STU FIIT

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Sentiment Analysis of Social Network Posts in Slovak

Rastislav Krchňavý
Supervisor: Marián Šimko
Consultant: Matej Hruška

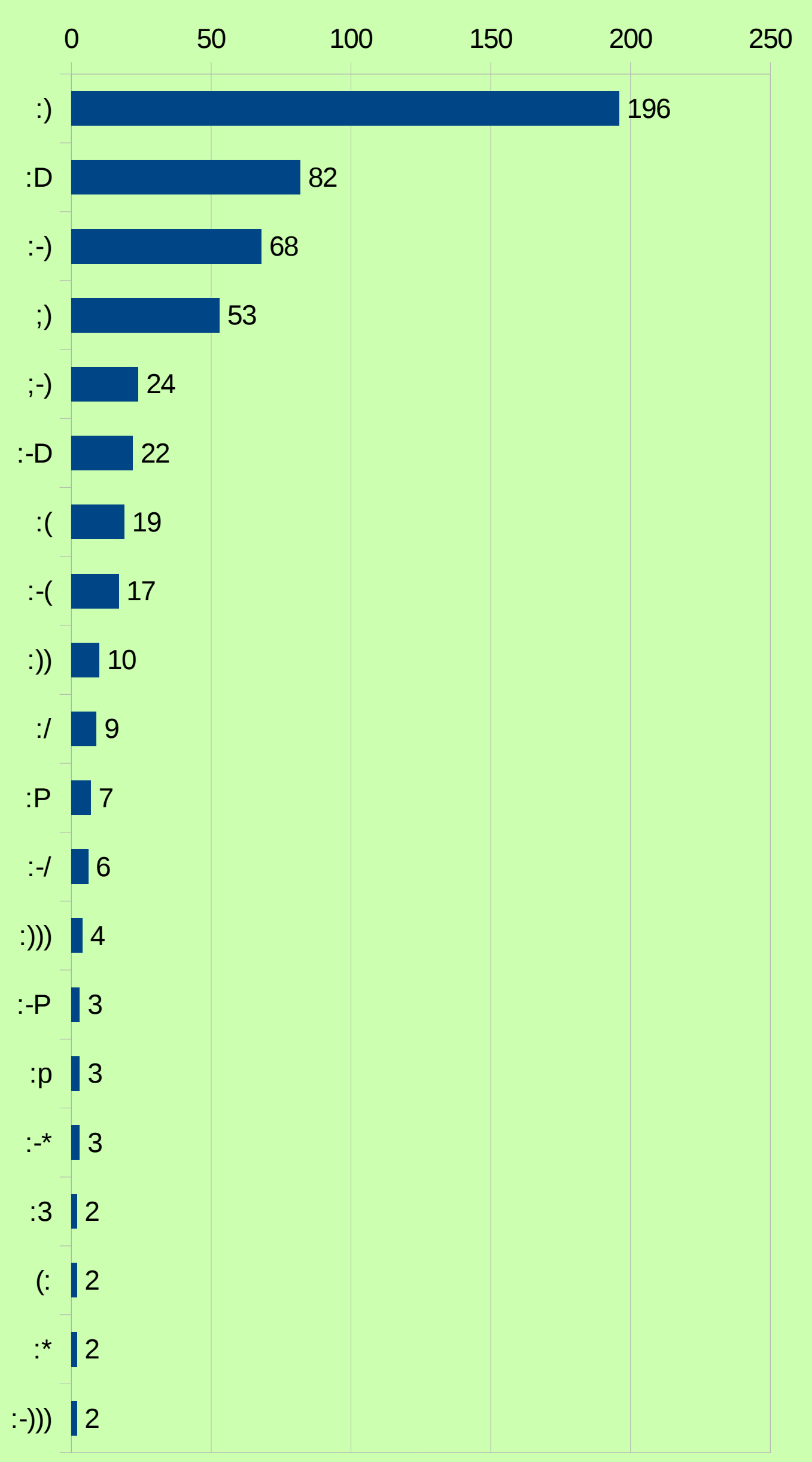
Motivation

- Slovak language
  - currently there is no existing solution for Slovak language
  - special rules
- Social networks
  - specific language
  - variable length
- Multiple domains
  - finance, living, retail, gastronomy, telco
- Existing approaches for another language
  - in English over 80 % in 2 classes (Pang, Lee, 2002)
  - in Czech over 70 % in 3 classes (Koktan, 2012)
- Humans agree in 79 % (Onegva, 2010)
- Main purpose – data analysis

Data

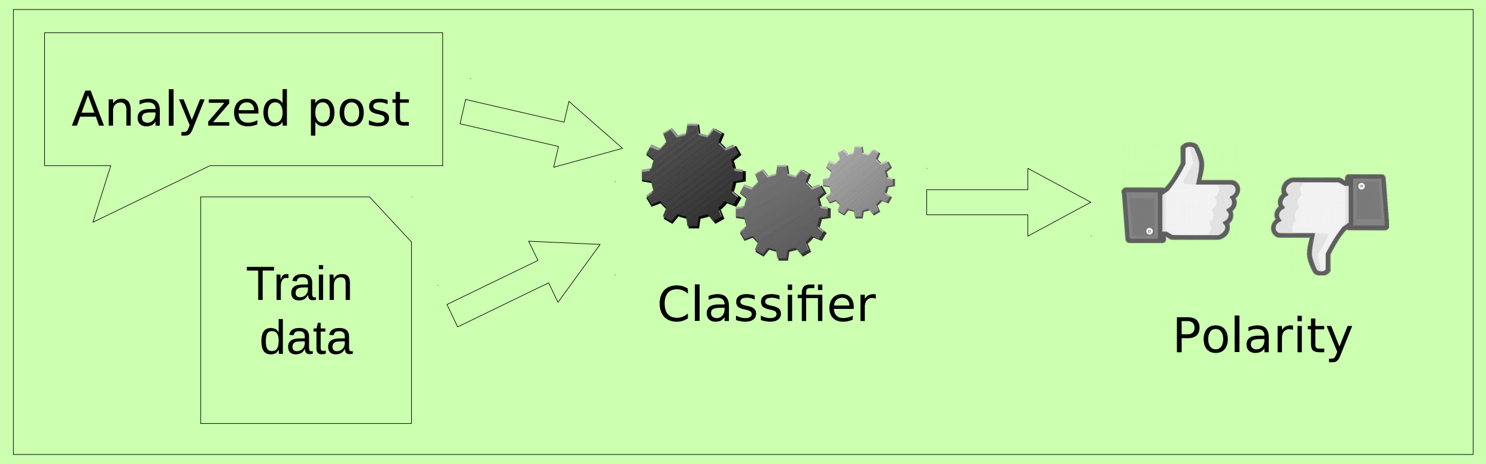
- Texts
  - Manually annotated by Seesame
  - Categorized into 5 classes from strongly negative to strongly positive
  - Over 1500 Facebook posts
- Lexicons
  - Data from Slovak Sentiment Lexicon project
  - Automatically translated lexicon from MPQA project
- Preprocessing
  - List of most used emoticons
  - Slovak National Corpus for lemmatizing and grammar categories detection

Emoticons in dataset



Our Method

- Preprocessing – converting plaint text to features
- Segmentation, emoticon extraction, lemmatization, removing stop words etc.
- Number of classes from 2 to 5
- Building the classifier – machine learning approach (Naïve Bayes, Maximum Entropy, Support Vector Machines) or lexicon based approach
- Decision based on train data
- Quality and quantity of data



Experiment & Results

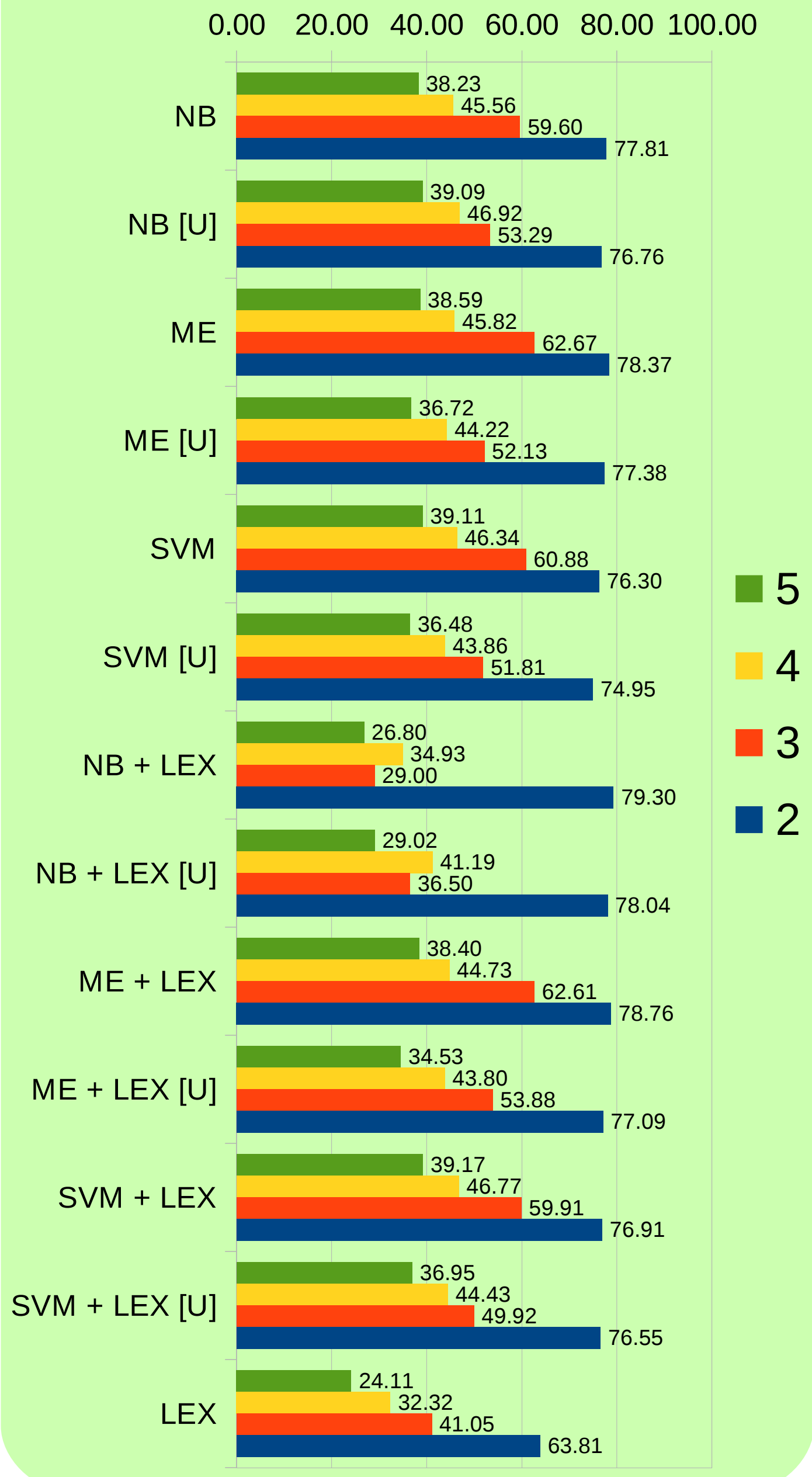
- For machine learning classifier we use Natural Language Toolkit in Python
- K-fold cross validation
- We measure accuracy, precision and recall
- Undersampling – make train set for each class same size
- Combinations of machine learning and lexicon based approaches
- Lots of parameters in pre-processing
- Comparable with state-of-the-art in world languages

Best results for different classes

Classes	2 [-2,-1],[1,2]	3 [-2,-1],0,[1,2]	4 -2,-1,1,2	5 -2,-1,0,1,2
Accuracy	79.30 %	62.67 %	46.92 %	39.17 %
Method	NB + LEX	ME	NB [U]*	SVM + LEX

* [U] – undersampled

Detailed accuracy



Seesame

- Seesame is Slovak PR agency
- They maintain Facebook profiles of various companies
- The companies wants to know the society opinion on new products
- They use our sentiment analyzer for detecting sentiment from Facebook comments
- Special thanks for making the main dataset of posts

SEESAME
COMMUNICATION EXPERTS

Contribution

- Method for sentiment analysis of Slovak texts
- Adopted for specific social network-based content (wall posts, comments)
- Annotated datasets for further research
- Comparison and evaluation of different approaches and classifiers
- Experimenting with classifier combination
- Evaluated on real muti-domain data from Facebook
- Over 79 % accuracy in 2 class classification – comparable to state-of-the-art
- Deployed as web service, online tool for analyzing .csv, .xsl files, and web page
- Ongoing evaluation with Seesame end-users
- Real world impact: potential improvement PR and marketing of brands maintained by Seesame