ktu kaunas university of technology 1922

Research Council of Lithuania

Co-funded by
the Prevention of and Fight against Crime
Programme of the European Union

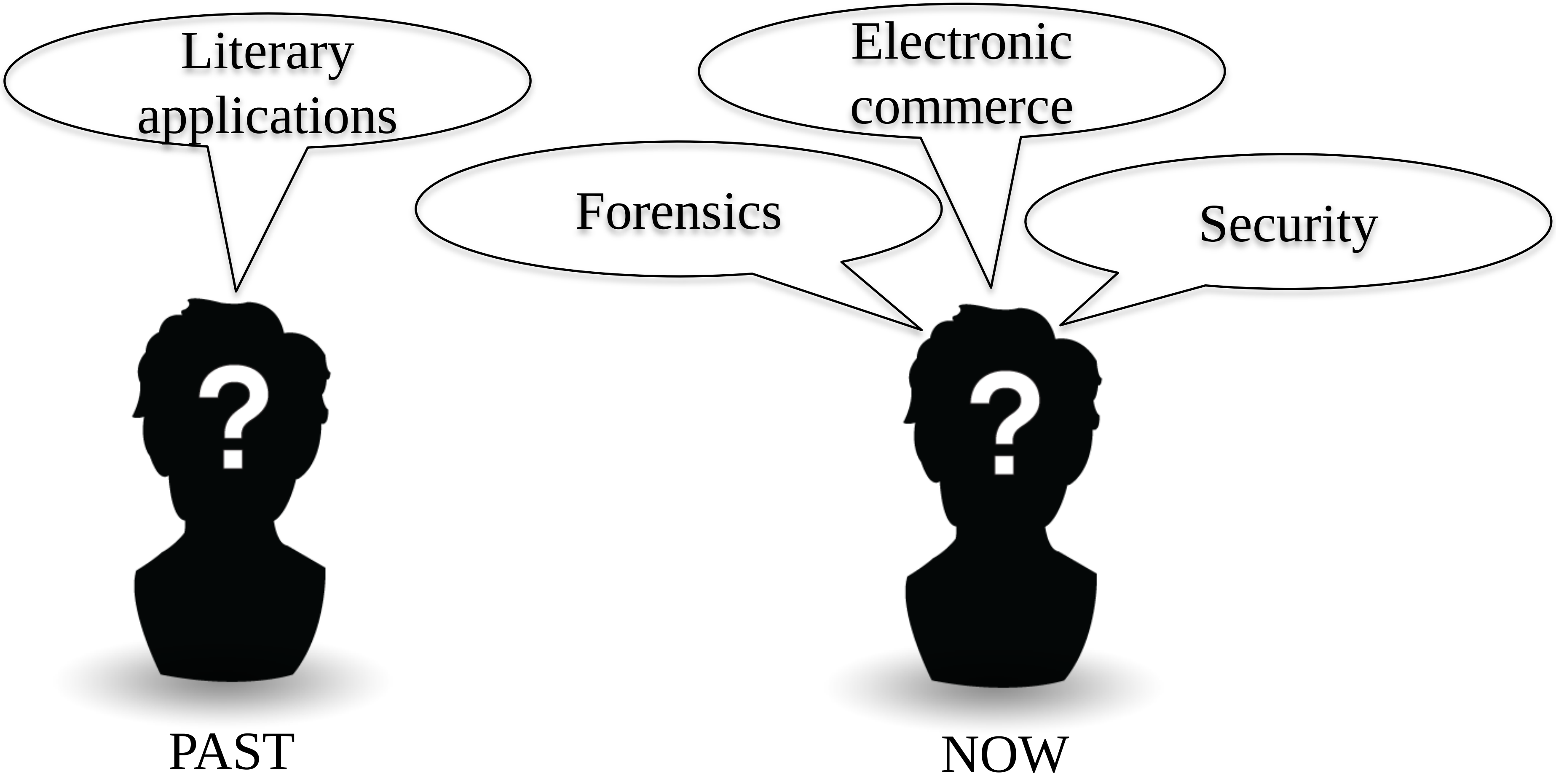
Authorship Attribution Approaches for the Lithuanian Language Using Internet Comments

JURGITA KAPOČIŪTĖ-DZIKIENĖ

Authorship Attribution (AA)

- AA – a task of identifying who from a set of candidate authors is an actual author of a given anonymous text
- Based on existing "human stylome" notion (*Van Halteren, 2005*)
- AA is one of the oldest problems, highly topical nowadays

Applications



Literary applications

PAST

Electronic commerce

Forensics

Security

NOW

Related works (1/2)

- The research is done with:
  - e-mails, web forum messages, online chats, Internet blogs, tweets

    /short texts, non-normative language/
  - hundreds (*Luyckx and Daelemans, 2008*) or thousands authors (*Koppel et al., 2011*); (*Narayanan et al., 2012*)

    /larger candidate author sets/

Related works (2/2)

- Concept of idiolect for the first time discussed in 1971 (*Pikčilingis, 1971*)

- Lots of descriptive linguistic works (e.g. *Žalkauskaitė, 2012*; *Venčkauskas et al., 2015*)

- AA research with:

Reference	**Texts**	**Authors**	**Methods**
(*Kapočiūtė-Dzikienė et al., 2015*)	Parliamentary transcripts forum posts	100	Machine learning
(*Kapočiūtė-Dzikienė et al., 2015a*)	Fiction texts	3; 5; 10; 20;50;100	Machine learning
(*Kapočiūtė-Dzikienė et al., 2015b*)	Internet comments	1,000	Similarity-based

The corpus*

- Internet comments from www.delfi.lt (topics “in Lithuania” and “Abroad”) in January 2015 – August 2015

- To get clean corpus none of these authors were included:
  - if the same pseudonym was used for writing from different IP addresses (dynamic IP problem)
  - if different pseudonyms were used under the same IP address

- Replies, meta-information and non-Lithuanian alphabetic letters (except punctuation marks and digits) were filtered out

- No texts shorter than 30 symbols (excluding white-spaces)

- Texts with out-of-vocabulary words, missing diacritics, etc.
  /spoken non-normative Lithuanian language/

* At: http://dangus.vdu.lt/~jkd/wp-content/uploads/2015/04/INT_KOMENTARAI_INDV2.7z

Corpus statistics

Number of authors	**10**	**100**	**1,000**
Number of texts	14,443	63,131	155,078
Number of tokens (words/digits)	289,462	1,511,823	4,068,231
Avg. text length in tokens	20.042	23.947	26.233
Random baseline	0.001	0.002	0.003
Majority baseline	0.018	0.018	0.018

Main research directions

- Methods:
  - Machine Learning (**ML**): Naïve Bayes – **NB**; Support Vector Machine – **SVM**)
  - Similarity-Based – **SB** (cosine similarity measure)

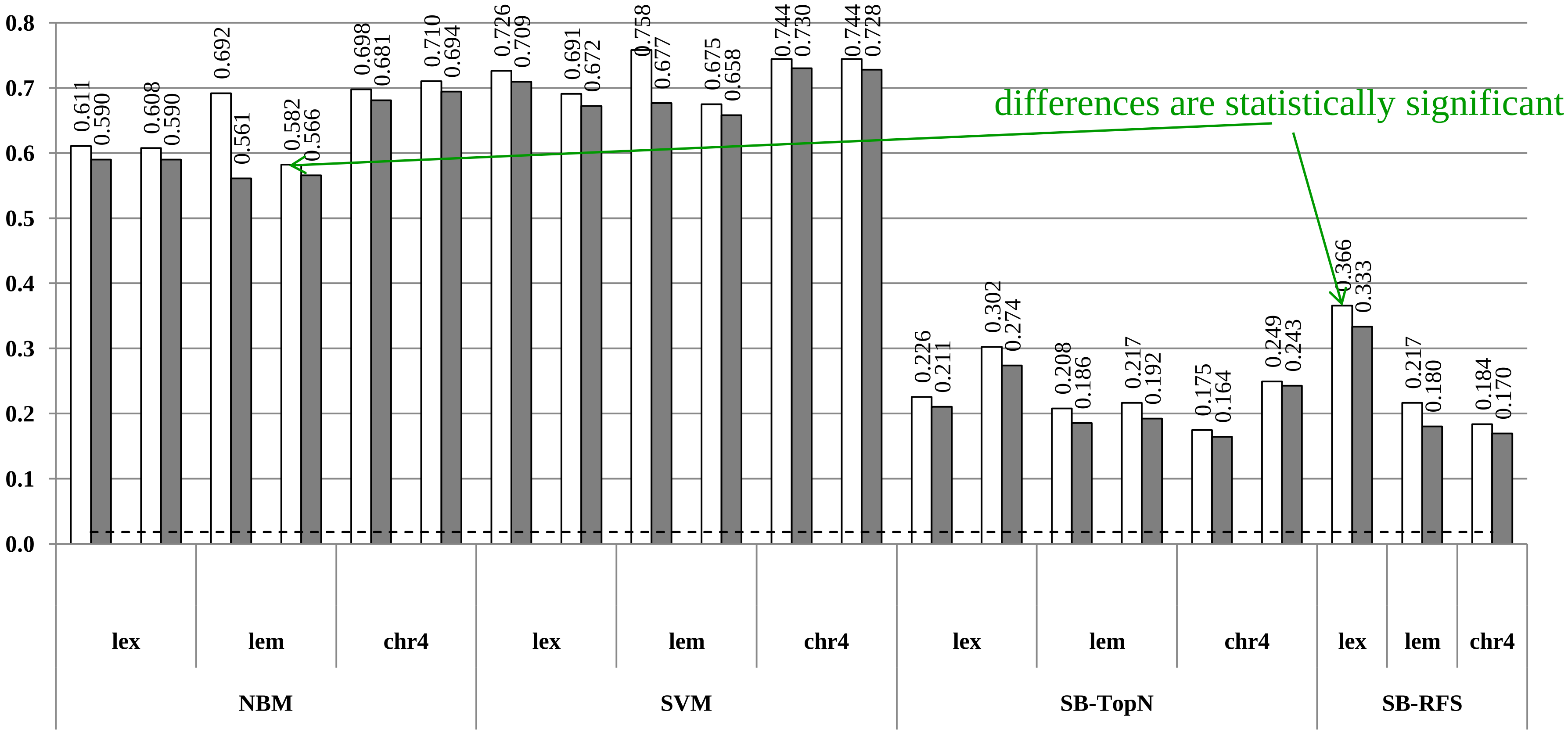
- Feature types:
  - **lex** – lexical
  - **lem** – lemmas – using *Lemuoklis* (*Zinkevičius*, *20000*)
  - **chr4** – word-level character tetra-grams

- Dimensionality reduction techniques:
  - Entire feature set
  - Feature ranking (using chi-squared) and **TopN** (*N*=30 thousand)
  - Random feature set of *N* features in *K* iterations (*K*=20) – **RFS** (offered by *Koppel et al.*, *2011*)

Experimental setup

- Stratified dataset
- 80%/20% for training/testing, respectively
- Calculated accuracies/f-score values
- McNemar’s (*McNemar, 1947*) test for statistical significance (α = 0.05)

10 candidate authors*



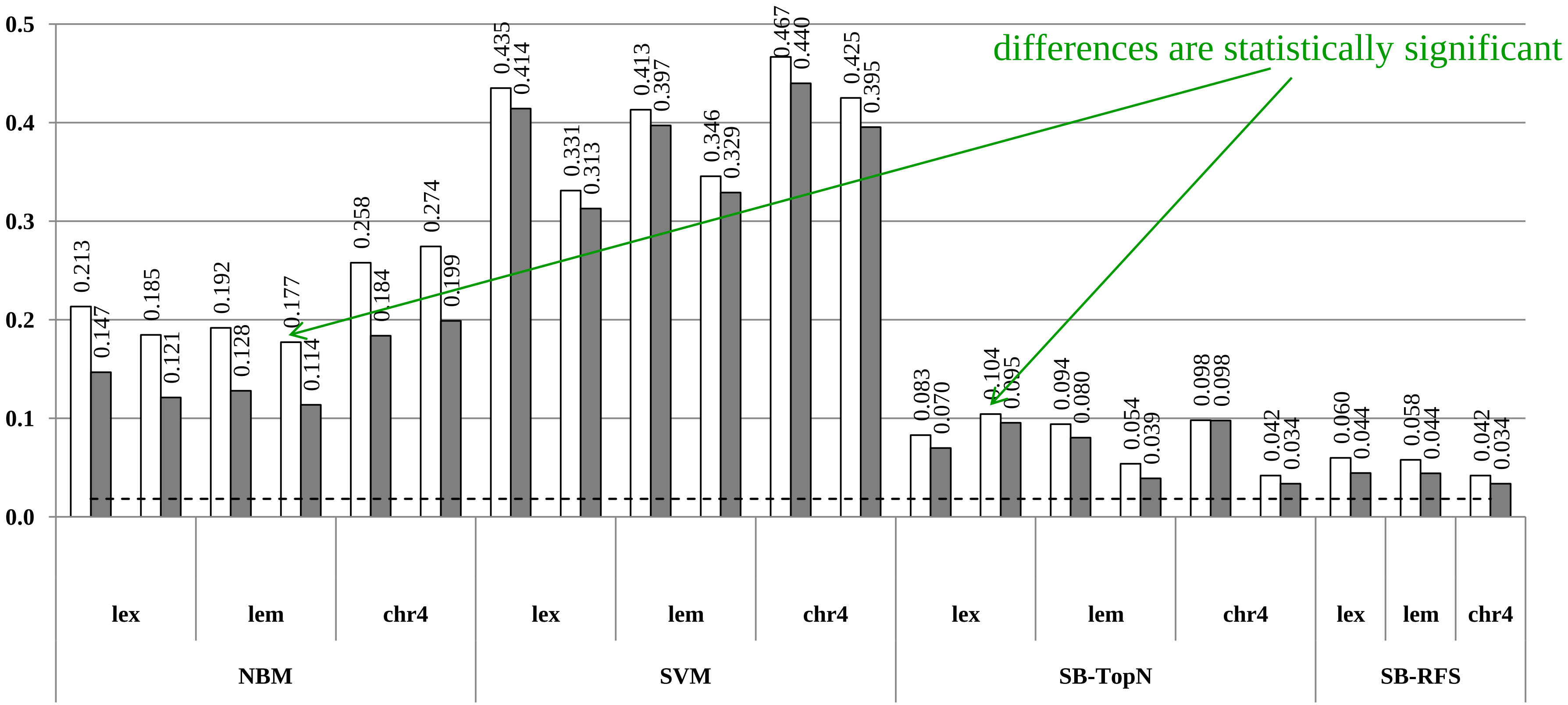
differences are statistically significant

* *Accuracy* and *f-score* in white and gray columns, respectively.

The first two columns – results on the entire feature set; the second two – on *N*=30,000 features

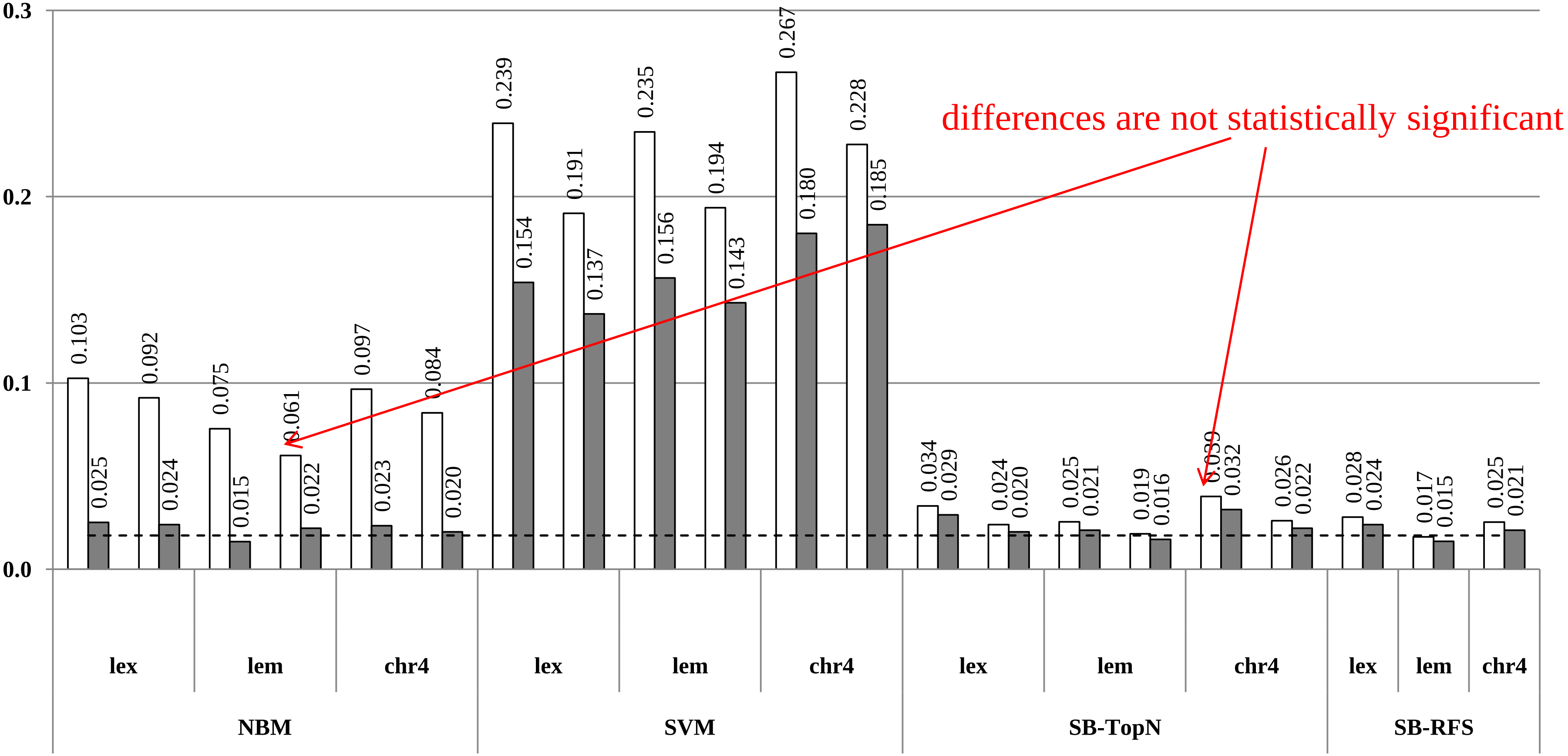
The dashed line indicates higher of random and majority baselines

100 candidate authors



differences are statistically significant

1,000 candidate authors



differences are not statistically significant

Summary of results

- Not all results exceed random and majority baselines:
  - Lemmatizer is not adjusted to cope with the non-normative language
- **ML** > **SB**
- **chr4** > **lex** > **lem** (on the larger author sets)
- **Entire feature set** > **RFS** (*N* = 30,000; *K* = 20) > **TopN** (*N* = 30,000)

Conclusions and future work

- Conclusions:
  - The first comparative AA results (in terms of methods, feature types, dimensionality reduction techniques) for the Lithuanian language using:
    - Internet comments
    - a big author set (i.e., 1,000 candidate authors)
  - The best results (especially on the larger author sets) with ML + chr4

- Future work:
  - Error analysis and improvements
  - Expanded number of candidate authors
  - Different types of non-normative texts (e.g., social networks data)

Researchers who also worked on this topic

Ligita Šarkutė
KUT

Andrius Utka
VMU

Algimantas Venčkauskas
KUT

Robertas Damaševičius
KUT

Research Council of Lithuania

Automatic extraction of style applied to individual authors and groups of authors (ASTRA) (LIT-8-69)
http://dangus.vdu.lt/~jkd/eng/

Co-funded by
the Prevention of and Fight against Crime
Programme of the European Union

Lithuanian Cybercrime Centre of Excellence for Training, Research and Education (HOME/2013/ISEC/AG/INT/400005176)

UNIVERSITAS VYTAUTI MAGNI CAUNAE

ktu
kaunas
university of
technology
1922

Research
Council of
Lithuania

Co-funded by
the Prevention of and Fight against Crime
Programme of the European Union

Thank you for your attention!

jurgita.k.dz@gmail.com