Hateful Word in Context Classification

Sanne Hoeken[1], **Sina Zarrieß**[1] and **Özge Alaçam**[1,2]
[1]Computational Linguistics, Department of Linguistics, Bielefeld University, Germany
[2]Center for Information and Language Processing, LMU Munich, Germany
{sanne.hoeken, sina.zarriess, oezge.alacam}@uni-bielefeld.de

Abstract

Hate speech detection is a prevalent research field, yet it remains underexplored at the level of word meaning. This is significant, as terms used to convey hate often involve non-standard or novel usages which might be overlooked by commonly leveraged LMs trained on general language use. In this paper, we introduce the Hateful Word in Context Classification (**HateWiC**) task and present a dataset of ∼4000 WiC-instances, each labeled by three annotators. Our analyses and computational exploration focus on the interplay between the subjective nature (context-dependent connotations) and the descriptive nature (as described in dictionary definitions) of hateful word senses. HateWiC annotations confirm that hatefulness of a word in context does not always derive from the sense definition alone. We explore the prediction of both majority and individual annotator labels, and we experiment with modeling context- and sense-based inputs. Our findings indicate that including definitions proves effective overall, yet not in cases where hateful connotations vary. Conversely, including annotator demographics becomes more important for mitigating performance drop in subjective hate prediction.

1 Introduction

This paper introduces the Hateful Word in Context Classification (HateWiC) task, which aims to determine the hatefulness of a word within a specific context, as illustrated in Figure 1. We argue that hateful word senses are not enough in focus within Hate Speech Detection (HSD) research, and not descriptive only, but highly subjective, asking for another approach than other lexical semantic tasks like Word Sense Disambiguation (WSD).

Hateful senses are not enough in focus within HSD research. The current focus of HSD research predominantly revolves around the classification of entire utterances, such as social media posts (Waseem and Hovy, 2016; Davidson et al., 2017). Within these utterances, lexical cues frequently play a significant role in the decision-making process. Yet, the computational modeling of context-specific hateful word meanings remains largely unexplored, with a few exceptions in this direction (Dinu et al., 2021; Hoeken et al., 2023b).

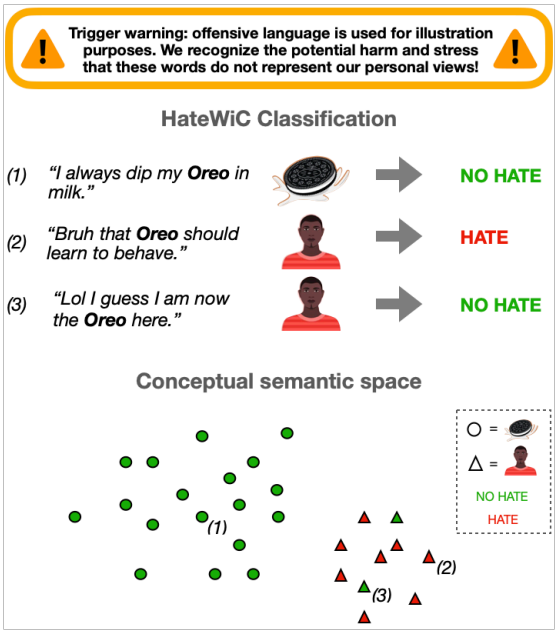

Figure 1: Illustration of the HateWiC Classification task and a conceptual semantic space that underlies the targeted phenomenona of hate-heterogeneous word senses, highlighting the distinction between the descriptive aspects (e.g. cookie or person) and hateful connotation.

LMs commonly employed in HSD systems demonstrate effective word meaning modeling (Nair et al., 2020), but they tend to lack sensitivity to domain-specific, non-standard or novel word senses (Kumar et al., 2019; Blevins and Zettlemoyer, 2020). This insensitivity becomes particularly critical in detecting hateful word meanings, that are used in unconventional or emerging contexts as the evolution of societal events gives rise to the continuous invention of novel expressions of

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 172–186
November 12-16, 2024 ©2024 Association for Computational Linguistics

hate (Qian et al., 2021). Words within the established lexicon, like *Oreo*, whose primary meaning may not have any negative connotations (a cookie), are repurposed to convey hate towards particular groups or individuals (e.g. based on ethnicity).

Hateful senses are not descriptive only. Following theoretic work by Frigerio and Tenchini (2019), hateful terms could be positioned along a meaning continuum from descriptive to expressive, closer to but not *at* the expressive outer end. The descriptive component comprises the truth-conditional attributes of a term, often recorded in dictionary definitions. The expressive component, i.e. the connotation of a term, concerns speakers' attitudes and emotions, making it highly context-specific and subjective. A word's sense definition could imply a hateful connotation, but this is not always the case, such as when used in a playful or self-identifying way (e.g. the third usage in Figure 1). Thus, a word's hateful connotation is not exclusively tied to its descriptive definition, a phenomena which we term as *hate-heterogeneous senses*, but depends on various contextual factors like conveyed content or the reader's identity. This aspect is often overlooked in HSD systems, typically developed using data reflecting a single (majority) perspective (Zampieri et al., 2019; Mathew et al., 2020).

Our contributions. In this study we address the gap in HSD by focusing on subjective hateful word meanings within context. We introduce the HateWiC dataset, a dataset of ∼4000 WiC-instances for which we collected three hatefulness ratings each. We design methods to classify sense representations and evaluate them both against the majority and the individual annotator's label. In doing so, we experiment with modeling descriptive and subjective aspects of hateful word senses by incorporating sense definitions (as also provided to annotators) and annotator information.[1]

2 Related Work

In this section, we discuss previous work on the key aspects of this study: HSD at the word level (2.1), incorporating subjectivity in HSD (2.2), and methods for modeling word senses (2.3).

[1] The code used for this study and the directly publicly available part of our data can be found at: https://github.com/SanneHoeken/HateWiC. The full HateWiC dataset will be open to public upon request and will be licensed under CC BY-NC 4.0.

2.1 Hate Speech Detection on Word Level

Although the main body of research into HSD has focused on the level of utterances, some studies have delved into hate speech on the lexical level. Prior to LMs, feature-based HSD systems (e.g. Lee et al. (2018)) often incorporated hate speech lexicons. Wiegand et al. (2018) demonstrated the induction of an abusive word lexicon in a non-contextualized setting. A specific subset of hateful terms *within* context is addressed by Hoeken et al. (2023b), who modelled slur detection employing a dimension-based method similar to the identification of gender bias in word embeddings (Bolukbasi et al., 2016). This approach, that requires a pre-given set of minimal pairs, is much more complex when tackling the broader spectrum of hateful terms, including words with both hateful and non-hateful meanings.

Qian et al. (2019) presented a framework aiming to predict the definition of hateful symbols, terms with a non-hateful surface form conveying hate, yet not covering the disambiguation between hate and non-hate. Mendelsohn et al. (2023) focused on a related phenomenon, dog whistles, examining whether GPT-3 can identify their covert meanings, surface them in text generation and detect them in real-world texts. Dinu et al. (2021) introduce the task of disambiguating pejorative word usage, presenting two small-scale datasets and evaluating several methods, with an MLP model classifying BERT embeddings (Devlin et al., 2019) as most effective approach. Muti et al. (2024) addressed pejorative word disambiguation as a preliminary step for misogyny detection in Italian texts.

Our study focuses on the disambiguation of words with hateful meanings, which, although overlapping with dog whistles and pejorative words, belong to distinct categories. Unlike hateful words, dog whistles are always intentionally ambiguous, concealing one meaning from the out-group which is not exclusively hateful. Pejorative words, encompass any negatively connoted terms that may not be hateful when not targeted at an individual or group. More importantly, unlike the single-perspective annotations employed in the aforementioned studies, our focus is on subjective hate speech annotation and it is conducted on a much larger scale.

2.2 Subjective Hate Speech Detection

Most existing datasets and methods in HSD adopt a single, majority perspective, ignoring the inher-

ent subjectivity influenced by diverse social and cultural factors (Zampieri et al., 2019; Founta et al., 2018). This approach has been shown to result in problematic biases, concerning e.g. ethnicity, gender, and political beliefs and highlight the need for new methodologies that account for the varying interpretations of hateful connotations (Davidson et al., 2019; Kumar et al., 2021; Sap et al., 2022).

Davani et al. (2022) took steps in this direction by training a model to predict individual annotations as subtasks, still ultimately aiming to predict the majority label. Kanclerz et al. (2022) addressed the task of predicting each individual annotator's label, by leveraging annotator's labeling statistics within the dataset. Alacam et al. (2024) study the incorporation of gaze features (on token- and sentence level) from human annotators for predicting their subjective hate ratings. Another more comprehensive approach is presented by Fleisig et al. (2023), who included annotators' demographics, preferences, and experiences as input, along with text. They utilized RoBERTa (Liu et al., 2019) to embed descriptions of these characteristics. Our research continues this line of work by predicting individual annotator labels and accounting for their demographics in the classification of hateful words.

2.3 Modeling Word Senses

Shifting the focus from modeling hateful utterances to the meaning of hateful words *within* utterances, touches upon various lexical semantic NLP tasks that involve the creation of word sense representations (Vulić et al., 2020a; Schlechtweg et al., 2020; Martelli et al., 2021). Approaches to these tasks often employ contextualized word embeddings extracted from pretrained (often BERT-based) LMs (Loureiro and Jorge, 2019; Martinc et al., 2020; Bommasani et al., 2020). Fine-tuning a model on particular data or tasks, such as WSD or sentiment classification, is performed to potentially inject relevant information into the resulting representations (Giulianelli et al., 2020; Hoeken et al., 2023a). Rachinskiy and Arefyev (2022) leveraged an effective WSD model developed by Blevins and Zettlemoyer (2020), which jointly optimizes two encoders for the context and gloss of a word sense, respectively. For the task of semantic change discovery, they extracted the representations of the context encoder of the WSD-finetuned model.

Recently, Giulianelli et al. (2023) introduced an innovative approach to computational sense representations. Their method adopts the definition-as-sense paradigm, utilizing definitions generated by a Flan-T5 model (Chung et al., 2022) fine-tuned on datasets of definitions with usage examples. Sentence embeddings of these generated context-specific definitions show promising results on lexical semantic similarity tasks. Despite these advancements focused on descriptive word senses, effective approaches for modeling highly connotative lexical phenomena remain unclear.

3 The HateWiC Dataset

We introduce the HateWiC dataset, which includes hate ratings for words within example usages along with their word sense definitions which may be hate-heterogeneous, as illustrated in Figure 1. We describe the dataset construction below.

3.1 Wiktionary Data

Data was scraped from the English Wiktionary in November 2023, an online dictionary where anyone can contribute to documenting and explaining words in use. Therefore, Wiktionary provides up-to-date insights from user perspectives and covers a broader range of terms from diverse domains than traditional dictionaries.

Each entry (word or multi-word expression) includes information such as definitions, example uses, and category labels that provide additional context about a word's use (e.g., 'British slang' or 'Archaic'). Using the Wiktionary API, we extracted all entries that had at least one word sense tagged with the categories Offensive and Derogatory and were also members of the category People, to gather the most relevant terms for hate speech detection purposes. For each of the resulting 1087 terms, we scraped all its sense definitions along with all labeled categories and example sentences (using the WiktionaryParser library). This resulted in 3500 senses and 4671 examples.

To suit the dataset for our HateWiC classification task, we manually excluded 642 examples due to historical spelling or other deficiencies, as detailed in Appendix A. After processing, the dataset comprised 4029 instances covering 1888 unique senses, averaging 4.88 examples per sense, and 826 unique terms, averaging 2.29 senses per term.

3.2 Annotation

The senses from the Wiktionary data include labels regarding their offensiveness or derogatory nature. However, these classifications do not represent the

Example	Term	Definition	Annotations	Binary labels	Majority label	Hate-heterogeneous sense	Agreement on binary
(1) "Me having an up to date style even though I've turned into a carrot cruncher."	carrot cruncher	Someone from a rural background.	Nh, Nh, Nh	0, 0, 0	0	True	True
(2) "you're a friggn' carrot cruncher and you support the bloody scally's."	carrot cruncher	Someone from a rural background.	Sh, Sh, Sh	1, 1, 1	1	True	True
(3) "The bugger's given me the wrong change."	bugger	A foolish person or thing.	Wh, Sh, Sh	1, 1, 1	1	False	True
(4) "He's a silly bugger for losing his keys."	bugger	A foolish person or thing.	Nh, Wh, Sh	0, 1, 1	1	False	False

Table 1: HateWiC examples with their annotations, illustrating the phenomena of annotator disagreement and hate-heterogeneous word senses (Nh = Not hateful, Wh = Weakly hateful, Sh = Strongly hateful)

diverse interpretations of these labels due to their subjective nature. In this study, we aim to survey and model different beliefs, following a descriptive data annotation paradigm as proposed in the framework by Rottger et al. (2022). This paradigm highlights the value of using crowd-sourced annotators from diverse backgrounds to encourage annotator subjectivity and mitigate bias, without relaying on a predefined detailed definition of hate speech. Specifically, we collected crowd-sourced annotations using Prolific with a link integration to Argilla. Argilla, an open-source platform launched on HuggingFace Spaces, is used to set up the annotation task on HateWiC data.

For each annotation instance, annotators are presented with an example sentence, the target term, and its sense definition. They are then prompted with the question: "*How would you rate the hatefulness of the meaning of the target term within the specific example text?*". Annotators respond by selecting from the labels: 'Not hateful', 'Weakly hateful', 'Strongly hateful' and 'Cannot decide'. An example of an annotation instance and the user interface are depicted in a screenshot provided in Appendix B. In the annotation guidelines (accessible on our repository), annotators are instructed to focus their evaluation on the specific usage of the term within the example sentence, rather than the overall connotation of the sentence, or the definition, which is only provided to aid in understanding the term's meaning. Additionally, we emphasize the subjective nature of their judgements.

We aimed for three annotations per instance, with each annotator labeling 250 instances.[2] Using Prolific's pre-screening filters, we selected annotators who indicated that their primary language is English. To improve the quality of the collected annotations, we excluded and replaced data from annotators who were too fast and/or failed control instances.[3] Prolific provides demographic information for each annotator, which can be connected to their annotations. The final pool of annotators, after exclusions, consisted of 48 individuals with diverse genders and ethnicities averaging 28 years old (more details in Appendix B).

3.3 Dataset Results

After excluding the 'Cannot decide' annotations[4], the dataset yielded 11902 individual annotations, of which 5708 (48.0%) hateful and 6194 (52.0%) not hateful (after converting to binary by merging 'Weakly hateful' and 'Strongly hateful'). After applying majority voting, out of the 3845 example sentences with a clear majority binary label, 1815 (47.2%) were classified as hateful and 2030 (52.8%) as not hateful, yielding a balanced dataset with respect to hatefulness.

Annotators agreed for 60% (i.e. 2414) of the binary classification with a Krippendorff's alpha of 0.45. For the three-class classification, agreement was 51.3% with a Krippendorff's alpha of 0.33. In comparison, Mathew et al. (2020) reported an agreement of 0.46 for a similar three-class task, and Vigna et al. (2017) 0.26 for their binary setting. The agreement scores underscore the inherent subjectivity of the task, motivating us to include individual demographics to our modeling.

The high degree of context dependency regarding hate becomes even more apparent when we examine the relationship between word senses (the descriptive aspects outlined in their definitions) and the hatefulness ratings assigned to examples of those senses. We identified 319 **hate-heterogeneous** sense definitions, i.e. unique definitions for which example sentences exist in the dataset with both hateful and non-hateful majority annotations. Two examples from the annotated data given in Table 1 illustrate this phenomenon. Both examples mention the term *carrot cruncher*

[2]The average reward per hour was £9.28.

[3]More than 2 out of 8 failed control instances and/or less than 45 min. completion time; median time was 90 min.

[4]The majority of the 514 'Cannot decide' annotations were found to concern deficient sentences upon closer analysis.

with the sense definition "Someone from a rural background." where (1) is unanimously annotated as not hateful and (2) is unanimously annotated as strongly hateful. This observation solidifies the idea, already implied by the inter-annotator agreement for individual labels (and exemplified by (4) in Table 1), that the hateful connotation of a word sense is not exclusively determined by its descriptive definition.

4 HateWiC Classification

Our HateWiC dataset enables the development and evaluation of computational methods for predicting whether the meaning of a target term is hateful within a specific context. Figure 2 provides an illustration of the primary methodological pipeline we present in this paper. We introduce various classification methods that differ with respect to the sense representations (outlined in 4.1) and incorporation of annotator information (4.2) as input to a classification model (4.3), or that leverage an instruction-tuned LLM (4.4).

4.1 Sense Representations

For representing the sense of a target term, we primarily follow a common procedure in lexical semantic NLP tasks and extract contextualized embeddings from pretrained LMs. To optimize effectiveness on the HateWiC task, we experiment with various encoder models and embedding types. Appendix C provides additional details on our employed methods.

Encoder models. We experiment with three different encoder models, each trained on different data or tasks. We use the pretrained **BERT** (base) model (Devlin et al., 2019) and **HateBERT** (Caselli et al., 2021), a re-trained BERT model on hate speech[5]. As third, we utilize a trained bi-encoder model for Word Sense Disambiguation (Blevins and Zettlemoyer, 2020), which we refer to as **WSD Biencoder**. The model comprises a contextualized word encoder and a gloss encoder initialized with BERT-base encoders. We train it on WordNet data (Miller et al., 1994), following the same procedure as detailed in (Blevins and Zettlemoyer, 2020), for 7 epochs with a batch size of 8. Following Rachinskiy and Arefyev (2022), the WSD-optimized contextualized word encoder is then used for obtaining embeddings.

Embeddings. The encoders are used to generate different word sense related representations. First, we compute **word in context (WiC)** embeddings. We feed the example sentence to the encoder model and extract the last hidden layer for the subword-tokenized position(s) that encode the target term (averaging over them in case of multi-subword target terms). Second, we test the incorporation of word sense definitions from Wiktionary. This **definition (Def)** embedding is obtained by averageing over all token embeddings, using the same procedure as for WiC embeddings but with the definition sentence as input. Third, considering that pre-given definitions may not be available in practical applications, we create **T5-generated definition (T5Def)** embeddings. We generate definitions using a FLAN-T5 Base (250M parameters) model developed by (Giulianelli et al., 2023)[6] which was fine-tuned on datasets of English definitions and usage examples. We prompt the model with the same template as it was trained on: "[SENTENCE] What is the definition of [TERM]?". Consequently, the generated generated are more context-specific than the Wiktionary definitions. These generated definitions are embedded the same way as the Def-embeddings.

4.2 Annotator Information

To address the subjective nature of the HateWiC classification task, highlighted by the inter-annotator agreement in our dataset, we incorporate this aspect into our modeling approaches. We experiment with a similar strategy as presented in Fleisig et al. (2023). For each individual annotation of a HateWiC instance, we concatenate an **annotator (Ann)** embedding to the corresponding sense embedding, that represent a description of annotator's demograpics. This description is embedded through the same procedure as the definition embeddings and follows this template:

> "Reader is [AGE], [GENDER] and [ETHNICITY]."

4.3 Classifying Embeddings

We test the effectiveness of (the concatenation of combinations of) the embeddings proposed above on our HateWiC classification task by using them as input to a classification model. To this end, we train and test a four-layer multi-layer perceptron (MLP) model (a classification algorithm also used in Dinu et al. (2021)) on the HateWiC dataset.

[5] https://huggingface.co/GroNLP/hateBERT

[6] https://huggingface.co/ltg/flan-t5-definition-en-base

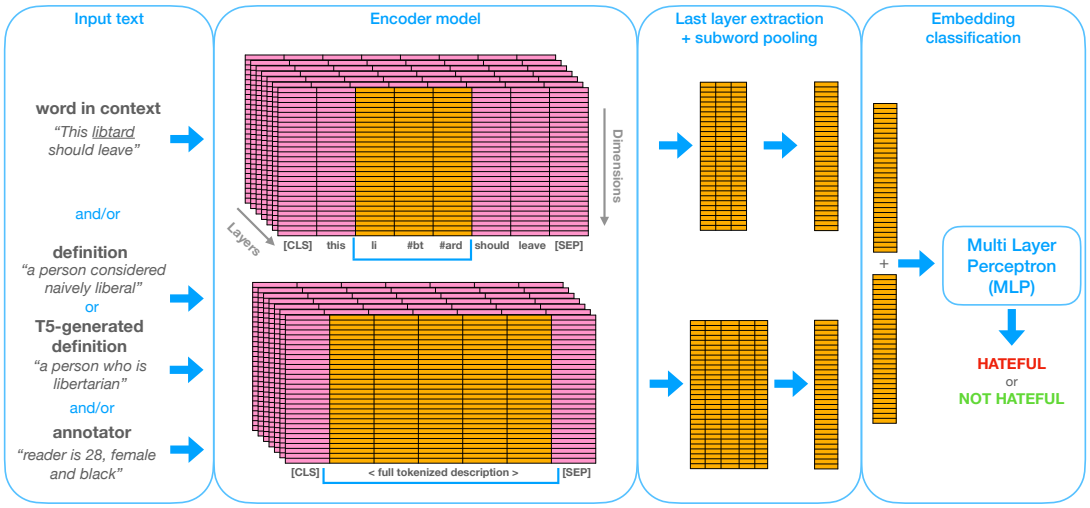

Figure 2: Illustration of our main HateWiC classification pipeline.

4.4 Classification with LLaMA 2

In addition to the encoder-LM based approaches above, we also experiment with a LLaMA 2 model (Touvron et al., 2023). Due to their instruction-tuning training regime, and huge amount of training data, foundation models like LLaMA 2 are proven to be superior to LMs on many zero-shot settings, yet subjective HSD and WSD are by nature very challenging tasks. We aim to see the abilities of an instruction-tuned LLM on this task as a (strong) baseline. We test zero-shot classification with a 7B-sized LLaMA 2 model[7]. We run the inference of this model using the *transformers* library. In our prompt, we input the example sentence and the target term and instruct the model to classify the meaning of the term as hateful or not hateful (complete template and configuration parameters are provided in Appendix C).

5 Evaluation Setup

We evaluate our proposed methods using various test setups on the HateWiC dataset (5.1). Additionally, we compare our methods with the work of Dinu et al. (2021), as described in 5.2.

5.1 HateWiC

Our HateWiC dataset includes three hate ratings for each example sentence, allowing evaluation on two distinct tasks that vary in terms of subjectivity inclusion. For both tasks, we utilize binary labels.

1. **Majority** label prediction: gold labels represent 4029 majority votes on each example.
2. **Subjective** label prediction: gold labels consist of all 12442 individual annotations: a rating per example and annotator.

We conduct evaluations for each task using a ten-fold cross-validation setup. For each fold, we divide the dataset into training, development, and test sets with an 80-10-10 ratio. We experiment with two variants:

1. **Random**: The data is randomly split based on example sentences, testing performance on sentences not seen during training (similar to common practice in WSD-like tasks (Dinu et al., 2021)), which is particularly relevant for individual annotator prediction where multiple instances of the same sentence occur.
2. **Out-of-Vocabulary (OoV)**: The data is split based on terms, testing performance on unseen terms, i.e. zero-shot capabilities.

5.2 Comparison with Dinu et al. (2021)

We also train and test on two small datasets of English tweets developed and used in Dinu et al. (2021). They collected these from existing hate speech datasets, focusing on tweets that mention one of the terms in a curated set of pejorative terms. Each tweet was labeled based on whether the term was used pejoratively. The first dataset, which we will refer to as **DINU1** comprised 1004 tweets covering 31 terms. The second, which we name **DINU2**, consisted of 301 tweets covering 11 terms. Their reported best method involved MLP classification of BERTweet (Nguyen et al., 2020) and BERT (base) embeddings (extracted as the sum of all model layers for the target word position) on

[7] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

DINU1 and DINU2, respectively. We aimed to use the same evaluation set-up as described in their paper, using five-fold cross-validation and reporting the average over accuracies *per term*.

6 Results

This section presents the results of our proposed methods on the HateWiC classification, evaluated using the above outlined setups.

6.1 Majority HateWiC Classification

Table 2 presents the accuracy results on HateWiC classification compared to the majority label. Overall, the performance values demonstrate the effectiveness of all methods, with only minimal differences (max. 2 %-points) between BERT, HateBERT and WSD biencoder models. Training BERT-based models on different types of information regarding hatefulness or word senses does not seem to have a substantial effect.

Embeddings	**BERT**		**HateBERT**		**WSD bien.**	
	Random	**OoV**	**Random**	**OoV**	**Random**	**OoV**
WiC	0.75	0.73	0.75	0.71	0.76	0.73
Def	0.77	0.75	0.78	0.73	0.78	0.73
T5Def	0.70	0.67	0.70	0.67	0.72	0.69
WiC+Def	0.78	0.77	<u>0.80</u>	0.77	0.79	**0.78**
WiC+T5Def	0.75	0.74	0.76	0.73	0.76	0.73

Table 2: Accuracy on HateWiC classification compared to the **majority** label, with different input embeddings, tested on a random data split (best underlined) and a test split with OoV terms only (best in bold).

Def-embeddings achieve slightly higher accuracies than WiC-embeddings , and a combination of the two yields the best results. For a test set with OoV terms only, all embedding types show only a slight drop in performance. WiC+Def-embeddings exhibit the smallest decline on the zero-shot setting and achieve 2-5 % higher accuracy than WiC- and Def-embeddings. This indicates that definitions provide valuable information, performing better on their own than word information alone, and the combination of both is most effective, especially for OoV-terms. T5-generated definitions demonstrated the lowest accuracy on their own but perform equally or slightly better than WiC-embeddings when concatenated. An evaluation of T5-generated definitions compared to Wiktionary definitions showed a very low SacreBLEU score of 3.822 (in range 0 to 100), possibly explaining the differences in performance between them.

The distinction between context-independent Def-embeddings and context-specific WiC- and T5Def-embeddings becomes more clear upon examining their performance across hate-homogeneous and hate-heterogeneous instances (as defined in Section 3.3), presented in in Table 3. In the case of hate-heterogeneous instances, we observe an accuracy drop of up to 47% when using Def-including embeddings compared to the homogeneous instances. This drop is limited to 24-29% for the other embeddings, showcasing their superior ability in handling less descriptive scenarios. We define hate-homogeneous here as instances where definitions have example sentences in the dataset with either hateful or non-hateful (majority) annotations whereas hate-heterogeneous have both (as detailed in Section 3.3).

HateBERT embeddings	**Hate-homogeneous**	
	True	**False**
WiC	0.82	<u>0.55</u>
Def	<u>0.91</u>	0.44
T5Def	0.76	0.52
WiC+Def	<u>0.91</u>	0.49
WiC+T5Def	0.84	<u>0.55</u>

Table 3: Accuracy on HateWiC classification compared to the **majority** label w.r.t. hate homogeneity of the sense definition (best underlined).

LLaMA 2 result. The accuracy score on the HateWiC classification using a LLaMa 2 model, following the zero-shot experimental setup detailed in Section 4.4, is 0.68. Unlike the superior performance on many downstream tasks, the LLaMA model falls short compared to the aforementioned models on our HateWiC task. This outcome highlights the subjective nature of the task, indicating that general-purpose models struggle to fully grasp its nuances and perform well on it.

6.2 Subjective HateWiC Classification

Performance of our designed methods on predicting individual annotation labels, which showed considerable variation in Section 3.3, are presented in Table 4. Overall, accuracy values are slightly lower (by 2-5 %-points) compared to predicting the majority label, but remain robust. The results exhibit the same patterns in terms of different models, test data setups, and tested embedding types. Adding the Annotator embedding has a minimal effect, generally resulting in equal or slightly improved performance compared to the same type of embedding without concatenated annotator information.

To better understand the impact of subjectivity, we more closely examine instances where subjec-

Embeddings	BERT		HateBERT		WSD bien.	
	Random	OoV	Random	OoV	Random	OoV
WiC	0.71	0.69	0.71	0.69	0.72	0.70
Def	0.74	0.71	0.75	0.73	0.74	0.71
T5Def	0.68	0.65	0.68	0.67	0.68	0.67
WiC+Def	0.75	**0.74**	0.75	0.73	0.75	0.73
WiC+T5Def	0.72	0.70	0.72	0.71	0.73	0.69
WiC+Ann	0.72	0.69	0.72	0.69	0.72	0.70
Def+Ann	0.74	0.72	0.76	0.72	0.75	0.72
T5Def+Ann	0.69	0.67	0.69	0.65	0.69	0.68
WiC+Def+Ann	0.75	0.73	0.75	**0.74**	0.75	**0.74**
WiC+T5Def+Ann	0.72	0.71	0.73	0.71	0.73	0.72

Table 4: Accuracy on HateWiC classification compared to the **individual** annotator label, with different input embeddings, on a random data split (best underlined) and a test split with OoV terms only (best in bold).

HateBERT embeddings	Majority annotation		Hate-homogeneous	
	True	False	True	False
WiC	0.77	0.40	0.77	0.55
Def	0.81	0.36	0.83	0.51
T5Def	0.72	0.42	0.72	0.53
WiC+Def	0.83	0.36	0.83	0.55
WiC+T5Def	0.78	0.39	0.78	0.56
WiC+Ann	0.77	0.49	0.77	0.59
Def+Ann	0.82	0.44	0.82	0.60
T5Def+Ann	0.73	0.47	0.72	0.59
WiC+Def+Ann	0.80	0.44	0.81	0.58
WiC+T5Def+Ann	0.77	0.48	0.78	0.58

Table 5: Accuracy on HateWiC classification compared to the **individual** label w.r.t. annotator agreement with the majority label and hate homogeneity of the sense definition (best underlined).

tivity is most apparent (and thus potentially harmful when methods fail). In Table 5 we report performance results not only with respect to the hate homogeneity of word senses, but also to annotator agreement, i.e. whether the annotator agreed with the majority. We present results for HateBERT embeddings in an evaluation setting with random test data split, but similar patterns are observed for BERT and WSD Biencoder embeddings, as well as on on a test data split with OoV terms only.

For sentence annotations where the annotator disagreed with the majority label or the sense definition is hate-heterogeneous, the performance of all embeddings drops significantly. This effect is most pronounced for definition-including embeddings (Wiktionary), less so for T5-generated, which aligns with their more context-specific nature. Specifically, there is an accuracy drop of up to 47% in cases of annotator disagreement, and up to 32% in cases of hate-heterogeneous definitions. However, incorporating annotator information mitigates this effect by up to 11%. Annotator information contributes to the cases where the subjective annotation deviates from the majority label, these cases also align with sense definitions that exhibit both hateful and non-hateful labeled sentences.

6.3 Results on DINU Data

The DINU1 and DINU2 evaluation datasets do not provide sense definitions or information on annotators, thereby limiting our testing to our methods that do not require this information. Table 6 presents the results on both DINU1 and DINU2. Our methods, except for those including T5Def-embeddings only, demonstrate improvements over the best-performing methods proposed by Dinu et al. (2021). These improvements are particularly substantial (by 8%) for the larger DINU1 dataset. Consistent with trends observed for the HateWiC dataset, the concatenation of WiC and T5-generated definition embeddings yields the best performance across both DINU sets, underscoring the potential of incorporating automatically generated definitions in the absence of dictionary definitions for HateWiC classification.

Model	Embedding	DINU1	DINU2
BERT	WiC	0.89	0.83
	T5Def	0.81	0.79
	WiC+T5Def	0.90	0.83
HateBERT	WiC	0.87	0.83
	T5Def	0.83	0.80
	WiC+T5Def	0.90	0.84
WSD Bienc.	WiC	0.90	0.82
	T5Def	0.80	0.79
	WiC+T5Def	0.90	0.84
Best Dinu		0.82	0.83

Table 6: Accuracy of our methods on the DINU datasets compared the accuracy of the best performing method as reported in Dinu et al. (2021) (best underlined).

7 Discussion

Our study offers valuable insights into the detection of hate speech through the lens of lexical semantics, introducing the HateWiC dataset and presenting classification experiments. The negligible difference observed in our experimental outcomes between HateBERT and general (WSD) models not only questions the efficacy of extensive training on hate speech data for accurately capturing hateful semantics, but also underscores the necessity of a more nuanced approach beyond the existing lexical semantic methods for tasks like HateWiC classification. Our results demonstrate the impact of incorporating sense definitions and annotator characteristics on model performance, particularly in scenarios involving out-of-vocabulary (OoV) terms

or high subjectivity.

To define or not define? Hateful terms, according to lexical semantic theory, primarily contain an expressive component but not exclusively. Incorporating sense definitions into our methods, to encompass the descriptive component of hateful terms, yielded mixed results. Overall, embedded Wiktionary definitions proved highly effective, outperforming Word in Context (WiC) embeddings alone. T5-generated definitions demonstrated the lowest accuracy on their own but performed equally or slightly better than WiC-embeddings only when concatenated with WiC-embeddings. However, in cases with more variation in the subjective ratings, the performance of all embeddings dropped significantly but most pronounced for Wiktionary definition embeddings, though to a lesser extent for T5-generated definitions (with a drop difference of up to 23%). This highlights the usefulness of automatically generating context-specific definitions for subjective lexical semantic tasks like HateWiC classification. Future research will focus on more advanced definition generation techniques, possibly leveraging larger models or fine-tuning on Wiktionary definitions, while avoiding overreliance on dictionary definitions as the ultimate standard.

To individualize anyway? The low inter-annotator agreement in our dataset underscores the importance of considering individual annotator perspectives in hate speech detection. Our experiments incorporating annotator information in our computational methods proved beneficial, particularly in cases of annotator disagreement or hate-heterogeneous definitions, where including annotator information mitigated accuracy decline by up to 11%-points. This highlights the value of personalizing models to account for subjectivity in annotations. Future research could explore additional annotator information and conduct ablation experiments to identify the most effective aspects for HateWiC classification.

To consider as well? Our study paves the way to obtaining deeper insights into the relationship between hateful and non-hateful word senses. For instance, whether certain semantic relations (e.g. metaphorical, metonymical), categories (e.g. food, animals), or attributes (e.g. color, material) are more likely to distinguish between hateful and non-hateful senses. And even next-level, whether these discriminators are language-specific or show cross-language parallels. Identifying such consistencies between (non-)hateful senses could enhance the (automatic) discrimination between them.

8 Conclusion

This paper introduces the Hateful Word in Context Classification (HateWiC) task, addressing the underexplored area of subjective hateful word meanings within specific contexts. We present the HateWiC dataset, comprising about 4000 WiC-instances, each annotated with three hateful ratings. Our study focused on the interplay between descriptive and subjective aspects of hateful word senses. We addressed the prediction of both majority and individual annotator labels. We experimented with different types of inputs to our classification system, including sense definitions and annotator demographics. We demonstrated the impact of these factors on model performance, particularly in cases involving out-of-vocabulary terms or high subjectivity. The incorporation of established sense definitions proved highly effective overall but demonstrating diminished performance in less descriptive scenarios. Conversely, including annotator characteristics proved beneficial, particularly in cases of annotator disagreement or hate-heterogeneous definitions. These findings underscore the value of personalizing models to account for subjectivity in annotations. Furthermore, our results suggest the potential usefulness of automatically generating definitions for subjective lexical semantic tasks like HateWiC classification.

Limitations

Although the Wiktionary data we utilize offers insights from user perspectives for a wide array of terms, its quality may be lower compared to expert-curated dictionaries. The provided information may contain inaccuracies, as users might not have the necessary expertise, and inconsistency in documentation could exist. However, the collaborative nature of Wiktionary allows for censorship by consensus and adherence to Wiktionary policies, mitigating some of these concerns.

A constraint of our evaluation set-up lies in its reliance on binary labels. Hate speech is a multifaceted phenomenon, and a more nuanced class scheme may offer a more comprehensive understanding in future research.

Ethics Statement

Our study includes demographic data of annotators that concern Prolific prescreening responses which are all with annotator's consent, self-reported, and are not provided with any direct identifiers like name or address. All prescreening questions, except for age and country of residence, are optional for participants to answer, and most personal questions have a 'Rather not say' option. By incorporating demographic information from annotators, we aim to enhance the understanding and prediction of how different groups perceive hate speech. This approach will ultimately lead to more robust and inclusive classification systems. However, the inclusion of demographic data raises privacy concerns, particularly the risk of re-identifying annotators. To address this, we have made our dataset available only upon request, under the CC BY-NC 4.0 license. This measure allows us to better control access to the information, ensuring it is used responsibly, ethically, and exclusively for non-commercial purposes.

Acknowledgements

The authors acknowledge financial support by the project "SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), which is funded by the program "Netzwerke 2021" of the Ministry of Culture and Science of the State of North Rhine-Westphalia, Germany.

References

Özge Alacam, Sanne Hoeken, and Sina Zarrieß. 2024. Eyes don't lie: Subjective hate annotation and detection with gaze. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. 2021. A computational exploration of pejorative language in social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 6715–6726, Singapore. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Aldo Frigerio and Maria Paola Tenchini. 2019. Pejoratives: a classification of the connoted terms. *Rivista Italiana di Filosofia del Linguaggio*, 13(1).

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Leopold Hess. 2021. *Slurs: Semantic and Pragmatic Theories of Meaning*, page 450–466. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Sanne Hoeken, Özge Alacam, Antske Fokkens, and Pia Sommerauer. 2023a. Methodological insights in detecting subtle semantic shifts with contextualized and static language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3662–3675, Singapore. Association for Computational Linguistics.

Sanne Hoeken, Sina Zarrieß, and Ozge Alacam. 2023b. Identifying slurs and lexical hate speech via lightweight dimension projection in embedding space. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 278–289, Toronto, Canada. Association for Computational Linguistics.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*, SOUPS'21, USA. USENIX Association.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Ho Suk Lee, Hong Rae Lee, Jun U. Park, and Yo Sub Han. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, 113:22–31. Publisher Copyright: © 2018 Elsevier B.V.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the Workshop on Human Language*

Technology, HLT '94, page 240–243, USA. Association for Computational Linguistics.

Arianna Muti, Federico Ruggeri, Cagri Toraman, Alberto Barrón-Cedeño, Samuel Algherini, Lorenzo Musetti, Silvia Ronchi, Gianmarco Saretto, and Caterina Zapparoli. 2024. PejorativITy: Disambiguating pejorative epithets to improve misogyny detection in Italian tweets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12700–12711, Torino, Italia. ELRA and ICCL.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2019. Learning to decipher hate symbols. 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 3006–3015, Minneapolis, Minnesota. Association for Computational Linguistics.

Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. Lifelong learning of hate speech classification on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2304–2314, Online. Association for Computational Linguistics.

Maxim Rachinskiy and Nikolay Arefyev. 2022. GlossReader at LSCDiscovery: Train to select a proper gloss in English – discover lexical semantic change in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 198–203, Dublin, Ireland. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Italian Conference on Cybersecurity*.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020a. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020b. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 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, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

A Wiktionary Data Processing

Our data was scraped from the English Wiktionary comprising entries with information on definitions, example uses, and category labels that provide additional context about a word's use. We scraped all sense definitions along with all labeled categories and example sentences of the selected terms using the WiktionaryParser library. This library method did not split the examples over the set of sense definitions (i.e. provided all examples in one bundle), so we manually matched the right examples with the right sense definitions, through look up on the Wiktionary website, afterwards.

To suit the dataset for the envisioned task we manually excluded 642 examples that were either written in historical spelling or not single in-the-wild usages of the term. The latter concerned usages, like the examples below (with the target term in bold), that were (a) dictionary-typical nominal phrases and not sentences, (b) concerned meta-level discussions of the target term or (c) dialogues or other indirect uses of the target term.

> (a) "a bird **feeder**"
>
> (b) "A '**lot lizard**' was somebody who walked the sales lot and looked at every car and still didn't buy."
>
> (c) "Threads on the social media giant Reddit occasionally discuss or condemn "**transtrenders**" [. . .]"

Finally, we slightly edited some type of instances that concerned non-exact matches between word form of the term and its occurrence in the example. For compounds or multi-word expressions, this mismatch often concerned the (non-)use of a whitespace or hyphen between compound parts (e.g. the term baby face occurred also as babyface or baby-face in examples). This type of mismatches was solved by applying a simple rule-based replacement strategy to the example sentences.

Other types of non-exact word form matches were mainly caused by inflection (e.g. plural forms for nouns) and some by misspellings. These cases were left unchanged for the final dataset as removing could influence the meaning.

We also created groupings to aggregate category labels, consolidating the 585 unique Wiktionary labels present in our dataset into a manageable set of usage tags. This enrichment potentially provides useful information for future analyses on usages of hateful terms.

B Annotation Details

Figure 3 displays the user interface for annotation, with an example of an annotation instance.

Below, we report the distribution of our annotators with respect to age, gender, and ethnicity. It is important to note that we use the categories as provided through the Prolific provided presecreening responses, which are simplified groupings intended to give a general overview. As detailed in the Ethics Statement, we acknowledge that this categorization does not fully capture the complexity and diversity of individual identities and may include sensitive terminology.

The final pool of 48 annotators, after exclusions, had an average age of 28 (ranging from 20 to 60) and included 26 females, 28 males, and 1 unspecified gender. Based on simplified ethnicity categories, 21 identified as White, 19 as Black, 4 as Asian, 3 as Mixed, and 1 as Other.

C Method Details

Finding target term sentence positions. For all WiC-embeddings, to find the indices of (the subwords that form) the target word in an example sentence that concerned a non-exact wordform match between target term and example mention (due to inflection or misspellings), we applied two subsequent strategies: 1) we tried to replace the target term with its plural form (through simple rules) and if this plural formation did not result in a match, 2) we tried to find the most similar word in the example sentence (using the difflib library) and replaced that wordform with the target term (as this most

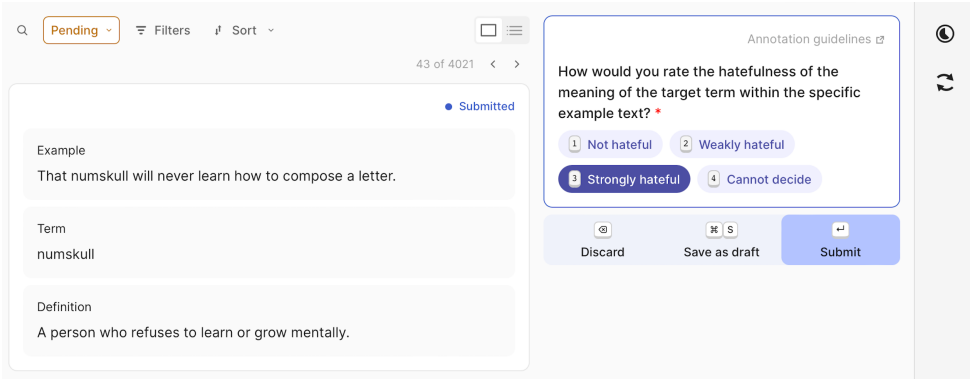

Figure 3: User interface for annotation

Embeddings	BERT		
	Last	All	LastFour
WiC	0.75	0.75	0.75

Table 7: Accuracy on HateWiC classification compared to the **majority** label, with BERT input embeddings consisting of different layer combinations, on the random data test split.

often concerned a misspelling).

Model layer configurations. We also tested the extraction of different layer configurations, since the effectivity of different configurations has shown to differ within lexical semantic tasks (Vulić et al., 2020b). We tested for BERT WiC-embeddings the extraction of: all layers (12 for BERT), last four layers or last layer only. The results in Table 7, demonstrate no effect of layer configuration on the method performance.

MLP classificaton model. The multilayer perceptron model used for classification consisted of four hidden layers with dimensionality 300, 200, 100 and 50, respectively. For training we used the MLPClassifier module from the sklearn libaray and we set the initial learning rate to 0.0005 the maximal number of training iterations to 10. These parameters were selected after a grid search on our development dataset, using sklearn's GridSearchCV module, applied to the following parameter grid: {'hidden_layer_sizes':[(300, 200, 100, 50), (200, 100, 50), (100, 50)], 'learning_rate_init':[0.0005, 0.001, 0.005], 'max_iter': [10, 20, 40, 80, 100, 200]}.

LLaMA 2. The following prompt template was used for leveraging LLaMA 2 for HateWiC Classification.

Instruction:
Given the following sentence that mentions a particular term, classify whether the meaning of that term expresses hate towards a person or group within that specific sentence. Respond with exactly one of the following corresponding labels without an explanation:
"HATEFUL"
"NOT HATEFUL"

Input:
Sentence: [SENTENCE]
Term: [TERM]

Response:

We use the *pipeline* module from the *transformers* library for running the 'text inference' task, where we set the number of return sequences to 1 and the max new tokens to 10; we used the default settings for the remaining parameters.

D Dimension Projection

We also tested the dimension approach of Hoeken et al. (2023b), adapted to our task. In their method for slur detection, they create a "hate dimension" by computing the average over difference vectors between representations of 10 minimal pairs of slurs and non-hateful equivalents (e.g. 'hillbillies' - 'rural people'). Unlike slurs, which generally carry derogatory connotations regardless of context (Hess, 2021), the hateful connotations of other hateful terms are less clear-cut (Frigerio and Tenchini, 2019). This was also illustrated in the conceptual semantic space in Figure 1. Consequently, we did not expect an effective dimension hate dimension to be extractable using pretrained models that encode general word semantics. Additionally, pre-establishing a set of minimal pairs is hardly

feasible for similar reasons.

Our approach. For our task, instead of using a pre-established list of word pairs, we derived this list from the training data. We calculated the cosine similarities between all possible pairs of positive and negative embeddings, i.e. sense representations of hateful and non-hateful training examples, respectively. We then selected pairs with a similarity above a certain threshold to create the dimension, trough the same computation procedure as Hoeken et al. (2023b). After testing a range of thresholds ([0.7, 0.75, 0.8, 0.85, 0.9, 0.95]) on the development set, we set the similarity threshold to 0.9 for testing. Following Hoeken et al. (2023b), we classified positive cosine similarity values between the hate dimension vector and the contextualized word sense representation as hateful, and negative values as non hateful.

Embeddings	**BERT**		**HateBERT**		**WSD bien.**	
	Random	**OoV**	**Random**	**OoV**	**Random**	**OoV**
WiC	0.52	0.53	0.44	0.43	0.44	0.44
Def	0.44	0.43	0.49	0.49	0.32	0.33
WiC+Def	0.49	0.49	0.44	0.45	0.49	0.48

Table 8: Accuracy on HateWiC classification compared to the **majority** label, with **dimension projection** and different input embeddings, tested on a random data split and OoV terms only.

Results. The results of this approach on our HateWic dataset are presented in Table 8, demonstrate low accuracy scores (max. 0.52) and confirm our expectations that a dimension approach as currently implemented is not effective for HateWiC classification.