SMS Sentiment Classification based on Stylometric Features, Emoticons, Informal abbreviations and other Text Features

Branislava Šandrih

branislava.sandrih@fil.bg.ac.rs

University of Belgrade, Faculty of Philology, Serbia

JeRTeh – Society for Language Resources and Technology

Motivation

- Sentiment analysis / Opinion mining / Sentiment classification:
  - contextual mining of text which identifies and extracts subjective information
- Analysis of social media is usually restricted to just basic sentiment analysis and count based metrics
  - what about SMS messages? They are even shorter!

Challenge

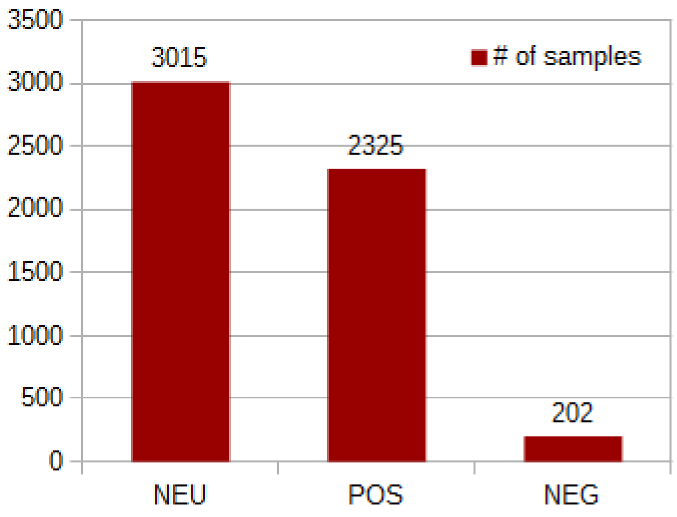
- Restrictions:
  - – 160 characters
    - 70 if diacritics are used
  - – Small keyboards, hard to type
    - messages contain only most important information
  - – Need to express attitude, mood, voice tone, facial expression, gesture…
    - the only available tool: characters!?

Workaround

- Authors
  - use sh-s for common used phrases
  - EMPHASIZE IMPORTANT INFORMATION WITH UPPERCASE
  - do not type whole ws
  - omit diacritics
  - excessively use emoticons :) :(:-P
- Consequence?
  - Hard to analyze using standard approaches

A different approach

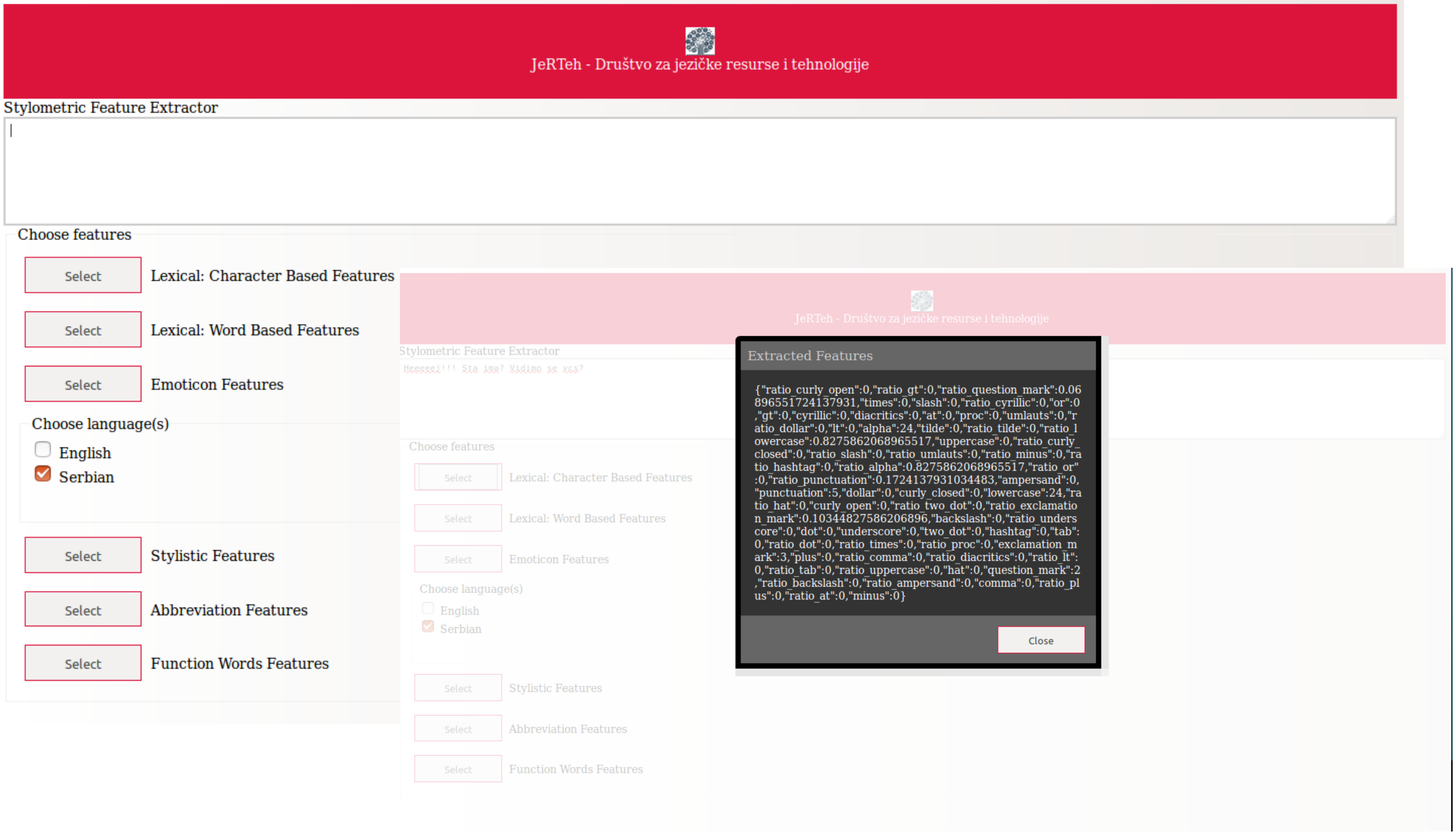
- But first dataset
  - modest ~ 5,500 SMS messages in Serbian/German/English, Cyrillic + Latin
  - hard to gather, because SMS are too personal!
  - Manual annotation



Features

- Lexical
  - Character based
    - counts of lowercase and uppercase letters, total # of characters, ratios…
  - Word based
    - average sentence length, average length of tokens etc.
- Stylistic
  - sentence starts with uppercase, spaces after punctuation etc.
- Emoticons
  - kiss, confused, heart etc.
- Abbreviatons
  - ty, tnx, k, fb, cu, u, l8r etc.

API and Web Interface



Heeeeej!!! Sta ima? Vidimo se vcs?

{"ratio_curly_open":0,"ratio_gt":0,"ratio_question_mark":0.06896551724137931,"times":0,"slash":0,"ratio_cyrillic":0,"or":0,"gt":0,"cyrillic":0,"diacritics":0,"at":0,"proc":0,"umlauts":0,"ratio_dollar":0,"lt":0,"alpha":24,"tilde":0,"ratio_tilde":0,"ratio_lowercase":0.8275862068965517,"uppercase":0,"ratio_curly_closed":0,"ratio_slash":0,"ratio_umlauts":0,"ratio_minus":0,"ratio_hashtag":0,"ratio_alpha":0.8275862068965517,"ratio_or":0,"ratio_punctuation":0.1724137931034483,"ampersand":0,"punctuation":5,"dollar":0,"curly_closed":0,"lowercase":24,"ratio_hat":0,"curly_open":0,"ratio_two_dot":0,"ratio_exclamation_mark":0.10344827586206896,"backslash":0,"ratio_underscore":0,"dot":0,"underscore":0,"two_dot":0,"hashtag":0,"tab":0,"ratio_dot":0,"ratio_times":0,"ratio_proc":0,"exclamation_mark":3,"plus":0,"ratio_comma":0,"ratio_diacritics":0,"ratio_lt":0,"ratio_tab":0,"ratio_uppercase":0,"hat":0,"question_mark":2,"ratio_backslash":0,"ratio_ampersand":0,"comma":0,"ratio_plus":0,"ratio_at":0,"minus":0}

Results

- Playing with features, incrementally adding:
  - Lexical
  - Lexical + Stylistic
  - Lexical + Stylistic + Emoticons
  - Lexical + Stylistic + Emoticons + Abbreviations
- Accuracy 94.4% in the last case

Conclusion

- For short messages, sentiment classification can be performed by exploring stylometry and other important characteristics

NOKIA
TY!