otherwise it traverses all categories in the module. Again, several traversals are possible if the annotators choose to annotate several action recommendations. After the `c_act` module, there is a single category `c_contr`, which also allows multiple answers, followed by the last module, `c_emo`. Analogous to previous modules, its dependent categories `c_emo_2a` to `c_emo_3` are only shown if the annotators state that there is an explicit expression of emotions in `c_emo_1`.

Appendix C: Inter-Annotator Agreement

The 100 user comments selected for calculating the inter-annotator agreement were collected from 5 different articles in German and French respectively. Each user comment is annotated by two annotators. We use Brennen annd Prediger's Kappa (κ) and the percentage agreement (agg) to calculate the inter-annotator agreement. The two metrics are calculated for each category, where each individual class per category is treated as a binary *yes/no* decision. Some categories are dependent of other categories, i.e., if an annotator annotates `c_ne_2`$= 1$ (*determinant*), then all following categories `c_ne_`$\{3-7\}$ (*dependants*) are follow-up questions based on the choice taken in `c_ne_2` (e.g., *is the agent a group or an individual?*). Thus, we only compare the annotations of two annotators on these dependent categories, if they share the same determinant annotation. The determinants are always compared to each other. The determinant $\rightarrow$ dependants groups in the annotations are `c_ne_2` $\rightarrow$ `c_ne_`$\{1|3-7\}$, `c_pe_2` $\rightarrow$ `c_pe_`$\{1|3-5\}$ and `c_act_1` $\rightarrow$ `c_act`$(\_\{2a|2b|3a|3b\})$.

The average agreement and κ per category is reported in Table 6.

Appendix D: Sample Annotations

We show one example annotation from each country in the corpus to give an intuition how the different annotation modules and categories are applied.

> Und Dumm-Michel darf diese Migranten finanzieren. Unglaublich!! (And stupid Michel

Category	DE		FR	
	agg	κ	*agg*	κ
c_usmod	1.00	1.0	–	–
c_refn	0.89	0.87	0.83	0.79
c_refc	0.96	0.94	0.93	0.90
c_topic	0.90	0.89	0.91	0.90
c_amp	0.96	0.94	0.88	0.78
c_ne_1	0.96	0.95	0.96	0.95
c_ne_2	0.96	0.95	0.95	0.95
c_ne_3	0.99	0.99	0.99	0.99
c_ne_4	0.99	0.99	0.99	0.99
c_ne_5	0.99	0.99	0.99	0.99
c_ne_6	0.99	0.99	0.99	0.99
c_ne_7	0.99	0.99	0.99	0.99
c_pe_1	0.84	0.80	0.83	0.77
c_pe_2	0.99	0.98	0.98	0.98
c_pe_3	0.99	0.99	1.0	1.0
c_pe_4	0.99	0.99	0.99	0.99
c_pe_5	1.0	1.0	1.0	1.0
c_act	0.96	0.96	0.93	0.93
c_act_1	0.95	0.98	0.92	0.92
c_act_2a	0.98	0.95	0.97	0.97
c_act_2b	0.96	0.95	0.89	0.88
c_act_3a	0.98	0.98	0.97	0.97
c_act_3b	0.93	0.92	0.90	0.89
c_contr	0.96	0.96	0.94	0.93
c_emo_1	0.97	0.97	0.83	0.77
c_emo_2a	0.99	0.98	0.94	0.94
c_emo_2b	0.98	0.97	0.82	0.76
c_emo_3	0.98	0.97	0.83	0.77

Table 6: Average agreement (*agg*) and Brennen and Prediger's Kappa (κ) across all classes in a given category for the German (DE) and French (FR) side.

has to finance these migrants. Incredible!!)

The above sample instance is a German user comment replying to another comment written under an article talking about new migrants arriving in Berlin. It contains an amplifier (*!!*) (i.e.,c_amp= 1). It contains an explicit negative evaluation of migrants as a group (c_ne_1= 1, c_ne_2= 1, c_ne_3= 2), and the reason for the evaluation is them being a financial burden (c_ne_4= 5). The German population (*Dumm-Michel*) is the victim of this behaviour (c_ne_5= 1). There is no sarcasm (c_ne_6= 0) or swearwords (c_ne_7= 0). There is no positive evaluation (c_pe_1= 0), which is why all follow-up categories in the c_pe module are skipped during annotation. Similarly, there is no action recommendation (c_act= 0), contrasting (c_contr= 0) or expression of emotion (c_emo_1= 0).

> Qu'ils les renvoient en Asie. La frontière est proche. (They should send them back to Asia. The border is close.)

The above sample instance is a French user comment referring to an article talking about new refugee camps opening on Lesbos and Chios (Greece). It contains no amplifier (c_amp= 0), no negative evaluation (c_ne_1= 0) and no positive evaluation (c_pe_1= 0). It does contain an explicit action recommendation (c_act= 1), namely a negative violence-free treatment (*Qu'ils les renvoient en Asie.*) (c_act_1= 3). The suggested agent of the action recommendation is unclear (c_act_2a= 99) and the victims are migrants (c_act_3a= 1) as a group (c_act_3b= 2). There is no contrasting (c_contr= 0) or expression of emotion (c_emo_1= 0).

Appendix E: Class Mapping

For the task-specific classification tasks, we select a subset of the M-Phasis categories and their classes. In Table 7 we list the mapping of M-Phasis categories and classes to (sub-)tasks and their classes.

Sub-Task	Class	Class Description	Original Category and Classes
E-1	0	negative	c_ne_1= {1|2}
E-1	1	positive	c_pe_1= {1|2}
E-1	2	none	c_ne_1= 0 & c_ne_0= 0
E-2	0	implicit	c_ne_1= 2
E-2	1	explicit	c_ne_1= 1
E-3a	0	migrant	c_ne_2= {1|111}
E-3a	1	politician	c_ne_2= {2|211 − 225}
E-3a	2	population	c_ne_2= 3
E-3a	3	discussants	c_ne_2= 13
E-3a	4	other	c_ne_2= {4 − 12}
E-3b	0	passivity	c_ne_4= 1
E-3b	1	conspiracy	c_ne_4= 2
E-3b	2	ignorance	c_ne_4= 3
E-3b	3	criminal behavior	c_ne_4= 4
E-3b	4	financial burden	c_ne_4= 5
E-3b	5	incompatibility	c_ne_4= 6
E-3b	6	invasion	c_ne_4= 7
E-3b	7	character trait	c_ne_4= 20
E-3b	8	political conviction	c_ne_4= 30
E-3c	–	–	same mapping as E-3a but with c_ne_5
A-1	0	no	c_act= 0
A-1	1	yes	c_act= {1|2}
A-2	0	implicit	c_act= 1
A-2	1	explicit	c_act= 2
A-3a	–	–	same mapping as E-3a but with c_act_2a
A-3b	0	positive treatment	c_act_1= 1
A-3b	1	adaption	c_act_1= 2
A-3b	2	negative violence-free	c_act_1= 3
A-3b	3	physical violence	c_act_1= 4
A-3b	4	elimination	c_act_1= 5
A-3c	–	–	same mapping as E-3a but with c_act_3a

Table 7: The mapping of sub-task classes to their corresponding original category and class(es).

Module	Category	Description	Class Code	Class Description	#Samples (DE)	#Samples (FR)
article	n_2	type of news piece	1	emphasizing facts	3,626	3,030
			2	emphasizing an opinion	1,857	1,960
	n_3	topic of news piece	0	migration not a topic	571	917
			1	management of immigration	2,370	1,346
			2	security and safety	873	888
			3	justice	63	0
			4	integration and cohabitation	22	527
			5	culture and religion	131	74
			6	education	0	0
			7	labor market and economy	530	105
			8	social issues	371	113
			9	health aspects	93	178
			10	environment	14	0
			11	media coverage on migration	246	0
			99	cannot tell	219	842
	n_4	mentioned agents	1	migrants	4,538	4,724
			111	residents of other countries	369	400
			2	in the area of politics	250	307
			211	CDU/CSU - LR	1,135	17
			212	SPD - PS	591	8
			213	Bündnis 90/Die Grünen - Les Verts	396	0
			214	Left-wing politicians	107	61
			215	FDP - En Marche!	0	0
			216	AfD - RN/FN	833	773
			217	government	833	773
			218	opposition	0	0
			219	left-wing political camp	11	0
			220	right-wing political camp	52	0
			221	left-wing extremists	45	25
			222	right-wing extremists	911	0
			223	political and public institutions	5,478	1,669
			224	states	3,325	670
			225	foreign politician/party/government	1,527	2,188
			3	the German/French population	580	403
			4	media	406	865
			5	civil society actor	1,294	3,526
			6	religious actors	452	232
			7	scientific actors	68	32
			8	police	541	798
			9	courts	568	343
			10	military	244	0
			12	abstract entities (values, practices etc.)	0	0
			13	discussant	0	0
			99	cannot tell	193	86

Table 8: Annotation modules for articles and their respective categories. Labels for each category are given with the corresponding number of German (DE) and French (FR) user comments that are part of the comment thread of an article with the given label.

Module	Category	Description	Label Code	Label Description	#Samples (DE)	#Samples (FR)
meta	c_usmod	type of comment	0	moderating comment	76	27
			1	user comment	4,745	3,910
	c_refn	reference to news article	0	makes no reference to news article	874	1,079
			1	approval of the article	132	118
			2	refusal of the article	438	252
			3	ambivalent	121	42
			4	establishes reference, without evaluating it	3,102	2,377
			99	cannot tell	79	49
	c_refc	reference to comment	0	does not refer to another comment	2,757	2,375
			1	agreement	829	863
			2	disagreement	1,160	677
			99	cannot tell	0	1,075
	c_topic	topic of comment	same as n_3			
	c_amp	amplifier	0	no	4,230	3,338
			1	yes	512	576
			99	cannot tell	4	1
c_ne (negative evaluations)	c_ne_1	negative evaluation	0	no	1,351	744
			1	yes, explicit	2,148	2,093
			2	yes, implicit	1,083	1,024
			99	cannot tell	164	1,077
	c_ne_2	agent		same as n_4		
	c_ne_3	level of generalization	1	case-specific	997	1,328
			2	generalized entity	2,218	1,808
			99	cannot tell	180	33
	c_ne_4	reason for the evaluation	1	passivity	604	604
			2	conspiracy or hypocrisy	778	604
			3	ignorance	378	301
			4	criminal behaviour	243	309
			5	financial burden	148 6 143	
			6	incompatibility	23	117
			7	invasion	75	185
			8	illness	12	39
			20	character traits	350	289
			30	political conviction	440	151
			99	cannot tell	344	271
	c_ne_5	victim of behavior		same as n_4		
	c_ne_6	irony or sarcasm		same as c_amp		
	c_ne_7	swearwords		same as c_amp		
c_pe (positive evaluation)	c_pe_1	positive evaluation		same as c_ne_1		
	c_pe_2	agent		same as c_ne_2		
	c_pe_3	level of generalization		same as c_ne_3		
	c_pe_4	reason for the evaluation	1	efficiency	118	248
			2	honesty	30	32
			3	seeing things through	77	111
			4	exemplary behavior	69	83
			5	financial advantages	17	19
			6	cultural enrichment	2	52
			20	character traits	30	64
			30	political conviction	62	40
			99	cannot tell	28	44
	c_pe_5	irony or sarcasm		same as c_ne_6		
c_act (action)	c_act	action recommendation	0	no action	3,708	2,647
			1	explicit action	834	817
			2	implicit action	203	451
	c_act_1	action	1	positive treatment	172	100
			2	call for change/adaption	591	446
			3	negative but violence free treatment	229	552
			4	physical violence	15	60
			5	elimination/killing	8	90
			99	cannot tell	22	20
	c_act_2a	agent		same as n_4		
	c_act_2b	level of generalization		c_ne_3		
	c_act_3a	victim		same as n_4		
	c_act_3b	level of generalization		c_ne_3		
c_contr (contrasting)	c_contr	contrasted groups	0	none	3,760	2,030
			1	elite vs. the people	191	638
			2	globalism vs. states	5	41
			3	right-wing vs. left-wing camps	40	58
			4	less advantaged citizens vs. migrants	24	28
			5	french vs. migrants	1	483
			6	germans vs. migrants	134	4
			7	french vs. other political actors abroad	1	88
			8	germans vs. other political actors abroad	134	0
			9	europeans/westerners vs. others	75	186
			10	pro-migrants vs. anti-migrants	46	95
			11	good migrants vs. bad migrants	55	10
			12	present vs. past	76	28
			99	cannot tell	200	222
c_emo (emotion)	c_emo_1	explicit expression of emotion	0	none	4,577	3,429
			1	negative emotion	110	365
			2	positive emotion	38	105
			3	expression of amusement/ridiculing	21	0
			99	cannot tell	21	16
	c_emo_2a	trigger for emotion		same as n_4		
	c_emo_2b	level of generalization		same as c_ne_3		
	c_emo_3	irony or sarcasm		same as c_ne_6		

Table 9: Annotation modules for comments and their respective categories. Labels for each category with their corresponding number of German (DE) and French (FR) comments.