

# Rozpoznawanie własności mówcy na podstawie treści wypowiedzi na przykładzie posłów na Sejm



Piotr Przybyła, Paweł Teisseyre

IPI PAN

27 listopada 2014

# Plan prezentacji

---

## Wstęp

## Dane

Korpus

Mówcy

## Klasyfikacja

Cechy

Selekcja zmiennych

Predykcja

## Wyniki

Wydajność klasyfikacji

Istotne zmienne

## Podsumowanie



UNIA EUROPEJSKA  
EUROPEJSKI  
FUNDUSZ SPOŁECZNY



Project co-financed by European Union within the framework of European Social Fund

# Wprowadzenie

---

## Cel:

- Identyfikacja własności posłów, takich jak:
  - płeć,
  - wykształcenie,
  - przynależność partyjna,
  - roku urodzeniana podstawie ich wypowiedzi w polskim Sejmie.

# Wprowadzenie

---

## Problemy związane z klasyfikacją tekstów:

- ekstrakcja cech wymaga użycia wiedzy i narzędzi lingwistycznych
- duża liczba cech  $p$  w porównaniu z liczbą rekordów  $n$  (u nas:  $p = 14.259.209$ ,  $n = 8780$ ) wymaga:
  - redukcji wymiaru danych i/lub selekcji cech,
  - zastosowania metod przystosowanych do sytuacji danych wysokowymiarowych.

# Wprowadzenie

---

## Podobne problemy:

- Bardzo popularne zagadnienie: klasyfikacja płci/wieku na podstawie **nieformalnych wypowiedzi** (blogi, Twitter, posty na forum, czaty, itp.), np. praca Argamon et. al. (2007).
- Najsilniejsze cechy powstają na podstawie specyficznych słów, np. "mom", "cried", "freaked", "gosh".

## Pytania:

- Czy jest możliwa efektywna klasyfikacja postów na podstawie ich bardzo **formalnych** wypowiedzi?
- Jakie cechy są najbardziej istotne?
- Które metody klasyfikacji/regresji/selekcji cech działają efektywnie?
- W jakim stopniu odmiana czasownika w języku polskim ułatwia zadanie klasyfikacji płci?

## Korpus sejmowy

---

Korpus wypowiedzi sejmowych, opisany w Ogrodniczuk (2012) i dostępny via CLIP <http://clip.ipipan.waw.pl/PSC>.

- 6 kadencji (lata 1991-2011), podzielonych na posiedzenia, podzielonych na wypowiedzi z przypisanym autorem.
- Oznaczony automatycznie (format NKJP):
  - analiza morfosyntaktyczna (*Morfeusz SGJP*),
  - ujednoznacznianie (*PANTERA*),
  - grupy składniowe (*Spejd*),
  - nazwy własne (*NERF* + *Quant* (z RAFAELa)),
- 300 milionów segmentów,
- pominęliśmy interpelacje i pytania (jedynie posiedzenia).

# Przykład

```
<div xml:id="div-146">
  <u xml:id="u-146.1" who="#WicemarszałekDariuszWojcik">Panie pośle, w art. 63 ust. 2, którym się
pan posłużył, stwierdzono, że dopuszcza się możliwość dyskusji na temat podjęty w interpelacji. Pan
po prostu nie przeczytał ust. 2 w art. 63.</u>
  <u xml:id="u-146.2" who="#PosełMarekMuszynski">(Jest jednak określony czas.)</u>
  <u xml:id="u-146.3" who="#WicemarszałekDariuszWojcik">Panie pośle, w ust. 2 nie określa się
czasu. Czas jest określony w ust. 1 - 10 minut na odpowiedź, 5 minut na dodatkową odpowiedź na
pytania zadane z sali, a o możliwości dyskusji mówi ust. 2.</u>
  <u xml:id="u-146.4" who="#komentarz">(Oklaski)</u>
  <u xml:id="u-146.5" who="#WicemarszałekDariuszWojcik">Regulamin jest więc przestrzegany.</u>
  <u xml:id="u-146.6" who="#WicemarszałekDariuszWojcik">Ja natomiast z innego powodu apeluję o
to, aby przestrzegać tematyki interpelacji, zadawać krótkie i precyzyjne pytania i udzielać takich
właśnie odpowiedzi. Będziemy mieli bowiem kłopoty ze zrealizowaniem punktów porządku dziennego, a
pozostały jeszcze do omówienia bardzo istotne sprawy, które nie powinny czekać, bo społeczeństwo
oczekuje od nas szybkich działań.</u>
  <u xml:id="u-146.7" who="#WicemarszałekDariuszWojcik">Chciałbym zapytać pana posła, czy po moim
wyjaśnieniu podtrzymuje pan swój wniosek formalny?</u>
</div>
<div xml:id="div-147">
  <u xml:id="u-147.1" who="#PosełMarekMuszynski">W takim razie, jeśli można, zgłaszam wniosek o
przerwanie dyskusji na ten temat.</u>
  <u xml:id="u-147.2" who="#komentarz">(Poruszenie na sali)</u>
</div>
<div xml:id="div-148">
  <u xml:id="u-148.1" who="#WicemarszałekDariuszWojcik">Proszę państwa, czy ktoś chciałby zabrać
głos w sprawie wniosku formalnego?</u>
  <u xml:id="u-148.2" who="#WicemarszałekDariuszWojcik">Proszę bardzo, pan poseł Marek Borowski.</
u>
```

## Znakowanie – tekst na wejściu

---

*W białym płaszczu z podbiciem koloru krwawnika, posuwistym krokiem kawalerzysty, wczesnym rankiem czternastego dnia wiosennego miesiąca nisan pod krytą kolumnadę łączącą oba skrzydła pałacu Heroda Wielkiego wyszedł procurator Judei Poncjusz Piłat.*



## Znakowanie – segmentacja

---

*W białym płaszczu z podbiciem koloru krwawnika , posuwistym krokiem kawalerzysty , wczesnym rankiem czternastego dnia wiosennego miesiąca nisan pod krytą kolumnadę łączącą oba skrzydła pałacu Heroda Wielkiego wyszedł procurator Judei Poncjusz Piłat .*

# Znakowanie – tagowanie

---

w:prep:loc:nwok biały:adj:sg:loc:m3:pos płaszcz:subst:sg:loc:m3  
z:prep:inst:nwok podbić:ger:sg:inst:n:perf:aff kolor:subst:sg:gen:m3  
krwawnik:subst:sg:gen:m3 ,:interp posuwisty:adj:sg:inst:m3:pos  
krok:subst:sg:inst:m3 kawalerzysta:subst:sg:gen:m1 ,:interp  
wczesny:adj:sg:inst:m3:pos ranek:subst:sg:inst:m3 czternasty:adj:sg:gen:m3:pos  
dzień:subst:sg:gen:m3 wiosenny:adj:sg:gen:m3:pos miesiąc:subst:sg:gen:m3  
nisan:ign pod:prep:acc:nwok kryty:adj:sg:acc:f:pos kolumnada:subst:sg:acc:f  
łączyć:pact:sg:inst:f:imperf:aff oba:num:pl:nom:n:congr  
skrzydło:subst:pl:nom:n pałac:subst:sg:gen:m3 herod:subst:sg:gen:m1  
wielki:adj:sg:gen:m1:pos wyniść:praet:sg:m1:perf procurator:ign  
Judea:subst:sg:gen:f Poncjusz:subst:sg:nom:m1 piłat:subst:sg:nom:m1 .:interp

## Znakowanie – płytkie parsowanie

---

(PrepNG W (NGa białym płaszczu)) (PrepNG z (NGg podbiciem koloru krwawnika)) , (NGg (NGa posuwistym krokiem) kawalerzysty) , (NGg (NGa wczesnym rankiem) (NGa czternastego dnia) (NGa wiosennego miesiąca)) nisan (PrepNG pod (NGa krytą kolumnadę)) łączącą (NumGz oba (NGg skrzydła pałacu (NGg Heroda Wielkiego))) wyszedł procurator (NG Judei) (NGs Poncjusz Piłat) .

## Znakowanie – nazwy własne

---

W białym płaszczu z podbiciem koloru krwawnika , posuwistym krokiem kawalerzysty , wczesnym rankiem czternastego dnia wiosennego miesiąca nisan pod krytą kolumnadę łączącą oba skrzydła pałacu (persName Heroda) Wielkiego wyszedł procurator (persName (forename Judei) (surname Poncjusz) (surname Piłat)) .

# Mówcy

## Posłowie IV kadencji 2001 - 2005



**Bronisław Cieślak**

Data i miejsce urodzenia: 08.10.1943, Kraków  
Stan cywilny: żonaty

Wykształcenie: średnie ogólne  
Ukończona szkoła: Liceum Ogólnokształcące

## Klub Parlamentarny Sojuszu Lewicy Demokratycznej

Staż parlamentarny: poseł III kadencji  
Data ślubowania: 19.10.2001  
Lista: Sojusz Lewicy Demokratycznej - Unia Pracy  
Okręg wyborczy: 13 Kraków  
Liczba głosów: 24252

Automatyczne pozyskiwanie ze strony sejmu skryptem w Pythonie.

## Wydobywane własności:

- rok urodzenia (ciągły, uwzględniając dzień i miesiąc), 1920-1985,  $\mu=1953$ ,  $\sigma=9,4$
- płeć (na podstawie imienia), kobiety: 17%
- partia (komitet wyborczy), 4-6 na kadencję
- wykształcenie, uproszczone do:
  - podstawowe (1,5%),
  - średnie zawodowe (14%),
  - średnie ogólne (1,7%),
  - policealne (2,5%),
  - wyższe (71%),
  - doktorat (9%).

## Budowa zadania klasyfikacji

---

1. Grupowanie wypowiedzi według posta (kadencje osobno),
2. Podział grupy na podgrupy po 100 wypowiedzi,
3. Każda podgrupa pojedynczym *przypadkiem uczącym*, traktowanym jak ciągły tekst,  
⇒ spośród 2.909 postów tylko 2.133 miało min. 100 wypowiedzi,  
⇒ powstało 8.780 przypadków uczących, 2030-16.720 słów, średnio 7.937
4. Ekstrakcja wektora cech dla każdego tekstu; przypisanie własności autora do tekstu,
5. Klasyfikacja.

! Nie korzystamy ze związku między wypowiedziami o wspólnym autorze.

# Cechy statystyczne

---

- średnia długość słowa w znakach @avgWordLength,
- średnia długość zdania w słowach @avgSenLength,
- udział słów dłuższych niż 6 liter @longWords,
- udział słów dłuższych niż 8 liter @longlongWords,
- udział słów dłuższych niż 10 liter @longlonglongWords,
- średnia liczba grup składniowych na zdanie @groupsPerSen,
- średnia długość grup składniowych @groupsLength,
- średnia liczba nazw własnych na zdanie według typów:
  - liczb @oneNumsPerSen,
  - osób @onePersonsPerSen,
  - miejsc @onePlacesPerSen,
  - dat @oneDatesPerSen.

# Cechy leksykalne

---

Częstości (liczność/długość tekstu) tokenów:

- form powierzchniowych,
- lematów (pref. \$),
- interpretacji (pref. :),
- bigramy:
  - form powierzchniowych (pref. #),
  - lematów (pref. # \$),
  - interpretacji (pref. # :),
  - części mowy (pref. # ;).

*czarne koty* ⇒

- czarne
- koty
- \$czarny
- \$kot
- :adj:pl:nom:m2:pos
- :subst:pl:nom:m2
- #czarne\_koty
- # \$czarny\_kot
- #:adj:pl:nom:m2:pos\_  
subst:pl:nom:m2
- #;adj\_subst

⇒ 14.259.202 różnych tokenów, 37.916 w co najmniej 10 przypadkach



# Przekształcony tekst

---

- To \$to :qub #BEG\_To #\$BEG\_to #:BEG\_qub #;BEG\_qub
- dlatego \$dlatego :adv #To\_dlatego #\$to\_dlatego #:qub\_adv #;qub\_adv
- , \$, :interp #dlatego\_, #\$dlatego\_, #:adv\_interp #;adv\_interp
- Wysoki \$wysoki :adj:sg:voc:m3:pos #,\_Wysoki #\$\_wysoki  
#:interp\_adj:sg:voc:m3:pos #;interp\_adj
- Sejmie \$sejm :subst:sg:voc:m3 #Wysoki\_Sejmie #\$wysoki\_sejm  
#:adj:sg:voc:m3:pos\_subst:sg:voc:m3 #;adj\_subst
- , \$, :interp #Sejmie\_, #\$sejm\_, #:subst:sg:voc:m3\_interp #;subst\_interp
- na \$na :prep:loc #,\_na #\$\_na #:interp\_prep:loc #;interp\_prep
- początku \$początek :subst:sg:loc:m3 #na\_początku #\$na\_początek  
#:prep:loc\_subst:sg:loc:m3 #;prep\_subst
- zacytował \$zacytować :praet:sg:m1:perf #początku\_zacytował  
#\$początek\_zacytować #:subst:sg:loc:m3\_praet:sg:m1:perf #;subst\_praet
- em \$być :aglt:sg:pri:imperf:wok #zacytował\_em #\$zacytować\_być  
#:praet:sg:m1:perf\_aglt:sg:pri:imperf:wok #;praet\_aglt

Pliki wczytywane i przeliczane do DocumentTermMatrix przez pakiet

# Selekcja cech

---

## Cechy:

- 14, 259, 202 (cechy leksykalne)
- 7 (cechy statystyczne)

## Selekcja zmiennych:

1. Odrzucenie tokenów występujących w mniej niż 10 przypadkach uczących (co daje 37, 916 cech).
2. Wybór cech z najwyższą wartością informacji wzajemnej (dla klasyfikacji) i największą korelacją (dla regresji).
3. Budowa modelu przy użyciu: 500, 5000, 10000 najbardziej istotnych cech (+wszystkie).
4. Cechy statystyczne są zawsze włączone do modelu.

Na koniec ocena istotności z lasu losowego.

## Selekcja cech

---

### Miary wykorzystane do selekcji zmiennych (pakiet FSelector w R):

- Przyrost informacji:

$$IG(y, x) := \frac{H(y) - H(y|x)}{H(x)},$$

gdzie  $H$  oznacza entropię.

- $1 \geq IG(y, x) \geq 0$ ,
  - $IG(y, x) = 0$  jeżeli  $x$  i  $y$  są **niezależne**.
- Współczynnik korelacji liniowej:

$$COR(y, x) := \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

# Selekcja cech

---

## Miara oceny istotności cech:

- Miara oparta na lasach losowych (pakiet RandomForest w R):

$$VI(x) := \sum_{(\text{drzewa})} \sum_{(\text{węzły zawierające } x)} IG(y, x > t)$$

gdzie:

- $x > t$  jest testem użytym w danym węźle do podziału danych.
- $IG(y, x > t)$  jest przerostem informacji  $y$  związanym z podziałem danych.
- Miara przyjmuje duże wartości, jeżeli:
  - zmienna  $x$  była często wykorzystywana do podziałów.
  - użycie zmiennej  $x$  do podziału zmniejszyło entropię  $y$ .

# Predykcja

---

## 3 problemy:

- klasyfikacja binarna (płeć)
- klasyfikacja wieloklasowa (wykształcenie, partie)
- regresja (rok urodzenia)

# Predykcja

---

## **Model klasyfikacyjne:**

- regresja logistyczna z regularyzacją (LOGISTIC, pakiet `glmnet`)
- drzewa decyzyjne (CART, pakiet `rpart`)
- lasy losowe (RF, pakiet `randomForest`)
- maszyny wektorów podpierających (SVM, pakiet `e1071`)
- naiwny klasyfikator Bayesa (NB, pakiet `e1071`)
- metoda najbliższych sąsiadów (KNN, pakiet `class`)

# Predykcja

---

## Model regresyjne:

- drzewa regresyjne (CART, pakiet `rpart`)
- lasy losowe (RF, pakiet `randomForest`)
- maszyny wektorów podpierających (SVM, pakiet `e1071`)
- regresja liniowa z regularyzacją (LASSO, pakiet `glmnet`)
- metoda najbliższych sąsiadów (KNN, pakiet `class`)
- regresja składowych głównych (PCR, pakiet `pls`)
- regresja częściowych najmniejszych kwadratów (PLSR, pakiet `pls`)

# Predykcja

---

## Kroswalidacja 5-krotna, ocena według wskaźników:

- Dokładność klasyfikacji:

$$Accuracy := \frac{\#\{y_i = \hat{y}_i\}}{n}$$

- Czułość (recall):

$$Recall := \frac{\#\{y_i = +, \hat{y}_i = +\}}{\#\{y_i = +\}}$$

- Precyzja (precision):

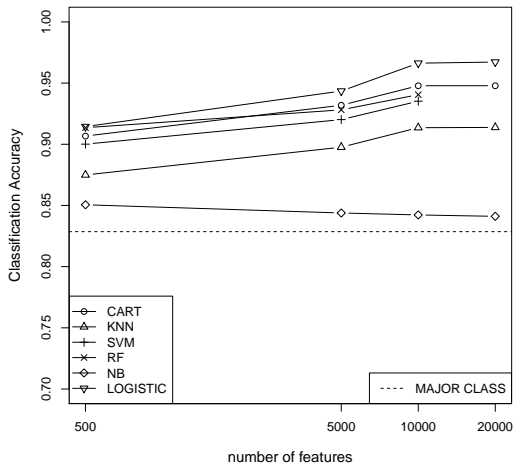
$$Precision := \frac{\#\{y_i = +, \hat{y}_i = +\}}{\#\{\hat{y}_i = +\}}$$

- Spierwiastkowany błąd średniokwadratowy (RMSE):

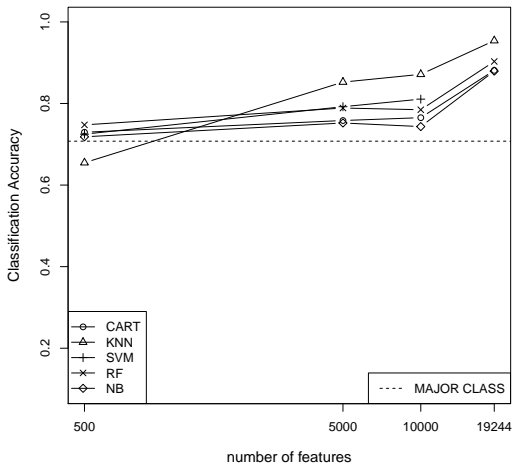
$$RMSE := \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$



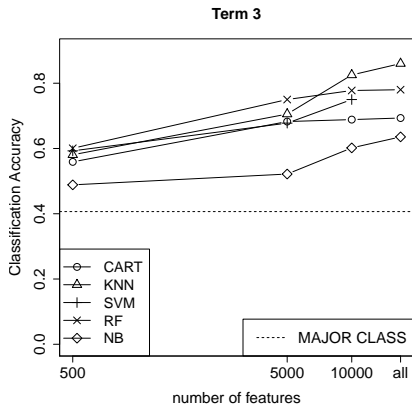
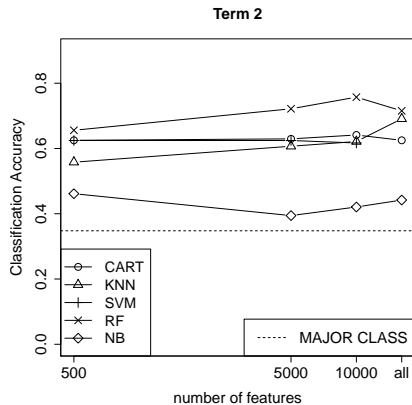
# Płeć



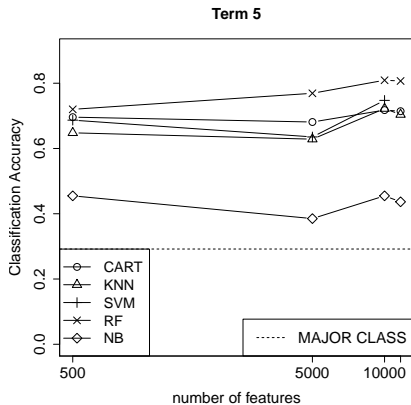
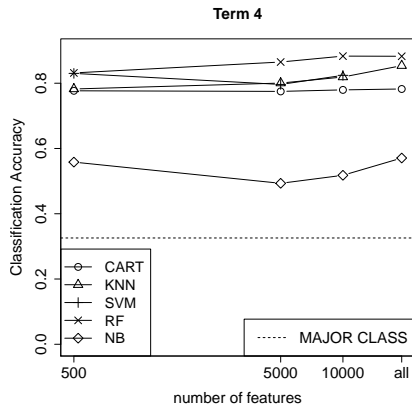
# Wykształcenie (4 kadencja)



## Partie (2, 3 kadencja)



## Partie (4, 5 kadencja)

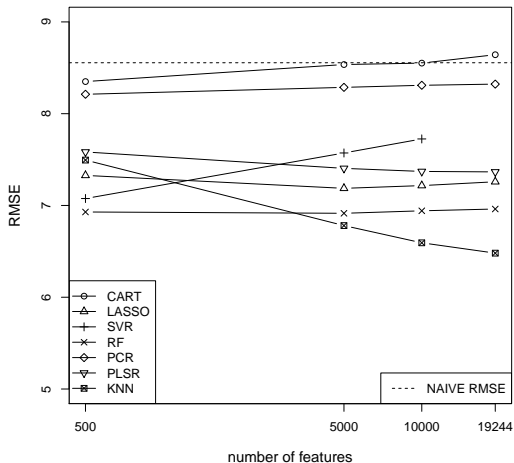


# Partie

---

Partia		2 kad.	3 kad.	4 kad.	5 kad.	6 kad.
PSL	P:	91%	88%	98%	100%	100%
	R:	58%	84%	67%	83%	82%
SLD	P:	64%	79%	81%	82%	92%
	R:	92%	86%	98%	82%	64%
PIS	P:			99%	69%	74%
	R:			87%	93%	92%
Samoobrona	P:			93%	100%	
	R:			89%	76%	
PO	P:			98%	75%	82%
	R:			74%	65%	77%
Największa partia		SLD	AWS	SLD	PIS	PO

# Rok urodzenia



# Podsumowanie

---

Zadanie	Metoda	Liczba cech	Dokładność	Poziom odniesienia
Płeć	LOGISTIC	wszystkie	96,72%	82,85%
Wykształcenie (4 kad.)	KNN	wszystkie	95,40%	70,76%
Partia (2 kad.)	RF	10000	75,74%	34,77%
Partia (3 kad.)	KNN	wszystkie	86,02%	40,66%
Partia (4 kad.)	RF	10000	88,31%	32,57%
Partia (5 kad.)	RF	10000	80,91%	29,19%
Partia (6 kad.)	RF	wszystkie	81,83%	39,41%
Rok urodzenia	KNN	wszystkie	6,48	8,55

## Istotne zmienne – płeć

---

- najsilniejsze 5 cech to interpretacje (i ich bigramy), przesądzające o rodzaju: `:praet:sg:m1:imperf :praet:sg:f:imperf`,  
`#:praet:sg:f:imperf_qub`, `:aglt:sg:pri:imperf:nwok`,  
`#:praet:sg:m1:imperf_qub`
- dalej także formy powierzchniowe powstałe z *chciałam*, *chciałabym*,
- słowa typowe dla kobiet (oraz ich lematy): *dzieci*, *państwo*, *kobieta*, *dziecko* i *rodzina*,
- kobiety używają dłuższych słów, wszystkie cechy: `@avgWordLength`, `@longWords`, `@longlongWords` i `@longlonglongWords`,
- mężczyźni częściej stosują nazwy własne: `@nePersonsPerSen` i `@nePlacesPerSen`.



# Istotne zmienne – płeć

Feature	VI	Average feature value for genders	
		F	M
:praet:sg:m1:imperf	131.02	3.2	8.71
:praet:sg:f:imperf	119.95	7.62	3.44
#:praet:sg:f:imperf_qub	103.91	2.1	0.06
:aglt:sg:pri:imperf:nwok	68.47	4.56	2.11
#:praet:sg:m1:imperf_qub	66.35	0.01	2.36
chciała	65.69	1.48	0
m	62.98	5.28	2.79
#chciała_by	59.01	1.16	0
chciał	37.52	0	1.88
#:praet:sg:f:imperf_aglt:sg:pri:imperf:nwok	33.5	0.52	0
#:beg_praet:sg:f:imperf	32.1	0.63	0
:praet:sg:f:perf	31.77	5.3	3.44
#chciała_by	27.15	0	1.46
#:praet:sg:f:perf_aglt:sg:pri:imperf:nwok	24.79	0.35	0
\$dziecko	20.46	2.38	0.4
#beg_chciała	19.11	0.41	0
dzieci	14.15	0.92	0.14
@nePersonsPerSen	13.6	0.06	0.08
#:beg_praet:sg:m1:imperf	13.23	0	0.88
\$kobieta	12.42	0.62	0.07
państwo	11.16	1.13	0.49
@longWords	10.98	433.74	424.37
@nePlacesPerSen	10.5	0.09	0.1

## Istotne zmienne – partia (3. kadencja)

---

- najmocniejsza cecha: \$? posłowie partii opozycyjnych zadawali pytania dwukrotnie częściej (AWS vs SLD),
- niektóre partie są rzadko wspominane przez pozostałe (PSL, UW),
- opozycja częściej odnosiła się do *rządu*, *ministrów*, a także używała słowa *pan* i wołacza (:subst:sg:voc:m1).

## Istotne zmienne – partia (3. kadencja)

		Average feature value for party affiliation				
Feature	VI	AWS	PSL	ROP	SLD	UW
\$?	5.07	6.43	8.21	6.36	12.22	2.92
#\$polski_stronnictwo	4.96	0	2.44	0	0	0
?	4.91	6.43	8.21	6.36	12.22	2.92
\$stronnictwo	4.88	0.05	2.48	0	0	0.03
rząd	4.78	0.92	3.04	0.67	3.17	0.82
\$minister	4.73	4.86	6.72	3.75	9.78	3.1
#\$stronnictwo_ludowy	4.65	0	2.45	0	0	0
#\$unia_wolności	4.54	0.05	0.01	0.2	0.08	1.4
\$ludowy	4.4	0.06	2.65	0	0.02	0
wolności	4.29	0.15	0.14	0.21	0.13	1.68
#\$unii_wolności	4.11	0.02	0	0.11	0.03	1.08
minister	3.84	0.88	1.77	0.64	3.05	0.63
stronnictwa	3.79	0.05	1.96	0	0	0.02
\$pan	3.75	14.36	14.38	15.94	22.81	10.52
:subst:sg:voc:m1	3.75	12.96	13.8	11.16	19.74	8.19

## Istotne zmienne – wykształcenie

---

- pierwsze 10 pozycji to cechy pochodne od wyrażenia *klub parlamentarny Samoobrona Rzecz(y)pospolitej Polskiej*,
- żaden doktor nie użył ani razu słowa *samoobrona*,
- następnie cechy związane z długością słów, choć niemonotonicznie z długością edukacji (doktorzy nie zawsze lepiej od magistrów, absolwenci szkół zawodowych gorzej od podstawowych),
- z czasem edukacji rośnie również złożoność zdań: @groupsPerSen i @groupsLength.

# Istotne zmienne – wykształcenie

Feature	VI	Average feature value for education					
		A	B	C	D	E	F
\$samoobrona	5.82	3.45	2.89	1.14	1.25	0.25	0
#\$parlamentarny_samoobrona	4.57	1.95	1.13	0.9	1.05	0.11	0
#\$samoobrona_rzeczpospolita	4.19	1.79	1.02	0.76	0.31	0.08	0
samoobrona	4.14	2.73	1.76	0.9	1.13	0.16	0
#:substsg:nom:f_substsg:gen:f	3.7	3.26	2.08	1.46	1.66	0.84	0.86
#rzeczypospolitej_polskiej	3.68	2.04	1.61	0.89	0.35	0.33	0.37
#samoobrona_rzeczypospolitej	3.25	1.58	0.79	0.65	0.26	0.07	0
#\$rzeczpospolita_polski	3.09	2.28	1.64	0.91	0.49	0.38	0.42
rzeczypospolitej	2.67	2.73	2.02	1.17	0.8	0.61	0.57
\$polski	2.67	8.59	11.14	12.9	6.26	7.11	8.79
@longlonglongWords	2.63	110.14	101.96	115.1	113.09	118.61	118.64
@longlongWords	2.63	236.33	224.24	243.52	235.66	246.23	244.62
@avgWordLength	2.37	5.93	5.84	6.04	5.95	6.02	6
polskiej	2.12	2.91	2.67	2.17	0.88	1.1	1.45
#:ger_subst	2.12	17.55	14.61	17.93	16.54	18.07	16.34
\$nasz	2.1	3.03	4.23	2.92	3.95	2.7	3.97
\$rzeczpospolita	2.06	3.01	2.16	1.29	1.05	0.74	0.64
#:fin_interp	2.06	19.02	20	15.79	16.34	15.29	16.1
:ign	2.04	42.79	43.59	46.58	39.22	45.16	43.96
@longWords	1.85	412.24	402.98	429.33	416.61	429.07	426.21
@groupsPerSen	1.83	6.97	7.14	7.36	7.47	7.91	8.2
panie	1.81	15.66	11.06	11.12	9.01	8.6	7
#:adj_interp	1.78	43	44.63	45.74	44.97	48.84	50.36

## Istotne zmienne – rok urodzenia

---

Dużo cech trudnych do interpretacji. Poza tym:

- u starszych częstsze odwołania do *emerytury* i *emerytów*,
- u młodszych cechy pochodne od nazw partii PO i PiS, posiadających średnio nieco młodszych członków,
- inne słowa: *naprawdę*, *wniosek* (młodzi); *sprawa*, *zagadnienie* (starsi).

# Istotne zmienne – rok urodzenia

Feature	VI	Weighted average of age
\$sprawiedliwość	3687.38	45.68
:ign	2551.71	51.23
#\$być_być	1535.96	75.86
#:prep_ign	1342.25	51.3
\$emeryt	1313.81	63.47
#,_które	1148.63	49.86
\$emerytura	1062.29	60.07
#:fin:sg:pri:imperf_interp	939.72	51.1
#:adv_comp	890.08	46.07
sprawą	830.81	60.88
\$który	796.65	50.14
:fin:sg:pri:imperf	704.13	50.86
#:adv:pos_qub	698.66	48.58
które	634.39	49.89
#:ign_interp	607.03	51.19
\$sejm	585.92	50.64
#:ign_interp	568.03	51.19
emerytów	537.03	66.31
sprawiedliwości	523.88	45.16
\$naprawę	518.18	45.14
:pcon:imperf	515.56	51.09
tej	504.28	50.16

# Wnioski

---

1. Dla każdego zadania wyniki klasyfikacji są lepsze niż poziom odniesienia.
2. Wszystkie rodzaje zmiennych (leksykalne oraz statystyczne) są przydatne do predykcji.
3. Zwykle duża liczba cech jest potrzebna aby osiągnąć lepsze wyniki predykcji.
4. Zazwyczaj najlepszą metodą klasyfikacji są lasy losowe.



# Literatura

---

- Ogrodniczuk M., (2012). The Polish Sejm Corpus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012.
- Argamon, S., Koppel, M., Pennebaker, J. W., Schler, J. (2007). Mining the Blogosphere: Age, gender and the varieties of self-expression. First Monday. <http://ojphi.org/ojs/index.php/fm/article/view/2003/1878>
- Dahllöf, M. (2012). Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability. *Literary and Linguistic Computing*, 27(2), 139–153. doi:10.1093/llc/fqs010
- Koppel, M. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4), 401–412. doi:10.1093/llc/17.4.401
- Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., Menczer, F. (2011). Predicting the Political Alignment of Twitter Users. In 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing (pp. 192–199). IEEE. <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6113114>

# Pytania

---

?

Przybyła, P, Teisseyre, P. (2014) *Analysing Utterances in Polish Parliament to Predict Speaker's Background*. Journal of Quantitative Linguistics.