Hate Speech Detection: the LiLaH Project

Walter Daelemans
CLiPS, University of Antwerp
walter.daelemans@uantwerpen.be
Outline

• The LiLaH project
• Datasets and lexical resources
• Approaches to (implicit) hate speech detection
  • Stylometric and emotion-based features
  • Hateful metaphors
  • The role of context in hate speech detection

• Antwerp Text Mining Center (TEXTUA)
  • A core facility of the University of Antwerp
The LiLaH project

• LiLaH: The **Linguistic Landscape of Hate Speech in Social Media**
• FWO + SSF project
• Cooperation:
  • University of Antwerp (Belgium) [*Ilia Markov, Walter Daelemans*]
  • University of Ljubljana (Slovenia) [*Darja Fišer*]
  • Jozef Stefan Institute (Slovenia) [*Nikola Ljubešić*]
• Hate speech against migrants and LGBT community
• Website: [lilah.eu](http://lilah.eu)
Datasets

• **Languages:**
  • English, Slovene, Croatian: the FRENK datasets (Ljubešić et al., 2019)
  • Dutch: the LiLaH dataset

• **Data collection:**
  • Facebook comment threads to *mainstream media posts*
  • Hate speech directed towards LGBT community and migrants
  • (Keyword-based retrieval)

• **Annotation:**
  • Annotation guidelines, trained annotators (8 annotators for English, Slovene, Croatian; 3 for Dutch)
  • Moderate inter-annotator agreement

• Identical sampling procedures and annotation guidelines, producing comparable data across languages
### Type of hate speech

<table>
<thead>
<tr>
<th>Type of hate speech</th>
<th>Target of hate speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>acceptable speech</td>
<td>migrants or LGBT</td>
</tr>
<tr>
<td>inappropriate speech (swearing, cursing, etc.)</td>
<td>individuals or groups related to migrants or LGBT (NGOs, public bodies, etc.)</td>
</tr>
<tr>
<td>offensive speech aimed at an individual</td>
<td>journalists and media</td>
</tr>
<tr>
<td>violence-inciting speech aimed at an individual</td>
<td>other commentators in the discussion thread</td>
</tr>
<tr>
<td>offensive speech aimed at someone because of their background</td>
<td>someone else</td>
</tr>
<tr>
<td>violence-inciting speech aimed at someone because of their background</td>
<td></td>
</tr>
</tbody>
</table>

### Number of posts and comments

<table>
<thead>
<tr>
<th>Language</th>
<th>Migrants</th>
<th>LGBT</th>
<th>Migrants</th>
<th>LGBT</th>
<th>Migrants</th>
<th>LGBT</th>
<th>Migrants</th>
<th>LGBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>16</td>
<td>14</td>
<td>30</td>
<td>93</td>
<td>57</td>
<td>22</td>
<td>194</td>
<td>64</td>
</tr>
<tr>
<td>Slovene</td>
<td>5.855</td>
<td>5.906</td>
<td>6.545</td>
<td>4.571</td>
<td>5.430</td>
<td>5.787</td>
<td>31.571</td>
<td>5.094</td>
</tr>
<tr>
<td>Croatian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
New lexical resource

• The LiLaH emotion lexicon of Croatian, Dutch and Slovene (Ljubešić et al., 2020)

• Based on manually corrected automatic translations of the EmoLex lexicon (NRC emotion lexicon, Mohammad and Turney, 2013)

• 14,182 emotion-conveying words and their associations with two sentiments and eight emotions:

• Uniform translation methodology across the three languages

• Download: [http://hdl.handle.net/11356/1318](http://hdl.handle.net/11356/1318)

<table>
<thead>
<tr>
<th>Entry</th>
<th>Sentiment</th>
<th>Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>positive</td>
<td>anger</td>
</tr>
<tr>
<td>good</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Corpus-based demographic analysis of hate speech producers

• Annotation task: **age** and **gender** of anonymized content creators

• Model the chance of a message being hateful based on the author’s socio-demographic profile

• Languages: Dutch, English, Slovene, Croatian

• General trends:
  • Men are more likely to produce hateful content than women
  • Older people produce more hateful content
  • Hateful content produced by men increases with age more than it does with women
Style and emotion in hate speech

• On these datasets, we investigate two linguistic phenomena:
  • **Writing style** of hateful social media content: function word usage
  • Emotion expression in hateful messages: **type of emotions** expressed and the frequency of emotion-conveying words

• Languages: English, Slovene, and Dutch

• Binary hate speech detection setting (hate speech vs. non-hate speech)

• In-domain (training-test partitions by post boundaries) vs. cross-domain hate speech detection settings
  • Cross-domain datasets:
    • English HASOC (Mandl et al., 2019): Twitter and Facebook
    • Dutch Ask.fm (Van Hee et al., 2015): cyberbullying posts
Style and emotion in hate speech

• Part-of-speech (POS) tags: morpho-syntactic patterns
• Function words (FWs): stylometric patterns
• Emotion-based: types of emotions expressed and the frequency of emotion-conveying words (NRC and LiLaH emotion lexicons)

• Mental illness on parade
  • POS:
    • ADJ NOUN PREP NOUN
  • POS & FWs:
    • ADJ NOUN on NOUN
  • POS & FWs & emotions:
    • ADJ illness on NOUN 2 fear sadness

• Term frequency (tf) weighting scheme
• Liblinear Support Vector Machines (SVM)
Results

• Stylometric and emotion-based features are robust indicators of hate speech
  • For all languages studied
  • In domain transfer contexts

• Comparable results to deep learning approaches (CNN, LSTM, BERT)

• Error analysis:
  • Stylometric and emotion-based approach better at implicit hate speech instances
  • Uncorrelated predictions

• Majority-voting ensemble of SVM, CNN, and multilingual BERT significantly outperforms best-performing individual models
English hate speech detection

- **SVM**: stylometric, emotions, hateful words from the POW lexicon (De Smedt et al., 2020)
- Monolingual **pre-trained language models**:
  - BERT (Devlin et al., 2019)
  - RoBERTa (Liu et al., 2019)
- **Ensemble**: majority voting (selecting the most often predicted label)
English: in-domain and cross-domain results

- Near SOTA performance (SVM)
- Cross-domain drop (around 10 F1 points)
- Ensemble performs best
- Results on OLID higher than the OffensEval-2019 winner
- Stylometric and emotion-based features are useful cues

<table>
<thead>
<tr>
<th>Model</th>
<th>FRENK, F1</th>
<th>OLID, F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>77.0</td>
<td>78.3</td>
</tr>
<tr>
<td>BERT</td>
<td>78.3</td>
<td>82.2</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>78.5</td>
<td>80.0</td>
</tr>
<tr>
<td>Ensemble</td>
<td>79.7*</td>
<td>83.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>OLID–FRENK, F1</th>
<th>FRENK–OLID, F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>67.7</td>
<td>69.0</td>
</tr>
<tr>
<td>BERT</td>
<td>69.4</td>
<td>72.1</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>69.2</td>
<td>72.4</td>
</tr>
<tr>
<td>Ensemble</td>
<td>69.7*</td>
<td>73.6*</td>
</tr>
</tbody>
</table>
Hateful metaphors

• Dutch LiLaH corpus
• Annotation task: identify hateful metaphors, determine semantic source domains
• Conceptual metaphor theory (Lakoff and Johnson, 1980): metaphorical expressions can be traced back to more abstract metaphor scheme
  • e.g., ... people are wild beasts (source domain: animals)
• Source domains found in the hateful metaphors (ordered by frequency in descending order):

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Animals</td>
<td>11. Carnival and circus</td>
</tr>
<tr>
<td>2. Dirt and personal hygiene</td>
<td>12. Home and kitchen linen</td>
</tr>
<tr>
<td>4. Disease and illness</td>
<td>14. Religious mythology</td>
</tr>
<tr>
<td>5. History</td>
<td>15. Sand</td>
</tr>
<tr>
<td>6. Fiction</td>
<td>16. Tourism</td>
</tr>
<tr>
<td>7. Food</td>
<td>17. Machines</td>
</tr>
<tr>
<td>8. Mental conditions</td>
<td>18. Physical conditions</td>
</tr>
<tr>
<td>10. Children</td>
<td>20. Other</td>
</tr>
</tbody>
</table>
Hateful metaphors

• Models:
  • Liblinear SVM, word unigrams & 2-grams, tf-idf weighting scheme
  • BERTje (de Vries et al., 2019)
  • RobBERT (Delobelle et al., 2020)

• Encoding methods:
  • … people are ANI wild beasts ANI + number of metaphorical expressions (SVM)
  • … people are <ani> wild beasts </ani> (BERTje and RobBERT)

• Findings:
  • Improvement: 5 F1 points (type), 3 F1 points (target)
  • Improvement for SVM higher than for BERTje/RobBERT (2 F1 points, type)
  • Improvement: implicit hate speech instances
The role of context in hate speech detection

- Dutch LiLaH corpus
- Previous work: preceding context / post information (Pavlopoulos et al., 2020)
- Annotation task:
  - Implicit context dependency (decision on the label depends on the context, not present in the data, world knowledge)
    - comment: The Leftist media lives up to its name
  - Explicit context dependency (decision on the label depends on the context, present in the data) → ID of the comment/post
    - comment: Go back home!  context: post about migrants

- BERTje (de Vries et al., 2019): message + [SEP] token + context
- Preliminary findings:
  - Better than preceding context / post information
  - Significant improvement for target identification
  - Improvement: implicit hate speech instances
Future work

• Implicit hate speech detection:
  • Automatically detect hateful metaphors and use to enhance implicit hate speech detection
  • Automatically detect relevant context and use to enhance implicit hate speech detection
Antwerp Text Mining Center (TEXTUA)

- Core facility of the University of Antwerp
  - Coordinator: Dr. Pieter Fivez
  - PIs: Walter Daelemans (NLP), Mike Kestemont (DH), David Martens (XAI), José Oramas (Image Processing)
- [https://www.uantwerpen.be/textua](https://www.uantwerpen.be/textua)

- Goal: support interdisciplinary text mining research in all research fields that have large text data (medicine, physics, political sciences, communication sciences, ...)
  - UAntwerp, Flanders, Europe, ... the world
Services

• Consultancy on data, software and methodology
• Data collection, conversion, preprocessing, text from image
• Data annotation and enrichment
• Software adaptation, development
• Education (targeted courses)

• Simple routine tasks: flat rate
• Joint strategic research projects (+ explanation): real cost + overhead

Get your free one-hour in-depth problem analysis!