Insulting and Negative Creative Comparisons

Master Thesis

Maja Geulig
Supervisor: Dr. Michael Wiegand
14.11.2019

Finalists’ Colloquium
Institute for Computational Linguistics
Heidelberg University
Table of contents

1. Introduction
2. Definition of Terms
3. Related Work
4. Thesis Outline
   Data Collection
5. Project Status and Future Work
Introduction
Is Social Media Creating an Insult-Culture? Twitter clamps down on abusive speech

"MEN ARE SCUM": INSIDE FACEBOOK'S WAR ON HATE SPEECH

France online hate speech law to force social media sites to act quickly

Germany Moves to Tighten Gun and Hate Speech Laws After Far-Right Attacks

Chrome extension from Google wants to filter out toxic comments

"Tune" lets you automatically hide Facebook, YouTube and Disqus.

Why America needs a hate speech law

For Facebook Content Moderator: Traumatizing Material Is A Job Hazard

UN takes aim at Trump over 'dehumanising' towards immigrants

YouTube disables comments on hate speech hearing livestream
enormous amount of text content generated on the internet each day
insulting language is a pressing issue for social media etc.
basic word filters cannot cover many types of insulting language
manual filtering is time-consuming and psychologically demanding for moderators

→ increasing demand for systems for insulting language detection
Role of Task-specific Datasets

- Datasets are the foundation of research:
  - Allow development of new classification approaches
  - Comparisons between different systems
  - Evaluation of performance
  - Analysis & directions of future research

- The field is (relatively) young and not all types of insulting language have been documented in datasets

- Creating datasets for insulting language is very time-consuming

- Skewed distribution in random samples of data (3-4% on Twitter)\(^1\)

→ Creating a new dataset for a previously undocumented specific type of insulting language opens new directions of research & insights into the performance of existing systems

\(^1\)Founta et al. (2018)
The thesis aims to create a **new dataset** of implicitly insulting language in the form of creative comparisons.

This dataset will be used to **analyse** this specific subtype of insulting language.

Additionally, the performance of **state-of-the-art classifiers** for insulting language will be tested on the dataset.
Definition of Terms
other terms used: hate speech, offensive/abusive/toxic language, profanity, ...

**Insult**
= “disparages a person or a group on the basis of some characteristic.”

**Explicit Insult**
= individual words themselves have an unambiguously offensive nature

**Implicit Insult**
= does not contain offensive words but still perceived as insult → irony, negative stereotypes, jokes, figurative language, comparisons

---

\(^2\)Schmidt & Wiegand (2017)
\(^3\)Nockleby (2000)
Explicit Insult

= You are a piece of scum.

He reminds me of a spoiled brat without a properly functioning brain.

Go away, you pervert sleazebag.

Implicit Insult

= I haven’t had an intelligent conversation with a woman in my whole life. (Negative Stereotype)

Why aren’t there any Mexicans on Star Trek? Because they do not work in the future either. (Joke + Stereotype)

You are as useful as an umbrella full of holes. (Comparison)
Comparison
= act of evaluating two or more things by determining the relevant characteristics of each thing to be compared + which characteristics of each are similar/different to the other, and to what degree

Simile: subset of comparisons which compare two very different things

Insulting Comparison
= expression which is disrespectful or scornful; may be accurate, but at the same time abusive

Negative Comparison
= either contains words with negative meaning, or the wording expresses negative meaning; for the purposes of this thesis also not insulting.

---

Comparison

= This person is as tall as a tree.

Insulting Comparison

= You are like an inbred.

Negative Comparison

= You are as pale as a ghost.
Subject of the Thesis

**Dataset**: 2 Classes/Labels (disjoint sets)

1. Implicitly Insulting Comparisons
2. Non-insulting Negative Comparisons

**Not in the Dataset**: other types of insults or comparisons

- general insults
- explicitly insulting comparisons
- neutral comparisons
- positive comparisons
**Dataset**: 2 Classes/Labels (disjoint sets)

1. Implicitly Insulting Comparisons: *You eat like a pig.*
2. Non-insulting Negative Comparisons: *You are as pale as a ghost.*

**Not in the Dataset**: other types of insults or comparisons

- general insults: *Shut up, you asshole.*
- explicitly insulting comparisons: *You look like a faggot.*
- neutral comparisons: *Your car is as green as an olive.*
- positive comparisons: *You are as radiant as the sun.*
Related Work
Related Areas of Research

- Hate Speech Detection
- Linguistic Perspectives on Insulting Language
- (Ironic) Simile and Sarcasm Detection
- **Polarity Detection in Similes**
- Dataset Creation for Skewed Distributions
- Crowdsourcing Task Design
Learning to Recognize Affective Polarity in Similes (Qadir et al. 2015)

- build a classifier for recognising polarity in similes on Twitter
- create a dataset of 1,500 positive, neutral and negative similes
- labels annotated through crowdsourcing (Amazon Mechanical Turk)
- dataset contains 524 negative similes
- negative label also includes instances of insulting language

<table>
<thead>
<tr>
<th>Label</th>
<th>Dataset Instance</th>
<th>Example Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative</td>
<td>PERSON</td>
<td></td>
</tr>
<tr>
<td>negative</td>
<td>PERSON</td>
<td></td>
</tr>
<tr>
<td>negative</td>
<td>PERSON</td>
<td></td>
</tr>
<tr>
<td>neutral</td>
<td>IT</td>
<td></td>
</tr>
<tr>
<td>positive</td>
<td>PERSON</td>
<td></td>
</tr>
</tbody>
</table>
### Label | Dataset Instance | Example Sentence
---|---|---
negative | PERSON || look || crackhead | You look like a crackhead.
negative | PERSON || sound || die whale | You sound like a dying whale.
negative | PERSON || be || lose puppy | You look like a lost puppy.

**Additional annotation in the context of this thesis:**

- subset of 359 similes that relate to a person, a person’s belongings or attributes
- manual annotation of existing data
- insulting: 274 (76.32%), of which 89 (24.79%) are explicitly insulting
- non-insulting negative: 86 (23.96%)

→ creating the new dataset represents existing phenomena
Thesis Outline
1. Development of Data Collection Methods
Decided on crowdsourcing for data collection, the design of the surveys tasks (dev surveys) and the data to collect.

2. Data Collection
3 Step Annotation Procedure

1. annotators creatively invent comparisons (insulting & negative in separate tasks) using patterns

2. invented instances are re-labelled by different annotators

3. a subset of similar / problematic instances are labelled again in a consistency task

3. Data Analysis

4. Experiments & Evaluation

5. Writing
1. Data Collection for Creative Comparisons

- using crowdsourcing avoids problems of skewed distribution & biases in existing data
- annotators are selected through Prolific Academic
- UK residents who are native speakers of English
- no linguistic background: clear & concise explanation of relevant concepts needed
- annotators are asked to invent examples of natural language

---

5https://www.prolific.co/
1. Crowdsourcing Creative Comparisons: Challenges

- task design needs to be compact & guidelines concise
- development phase showed improvements when separating insulting and negative comparisons into different tasks
- free generation is too demanding, but annotators work well when provided with patterns
  - range of patterns iteratively developed throughout the development phase, based on annotator responses
  - initially multi-slot patterns: *Your [X] is as [Y] as [Z]*.
  - abandoned in favour of single-slot patterns, e.g.
    - *Your voice is like [X]*.
    - *You talk like [X]*.
    - *You are as polite as [X]*.
1. Crowdsourcing Creative Comparisons: Challenges

- for generating non-insulting negative comparisons, providing a ’situational frame’ is helpful
- for generating insulting comparisons, avoiding explicit swearwords is often a challenge
- noise phenomena: answers with no comparison structure, nonsensical comparisons, context-dependent comparisons
- some comparisons are actually fixed expressions: when prompted, a high number of annotators give similar answers
1. Crowdsourcing Creative Comparisons: Design Example

Figure 1: Example for a task to generate negative comparisons

Situation: You have just overheard an argument between two of your friends, and now you are talking to one of them. You think that she behaved too aggressively during the argument. You want to comment on her words and behaviour, without offending her.

Please write full sentences that follow the format provided. Replace the forms given in brackets [ ] with your own choice of words. Please keep your sentences short.

13. Your attitude is like [X].
2. Re-labelling Comparisons

- goal: each comparison in dataset should have the label INS (insulting) or NEG (non-insulting negative)

- instances are generated in separate survey tasks, so they already have an assigned label

- however, instances are generated by many different annotators with inconsistent views on what constitutes INS/NEG boundary

- label assignment is difficult and somewhat inconsistent
  → all generated instances are re-labelled in manual classification task performed by 5 different annotators
  → label assigned through majority vote of 3
3. Checking Label Consistency for Similar Instances

- distinguishing between INS / NEG labels is a difficult task for human annotators
- dataset was developed & labels assigned iteratively: decisions based on single instance without any context
- semantically similar instances should receive the same label (label consistency)
  → instances grouped by similarity
  → similarity groups annotated again in consistency task design
  → labels reassigned for selected instances
3. Checking Label Consistency for Similar Instances: Example

<table>
<thead>
<tr>
<th>Group 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Your lips are like a cracked glacier.</td>
</tr>
<tr>
<td>B. Your lips are like cracked ice.</td>
</tr>
<tr>
<td>C. Your lips are like a cracked Iceberg.</td>
</tr>
</tbody>
</table>

**How would you label the whole group?**

- Negative
- Insulting

**Which comparison in the group is inconsistent with the others?**

- A
- B
- C
- NONE

**Figure 2:** Example of a Similarity Group in the Consistency Task
Removing Instances

Instances were removed from the dataset for the following reasons:

1. **Explicit Insult**: instances contains offensive words
2. **No Comparison Structure**: colloquial usages of *like* as hedge/quotative
3. **(Near-)Duplicate**: (near-)duplicates of existing instances
4. **Context-Dependent**: comparison requires knowledge about the speaker or specific world knowledge
5. **Other Label**: no majority agreement on label or agreement that comparison is positive/neutral
Removing Instances: Examples

Instances were removed from the dataset for the following reasons:

1. **Explicit Insult:** You seem like a demented idiot.
2. **No Comparison Structure:** Your progress is like glacial.
   Your clothes are like less beautiful.
3. **(Near-)Duplicate:** You are as thin as a rake.
4. **Context-Dependent:** Your reaction reminds me of how I felt.
   Your progress is like Brexit.
5. **Other Label:** Your smile is like spring sunshine.
Pattern Distributions

- comparisons are generated using pattern prompts
- to ensure distribution of instances across patterns, each pattern is limited to 20 instances total (max. 10 for each label)
- during the data collection process some patterns show a strong bias towards only one label
  → biased patterns are removed from the dataset
- fixed range of unbiased patterns for controlled dataset design: prevent classifiers from only learning patterns instead of INS/NEG labels
Examples of Removed Patterns

Biased towards negative comparisons:

• You are as sad as
  ... a wilted lettuce. / ... a weeping willow / ... a rain cloud.
• You are as organised as
• You are as pale as

Biased towards insulting comparisons:

• Your make-up reminds me of
  ... a clown. / ... an old lady. / ... crayons.
• You are as useful as
• You are as intelligent as
Project Status and Future Work
## Dataset Version 11

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances</td>
<td>1004</td>
</tr>
<tr>
<td>Patterns</td>
<td>71</td>
</tr>
<tr>
<td>Annotators</td>
<td>98</td>
</tr>
<tr>
<td>Surveys Used</td>
<td>18 (+ 8 Annotation Tasks)</td>
</tr>
<tr>
<td>Average Length of Comparison</td>
<td>22.53</td>
</tr>
<tr>
<td>Total Tokens</td>
<td>9372</td>
</tr>
<tr>
<td>Total Tokens (without Pattern)</td>
<td>5302</td>
</tr>
</tbody>
</table>

### Most frequent tokens

<table>
<thead>
<tr>
<th>Token</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>'old'</td>
<td>26</td>
</tr>
<tr>
<td>'person'</td>
<td>18</td>
</tr>
<tr>
<td>'child'</td>
<td>14</td>
</tr>
<tr>
<td>'day'</td>
<td>13</td>
</tr>
<tr>
<td>'man'</td>
<td>12</td>
</tr>
<tr>
<td>'dog'</td>
<td>12</td>
</tr>
<tr>
<td>'car'</td>
<td>11</td>
</tr>
<tr>
<td>'night'</td>
<td>10</td>
</tr>
<tr>
<td>'cat'</td>
<td>9</td>
</tr>
<tr>
<td>'seen'</td>
<td>9</td>
</tr>
<tr>
<td>'time'</td>
<td>9</td>
</tr>
<tr>
<td>'monkey'</td>
<td>8</td>
</tr>
<tr>
<td>'white'</td>
<td>8</td>
</tr>
<tr>
<td>'paper'</td>
<td>8</td>
</tr>
<tr>
<td>'shop'</td>
<td>8</td>
</tr>
<tr>
<td>'need'</td>
<td>8</td>
</tr>
<tr>
<td>'clown'</td>
<td>8</td>
</tr>
<tr>
<td>'pig'</td>
<td>8</td>
</tr>
<tr>
<td>'got'</td>
<td>7</td>
</tr>
<tr>
<td>'ghost'</td>
<td>7</td>
</tr>
</tbody>
</table>
## Dataset Version 11: Examples

<table>
<thead>
<tr>
<th>ID</th>
<th>Survey</th>
<th>Label</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>0093</td>
<td>NEG</td>
<td>INS/5</td>
<td>You talk like a monkey with a mouth full of nuts.</td>
</tr>
<tr>
<td>0574</td>
<td>NEG</td>
<td>NEG/5</td>
<td>You talk like an express train.</td>
</tr>
<tr>
<td>0162</td>
<td>INS</td>
<td>INS/5</td>
<td>Your face is like a squashed tomato.</td>
</tr>
<tr>
<td>1593</td>
<td>NEG</td>
<td>NEG/4</td>
<td>Your face is like a white sheet of paper.</td>
</tr>
<tr>
<td>0015</td>
<td>INS</td>
<td>NEG/5</td>
<td>You reacted like a child who lost a balloon.</td>
</tr>
<tr>
<td>1614</td>
<td>INS</td>
<td>INS/5</td>
<td>You reacted like a virgin.</td>
</tr>
<tr>
<td>0357</td>
<td>INS</td>
<td>INS/5</td>
<td>Your voice is like nails on a chalkboard.</td>
</tr>
<tr>
<td>1994</td>
<td>NEG</td>
<td>NEG/5</td>
<td>Your voice is like a whisper.</td>
</tr>
</tbody>
</table>

**Additional Information:** Annotator ID, Survey ID, Unedited input text
# Thesis Outline: Future Work

<table>
<thead>
<tr>
<th>1. Development of Data Collection Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Data Collection</td>
</tr>
<tr>
<td><strong>2. Data Analysis – 2 weeks</strong></td>
</tr>
<tr>
<td>Analyse created dataset to identify biases and distribution of patterns, labels, specific semantic fields.</td>
</tr>
<tr>
<td><strong>3. Experiments &amp; Evaluation – 5 weeks</strong></td>
</tr>
<tr>
<td>Choose classifier(s) that represent state-of-the-art performance for insulting language.</td>
</tr>
<tr>
<td>Implement &amp; evaluate performance on dataset.</td>
</tr>
<tr>
<td><strong>4. Writing – 5 weeks + 2 weeks corrections/buffer</strong></td>
</tr>
<tr>
<td>Write the thesis.</td>
</tr>
<tr>
<td>Planned deadline: 17.02.2020.</td>
</tr>
</tbody>
</table>
• the aim of the thesis is to create and analyse a dataset of implicitly insulting comparisons and non-insulting negative comparisons
• data has been collected through crowdsourcing in a 3-step annotation procedure
• the new dataset contains 1004 creative comparisons collected from almost 100 annotators
• the remaining steps are a thorough analysis of the data, experiments for the performance of classifiers and the writing process
Thank you for your attention.

Questions and feedback are very welcome.


This slide only lists works cited directly within the presentation, and does not represent an overview of related work.


Ng, Alfred (12-03-2019). Chrome extension from Google wants to filter out toxic comments. In: CNET. https://tinyurl.com/cnetxbox


Warren, Tom (14-10-2019). Microsoft unveils XBox content filters to stop the swears and toxicity. In: The Verge.
