Vitriol on Social Media: Curation and Investigation

Xing Zhao and James Caverlee
Department of Computer Science and Engineering
Texas A&M University, USA
What we were promised
What we were promised
What we were promised
Antisocial Behaviors Online
Antisocial Behaviors Online

Hate Speech  Attacking on a target’s race, religion, ethnic origin, sexual orientation, etc. (Banks’10, Warner’12, Burnap’15, Nobata’16, Serra’17)
Antisocial Behaviors Online

**Hate Speech** Attacking on a target’s race, religion, ethnic origin, sexual orientation, etc. (Banks’10, Warner’12, Burnap’15, Nobata’16, Serra’17)

**Flaming** A hostile and insulting interaction between persons, often involving the use of profanity. It can also be the swapping of insults back and forth or with many people teaming up on a single victim. (Kayany’98, Lee’05)
Antisocial Behaviors Online

**Hate Speech**  Attacking on a target’s race, religion, ethnic origin, sexual orientation, etc. (Banks’10, Warner’12, Burnap’15, Nobata’16, Serra’17)

**Bullying**  Using strength or influence to harm or intimidate those who are weaker, often including persistent and targeted behaviors to induce harm in another. (Lieberman’11, Macbeth’13)

**Flaming**  A hostile and insulting interaction between persons, often involving the use of profanity. It can also be the swapping of insults back and forth or with many people teaming up on a single victim. (Kayany’98, Lee’05)
Antisocial Behaviors Online

**Hate Speech** Attacking on a target’s race, religion, ethnic origin, sexual orientation, etc. (Banks’10, Warner’12, Burnap’15, Nobata’16, Serra’17)

**Bullying** Using strength or influence to harm or intimidate those who are weaker, often including persistent and targeted behaviors to induce harm in another. (Lieberman’11, Macbeth’13)

**Flaming** A hostile and insulting interaction between persons, often involving the use of profanity. It can also be the swapping of insults back and forth or with many people teaming up on a single victim. (Kayany’98, Lee’05)

**Trolling** Unleashing one or more cynical or sarcastic remarks on an innocent by-stander. (Hardaker’10, Cheng’15, Cheng’17)
Our Interest: Vitriol
Our Interest: Vitriol

- **Hate Speech?** Nothing related to recipient’s race, religion, ethnic origin, or sexual orientation
Our Interest: Vitriol

- **Hate Speech?** Nothing related to recipient’s race, religion, ethnic origin, or sexual orientation

- **Bullying?** Ordinary -> Famous
Our Interest: Vitriol

- **Hate Speech?** Nothing related to recipient’s race, religion, ethnic origin, or sexual orientation
- **Bullying?** Ordinary -> Famous
- **Flaming?** Individual and unidirectional behavior
Our Interest: Vitriol

- **Hate Speech?** Nothing related to recipient’s race, religion, ethnic origin, or sexual orientation
- **Bullying?** Ordinary -> Famous
- **Flaming?** Individual and unidirectional behavior
- **Trolling?** Straightforward, savage, and crude
Our Interest: Vitriol
Our Interest: Vitriol

• a persistent coarsening of the discourse that leads to a caustic, corrosive, and negative experience on social media;
Our Interest: Vitriol

• **a persistent coarsening** of the discourse that leads to a caustic, corrosive, and negative experience on social media;

• needs not rise to the level of hate speech, is not similar to flaming and bullying, but is much more severe than trolling;
Our Interest: Vitriol

• a persistent coarsening of the discourse that leads to a caustic, corrosive, and negative experience on social media;

• needs not rise to the level of hate speech, is not similar to flaming and bullying, but is much more severe than trolling;

• is existing on every social media and content-based platform.
Research Questions of this work

RQ 1. How can we define Vitriol?

RQ 2. How to curate vitriol dataset?

RQ 3. What are characteristics of Vitriol?

RQ 4. Can we detect vitriol on social media?
RQ 1. How can we define Vitriol?

Oftentimes it is personal,
RQ 1. How can we define Vitriol?

off-topic,
RQ 1. How can we define Vitriol? and unidirectional.
RQ 1. How can we define Vitriol?

- We define vitriol as a broad antisocial behavior online, which is:
RQ 1. How can we define Vitriol?

- We define vitriol as a broad antisocial behavior online, which is:
  - **Personal**: the post should target another user, rather than just “shouting to the wind”;
RQ 1. How can we define Vitriol?

- We define vitriol as a broad antisocial behavior online, which is:

  - **Personal**: the post should target another user, rather than just “shouting to the wind”;

  - **Context-free**: the post should ignore the substance of what the recipient cares about (the context); and
RQ 1. How can we define Vitriol?

We define vitriol as a broad antisocial behavior online, which is:

- **Personal**: the post should target another user, rather than just “shouting to the wind”;

- **Context-free**: the post should ignore the substance of what the recipient cares about (the context); and

- **Unilateral**: the post should be **one-way** from a vitriolic user (anonymous) to a target user (popular account).
RQ 2. How to curate vitriol dataset?

- Why is it hard to detect?
RQ 2. How to curate vitriol dataset?

- Why is it hard to detect?

<table>
<thead>
<tr>
<th>Vitriol</th>
<th>Non-Vitriol</th>
</tr>
</thead>
<tbody>
<tr>
<td>@WolfForPA Then if you want to switch back to produce the soil ruined. Nice going, moron</td>
<td>@josel767 @rosariolopezn And remember kids, you’ll always be shit, but you wanna be the best shit to have ever been created [emoji]</td>
</tr>
<tr>
<td>@RCorbettMEP Your peddling fear mongering bull shit. You don’t mention Fracki that’s a serious ecological risk. You arrogantly assume the ..</td>
<td>@Rival Laxno @HypeWicked @VillainGoofys Holy shit theirs been hella beef today</td>
</tr>
</tbody>
</table>
RQ 2. How to curate vitriol dataset?

- Why is it hard to detect?

<table>
<thead>
<tr>
<th>Vitriol</th>
<th>Non-Vitriol</th>
</tr>
</thead>
<tbody>
<tr>
<td>@WolfForPA Then if you want to switch back to produce the soil ruined. Nice going, moron</td>
<td>@josel767 @rosariolopezn And remember kids, you’ll always be shit, but you wanna be the best shit to have ever been created [emoji]</td>
</tr>
<tr>
<td>@RCorbettMEP Your peddling fear mongering bull shit. You don’t mention Fracki that’s a serious ecological risk. You arrogantly assume the ..</td>
<td>@Rival Laxno @HypeWicked @VillainGoofys Holy shit theirs been hella beef today</td>
</tr>
</tbody>
</table>

**HARD!** Many previous methods for extracting abusive language have focused only on **content-based features**, and yet, some profanity can be a **banter** or just **joking**.
RQ 2. How to curate vitriol dataset?

- Vitriol Curation Frame
RQ 2. How to curate vitriol dataset?

- Vitriol Curation Frame

Potential Vitriol:
Tweets: 3,336,477
User: 1,720,420
RQ 2. How to curate vitriol dataset?

- Vitriol Curation Frame
RQ 2. How to curate vitriol dataset?

- Vitriol Curation Frame

- Content Level
- Tweets Level
- Users Level

- Content filter
- Format
- Originality
- Activeness
- Unilaterality
RQ 2. How to curate vitriol dataset?

- Vitriol Curation Frame

- Content Level
- Tweets Level
- Users Level

Tweets: 14,001
User: 926
RQ 2. How to curate vitriol dataset?

- Vitriol Curation Frame

Note:
1. Filters are strict to surface the real vitriolic tweets
2. False negatives exist since we care about “precision” more than “recall”.
RQ 2. How to curate vitriol dataset?

- Vitriol Dataset

**Dataset of Vitriol**
- Duration: June 30th 2017 to September 14th 2017
- Validation: precision > 95% by sample annotations.

<table>
<thead>
<tr>
<th></th>
<th>Tweets</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential Vitriolic</td>
<td>3,336,477</td>
<td>1,720,281</td>
</tr>
<tr>
<td>Potential Targeted</td>
<td>2,883,092</td>
<td>1,374,420</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitriolic</td>
<td>14,001</td>
<td>926</td>
</tr>
<tr>
<td>Targeted</td>
<td>11,938</td>
<td>3,188</td>
</tr>
</tbody>
</table>
RQ 3. What are characteristics of Vitriol?

- RQ 3-1. Vitriolic Posts
RQ 3. What are characteristics of Vitriol?

- **RQ 3-1. Vitriolic Posts**

- **Emotions**

![Bar chart showing the score of likelihood for different emotions in non-vitriolic and vitriolic posts.](chart.png)
RQ 3. What are characteristics of Vitriol?

- RQ 3-1. Vitriolic Posts

- Emotions

- Social Tendencies
RQ 3. What are characteristics of Vitriol?

- RQ 3-2. Vitriolic Users
RQ 3. What are characteristics of Vitriol?

- RQ 3-2. Vitriolic Users

**Popularity**

![Box plots showing follower count and friend count for non-vitriolic and vitriolic users.](image)
RQ 3. What are characteristics of Vitriol?

- RQ 3-2. Vitriolic Users

**Popularity**

**Activeness**
RQ 3. What are characteristics of Vitriol?

- RQ 3-3. Vitriolic Recipients (categories)

Expected, considering the divisiveness of politics, news and opinions among many people.
RQ 3. What are characteristics of Vitriol?

- RQ 3-3. Vitriolic Recipients (individuals)

Unsurprisingly, most of these users are composed of politicians and news media accounts.
RQ 4. Can we detect vitriol on social media?

- To build our classifier to distinguish vitriolic tweets from others, we adopt 4 categories of features which can help us to characterize vitriol:

Read paper for more details about experimental settings.
RQ 4. Can we detect vitriol on social media?

- To build our classifier to distinguish vitriolic tweets from others, we adopt 4 categories of features which can help us to characterize vitriol:

  LP: Language Pattern (# = 25)
  Part-of-Speech Tagging such as common noun, verb, adjective, emoji, etc.

Read paper for more details about experimental settings
RQ 4. Can we detect vitriol on social media?

- To build our classifier to distinguish vitriolic tweets from others, we adopt 4 categories of features which can help us to characterize vitriol:

  **LP: Language Pattern (# = 25)**
  Part-of-Speech Tagging such as common noun, verb, adjective, emoji, etc.

  **CS: Communication Sentiment (# = 14)**
  Sentiment Analysis such as anger, joy, openness, sadness, etc.

Read paper for more details about experimental settings
RQ 4. Can we detect vitriol on social media?

- To build our classifier to distinguish vitriolic tweets from others, we adopt 4 categories of features which can help us to characterize vitriol:

  **LP: Language Pattern (# = 25)**
  Part-of-Speech Tagging such as common noun, verb, adjective, emoji, etc.

  **CS: Communication Sentiment (# = 14)**
  Sentiment Analysis such as anger, joy, openness, sadness, etc.

  **CR: Content Relevance (# = 100)**
  Latent features from Doc2Vec for learning a distributed representation

Read paper for more details about experimental settings
RQ 4. Can we detect vitriol on social media?

To build our classifier to distinguish vitriolic tweets from others, we adopt 4 categories of features which can help us to characterize vitriol:

**LP: Language Pattern (# = 25)**
Part-of-Speech Tagging such as common noun, verb, adjective, emoji, etc.

**CS: Communication Sentiment (# = 14)**
Sentiment Analysis such as anger, joy, openness, sadness, etc.

**CR: Content Relevance (# = 100)**
Latent features from Doc2Vec for learning a distributed representation.

**LT: Latent Topic (# = 10)**
Latent topics from LDA model, which allows both LDA model estimation from a training corpus and inference of topic distribution on new, unseen documents.

Read paper for more details about experimental settings.
RQ 4. Can we detect vitriol on social media?

- **Experiment Setting**
  - **Data Splitting:** 80% Training, 20% Testing
  - **Procedure:** 5-fold cross validation
  - **Classifier:** Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Multi-layer Perceptron (MLP)
  - **Evaluation Measures:** F-1, AUC
**RQ 4. Can we detect vitriol on social media?**

- The F-1 score for Vitriol vs. Non-Vitriol

<table>
<thead>
<tr>
<th></th>
<th>Log-Reg</th>
<th>SVM</th>
<th>RF</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>0.8290</td>
<td>0.8323</td>
<td>0.8424</td>
<td>0.8384</td>
</tr>
<tr>
<td>LP</td>
<td>0.7636</td>
<td>0.8053</td>
<td>0.8028</td>
<td>0.8061</td>
</tr>
<tr>
<td>CR</td>
<td>0.8486</td>
<td>0.8386</td>
<td>0.8492</td>
<td>0.8597</td>
</tr>
<tr>
<td>LT</td>
<td>0.5761</td>
<td>0.5861</td>
<td>0.7265</td>
<td>0.6642</td>
</tr>
<tr>
<td>LP-CS</td>
<td>0.8564</td>
<td>0.8832</td>
<td>0.8886</td>
<td>0.8865</td>
</tr>
<tr>
<td>LP-CR-LT</td>
<td>0.8780</td>
<td>0.8810</td>
<td>0.8832</td>
<td><strong>0.8978</strong></td>
</tr>
<tr>
<td>all</td>
<td>0.8983</td>
<td>0.9007</td>
<td>0.9050</td>
<td><strong>0.9200</strong></td>
</tr>
</tbody>
</table>
### RQ 4. Can we detect vitriol on social media?

#### The F-1 score for Vitriol vs. Non-Vitriol

<table>
<thead>
<tr>
<th></th>
<th>Log-Reg</th>
<th>SVM</th>
<th>RF</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>0.8290</td>
<td>0.8323</td>
<td>0.8424</td>
<td>0.8384</td>
</tr>
<tr>
<td>LP</td>
<td>0.7636</td>
<td>0.8053</td>
<td>0.8028</td>
<td>0.8061</td>
</tr>
<tr>
<td>CR</td>
<td>0.8486</td>
<td>0.8386</td>
<td>0.8492</td>
<td>0.8597</td>
</tr>
<tr>
<td>LT</td>
<td>0.5761</td>
<td>0.5861</td>
<td>0.7265</td>
<td>0.6642</td>
</tr>
<tr>
<td>LP-CS</td>
<td>0.8564</td>
<td>0.8832</td>
<td>0.8886</td>
<td>0.8865</td>
</tr>
<tr>
<td>LP-CR-LT</td>
<td>0.8780</td>
<td>0.8810</td>
<td>0.8832</td>
<td><strong>0.8978</strong></td>
</tr>
<tr>
<td>all</td>
<td>0.8983</td>
<td>0.9007</td>
<td>0.9050</td>
<td><strong>0.9200</strong></td>
</tr>
</tbody>
</table>

*Best classifier overall*
RQ 4. Can we detect vitriol on social media?

- The F-1 score for Vitriol vs. Non-Vitriol

<table>
<thead>
<tr>
<th>Feature</th>
<th>Log-Reg</th>
<th>SVM</th>
<th>RF</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>0.8290</td>
<td>0.8323</td>
<td>0.8424</td>
<td>0.8384</td>
</tr>
<tr>
<td>LP</td>
<td>0.7636</td>
<td>0.8053</td>
<td>0.8028</td>
<td>0.8061</td>
</tr>
<tr>
<td>CR</td>
<td>0.8486</td>
<td>0.8386</td>
<td>0.8492</td>
<td>0.8597</td>
</tr>
<tr>
<td>LT</td>
<td>0.5761</td>
<td>0.5861</td>
<td>0.7265</td>
<td>0.6642</td>
</tr>
<tr>
<td>LP-CS</td>
<td>0.8564</td>
<td>0.8832</td>
<td>0.8886</td>
<td>0.8865</td>
</tr>
<tr>
<td>LP-CR-LT</td>
<td>0.8780</td>
<td>0.8810</td>
<td>0.8832</td>
<td>0.8978</td>
</tr>
<tr>
<td>all</td>
<td>0.8983</td>
<td>0.9007</td>
<td>0.9050</td>
<td>0.9200</td>
</tr>
</tbody>
</table>

Best classifier overall: MLP

Best individual feature: LP-CR-LT
RQ 4. Can we detect vitriol on social media?

- The F-1 score for Vitriol vs. Non-Vitriol

<table>
<thead>
<tr>
<th></th>
<th>Log-Reg</th>
<th>SVM</th>
<th>RF</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>0.8290</td>
<td>0.8323</td>
<td>0.8424</td>
<td>0.8384</td>
</tr>
<tr>
<td>LP</td>
<td>0.7636</td>
<td>0.8053</td>
<td>0.8028</td>
<td>0.8061</td>
</tr>
<tr>
<td>CR</td>
<td>0.8486</td>
<td>0.8386</td>
<td>0.8492</td>
<td>0.8597</td>
</tr>
<tr>
<td>LT</td>
<td>0.5761</td>
<td>0.5861</td>
<td>0.7265</td>
<td>0.6642</td>
</tr>
<tr>
<td>LP-CS</td>
<td>0.8564</td>
<td>0.8832</td>
<td>0.8886</td>
<td>0.8865</td>
</tr>
<tr>
<td>LP-CR-LT</td>
<td>0.8780</td>
<td>0.8810</td>
<td>0.8832</td>
<td><strong>0.8978</strong></td>
</tr>
<tr>
<td>all</td>
<td>0.8983</td>
<td>0.9007</td>
<td>0.9050</td>
<td><strong>0.9200</strong></td>
</tr>
</tbody>
</table>

- Best classifier overall: **MLP**
- Best individual feature: **Log-Reg**
- The performance of our classifier avoiding bias introduced by the curation strategies (content filter): **LP-CR-LT**
RQ 4. Can we detect vitriol on social media?

- The F-1 score for Vitriol vs. Non-Vitriol

<table>
<thead>
<tr>
<th></th>
<th>Log-Reg</th>
<th>SVM</th>
<th>RF</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>0.8290</td>
<td>0.8323</td>
<td>0.8424</td>
<td>0.8384</td>
</tr>
<tr>
<td>LP</td>
<td>0.7636</td>
<td>0.8053</td>
<td>0.8028</td>
<td>0.8061</td>
</tr>
<tr>
<td>CR</td>
<td>0.8486</td>
<td>0.8386</td>
<td>0.8492</td>
<td>0.8597</td>
</tr>
<tr>
<td>LT</td>
<td>0.5761</td>
<td>0.5861</td>
<td>0.7265</td>
<td>0.6642</td>
</tr>
<tr>
<td>LP-CS</td>
<td>0.8564</td>
<td>0.8832</td>
<td>0.8886</td>
<td>0.8865</td>
</tr>
<tr>
<td>LP-CR-LT</td>
<td>0.8780</td>
<td>0.8810</td>
<td>0.8832</td>
<td><strong>0.8978</strong></td>
</tr>
<tr>
<td>all</td>
<td>0.8983</td>
<td>0.9007</td>
<td>0.9050</td>
<td><strong>0.9200</strong></td>
</tr>
</tbody>
</table>

Best classifier overall

Best individual feature

The performance of our classifier avoiding bias introduced by the curation strategies (content filter)

Best performance overall
Conclusion

• This work is an initial attempt at formally studying vitriol on social media.

• RQ 1. We formally defined the vitriol behavior on social media.

• RQ 2. We designed a vitriol curation framework as an initial step to extract vitriolic posts from social media with high confidence.

• RQ 3. We investigated a large collection of vitriolic posts sampled from Twitter, and examined user-level, post-level, and recipient-level characteristics.

• RQ 4. We built a corresponding classifier for distinguishing vitriol on social media.
Conclusion

• This work is an initial attempt at formally studying vitriol on social media.
Conclusion

• This work is an initial attempt at formally studying vitriol on social media.

• RQ 1. We formally defined the vitriol behavior on social media.
Conclusion

• This work is an initial attempt at formally studying vitriol on social media.
• RQ 1. We formally defined the vitriol behavior on social media.
• RQ 2. We designed a vitriol curation framework as an initial step to extract vitriolic posts from social media with high confidence.
Conclusion

• This work is an initial attempt at formally studying vitriol on social media.

• RQ 1. We formally defined the vitriol behavior on social media.

• RQ 2. We designed a vitriol curation framework as an initial step to extract vitriolic posts from social media with high confidence.

• RQ 3. We investigated a large collection of vitriolic posts sampled from Twitter, and examined user-level, post-level, and recipient-level characteristics.
Conclusion

• This work is an initial attempt at formally studying vitriol on social media.
• RQ 1. We formally defined the vitriol behavior on social media.
• RQ 2. We designed a vitriol curation framework as an initial step to extract vitriolic posts from social media with high confidence.
• RQ 3. We investigated a large collection of vitriolic posts sampled from Twitter, and examined user-level, post-level, and recipient-level characteristics.
• RQ 4. We built a corresponding classifier for distinguishing vitriol on social media.
Future work

• Expand our vitriol dataset by using well-designed statistical methods to relax our requirements.

• Temporally track the behaviors of our vitriolic users, e.g., to explore how many of them have been ultimately suspended.

• Study vitriol from the perspective of the users who are the targets of vitriol; are there strategies to incite or minimize the number of vitriolic attacks?
Thank you!
Q&A