Detection of abusive speech for mixed sociolects of Russian and Ukrainian languages

Bohdan Andrusyak
Mykhailo Rimel
Roman Kern


TU
Graz
Graz University of Technology


Grenoble INP


KNOW
Center

Ukrainian alphabet

А а Б б В в Г г Ґ ґ Д д Е е
Є є Ж ж З з И и І і Ї ї Й й
К к Л л М м Н н О о П п Р р
С с Т т У у Ф ф Х х Ц ц Ч ч
Ш ш Щ щ Ь ь Ю ю Я я

Russian alphabet

А а Б б В в Г г Д д Е е
Ё ё Ж ж З з И и Й й К к
Л л М м Н н О о П п Р р
С с Т т У у Ф ф Х х Ц ц
Ч ч Ш ш Щ щ Ъ ъ Ы ы Ь ь
Э э Ю ю Я я

Letters with same phonetics

І і <=> И и
И и <=> Ы ы
Є є <=> Э э

What is surzhyk?

Czech:

- Myslím, že nejlepší bar je ten blízko kina

Ukrainian:

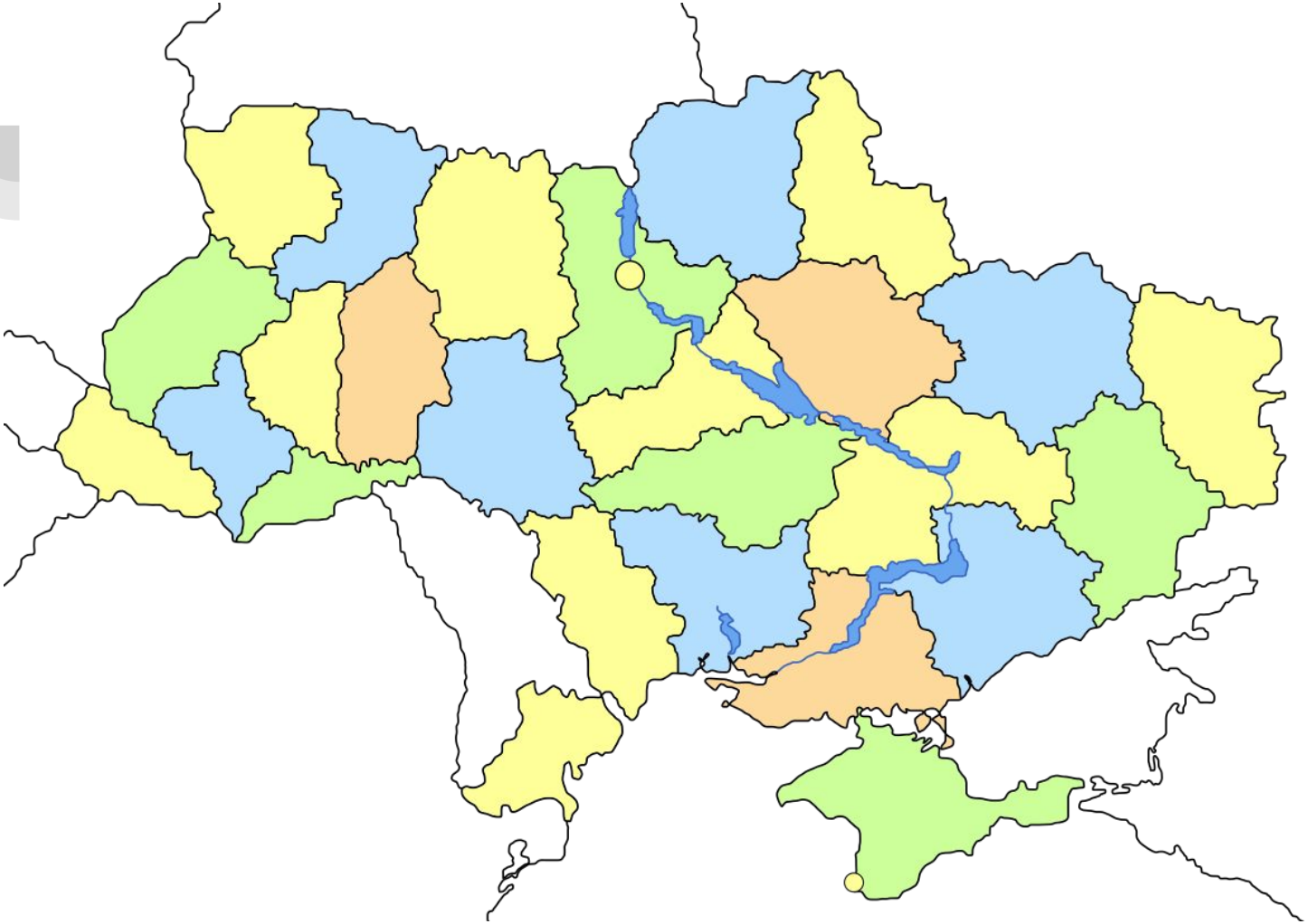
- Я думаю, що найкращий бар це той що біля кінотеатру

Russian:

- Я считаю, что лучший бар это то что у кинотеатра

Surzhyk:

- Я счітаю, што найкращий бар цей той што біля кінотетра



Problems of abusive speech in social media

- Bullying
- Destroying ability to have an conversation
- Incitement to hatred

Abusive words in Ukrainian and Russian

Due to the peculiarities of word formation in Russian and Ukrainian languages, it is practically impossible to define a finite list of abusive word.

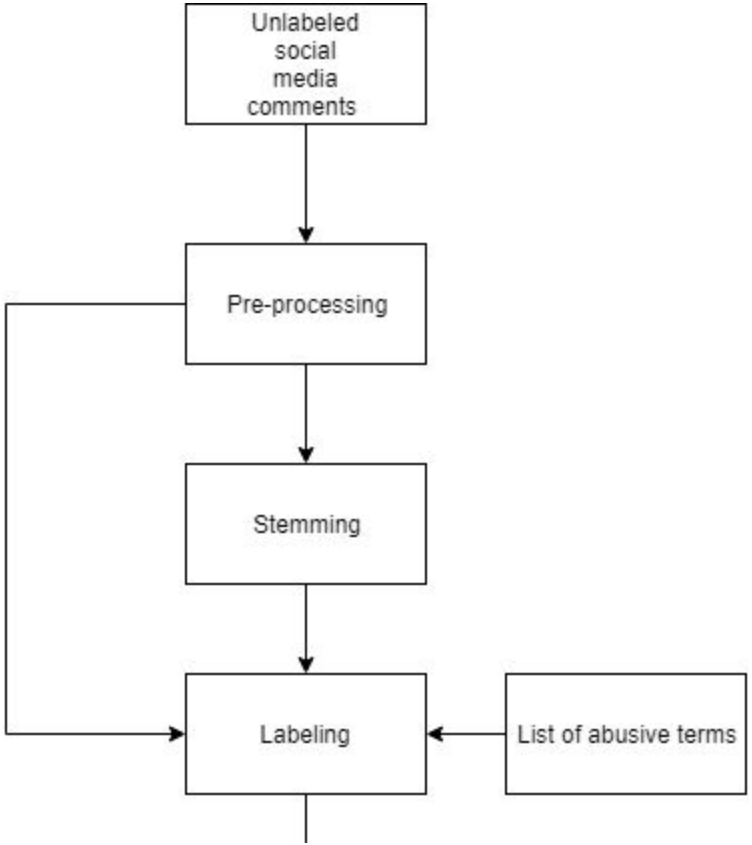
- пизда (noun) + рыло (noun) = пиздорылый (adjective)
- автобус (noun) + хуй (noun) = хуйобус (noun)

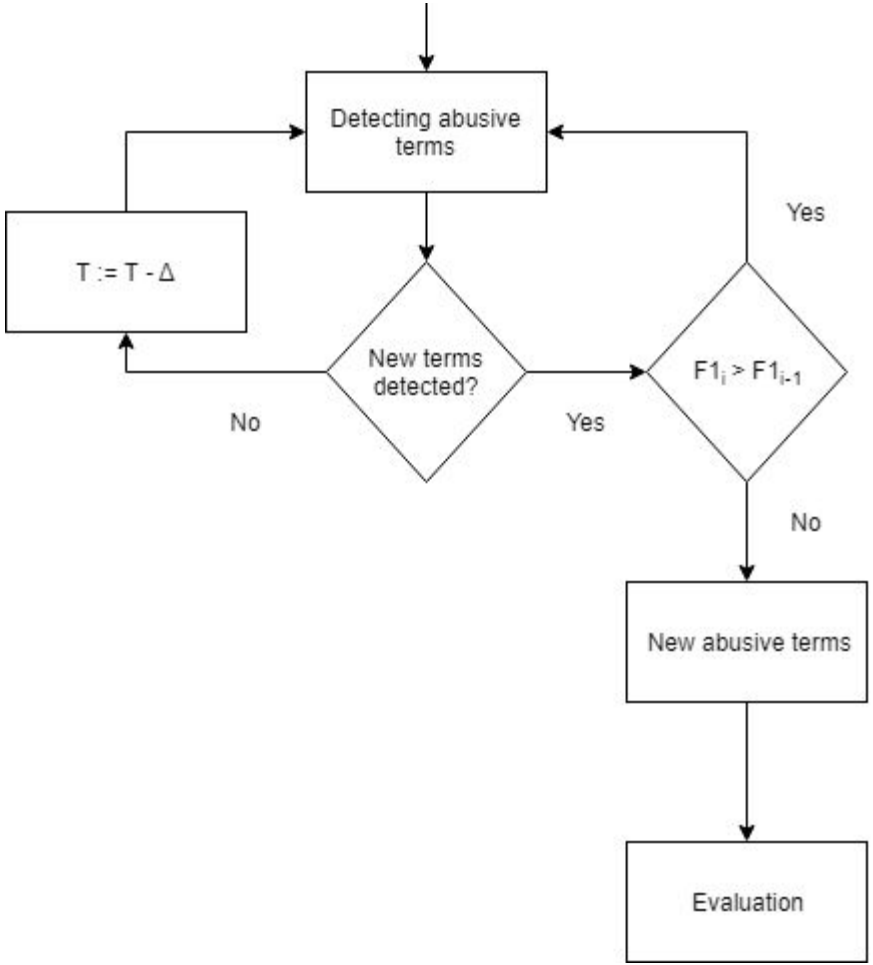
Use of surzhyk creates great variety of abusive words.

- пИзда, пЫзда, пЕзда, пЁзда, пІзда, пЄзда

Uses try to mask abusive words:

- пи3да, пзда, п..зда, пизд@


Unlabeled social media comments
Pre-processing
Stemming
Labeling
List of abusive terms


Detecting abusive terms
T := T - Δ
New terms detected?
No
Yes
$F1_i > F1_{i-1}$
Yes
No
New abusive terms
Evaluation

Pre-processing

- Remove comments written in Latin alphabet
- Remove from comments:
  - Punctuation
  - Numbers
  - Emoticons
  - Other non-alphabetic symbols
- Stemming:
  - Both options were tested, with stemmed and unstemmed words

Initial Labeling

- Initial labels are based on a provided list of abusive words
- Comment is labeled as abusive if at least one word matches word from list of abusive words

Finding new abusive words

1. Comments are separated in two groups “good” and “bad”.
2. Calculate for each word in dataset likelihood of being in “good” and “bad” group.
3. Calculate for each word:
    - Relative distance
    - Log odds ratio
4. Word with Relative Distance or Log Odds Ratio higher than threshold T, are considered to be abusive and are added to the initial list of abusive words
5. If no words have RD or LOR higher than threshold T, T is lower by delta
6. Comments are relabled with expanded list of abusive words

Formulas for Relative Distance and Log Odds Ratio

$$P(x) = \frac{p_b(x) - p_g(x)}{max(p_b(x), p_g(x))}$$

$$L(x) = \frac{\frac{p_g(x)}{1-p_g(x)}}{\frac{p_b(x)}{1-p_b(x)}}$$

Example

Good	Bad
на 0.95	в 0.95
в 0.93	на 0.9
кіно 0.8	**пизда** 0.89
пизда 0.05	кіно 0.03

Word	Relative distance
пизда	0.94
на	0.02
в	0.02
кіно	-0.96

Experimental setup

- Social Media: YouTube
- Topic: Euromaidan, Revolution of Dignity
- 329 videos
- over 50.000 comments
- 2.000 comments manually labeled for evaluation
  - 32.7% as abusive
- Intial list of abusive terms over 600 words
- Micro list of 5 words

Results

Method	P	R	F_1	$Words$
SD	0.875	0.510	0.644	—
SD, RD	0.736	0.629	0.678	8
SD, LOR	0.678	0.629	0.652	11
STM, SD	0.742	0.629	0.681	—
STM, SD, RD	0.667	0.675	0.671	10
STM, SD, LOR	0.474	0.741	0.578	13
MSD	0.857	0.158	0.268	—
MSD, RD	0.588	0.463	0.518	44
MSD, LOR	0.303	1.000	0.465	573
STM, MSD	0.897	0.231	0.368	—
STM, MSD, RD	0.684	0.344	0.458	17
STM, MSD, LOR	0.308	0.920	0.462	145

- SD - seed dictionary
- RD - relative distance
- LOR - log odds ratio
- STM - stemmed dictionary
- MSD - micro seed dictionary
- Words- number of new abusive terms added to the seed dictionary

Links to data

All comments:

- https://github.com/bohdan1/AbusiveLanguageDataset/blob/master/data.csv

Labeled comments:

- https://github.com/bohdan1/AbusiveLanguageDataset/blob/master/labled.csv

List of abusive terms:

- https://github.com/bohdan1/AbusiveLanguageDataset/blob/master/bad_words.txt

Micro list of abusive terms:

- https://github.com/bohdan1/AbusiveLanguageDataset/blob/master/bad_words_seed.txt