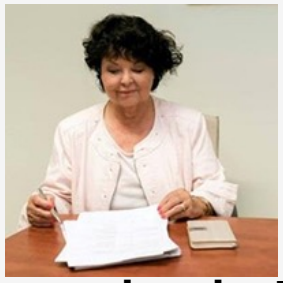
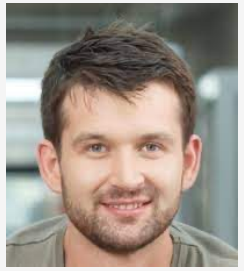
LOD-connected offensive language ontology and tagset enrichment



Barbara Lewandowska-Tomaszczyk
Department of Language and Communication,
State University of Applied Sciences

Chaya Liebeskind
Department of Computer Science,
Jerusalem College of Technology



Slavko Žitnik
University of Ljubljana,
Faculty for computer and information science

Jelena Mitrović
University of Passau, Germany and
The Institute for Artificial Intelligence Research and Development of Serbia

Anna Bączkowska
Department of Glottodidactics and Natural Language Processing, University of Gdansk

Giedre Valunaite Oleškevičiene
Institute of Humanities,
Mykolas Romeris University

Nexus Linguarum

Working group 4, Task 4.1, Use case 4.1.1

Research aims

To identify and critically review existing offensive language tagsets.

To create an ontology basis that is proposed as schema for offensive language identification.

Methodology

Over 60 available corpus data sets are scrutinized and the relevant tagging schemas originally applied compared, while making an attempt to explain semantic differences between particular concepts of the category **OFFENSIVE** in English.

We adopt **a finite set of classes** that cover aspects of offensive language representation, based on the categories originally proposed by Zampieri et al.

Particular offensive words were tested by means of **Sketch Engine (SE) and Thesaurus tools** on a large web-based corpus (19 billion items).

The schemata are juxtaposed and discussed with reference to non-contextual word embeddings **FastText, Word2Vec, and Glove**, originally on a smaller number of offensive terms, in the second phase with four additional offence-related items added, and eventually in the final phase – with an additional inclusion of the item taboo.

The proposed schema enables further comparative research and an effective use of corpora of languages other than English. It will also be applied in building an enriched tagset to be trained and used on new data, with the application of recently developed LLOD techniques

Terminological problem

The main problems for automatic identification of offensive language (as well as other related phenomena e.g. hate-speech, abusive language) has been a lack of consensus concerning the definitions of such phenomena.

The tagsets used for annotating such texts vary.

Proposed concepts

Based on the Sketch Engine data and collocate frequencies we propose the **concept of offensive language as a superordinate category in our system** to cover instances of language which upsets or embarrasses people because of its insulting character.

In terms of socio-cultural standards, **offensive language identifies** a number of hierarchically arranged subcategories such as taunts, directed to ridicule the addressee, references to handicaps, squalid language, which includes allusions to sexual fetishes, slurs which are directed to attack certain culture or ethnicity, homophobia, racism, extremism, crude language which refers either to sexual matters or excrements, disguise which may carry ambiguous meaning, or else direct insults, which contain so-called four-letter words, or provocative language which may cause anger as well the use of taboo-words.

We propose to consider **abusive language to be viewed as a constituent of the superordinate category of offensive language**, characterized in legal terms as harsh, violent, profane, or derogatory language which is directed to violate the dignity of an individual, including profanity and slurs of racial, ethnic, or sexist manner

Following concepts

Hate speech is a similarly fuzzy concept in the previous studies, causing familiar problems in the inter-annotator agreement.

We constrain **hate speech** and define it as a group-targeted or an individual-targeted offense, based on one or more of the identifiable negative stereotypes referring to ethnicity, gender, religion, or ridiculed properties attributed to this group.

Harassment overlaps as part of a larger cyberbullying category.

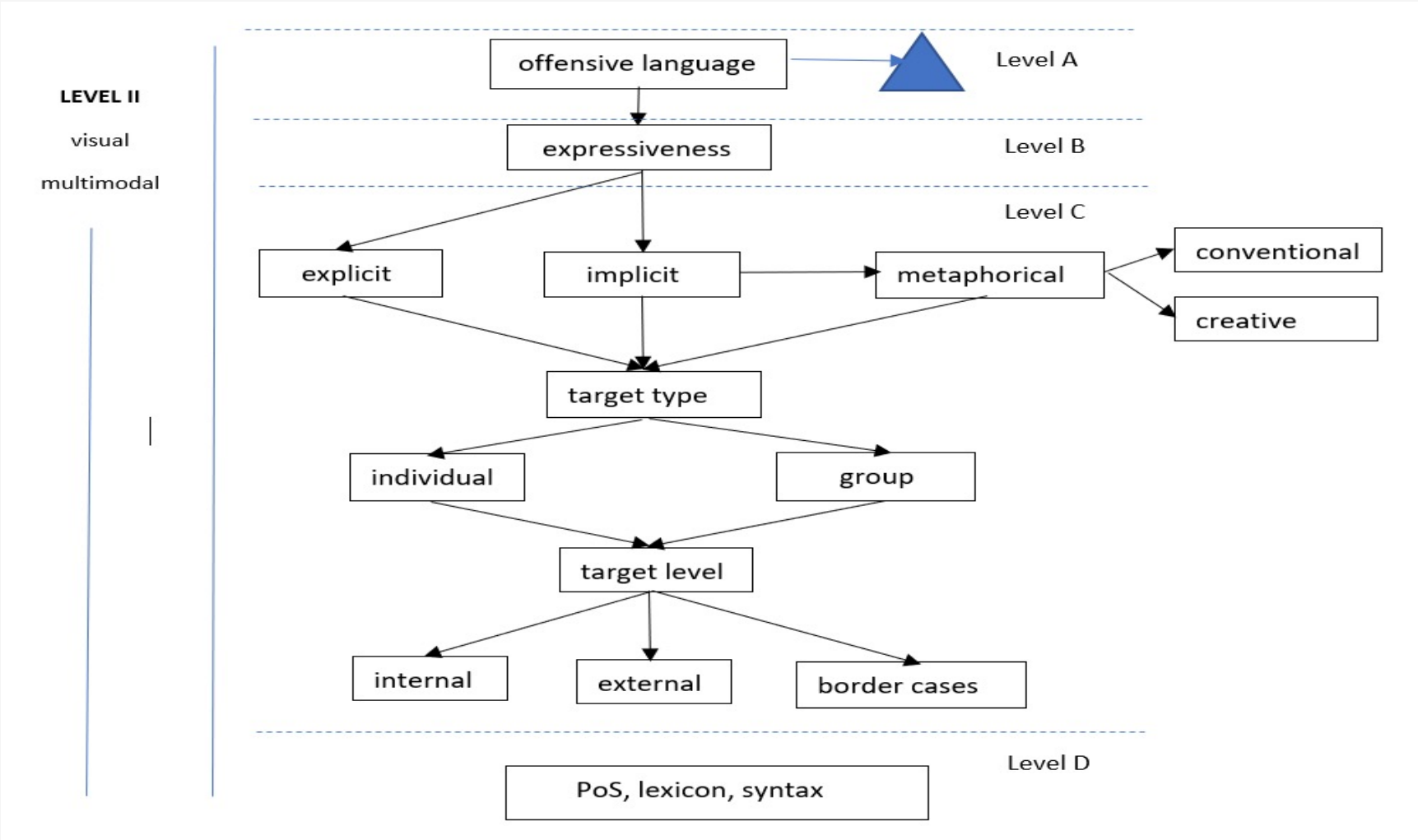
Cyberbullying refers to the online intimidating and threatening content and embraces not only harassment, possible hate speech and other offensive modes but also to massive behaviour of individual or group attackers targeted towards discrimination or exclusion of particular individuals from their groups by various kinds of offence, defaming, deceit, etc., both online as well as by means of the devirtualization of the offence from online to offline real world spaces

Extended offensive language tagset

The extended ontological tagset presented identifies two basic levels

Level I covering sub-levels (A, B) and Level II - sublevels (C, D) – for further multimodal uses, which additionally includes subcategories connected to visual elements (in social media datasets) and, considered for further extension, prosodic elements of speech parameters.

Ontology and Methodology of offensive language (OF)



Sublevels of Level I

Sublevel A: offensive vs. non-offensive. We distinguish offensive from non-offensive language The non-offensive cases are beyond the scope of our research.

Sublevel B: targeted vs. non-targeted. The question of Target is - if there is no identifiable addressee of offense the language is considered an example of self-expression, having, e.g., an exclamatory function (e.g. swear words used to express anger, frustration, pain, etc.).

Sublevels of Level II

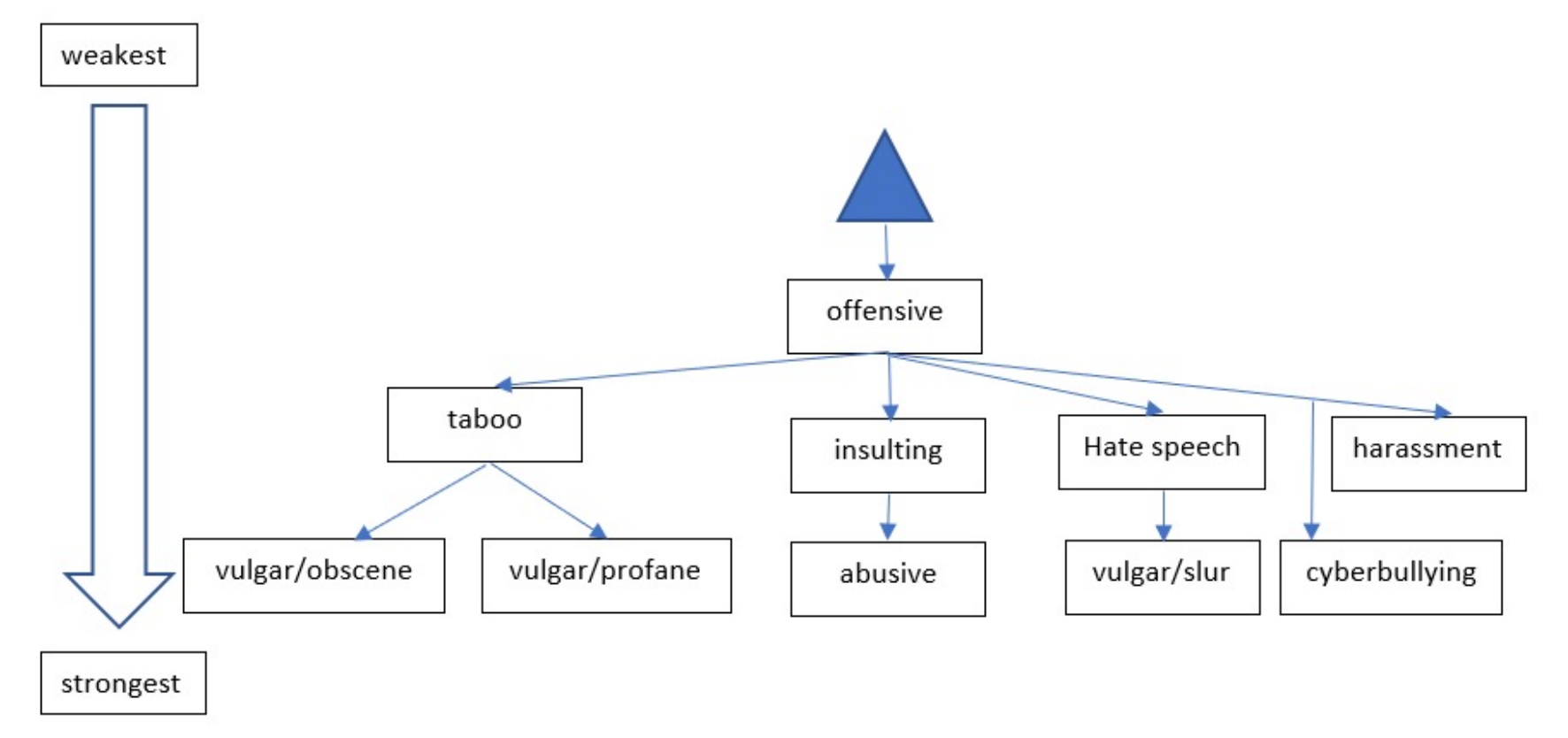
Sublevel C: implicit vs. explicit language. Targeted offensive terms are further divided into implicit or explicit. While implicitness may be encoded by, for example, sarcasm and irony, in which offense is not straightforward, explicitness entails more direct forms of verbal attack.

Sublevel D requires an analysis of morphosyntactic features, i.e. aspects at the word (parts of speech tagging, lexical analysis) and sentence level.

Importantly, sublevel C and sublevel D comprise not only verbal offense but also visual and multimodal forms that combine gestures, proxemics, kinesics, gaze, as well as paralinguistic/prosodic features.

Further we applied Sketch Engine tools: Thesaurus (Th) and Word Sketch (WS) and Word Sketch Difference (WSD) to produce the Typology of *offensive language.*

Typology of offensive language

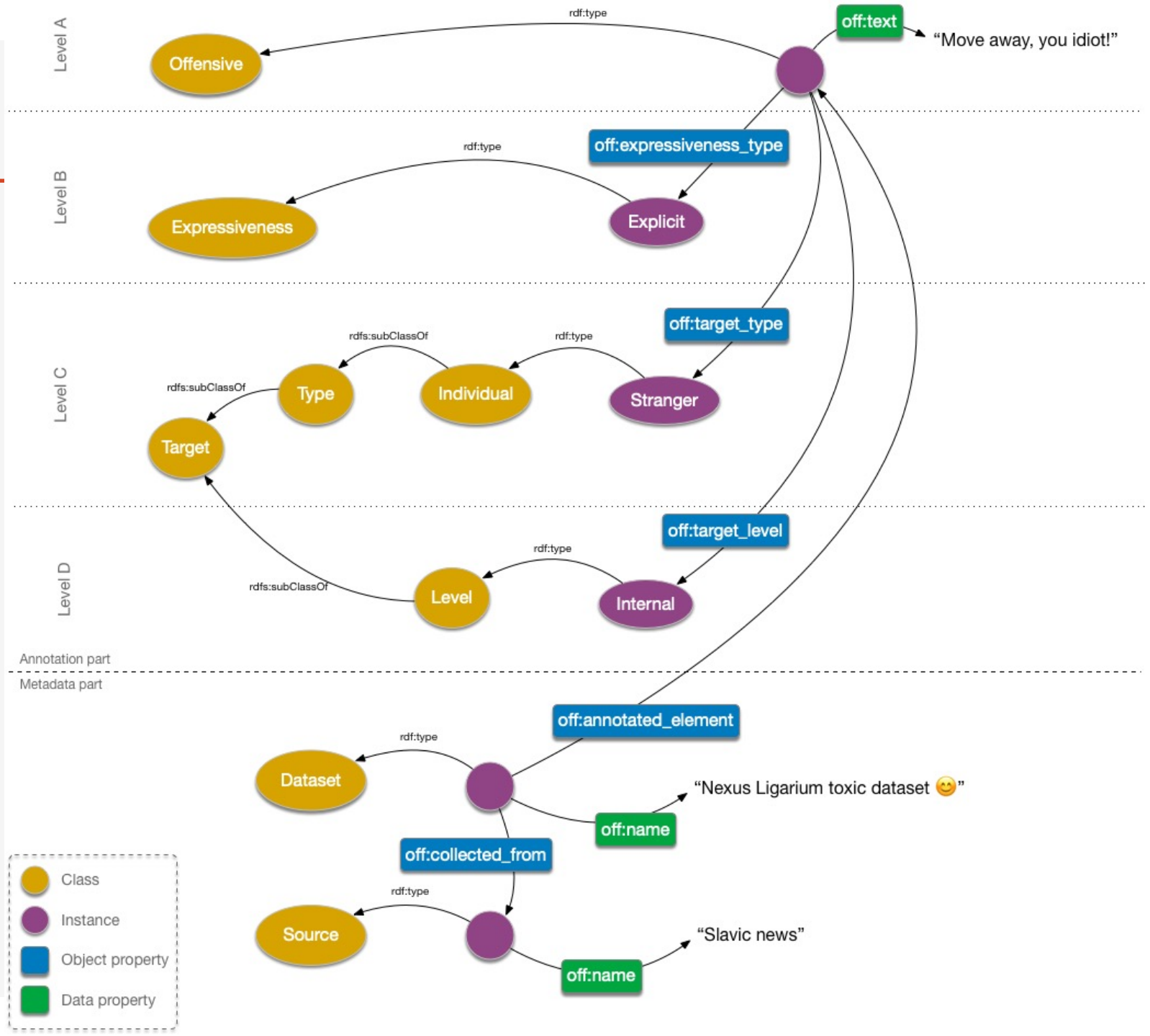


Comment on typology

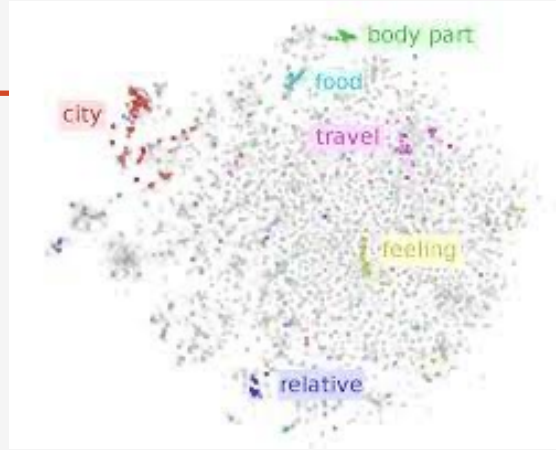
From this study it has transpired that offensive is the weakest term, abusive is the strongest, and insulting stands mid-way.

Based on the above categorization we encoded the schema into a generic ontology.

Proposed ontology schema along with an example



Validation of offensive language terms using pre-trained word embeddings



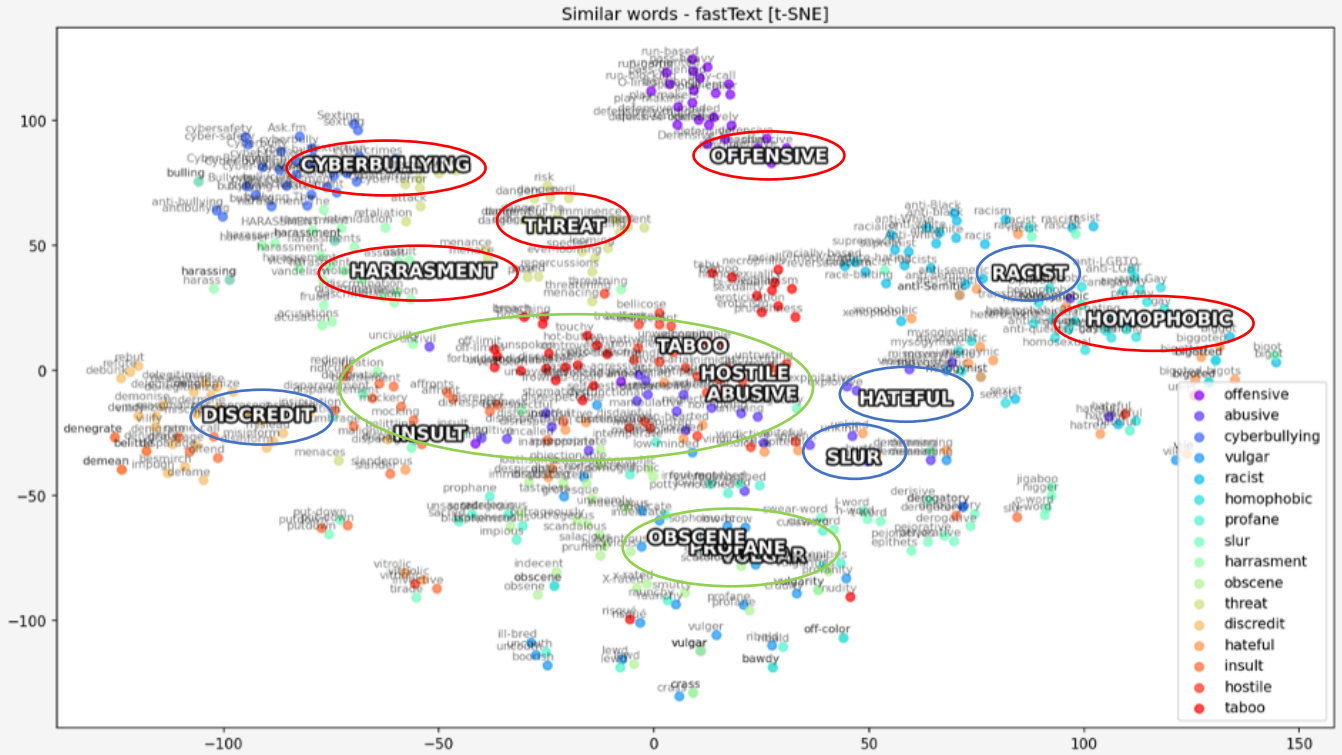
- We support our categorization using 3 pre-trained word embeddings:
  - **Word2Vec**
    - Pre-trained Google News corpus having 3 million 300-dimensional English word vectors
  - **Glove**
    - Pre-trained Wikipedia and Gigaword corpus having 0.4 million 300-dimensional word vectors
  - **FastText**
    - Pre-trained CommonCrawl and Wikipedia having 2 million 300-dimensional English word vectors

Validation of offensive language terms using pre-trained word embeddings

- **16 keywords** in their lemma forms for the analysis were selected
  - offensive, abusive, cyberbullying, vulgar, racist, homophobic, profane, slur, harrasment, obscene, threat, discredit, hateful, insult, hostile and taboo.
  - **30 neighbouring words** were retrieved for each keyword
    - Words that contain the keyword or its lemma, or else a stem as a substring were omitted
- Multiple types of visualization techniques were performed (**PCA, MDS, t-SNE**)
  - Due to visualization techniques' limitations
    - For example, t-SNE visualization:
      1. cluster sizes seem to play no particular role
      2. neither do inter-cluster distances
      3. random noise may present some non-random significance

Word embeddings versus linguistic analysis categories (t-SNE)

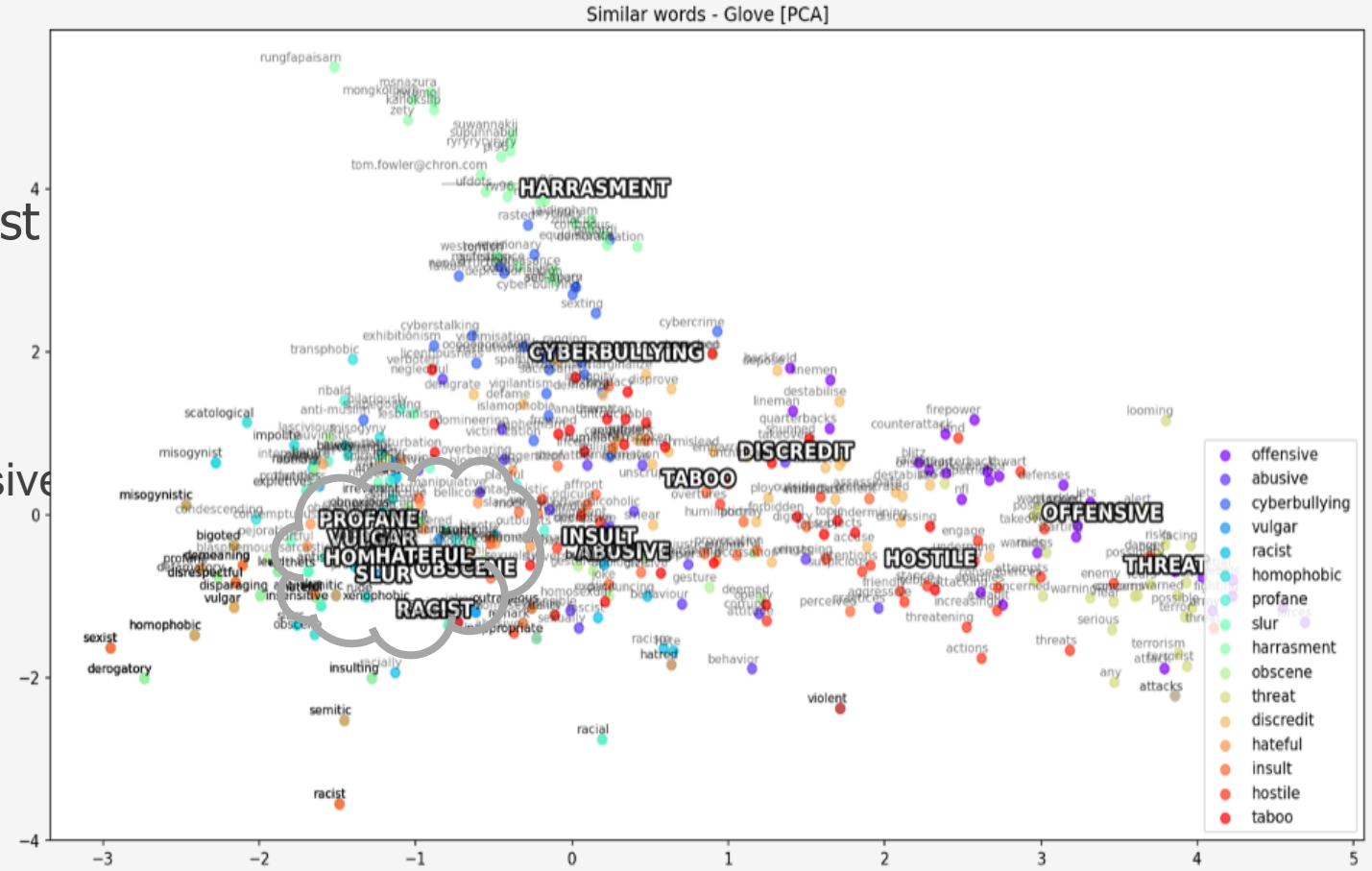
- **FastText**
  - Based on character n-grams
  - Shows the most discrete clusters



- Discrete items
- Relatively well-delineated
- Concepts with the least clear-cut and spread widely
- Concepts which show overlap with other elements:
  - insult (intermingles with abusive and slur), abusive (intermingles with hostile, insult, hateful), profane (intermingles with obscene and vulgar)

Detailed analysis of word embeddings (PCA)

- **Glove**
  - **Cloud** of concepts
    - profane, vulgar, hateful, slur, obscene, homophobic and racist
    - Remaining terms show distancing between the represented notions
- Certain overlapping of taboo, hostile, abusive slur, hateful, obscene, profane, vulgar, and insult
  - Could be also supported by:
    - FastText
    - SketchEngine word sketch representations



Future work

- In-depth linguistic analysis of ***offensive*** language categories
- Training of a new set of non-contextual word embeddings ***(Word2Vec, FastText)***
- BERT-based finetuning ***(HateBERT)*** on specific categories and cross comparison

	1-severe_toxic	2-hate_speech	7-abusive	17-covertly-aggressive	15-offensive	29-sexism	6-cyberbullying	18-spam	9-religious	19-harrasment	1-obscene	1-insult	9-homophobic	9-racist	27-vulgar	1-threat	3-profane	AVG
1-severe_toxic	**28.5**	68.3	82.8	26.3	49.2	29.1	36.5	7	2.8	39.6	65.5	62.6	14.1	36.7	67.6	11.4	54.4	40.1
2-hate_speech	36.3	**97.7**	82.4	25.5	49.6	29.1	50.6	1.7	2.9	43.3	69	67.5	14.6	39.9	61.8	14	55.6	43.6
7-abusive	12.3	56.8	**87.6**	29.7	63	39.1	31.6	10.5	3	42.8	45.9	43.9	14.7	40.5	73	4.4	42	37.7
17-covertly-aggressive	13.9	35.1	30.8	**90.2**	33.8	30.3	25.6	0.9	2.3	21.5	32.3	34.5	6.8	8.5	33.4	4.6	24.5	25.2
15-offensive	5.1	54.7	80.9	38.2	**82.2**	48.2	26.1	8.2	3	43.5	24.9	23.7	14.6	39.7	73.4	1.8	37.1	35.6
29-sexism	4.3	53.8	46.4	33.5	42.8	**88.6**	30.3	10.4	2.8	39.9	16	15.8	13.5	32.5	35.9	1.1	27.5	29.1
6-cyberbullying	4.9	48.8	58.8	27.7	51	48.3	**33.8**	30.5	3	41.1	22.1	21.6	14.8	41.4	57.7	1.7	28.9	31.5
18-spam	1.1	14.6	27.4	10.1	19.7	13.9	6.4	**53.9**	2.7	18.5	5.1	4.3	12.5	28	23.8	0.3	13.2	15.0
9-religious	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0.0
19-harrasment	37.4	69.8	36.7	15.7	32.2	20.2	40.2	3.7	2.9	**75.4**	54.2	55.8	13.5	33.7	29.6	18	44.1	34.3
1-obscene	1.9	40.7	50.2	22.4	49.6	41.6	11.5	31	3	41.2	**10.8**	10.2	14.9	41.7	72.5	0.7	27.1	27.7
1-insult	2.1	41.3	50.7	25.3	50.1	42.4	11.5	30.5	3	41.4	11.5	**10.8**	14.9	41.7	72.3	0.7	27.5	28.1
9-homophobic	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0.0
9-racist	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0.0
27-vulgar	4.9	59.4	72.4	37.1	56.8	43.2	26.1	4	3	41.5	23.5	23.1	13.8	33.4	**97.2**	1.7	38.5	34.1
1-threat	3.5	50.6	62.9	28.3	49.5	38.8	14.5	17.8	2.9	41.4	17.2	16	14.1	36.7	70.4	**1.3**	34.6	29.4
3-profane	65.4	72.7	86.3	6.5	29.4	13	46.1	1.3	2.7	32	75.6	70.7	13.4	32.9	56.3	26.3	**82.8**	42.0

HateBERT preliminary results