Automatic detection of hate speech on social media

Ben Verhoeven
CLiPS, University of Antwerp

VivA Haatspraaksimposium
Johannesburg
17 June 2015

Why me?

- Computational linguist
  - Study and model language with computers
    - Language technology to deal with abundance of online text
- CLiPS, University of Antwerp
  - Projects on online safety
    - DAPHNE
    - AMiCA

What I will talk about

- Case study on racism detection
- Related research on ‘social’ cybersecurity
  - Detection of cyberharassment
- Introduction on text categorization

Text categorization

- Given a text and a predefined set of classes, predict the class the text belongs to
- Many applications
  - Spam
  - News article topics
  - Racism

    ...
- Methods
  - Handcrafted approach (rules)
  - Machine learning (since nineties)

Text categorization

- Many aspects of such systems
  - What data?
  - Class representation
  - Document representation
  - Supervised machine learning method

Data

- Text documents, preferably …
  - A lot!
  - Not too short (more words is more information)
  - Not too noisy
  - Already labeled appropriately for your task

Class representation

- Binary
  - Spam vs. ‘ham’
  - Man vs. woman writer
  - Racist vs. non-racist content

- Multi-class
  - Genres: blogs, news, jokes, novels, ...
  - Topics of reviews: books, phones, movies, …
  - Sort of discrimination: racism, sexism, ageism, …

Document representation

E.g. “I am smart, I am good.”

- Simplest method: to represent text as the distribution of words appearing in it
  - » I 2
  - » Am 2
  - » Smart 1
  - » Good 1
- N-grams: chunks of *n* consecutive items
  - trigrams: [I am smart], [am smart ,], [smart , I], [, I am], [I am good], [am good .]

Supervised machine learning

- Learning
  - Extract information from data
  - ‘Learn’ a model
- Machine
  - Automatic, with a computer
- Supervised
  - Labelled training data available
  - Predict output for new data

Supervised machine learning

- Different algorithms
  - Lazy learning
    - Learning: Store data in memory
    - Classification: Compare new data to data in memory
  - Eager learning
    - Learning: Abstract model from data
    - Classification: Apply abstracted model to new data

    - e.g. decision tree

Lazy learning

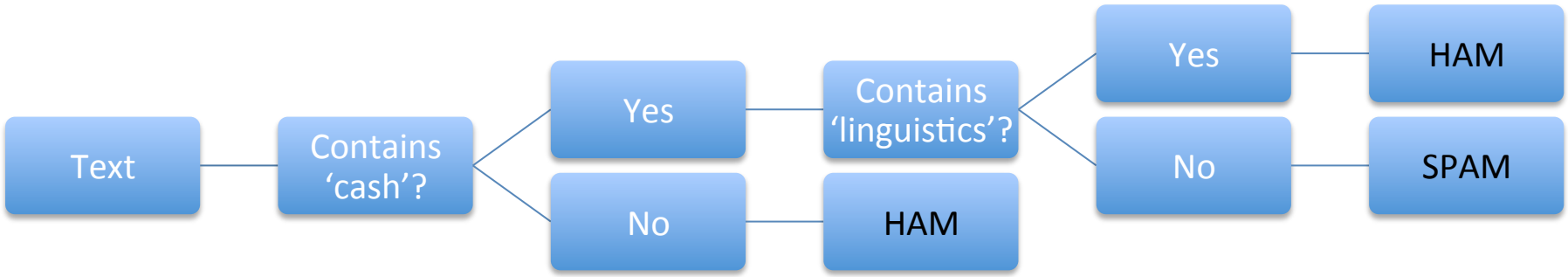
Compare each new instance to the others and find its nearest neighbour(s), assign the same class

- Analogy

This 'rule of nearest neighbour' has considerable elementary intuitive appeal and probably corresponds to practice in many situations. For example, it is possible that such medical diagnosis is influenced by the doctor's recollection of the subsequent history of an earlier patient whose symptoms resemble in some way those of the current patient. (Fix & Hodges, 1952, p.43)

Eager learning

- A decision tree is an abstract model of the patterns in the data, learned by the machine
- Example for spam detection



Evaluation

- Test your trained model on an independent test set

- Compare accuracy and more complex measures
- Look at well-performing features
- Where does it go wrong and why?

Language
Problematic for machine learning

Language is not an exact science

- Ambiguity at multiple levels
- Naive assumptions are false
- Humor
- Irony & sarcasm

Language ambiguity

- Lexical
  - Polysemy: word with different meanings (“book”)
  - Context matters (“fall terribly” vs. “terribly interesting”)
- Structural
  - Metaphor (“Her smile was like sunshine”)
  - Implication (“A bus!” could mean “Watch out!”)
  - Co-reference (“I am here”)

Naive assumptions

- Word order does matter
  - The woman hit the man.
  - The man hit the woman.
- Word occurrences are not independent
  - Syntax has rules, certain word forms demand others (Determiner + Noun)
  - Phrases, proverbs (“Kind regards”, “Raining cats and dogs”)
  - Whole idea of distributional semantics: a word can be understood by the company it keeps

Humor & Irony/Sarcasm

- The words aren't meant literally or even seriously
- Words might have different (emotional) meanings when used by different people

Online safety (for children)

Automatic Monitoring for Cyberspace Applications (AMiCA)

- Current project at CLiPS, UAntwerpen
  - Cyberharassment
  - Cyberpaedophilia
  - Suicide/depression

Cyberharassment

- The whole of harassing interactions online
  - Hate speech is a subset
  - Also contains
    - Sexual harassment (not paedophilia)
    - Bullying
  - Yet, not all insults are harassment

Cyberharassment

- Detection experiments using text categorisation
  - Harassment: binary
  - On text level: multi-class
    - Threat or Blackmail
    - Insult
    - Curse/Exclusion
    - Defamation
    - Sexual talk
    - Defense
    - Encouragement
    - Sarcasm

Cyberharassment

- Results
  - Harassment: ±55%
  - Text classes: 20 - 55%
- Still a lot of work, though, given the low frequency of such events, these results are not bad

Case study: racism detection

- MA student project at University of Antwerp this semester
  - Three students
  - Supervision: Walter Daelemans & myself
  - Results are fresh (presented 15 June 2015)

- Task: detect racism in user-generated concent

Definition

- Definition of racism needed for class labels
- Belgian law?
  - Discrimination and inciting hate is illegal, insulting isn't
- Common sense definition
  - Insults
    - Skin color
    - Ethnicity
    - Religion
  - Comparisons or generalizations

Data Collection

- All sensitive data
  - Extremely hard to get your hands on
- Regularly encountered problems
  - Privacy
  - Copyright
  - Sparsity
  - Not representative
  - High cost to label
- All scientific results depend on availability of representative data

Data Collection

- Interfederal Center for Equal Opportunities (CGKR)
  - Racism “hotline”
  - Referred us to 2 public Facebook pages with high ratio of racist posts
- Extract (sub)comments on first 100 posts on both pages
  - 5759 texts
  - Typically quite short
  - Unlabelled

Data Collection

- Interfederal Center for Equal Opportunities (CGKR)
  - Racism “hotline”
  - Referred us to 2 public Facebook pages with high ratio of racist posts
- Extract (sub)comments on first 100 posts on both pages
  - 5759 texts
  - Typically quite short
  - Unlabeled
- A separate test set of 620 texts was collected later for result validation

Annotation

Adding class labels to instances

Four labels:

1. Racist: the comment is insulting according to our racism definition.
   e.g.“Weg met alle niet Westerse buitenlanders”
2. Context: the comment itself wasn’t racist, but agreed with a previous racist post.
   e.g. “Ik ben het volledig met je eens”
3. Non-Racist: Default.
4. Invalid: the comment didn’t contain any text or wasn’t in Dutch.

Annotation

- Two annotators annotated the whole set
  - Agreement of ±80% on racist posts
    - Common sense definition seems to capture a pattern
    - One annotator was more sensitive than the other
  - Third annotator was the tie-breaker to create a gold-standard list of labels
    - Statistics
      - niet-racistisch: 4438
      - racistisch: 924
      - ongeldig: 335
      - context: 62

Annotation

- Ideally, no annotation needed
  - Data with labels
    - Report by users on social media
    - Deleted by moderator
- Advantage
  - Costs less
  - Users/moderators define the problem
    - Real-world setting
  - Easily updated/maintained

Features

- Word counts capture
  - Insults
  - Religious, racial, cultural terms, …
  - Pronouns (us/them)
- Word ngrams capture
  - Distancing (e.g. “that islam”)
  - Focus on self (e.g. “our culture”, “our country”)
- These are important features
  - Especially when combined

Features
Beyond mere word counts

- Stylistic
  - Average word/sentence length
  - Vocabulary richness
  - Punctuation marks & emoticons
  - Flooding: e.g. *I'm soooo tired*
- Content
  - Linguistic Inquiry and Word Count dictionaries
    - LIWC - James Pennebaker
    - Capture sociological & psychological factors
  - Self-created racism dictionaries

Racism dictionaries

- Data-based
  - Manually extracted all relevant terms from the training data
- Placed them into categories
  - Racist
  - Neutral
  - Skin color
    - Brown
    - Black
  - Nationality
    - North-African
    - Eastern-European
    - Belgian
  - Religion
    - Islam
    - Judaism
  - Culture
  - Clothing
  - Animals
  - Diseases
  - Immigrant
  - Natives
  - Criminal
  - Insults
  - Race
  - Country
  - Stereotype

Racism dictionaries

- Pro's
  - Abstracts over single words to categories
  - Can be used for sociological correlation research
    - What kind of people use what kinds of insults?
- Con's
  - Currently heavily biased towards
    - a Belgian context
    - Anti-Islam (source Facebook pages)
  - Quite small
    - 197 words in 23 categories
  - Contains ambiguous or context-dependent words

Classification

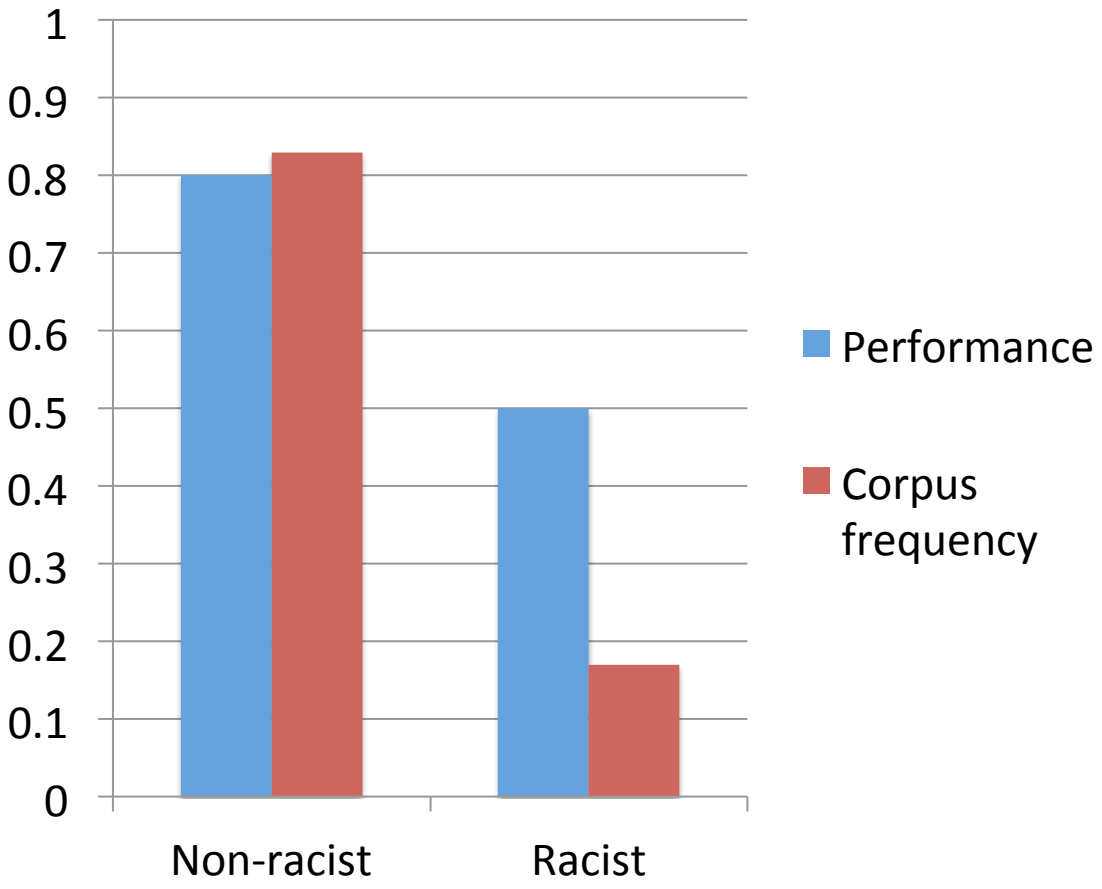
- What to do with label ‘context’?
  - Very infrequent
  - Semantically not racist, so we don’t want to label it that way
    - Maybe have a separate classifier look for reinforcements/encouragements of racism?
      - Similar to our cyberbullying approach

General results

- General baselines for comparison
  - Random baseline: 0.50
  - Weighted random baseline: 0.71
  - Majority baseline: 0.83

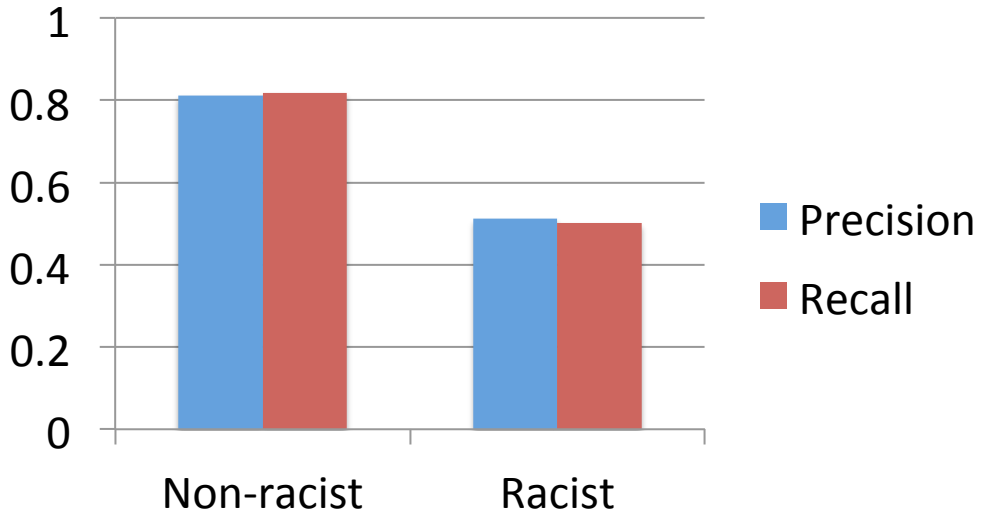
- Best-performing system
  - Within training set
    - F-score: 70%
  - On test set
    - F-score: 66%
    - Quite robust

Per-class results



Precision & Recall

- Precision: how many of our as ‘racist’ detected cases were in fact racist?
- Recall: how many of the existing ‘racist’ cases did we detect?



Precision & Recall

Practical Implications

- Building a system to detect and automatically delete the most severe transgression without human intervention?
  - Optimise for high precision because you don't want to delete too much.

- Building a system to support a human moderator in screening user-generated content.
  - Optimise for high recall because you don't want to miss anything. Main job is to reduce the moderator's work.

Evaluation

- Despite extra effort, simple word count approach works best
- However, serious limitations!
  - See language issues above
  - Naming function
    - E.g. “nigger is a bad word”
  - In-group use
    - E.g. ‘bitch’ is sometimes fine within a social group

Conclusion

- It's possible!
  - 66% on general classification
  - 50% on racist text
- Preliminary results
  - Small dataset
    - Not representative
  - Limited knowledge-based features
- Language is difficult for machine learning

Requirements to do this for Afrikaans?

Research

- (Labeled) data
- Computational linguist
  - Natural language processing
  - Machine learning
- Time (~money)

Development

- Software engineer

What steps to undertake?

- (Data collection)
- Assume a definition of racism
- Research what kind of system works best
  - Features
  - Algorithms
- Develop your moderation tool
  - Optimise for your purpose

- Join forces!

Thank you for your attention!

Acknowledgements:
- Walter Daelemans
- Lisa Hilte
- Elise Lodewyckx
- Stéphan Tulkens
- Gerhard van Huyssteen

CLiPS, University of Antwerp
www.clips.uantwerpen.be
@clipsua