Transfer Learning for Hate Speech Detection in Social Media

This document is accompanying the submission *Transfer Learning for Hate Speech Detection in Social Media*. The information in this document complements the submission, and it is presented here for completeness reasons. It is not required for understanding the main paper, nor for reproducing the results.

Supplemental figures

This section presents the supplemental figures mentioned in the main text.

- Figs. 6a and 6b show the confusion matrixes made by Davidson baseline and t-HateNet on the DAVIDSON data set;
- Fig. 6c Prediction performances with limited amounts of training data on the DAVIDSON dataset;
- Fig. 7a shows the results of the grid search for the optimal size of the hidden state vector;
- Fig. 7b shows the results of the grid search for the optimal learning batch size;
- Figs. 8b and 8c depict the Map of Hate constructed by HateNet using general embeddings and task-specific embeddings, respectively. Visibly, when using general-purpose embeddings, the tweets belonging to different classes do not appear distinguishably separated, whereas they appear clustered when using task-specific embeddings. This highlights the impact of performing domain adaptation for textual embedding;

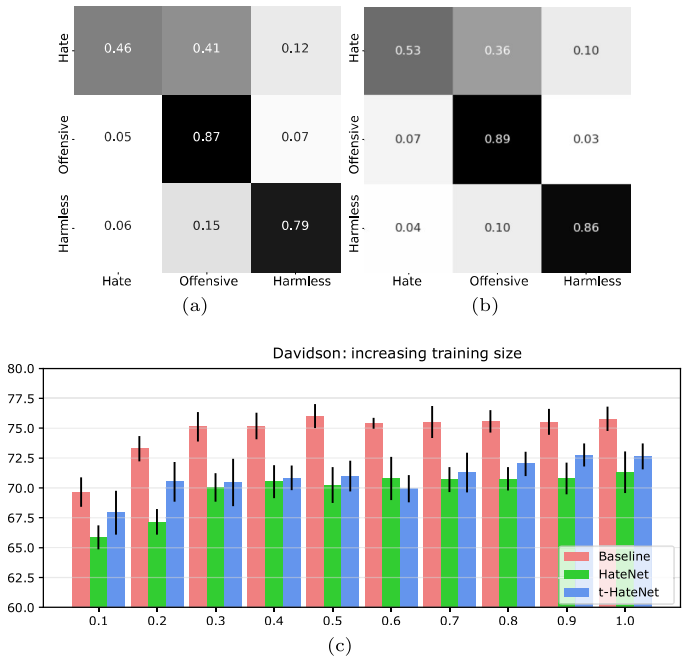

Fig. 6: **(a)(b)** Confusion matrix for one prediction made by the baseline (a) and by t-HateNet (b) on the DAVIDSON data set. **(c)** Prediction performances with limited amounts of training data on the DAVIDSON. The x-axis show the percentage of the training set used for training, the y-axis shows the macro-F1 measure. Each bar shows the mean value over 10 runs, and the standard deviation.

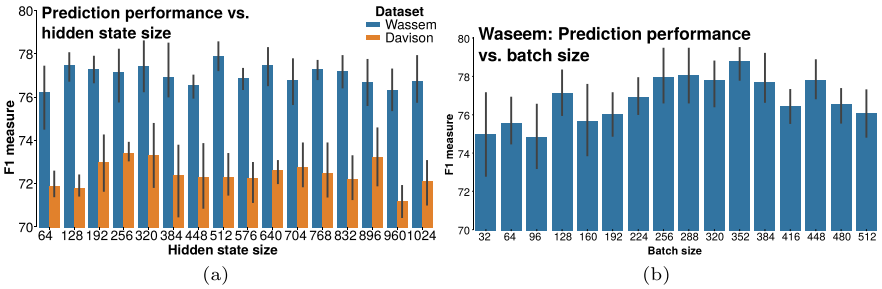

Fig. 7: Grid search for optimal hyper-parameter values: the hidden state size of the bi-LSTM (a) and the learning batch size (b).

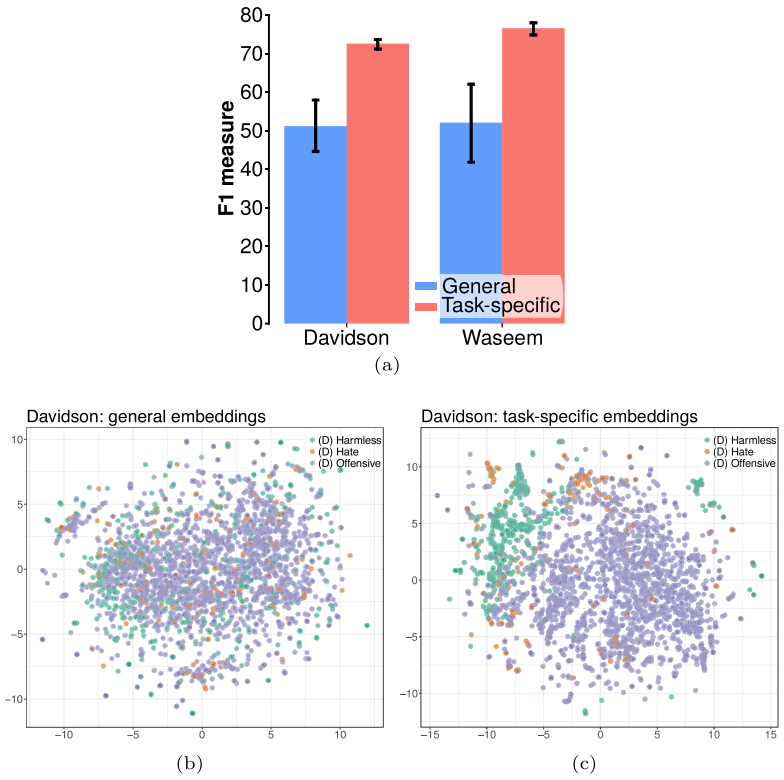

Fig. 8: **The impact of building task-specific word embeddings. (a)** The prediction performance (mean macro-F1 and standard deviation) of HateNet with general-purpose, and with task-specific embeddings. **(b)(c)** The Map of Hate constructed using general-purpose embeddings (b) and using task-specific embeddings (c).