HATE SPEECH CLASSIFICATION

PAWEŁ FLIS, KAROL KRUPA, RAFAŁ CHABASIŃSKI

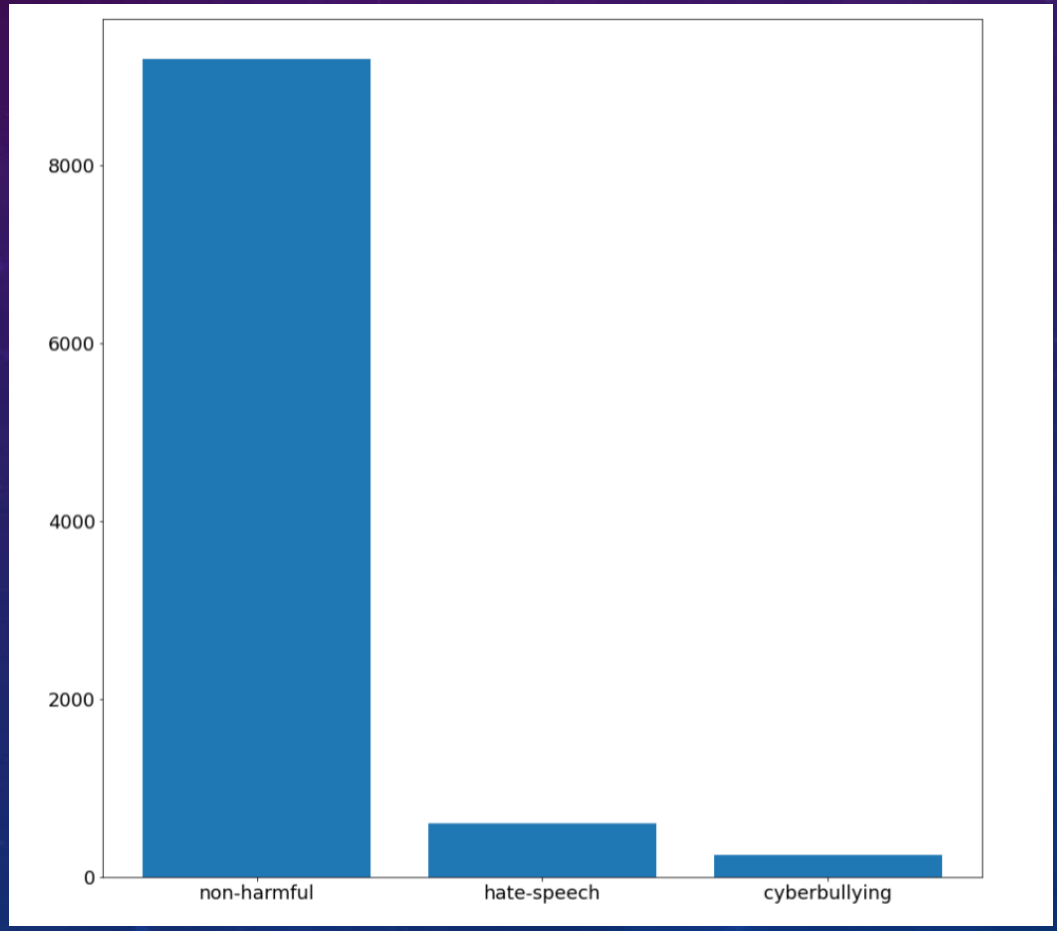
TASK, DATA

```
#Woronicza 17 poseł Halicki oburzony za Bolka.Naprawdè taki tępy czy tylko udaje idiotę?
„Ta aktorka ma 20 lat?!?! Jaka stara!!” \n\nChyba musimy się już do grobu pakować roczniku 98 😂👌
RT @anonymized_account „Ta aktorka ma 20 lat?!?! Jaka stara!!” \n\nChyba musimy się już do grobu pakować roczniku 98 😂👌
Ahnherr der Schtwätzer wykonawcy Von Spar\nhttps://t.co/SOtenSqIr0
@anonymized_account @anonymized_account @anonymized_account Bierze cie cie pod chuj a ty sie produkujesz
@anonymized_account @anonymized_account @anonymized_account Jak narazie to masz przywidzenia co nie zmienia faktu że cały czas jesteś idiotą.
@anonymized_account Kiedy do licznika dojdą bilety z fan clubow?
@anonymized_account A kto prowadzi zespół ? Będzie podany skład z tego meczu ?
@anonymized_account Główny powód to brak kasy, trzeba dać bogatym 500+,300+ i być bez godności i honoru
Zrobiłam takie Cv że ohohohoho
Świętować uchwalenie Konstytucji 3 maja i łamać Konstytucję RP obecnie obowiązującą?!\n#3Maja - dzień hipokryzji.
RT @anonymized_account Świętować uchwalenie Konstytucji 3 maja i łamać Konstytucję RP obecnie obowiązującą?!\n#3Maja - dzień hipokryzji.
```

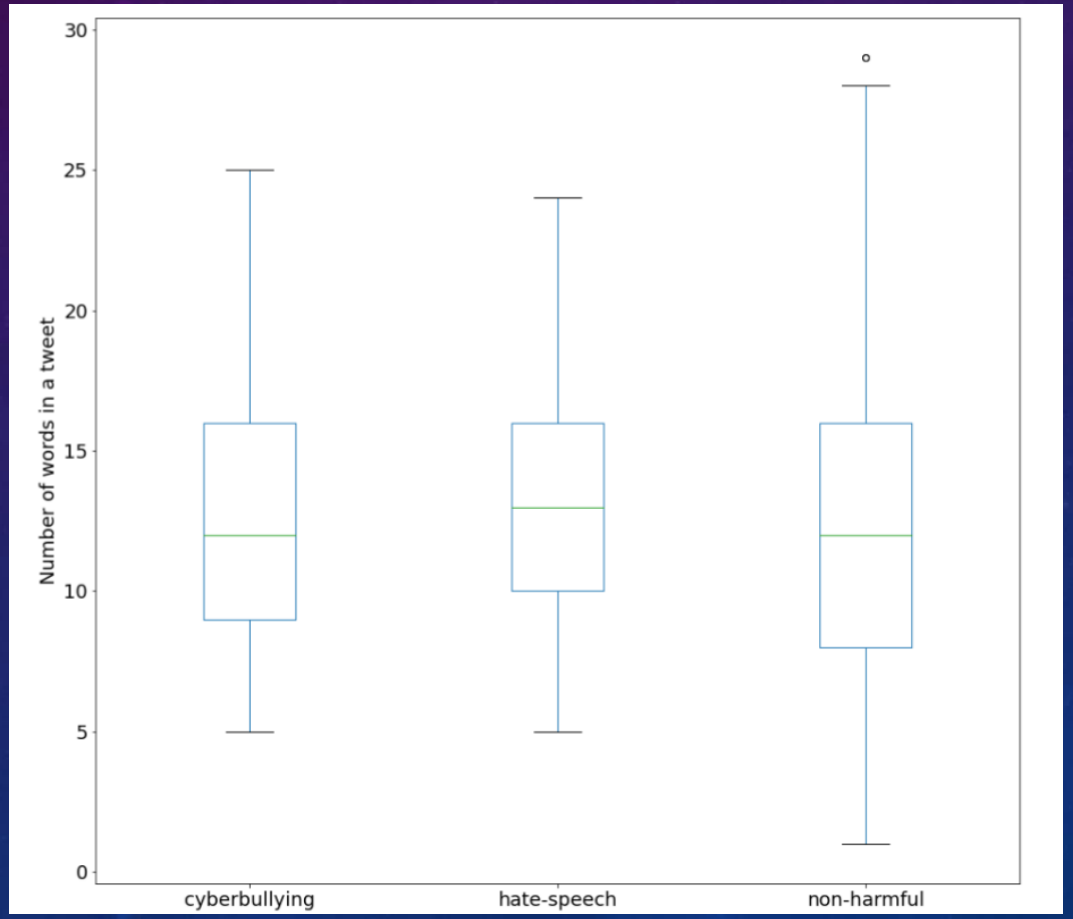
Poleval, Task 6, example test data

Goal: classify the tweets into cyberbullying, hate-speech and non -harmful

DATA DISTRIBUTION BY CLASSES



WORDS COUNTS FOR DIFFERENT CLASSES

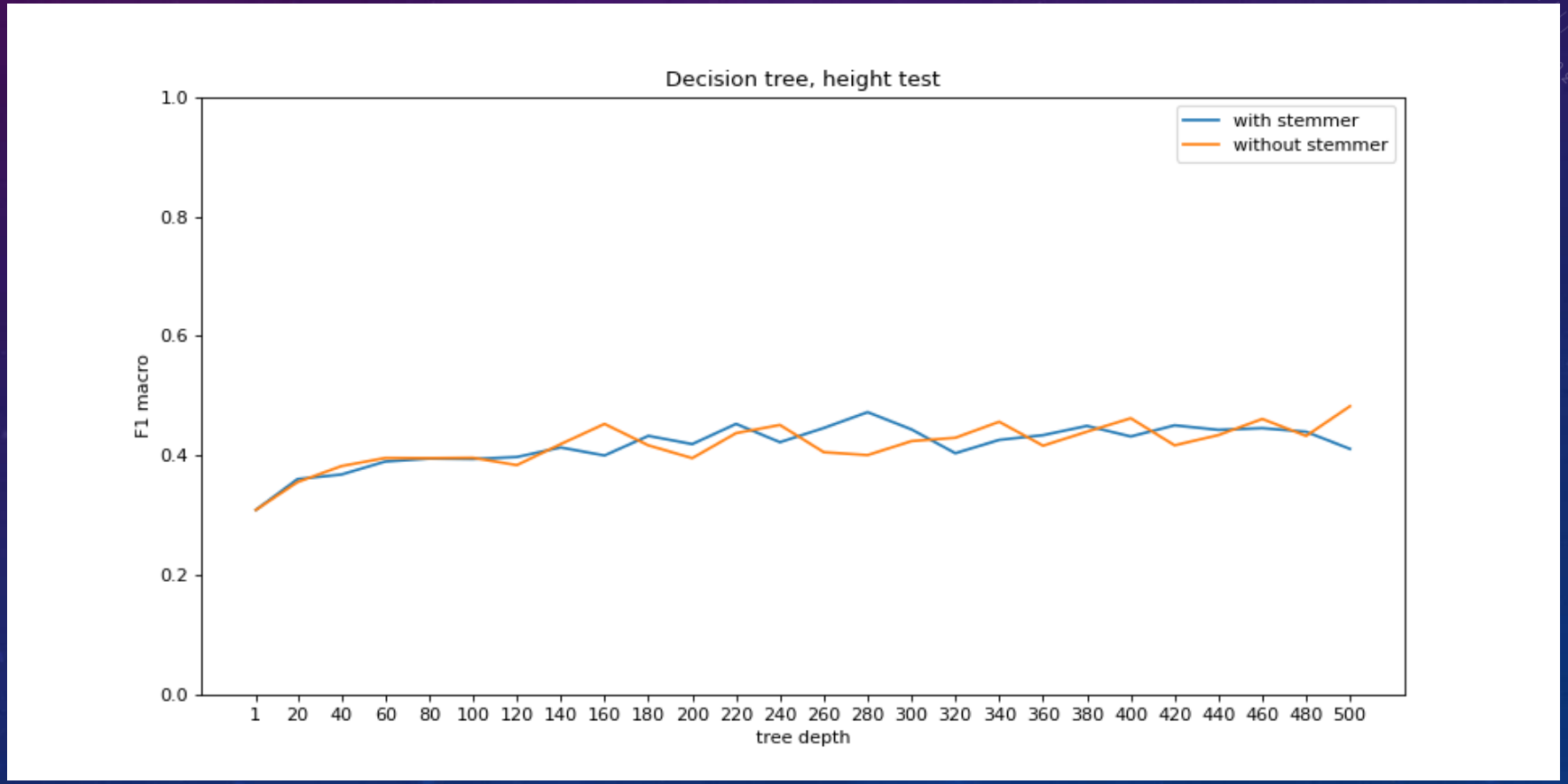


WORD CLOUD

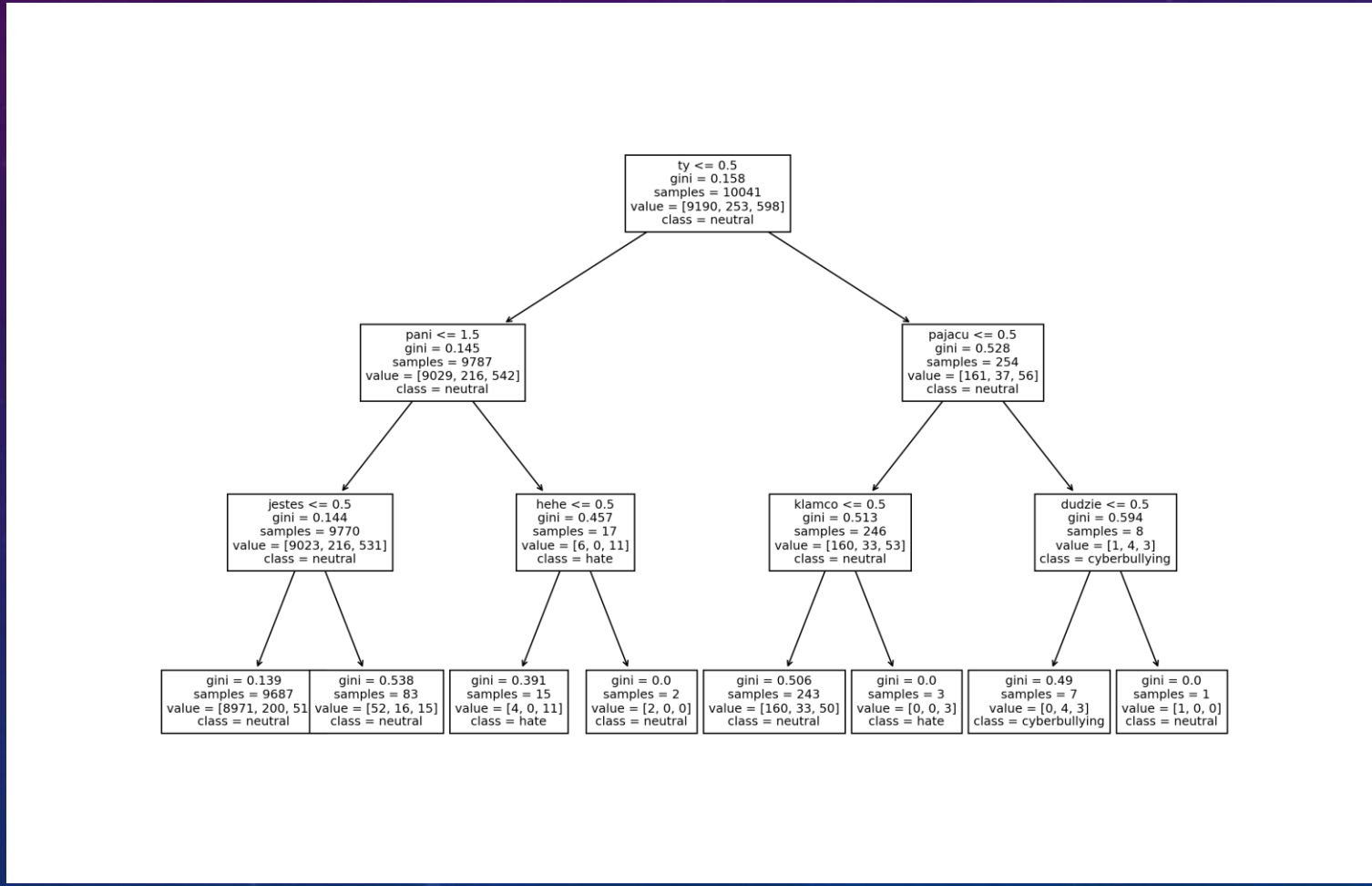

wiem
nrt
chyba
xd
nhttps
tco
jesteś
trochę
masz
kurwa
pewnie
mieć
czas
chodzi
mecz
meczu
siebie
dalej
nikt
dzieci
wisła
nie
mówi
jak
końcu
kiedyś
tyle
jakieś
ludzie
mamy
widzę
robi
lepiej
chce
skoro
jutro
mogę
dobra
problem
raczej
polski
pewno
wczoraj
takich
wisły
samo
proszę
pi
raz
tt
n
p
w
piątek
trener
powiedział
czego
życie
źle
chociaż
grać
dawno
kibiców
ważne
ludzi
muszę
szybko
ok
bym
da
będziesz
prezes
niby
jasne
zawodników
mieli
dobre
na
rok
swoich
wie
wiadomo
brawo
dzień
myślę
iść
wg
co
tyś
wiec
po
kogo
polska
sumie
legia
idzie
czym
ja
wiśle
akurat
n nhttps
widać
mln
brak
wygląda
oczywiście
woroniczа17
potem
zaraz
chciał
pracy
zrobił
info
ale
prawda
koniec
dopiero
takim
chcesz
tvn
fajnie
dwóch
kontrakt
czy
ty
macie
super
zrobić
ej
kim
klubu
czasu
wiesz
się
ogóle
miałam
takiego
dni
znaczy
jaka
robić
niema
jednym
zamiast
przynajmniej
razem
spokojnie
dzięki
pytanie
sędzia
pieniądze
panie
polacy
temu
pisu
razie
serio
tv
chcę
legii
lidze
domu
dobry
var
tak
jest
grał
chuj
ciekawe
szkoda
kibice
tez
tą
rozumiem
miała
tacy
polsce
kasy
innego
mecze
człowiek
drugi
np
temat
oby
hehe
później
niestety
myślałem
prostu
powiedzieć
sezonie
możesz
prawo
tydzień
ostatnio
wczoraj
przynajmniej
razem
zamiast
spokojnie
dzięki
pytanie
sędzia
pieniądze
panie
polacy
temu
pisu
razie
chcę
legii
lidze
domu
dobry
var
tak
jest
grał
chuj
ciekawe
szkoda
kibice
tez
tą
rozumiem
miała
tacy
polsce
kasy
innego
mecze
człowiek
drugi
np
temat
oby
hehe
później
niestety
myślałem
prostu
powiedzieć
sezonie

DECISION TREE BUILDING

No.	Stemmed data	Best tree height	F1 macro	F1 micro
1	true	280	0.473	0.868
2	false	500	0.483	0.87



EXAMPLE DECISION TREE

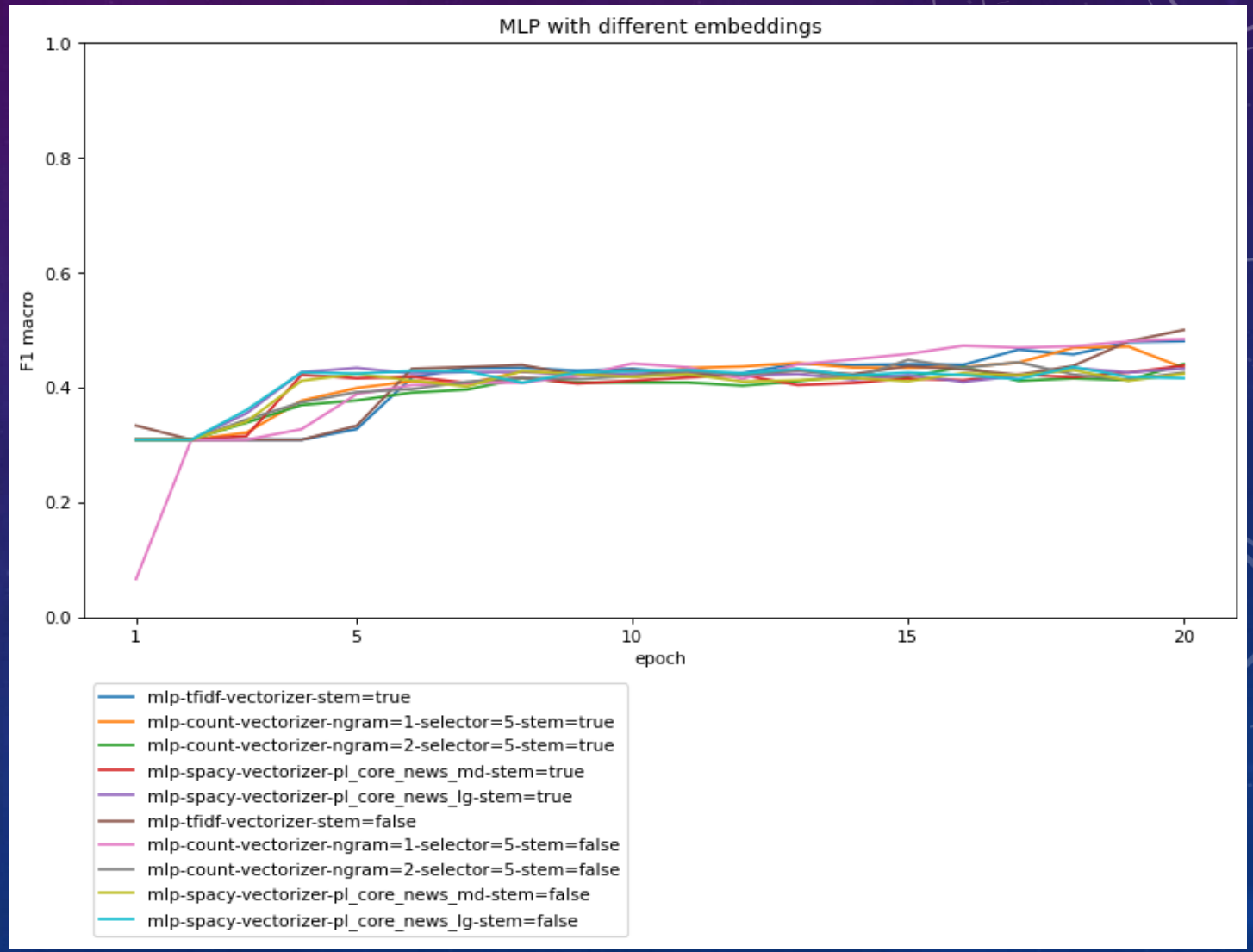


SVM AND NAIVE BAYES

No.	model	ngram	selector_percentile	F1 macro (S)	F1 micro (S)	F1 macro	F1 micro
1	SVM RBF	1	5	0.365	0.873	0.350	0.871
2	SVM RBF	1	10	0.367	0.864	0.369	0.861
3	SVM RBF	2	5	0.399	0.869	0.403	0.871
4	SVM RBF	2	10	0.351	0.864	0.356	0.866
5	SVM Poly	1	5	0.398	0.875	0.366	0.873
6	SVM Poly	1	10	0.322	0.867	0.321	0.867
7	SVM Poly	2	5	0.339	0.87	0.339	0.869
8	SVM Poly	2	10	0.322	0.868	0.322	0.868
9	**SVM Linear**	**1**	**5**	**0.409**	**0.877**	**0.375**	**0.874**
10	SVM Linear	1	10	0.379	0.871	0.357	0.868
11	SVM Linear	2	5	0.389	0.872	0.389	0.872
12	SVM Linear	2	10	0.353	0.869	0.365	0.871
13	Naive Bayes	1	5	0.322	0.868	0.322	0.868
14	Naive Bayes	1	10	0.344	0.87	0.349	0.87
15	Naive Bayes	2	5	0.322	0.868	0.322	0.868
16	Naive Bayes	2	10	0.364	0.87	0.364	0.87

MULTILAYER PERCEPTRON

Linear layer Input x 200
Dropout 0.2
ReLU
Linear layer 200 x 50
Dropout 0.2
ReLU
Linear layer 50 x 3
Softmax



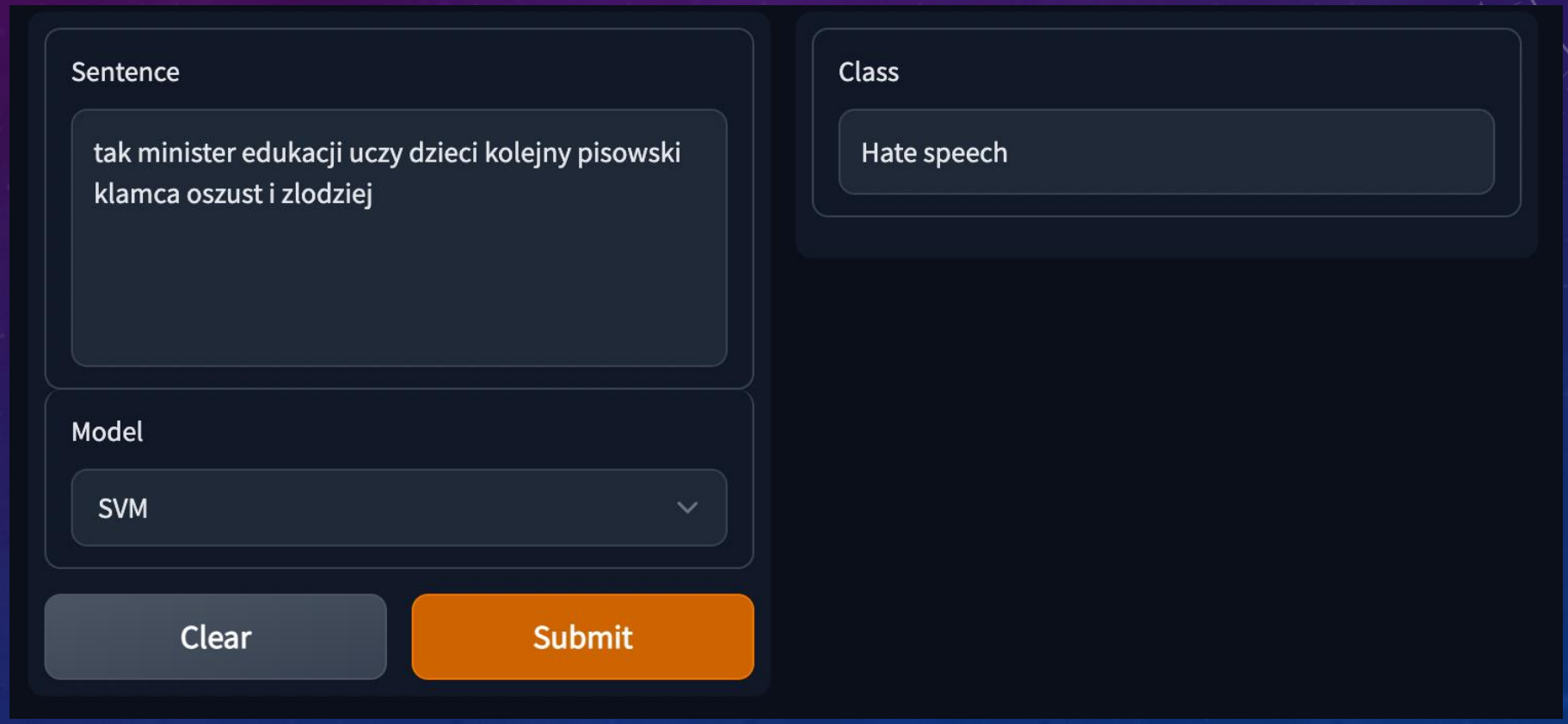
MULTILAYER PERCEPTRON RESULTS

No.	Vectorizer	F1 macro (S)	F1 micro (S)	F1 macro	F1 micro
1	CountVectorizer, ngram=1, percentile=5	0.472	0.874	0.485	0.876
2	CountVectorizer, ngram=2, percentile=5	0.441	0.874	0.427	0.877
3	**Tfidf**	**0.481**	**0.815**	**0.500**	**0.828**
4	Spacy, pl_core_news_md	0.438	0.816	0.430	0.787
5	Spacy, pl_core_news_lg	0.435	0.800	0.436	0.803

DEEP LEARNING

	micro-average F-score	macro-average F-score
base-075-5-3	88.10	57.24
large-v2-075-5-3	90.00	58.63
large-v2-1-1-1	90.70	51.48
large-v2-1-15-10	89.50	54.21

USER APPLICATION



THANK YOU!