Towards Tackling Hate Online Automatically

Nikola Ljubešić[1], Darja Fišer[2,1], Tomaž Erjavec[1]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2]Department of Translation, University of Ljubljana

SS22 Colloquium on Intolerant and Abusive Content Online
Auckland, New Zealand
30 June 2018

Overview

1. The FRENK project
2. Data harvesting (Facebook)
3. Filtering the data by topic (migrants, LGBT)
4. Manual data annotation (PyBossa)
5. Automating the identification process

FRENK

The FRENK Project

- Slovene basic research project "Resources, methods, and tools for the understanding, identification, and classification of various forms of socially unacceptable discourse in the information society" (2017 — 2019)
- Primary project goal: Interdisciplinary treatment of linguistic, sociological, legal and technological dimensions of different forms of socially unacceptable discourse (SUD)
- Partners
  - Dept. of Knowledge Technologies, Jožef Stefan institute (lead)
  - Faculty of Arts (linguistics)
  - Faculty of Social Sciences (social sciences)
  - The Peace Institute (law)

State of the art in automated hate speech detection

- Usage of supervised machine learning: computer is given (as) many (as possible) examples of hate speech and non-hate speech, a classifier is trained on these examples
- To obtain these examples, annotation campaigns have to be run
  1. Classification schema / typology
  2. Annotation guidelines
  3. Annotator training
- In most (all?) cases ad-hoc treatment of these three components
  1. Not well-defined / well-argued typology
  2. No or very basic annotation guidelines
  3. Untrained students (or paper authors?) at disposal used for data annotation
- FRENK tries to address all the above issues

Harvesting

Harvesting the data from Facebook

- Facebook has the Graph API - we can communicate with Facebook (data) via computer programs
- Collecting all posts and comments on Facebook pages of three popular daily newspapers (alexa.com)

	# of posts	# of comments
24urcom	8,375	126,983
RTV.SLOVENIJA	12,192	12,998
SiOL.net.Novice	20,257	57,406
Nova24TV	9,848	83,728

Filtering

Filtering the data for topics of interest

- Two topics (targets) of interest:
  - Migrants / Islamophobia
  - LGBT / Homophobia
- Want to (semi-)automate the filtering process
- Application of supervised machine learning
  - Identify examples of each topic via keyword search (100 posts per topic)
  - Use these exemplary documents to train classifiers for each topic – for each post the classifier predicts whether the post is on the topic of migrants, LGBT, or other
- Results of automatic classification are not perfect, but good enough for pre-filtering the data

	Precision	Recall
Migrants	0.80	0.66
LGBT	0.86	0.53
Other	0.75	0.97

Amount of data after filtering

	# of posts	# of comments
24urcom	8,375	126,983
Migrants	178	16,849
LGBT	17	2,252
SiOL.net.Novice	20,257	57,406
Migrants	98	3,205
LGBT	12	456
Nova24TV	9,848	83,728
Migrants	684	23,174
LGBT	65	2,037

Annotation

Annotation schema and guidelines: SUD type

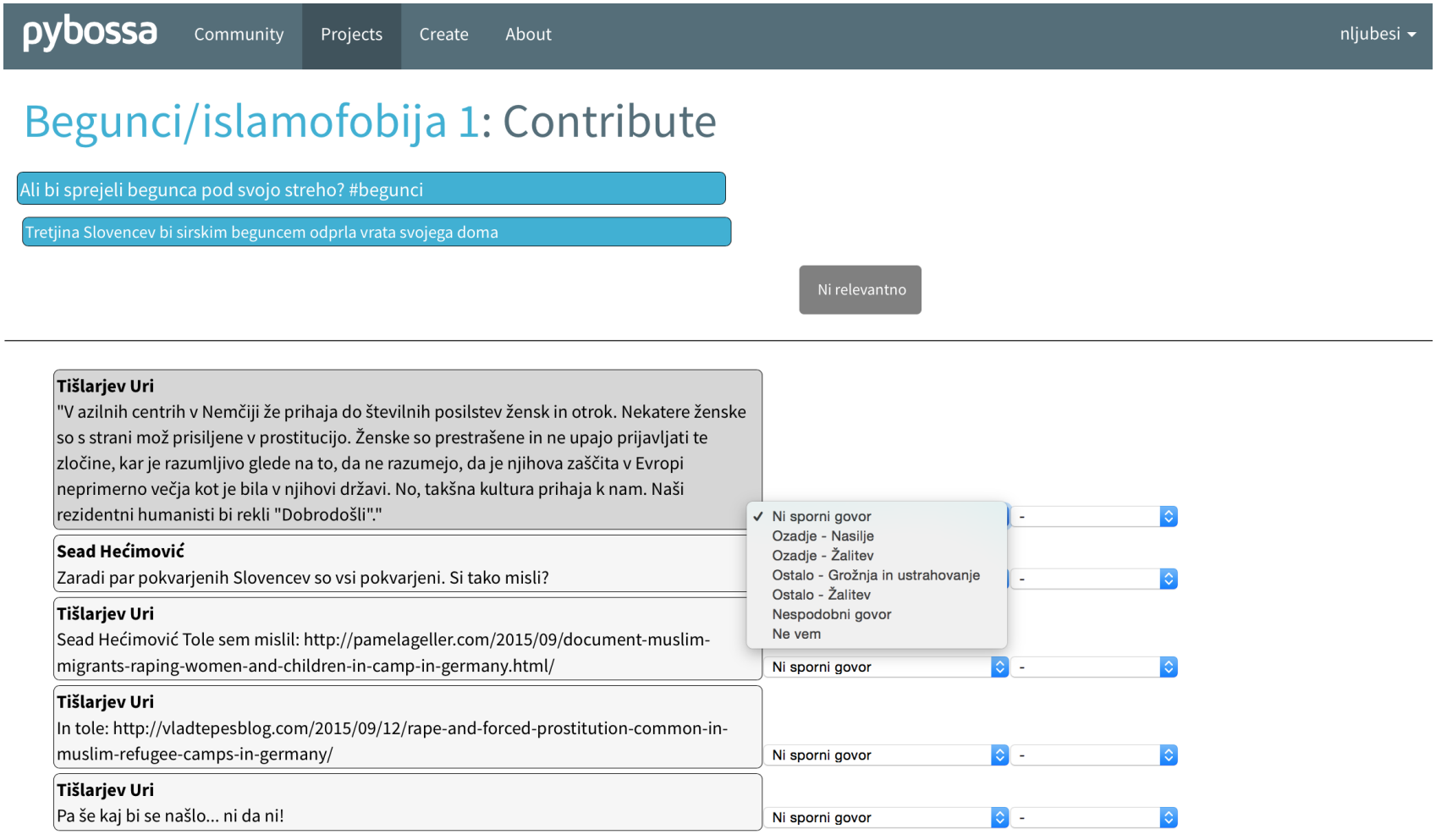
Decision tree for SUD type

```
Background based SUD?
    YES: are there elements of violence?
         YES: background, violence
         NO:  background, hate
    NO:  SUD towards individuals and groups?
         YES: elements of violence?
              YES: other, threat
              NO:  other, hate
         NO:  is the speech unacceptable?
              YES: unacceptable speech
              NO:  acceptable speech
```

Annotation schema and guidelines: SUD target

```
Migrants / LGBT
Related to migrants / LGBT
Journalists or media
Another commenter
Other
```

Annotation in PyBossa - a tool for crowdsourcing



Initial annotation campaign

- Annotators: bachelor and master students from the Faculties of Arts and Social Sciences, University of Ljubljana
- 33 annotators, 16/17 per topic
- Each annotator annotates the same data, 16/17 annotations per instance
- Training session, 5 hours
- Annotation guidelines on 8 pages
- Communication via mailing list

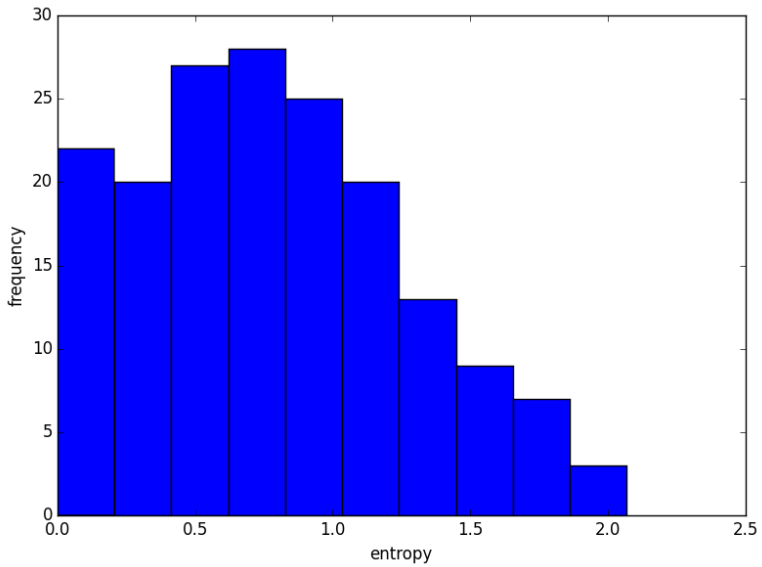
Distribution of responses

Migrants	acceptable	47.57 %
	background, hate, migrants	23.51 %
	other, hate, commenter	6.19 %
	background, violence, migrants	4.69 %
	other, hate, journalist	4.2 %
	other, hate, other	2.56 %
	other, hate, related	1.96 %
	background, hate, related	1.83 %
LGBT	acceptable	63.77 %
	background, hate, lgbt	17.57 %
	other, hate, commenter	5.44 %
	other, hate, other	4.22 %
	background, hate, related	2.43 %
	other, hate, related	1.47 %
	unacceptable, no target	0.88 %
	do not know	0.76 %

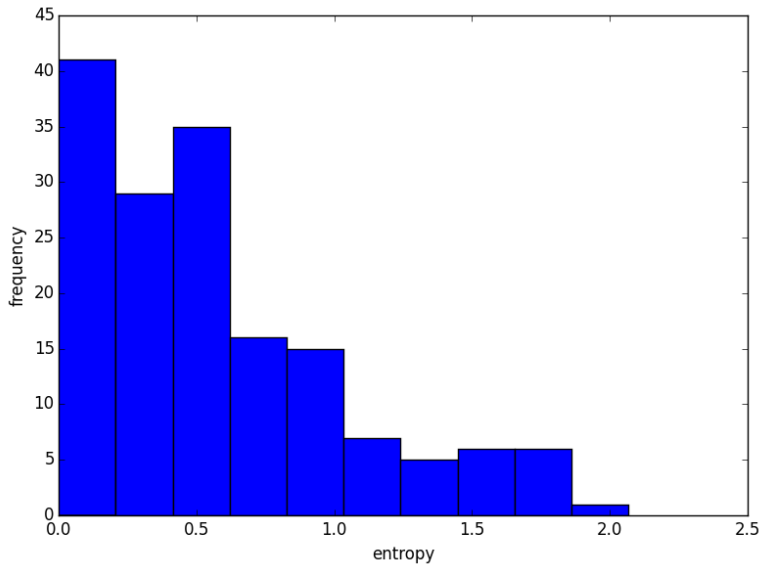
Entropy of response distributions

Entropy: measure of uncertainty.
Lower is better.
If every annotator gave the same response, entropy is 0.

Migrants



LGBT



Easy examples

acceptable

If I myself had enough for a decent life, I'd take in or at least help one of our families

background, violence, migrants

The media show only how they are in need and such... I wonder how many of those that would open their door to them now would help them if they physically or psychologically harassed them ... or their relatives ... they are not so terribly in need as the media show! They are like the Trojan horse! Seal the borders with a wall and shoot those that come near!

Hard examples

unacceptable, other 5; acceptable 3; background, hate, migrants 2; other, hate, commenter 2; ...

DON'T EAT SHIT

other, hate, related 5; background, hate, related 3; other, hate, journalist 2; unacceptable, other 2; ...

We have proof that monkeys are not only in parliament..

Automation

Two current approaches in machine learning

Traditional methods

- Linear regression, Logistic regression, Decision trees, Support vector machines...
- Text representation through manually defined variables, mostly specific words or sequences of words (n-grams)

Deep learning methods

- "AI hype", drastic improvements in image and audio processing, varying in text processing, data hungry!
- Text representation through distributed word representations fed into a neural network (matrix multiplications)
- Each word is represented through a sequence of numbers, representations of "cat" and "dog" are much more similar than of "cat" and "car"

Two use cases

GermEval 2018 shared task

- This years' shared task at the German NLP conference, 20+ teams on board (a lot!)
- 5,000 training examples
- Traditional methods: ~75% accuracy
- Deep learning methods: ~75% accuracy

Dataset of deleted comments from a website

- Croatian, 24sata.hr, obtained from the pubilsher
- 500,000 training examples
- Traditional methods: ~85% accuracy
- Deep learning methods: ~95% accuracy

Conclusion

- FRENK – interdisciplinary project, trying to improve the problem definition and data annotation deficiencies of current projects
- Data harvesting: easy
- Data selection: medium, but crucial, question of sample representativeness
- Data annotation: hard, very costly, both in terms of annotator training and the annotation itself (if done properly)
- (Semi-)Automation: possible, but very challenging
  - Accuracy depends on the amount of training data
  - Good results can be expected on a small number of classes
  - Training data very situational, topic- and target-dependent

Towards Tackling Hate Online Automatically

Nikola Ljubešić[1], Darja Fišer[2,1], Tomaž Erjavec[1]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2]Department of Translation, University of Ljubljana

SS22 Colloquium on Intolerant and Abusive Content Online
Auckland, New Zealand
30 June 2018