Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions

Eric Holgate[†] **Isabel Cachola**[‡] **Daniel Preoţiuc-Pietro**[◇] **Junyi Jessy Li**[†]
[†]Department of Linguistics, [‡]Department of Mathematics,
The University of Texas at Austin
{holgate@,isabelcachola@,jessy@austin.}utexas.edu
[◇]Computer and Information Science, University of Pennsylvania
danielpr@sas.upenn.edu

Abstract

Vulgar words are employed in language use for several different functions, ranging from expressing aggression to signaling group identity or the informality of the communication. This versatility of usage of a restricted set of words is challenging for downstream applications and has yet to be studied quantitatively or using natural language processing techniques. We introduce a novel data set of 7,800 tweets from users with known demographic traits where all instances of vulgar words are annotated with one of the six categories of vulgar word use. Using this data set, we present the first analysis of the pragmatic aspects of vulgarity and how they relate to social factors. We build a model able to predict the category of a vulgar word based on the immediate context it appears in with 67.4 macro F1 across six classes. Finally, we demonstrate the utility of modeling the type of vulgar word use in context by using this information to achieve state-of-the-art performance in hate speech detection on a benchmark data set.

1 Introduction

Vulgarity is a common element of conversation (Jay, 2009; Mehl et al., 2007) and is used even more frequently in social networks such as Twitter (Wang et al., 2014). Understanding the motivation behind the choice to be vulgar and the way in which vulgarity is manifested in naturally occurring environments is of interdisciplinary interest. Pragmatic functions that dictate patterns of vulgarity usages may interact with speaker cultural background and demographics. This makes them appealing—and challenging—to model in NLP applications

Yet, to date, there has been no empirical study on the type of vulgar word usage. Research in linguistics and psychology has identified several types of usage for vulgar words (Andersson and Trudgill, 1990; Pinker, 2007; Wang, 2013). These range from use as an intensifier for an opinion or emotion, to offend others, or simply as a way of speaking or to signal the level of (in)formality in a conversation (Pinker, 2007). Table 1 shows example tweets with the six general functions of vulgar word usage.

Function	**Tweet**
Express aggression	<USER> You are an **ass** Your industry is full of assholes and you do nothing to improve (...)
Express emotion	There are so many things I want to do, But investing in equipment is a pain in the **ass**
Emphasise	today is a good **ass** day <URL>
Auxiliary	Wish <USER> could save my **ass** on these exams like he used to
Signal Group Identity	Now this is a group of **ass** kickers!
Non-vulgar	Kick **Ass** 2 - Red Band Trailer <URL>

Table 1: Examples of tweets containing the vulgar word **ass** with six different functions.

We notice that in one of the examples, the vulgar word *ass* is used to verbally abuse another user, while the same word can also be employed to emphasize a feeling ('good *ass* day') or to express an emotion ('pain in the *ass*'). Hence, explicitly modeling vulgar words use is expected to positively impact the performance of practical tasks such as hate speech detection or the way in which profanity filtering is performed.

The goal of our study is to present a comprehensive and multi-faceted analysis of the types of vulgar word usage. To this end, this paper presents:

1. The first data set of written utterances that contain vulgar words, where each vulgar word is labeled for one of six functions of use[1]
2. A quantitative analysis of vulgar word usage across different user demographic traits
3. A machine learning approach to predicting one of six types of vulgar word usage from context
4. Experiments demonstrating that modeling the type of vulgar word usage in context can im-

[1] https://github.com/ericholgate/VulgarFunctionsTwitter

prove predictive performance of the hate speech prediction task on a benchmark data set

Our novel data set contains 7,800 tweets with 8,524 vulgar word labels annotated for one of six functions by seven annotators. We find that the way in which vulgarity is used interacts with user demographic variables such as age or political ideology. We then build a model for predicting the usage type of each vulgar word in the tweet using the tweet context. Finally, we explicitly model the vulgar word usage type in the task of hate speech detection to discriminate between hate speech and tweets including profanity, demonstrating an improvement in predictive accuracy of 3.7 F1. This demonstrates that using insights into vulgar word usage developed in linguistics and psychology, we can achieve quantitative improvements on downstream NLP applications and inform the way models are built and tailored to the task.

2 Related Work

Vulgar language and its uses and pragmatic functions have been studied in several linguistic and psychological studies and the phenomenon has many names. In this paper, we will use *vulgarity*, *profanity*, and *swear/curse words* interchangeably.

Vulgar words were found to be very versatile, with a vulgar word being able to perform different interpersonal functions according to different contexts. Four types of usage are identified in Andersson and Trudgill (1990), including abusive (intended to harm the hearer), expletive (used to express emotions; not directed towards others), humorous (looks like abusive swearing, but has the opposite function) and auxiliary (swearing as a way of speaking, often or always non-emphatic). The five functions of swear words suggested in Pinker (2007) are: dysphemistic (conveying negative sentiment), idiomatic (signaling informality or simply used as a manner of speaking), abusive (intending to offend or harm), emphatic (intending to stress a claim or intensify emotive content) and cathartic (communicating pain). Finally, Wang (2013) identifies four pragmatic roles for profanity with a considerable degree of overlap with Pinker: emoting, emphasizing, aggressing, and group identity signaling.

For the scope of this study, we aim to cover all the functions identified by past research that can be identified from text with restricted content and context such as tweets. Thus, we dropped the cathartic function from Pinker (2007), which is an instantaneous reaction more specific to speech to relieve the effect of physical pain (Stephens et al., 2009). This would thus be very rarely – if ever – expressed through social media and would be very hard to annotate with textual content alone while lacking the broader context of its utterance. We also considered the abuse and aggression functions as equivalent across categorizations, as they imply a face-threatening act (Brown and Levinson, 1987). We considered the auxiliary and idiomatic categories as equivalent across classes, but maintained signaling group identity as in Wang (2013). We also created a non-vulgar classification in case the vulgar word is used in a non-vulgar context (e.g., a name that doubles as a vulgar term).

Due to their affective impact, vulgar words are often used as *strange* synonyms by substituting each other in context or idioms, even when they have no affinity in syntax or meaning (Quang, 1971; Pinker, 2007) (e.g., for God's sake – for fuck's sake; 'I don't give a damn/fuck/shit). This heavily contributes to vulgar word volatility across different functions and higher ambiguity in context. However, this type of usage can allow computational approaches that model the immediate context around a word to generalize across words to functions. To date, there has been no research on quantitatively modeling the **function** of vulgar words in context.

The overall frequency of usage of vulgar words has been quantitatively studied in social media and online communities. For example, Wang et al. (2014) estimates that vulgar posts constitute upwards of 1.15% of tweets and examines vulgar token frequency and how it varies with time, geolocation, and gender. An analysis of profanity across gender and age also appears in Gauthier et al. (2015).

In fact, gender is the most studied sociodemographic factor in relation to the use of profanity. Many studies have shown that male-identifying users employ vulgar terms more frequently than female-identifying users (e.g. Selnow, 1985; Wang et al., 2014).

Jay and Janschewitz (2008) demonstrate that profanity is moderated by pragmatic or contextual factors that go beyond gender, including occupation, social status, and even the nature of the relationship between interlocutors – though this last point, proves difficult to explore via Twitter where the identity of the audience is at least partially obfuscated. Other social factors such as age, religiosity or social status have also been shown to vary with vulgar frequency (McEnery, 2004), as has political ideology (Sylwester and Purver, 2015; Preoţiuc-Pietro et al., 2017). It is thus likely that sociodemographic factors also influence the functions with which vulgar words are used.

This study expands the scope of this type of research, by going beyond simple frequency of usage to pragmatic functions of vulgar words and how they are used differently by different sociodemographic groups.

3 Data

We use social media as our data source as this contains a high level of expression of thoughts, opinions and emotions (Java et al., 2007; Kouloumpis et al., 2011) and represents a platform for observing written interactions and conversations between users (Ritter et al., 2010).

Social media and Twitter in particular provide vast volumes of text which are more informal and less curated compared to other domains such as newswire. An additional advantage of Twitter data is that it allows us to study the sociodemographic context.

3.1 Identifying Vulgar Tweets

We use the corpus of tweets utilized to construct the Vulgar Twitter corpus introduced in prior work (Cachola et al., 2018). Every tweet in this corpus contains at least one vulgar term. We then annotate each instance of a vulgar token for type of use. Note that we use the full dataset of 7,800 tweets which contains 1K more tweets[2] than the released version of the Vulgar Twitter corpus.

The Vulgar Twitter corpus was constructed by identifying tweets containing vulgarity through use of the vulgarity lexicon available at `www.noswearing.com`. A total of 82 tokens were removed from this list as they were deemed not to be unambiguously vulgar after manual inspection.[3] Regular expressions were utilized to identify common intentional spelling variations (e.g., vowel reduplication such as *fuuuuuck* or self-censorship such as *a$$*).

For the complete description of the composition and construction of the Vulgar Twitter corpus, we refer the interested reader to the original paper (Cachola et al., 2018).

3.2 Data Sampling

The Vulgar Twitter corpus overlaps with Preoţiuc-Pietro et al. (2017) which allows us to consider the relationship between sociodemographic features and vulgar functionality. The tweets are compiled from up to 3,200 most recent tweets (per Twitter Developer API) of 4,132 twitter users who provided sociodemographic information via self-report in an online survey (Preoţiuc-Pietro et al., 2017). This sociodemographic data has been utilized in our previous research Cachola et al. (2018); for a full description of its collection, we refer the reader to Preoţiuc-Pietro et al. (2017).

3.3 Demographic Variables and Coding

Demographic information (including gender, age, level of education, level of annual income, faith, an political ideology) was was self-reported via online survey. To control for cultural variation, data was only solicited from residents of the United States. All demographic variables are ordinal with the exception of gender, which is binary.

- **Gender:** a binary[4] variable (Female as 1; Male as 0).

- **Age:** an ordinal, integer-valued variable [13-90].

- **Income:** An ordinal variable [1-8]; the lowest level (1) stands for '< $20,000' and the highest (8) stands for '> $200,000'.

- **Education:** An ordinal variable [1-6]; the lowest level (1) stands for 'no high school degree' and the highest (6) stands for 'Advanced Degree (e.g., Ph.D.)'.

- **Political ideology:** an ordinal variable on the liberal-conservative spectrum (a common form of representation of US political ideology (Ellis and Stimson, 2012)). Reporting options ranged from 'Very Conservative'

[2] These tweets are excluded due to low sentiment agreement in Cachola et al. (2018).

[3] These terms were largely anatomical words or general verbs like *penis*, *vagina*, and *blow*, but some identity descriptors like *gay*, *queer*, and *lesbian* were also excluded after manual review of a large sampling of uses revealed they were not overwhelmingly employed as slurs.

[4] Users were asked to identify their gender as Male, Female, or via an open field. Users who did not respond as either Female or Male were excluded from data collection as there was not a sufficient population to be confident that data would be representative.

Function	**Definition**	**Freq.**
Express aggression (**Agr**)	The word is used in order to harm the person or group the tweet is about.	15.2%
Express emotion (**Emo**)	The word is used to express emotions (positive or negative) related to the users internal states, exclamations, feelings or attitudes towards an object. If removing the vulgar term, the expressed emotion is lacking.	24.8%
Emphasise (**Emp**)	The word is used to emphasize a statement or feeling.	29.8%
Auxiliary (**Aux**)	The use of this word is simply a manner of speaking and does not fit any of the above descriptions. Descriptions of external emotions (those of someone else) fall into this category.	17.0%
Signal Group Identity (**Sig**)	This word is used as a marker of identity in a specific social group.	4.7%
Non-vulgar Use (**Non**)	The use of this word is not vulgar (e.g., named entities that involve vulgar words).	8.2%

Table 2: Functions of vulgar words, their definition presented to the annotators and their frequency in the final data set.

	Agr	**Emo**	**Emp**	**Aux**	**Sig**	**Non**
Agr	0.63	0.11	0.09	0.07	0.10	0.01
Emo	0.07	0.59	0.20	0.13	0.01	0.01
Emp	0.04	0.18	0.68	0.07	0.01	0.02
Aux	0.07	0.16	0.15	0.56	0.03	0.03
Sig	0.17	0.06	0.07	0.11	0.57	0.02
Non	0.02	0.04	0.04	0.10	0.02	0.77

Table 3: Confusion matrix between aggregated function (row) and individual annotations (column), normalized by row.

(1) to 'Moderate' (4) to 'Very Liberal' (7). Two additional responses, 'Other' (8) and 'Apathetic'(9) were included to cover the full breadth of the ideological spectrum, but users selecting these options were excluded from our dataset (1,290 in total) in order to maintain an ordinal scale.

- **Faith:** an ordinal variable [1-6]; users were asked to report the average number of religious services attended. Available responses ranged from 'Never' (1) to 'Multiple times per week' (6).

3.4 Data Processing

We follow the same preprocessing procedure as in Cachola et al. (2018). URL's and usernames are replaced by <URL> and <USER> tokens respectively to protect user privacy. Punctuation is then removed and all words are lowercased.

3.5 Annotation

We have collected annotations via Amazon Mechanical Turk (MTurk) for vulgar word usage type for 8,524 instances of vulgar words across the 7,800 tweets present in the Vulgar Twitter corpus (Cachola et al., 2018).

The task guidelines follow previous research from linguistics and psychology described in Section 2. For generality, we use a union of the different classes proposed and grouped classes where it was possible. The final guidelines include six different functions of vulgar words described in Table 2.

For quality control, we asserted the following qualifications on MTurk: locale=US, approval rate $>$90%, number of HITs approved $>$100. Further, we removed all ratings from users that have a Cohen's Kappa of lower than 0.2 when compared to the majority rating of the other six annotations resulting in the removal of 8,430 ratings (14% of the total number) from 150 out of 663 users. These users were banned and annotations were recollected until 7 ratings were obtained for all instances.

We measured inter-annotator agreement using Krippendorf's Alpha as this can handle cases where each item was labeled by different groups of users. The overall Krippendorfs Alpha is 0.506 despite there being a large number of classes (6). This alpha value (0.506) is regarded as a moderate level of agreement (Artstein and Poesio, 2008). To reduce uncertainty, we aggregate our labels across seven different annotators. In cases where no majority class emerged from the seven annotations (10.6% of the instances), the tie was broken by one of the authors of the paper, who have significant training and experience in linguistic annotation.

The distribution of the final vulgar word type is presented in the last column from Table 2. Table 3 shows the confusion matrix between aggregated function (row) and individual annotations (column); each cell is normalized by the row sum. Some patterns in disagreements include: (1) Vulgar words used to signal group identity are sometimes confused with aggression as annotating these may require additional social context about the user (e.g. if they are female, African-American, etc.) or about social relationships (e.g. *"lmao yeah cause a bitch can't sing"*). (2) Emotion confused with auxiliary usage in idioms or where there is a lack on context about what is

Word	Rank	Entropy	Agr	Emo	Emp	Aux	Sig	Non
bitchy	35	1.547	3	5	2	3	2	0
dicks	33	1.442	7	2	0	6	2	2
bastard	26	1.307	18	3	1	7	1	3
fuck	4	1.272	246	345	190	75	0	0
pussy	21	1.256	17	3	2	33	1	7
ass	5	1.250	116	45	222	352	7	5
hell	2	1.220	16	242	602	71	0	238
dick	9	1.208	36	4	5	87	0	67
bitch	7	1.194	296	23	20	60	110	3
shit	1	1.170	59	555	200	488	1	1

Table 4: Top vulgar words sorted by entropy. Higher entropy indicates a more evenly distributed usage is across functions (maximum entropy over six values = 1.791, minimum entropy over all functions = 0). Rank represents the rank of the word by frequency in the data set.

the author's intent or target (e.g. *"Stop cryin.. Damn you got the foul"*). (3) Auxiliary use of vulgar words in an emotional tweet (e.g, *"ok knicks. we're winning. dont fuck it up."*). (4) Short tweets lacking context drive confusion about the target of vulgarity or if an emotion is expressed (e.g. *"Fuck yeah"*).

4 Analysis

We start with a quantitative analysis of our data. First, we examine the extent to which the same vulgar word is used for different functions. Then, we identify if sociodemographic factors impact the functions with which vulgar words are used.

4.1 Vulgar Word Analysis

To quantitatively measure which vulgar words are most used with different functions, we first compute its distribution over the six functions in our entire data set. Then, we compute the entropy as a measure of how evenly distributed the distribution over functions of each word is. To avoid uncertainly associated computing statistics over distributions with low counts, we keep only words that appear more than 10 times in our data set (43 words) after collapsing variants of the same word (e.g. fuck – f*ck – fuuuck).

The average entropy of all vulgar words is $\mu = 0.835$ ($\sigma = 0.36$), with 0 being the minimum entropy (i.e., all words are used with one function) and 1.791 being the maximum entropy (i.e., all words are used with the same frequency with all six functions). The words with the highest entropy are presented in Table 4.

The table shows that four of the most frequent five words are in the top ten words by entropy, with all of them having significant numbers of occurrences in at least three vulgar functions. Actually, the average entropy of words used at least 100 times in our data set (15 words) is 0.930 compared to 0.835 for words used at least 10 times.

We see that all words in the table are used significantly with three or more functions. On average, in the entire data set, each word is used at least once with $\mu = 4.00$ functions ($\sigma = 1.34$).

This highlights both the challenges in modeling vulgar word functions and the opportunity of using the function to improve practical applications.

In contrast, Table 5 shows the vulgar words which are most likely to be used with each of the six functions.

4.2 Demographic Analysis

Sociodemographic factors may impact the distribution with which each function of vulgar words is used. To measure this, we compute for each user a normalized distribution over the functions of vulgar words used in our data set. Then, we compute Pearson correlation where the dependent variable is the fraction of each vulgar word function and the independent variables are the user sociodemographic trait values. Following previous work (Schwartz and et al., 2013; Preoţiuc-Pietro et al., 2017), for all analyses we consider gender and age basic traits and control for potential data skew by introducing both variables as controls in partial correlation. When studying age and gender, we use the other trait as the control. Since we are running 36 tests at once without pre-stated hypotheses, we correct the correlations for multiple comparisons using Bonferroni correction. Results of these analyses are presented in Table 6.

The results show several vulgar word functions are specific of age. Younger users of Twitter are more likely to use vulgar words to signal group identity and to express emotion. Older age is more likely to be related to use of words that are vulgar with non-vulgar functions. These correlations show that there are differences in how younger generations are using vulgar words, even if tweets were posted in the same time interval, signaling a diachronic change in usage.

The analysis shows that the only significant correlation with the other five demographic variables is between both political ideology and faith and using vulgar words for emphasis and in non-vulgar functions. Liberals are more likely to use vulgar words for emphasis and less likely to use them with non-vulgar functions. Previous research

Aggression		Express Emotion		Emphasis		Auxiliary		Signal Group Identity		Non-Vulgar	
Word	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**
cunt	86.9%	pissed	84.4%	fucking	84.7%	asses	73.9%	bitches	88.9%	mick	100%
asshole	86.3%	bullshit	64.2%	fuckin	84.0%	shitting	69.2%	nigga	85.7%	cracker	97.5%
asshit	83.0%	fucked	61.3%	goddamn	70.0%	arse	69.2%	slut	26.0%	dyke	92.8%
faggot	81.8%	shitty	52.6%	damn	62.3%	cock	62.9%	whore	25.0%	coon	92.8%
fag	73.3%	shit	42.5%	hell	51.4%	pussy	52.3%	hoe	23.8%	ho	88.3%

Table 5: Vulgar words most used with each of the six functions.

Trait	Agr	Emo	Emp	Aux	Sig	Non
Gender	.011	-.005	.004	-.044	.051	.011
Age	-.013	-.085*	-.046	-.036	-.100**	.227**
Education	-.030	.009	-.006	-.007	-.032	.027
Income	.037	.002	.027	-.035	-.045	.032
Faith	-.031	-.047	-.112**	-.066	.014	.224**
Political Ideology	.009	.050	.092**	-.022	.003	-.124**

Table 6: Pearson correlation between user demographic traits and usage of the different functions of vulgar words. All correlations are significant at (*) $p < .05$, (**) $p < .01$, two-tailed t-test, **Bonferroni corrected** for multiple comparisons. Results for education, income and religiosity are controlled for age and gender.

showed that liberals are more likely to use more vulgarity overall in social media (Sylwester and Purver, 2015; Preoţiuc-Pietro et al., 2017) and are perceived by others to use more frequently than they do vulgar words (Carpenter et al., 2016), but this analysis shows this is especially due to vulgar word use to emphasise. The results are reversed for faith, which is known to be strongly correlated to conservative political ideology. Controlling for faith and political ideology with partial correlation does not alter the significance of this result.

Intriguingly, all other traits (gender, education and income) are not significantly correlated with an increased usage in any of the functions.

The vulgar word functions of aggression and auxiliary usage, which are more standard and traditional usages of vulgar words, do not show any significant differences with any sociodemographic trait.

5 Modeling Vulgar Word Use

The previous section showed that the same vulgar words can be used with several different functions. In this section, we use machine learning approaches to explicitly predict the function of a vulgar word given the tweet it appears in as context.

5.1 Method

We use logistic regression[5] to build six one vs. all binary classifiers for each of the six functions using information from the immediate lexical and syntactic context surrounding the word and general usage of the word in training data.

5.2 Features

We use the following feature types in our experiments:

Intention Distribution –We include six features encoding the distribution over intentional classes of the target word in training data, as some words use only several functions and some more predominantly than others.

Tweet Content –We derive a tweet-level representation of the entire content of the tweet by averaging vector representations of its constituent words. We utilize 200-dimensional GloVe embeddings pre-trained on 2B tweets (Pennington et al., 2014).

Sentiment Content –We include two features which represent the number of positive and negative valence words in the tweet, normalized by tweet length. For this feature group, we utilize the opinion lexicon introduced in Hu and Liu (2004).

Part of Speech Context –We encode the part of speech of the target word, the previous word and the next word as one-hot vectors as we expect syntactic information to be an indicator of different functions in context. We extract parts of speech using the Twitter version of the Stanford POS tagger which demonstrated good results on tagging tweets and uses the finer grained Penn Treebank tagset (Derczynski et al., 2013).

Brown Clusters –Finally, we include two one-hot feature groups which indicate the Brown Cluster (Brown et al., 1992) membership of word immediately before and immediately after the vulgar term.

[5] In preliminary experiments, we attempted to utilize a BiLSTM to encode tweet context, but it did significantly worse than the logistic regression model, possibly due to many parameters and classes compared to the size of the training data.

Method	Precision	Recall	F1
Most Frequent Class	5.05	16.6	7.76
All Features	**68.8**	**66.4**	**67.4**
– Intention Distribution	58.3	53.8	55.3
– Tweet Content	67.9	64.0	65.6
– Sentiment Content	68.6	66.3	67.3
– Part of Speech Context	67.8	64.6	65.9
– Brown Clusters	68.6	64.9	66.3

Table 7: Performance statistics for our baseline, predictive model with all features and with ablating each feature group. Precision, Recall and F1 score are macro-averaged across the six classes.

Brown Clusters are obtained by hierarchical clustering tokens based on contexts in which they immediately co-occur. We use the precomputed cluster representations as seen in Turian et al. (2010). We also experimented with Twitter-specific clusters (Owoputi et al., 2012), but found they did not perform as well on our development set.

Additionally, we experimented with a personal pronoun indicator feature in a three word window around the target, a one-hot lexical feature encoding the target vulgar item, and NRC emotion scores (Mohammad and Turney, 2013), but found there to be no improvement in performance as a result. We did not experiment with using the demographic variables as features as these are generally unavailable for use in predictive systems.

5.3 Experimental Results

We split our data into a training set of 6,883 tweets and a testing set of 1,087 tweets, and held out a set of 554 tweets as a validation set on which to test different hyperparameter settings.

Performance statistics for our predictive model are presented in Table 7, as well as ablation experiments for each feature group.

Our predictive model vastly outperforms the most frequent baseline, which uniformly selects the most frequent class overall (emphasis) and scores very low due to the very even distribution over functions. Our best model achieves a macro-averaged F1 score of 67.4 across the six classes.

In the ablation experiments, we see by removing one feature group at a time, which feature type adds most predictive value over others. Withholding the intention distribution feature group from the model shows the greatest loss in performance (12.1 macro F1). This is somewhat expected, as this feature gives a prior distribution over functions for the target word based on training data and, as most words are rarely used with some functions, allows the model to downweigh them. However, even with no word function prior, the predictive performance is still relatively high (55.3 macro F1 across six classes), showing that only the content and context is substantially predictive for the function of a vulgar word

Removing tweet content features or part-of-speech context introduce a similar drop in predictive performance, showing that the overall tweet content and the local syntactic context of the mention play complimentary roles in inference. The sentiment feature groups are the least informative, yielding an negligible increase in performance of only 0.1 macro F1.

The predictive model's F1 scores by class is as follows: *Aggression* – 65.6, *Emotion* – 62.4, *Emphasis* – 76.4, *Signal Group Identity* – 56.5, *Auxiliary* – 62.2, *Not Vulgar* – 81.4.

The highest predictive performance is obtained for vulgar words used in a non vulgar context, which is due to the different tweet content of these tweets and the restricted sets of words which are usually used as non-vulgar. The emphasis functions is the second most accurately predictable using our model, due to the very distinctive syntactic patterns of usage of this function (usually as an adjective). The least predictable function is signaling group identity. We observed that this function is usually used as part of larger conversational context and often relies on a shared social context.

6 Hate Speech Prediction

Finally, we aim to show that modeling the function of vulgar words explicitly has practical implications by using this in a downstream application.

6.1 Task

Automatic hate-speech detection on social media is the task defined as generally identifying abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation (Warner and Hirschberg, 2012) with a clear intention to incite harm, or to promote hatred (Zhang and Luo, 2018). Several data sets and approaches to automatic hate speech detection have been recently proposed (Djuric et al., 2015; Burnap and Williams, 2015; Waseem and Hovy, 2016; Nobata et al., 2016; Davidson et al., 2017).

The task of predicting hate-speech is challenging for natural language processing using lexical information as it aims to predict the intention of the message and several words used in conveying hate speech can have other common uses (Davidson et al., 2017; Malmasi and Zampieri, 2018).

Hate speech is very often confused with offensive language, as highlighted in the error analyses of past hate speech detection papers (Davidson et al., 2017). Quantitative analysis of the machine learning models suggest that obscene words are very informative for both the hate speech and offensive classes of tweets (Malmasi and Zampieri, 2018), hinting that the functions of vulgar words usage are a major source of ambiguity.

Our hypothesis is that explicitly modeling the function a vulgar word has in context will benefit the hate speech prediction task, by differentiating between aggression and other usages.

6.2 Experiments

Data. We use the dataset introduced in Davidson et al. (2017) as this is publicly available, contains tweets collected using vulgar words and explicitly differentiates between offensive tweets and tweets containing hate speech. The three classes in this data set are hate speech, offensive, and neither.

Another public dataset on tweets, introduced in Waseem and Hovy (2016), focuses on specific forms of hate speech (sexist and racist), but is collected with a restricted set of keywords, has low coverage of vulgar words and does not explicitly distinguish between hate speech and other offensive language. Other datasets for this task are not publicly available e.g., Nobata et al. (2016).

Setup. In order to directly measure the impact on predicting performance introduced by explicitly modeling the function of the vulgar word in the tweet, we follow the same methodology to identify hate speech as described in Davidson et al. (2017), as implemented through the openly available code provided by the authors.[6] We thus train three one-vs-all logistic regression classifiers with L2 regularization as implemented in scikit-learn (Pedregosa et al., 2011). Features used in the model include unigram to trigram TF-IDF weighted word features, Part-of-Speech unigram to trigrams, reading level, sentiment words, Twitter specific features (e.g., hashtags, mentions, retweets, and URLs) as well as generic tweet-level features (e.g., number of characters, words, and syllables in each tweet).

	Method	
Class	Davidson et al.	+ vulgar features
Hate Speech	33.6	**39.7**
Offensive	92.1	**93.5**
Neither	82.2	**85.7**
Average	69.3	**73.0**

Table 8: F1 score per class and the macro average.

Vulgar Function Features. We directly and explicitly include the function of the vulgar word present in the tweet by introducing six new features to the hate speech detection model which represent the scores with which the vulgar word is associated with the six functions. If multiple vulgar words exist in a tweet, we use the average predictions over the six functions.

Metrics. We run the model from Davidson et al. (2017) using the provided code on 10-fold cross validation and report the average F1 score for each class as well as the macro-averaged F1 score across all ten folds. Using the available code, we could not reproduce exactly results presented in the Davidson et al. (2017) paper. For predicting the function of the vulgar words from context, we use our best predictive model described in Section 5. We also re-scale our six function features by multiplying them with a large exponent in order to make them significant in model training.

Results. As presented in Table 8, the addition of the six vulgar function features improves the F1 score for each of three classes to up to 6.1 F1 for the hate speech class, which had the lowest performance. This results in an improvement of the macro-F1 score for the entire classification task of 3.7 in F1. This demonstrates the importance of the proposed vulgar function modeling task in detecting hate speech.

7 Conclusion

This paper presents the first empirical study on the pragmatic functions of vulgar words. We created a novel, freely available data set of 7,800 vulgar tweets having 8,524 instances of vulgar words labeled for one of six functions by seven annotators and expert adjudication. We quantitatively showed, leveraging research in linguistics and psychology, that vulgar words are frequently used with different functions and, in the first quantitative analysis on this topic, uncovered that vulgar words are used with different functions by younger users to signal group identity and for expressing emotions.

We have built the first machine learning model

[6] https://github.com/t-davidson/hate-speech-and-offensive-language

for predicting vulgar word function from context, achieving a performance of 67.4 macro F1, demonstrating the practical feasibility of this task. We showed the usefulness of this task, by integrating predicted vulgar word function in the downstream task of hate speech detection, achieving an improvement of 3.7 in F1 on a benchmark data set.

This study showed that modeling pragmatic function is of practical importance. Future work will use this linguistic information to inform more complex machine learning models, e.g., deep neural networks, in an attempt to increase predictive gains. As two of the most used functions of vulgar words relate to expressing sentiment or emotions, we will also explore collecting sentiment annotations for joint sentiment and vulgar word function inference and use this to improve the task of sentiment analysis using multi-task methods.

References

Lars-Gunnar Andersson and Peter Trudgill. 1990. *Bad language*. Penguin.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some Universals in Language Usage*, volume 4. Cambridge University Press.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Pete Burnap and Matthew L Williams. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, 7(2):223–242.

Isabel Cachola, Eric Holgate, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Expressively Vulgar: The Socio-Dynamics of Vulgarity and its Effects on Sentiment Analysis in Social Media. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING, pages 2927–2938.

Jordan Carpenter, Daniel Preoţiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret Kern, Anneke Buffone, Lyle Ungar, and Martin Seligman. 2016. Real Men don't say 'cute': Using Automatic Language Analysis to Isolate Inaccurate Aspects of Stereotypes. *Social Psychological and Personality Science*, 8.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM, pages 512–515.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter Part-of-Speech Tagging for all: Overcoming Sparse and Noisy Data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP, pages 198–206.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW, pages 29–30.

Christopher Ellis and James A Stimson. 2012. *Ideology in America*. Cambridge University Press.

Michael Gauthier, Adrien Guille, A Deseille, and Fabien Rico. 2015. Text Mining and Twitter to Analyze British Swearing Habits. *Handbook of Twitter for Research*.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD, pages 168–177.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65.

Timothy Jay. 2009. The Utility and Ubiquity of Taboo Words. *Perspectives on Psychological Science*, 4(2):153–161.

Timothy Jay and Kristin Janschewitz. 2008. The Pragmatics of Swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Proceedings of the 5th International AAAI Conference on Web and Social Media*, ICWSM, pages 538–541.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Tony McEnery. 2004. *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. Routledge.

Matthias R Mehl, Simine Vazire, Nairán Ramírez-Esparza, Richard B Slatcher, and James W Pennebaker. 2007. Are Women Really more Talkative than Men? *Science*, 317(5834):82–82.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW, pages 145–153.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. 2012. Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP, pages 1532–1543.

Steven Pinker. 2007. *The stuff of thought: Language as a window into human nature*. Penguin.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 729–740.

Phuc Dong Quang. 1971. English Sentences without Overt Grammatical Subject. pages 3–10.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 172–180.

H Andrew Schwartz and et al. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PloS ONE*, 8(9).

Gary W Selnow. 1985. Sex differences in uses and perceptions of profanity. *Sex Roles*, 12(3-4):303–312.

Richard Stephens, John Atkins, and Andrew Kingston. 2009. Swearing as a Response to Pain. *Neuroreport*, 20(12):1056–1060.

Karolina Sylwester and Matthew Purver. 2015. Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. *PLoS ONE*, 10(9).

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word Representations: A Simple and General Method for Semi-Supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 384–394.

Na Wang. 2013. An Analysis of the Pragmatic Functions of ‘swearing’ in Interpersonal Talk. *Griffith Working Papers in Pragmatics and Intercultural Communication*, 6:71–79.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in English on Twitter. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing*, CSCW, pages 415–425.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26.

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*, NAACL, pages 88–93.

Ziqi Zhang and Lei Luo. 2018. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *Semantic Web*.