Wykrywanie agresywności z polskojęzycznych postów użytkowników sieci socjalnych

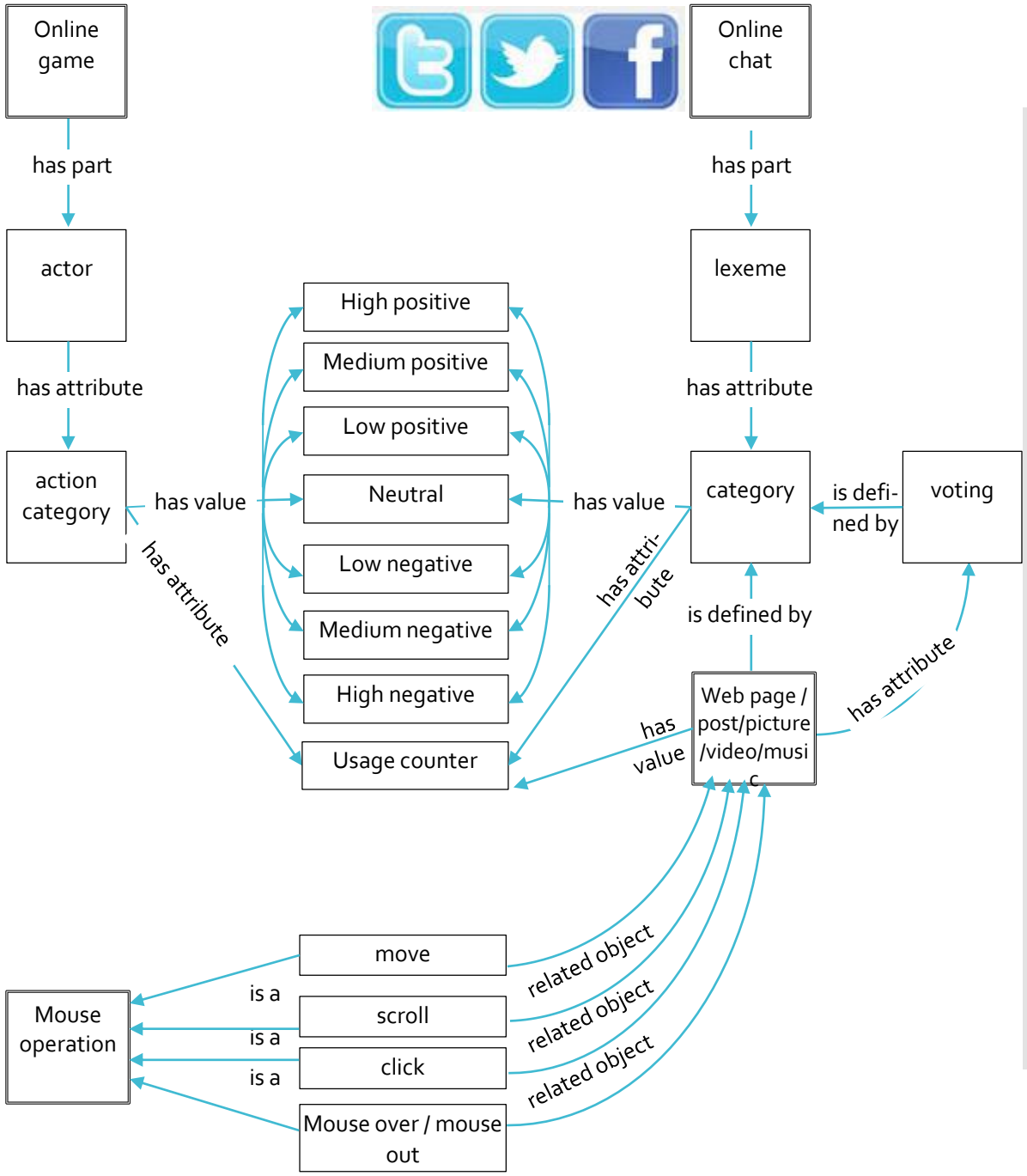
dr German Budnik (*german.budnik@uwb.edu.pl*)

Wydział Ekonomiczno-Informatyczny w Wilnie

Uniwersytet w Białymstoku

Motywacja

Motywacja
Dyskretyzacja danych o zachowaniu



Motywacja
Główna hipoteza badawcza

- Zebrane dane o zachowaniu osób w środowiskach wirtualnych, pozwolą wnioskować o wzorcach czynności agresywnych
- Uzasadnienie wybrania
  - Aktualność
  - Wyróżnienie

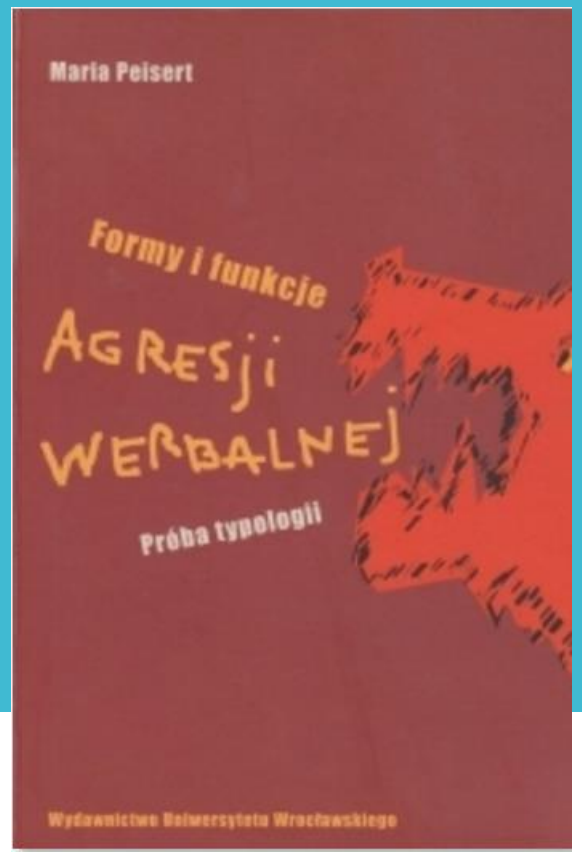
Motywacja
Zapotrzebowanie i aktualność

1. Złożone problemy społeczne spowodowane marginalnym zachowaniem osób (akty agresji, zachowanie aspołeczne itp.);
2. Zgłoszony konkurs Agencji Zaawansowanych Projektów Badawczych w Obszarze Obronności w zakresie DETECTION AND COMPUTATIONAL ANALYSIS OF PSYCHOLOGICAL SIGNALS [DARPA, 2014];
3. Nowoczesne prace w obszarze klasyfikacji i interpretacji czynności w środowiskach wirtualnych [Ho et al., 2014], [Gutschmidt, 2013];
4. Istnienie nowoczesnych metod implementacji zmiany zachowania z wykorzystaniem oprogramowania inteligentnego [Klein, 2014].

Kategorie agresywności

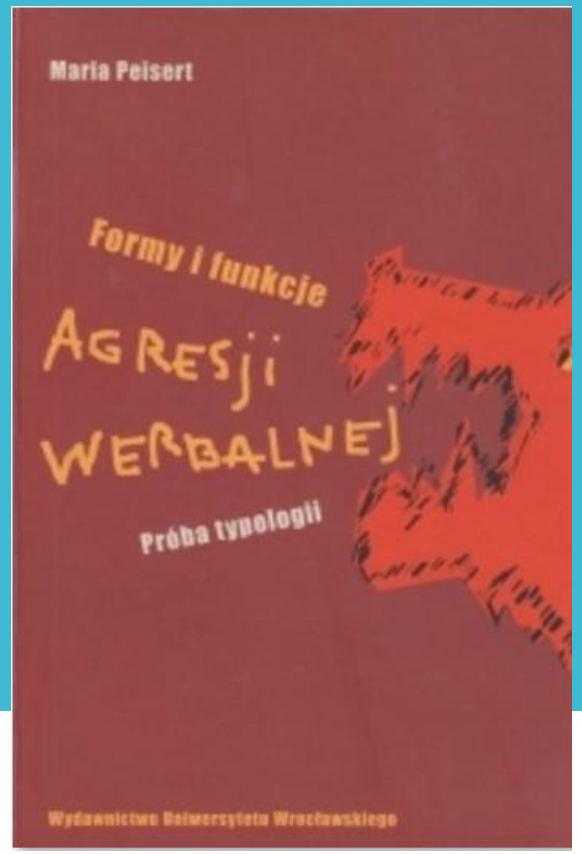
Active *Czynna*	Passive *Bierna*	Physical *Fizyczna*	Verbal *Słowna*	Without physical or verbal contact *Bez kontaktu fizycznego lub słownego*
C1.krzyki C2.groźby C3.wyśmiewanie C4.formy agresji fizycznej	B1.nieodzywanie się B2.naburmuszona mina B3.świadome 'granie na nerwach'	F1.bicie F2.popychanie F3.podcinanie F4.plucie F5.niszczenie własności F6.zmuszanie do pełnienia czynności	S1.obrażanie S2.groźba S3.szantażowanie	E1.grymasy E2.wrogie gesty E3.isolacja –zamykanie w pomieszczeniach

Funkcje agresywnych działań językowych



- Prof. **Maria Peisert** (Uniwersytet Wrocławski)
- „Dążność do poniżenia osoby oponenta, zapchnięcie jej na niższe poziomy hierarchii społecznej to jedna z najważniejszych funkcji działań językowych, które można włączyć w zakres pojęcia *agresja*."

Wyrządzenie szkody odbiorcy przez pomniejszenie jego statusu



1. Umieszczenie imienia oponenta w nieprzystojnym otoczeniu leksykalnym .
2. Metaforyczne przeniesienie na oponenta
   - nazw zwierząt, z którymi wiążą się negatywne konotacje semantyczne, np. *szczur* ,
   - nazw części ciała człowieka terminami należącymi do świata zwierząt, np. *r$$$j* .
3. Oskarżanie o naruszenie norm społecznych, np. *dz$$$ka*, *pe$$$ał* .
4. Zarzucenie oponentowi braku wiedzy, inteligencji, np. *idiota*, *bałwan*.

Wstępna analiza tekstów

- Oprogramowanie Instytutu Monitorowania Mediów pozwala na śledzenie informacji i opiniach o markach we wszystkich rodzajach mediów.

Wstępna analiza tekstów

- Zapytanie - szereg słów stosowanych dla metaforycznego określenia oponenta.
- 9191 postów w sieciach socjalnych (23-11-2015).
- 1134 przeczytano postów.
- 261- usunięto duplikatów

Wyniki analizy wstępnej

- 873 – podjęte klasyfikacji ręcznej w obiec modelu prof. M. Peisert
- Wykryte szablonowe wyrazy agresywne (*w formie podstawowej*):

1. <ty> <być> <w>
2. <wy> <być> <w>
3. <sam> <być> <w>
4. <być> <w>
5. <ty> <w>
6. <jak> <w>
7. <być> <jak> <w>
8. <co to za> <w>

* w – wyraz wulgarny

- Agresywnych postów – 140 ; Nieagresywnych postów – 733.

Analizator morfologiczny online

Rozpoznawanie części mowy w języku polskim i określanie ich właściwości.

Rozpoznawanie form podstawowych

```
<?xml version="1.0" encoding="utf-8"?>
<dokument>

  <zdanie id="1" kontekst="zwykly">
    <token id="1" slowo="Czasami" lemat="Czas" rodzajTokena="slowoPierwszaDuza" czescmowy="rzeczownik" przypade
    <token id="2" slowo="mam" lemat="mieć" rodzajTokena="slowoMale" czescmowy="czasownik" liczba="pojedyncza"
    <token id="2" slowo="mam" lemat="mieć" rodzajTokena="slowoMale" czescmowy="czasownik" liczba="pojedyncza"
    <token id="3" slowo="wrażenie" lemat="wrażony" rodzajTokena="slowoMale" czescmowy="czasownik" lewy="3" />
    <token id="3" slowo="wrażenie" lemat="wrazić" rodzajTokena="slowoMale" czescmowy="rzeczownik" przypadek="m
    <token id="3" slowo="wrażenie" lemat="wrazić" rodzajTokena="slowoMale" czescmowy="rzeczownik" przypadek="b
    <token id="4" slowo="," rodzajTokena="znak" />
    <token id="5" slowo="że" lemat="że" rodzajTokena="slowoMale" czescmowy="spojnik" typ="podrzednosciowy" />
    <token id="6" slowo="ten" lemat="ten" rodzajTokena="slowoMale" czescmowy="przymiotnik" stopien="podstawowy
    <token id="6" slowo="ten" lemat="ten" rodzajTokena="slowoMale" czescmowy="zaimek" typ="wskazujacy" przypade
    <token id="7" slowo="pan" lemat="pan" rodzajTokena="slowoMale" czescmowy="rzeczownik" przypadek="mianownik
    <token id="8" slowo="leci" lemat="lecieć" rodzajTokena="slowoMale" czescmowy="czasownik" liczba="pojedyncz
    <token id="8" slowo="leci" lemat="lecieć" rodzajTokena="slowoMale" czescmowy="czasownik" liczba="pojedyncz
    <token id="8" slowo="leci" lemat="lecieć" rodzajTokena="slowoMale" czescmowy="czasownik" liczba="pojedyncz
    <token id="8" slowo="leci" lemat="lec" rodzajTokena="slowoMale" czescmowy="rzeczownik" przypadek="dopelnia
```

Główne funkcje stworzonego prototypu

- Połączenie się z Analizatorem Morfologicznym dla przetwarzania tekstów
- Analiza i klasyfikacja przetworzonych tekstów
  - Wczytywanie wyrazów wulgarnych
  - Uwzględnienie zdań złożonych, fragmentując ich na zdania proste
  - Rozpoznawanie szablonowych wyrazów agresywnych
  - Uwzględnienie negacji w zdaniach prostych
  - Wpisywanie wyniku klasyfikacji

Połączenie się z Analizatorem Morfologicznym dla przetwarzania tekstów

- Interfejs REST API
- 1 post (wiadomość tekstowa) ~ 11 sekund
- 873 postów

Uwzględnienie zdań złożonych, fragmentując ich na zdania proste

- Kilka zdań (token </zdanie>)
- Znak (np. <,> / < ; >) i brak myślniku (< - >)
- < ale >
- < że > / < ze >
- < a > / < i >

Rozpoznawanie szablonowych wyrazów agresywnych

- Dla szablonowych wyrazów agresywnych
  1. <ty> <**być**> <w>
  2. <wy> <**być**> <w>
  3. <sam> <**być**> <w>
  4. <**być**> <w>
  5. <**być**> <jak> <w>
- podstawowa forma <**być**> uwzględniana, jeżeli słowem oryginalnym nie jest
  - <jestem>
  - <jest>
  - <jesteśmy>

Logika prototypu klasyfikującego

1. <ty> <być> <w>
2. <wy> <być> <w>
3. <sam> <być> <w>
4. <być> <w>
5. <ty> <w>
6. <jak> <w>
7. <być> <jak> <w>
8. <co to za> <w>

Przykład 1

Post oryginalny		Szablonowy wyraz agresywny
@justinbieber ja i @dallaspade chciałyśmy ci powiedzieć że jesteś i$$$ą.		Być W

```
<?xml version="1.0" encoding="utf-8"?>
<dokument>

  <zdanie id="1" kontekst="zwykly">
    <token id="1" slowo="ja" lemat="ja" rodzajTokena="slowoMale"
    <token id="2" slowo="i" lemat="i" rodzajTokena="slowoMale" cz
    <token id="2" slowo="i" lemat="i" rodzajTokena="slowoMale" cz
    <token id="3" slowo="@dallaspade" rodzajTokena="nieznany" />
    <token id="4" slowo="chciałyśmy" lemat="chcieć" rodzajTokena=
    <token id="5" slowo="ci" lemat="ty" rodzajTokena="slowoMale"
    <token id="5" slowo="ci" lemat="ty" rodzajTokena="slowoMale"
    <token id="6" slowo="powiedzieć" lemat="powiedzieć" rodzajTok
    <token id="7" slowo="że" lemat="że" rodzajTokena="slowoMale"
    <token id="8" slowo="jesteś" lemat="być" rodzajTokena="slowoM
    <token id="8" slowo="jesteś" lemat="być" rodzajTokena="slowoM
    <token id="9" slowo="i[redacted]ą" lemat="i[redacted]a" rodzajTokena="slo
    <token id="10" slowo="." rodzajTokena="znak" />
  </zdanie>
</dokument>
```

Przykład 2

Post oryginalny		Szablonowy wyraz agresywny
… Nie ma wątpliwości , że jesteś fanatycznym i$$$ą…		Być W

```
53      <token id="37" slowo="Nie" lemat="nie" rodzajTokena="slo
54      <token id="38" slowo="ma" lemat="mój" rodzajTokena="slow
55      <token id="38" slowo="ma" lemat="mieć" rodzajTokena="slo
56      <token id="38" slowo="ma" lemat="mieć" rodzajTokena="slo
57      <token id="38" slowo="ma" lemat="mieć" rodzajTokena="slo
58      <token id="39" slowo="wątpliwości" lemat="wątpliwość" ro
59      <token id="40" slowo="," rodzajTokena="znak" />
60      <token id="41" slowo="że" lemat="że" rodzajTokena="slowo
61      <token id="42" slowo="jesteś" lemat="być" rodzajTokena="
62      <token id="42" slowo="jesteś" lemat="być" rodzajTokena="
63      <token id="43" slowo="fanatycznym" lemat="fanatyczny" ro
64      <token id="44" slowo="i$$$ą" lemat="i$$$a" rodzajTokena=
65      <token id="45" slowo="..." rodzajTokena="znak" />
66  </zdanie>
```

Przykład 3

Post oryginalny		Szablonowy wyraz agresywny
Chcę powiedzieć, że lubię być namiętnym i szczerym, ale również lubię dobrze się bawić i działam jak i$$$a.		Być W *false positive* !

```
<token id="13" slowo="lubię" lemat="lubić" rodza
<token id="13" slowo="lubię" lemat="lubić" rodza
<token id="14" slowo="dobrze" lemat="dobrzeć" ro
<token id="14" slowo="dobrze" lemat="dobrzeć" ro
<token id="14" slowo="dobrze" lemat="dobrzeć" ro
<token id="14" slowo="dobrze" lemat="dobro" rodz
<token id="15" slowo="się" lemat="się" rodzajTok
<token id="16" slowo="bawić" lemat="bawić" rodza
<token id="17" slowo="i" lemat="i" rodzajTokena=
<token id="17" slowo="i" lemat="i" rodzajTokena=
<token id="18" slowo="działam" lemat="działać" r
<token id="18" slowo="działam" lemat="działać" r
<token id="18" slowo="działam" lemat="dziać" roo
<token id="19" slowo="jak" lemat="jak" rodzajTok
<token id="19" slowo="jak" lemat="jak" rodzajTok
<token id="19" slowo="jak" lemat="jak" rodzajTok
<token id="20" slowo="i[redacted]ta" lemat="i[redacted]ta" roo
<token id="21" slowo="." rodzajTokena="znak" />
```

Podsumowanie wyników klasyfikacji komputerowej

	Agresywnych	Nieagresywnych
# postów	140	733
# false positive	13	83
% false positive	9.3%	11,3%
# true positive	127	650
% true positive	90,7%	88,7%

Omówienie wyników

- Wada – 1 osoba klasyfikująca posty
- Możliwości dla polepszenia pracy klasyfikatora
  - nieprawidłowa pisownia wulgaryzmów
  - rozpoznanie zdania prostego z wymienieniem cech (np. *TY kłamliwa, $$$, mała <WULGARYZM>*.)
  - rozpoznanie zdania złożonego jako prostego (np. *nie jesteś kłamcą a zwykłym <WULGARYZM>*)
  - rozpoznanie jednego zdanie rozdzielonego znakiem - wygląda jako kilka zdań
  - polepszenie pracy analizatora morfologicznego
- Przyczyny *false positive*:
  - żarty (np. *ładny jak na szczura*)
  - wielu semantyczne słowo
  - cytowanie

Porównanie wyników

Modeling the Detection of Textual Cyberbullying

Karthik Dinakar[+]
karthik@media.mit.edu

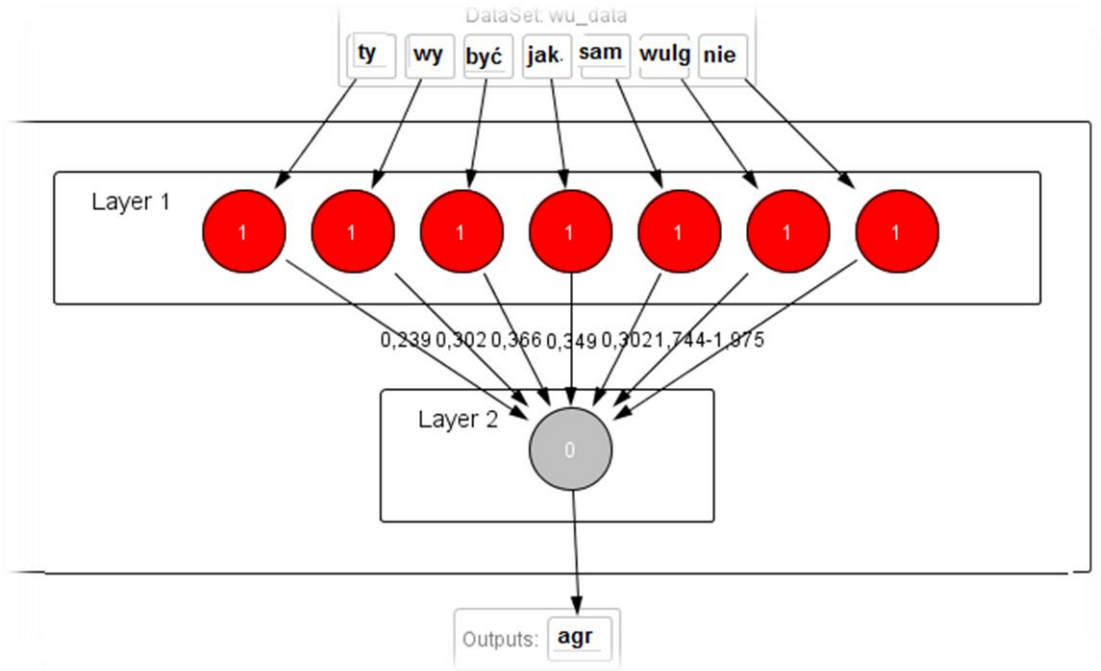
Roi Reichart[*]
roiri@csail.mit.edu

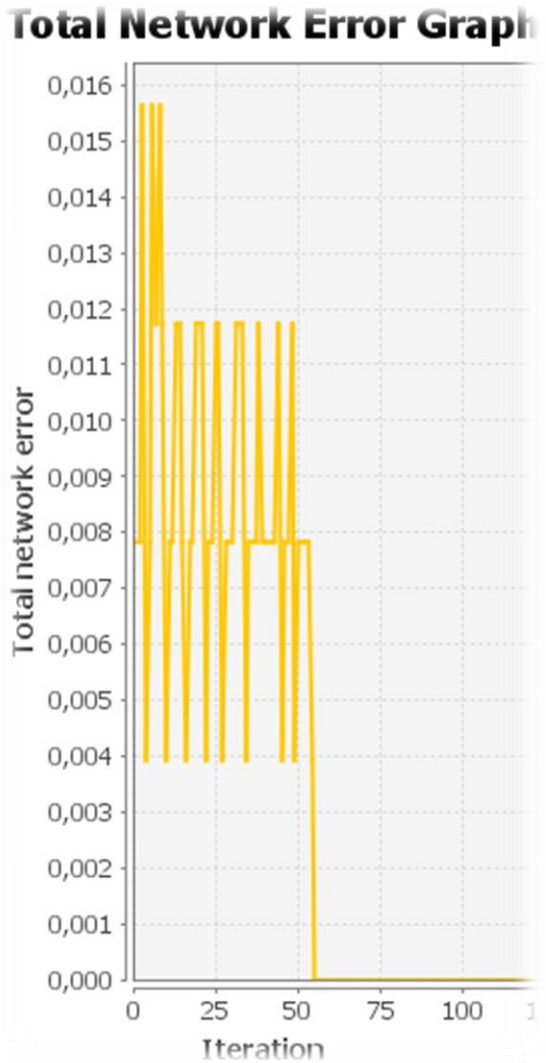
Henry Lieberman[+]
lieberman@media.mit.edu

[+]MIT Media Lab, [*]Computer Science & Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

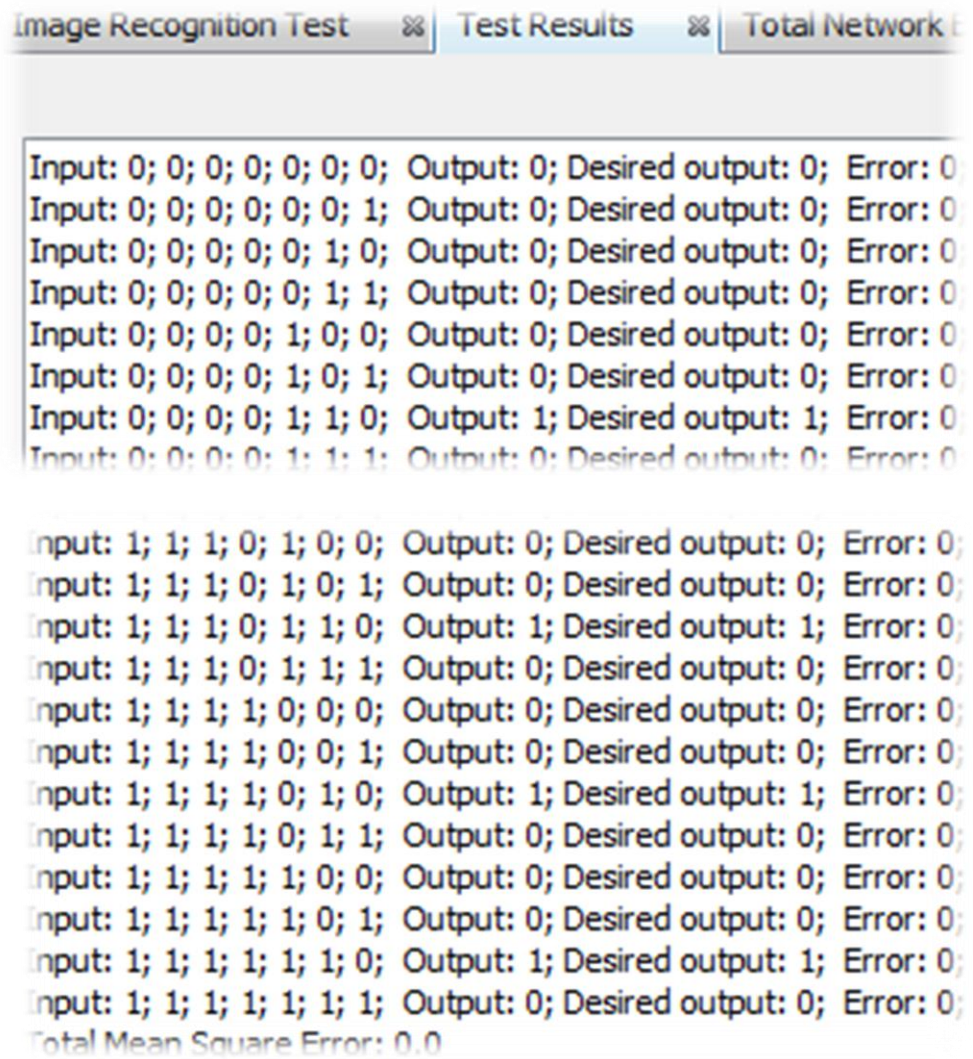
	Naïve Bayes		Rule-based Jrip		Tree-based J48		SMO (SVM)	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Sexuality	66%	0.657	**80.20%**	0.598	63.40%	0.573	66.70%	**0.79**
Race	66%	0.789	**68.30%**	0.789	63.50%	0.657	66.70%	**0.718**
Intelligence	72%	0.467	**70.39%**	0.512	70%	0.568	72%	**0.7723**

Neuroph Studio: perceptron - model





Neuroph Studio: perceptron - model



Porównanie wyników pracy prototypu oprogramowania i perceptron modelu

Podziękowania

- Profesor W. Homenda (PW)
- Profesor M. Muraszkiewicz (PW)
- Dr P. Sołdacki (astrafox.pl)
- Instytut Monitorowania Mediów (www.imm.com.pl)