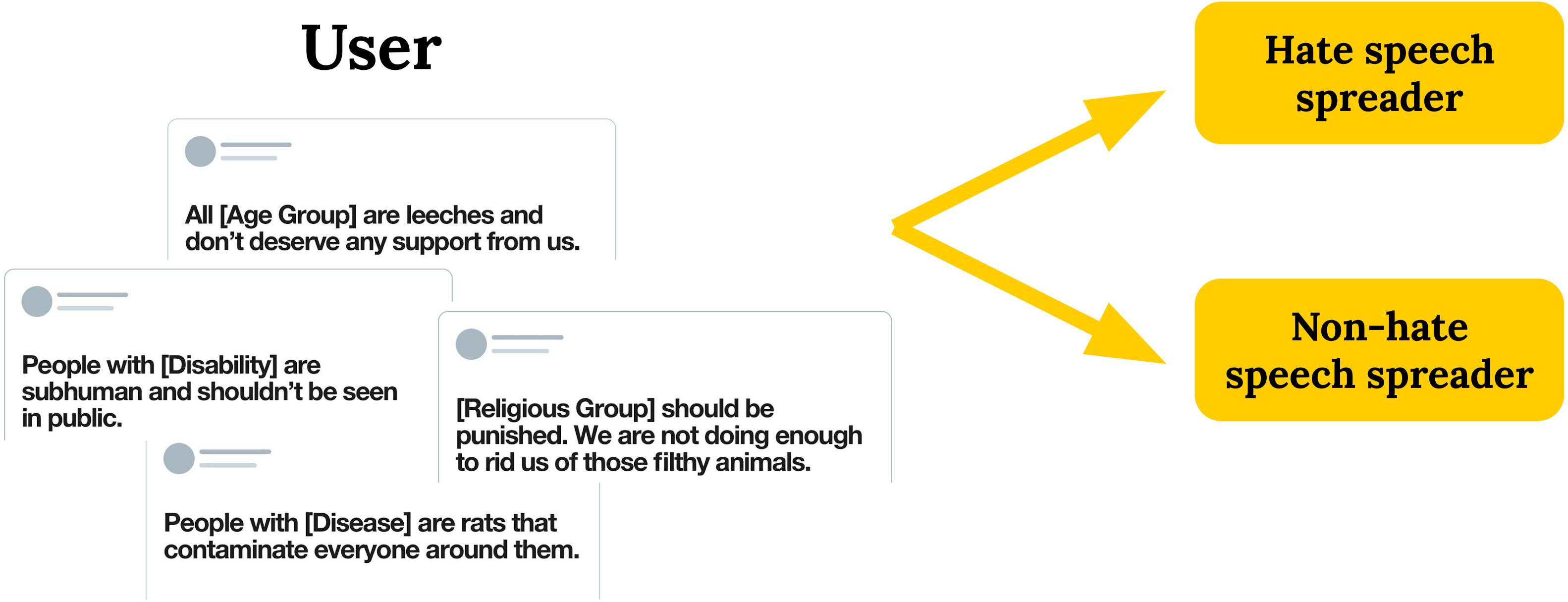
Hate Speech Spreader Detection

Miki Mizutani, Anna Fabris, Leonidas Gee

Hate Speech Spreader Detection



User

All [Age Group] are leeches and don’t deserve any support from us.

People with [Disability] are subhuman and shouldn’t be seen in public.

[Religious Group] should be punished. We are not doing enough to rid us of those filthy animals.

People with [Disease] are rats that contaminate everyone around them.

Hate speech spreader

Non-hate speech spreader

Dataset:
PAN-AP-2021 En

- 60000 tweets from 300 different Twitter users.
- Uniformly balanced with 150 hate speech spreaders users and 150 non-hate speech spreaders.
- Partial preprocessing carried out.
- Access to only 40000 tweets from 200 users.

Preprocessing

- Remove the **dataset specific terms** and other html leftovers.
- Expand contractions (e.g. **you're** → **you are**).
- Normalise sequences of **at least 3** repeated characters with a maximum of two letters (e.g. hiiiiii → hii).
- Remove **numbers** and **punctuations**.
- Transform **emojis** into their aliases.
- Remove extra **white spaces** and any **left or right spacing**.

Model

SVM

TF-IDF with no stop words removal.

BiGRU

GloVe embeddings with random vectors for out-of-vocabulary tokens.

BiLSTM

GloVe embeddings with random vectors for out-of-vocabulary tokens.

BERTweet

Embeddings from pre-trained BERTweet on sentiment analysis.

Hyperparameter Tuning

SVM

‘C’ (between 0 and 1 → **0.59**),

‘kernel’ (‘poly’, **‘rbf’**, ‘sigmoid’)

BiLSTM

‘LSTM units’ (between 64 and 320 with step 64 → **128**)

BiGRU

‘GRU units’ (between 64 and 320 with step 64 → **256**)

BERTweet

Learning Rate = 2e-05

Batch = 32

Training

SVM

Trained using 5-fold cross validation.

BiLSTM

Trained for 10 epochs with early stopping.

BiGRU

Trained for 10 epochs with early stopping.

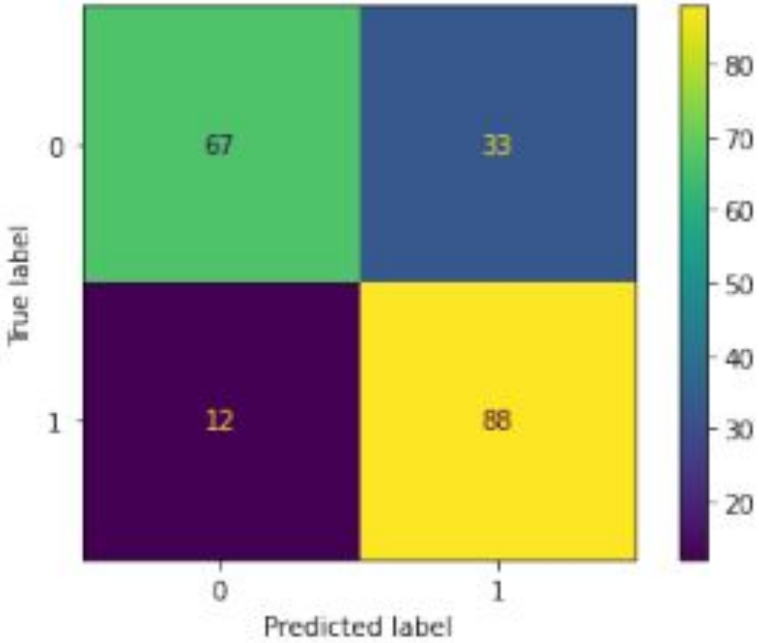
BERTweet

Fine-tuned using 10 epochs with early stopping.

Results

Method	Accuracy
TF-IDF + SVM	76.0
GLoVe + BiLSTM	64.0
GLoVe + BiGRU	67.0
BERTweet	78.0

Confusion Matrix of BERTweet



Potentials Improvements

- N-grams.
- CNNs.
- Combining BERT with SVM.
- Exploiting relationships between users.
- Modifying the majority voting threshold.