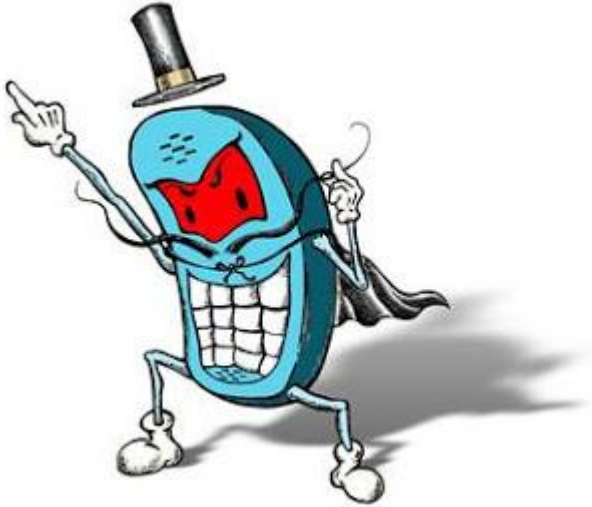
HATE BLOCKER WITH NATURAL LANGUAGE FRAMEWORK

Jan Schwarz

STRV

HATE BLOCKER



You are an ugly baby

HOW TO FIX IT

- Prevent evil users from using the app
- Prevent evil users from connecting with good users
- Prevent evil users from sending hateful messages

NATURAL LANGUAGE PROCESSING

STRV

WHAT IS NLP

- Semantics - meaning
- Sentiment - subjective information
- Entities - names, places, companies...
- Interaction between computers and human languages

NLP SUBTASKS

- Language identification
  - *Ninakuchukia*
- Tokenization - paragraphs, sentences, clauses, phrases, words…
  - *Mr. Jindřich Doležy joined STRV s.r.o. in 2016.*
  - *Mr || Jindřich Doležy joined STRV s || r || o || in 2016*
- Part of speech - noun, verb, preposition…
  - I want to fly like a fly.
  - Pronoun Verb Particle Verb Preposition Determiner Noun
- Lemmatization
  - I haven’t eaten for hours
  - I have eat for hour

NLP IN IOS

- NSLinguisticTagger
  - Language identification
  - Tokenization
  - Part of speech
  - Lemmatization
  - Named entity recognition
- Natural Language Framework
  - NSLinguisticTagger
  - Language models

NATURAL LANGUAGE FRAMEWORK VS. HATERS

STRV

DETECT HATEFUL WORDS

- Create a list of hateful words
- Analyze a message
- Ban messages that contain words from the blacklist

DEMO

STRV

ISSUES

- Potentially long list
- Ambiguity
  - Git - moron vs. versioning system
- Negation
  - You are a moron vs. You are not a moron
- (In)direct speech
  - Sara said you were a moron
- Context
  - You have a memory like an elephant vs. You have a bottom like an elephant
- ...

CAN WE DO BETTER?

TEXT CLASSIFIER

- Training dataset
- Build a model with CreateML
- Use Natural Language Framework to classify any text

TRAINING DATASET

- https://github.com/t-davidson/hate-speech-and-offensive-language
- 24783 labeled tweets

Hateful	Offensive	Neither	Class	Tweet
0	0	3	2	@mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should always take the trash out...
0	3	0	1	All I wanna do is get money and fuck model bitches!
3	0	0	0	@MarkRoundtreeJr: LMFAOOOO I HATE BLACK PEOPLE https://t.co/RNvD2nLCDR" This is why there's black people and niggers

WARNING:
THE FOLLOWING DEMO CONTAINS EXPLICIT LANGUAGE

STRV

ISSUES

- Training dataset
  - Hateful - 1430
  - Offensive - 19190
  - Neither - 4163
- CreateML is a blackbox

RECAP

- Natural Language Framework
  - Replacement of NSLinguisticTagger
  - MLTextClassifier, MLWordTagger
- Your model is just as accurate as your data is

- Demo: https://github.com/schwarja/hateblocker
- jan.schwarz@strv.com

THANK YOU

Jan Schwarz

STRV

QUESTIONS

STRV