Impact Of Content Features For Automatic Online Abuse Detection

Etienne Papegnies
Vincent Labatut, Richard Dufour, Georges Linares

Laboratoire Informatique d'Avignon

{firstname.lastname}@univ-avignon.fr

Task: Automatically Detect Abuse In An Online Community

- **Classify messages in two classes:**
  - **Abusive**
  - **Non-Abusive**
- **Abusive messages can be:**
  - **Straight Insults**
  - **Violations of the community usage guidelines**
- **Moderation done by hand is:**
  - **Expensive**
  - **Hard on the moderators**

Idea: tune preprocessing and detect impact of message

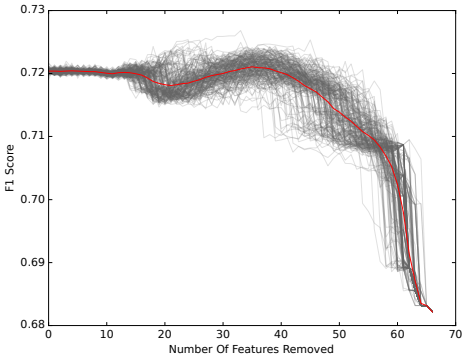
- We use specific preprocessing approaches for a target community
  - Meta Word for community jargon
  - Community-specific Deobfuscation
  - URL discrimination
- We develop a couple of features to detect impact of a message
  - Based only on messages from other users in response
  - These features are immune to intentional obfuscation

Results

Advanced preprocessing and new features improves classification score by 3.2 points

Data	Features	Preprocessing	Precision	Recall	*F* -Measure
iM+cM	**Classic set**	**Basic**	65.7	72.3	68.9
	Full set	**Advanced**	**68.3**	**76.4**	**72.1**

The new features and two others account for 15% Of classifier performance



Impact Of Content Features For Automatic Online Abuse Detection

Etienne Papegnies etienne.papegnies@univ-avignon.fr
Georges Linarès, Vincent Labatut and Richard Dufour (LI...
Pierre Gotab, Stanislas Oger (Nectar de Code)

Problem Description

Context

- In **Online communities** abuse is common.
- Community maintainers have to ensure moderation of user-gene tent
- → So users want to stay
- → Sometimes because the Law requires it
- Moderation is usually done by hand
- → It's expensive
- → It's hard for the moderators

Objective

- Develop an automatic system to assist moderation
- There is two tasks for this system
- → Automatically flag content for review by human moderators
- → Perform automatic moderation

challenges

- → Abbreviations, Typos
- → Images, URLs
- → Natural Language
- → Abuse can depend on context

Experiment and Results

SpaceOrigin Massively Multiplayer On-line Game

- Two types of player communications:
- → Messages from Internal mail system (iM)
- → Messages from Chat system (cM)
- Abusive messages reported by players then confirmed by moderators
- Non-Abusive messages are selected at random

Configuration	Abusive Messages	Non-Abusive Messages
iM+cM	779	1558
iM	111	222
cM	668	1336

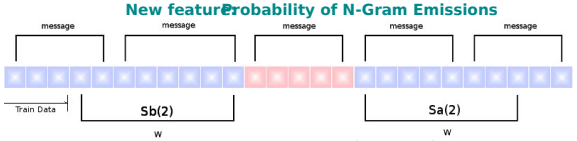
Features

- Classic Set
- → Morphological features
  - ∗ Message Length
  - ∗ Capitalized letters
  - ∗ Punctuation
- → Content Features
  - ∗ Bag Of Words
  - ∗ tf-idf weigths
  - ∗ Badwords
- Full Set
- → New content feature: PNE
- → PNE Applicability Criterion

Preprocessing

- Basic
- → Tokenization
- → Normalization
- Advanced
- → Elision Reversal
- → Stemming
- → Meta words for jargon
- → Word extraction for URLs
- → Community-Specific:
  - ∗ Deobfuscation
  - ∗ URL discrimination

New feature: Probability of N-Gram Emissions



$P_{i,i+1}$: Emission Probability of transition between i^{th} and $i+1^{th}$ n-grams in window W . Average Emission Probability S over W :

$$S = \frac{\sum_{i=0}^{W-1} P_{i,i+1}}{W} \quad (1)$$

$S_B(u), S_A(u)$: average probabilities before and after targeted message for user u. Final score $S(u)$ for user u is:

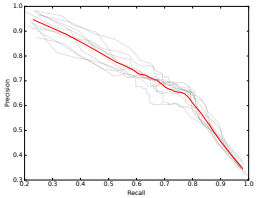
$$S(u) = S_A(u) - S_B(u) \quad (2)$$

Classification tasks results

Data	Features	Preprocessing	Precision	Recall	F -Measure
iM only	**Classic set**	**Basic**	66.9	72.8	69.7
	Full set	**Basic**	67.2	73.4	70.2
	Full set	**Advanced**	**69.6**	**76.2**	**72.8**
cM only	**Classic set**	**Basic**	65.2	71.6	68.2
	Full set	**Basic**	65.5	72.2	68.7
	Full set	**Advanced**	**67.6**	**75.9**	**71.5**
iM+cM	**Classic set**	**Basic**	65.7	72.3	68.9
	Full set	**Basic**	65.9	73.2	69.3
	Full set	**Advanced**	**68.3**	**76.4**	**72.1**

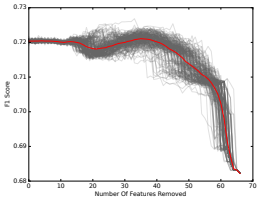
- Stacked Naive Bayes - SVM classifiers
- Results are averages for 10-fold cross-validation
- Advanced preprocessing increase performance by 2.8 points
- We have similar results for both message types
- The new features increase performance by 0.4 points
- → The new feature is immune to intentional obfuscation

Precision-Recall Curves



System can be tuned for automatic moderation or as a warning system by shifting the post-probability threshold.

Feature Selection



Drop at the end: 4 features account for 15% of classifier performance. Number of bad words, Average word length, PNE and it's Applicability criterion