

Racism detection in Dutch social media posts

An exploratory study

Stéphan Tulkens, Lisa Hilde, Elise Lodewyckx
Ben Verhoeven, Walter Daelemans

ATILA '15



Contents

1. How to define racism?
2. Data Collection
3. Annotation
4. Classification
5. Results

How to define racism?

1. Belgian law

How to define racism?

1. Belgian law

- Discrimination
- Inciting hate

How to define racism?

1. Belgian law
 - Discrimination
 - Inciting hate
2. Our definition

How to define racism?

1. Belgian law

- Discrimination
- Inciting hate

2. Our definition

- Insults and generalizations
 - Skin color, ethnicity, nationality
 - **Religion and culture**

Data collection

1. Interfederal Centre for Equal Opportunities
2. Two public social media pages
 - Training set: 5759 comments
 - Test set: 616 comments



Annotation guidelines

Four labels:

1. Racist

- “Weg met alle niet Westerse buitenlanders”
“Away with all non-Western foreigners”

2. Context

- “Ik ben het volledig met je eens”
“I totally agree with you”

3. Non-racist

4. Invalid

Annotations

Three annotators: A, B & C

1. Training data

- A and B annotated all posts
 - Agreement: 0.79, $\kappa = 0.60$
- C: tiebreaker

2. Test data

- A, B & C annotated the posts
 - Agreement: 0.77, $\kappa = 0.54$ (125 posts)
 - C has low overlap with both A and B

Gold standard

| | Train data | Test data |
|------------|------------|-----------|
| Non-racist | 4438 | 436 |
| Racist | 924 | 164 |
| Invalid | 335 | 9 |
| Context | 62 | 7 |

For automatic classification: only **two** labels are kept

Gold standard

| | Train data | Test data |
|------------|------------------|----------------|
| Non-racist | 4438 + 62 | 436 + 7 |
| Racist | 924 | 164 |
| Invalid | 335 | 9 |
| Context | 62 | 7 |

For automatic classification: only **two** labels are kept

Classifier

- Support Vector Machine algorithm
- Features:
 - Content-based
 - Stylistic

Content-based features

- Word (unigram and bigram) frequencies
- Dictionaries
 - LIWC dictionaries
Linguistic Inquiry and Word Count, Pennebaker
 - Racism dictionaries: manually extracted from train data

Categories

- Racist
- Neutral
- Skin color
 - Brown
 - Black
- Nationality
 - North-African
 - East-European
- Belgian
- Religion
 - Islam
 - Judaism
- Culture
- Clothing
- Animals
- Diseases
- Immigrant
- Natives
- Criminal
- Insults
- Race
- Country
- Stereotype

Stylistic features

- Average sentence and word length
- Vocabulary richness
- POS-tags
- Punctuation
- Character bigrams
- Chatspeak features: emoticons, etc.

Results

- Train set (tenfold cross-validation)
 - F-score 0.71 (+/- 0.05)
- Test set
 - F-score **0.66**
 - Robust
- Baselines:
 - Weighted random baseline: 0.71
 - Majority baseline: 0.83

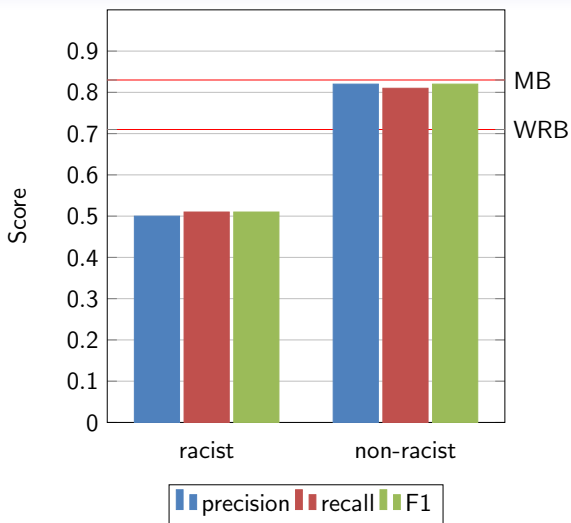


Figure: Precision, recall, and F1 for each class (test set)

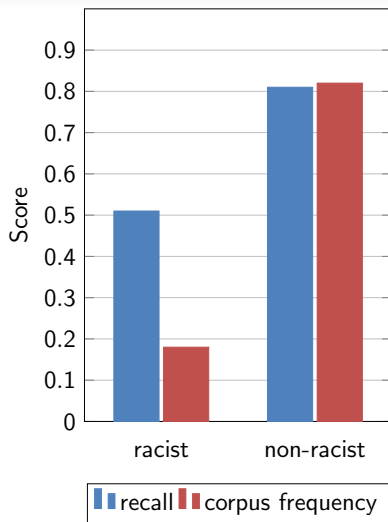


Figure: Recall vs. corpus frequency for each class

Most relevant features: interpretation

Words (unigrams)

- Us/them discourse
 - 'hunne'
'their'
- Insults, often concerning land of origin, religion...
 - 'ratten', 'zandbak', 'doctrine'
'rats', 'sandpit', 'doctrine'
- Islamic culture
 - 'moslim'
'Muslim'

Most relevant features: interpretation

Expressions (bigrams)

- Us/them discourse
 - 'onze cultuur', 'die islam'
'our culture', 'that Islam'
- Migration
 - 'terug naar', 'eigen land'
'back to', 'own country'

Relevance (current version of the) dictionaries?

1. Influence?
 - Predictable: derived from training data
 - Not much of a difference with or without dictionaries
2. Likely to generalize to unseen data?
 - Bound to our specific data

But: can be extended and optimized

Conclusion

1. Promising preliminary results:
 - Classifier reaches 0.66 f-score on test set
 - Quite robust
2. Important features:
 - Word counts (unigrams)
 - Word bigrams
 - Features concerning Islamic culture
3. Future work:
 - Optimization dictionaries
 - Experiments with word embeddings