# Label Propagation-Based Semi-Supervised Learning for Hate Speech Classification

**Ashwin Geet D'Sa**[1], **Irina Illina**[1], **Dominique Fohr**[1], **Dietrich Klakow**[2], **Dana Ruiter**[2]
[1]Université de Lorraine, CNRS, Inria, LORIA
[2]Spoken Language System Group, Saarland University

## Abstract

Research on hate speech classification has received increased attention. In real-life scenarios, a small amount of labeled hate speech data is available to train a reliable classifier. Semi-supervised learning takes advantage of a small amount of labeled data and a large amount of unlabeled data. In this paper, label propagation-based semi-supervised learning is explored for the task of hate speech classification. The quality of labeling the unlabeled set depends on the input representations. In this work, we show that pre-trained representations are label agnostic, and when used with label propagation yield poor results. Neural network-based fine-tuning can be adopted to learn task-specific representations using a small amount of labeled data. We show that fully fine-tuned representations may not always be the best representations for the label propagation and intermediate representations may perform better in a semi-supervised setup.

## 1 Introduction

Online hate speech is anti-social communicative behavior and targets minority sections of the society based on religion, ethnicity, gender, etc. (Delgado and Stefancic, 2014). It leads to threat, fear, and violence to an individual or a group. As monitoring these contents by humans is expensive and time-consuming, machine learning-based classification techniques can be used. The last few years have seen a tremendous increase in research towards hate speech classification (Badjatiya et al., 2017; Nobata et al., 2016; Del Vigna et al., 2017; Malmasi and Zampieri, 2018). The performance of these classifiers depends on the amount of available labeled data. However, in many real life scenarios there is a limited amount of labeled data and abundant unlabeled data. In need of data, different data augmentation techniques based on synonym replacement (Rizos et al., 2019), text generation (Rizos et al., 2019; Wullach et al., 2020), back translation (Aroyehun and Gelbukh, 2018), knowledge graphs (Sharifirad et al., 2018), etc, have been employed for up-sampling the training data in the field of hate speech classification. The performance gain by these techniques is small, and they fail to take advantage of the available unlabeled data.

Semi-supervised learning is a technique to combine a small amount of labeled data with a large amount of unlabeled data during training (Abney, 2007), intending to improve the performance of the classifiers. Label propagation (Xiaojin and Zoubin, 2002) is a graph-based semi-supervision technique analogous to the k-Nearest-Neighbours algorithm. It assumes that data points close to each other tend to have a similar label. These algorithms rely on the representation of data points to create a distance graph which captures their proximity.

Recently, pre-trained word embeddings such as Word2Vec, fastText, Global Vectors for Word Representation (GloVe) have been used for representing words for hate speech classification (Waseem et al., 2017; Badjatiya et al., 2017). Furthermore, pre-trained sentence embeddings such as InferSent, Universal Sentence Encoder, Embeddings from Language Models (ELMo) have been used for the task of hate speech classification (Indurthi et al., 2019; Bojkovskỳ and Pikuliak, 2019). These pre-trained sentence embeddings are generic representations and are unaware of task-specific classes. Transforming the pre-trained sentence embeddings to task-specific representations can be helpful for label propagation, and our work explores this direction. The contributions of this article are:

- evaluation of label propagation based semi-supervised learning for hate speech classification;
- comparison of label propagation on pre-trained and task-specific representations learned from a small labeled corpus.



*Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 54–59

ISBN 978-1-952148-66-8

The rest of the paper is organized as follows: Section 2 describes label-propagation based semi-supervised learning for hate speech classification. Section 3 describes the experimental setup. Results and discussions are presented in Section 4.

## 2 Semi-supervised Learning

In this section, we briefly describe sentence embeddings and the label propagation algorithm. This is followed by our methodology for semi-supervised training for hate speech classification.

### 2.1 Sentence Embeddings

Sentence embeddings are fixed-length vector representations that capture the semantics of the sentence. These embeddings are learned from large unlabeled corpora. Similar sentences are close to each other in this vector space and hence they are used as an input representation for various downstream tasks. We use the pre-trained Universal Sentence Encoder (USE) (Cer et al., 2018) to represent the tweets.

### 2.2 Label Propagation

Label propagation is a graph-based semi-supervised learning technique that uses the labels from the labeled data to transduce the labels to unlabeled data. Label propagation considers two sets: $(x_1, y_1) \dots (x_l, y_l) \in L$ as labeled set and $(x_{l+1}, y_{l+1}) \dots (x_n, y_n) \in U$ as unlabeled set, where $y_1 \dots y_l \in \{1 \dots C\}$, $\{x_1 \dots x_n\} \in \mathbb{R}^D$. Here, $C$ is the number of classes, and the labeled set $L$ consists of all the classes. The algorithm constructs a graph $G = (V, E)$, where $V$ is the set of vertices representing set $L$ and $U$, and the edges in set $E$ represents the similarity between two nodes $i$ and $j$ with weight $w_{ij}$. The weight $w_{ij}$ is computed such that nodes with smaller distances (similar nodes) will have larger weights. The algorithm uses a probabilistic transition matrix $T$:

$$T_{ij} = P(i \rightarrow j) = \frac{w_{ij}}{\sum_{k=1}^{n} w_{kj}}$$

The algorithm iteratively updates the labels $Y \leftarrow TY$, by clamping the labels of the labeled set, and until $Y$ converges.

### 2.3 Proposed Methodology

Figure 1 outlines the semi-supervised learning setup adopted in our study. We use a Multilayer Perceptron (MLP) for two purposes: (a) to learn task-specific representations; (b) to perform multi-class classification.

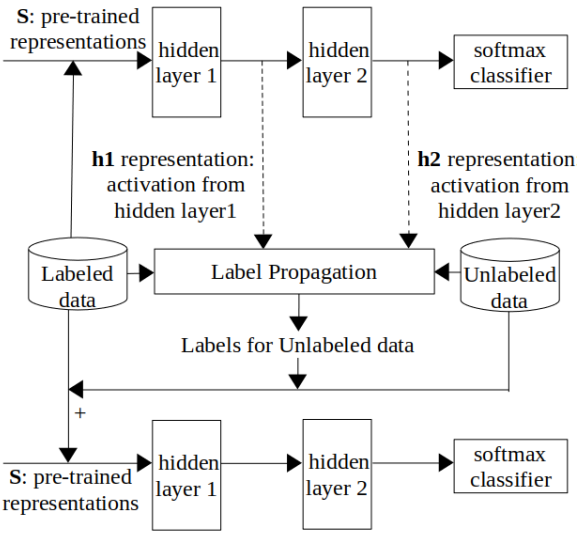


Figure 1: Block diagram for semi-supervised learning.

**Task-specific representations:** First, the pre-trained representation $S \in \mathbb{R}^D$ are transformed to task-specific representation $\hat{S} \in \mathbb{R}^{\hat{D}}$ with the MLP classifier trained using a small amount of available labeled set $L$. After training the MLP classifier, we pass the label agnostic pre-trained representation $S$ of a given sample from the labeled set $L$ and unlabeled set $U$ as input to the MLP. We consider the activation from outputs of the hidden layers (h1 and h2) as two different task-specific transformed representations $\hat{S}$. Since the MLP classifier is trained with labeled data, we expect the representations h1 and h2 to capture task-specific label information.

**Semi-supervised training:** The pre-trained representations $S$ or task-specific representations $\hat{S}$ will be used to represent data points in label propagation. We perform label propagation using the labeled set $L$ and unlabeled set $U$, to obtain the labels for the samples in $U$. Finally, the pre-trained embeddings of set $L$ and set $U$, along with original labels for $L$ and labels obtained from label propagation for $U$ will be used to train an MLP for hate speech classification.

## 3 Experimental Setup

### 3.1 Data Description

We consider two datasets, by Founta et al. (2018) and Davidson et al. (2017), containing tweets sampled from Twitter. Table 1 shows the statistics of these datasets. Both datasets have an imbalanced class distribution with 'hateful' as a minority class.

| Dataset | #Samples | Normal | Abusive | Hateful |
|---|---|---|---|---|
| Founta | 86.9K | 63% | 31% | 6% |
| Davidson | 24.7K | 17% | 77% | 6% |

Table 1: Dataset statistics for Founta et al. (2018) and Davidson et al. (2017)

**Twitter data by Founta et al. (2018):** A large part of this dataset is collected using random sampling. Since hate and abusive speech occurs in a very small percentage, the authors chose to perform boosted sampling. The dataset has four classes, namely 'normal','spam','abusive', and 'hateful'. We exclude the samples from the 'spam' class. This brings down the total number of samples in the dataset from 100K to 86.9K.

**Twitter data by Davidson et al. (2017):** It is collected based on the keywords from the hatebase[1] lexicon. The dataset is annotated into three classes, namely 'hate speech', 'offensive language' and 'neither'. Further in this paper, to map the class labels to Founta et al. (2018) dataset, we consider 'neither' class as 'normal' and 'offensive language' as 'abusive'. Since 'abusive speech' and 'offensive language' are strongly correlated (Founta et al., 2018), we use the terms interchangeably.

### 3.2 Data Processing

#### 3.2.1 Text Preprocessing

We remove all the numbers and punctuations except '.', ',', '!', '-', and apostrophe. The repeated occurrence of the same punctuation is changed to a single one. Hashtags are preprocessed by removing the '#' symbol and those with multiple words are split based on uppercase letters. For example, "#getBackHome" is processed into "get Back Home". This is done to ensure that the text tokenizer treats multiple words as a sequence of distinct words. However, the hashtags with multiple words without any uppercase is left as it is. Twitter user handles and the symbol 'RT' which indicates re-tweet are removed.

#### 3.2.2 Data Split

We split the datasets randomly into three portions 'training', 'validation', and 'test' sets, each containing 60%, 20%, and 20% respectively. To simulate the semi-supervised learning setup, we partition the training set into labeled and unlabeled set with a ratio of 1:4. The final labeled set consists of 12% of the entire dataset.

[1] https://www.hatebase.org

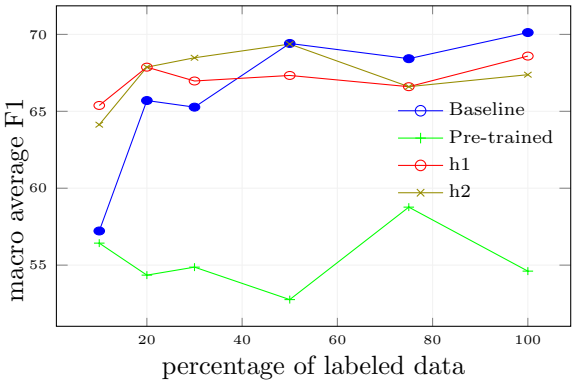


Figure 2: Performance on Founta et al. (2018) dataset.

### 3.3 Sentence Embeddings

We obtain pre-trained sentence embeddings from the transformer-based model of USE[2]. This model is trained on data from Wikipedia, web news, web question-answer pages, and discussion forums. Each sentence embedding is of 512 dimension.

### 3.4 Label Propagation

The label propagation API provided by scikit-learn library[3] is used. Euclidean distance is used to compute the distances between two data points. Based on the validation set, we have chosen 80 nearest neighbors and a maximum of 4 iterations.

### 3.5 Model and Representation Setup

We use an MLP with two hidden layers to perform classification and derive transformed representations. As shown in the Figure 1, the MLP model always has pre-trained representations $S$ as its input. The hidden layers have ReLU activation. The transformed representation $\hat{S}$ are taken from the 1st or 2nd hidden layers, referred as 'h1-representation' and 'h2-representation' in our experiments, respectively. We use Adam optimizer, early-stopping based on validation set, and maximum of 10 epochs. Both the hidden layers have 50 units. The model weights learnt while training the system with only the labeled data is used to initialize the classifier before training with labeled and unlabeled data.

## 4 Results and Discussion

Figures 2 and 3 show percentage macro-average F1 of classification on the test set in the semi-supervised approach. The amount of unlabeled

[2] https://tfhub.dev/google/universal-sentence-encoder-large/3

[3] https://scikit-learn.org/stable/modules/label_propagation.html

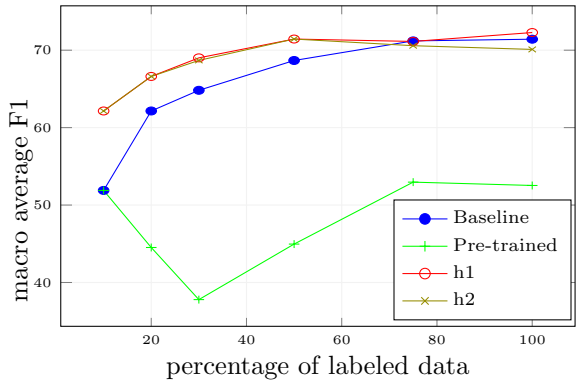


Figure 3: Performance on Davidson et al. (2017) dataset.

data is fixed, but the amount of labeled data is varied to 10%, 20%, 30%, 50% and 100% of the available labeled set. The 'Baseline' is obtained by training the MLP classifier with only the labeled set, and without using label propagation. The 'Pre-trained', 'h1', and 'h2' show the results of the classifier trained using labeled and unlabeled set, the respective representation was used to label the unlabeled set using label propagation. In all the four cases, pre-trained embeddings were used as input to the MLP classifier.

Results in Figure 2 and Figure 3 show that semi-supervised training using label propagation on pre-trained representations performs worse than the 'Baseline' classifier. This implies that label propagation to the unlabeled sets using pre-trained representations introduces significant noise in the classifier. To support this hypothesis, we analyse the intra-class and inter-class separations.

| | **pre-trained** | **h1** | **h2** |
|---|---|---|---|
| **intra-class distance** | | | |
| **'normal'** | 1.33 | 0.89 | 1.48 |
| **'abusive'** | 1.26 | 0.81 | 1.43 |
| **'hate'** | 1.25 | 0.87 | 1.63 |
| **inter-class distance** | | | |
| **'normal'-'abusive'** | 1.33 | 1.17 | 2.72 |
| **'normal'-'hate'** | 1.33 | 1.14 | 2.39 |
| **'abusive'-'hate'** | 1.33 | 1.14 | 2.99 |

Table 2: Average inter-class and intra-class euclidean distance across different representations for the train set of Founta et al. (2018) dataset.

Table 2 shows the average intra-class and inter-class euclidean distance across different representations on the training set, containing the labeled and unlabeled set. Ground truth labels are used for this analysis. From Table 2, we observe that

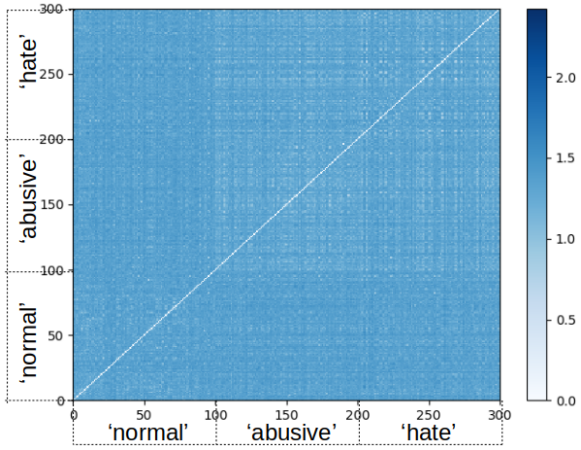


Figure 4: Color plot for distances between samples using the pre-trained representations.

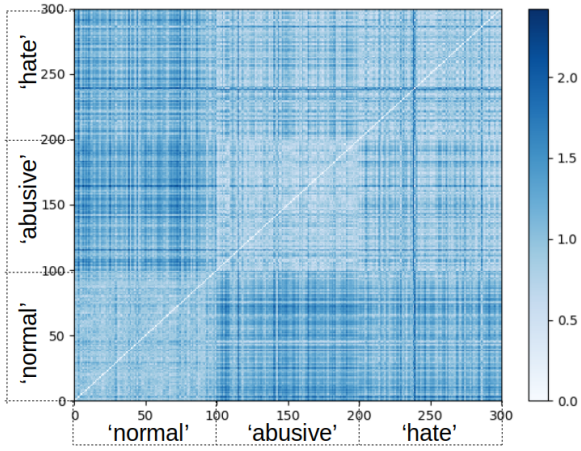


Figure 5: Color plot for distances between samples using the h1 representations.

the intra-class and inter-class distance are similar for pre-trained embeddings. This implies that the representations belonging to the samples from the same class were not close to each other and those from different classes were not far from each other. As hypothesised, h1 and h2 capture class information and hence have lower intra-class and higher inter-class distances.

Further, we qualitatively analyze the euclidean distances across pre-trained representation and h1 representation using color maps, shown in Figure 4 and Figure 5 respectively. We randomly consider 300 samples from Founta et al. (2018) dataset, of which the 1st set of 100 samples belongs to the labeled set of 'normal' class, the 2nd and the 3rd set of 100 samples belong to the labeled sets of 'abusive', and 'hate' classes respectively. From Figure 4, we observe that the pre-trained representation has similar distance across all the samples, thus appearing

in a similar color. This infers that pre-trained representation does not contain any task-specific label information, and are generic. However, as shown in Figure 5, the intra-class distance of the samples within 'normal' and 'abusive' classes is smaller than their inter-class distance of h1 representation, hence appearing as lighter-colored squares. From these figures, we observe that h1 representation captures task-specific label information.

Figure 2 and Figure 3 further show that semi-supervised training using label propagation on transformed representations ('h1' and 'h2') performs better than the 'Baseline' when amount of labeled data is small. The performance is comparable for larger labeled data. Thus, semi-supervised learning with label propagation using task-specific representations can have significant advantages when the available labeled samples are very few.

Furthermore, in a few cases, label propagation using the representations from the intermediate hidden layer 'h1' performed slightly better than label propagation using the representations from the final hidden layer 'h2'. This hints that fully fine-tuned representation may not always be the best performing representations in the k-Nearest Neighbors space.

## 5 Conclusion

In this article, we have explored label propagation-based semi-supervised learning for hate speech classification. We evaluated our approach on two datasets of hate speech on the multi-class classification task. We showed that label propagation using the pre-trained sentence embeddings reduces the performance achieved with only the labeled data. This is because pre-trained embeddings do not contain any task-specific information. We validated this by comparing the average intra-class and inter-class distances. Further, training an MLP classifier with a small amount of labeled data and using the activations of its hidden layers as task aware representations improved the performance of the label propagation and semi-supervised training. It also appears that the fully fine-tuned representations from the MLP may not be the best representations for the label propagation. We can conclude that semi-supervised learning based on label propagation helps to improve hate speech classification in very low resource scenarios and that the performance gain reduces with more amount of labeled data.

## Acknowledgments

This work was funded by the M-PHASIS project supported by the French National Research Agency (ANR) and German National Research Agency (DFG) under contract ANR-18-FRAL-0005.

## References

Steven Abney. 2007. *Semi-supervised learning for computational linguistics*. CRC Press.

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

Michal Bojkovskỳ and Matúš Pikuliak. 2019. Stufiit at semeval-2019 task 5: Multilingual hate speech detection on twitter with muse and elmo embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 464–468.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity*, pages 86–95.

Richard Delgado and Jean Stefancic. 2014. Hate speech in cyberspace. *Wake Forest L. Rev.*, 49:319.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. Fermi at semeval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 991–1000.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 107–114.

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault. 2017. Proceedings of the first workshop on abusive language online. In *Proceedings of the First Workshop on Abusive Language Online*.

Tomer Wullach, Amir Adler, and Einat Minkov. 2020. Towards hate speech detection at large via deep generative modeling. *arXiv preprint arXiv:2005.06370*.

Zhu Xiaojin and Ghahramani Zoubin. 2002. Learning from labeled and unlabeled data with label propagation. *Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University*.