Recognition of Emotions and Abusive Language in Texts : Machine Learning Models and Their Interpretability

Faculty of Mathematics and Information Science

WARSAW UNIVERSITY OF TECHNOLOGY

MI DATA LAB

APPLICA.AI

Recognition of Emotions and Abusive Language in Texts - Machine Learning Models and Their Interpretability

Anna Wróblewska - Ass. Prof. @ WUT, Senior Data Scientist @ Applica.ai
Filip Graliński @ Applica.ai

Paweł Pollak, Michał Mierzyński, Łukasz Dragan - former MSc Students @ WUT
Dawid Lipiński - Linguist @ Applica.ai
& Team @ Applica.ai

Anna

- 2008 - PhD - image recognition, Elka WUT
- Mainly R&D
  - SYNAT, semantic modeling, WUT
  - Controlled Natural Language, Cognitum
  - Semantic service for COP - common operational picture, WAT & ABG
- Allegro, Senior Data Scientist, 4 years
- Applica.ai, deep text modeling
- MIS WUT, data science, www.datascience.edu.pl
- Postgraduate Data Science WUT
- Industrial PhD programme, cooperation with industry: Applica.ai,

Liliana
Pięta
Allegro

Prof. Sylwia
Sysko-Romańczuk
Business School WUT

APPLICA.AI

FINDWISE
Paweł Wróblewski

Sylwia Grodecka
ARETE
Relationships
development

CENTRUM NAUKI
KOPERNIK

MI DATA LAB

Prof. Przemek Biecek, MIS WUT

Agenda

1. Use Cases
2. Annotating data challenges
3. Machine learning models
4. Evaluation & interpretability

Business Use Cases

We need to understand the real meaning of texts - semantics AUTOMATICALLY

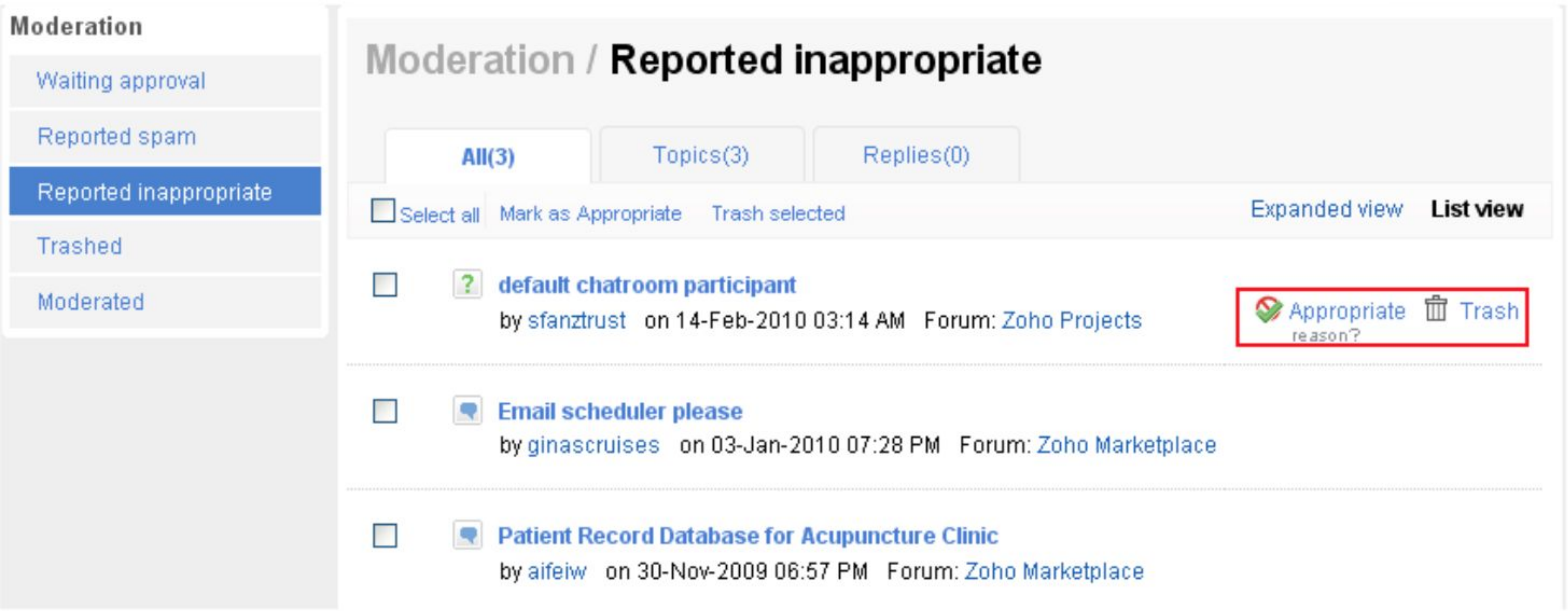
To better adjust content to the user or text/ads owners, to remove unwanted content, to understand intentions etc.

Use Cases

- Portal moderation: users comments
- RTB advertisements
- Behavioral personalisation
- Mood in social media, regarding any brand, event etc.
- Emotions, hate speech

APPLICA.AI

All new content is placed in 'in tray'
Moderators view in tray
Moderators mark content as 'read' OR they remove content



Moderation
Waiting approval
Reported spam
Reported inappropriate
Trashed
Moderated
Moderation / Reported inappropriate
All(3)
Topics(3)
Replies(0)
Select all
Mark as Appropriate
Trash selected
Expanded view
List view
default chatroom participant
by sfanztrust on 14-Feb-2010 03:14 AM Forum: Zoho Projects
Appropriate
Trash
reason?
Email scheduler please
by ginascruises on 03-Jan-2010 07:28 PM Forum: Zoho Marketplace
Patient Record Database for Acupuncture Clinic
by aifeiw on 30-Nov-2009 06:57 PM Forum: Zoho Marketplace

Real Time Bidding / Programmatic Ads

Owner of ads: particular sites, e.g. with appropriate emotions, brand safe

Behavioral personalisation

- Usual mood, texts / sites that are read by a user
- Adapt communication language to be understood better

  or manipulate :-(

- Ethical issues !

APPLICA.AI

		High	Low
O	Openness to Experience	*Imaginative*	*Conventional*
C	Conscientiousness	*Organized*	*Spontaneous*
E	Extraversion	*Outgoing*	*Solitary*
A	Agreeableness	*Trusting*	*Competitive*
N	Neuroticism	*Prone to stress*	*Emotionally stable*

Ways to work on use cases

- Sentiment, emotions
- Offensive language / hate speech
- Named entities, semantic relations between entities
- Synonyms, similar concepts etc.

Text sentiment

- Sentiment – positive, neutral, negative
- Deep sentiment / Emotion recognition – Emotions expressed in texts

Text	Emotions expressed
Nie spodziewałem się, że to będzie takie fajne! I did not expect that it would be so cool!	Joy / Surprise

Emotion theories

Plutchik (1960-1980)

1. **Anticipation**
2. **Joy**
3. **Sadness**
4. **Fear**
5. **Trust**
6. **Disgust**
7. **Surprise**
8. **Anger**

Ekman

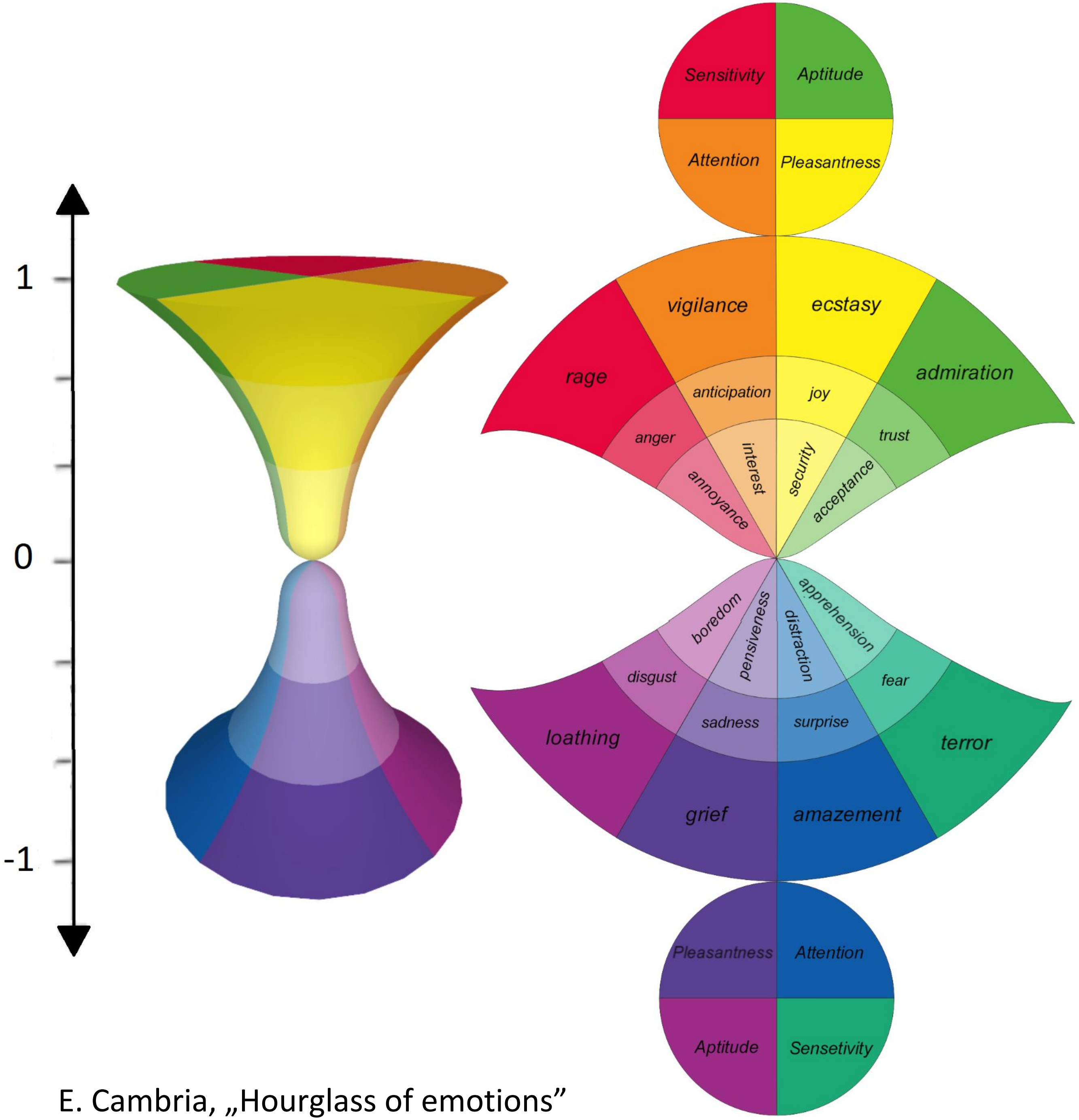
1. Joy
2. Sadness
3. Fear
4. Disgust (revulsion)
5. Anger
6. Surprise

Nakurama

+ Shame emotion

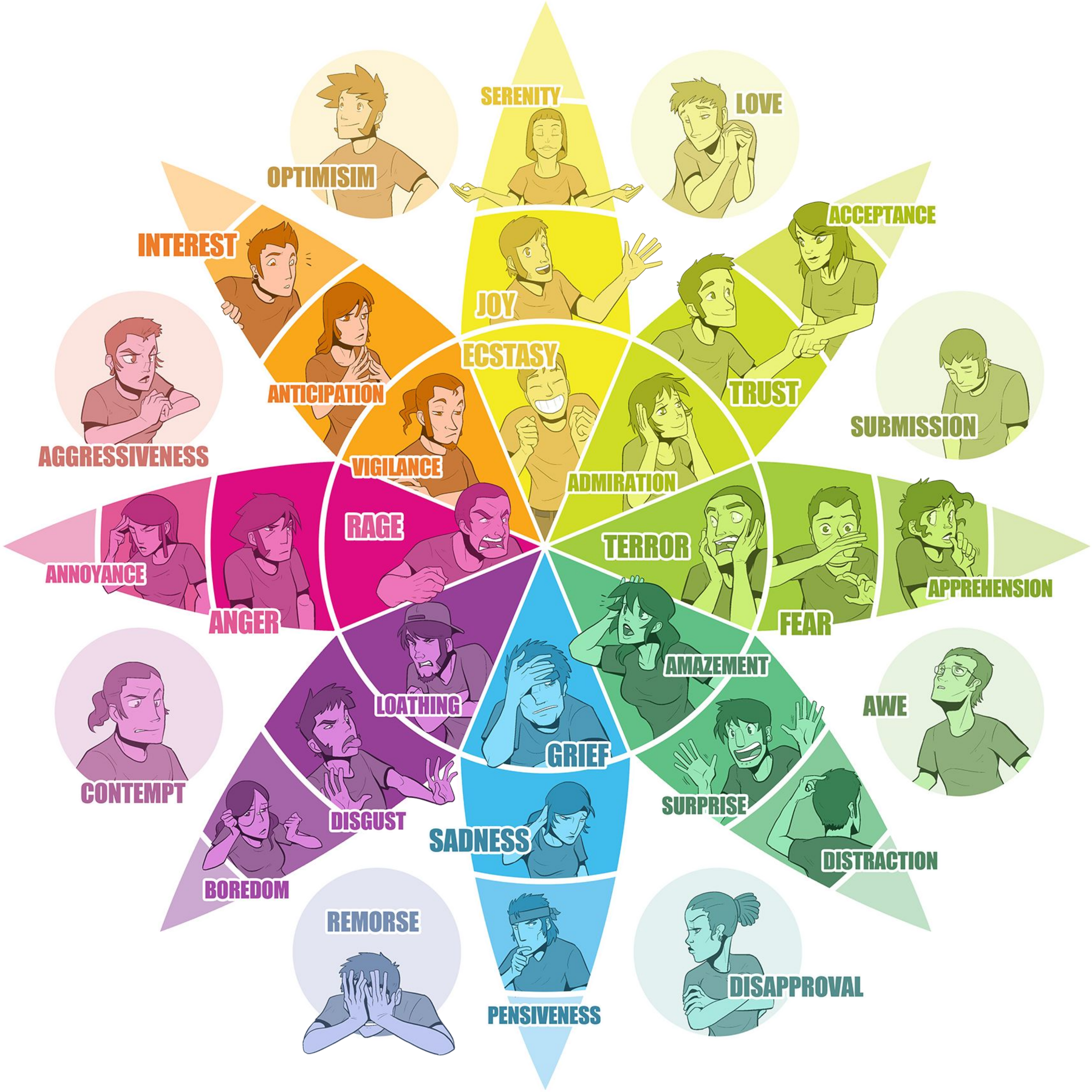
E. Cambria, ...: "The hourglass of emotions", LNCS, 2012


SADNESS
JOY
FEAR
DISGUST
ANGER
INSIDE OUT
INSIDE OUT
INSIDE OUT
INSIDE OUT
INSIDE OUT



E. Cambria, „Hourglass of emotions”

APPLICA.AI


SERENITY
LOVE
OPTIMISIM
ACCEPTANCE
INTEREST
JOY
AGGRESSIVENESS
ECSTASY
ANTICIPATION
TRUST
SUBMISSION
VIGILANCE
ADMIRATION
RAGE
TERROR
ANNOYANCE
APPREHENSION
ANGER
FEAR
AMAZEMENT
LOATHING
AWE
CONTEMPT
GRIEF
SURPRISE
DISGUST
SADNESS
DISTRACTION
BOREDOM
REMORSE
DISAPPROVAL
PENSIVENESS

APPLICA.AI

Abusive language

	Explicit	*Implicit*
Directed	"Go kill yourself", "You're a sad little f*ck" (Van Hee et al., 2015a), "@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga" (Davidson et al., 2017), "Youre one of the ugliest b*tches Ive ever fucking seen" (Kontostathis et al., 2013).	"Hey Brendan, you look gorgeous today. What beauty salon did you visit?" (Dinakar et al., 2012), "(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles" (Hine et al., 2017), "you're intelligence is so breathtaking!!!!!!" (Dinakar et al., 2011)
Generalized	"I am surprised they reported on this crap who cares about another dead n*gger?", "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!" (Nobata et al., 2016), "So an 11 year old n*gger girl killed herself over my tweets? ^_^ thats another n*gger off the streets!!" (Kwok and Wang, 2013).	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." (Burnap and Williams, 2015), "most of them come north and are good at just mowing lawns" (Dinakar et al., 2011), "Gas the skypes" (Magu et al., 2017)

Z. Wassem: Understanding Abuse: A Typology of Abusive Language Detection Subtasks, Proc. of Workshop on Abusive Lang. 2017, ACL

Hate speech

Language that is used to **expresses hatred** towards a **targeted group** or is intended to be derogatory, to humiliate, or to insult the members of the group

- GENERALIZED EXPLICIT or IMPLICIT

Some linguistics

Text processing & modeling

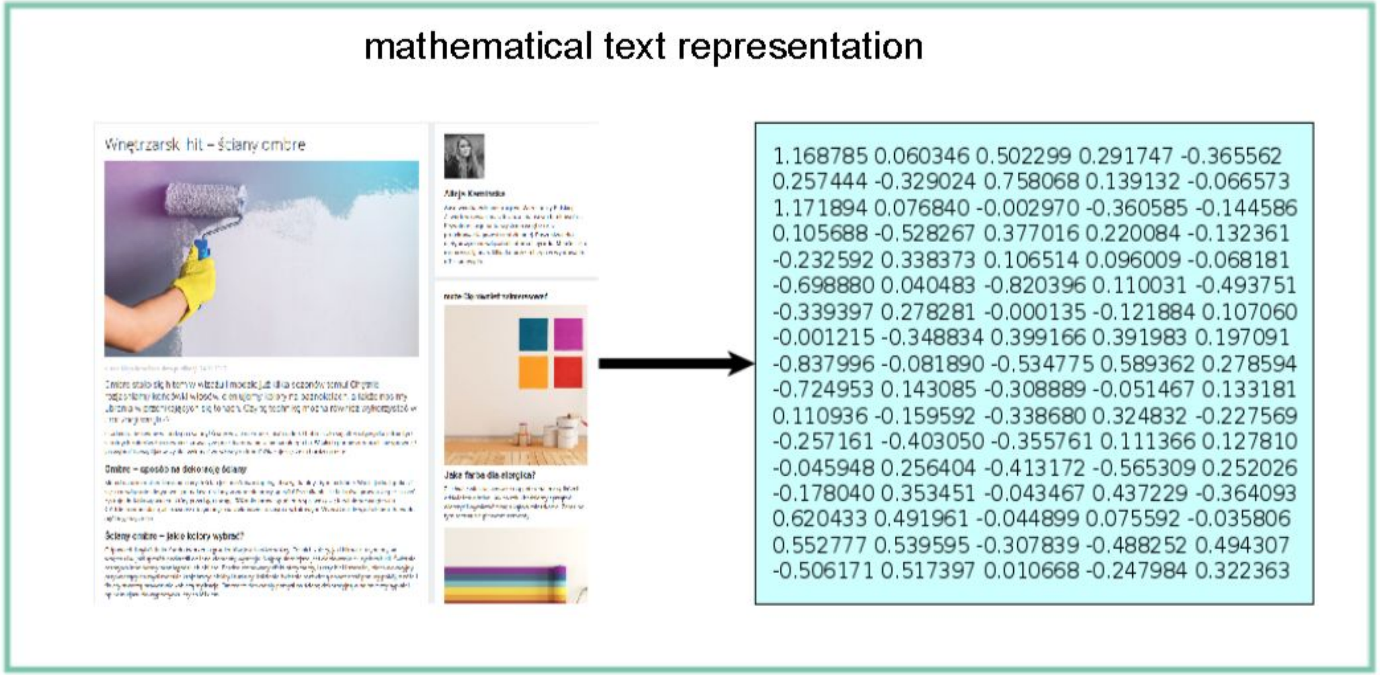
First to understand data structure

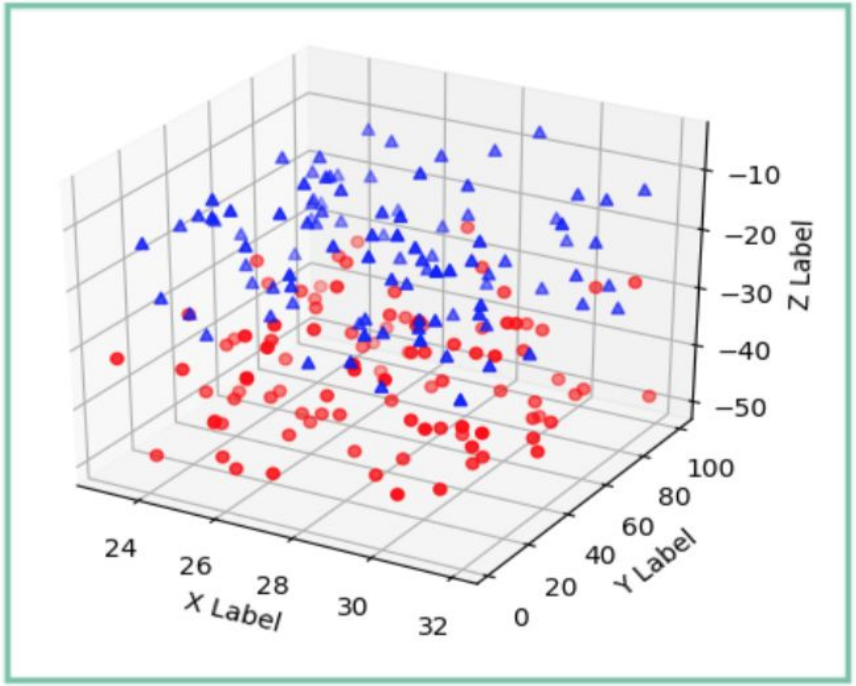
then meaning

Text processing pipeline

APPLICA.AI

Language modeling

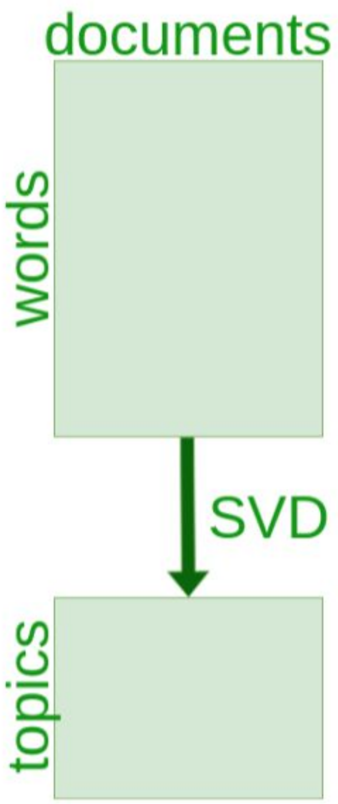




Language modeling

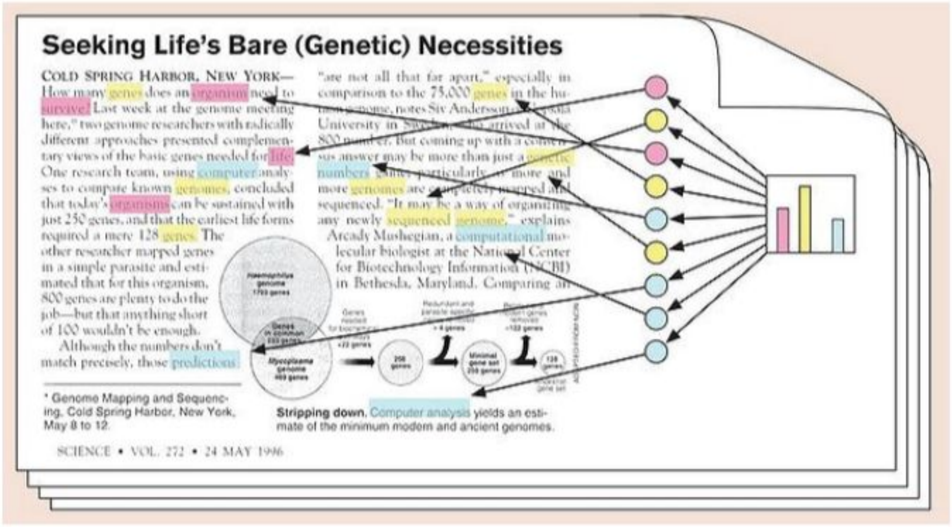
Latent semantic indexing

Groups words into topics using Singular Value Decomposition



Latent Dirichlet allocation

Represents documents as mixtures of topics that are probability distributions over words



https://medium.com/@connectwithghosh

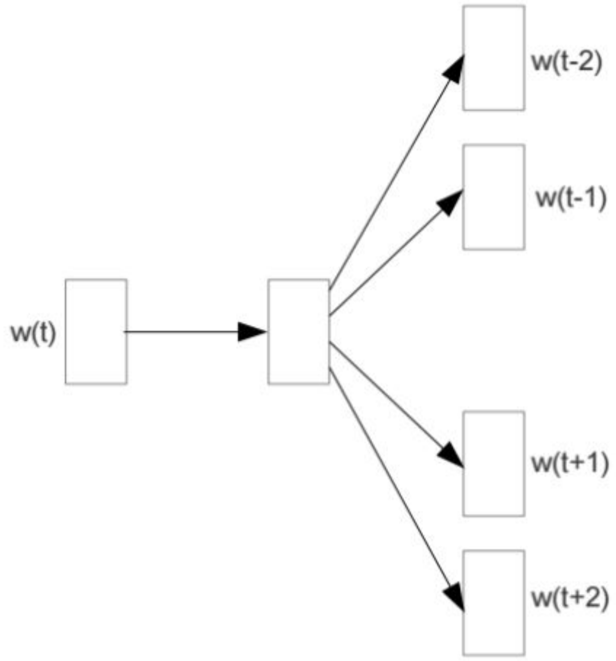
Language modeling

Distributional techniques main assumption:

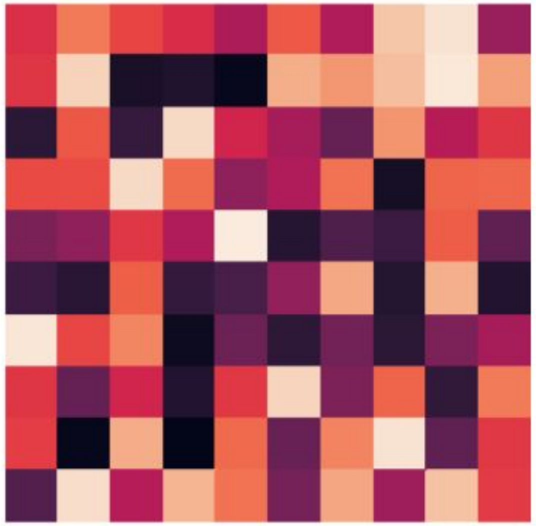
Any word is determined based on its context

Language modeling

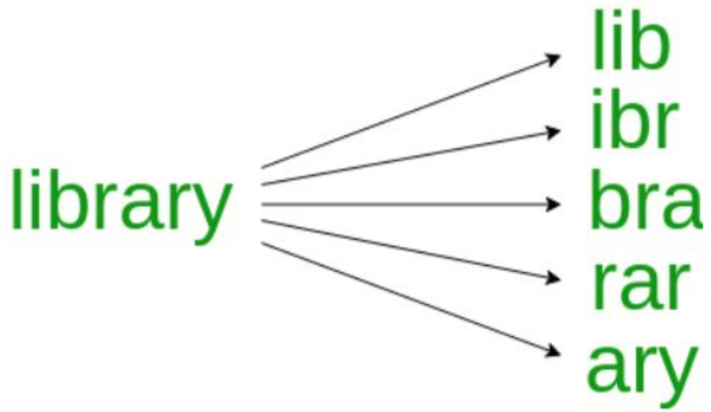
word2vec
Shallow feed-forward neural network

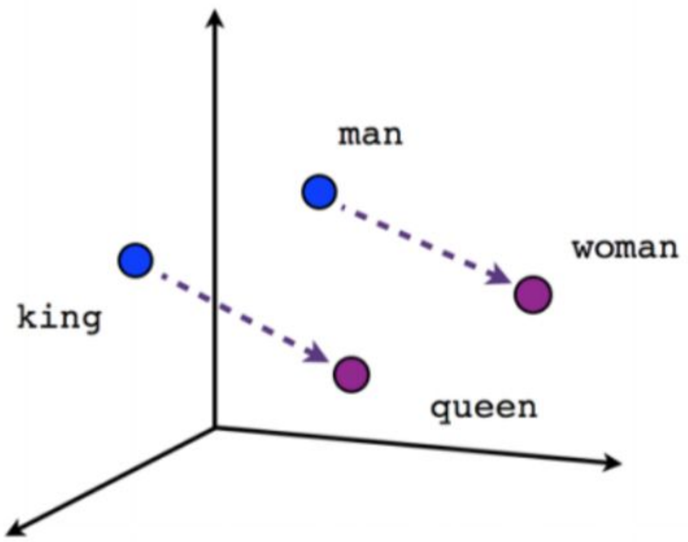


GloVe
Words co-occurrences matrix

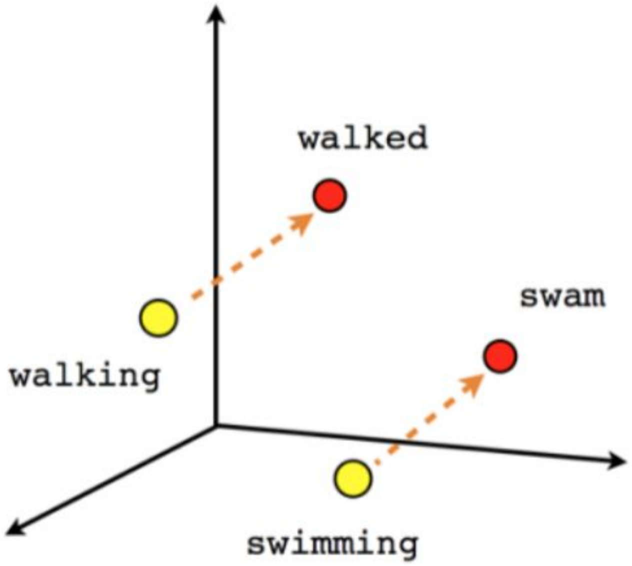


FastText
Words divided into character n-grams

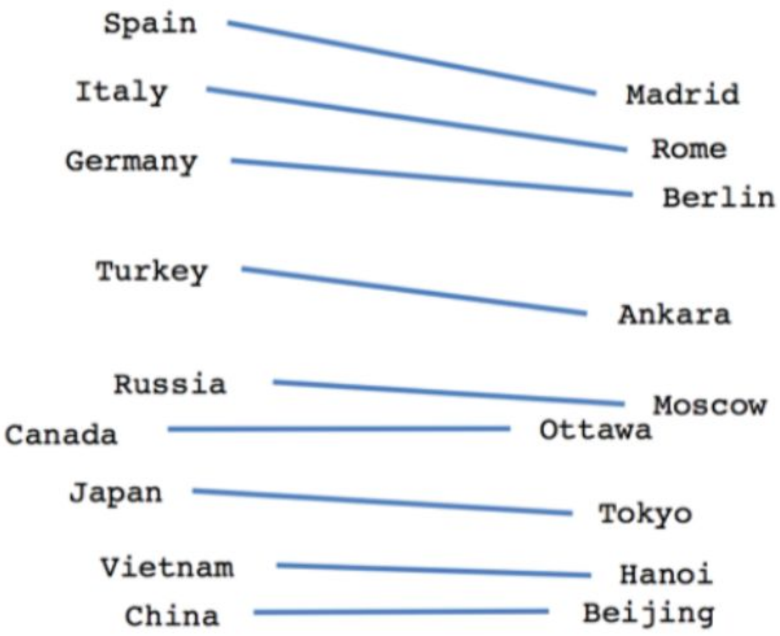




Male-Female



Verb tense

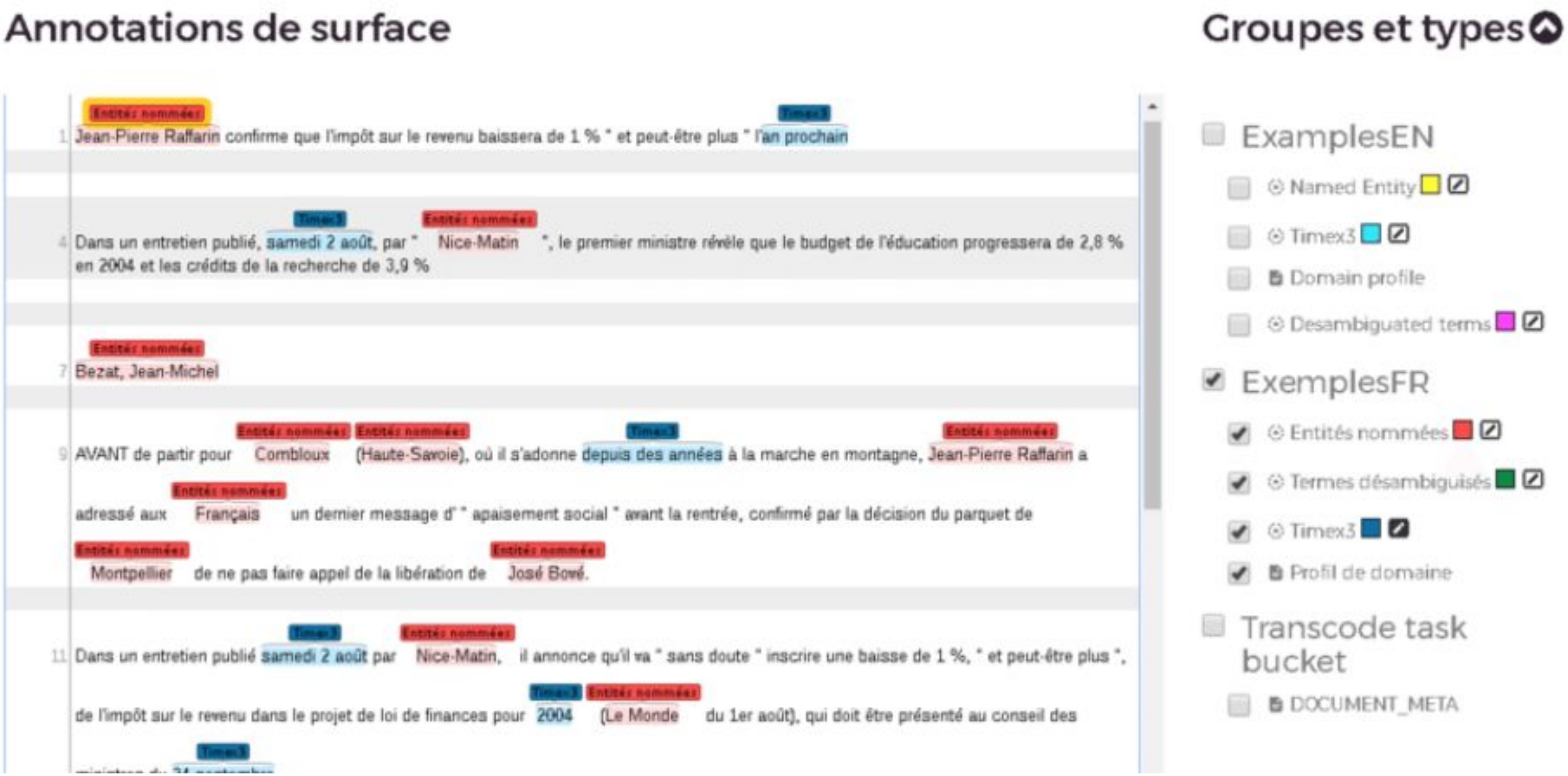


Country-Capital

Text annotation

- Even for linguists can be a very difficult task to differentiate between classes (emotions expressed etc.)
- Tedious, costly task
- Applica.ai has dedicated platform to annotation processes (organize works of linguists, get metrics of work, and annotated texts etc.)

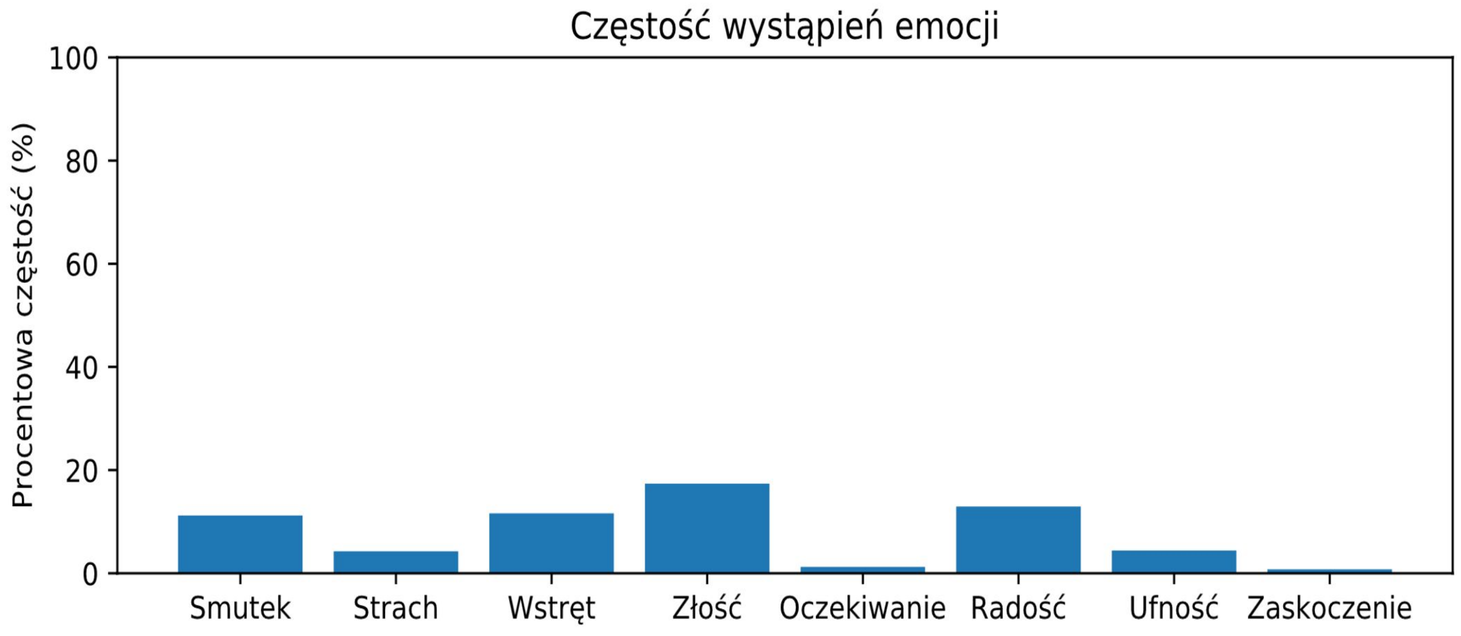
The challenge is to annotate less data and to get the best results



https://prodi.gy/demo?view_id=textcat

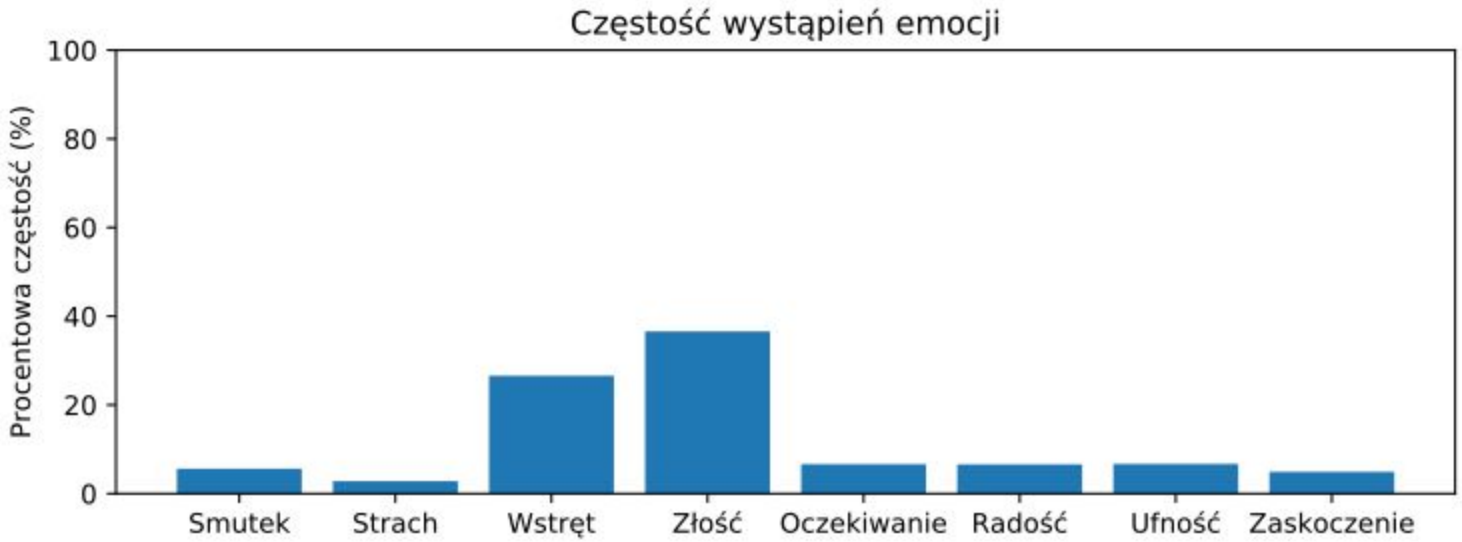
Source texts to annotate emotions

- Wordnet.pl by Wrocław University of Technology
  - Emotions assigned to lexical units, having exemplar sentences
  - + neutral sentences, without emotions
  - about 30 000 sentences



Source texts 2

- Comments and articles from information portals
- Annotation by linguists
- About 2 000 texts



Annotation for emotions interface

Tekst komentarza:

dieta musi być racjonalna wtedy jesteśmy zdrowe i ładnie wyglądamy, dieta to sposób żywienia. To bardzo wazne aby robić to z głową. Poczytajcie sobie książkę Odchudzanie z elementami fizjologii i biochemii. Tam są informacje , które diety są niezdrowe i dlaczego, jakie spustoszenie sieją w organizmie, ale przede wszystkim dowiesz się jak jeść i ćwiczyć, by zrzucać kilogramy i być zdrowym! ja mam po 8 miesiącach 14 kilo mniej :)

Emocja				
Smutek	0	1	2	3
Strach	0	1	2	3
Wstręt	0	1	2	3
Złość	0	1	2	3
Oczekiwanie	0	1	2	3
Radość	0	1	2	3
Ufność	0	1	2	3
Zaskoczenie	0	1	2	3

Text processing in the framework

Jestem dumny z adekwatności twojego zachowania do tej sytuacji.

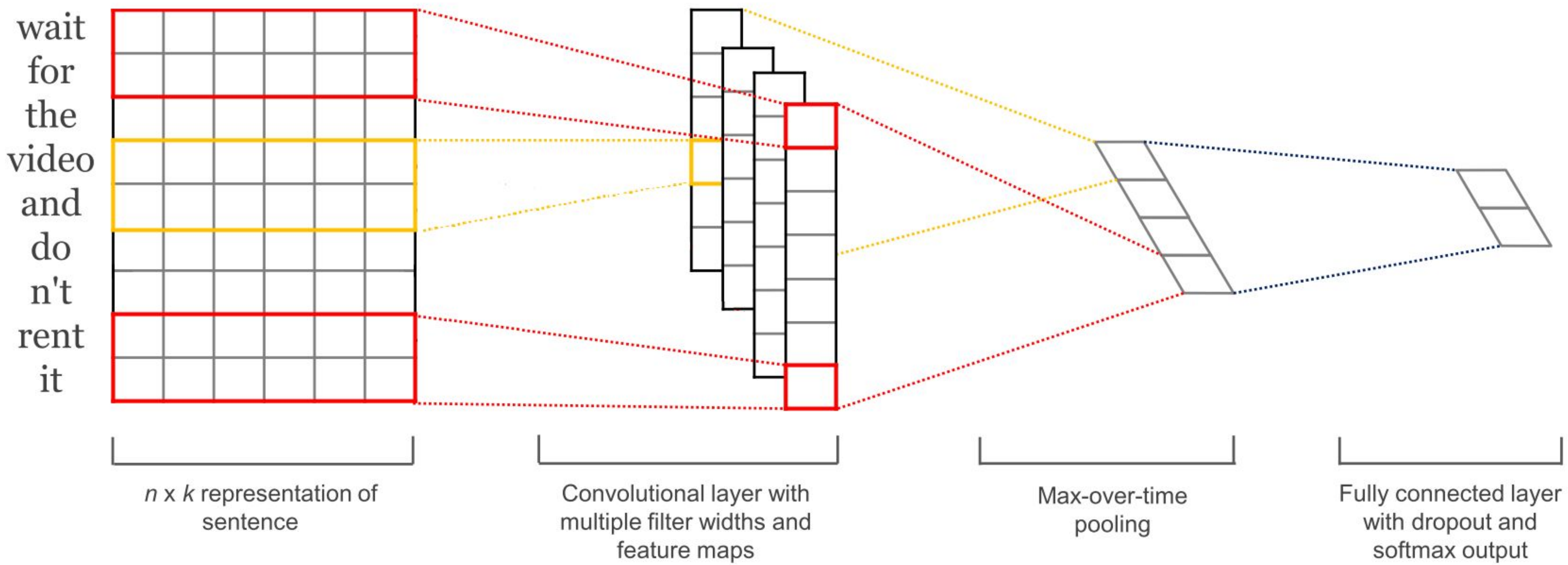
I am proud of the adequacy of your behavior to this situation.

↓

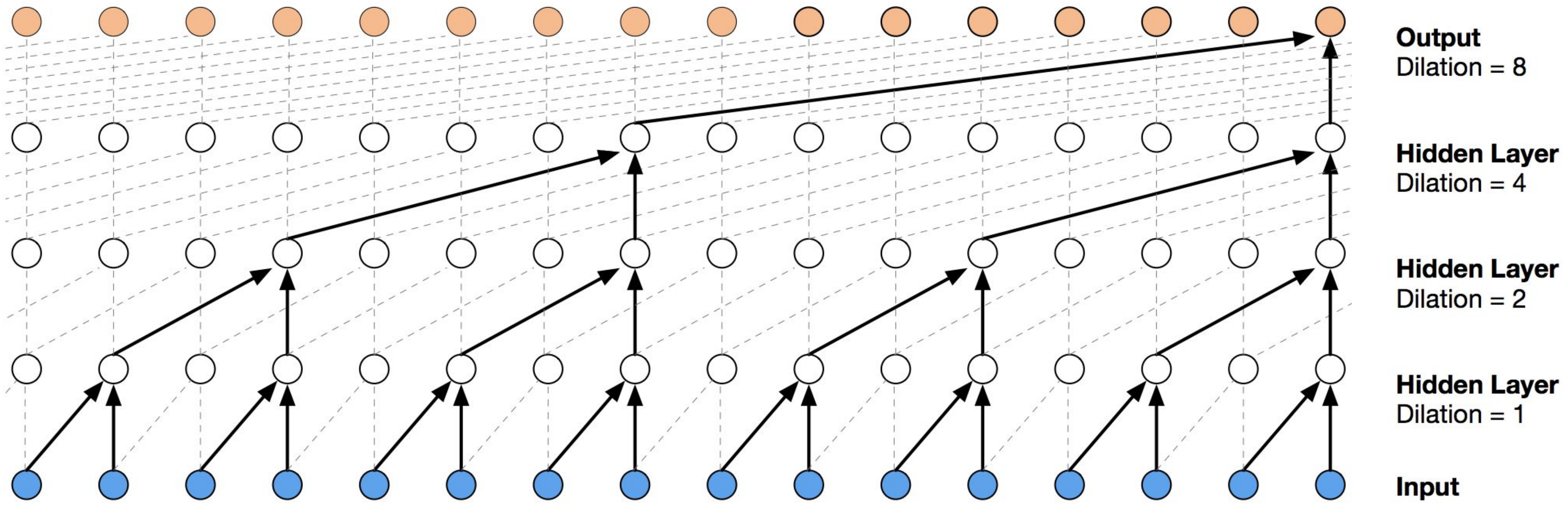
["być", "dumny", "z", "adekwatność", "twój", "zachowanie", "do", "ten", "sytuacja", "."]

LEMMAS - base forms

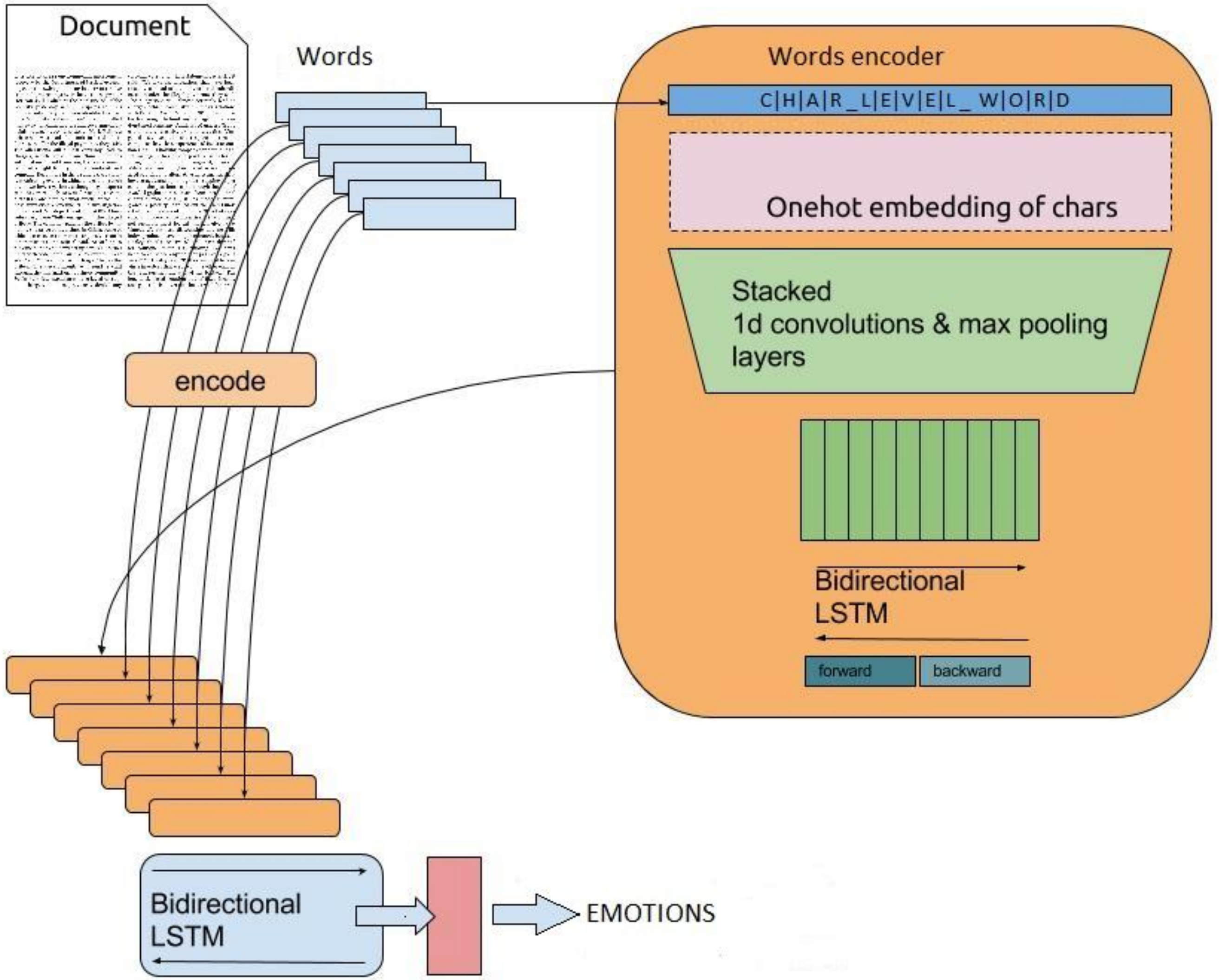
CNN - Convolutional Neural Network



TCN - Temporal Convolutional Network



LSTM

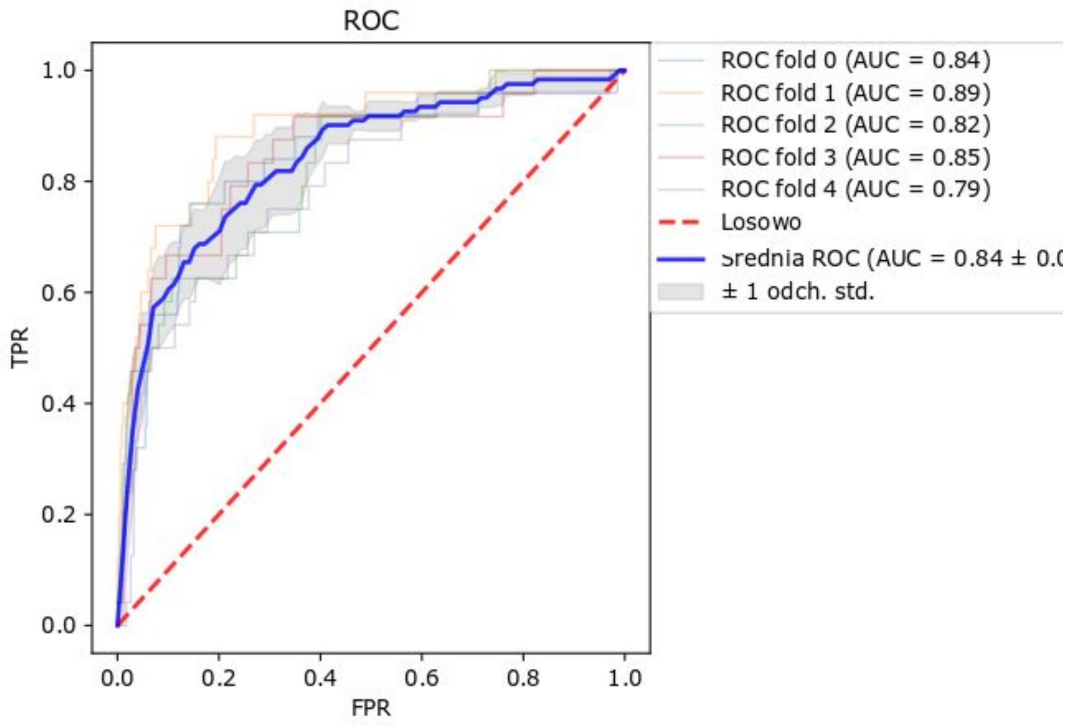

Document
Words
Words encoder
C|H|A|R_L|E|V|E|L_W|O|R|D
Onehot embedding of chars
Stacked
1d convolutions & max pooling
layers
Bidirectional
LSTM
forward
backward
encode
Bidirectional
LSTM
EMOTIONS

Evaluation measures

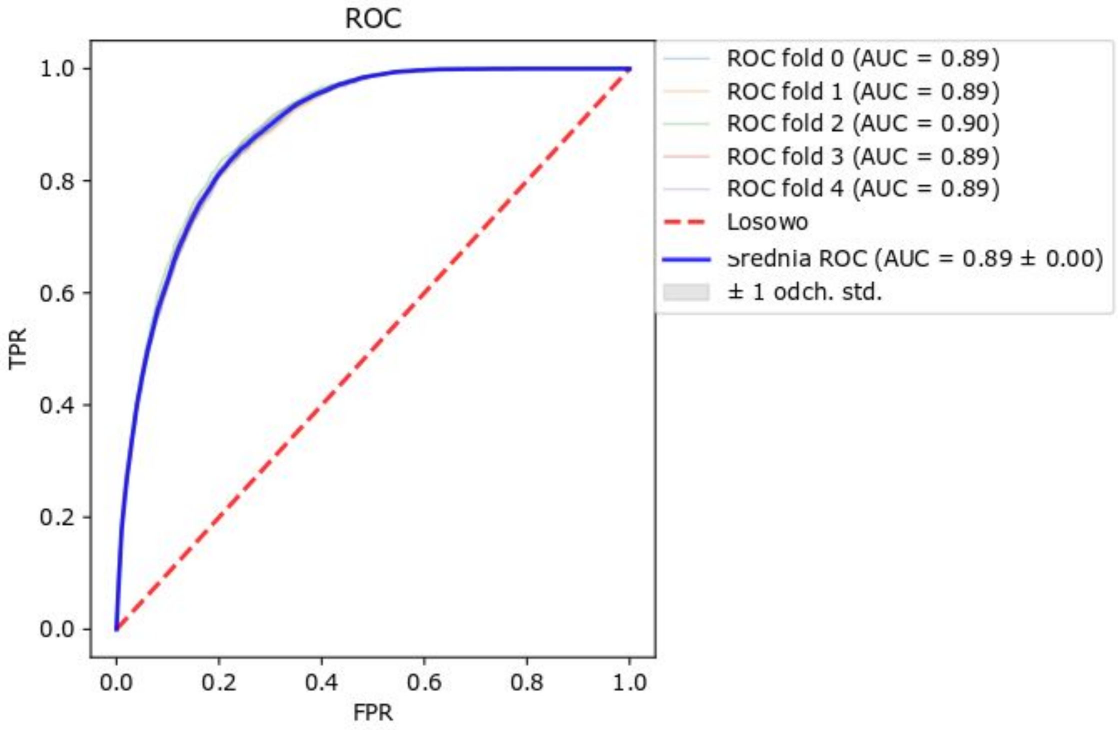
- AUC/ROC
  - Not dependent on threshold for decision making
- F1 - precision / recall
  - Depends on balance on test dataset
- Kappa Cohena
  - Agreement of 2 observers
  - But good model has not the same value of the metric
  - Random classifier has 0-value metric
- Precision@10%
  - Depends on the balance on test set

ROC/AUC - Joy

Internet comments

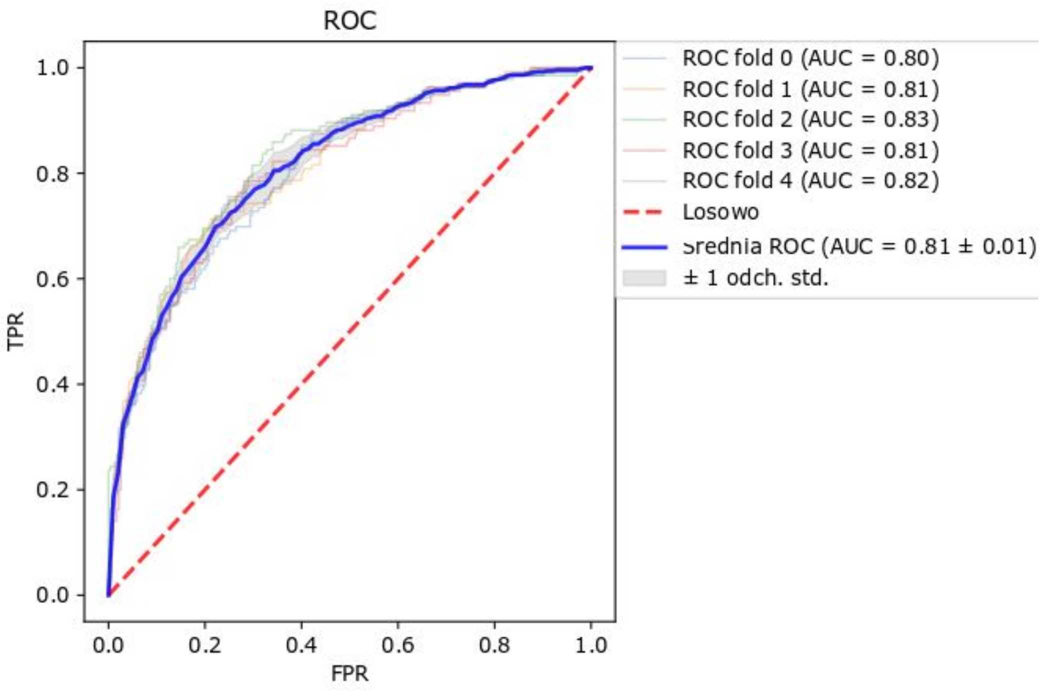


WordNet.pl

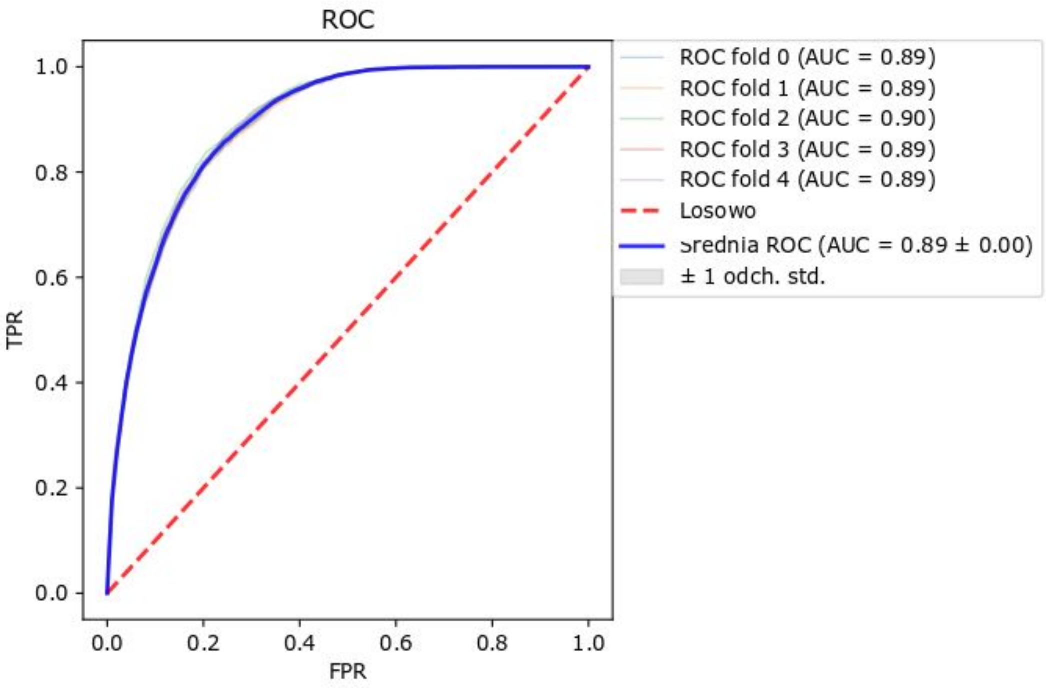


ROC/AUC - Anger

Internet comments



WordNet.pl



Best models for emotions

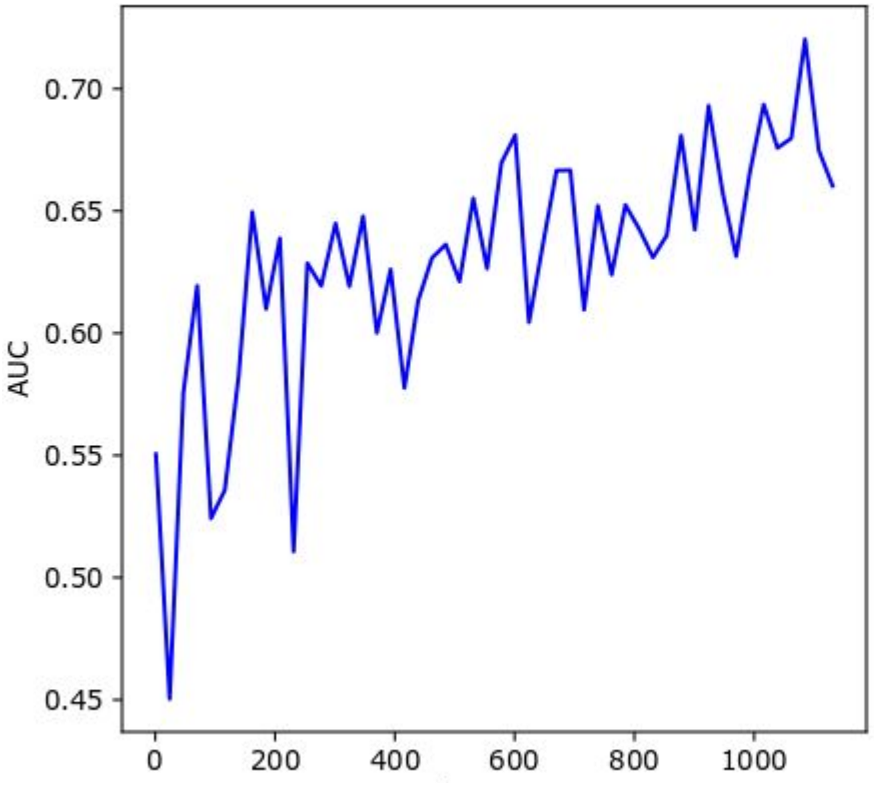
- CNN
- Models based on chars need bigger datasets

AUC
[std. dev. on folds]

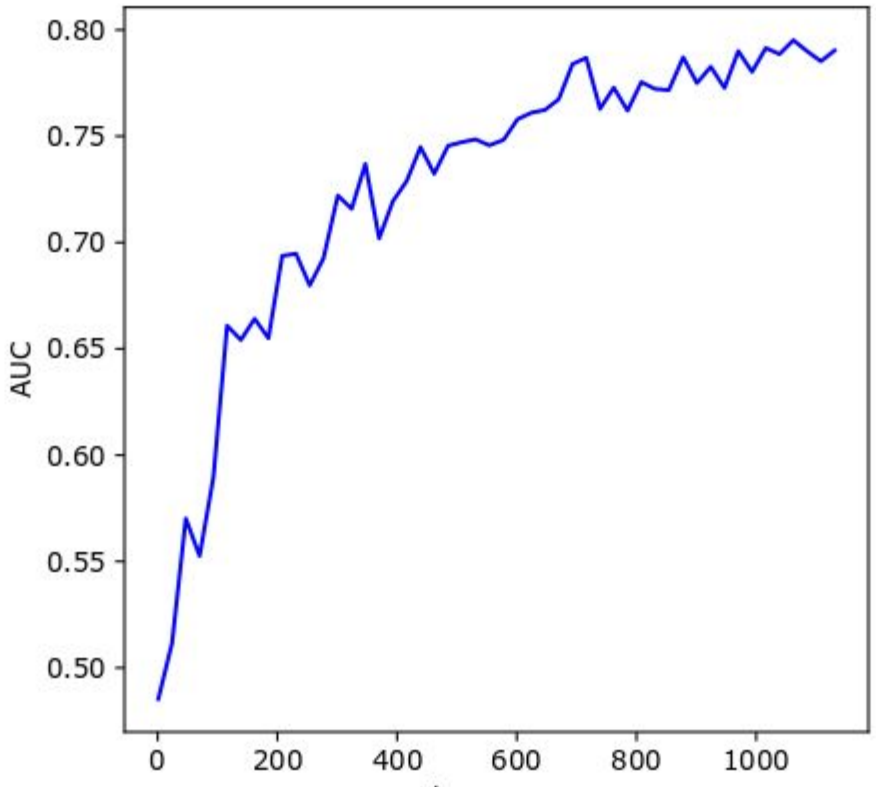
Emotion	Dataset	
	Comments	WordNet.pl
Sadness	0.657 [0.054]	0.840 [0.002]
Fear	0.791 [0.069]	0.846 [0.006]
Disgust	0.820 [0.042]	0.877 [0.005]
Anger	0.813 [0.009]	0.891 [0.003]
Anticipation	0.604 [0.058]	0.813 [0.007]
Joy	0.839 [0.034]	0.891 [0.003]
Trust	0.778 [0.043]	0.852 [0.005]
Surprise	0.695 [0.031]	0.803 [0.021]
Negative	0.830 [0.025]	0.948 [0.004]
Positive	0.822 [0.035]	0.903 [0.005]

Quality of classification results depends on size of training dataset - comments dataset

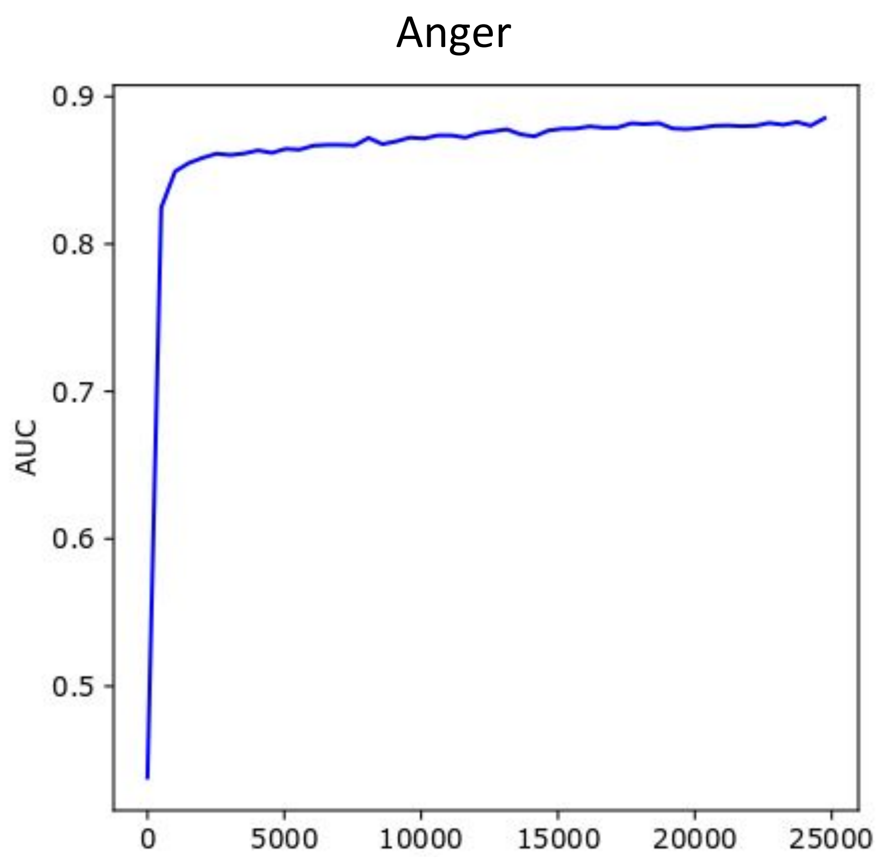
Sadness



Anger



Quality of classification results depends on size of training dataset - WordNet.pl dataset



APPLICA.AI

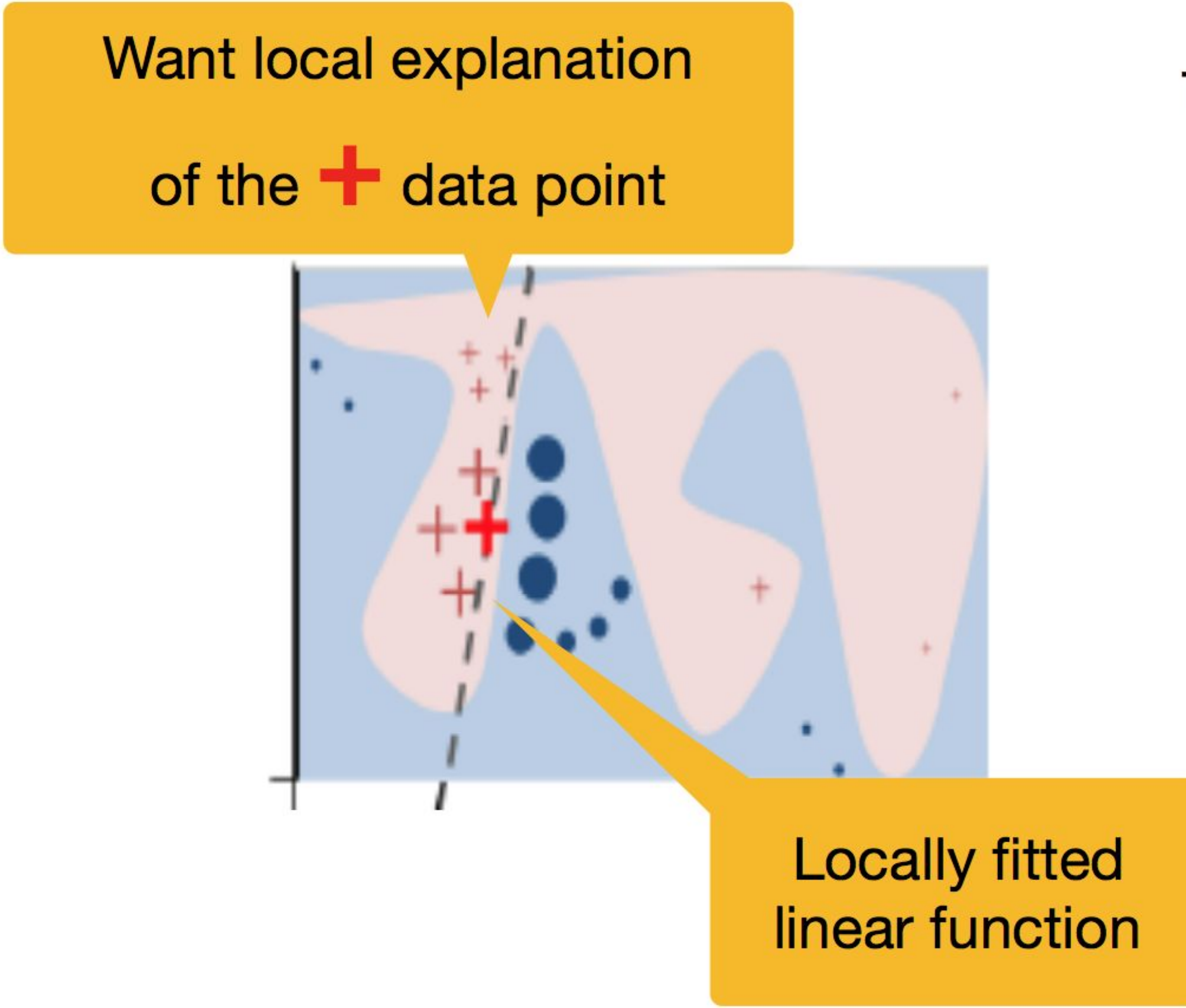
MI

Interpretability

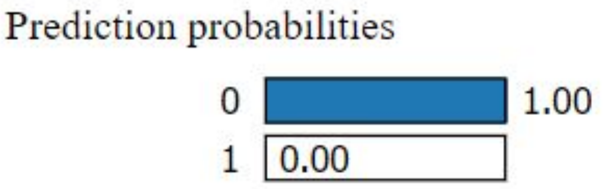
LIME

Ribeiro '2016

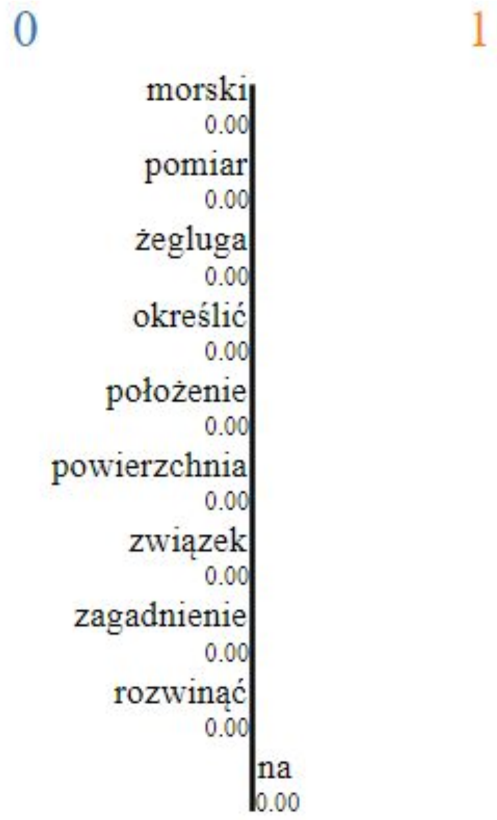
- Local Interpretable Model-agnostic Explanations



Interpretability of models



Model: Is any emotion expressed in the text or not?

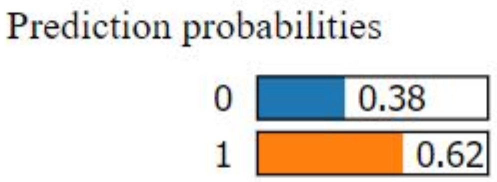


Text with highlighted words

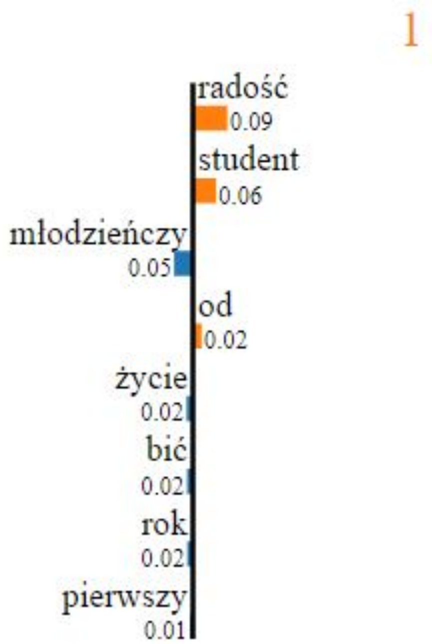
trygonometria powstać i rozwinąć się głównie w związek z zagadnienie pomiar na powierzchnia Ziemia oraz potrzeba żegluga morski (określić położenie i kierunek przy pomoc ciało niebieski) .

Trigonometry develops due to measurements on the surface of the earth….

Interpretability of models



Model: Is joy emotion expressed in the text or not?

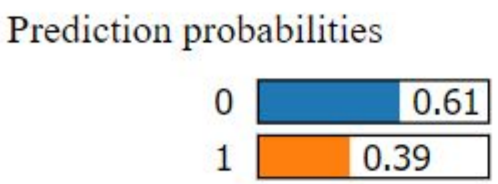


Text with highlighted words

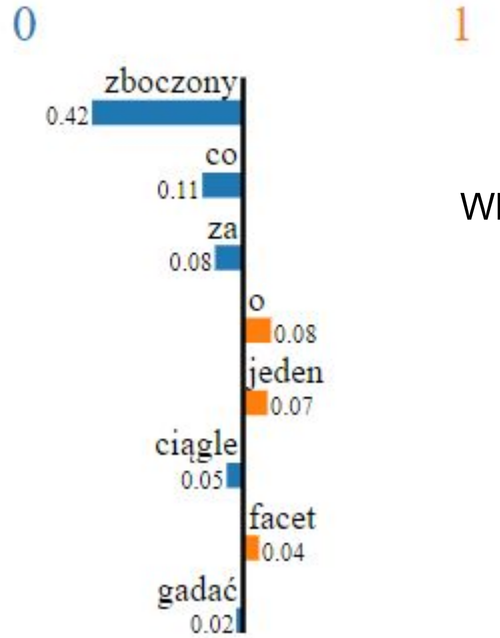
od student pierwszy rok bić młodzieńczy radość życie .

From students in the first year beats the youthful joy of life.

Interpretability of models



Trust

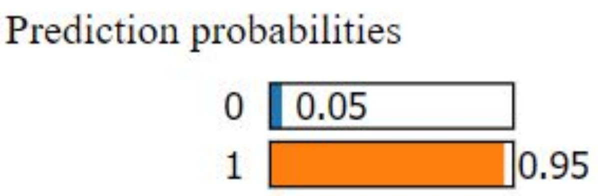


Text with highlighted words

co za zboczony facet , ciągle gadać o jeden .

What a perverted guy who is constantly talking about the only one.

Interpretability of models



Anger

Text with highlighted words

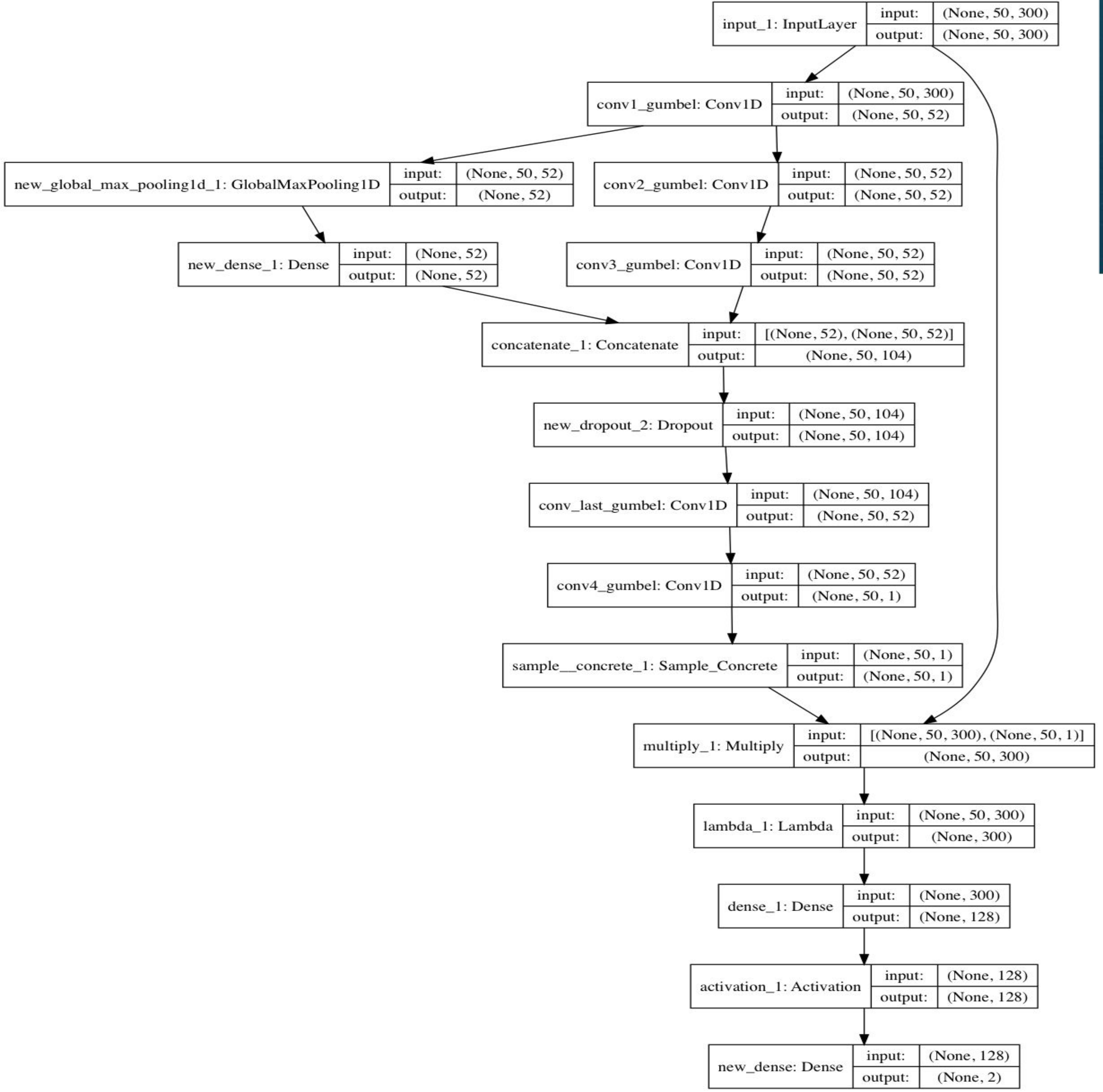
bać się ten polakożerczy fanatyk .

I'm afraid of this fanatic who has Poles for the prey.

L2X

Chen Jianbo, ...: An Information-Theoretic Perspective on Model Interpretation, ICML 2018

20% improvement in measures on our datasets



L2X for trust model - TP

Cieszę się z obrotności mojej córki.

I am happy for my daughter's agility.

Dobry nauczyciel powinien być sprawiedliwy !

A good teacher should be just!

Abusive Language - English datasets

Twitter1

- About 25k tweets
- Classes: offensive language 76% / non-offensive 19% / hate speech 5%
- *”Automated Hate Speech Detection and the Problem of Offensive Language”*, ICWSM 2017

Twitter 2

- About 13K tweets
- Classes: racist 12% / sexist 20% / non-offensive 68%
- *”Deep Learning for Hate Speech Detection in Tweets”*WWW 2017

Models

- LSTM
- CNN
- FastText classifier
- XGBoost

- Inputs are language model vectors - embeddings

Twitter1

APPLICA.AI

SEMANTYCZNA ANALIZA TEKSTU

Metoda	Precision	Recall	F1	Kappa Cohena	AUC
ICML 2018	0.910	0.900	0.900	-	-
LSTM + rand + GBDT	**0.942 [0.003]**	**0.942 [0.003]**	**0.942 [0.003]**	**0.838 [0.008]**	**0.971 [0.002]**
LSTM + Glove + GBDT	0.913 [0.005]	0.913 [0.005]	0.913 [0.005]	0.746 [0.014]	0.949 [0.005]
CNN + rand + GBDT	0.940 [0.003]	0.940 [0.003]	0.940 [0.003]	0.832 [0.009]	0.972 [0.003]
CNN + Glove + GBDT	0.907 [0.005]	0.907 [0.005]	0.907 [0.005]	0.727 [0.016]	0.941 [0.005]
fastText + rand + GBDT	0.919 [0.004]	0.919 [0.004]	0.919 [0.004]	0.772 [0.010]	0.956 [0.004]

Unbalanced datasets

- Reduction of the biggest class
- Scaling the smaller class
- Leave as in the distribution of domain
- Merging negative classes

Reducing the biggest class

Zbiór	Precision	Recall	F1	Kappa Cohena	AUC
twitter1 full	**0.911 [0.006]**	**0.898 [0.007]**	**0.904 [0.006]**	0.728 [0.020]	0.930 [0.010]
twitter1	0.887 [0.007]	0.869 [0.007]	0.878 [0.007]	**0.752 [0.015]**	**0.931 [0.010]**
twitter2 full	**0.832 [0.007]**	**0.822 [0.008]**	**0.827 [0.007]**	0.627 [0.018]	0.914 [0.006]
twitter2	0.811 [0.014]	0.794 [0.013]	0.803 [0.013]	**0.674 [0.022]**	**0.924 [0.007]**
polishData full	**0.822 [0.002]**	**0.820 [0.002]**	**0.821 [0.002]**	0.490 [0.009]	0.883 [0.003]
polishData	0.767 [0.004]	0.762 [0.005]	0.764 [0.004]	**0.567 [0.008]**	**0.890 [0.003]**

Weighting the smaller classes

APPLICA.AI

SEMANTYCZNA ANALIZA TEKSTU

Zbiór	Precision	Recall	F1	Kappa Cohena	AUC
twitter1 full	**0.911 [0.006]**	**0.898 [0.007]**	**0.904 [0.006]**	**0.728 [0.020]**	**0.930 [0.010]**
twitter1	0.891 [0.005]	0.889 [0.005]	0.890 [0.005]	0.648 [0.023]	0.923 [0.008]
twitter2 full	**0.832 [0.007]**	**0.822 [0.008]**	**0.827 [0.007]**	**0.627 [0.018]**	**0.914 [0.006]**
twitter2	0.804 [0.020]	0.802 [0.019]	0.803 [0.019]	0.522 [0.066]	0.911 [0.008]
polishData full	**0.822 [0.002]**	**0.820 [0.002]**	**0.821 [0.002]**	**0.490 [0.009]**	**0.883 [0.003]**
polishData	0.804 [0.001]	0.802 [0.002]	0.803 [0.002]	0.333 [0.008]	0.880 [0.002]

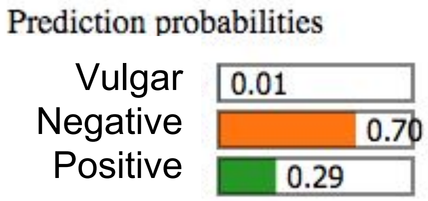
Merging classes

APPLICA.AI

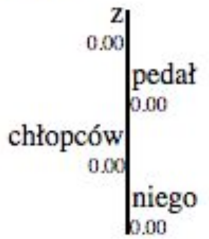
SEMANTYCZNA ANALIZA TEKSTU

Zbiór	Precision	Recall	F1	Kappa Cohena	AUC
twitter1 full	0.911 [0.006]	0.898 [0.007]	0.904 [0.006]	0.728 [0.020]	0.930 [0.010]
twitter1	**0.950 [0.005]**	**0.950 [0.005]**	**0.950 [0.005]**	**0.820 [0.016]**	**0.983 [0.002]**
twitter2 full	0.832 [0.007]	0.822 [0.008]	0.827 [0.007]	**0.627 [0.018]**	**0.914 [0.006]**
twitter2	**0.835 [0.007]**	**0.835 [0.007]**	**0.835 [0.007]**	0.617 [0.016]	0.889 [0.008]
polishData full	0.822 [0.002]	0.820 [0.002]	0.821 [0.002]	**0.490 [0.009]**	**0.883 [0.003]**
polishData	**0.824 [0.002]**	**0.824 [0.002]**	**0.824 [0.002]**	0.485 [0.007]	0.851 [0.002]

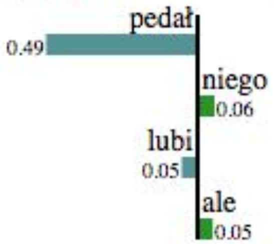
LIME - offensiveness depends on context



Not vulgar vulgar



Not positive positive

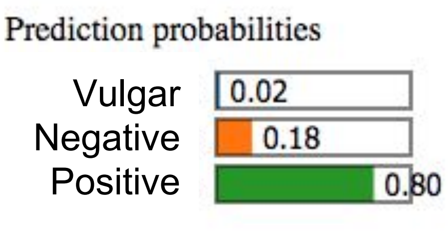


Text with highlighted words

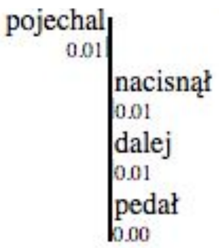
ale z niego pedał. Pewnie lubi chłopców

He is a queer! He probably likes boys.

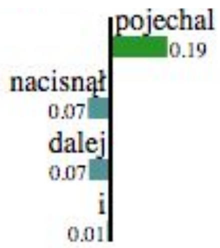
LIME - the same word but positive context



Not vulgar vulgar



Not positive positive



Text with highlighted words

nacisnął pedał i pojechal dalej

He pressed the pedal and drove away.

APPLICA.AI

MI

Interpretability - geval

GEVAL

Mann U-Whitney rank test

$$(X, Y, \hat{Y})^{+f} \quad (X, Y, \hat{Y})^{-f}$$

WORD	COUNT	+	−	ACC	χ^2	P-VALUE
THOUGH	343	254	89	0.7405	35.2501	0.00000
KNOW	767	619	148	0.8070	13.4284	0.00025
READING	72	57	15	0.7917	2.226	0.1357

Features

- Word
- Bigram
- Cartesian features

Source:

- Input file
- Output
- Expected - gold standard

APPLICA.AI

Twitter sentiment

Feature	Count	Acc	p-value
in<1>:though	343	0.74	0.00004
in<1>:no++idea	21	0.48	0.001
in<1>:count	16	0.44	0.002
in<1>:yeah	227	0.76	0.003
in<1>:know	767	0.81	0.004
in<1>:which	98	0.71	0.006
in<1>:what++the	23	0.56	0.007
in<1>:wouldn't	38	0.68	0.029
in<1>:Haven't	14	0.57	0.030
in<1>:can't++even	12	0.58	0.047
exp:1˜˜in<1>:sad	13	0.38	0.001
exp:1˜˜in<1>:though	72	0.67	0.002
exp:1˜˜in<1>:can't	160	0.73	0.002
exp:1˜˜in<1>:never	81	0.67	0.001
exp:1˜˜in<1>:miss	73	0.34	0.0000
exp:1˜˜in<1>:hate	43	0.35	0.0000
exp:1˜˜in<1>:but	549	0.73	0.0000
exp:1˜˜in<1>:not	395	0.71	0.0000
exp:1˜˜in<1>:no	196	0.64	0.0000
exp:1˜˜in<1>:wish	66	0.50	0.0000
exp:1˜˜in<1>:i	1067	0.77	0.0000
exp:0˜˜in<1>:you	958	0.77	0.0000
exp:1˜˜in<1>:sorry	39	0.41	0.0000
exp:1˜˜in<1>:want	157	0.64	0.0000
exp:1˜˜in<1>:doesn't	39	0.49	0.0000
exp:1˜˜in<1>:bad	52	0.56	0.0000

APPLICA.AI

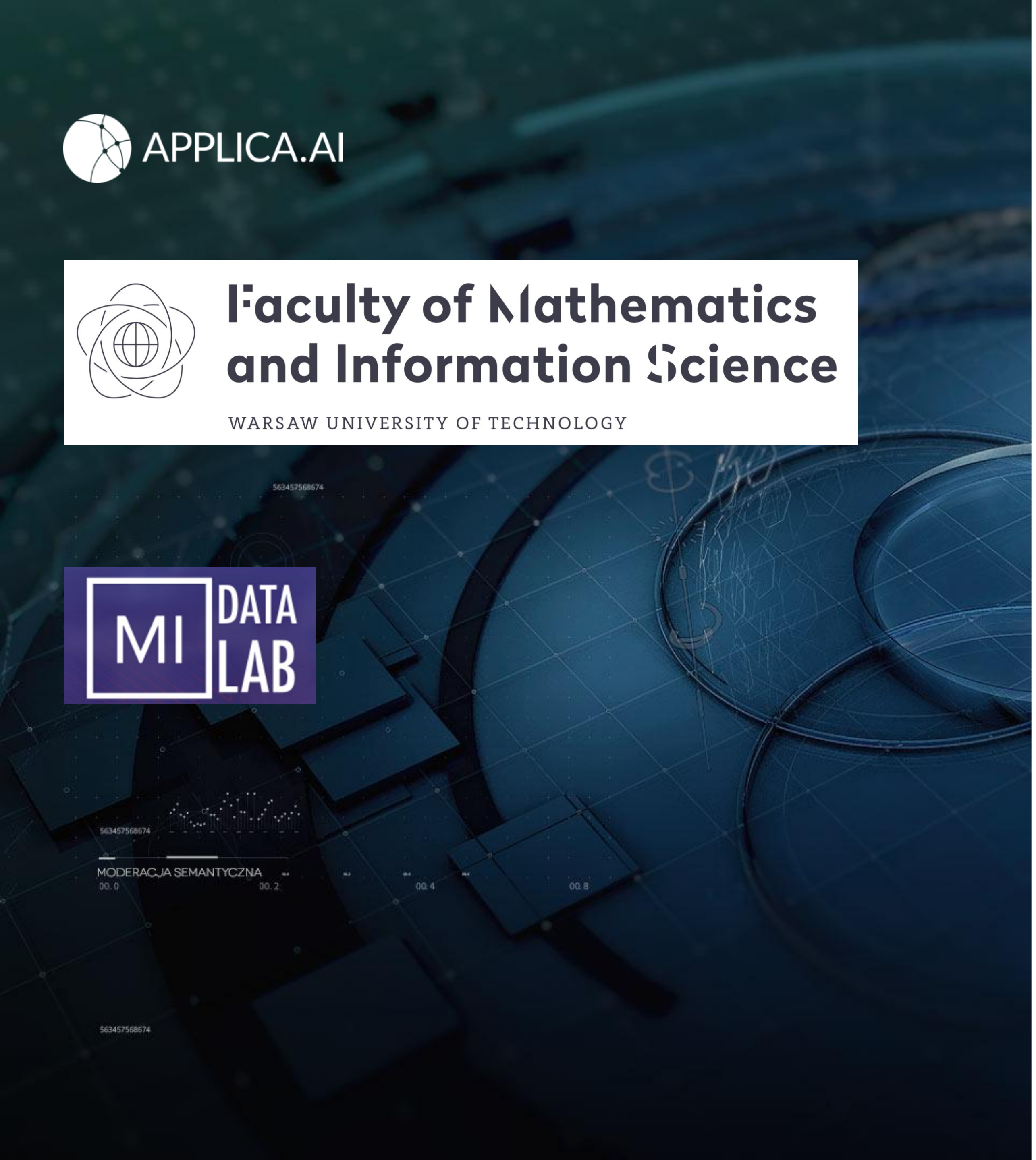
PRE-PROCESSING	OVERALL ACCURACY
NO PREPROCESSING	82.520%
REPLACING "DON'T" WITH "DO NOT" (TRAIN & TEST SAMPLES)	82.575%
CHANGING SENTENCES WITH "BUT" (ONLY TEST SAMPLES)	82.592%

IMDB

FEATURE	COUNT	ACC	P-VALUE
in<1>:now++?	20	0.60	0.003
in<1>:ever++happened	21	0.67	0.014
in<1>:stripped	44	0.75	0.013
in<1>:quite++interesting	28	0.71	0.017
in<1>:weird	513	0.896	0.024
in<1>:would++prefer	13	0.615	0.020
in<1>:objections	13	0.615	0.020
in<1>:DID++NOT	10	0.600	0.030

Challenges

- The most difficult task - to define the task :-)
  - Find resources
  - Instruction for linguists
  - Annotating
- Adjust ML to Polish language
  - Need rules for grammar to be correct or preprocessing
  - Extended stopwords lists and reinforcement words lists
  - Complex word interactions



Future works

- Hierarchical models: emotional / not emotional texts
- Transfer learning / concept language modeling
- Topic modeling + Emotion modeling
- Active learning to choose the best sample to annotate