A MACHINE LEARNING APPROACH TO CLASSIFICATION OF DRUG REVIEWS IN RUSSIAN

Ilseyar Alimova
Elena Tutubalina
Guzel Gafiyatullina
Kazan Federal University

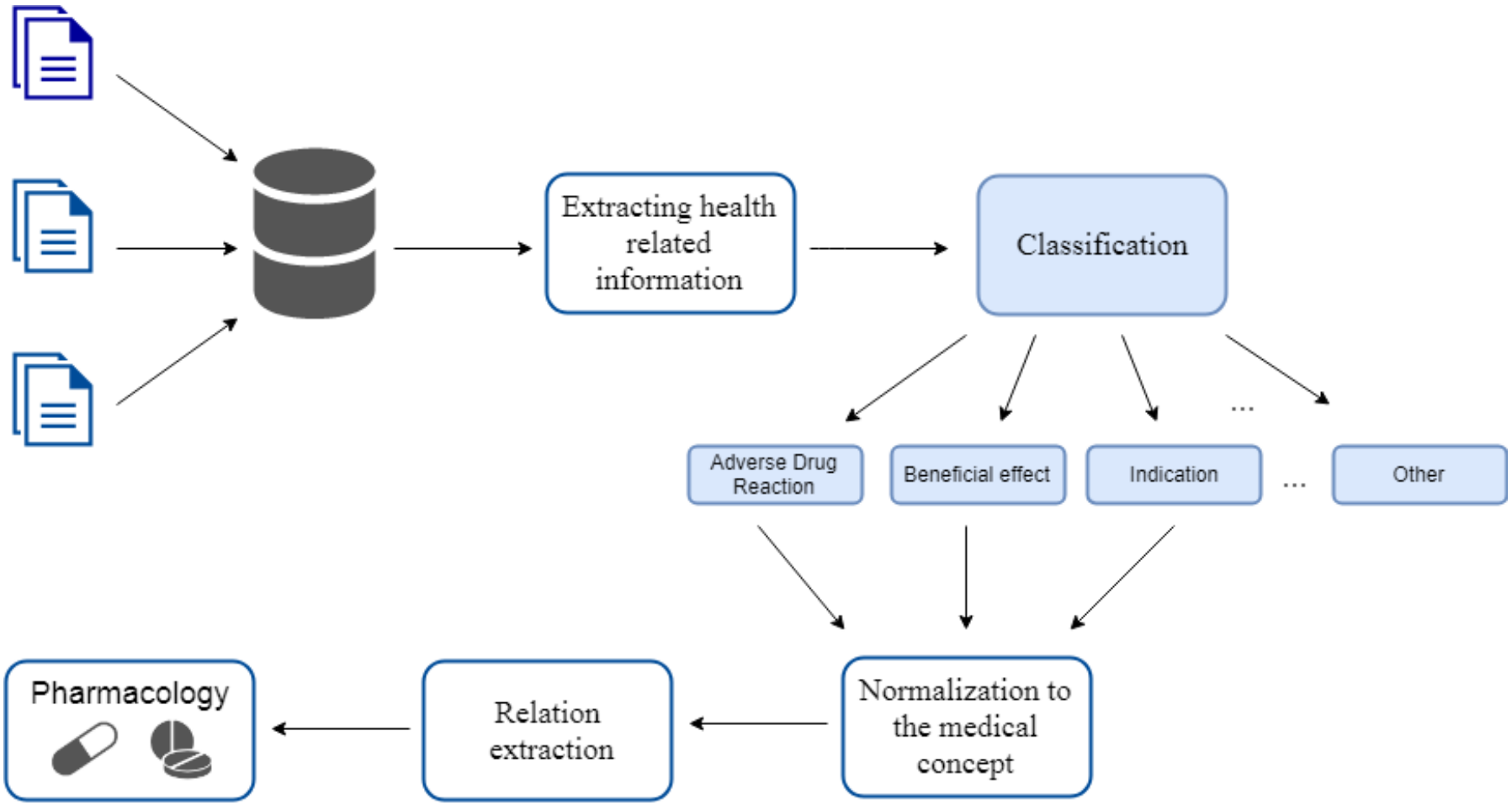
Julia Alferova
N.V. Sklifosovsky Research Institute of Emergency Care

Introduction

- Detection of side effects in the post-approval period
- Users write about side effects in social media
- Classify disease-related information in Russian

Text mining for pharmacology

Using social media



Research questions

- Classification of drug reviews on sentence level in Russian for four classes:
  - Indication
  - Beneficial side effect (BNE)
  - Adverse drug reaction (ADR)
  - Other
- Evaluation of features for drug reviews classification

Dataset

- contains 580 reviews from otzovik.com
- 5 748 sentences
- 646 - Indication sentences
- 335 - Beneficial effect
- 279 - Adverse Drug Reaction
- 4 488 - Other

Examples

Label	Example
ADR	Стала спокойной даже чересчур, на работе стала тупить, коллеги сказали что я какая то заторможенная, все время клонит в сон.
	I became calm even too much, at work began to blunt, colleagues said that I am a little bit inhibited, want to sleep all the time.
BNE	Прием мелаксена помог наладить сон.
	Taking of Melaksen helped to establish a sleep.

Examples

Label	**Example**
Indication	Я пользуюсь свечами по 1 000 000 МЕ как для профилактики, так и для лечения ОРВИ.
	I use candles for 1 000 000 ME for both prevention and treatment of cold.
Other	Время использования: месяц Стоимость: 150 руб.
	Duration of use: month. Cost: 150 rub.
	Общее впечатление : Снотворное донормил может вас удивить, но не всегда приятно
	General impression: sleeping pill Donormyl can surprise you, but not always pleasantly

Method

SVM with linear kernel with features:

- bow – Bag of Words
- pos – Part-of-speech tag
- emb – Word embedding
- sent – Sentiment
- pol – Polarity
- dis – Disease lexicon
- drug – Drug name presence
- pmi – Pointwise mutual information

Resources

- Ruscorpora vector representation from RusVectores resource
- RuSentiLex
- Texterra[1] polarity detection
- Manually collected disease lexicon
- Manually collected drug names
- PMI was counted on 80698 unlabeled reviews from forum tzovik.com

[1] https://api.ispras.ru/products

Results

5-fold cross-validation for 4 classes

Features	**av. F**
bow	.544
bow, pos	.544
bow, dis	.550
bow, w2v	.559
bow, pol	.547
bow, sent	.545
bow, drug	.544
bow, pmi	.538
bow, w2v, pol, dis, pmi	**.565**
all features	.564

Results

for 2 classes (Indication, BNE, ADR) versus (Other)

Features	**av. F**
bow	.718
bow, pos	.720
bow, dis	.727
bow, w2v	.724
bow, pol	.717
bow, sent	.718
bow, drug	.718
bow, pmi	.719
bow, dis, w2v, pos, pmi	**.733**
all features	.732

Word Embedding evaluation

- RusDrugReviewsVec trained on 127 840 unlabeled drug reviews from three resources: otzovik.com, meduniver.com, rusmedserve.ru
- Word embedding model from RusVectorizers resource
- Binary classification (*Indication*, *BNE*, *ADR*) vs. (*Other*)

	Dictionary	**Vector dim.**	**av. F1**
bow, ruwikiruscorpora	392339	300	.721
bow, web	267540	300	.722
bow, news	194058	300	.722
bow, ruscorpora	184973	300	**.724**
bow, ruscorpora 2017	173816	600	.722
bow, RusDrugReviewsVec	95013	100	**.724**

Lexicons evaluation

Resources	RuSentiLex	Disease lexicon	Drug name list	Ruscorpora word2vec model	All
unique unigrams	849	61	41	3976	6209
unigrams	13466	820	329	34108	65505

Conclusion

- The idea is applicable
- Necessary to develop special resources for Russian language
- Source code can be found at https://github.com/Ilseyar/adr_detection_russian