A Twitter Corpus and lexicon for abusive speech detection in Serbian

Danka Jokić, danka.jokic@afrodita.rcub.bg.ac.rs, Ranka Stanković, ranka@rgf.rs ,
Cvetana Krstev, cvetana@matf.bg.ac.rs , Branislava Šandrih, branislava.sandrih@fil.bg.ac.rs

Human Language Technology group at the University of Belgrade, Studenski trg 1, 11000 Belgrade, Serbia

1. Motivation and background

- **AbCoSER - abusive speech corpus in Serbian**
- **abusive speech lexicon**
  - automatic abusive speech detection systems for the Serbian language
- **Existing annotation schemes and abusive term definitions**
  - ✓ used for general data set to detect broad range of abusive topics
  - ✓ detection of abusive triggers
  - ✓ augmentation of the abusive language lexicon

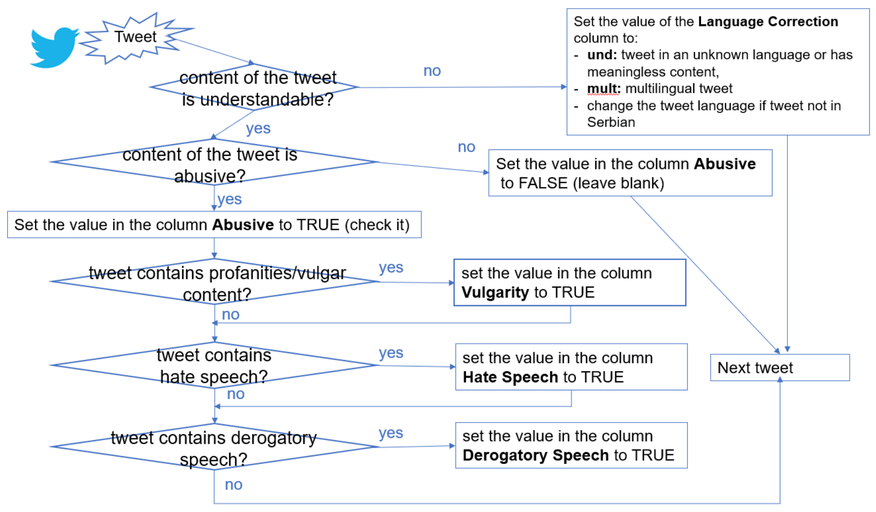
2. Data Acquisition and Annotation

- **Data collection**
  - ✓ tweets from Twitter written in Serbian, without URLs and no retweet
  - ✓ 111 user accounts
    - whose tweets previously reported as hate speech or
    - detected by seed words search
  - ✓ 194,348 tweets (after cleaning of 320,440)
  - ✓ language tag sparce and not reliable
  - ✓ randomly sampled 6,500 tweets, 64 English tweets removed,
  - ✓ Final set for research: 6,436 tweets
- **General Corpus annotation for classification of tweets**
  - ✓ Level 1 – binary classification
  - ✓ Level 2 - Profanity (PROF), Hate speech (HS), Derogatory speech (DS), Other (OTH)

	TWEET DATA		LEVEL 1	LEVEL 2			LANG
Index	Tweet text	Lang	Abusive	Vulgar	Hate	Derogatory	Correction
138451	@Stefan_Visoki Ali bitno da je nacija poruku shvatila nedvosmisleno: "oni nas mnogo jako jebu u mozak"	und	☑	☑	☐	☑	sr
82461	O miševima i mudima.	und	☑	☑	☐	☐	sr
176861	ali “digla dzevu u zari” idi bre kuvaj neki rucak nesto tako	und	☑	☐	☑	☐	sr
16907	Na koji broj se šalje SMS za lečenje Marijana Rističevića?	und	☑	☐	☐	☑	sr
153622	@Milan92551954 Рада Ђурић је говно, било где да ради, чак и у медију власти, остаће безлично говно.	sr	☑	☑	☐	☑	sr
230346	"Jbte, zarazio me" -Jbte, dopustio si. I to vam je cela priča, jer verovati nekom na reč odavno ne pije vodu..	sl	☑	☑	☐	☐	sr
256251	@KalasturaB Jesi kalaštura... Mrš od dece.	und	☑	☑	☐	☑	sr
2154	iz komentara sam zaključio da je čovek poljak jer često piše na poljskom, a moja izvanredna moć dedukcije zaključuje da je čovek serviser mašina, jer naravno da niko ne poseduje na desetine starih modela veš mašina	und	☐	☐	☐	☐	sr
90923	Да сам поднаслов књиге био бих "Догодовштине једног пушача на почетку трећег миленијума".	sr	☐	☐	☐	☐	sr

Listing

- ✓ Annotation guidelines in the form of the decision tree

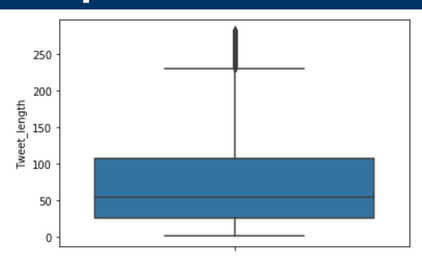


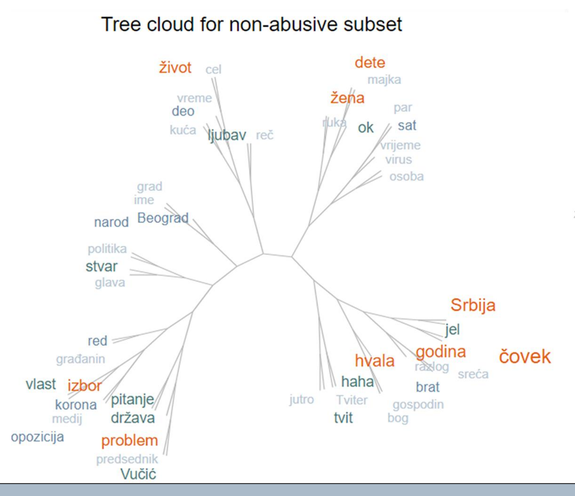
- **Annotation of abusive token spans for lexicon building**
  - ✓ abusive tweet in which two triggers were identified, classified into different categories of abusiveness and assigned abusiveness score.

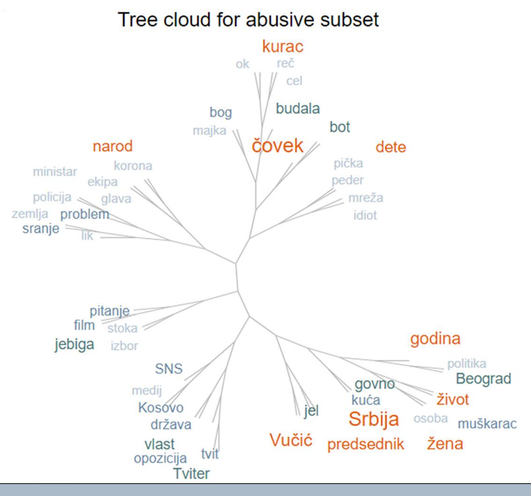
Tweet index	Tweet text	Abusive trigger (lemma)	Abusive class	Score
169879	E vala ja bih prišao da je udavim odmah, da se više ne bori za dah. Majku li vam jebem američku!!!!!!! (eng. Well, I would approach her to drown her right away, so she doesn't fight for her breath anymore. Fuck your American mother !!!!!!!)	udaviti (eng. drown)	Threat	4
169879		jabati (eng, fuck)	Offensive	4

3. 1 Analysis of the twitter corpus

1) distribution of tweets length
2) word frequencies
3) hashtags







3.2 Statistics of the corpus

- ✓ Cohen’s Kappa coefficient for inter-annotator agreement

Category/Subcategory of hate speech	The inter-annotator agreement, accuracy
Offensive/Non-offensive	$\kappa = 0.513$, $accuracy = 0.860$
Profanities	$\kappa = 0.612$, $accuracy = 0.956$
Hate speech	$\kappa = 0.263$, $accuracy = 0.949$
Derogatory speech	$\kappa = 0.370$, $accuracy = 0.895$

Table 1 The inter-annotator agreement per categories of abusive speech

- ✓ 1,416 abusive tweets labelled (out of 6,436)
- ✓ 472 PROF, 273 HS, 843 DS, 169 OTH.
- ✓ 637 tweets assigned to more than one abusive category

3.2 The lexicon of abusive speech

- ✓ triggers for the recognition of abusive language
- ✓ simple words, phrases and figurative speech as indicators.
- ✓ XML and RDF versions

XML version of dictionary

```
<LexicalEntry id="SR0001" lng="sr" pos="n" Probability="0.8">
  <lemma>lopov</lemma>
  <OffensCategories>
    <OffensCategory Severity="4" OffensLevel="0.7">
      <Examples type="Immoral or criminal activities">
        <Example beginIndex="0" endIndex="6" form="lopovu">
            lopovu je mesto u zatvoru</Example>
        <Example beginIndex="19" endIndex="25" form="lopovi">
            svi političari su lopovi</Example>
      </Examples>
    </OffensCategory>
    <OffensCategory Severity="3" OffensLevel="0.4">
      <Examples type="Derogatory words and insults">
        <Example beginIndex="12" endIndex="17" form="lopov"
            type="MWU">ružan kao lopov</Example>
      </Examples>
    </OffensCategory>
  </OffensCategories>
</LexicalEntry>
```

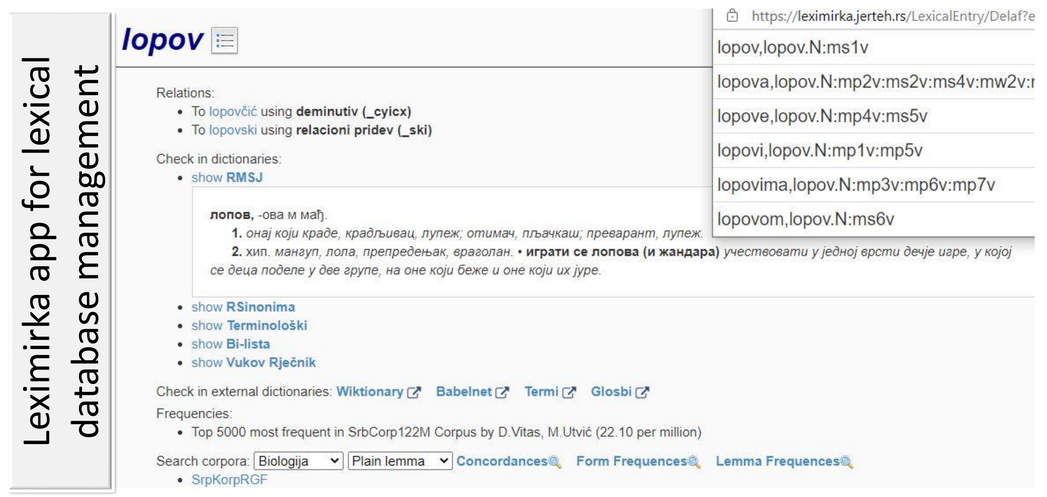
Ontolex Lemon RDF version of dictionary

```
:lopov a ontolex:lopov;
    dct:language <http://id.loc.gov/vocabulary/iso639-1/sr> ;
    lexinfo:partOfSpeech lexinfo:noun;
    ontolex:lexicalForm :lopov-form;
    ontolex:sense :lopov-sense.
:lopov-form a ontolex:Form;
    ontolex:writtenRep "lopov"@sr.
:lopov-sense skos:definition "onaj koji krade, kradljivac,
    lupež; otimač, pljačkaš; prevarant, lupež"@sr;
    ontolex:reference <https://www.wikidata.org/wiki/Q3562775>.
```

Ontolex Lemon (FrAC) RDF version with frequency and attestations

```
# subproperty definition for frequency in twitter corpus
:atvitkoFrequency rdfs:subClassOf frac:CorpusFrequency .
:atvitkoFrequency rdfs:subClassOf [
    a owl:Restriction ;
    owl:onProperty frac:corpus ;
    owl:hasValue <https://app.sketchengine.eu/#
        dashboard?corpname=user%2Franka%2Fatwitco>] .
# frequency assessment (in twitter corpus)
:lopov frac:frequency [
    a :atvitkoFrequency;
    rdf:value "17"^^xsd:int].
# usage examples as attestations
:lopov frac:attestation attestation_1324567;
attestation_1324567 a frac:Attestation ;
    cito:hasCitedEntity <https://app.sketchengine.eu/#
        dashboard?corpname=user%2Franka%2Fatwitco> ;
    rdfs:comment "Immoral or criminal activities" ;
    frac:locus :locus_2415677;
    frac:quotation "svi političari su lopovi." .
:locus_2415677 a :Occurrence ;
    nif:beginIndex 19 ;
    nif:endIndex 25.
```

Leximirka app for lexical database management



4. Discussion and conclusion

- **Moderate inter-annotator agreement possible causes:**
  - ✓ lack of the generally accepted definitions of abusive speech
  - ✓ individual bias of annotators
  - ✓ vague or incomplete annotation instructions
  - ✓ overlapping of abusive speech sub-categories
  - ✓ negation and emoticon change, usage of irony and sarcasm
- **Future research**
  - ✓ extend AbCoSER corpus with new tweets and other sources
  - ✓ improve models for the automatic classification of abusive tweets: hybrid approach (machine learning and lexical resources)
  - ✓ develop ontology for the classification of abusive data
  - ✓ use of VocBench for the lexicon and the ontology will continue
  - ✓ enrich the lexicon with triggers identified during the annotation of abusive token spans and use it to upgrade the AbCoSER corpus

Research is supported by COST Action CA18209-NexusLinguarum “European network for Web-centred linguistic data science” and ELEXIS project , grant number 731015