# Planning Report Group 13

Fabian Forsman, Thea Granström, Angelica Hörngren, Conrad Olsson
Chalmers University of Technology

## I. Problem

Today social media platforms such as X (previously Twitter) face a constant challenge with toxic and hate speech content, which can harm users and affect many people. While supervised models like fine-tuned BERT classifiers have shown strong performance, they require large annotated datasets, which are costly and time-consuming to produce. Our project therefore investigates whether unsupervised and semi-supervised methods can effectively classify toxic posts compared to other approaches. This is interesting because it evaluates the possibility of reducing annotation costs while maintaining good performance. It could also possibly detect new types of hate speech that would need to be annotated to be discovered in a fully supervised model. The results could highlight trade-offs between annotation effort, model performance, and generalization.

## II. Datasets

We plan to use the **Hate Speech and Offensive Language Dataset** [1], available via Kaggle, which contains almost 25 000 labeled tweets categorized as hate speech, offensive language, or neither. In addition, we will explore unlabeled tweet datasets to test unsupervised pre-training and semi-supervised approaches. If we have time, we may compare with alternative datasets such as **Toxic Comment Classification Challenge** [2] and **twitterhate** [3], also available on Kaggle.

## III. Available code

There is existing code for training transformer-based models to get sentence embeddings, which are dense vector representations of text. One approach this project can leverage from is Sentence-BERT (SBERT), for which a number of pre-trained models are available that can be applied directly in this project to extract feature vectors. One such model is *all-MiniLM-L6-v2*, which is very fast and lightweight while still providing good general-purpose embeddings for short texts such as tweets. This model outputs a 384-dimensional dense vector space [4]. Another model, *paraphrase-MiniLM-L6-v2*, is also lightweight and produces 384-dimensional embeddings, but it is specifically optimized for paraphrase detection. This makes it particularly suitable for grouping semantically similar expressions, for example distinguishing between "toxic" and "non-toxic" content [5]. For higher accuracy, but at the cost of increased computational requirements, the model *all-mpnet-base-v2* can be used, which outputs 768-dimensional embeddings and generally outperforms the MiniLM variants [6].

Once feature vectors are obtained, there are several clustering methods available in the *scikit-learn* library that can be used to group the embeddings. One such method is k-means, which is efficient and effective when the number of clusters is known in advance [7]. Another widely used method is DBScan, which is capable of finding clusters of arbitrary shape and can also identify outliers in the data, without setting the number of clusters beforehand [8]. These are just some of the available cluster methods.

If the project group chooses to explore semi-supervised methods there are already a lot of models available for this task as well. One approach is *MixText*, which is a semi-supervised learning method designed specifically for text. MixText uses data augmentation through *MixUp* in the embedding space and has open-source code that can be trained on datasets such as hate speech and offensive language [9].

In this project some of the available code might be utilized but each project group will decide what parts of the project they want to implement themselves. An example of this could be to implement a self-designed clustering method or fine-tuning the available sentence-embedding models.

## IV. Relevant papers

- **Attention is all you need** [10].
- **Deep Learning** [11].
- **Unsupervised Learning** [12].
- **Mean teachers are better role models** [13].
- **Text Classification Techniques: A Literature Review** [14]. Used to differentiate the Supervised Learning (SL), Semi-SL, and unsupervised learning with various text classification techniques, such as k-means.
- **Deep Learning for Hate Speech Detection in Tweets** [15]. Used for comparing different techniques of word embeddings for similar words.
- **The text mining handbook** [16]. A comprehensive theory and architecture of text mining.

## V. Evaluation

Extrinsic evaluation will be applied to the model, since although the model is unsupervised and does not use labels during its training, the outputs can still be assessed against the ground-truth labels from the dataset. Other labeled datasets can also be used for this purpose. The model will therefore be trained on unlabeled data but evaluated on how well its discovered structure, such as clusters, corresponds to the true labels (e.g., hate against non-hate). For example, tweets can be clustered into groups k = 2 and the resulting assignments

can then be compared to the true labels using metrics such as the Adjusted Rand Index (ARI).

## VI. Time plan

- Planning report: 9 October
- Project poster: 23 October
- Project report: 27 October

The project planning group will begin with writing the project planning report during week 40, and will during week 41 move on to data preprocessing and initial model training together. Once the model baseline has been established, the group will split up into the assigned project groups to test different approaches such as clustering methods, embedding models or semi-supervised techniques. These approaches will be evaluated on their performance during week 42. During week 43, the results of the groups will be assembled in the project poster and the project report, both of which should be complete by the end of that week.

## References


[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2017.

[2] Jigsaw/Google, "Toxic comment classification challenge." Kaggle, 2018. https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge [Accessed: Oct. 2, 2025].

[3] D. Kunnool, "twitterhate dataset." Kaggle, 2020. https://www.kaggle.com/datasets/dilnakunnool/twitterhate/data [Accessed: Oct. 2, 2025].

[4] Huggingface, "all-minilm-l6-v2."

[5] Huggingface, "sentence-transformers/paraphrase-minilm-l6-v2."

[6] Huggingface, "all-mpnet-base-v2."

[7] scikit learn, "Kmeans."

[8] scikit learn, "Dbscan."

[9] J. Chen, Z. Yang, and D. Yang, "Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[12] Z. Ghahramani, *Unsupervised Learning*, pp. 72–112. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[13] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2018.

[14] M. Thangaraj and M. Sivakami, "Text classification techniques: A literature review," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 13, pp. 117–135, 2018.

[15] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, (Geneva, Switzerland), pp. 759–760, International World Wide Web Conferences Steering Committee, 2017.

[16] R. Feldman and J. Sanger, *The text mining handbook*. Cambridge, England: Cambridge University Press, Dec. 2006.