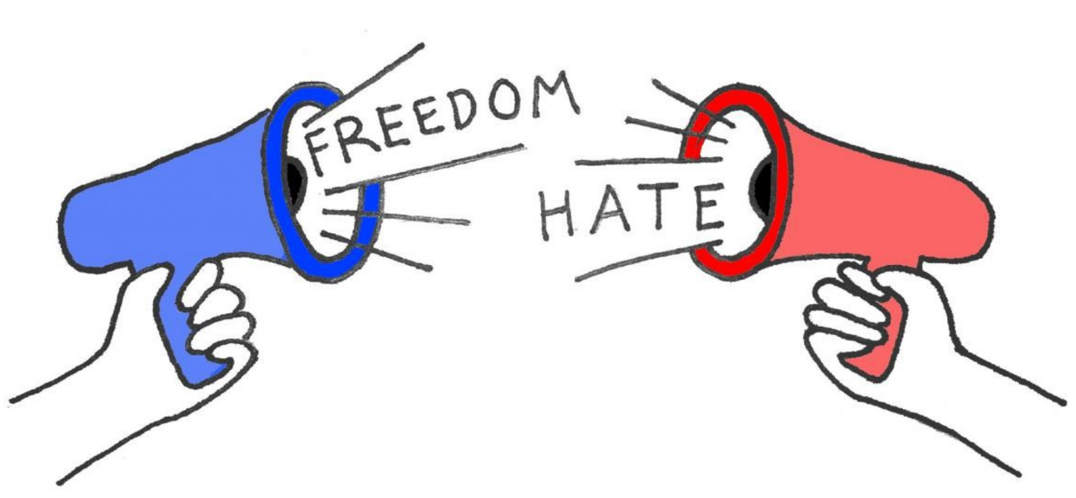
Univerza *v Ljubljani*
Fakulteta *za računalništvo*
in informatiko

Hate speech paraphraser

Author: Drejc Pesjak
Mentor: prof. dr. Zoran Bosnić
Comentor: prof. dr. Marko Robnik Šikonja

Introduction

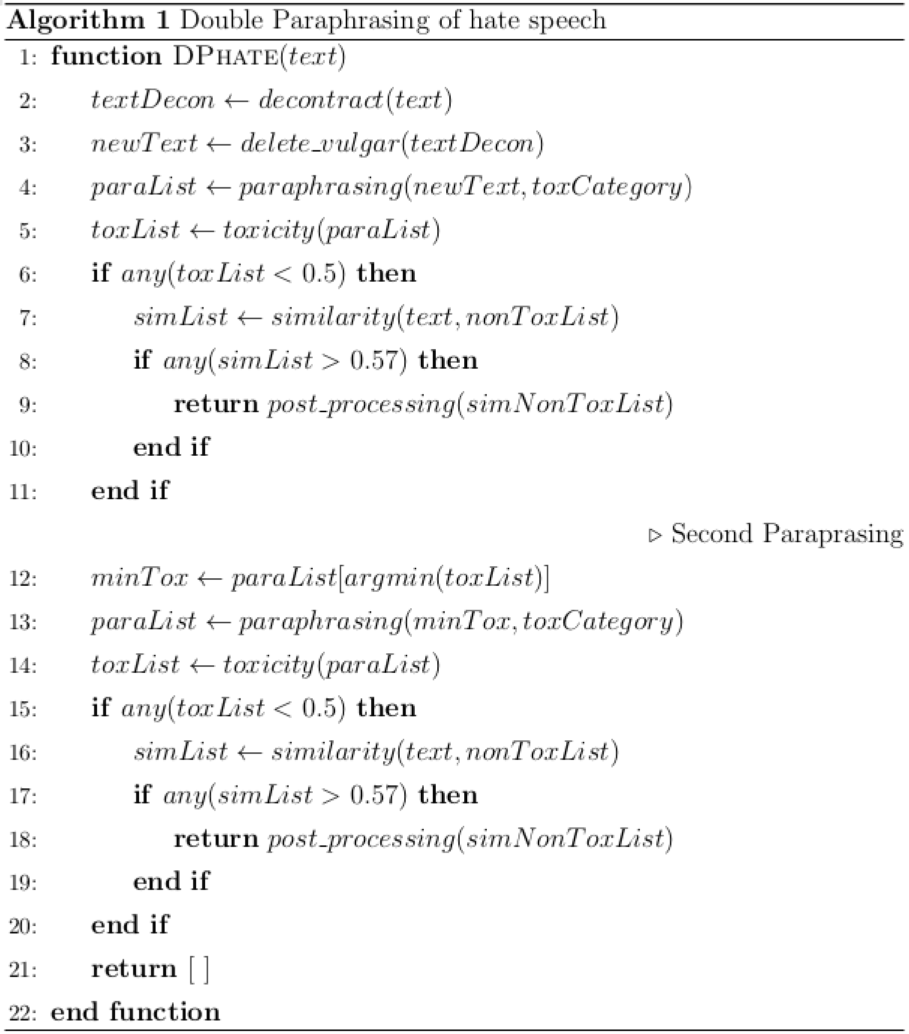
- Hate speech fills the world wide web.
  - Social media, news sites, forums …
- Moderators remove hate speech from these platforms.
  - With the aid of hate speech detection software.
- Such censorship is against freedom of speech.



The Proposed solution

- A system that takes a hateful comment as input and outputs a non hateful one.
  - Beneficial to all parties involved.

System DPhate



Algorithm 1 Double Paraphrasing of hate speech

```
function DPhate($text$)
    $textDecon \leftarrow decontract(text)$
    $newText \leftarrow delete\_vulgar(textDecon)$
    $paraList \leftarrow paraphrasing(newText, toxCategory)$
    $toxList \leftarrow toxicity(paraList)$
    if $any(toxList < 0.5)$ then
        $simList \leftarrow similarity(text, nonToxList)$
        if $any(simList > 0.57)$ then
            return $post\_processing(simNonToxList)$
        end if
    end if
                                                    ▷ Second Paraprasing
    $minTox \leftarrow paraList[argmin(toxList)]$
    $paraList \leftarrow paraphrasing(minTox, toxCategory)$
    $toxList \leftarrow toxicity(paraList)$
    if $any(toxList < 0.5)$ then
        $simList \leftarrow similarity(text, nonToxList)$
        if $any(simList > 0.57)$ then
            return $post\_processing(simNonToxList)$
        end if
    end if
    return [ ]
end function
```

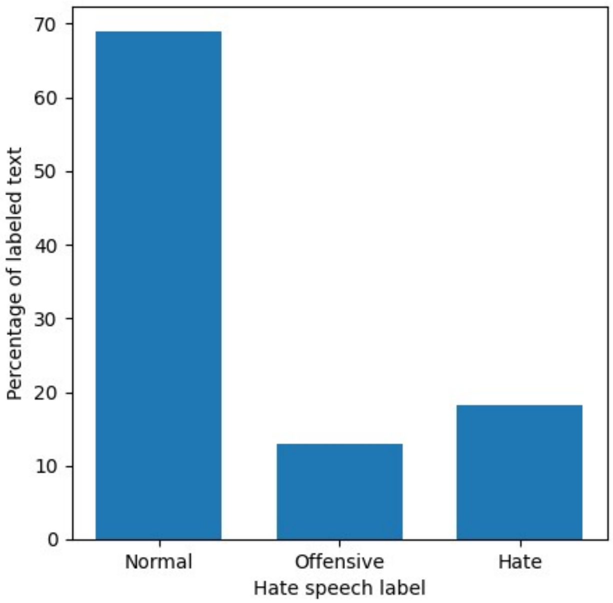
- Hatexplain

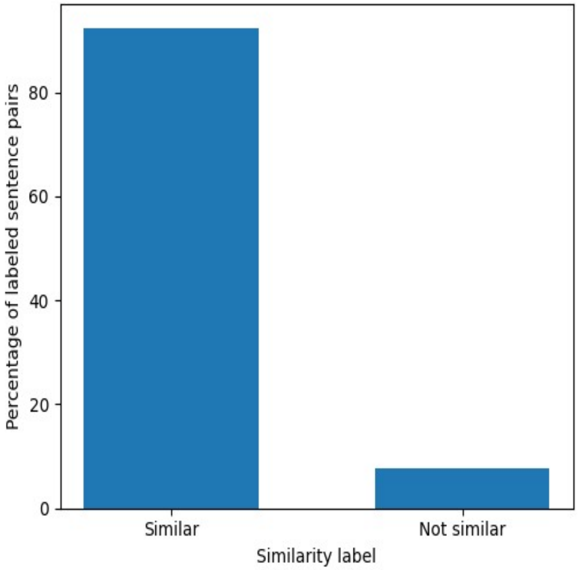
- PEGASUS
- Detoxify
- BERT embeddings + cosine similarity

Evaluation

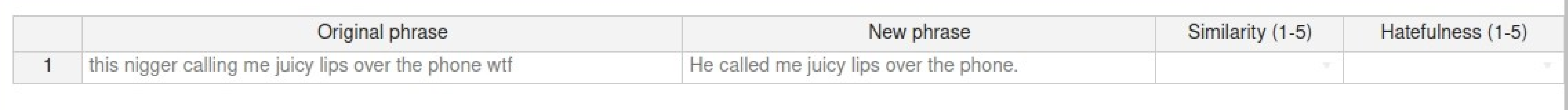
- Two measures of quality:
  - Hatefulness / toxicity
  - Semantic similarity
- Two approaches to evaluation:
  - An automatic system
  - Human evaluators

- Ensemble method, stacking:
  - 3 hate speech detection models
  - 1 meta learner (logistic regression)
- Used similarity model simCSE.
- 84.37% of input sentences got at least one acceptable transformation.



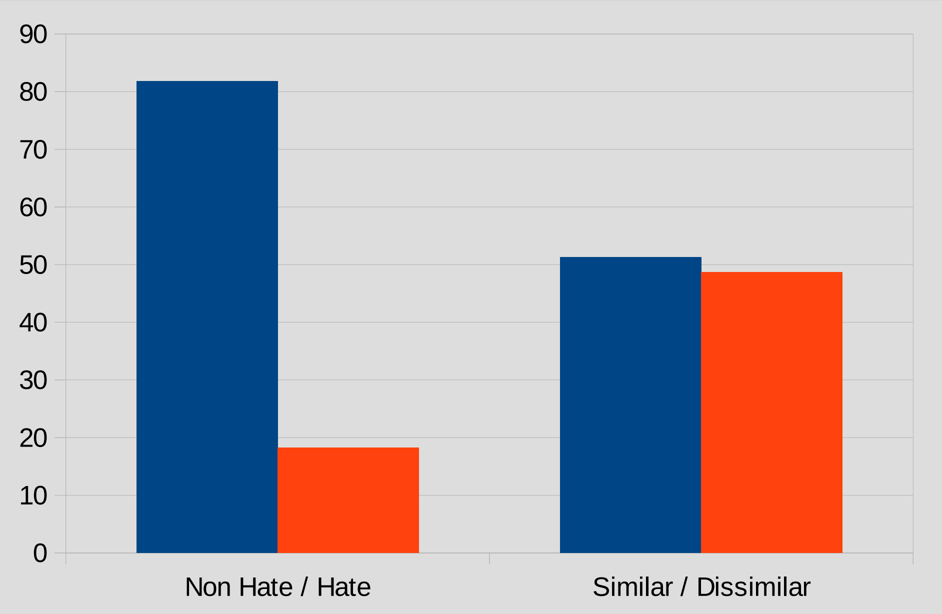


microWorkers



	Original phrase	New phrase	Similarity (1-5)	Hatefulness (1-5)
1	this nigger calling me juicy lips over the phone wtf	He called me juicy lips over the phone.		

- 61 people from English speaking countries evaluated 876 examples.
- Each was labeled by 3 different evalutors.
- 67.90% of input sentences got at least one acceptable transformation.



Good examples

“the kike shill psyop ricky vaughn has deleted his post where he stated white identity does not exist so i will post the screen shot here now ask ricky tkr get gassed faggot”

“I will post a screen shot of the deleted post where ricky vaughn stated white identity does not exist.”

Good examples

“that a retarded non argument the bible is also a best seller in communist china”

“In communist china, the bible is a best seller.”

Good examples

“ why do not you respond to my posts you white nigger lying sack of shit i called you out for being a fed 201 days before you published it on your website faggot Gemma ”

“ Why don't you respond to my posts, I called you out for being a fed 201 days before you published it on your website. ”

Conclusion

- Hate speech on social media, the problem of censorship and limits on freedom of speech.
- The proposed approach of removing hate from the hateful comments.
- System DPhate using double paraphrasing, with hate speech detection and semantic similarity measures.
- Results, where 85% or 68% of input sentences were improved.