Insulting and Negative Creative Comparisons

Master Thesis

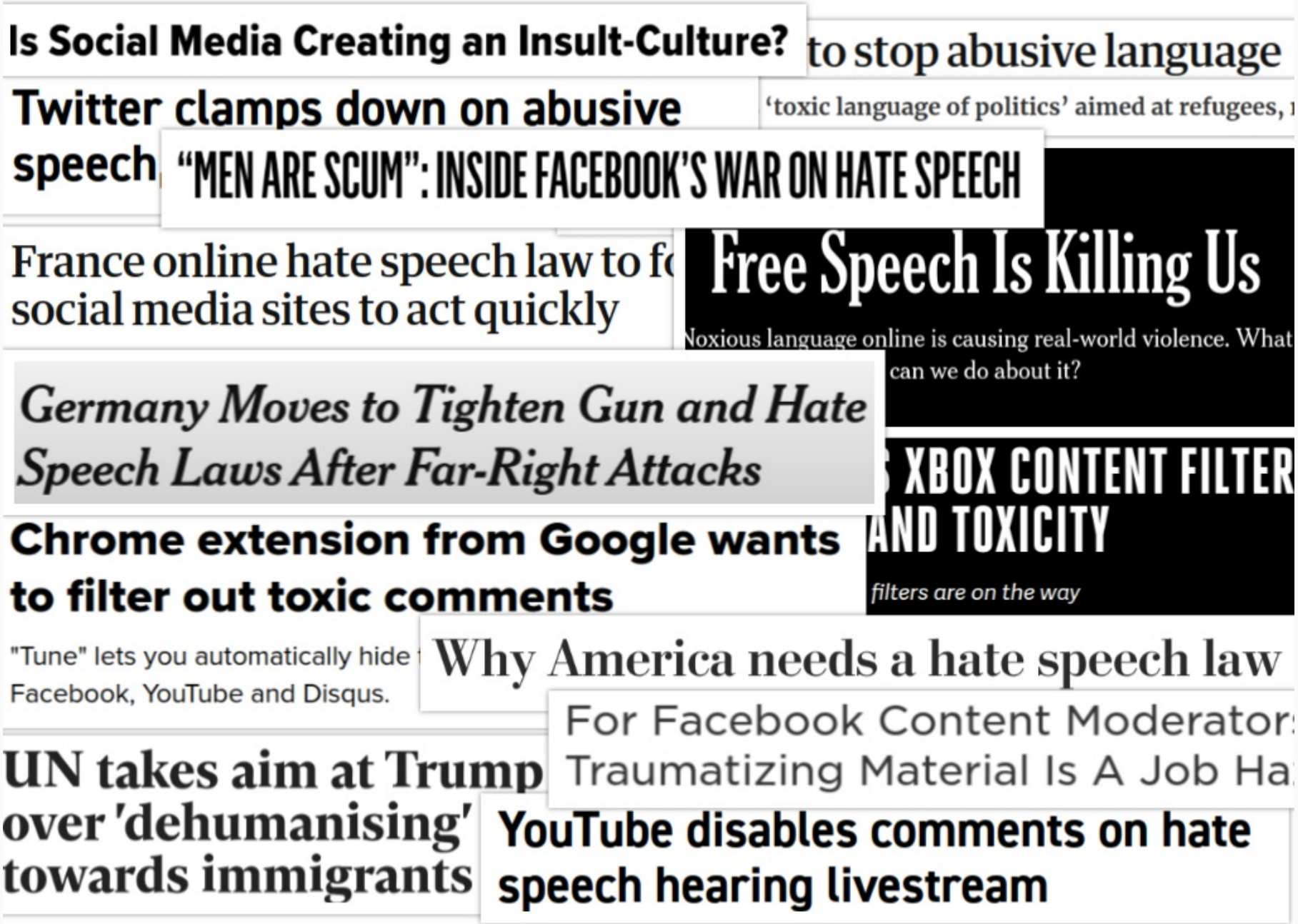
Maja Geulig
Supervisor: Dr. Michael Wiegand

14.11.2019

Finalists' Colloquium
Institute for Computational Linguistics
Heidelberg University

Table of contents

Introduction


Is Social Media Creating an Insult-Culture?
to stop abusive language
Twitter clamps down on abusive speech
'toxic language of politics' aimed at refugees,
"MEN ARE SCUM": INSIDE FACEBOOK'S WAR ON HATE SPEECH
France online hate speech law to f
social media sites to act quickly
Free Speech Is Killing Us
Noxious language online is causing real-world violence. What
can we do about it?
Germany Moves to Tighten Gun and Hate Speech Laws After Far-Right Attacks
XBOX CONTENT FILTER
AND TOXICITY
filters are on the way
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 Ha
UN takes aim at Trump over 'dehumanising' towards immigrants
YouTube disables comments on hate speech hearing livestream

Researching Insulting Language

- 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

$\rightarrow$ 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]

$\rightarrow$ 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)

Thesis Idea

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

Insulting Language

other terms used: hate speech, offensive/abusive/toxic language, profanity, ... [2]

Insult
= "disparages a person or a group on the basis of some characteristic." [3]

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)

Insulting Language: Examples

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)

Types of Comparisons

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 [4]

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**.

[4] Niculae & Yeneva (2013), Qadir et al. (2015)

Types of Comparisons: Examples

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

Subject of the Thesis: Examples

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

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.*
neutral	IT || smell || pizza	*It smells like pizza.*
positive	PERSON || look || princess	*You look like a princess.*

Learning to Recognize Affective Polarity in Similes (Qadir et al. 2015)

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

Thesis Outline: Current Status

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 [5]
- 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

[5] https://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

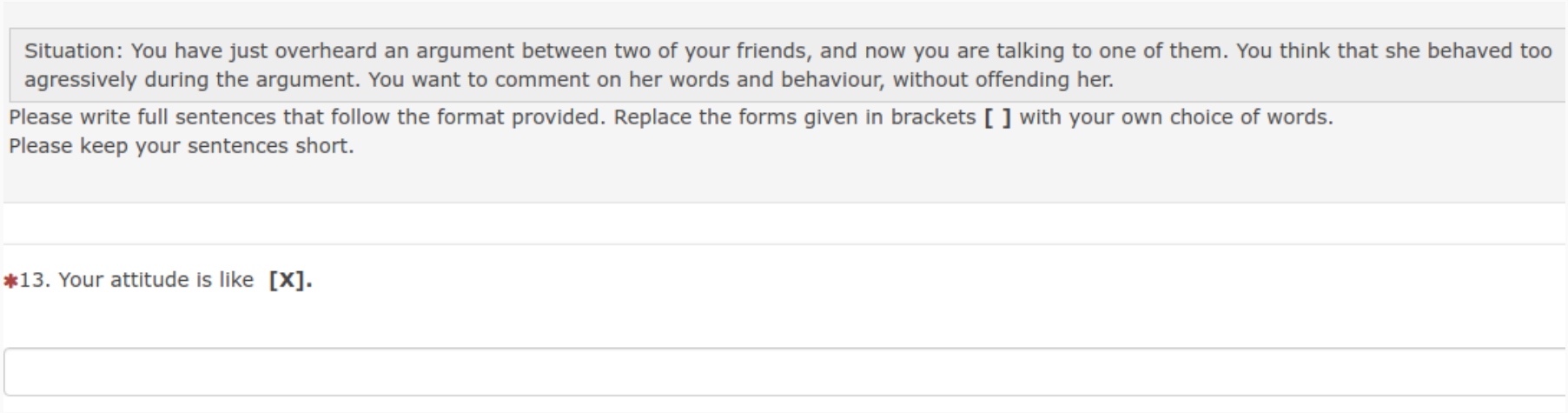


Figure 1: Example for a task to generate negative comparisons

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
  $\rightarrow$ all generated instances are re-labelled in manual classification task performed by 5 different annotators
  $\rightarrow$ 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)
  $\rightarrow$ instances grouped by similarity
  $\rightarrow$ similarity groups annotated again in consistency task design
  $\rightarrow$ labels reassigned for selected instances

3. Checking Label Consistency for Similar Instances: Example

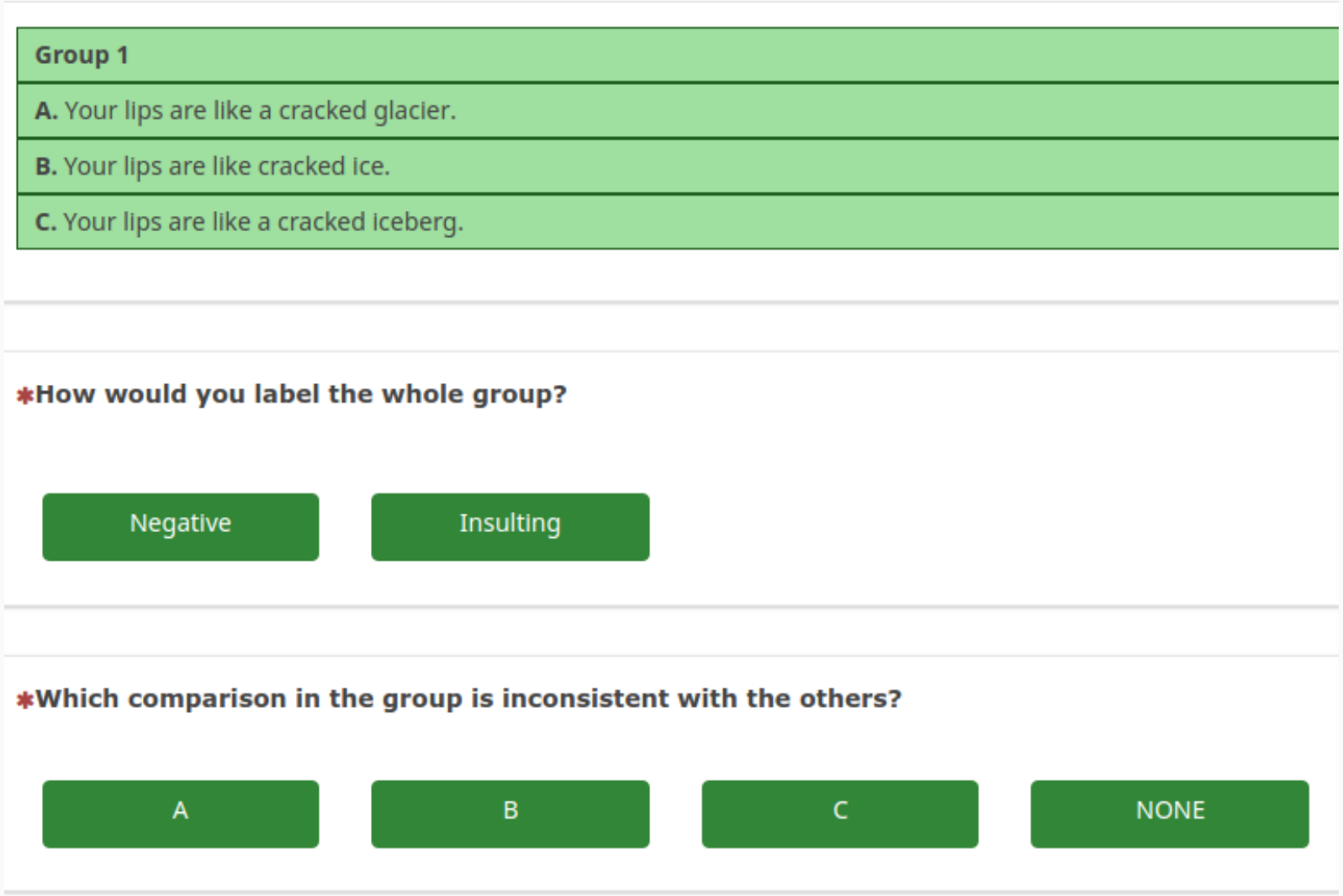


Figure 2: Example of a Similarity Group in the Consistency Task

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
  $\rightarrow$ 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

Feature	
Instances	1004
Patterns	71
Annotators	98
Surveys Used	18 (+ 8 Annotation Tasks)
Average Length of Comparison	22.53
Total Tokens	9372
Total Tokens (without Pattern)	5302

Most frequent tokens

'old': 26, 'person': 18, 'child': 14, 'day': 13, 'man': 12, 'dog': 12, 'car': 11, 'night': 10, 'cat': 9, 'seen': 9, 'time': 9, 'monkey': 8, 'white': 8, 'paper': 8, 'shop': 8, 'need': 8, 'clown': 8, 'pig': 8, 'got': 7, 'ghost': 7

Dataset Version 11: Examples

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

Additional Information: Annotator ID, Survey ID, Unedited input text

Thesis Outline: Future Work

1. Development of Data Collection Methods

2. Data Collection

2. Data Analysis – 2 weeks

Analyse created dataset to identify biases and distribution of patterns, labels, specific semantic fields.

3. Experiments & Evaluation – 5 weeks

Choose classifier(s) that represent state-of-the-art performance for insulting language.

Implement & evaluate performance on dataset.

4. Writing – 5 weeks + 2 weeks corrections/buffer

Write the thesis.

Planned deadline: 17.02.2020.

Summary

- 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.

References

Founta, A. M. et al. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. 12th International AAAI Conference on Web and Social Media, ICWSM 2018.

Niculae, V., & Yaneva, V. (2013). Computational considerations of comparisons and similes. 51st Annual Meeting of the ACL, (February), 8995.

John T. Nockleby. (2000). Hate Speech. In Leonard W. Levy, Kenneth L. Karst, and Dennis J. Mahoney, editors, Encyclopedia of the American Constitution, pages 12771279. Macmillan, 2nd edition

Qadir, A., Riloff, E., & Walker, M. (2015). Learning to Recognize Affective Polarity in Similes.

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing.

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

References: Newspaper Headlines (p. 2) i

All urls accessed 2019-11-11.

Agence France-Presse (09-07-2019). France online hate speech law to force social media sites to act quickly. In: The Guardian. https://tinyurl.com/guardianfrance

Blanchard, Paul (26-02-2016). Is Social Media Creating an Insult-Culture? In: The Huffington Post. https://tinyurl.com/blancard19

Gross, Terry (01-07-2019). For Facebook Content Moderators, Traumatizing Material Is A Job Hazard. In: NPR. https://tinyurl.com/nrpgross

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

Marantz, Andrew (04-10-2019). Free Speech is Killing Us. In: The New York Times. https://tinyurl.com/marantz19

References: Newspaper Headlines (p. 2) ii

Lima, Christiano (09-04-2019). YouTube disables comments on hate speech hearing livestream. In: Politico. https://www.politico.com/story/2019/04/09/youtube-hate-speech-congress-livestream-1264087

Scola, Nancy (27-06-2019). Twitter clamps down on abusive speech, in seeming shot at Trump. In: Politico. https://www.politico.com/story/2019/06/27/twitter-abusive-speech-trump-1559787

Stengel, Richard (29-10-2019). Why America needs a hate speech law. In: The Washington Post. https://www.washingtonpost.com/opinions/2019/10/29/why-america-needs-hate-speech-law/

Van Zuylen-Wood, Simon (26-02-2019). 'Men Are Scum': Inside Facebook's War on Hate Speech. In: Vanity Fair. https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech

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

References: Newspaper Headlines (p. 2) iii

https://www.theverge.com/2019/10/14/20913323/microsoft-xbox-live-content-filters-messages-party-chat-moderation

Waterson, Jim (04-11-2019). MPs pledge to stop abusive language during general election. In: The Guardian. https://www.theguardian.com/politics/2019/nov/03/mps-pledge-to-stop-abusive-language-during-general-election

- (09-04-2019). UNs Grandi slams toxic language of politics aimed at refugees, migrants. In: UN News. https://news.un.org/en/story/2019/04/1036391

- (30-10-2019). Germany Moves to Tighten Gun and Hate Speech Laws After Far-Right Attacks. In: The New York Times. https://www.nytimes.com/2019/10/30/world/europe/germany-gun-hate-speech-laws.html