Why Swear?

Analyzing and Inferring the Intentions of Vulgar Expressions

Eric Holgate†, Isabel Cachola†, Daniel Preotiuc-Pietro*, Junyi Jessy Li†

†University of Texas at Austin
*Bloomberg LP
Motivation

- **Positive_User**
  - EMNLP 2018 is the shit!
  - 2:48 PM - 6 May 2015

- **Negative_User**
  - The weather today is utter shit.
  - 2:48 PM - 6 May 2015

- **Abusive_User**
  - Don't @ me you piece of shit.
  - 2:48 PM - 6 May 2015
Motivation

1. Vulgarity is employed purposefully
2. Vulgarity is used for various pragmatic goals
3. Vulgarity is prevalent in daily communication
Motivation

- People use vulgarity for various pragmatic functions (intentions)
- Several linguists and psychologists have studied these roles
  - Anderson & Trudgill (1990): Four functions
  - Pinker (2007): Five functions
  - Wang (2013): Four functions
- We aim to build upon this and introduce the first computational approach to this problem.
- We introduce a data set of 8,524 instances of vulgar words annotated with one of six roles
Pragmatic Roles

**Aggression (15.2%)**:

- *The word is used in order to harm the person or group the tweet is about*
Pragmatic Roles

**Emotion (24.8%)**

- The word is used to express emotions (positive or negative) related to the user’s internal states, exclamations, feelings or attitude towards an object.
  - If the vulgar token is removed, the emotion is too.

![Twitter Emotion Example](image)
Pragmatic Roles

**Emphasis (29.8%)**

- The word is used to emphasize a statement or feeling
Pragmatic Roles

**Auxiliary Use (17.0%)**

- The use of this word is simply a manner of speaking.
  - Descriptions of external emotions (i.e., those of someone else) fall into this category
Pragmatic Roles

**Signaling Group Identity (4.7%)**

- The word is used to mark membership in a social group
  - This includes reappropriative usage of slurs

---

*Group Identity Signal*  
@Identifying_User

**Now this is a group of ass kickers!**

2:48 PM - 6 May 2015
Pragmatic Roles

Non-Vulgar (8.2%)

- The use of this word is not vulgar
  - i.e., named entities
Pragmatic Roles

1. Aggression
2. Emotion
3. Emphasis
4. Auxiliary Use
5. Signaling Group Identity
6. Non-Vulgar
Research Questions

1. Do demographic factors impact why users employ vulgarity?
2. Can we predict why users employ vulgarity?
3. Is modeling vulgar intent useful for NLP tasks?
We introduce a data set of 8,524 instances of vulgar words annotated with one of six roles:

- Across 7,800 tweets
- Sourced from 4,132 users with demographic info (Preotiuc-Pietro et al., 2017)
  - Gender, age, education, income level, faith, political ideology
- Vulgarity defined with a list from www.noswearing.com
  - Regular expressions include spelling variation and self-censorship e.g., `damnnnen` or `a$$`
Data

• Annotated for vulgar intention
  – MTurk
  – IAA - Krippendorf’s Alpha of 0.506
  – 7 annotations-instance
  – QC - Excluded annotators with <0.2 agreement with the majority of others
  – Majority vote aggregation, ties were split by one of the co-authors
• Available at: https://github.com/ericholgate/VulgarFunctionsTwitter
• Sentiment annotation for 6,800 tweets from the same corpus is also available
  (Cachola et al., 2018): https://github.com/ericholgate/vulgartwitter
Research Questions

1. Do demographic factors impact why users employ vulgarity?

2. Can we predict why users employ vulgarity?

3. Is modeling vulgar intent useful for NLP tasks?
Demographic Analysis

Pearson correlation

• Dependent variable
  – fraction of vulgar function use
• Controlled for age & gender
• Bonferroni corrected
  – account for multiple comparisons
Demographic Analysis

Pearson correlation

- Dependent variable
  - fraction of vulgar function use
- Controlled for age & gender
- Bonferroni corrected
  - account for multiple comparisons

![Pearson Correlation - Emotion Graph]
Demographic Analysis

Pearson correlation

- Dependent variable
  - fraction of vulgar function use
- Controlled for age & gender
- Bonferroni corrected
  - account for multiple comparisons
Demographic Analysis

Pearson correlation

- Dependent variable
  - fraction of vulgar function use
- Controlled for age & gender
- Bonferroni corrected
  - account for multiple comparisons
**Demographic Analysis**

**Pearson correlation**

- Dependent variable
  - fraction of vulgar function use
- Controlled for age & gender
- Bonferroni corrected
  - account for multiple comparisons
Demographic Analysis

Pearson correlation

- Dependent variable
  - fraction of vulgar function use
- Controlled for age & gender
- Bonferroni corrected
  - account for multiple comparisons

Pearson Correlation - Non-Vulgar

<table>
<thead>
<tr>
<th>Trait</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0.011</td>
</tr>
<tr>
<td>Age</td>
<td>0.227</td>
</tr>
<tr>
<td>Education</td>
<td>0.027</td>
</tr>
<tr>
<td>Income</td>
<td>0.032</td>
</tr>
<tr>
<td>Faith</td>
<td>0.224</td>
</tr>
<tr>
<td>Political Ideology</td>
<td>0.124</td>
</tr>
</tbody>
</table>
Demographic Analysis

- Younger users
  - [+ signal group identity
  - [+ emotion
  - [- non-vulgar
- Politically liberal users:
  - [+ emphasis
- Religious users
  - [- emphasis
- Gender, Education, & Income
  - No effects
Research Questions

Do demographic factors impact why users employ vulgarity?

Yes!
Research Questions

1. Do demographic factors impact why users employ vulgarity?
2. Can we predict why users employ vulgarity?
3. Is modeling vulgar intent useful for NLP tasks?
Prediction: Features

- Vulgar token features
  - Intention distribution from training data
- Global tweet features
  - Tweet content (average GloVe embeddings)
- Local word context
  - Sentiment
  - Part-of-Speech (trigrams around the target token)
  - Brown Clusters (previous and next token)
Predicting Vulgar Function

- Some words are used with predominately one function
- The most frequent, however, are distributed amongst all the functions
Predicting Vulgar Function

• Features
  – Vulgar token features
    • Intention distribution from training data
  – Global tweet features
    • Tweet content (average GloVe embeddings)
  – Local word context
    • Sentiment
    • Part-of-Speech (trigrams around the target token)
    • Brown Clusters (previous and next token)
Predicting Vulgar Function

- Vulgar instances centered around product reviews tend to be emotive or emphatic
- Conversational tweets are more frequently auxiliary
Features

- **Features**
  - Vulgar token features
    - Intention distribution from training data
  - Global tweet features
    - Tweet content (average GloVe embeddings)
  - Local word context
    - Sentiment
    - Part-of-Speech (trigrams around the target token)
    - Brown Clusters (previous and next token)
Intuition

- Many vulgar words can even be substituted for one another, even when they don’t have any concrete meaning:
  - Who the *hell/fuck*
  - I don’t give a *damn/shit/fuck*
- Pinker (2007) calls these *strange synonyms*
Predicting Vulgar Functions

- Logistic regression classification
  - six one vs. all binary classifiers
- Data Split:
  - Train: 6,883
  - Test: 1,087
  - Val: 554
- A BiLSTM-based approach did not yield improvement
Predicting Vulgar Functions

Ablation Study

• Intention distribution contributes unique information
• Other features are complementary
Research Questions

Can we predict why users employ vulgarity?

Yes!
Research Questions

1. Do demographic factors impact why users employ vulgarity?
2. Can we predict why users employ vulgarity?
3. Is modeling vulgar intent useful for NLP tasks?
Vulgar Intention and Hate Speech

• Hate Speech detection
  – Downstream task for vulgar function prediction
• Dataset introduced by Davidson et al. (2017)
  – 24,802 tweets labeled with one of the three classes:
    • hate speech
    • offensive
    • neither
  – all tweets contain vulgar words
Vulgar Intention and Hate Speech

• Logistic regression model from Davidson et al (2017) using:
  – TF-IDF weighted token features
  – POS unigram to trigrams
  – reading level metrics
  – sentiment information
  – Twitter features (hashtags, mentions, etc.)
  – generic tweet features (character, word and syllable counts)
• We add an explicit vulgarity feature group:
  – The predicted distribution over vulgar functions (6 features)
    • averaged if more than one vulgar token/tweet
Vulgar Intention and Hate Speech

Results

• The addition of vulgarity features yields improvement in all three classes
• These features are most influential for detecting hate speech, the class with the lowest accuracy
Research Questions

Is modeling vulgar intent useful for NLP tasks?

Yes!
Take Aways

• Vulgarity is used with several pragmatic functions
  – We can predict these from context
  – Vulgar intent is useful for downstream tasks like hate speech detection

• New data set focused on vulgarity functions

• Sociodemographic features impact vulgar role usage
Thank You!

We hope you thought this talk was damn interesting!

holgate@utexas.edu
## Annotation Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Agr</th>
<th>Emo</th>
<th>Emp</th>
<th>Aux</th>
<th>Sig</th>
<th>Non</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agr</td>
<td>0.63</td>
<td>0.11</td>
<td>0.09</td>
<td>0.07</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>Emo</td>
<td>0.07</td>
<td>0.59</td>
<td>0.20</td>
<td>0.13</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Emp</td>
<td>0.04</td>
<td>0.18</td>
<td>0.68</td>
<td>0.07</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Aux</td>
<td>0.07</td>
<td>0.16</td>
<td>0.15</td>
<td>0.56</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Sig</td>
<td>0.17</td>
<td>0.06</td>
<td>0.07</td>
<td>0.11</td>
<td>0.57</td>
<td>0.02</td>
</tr>
<tr>
<td>Non</td>
<td>0.02</td>
<td>0.04</td>
<td>0.04</td>
<td>0.10</td>
<td>0.02</td>
<td>0.77</td>
</tr>
</tbody>
</table>
### Most Frequent Terms by Function

<table>
<thead>
<tr>
<th>Aggression</th>
<th>Express Emotion</th>
<th>Emphasis</th>
<th>Auxiliary</th>
<th>Signal Group Identity</th>
<th>Non-Vulgar</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td><strong>Freq</strong></td>
<td><strong>Word</strong></td>
<td><strong>Freq</strong></td>
<td><strong>Word</strong></td>
<td><strong>Freq</strong></td>
</tr>
<tr>
<td>cunt</td>
<td>86.9%</td>
<td>pissed</td>
<td>84.4%</td>
<td>fucking</td>
<td>84.7%</td>
</tr>
<tr>
<td>asshole</td>
<td>86.3%</td>
<td>bullshit</td>
<td>64.2%</td>
<td>fuckin</td>
<td>84.0%</td>
</tr>
<tr>
<td>ass</td>
<td>83.0%</td>
<td>fucked</td>
<td>61.3%</td>
<td>goddamn</td>
<td>70.0%</td>
</tr>
<tr>
<td>faggot</td>
<td>81.8%</td>
<td>shitty</td>
<td>52.6%</td>
<td>damn</td>
<td>62.3%</td>
</tr>
<tr>
<td>fag</td>
<td>73.3%</td>
<td>shit</td>
<td>42.5%</td>
<td>hell</td>
<td>51.4%</td>
</tr>
</tbody>
</table>

- Some words are very consistently used with a specific function
  - For these words, this feature will be very predictive
- For words like ass, which are very diverse, this feature will be less informative.