Implicit Hate speech Detection Using a Hybrid Deep Learning Approach: Final Report

G018 (s2029442, s2087805, s2015345)

Abstract

Online hate speech continues to be a pervasive issue in digital spaces, impacting individuals and perpetuating harmful stereotypes. This study is concerned with the automatic detection of Hate Speech, with a focus on the more challenging classification of Implicit Hate Speech and its facets. The dataset used makes it possible to distinguish hate speech with fine-grained labels, and allows us to take a step-by-step approach to classification. Particular attention will be given to the analysis of the impact of cascading errors in the different stages of classification. The results from our two baseline models, a shallow SVM classifier and a deep pre-trained BERT model (Devlin et al., 2018), indicate the need for a more thorough study of possible approaches to this task. The model we are planning to develop integrates multiple techniques for enhanced performance. Specifically, the model incorporates BERT as its base architecture, coupled with data augmentation methods, and a Convolutional Neural Network (CNN) module. This combined approach aims to leverage the strengths of each component to achieve more robust and accurate hate speech detection and classification.

1. Introduction

Online hate speech continues to be a pervasive issue in digital spaces, such as social media platforms (SPMs), impacting individuals and perpetuating harmful stereotypes. The exposure to online HS can deepen prejudice and mislead the reader's opinion (Citron & Norton, 2011), leading to desensitization towards this form of verbal violence and lower evaluations of the victims (Soral et al., 2023).

The most common action against HS on SMPs is through content moderation, which consists of the removal of problematic user accounts and the hiding of sensitive or harmful user-generated content (Gerrard, 2018). The content to delete is flagged by moderators, which can be employees, users of the platform or, due to the sheer amount of data to screen, algorithms that follow the hate speech policy implemented by each social media company. Manually scrutinizing all flagged content is very time-consuming, the decisions made by human moderators tend to be subjective (Caplan, 2018), and continuous exposure to online hate can have negative mental health effects (Kwan et al., 2020).

While the current state-of-the-art for explicit HS detection has reached excellent results (Jahan & Oussalah, 2023), detecting implicit HS proves more challenging. Moreover, the topic of explicit versus implicit HS classification has not been thoroughly researched yet, mainly due to the scarcity of high quality annotated data. Detecting implicit HS, recognising its different forms and facets, and distinguishing it from the more noticeable forms of harmful content can be a very valuable task: while content deletion tends to be a good way to counteract explicit hate, implicit HS such as microaggressions, cyberbullying, or stereotyping, could be better counteracted and mitigated through interventions like Counter Narrative (CN) generation. Recognising implicit hate speech would make it possible to stop it early, either by offering a CN to people as they are typing the HS (before they post) (Katsaros et al., 2022), or by posting a CN publicly for everyone to see in order to diminish the spread of prejudiced views and harmful misinformation (Chung et al., 2021).

1.1. Research Question and Objectives

This study is concerned with automatic HS detection, in particular, we will focus on the more challenging task of detecting implicit HS. We will frame this as a classification problem, first of HS versus non-HS content, then of explicit versus implicit HS, and finally of different fine-grained types of implicit HS. In other words, this project aims to answer the following research question:

Can the use of a hybrid deep learning model and a multi-stage approach to classification improve the performance of implicit HS detectors?

To tackle this question, this work aims to achieve the following objectives:

1. Build a deep learning model to improve the performance of baseline classifiers on the detection of implicit and explicit HS in ElSherief et al.'s dataset (2021).
2. Examine the effect of cascading errors in the fine-grained classification of different types of (implicit) hate speech.
3. Show the importance of a multi-stage approach to classification for the overall accuracy of HS detection models and for possible content moderation strategies.

2. Related work

Prior research on identifying HS on SMPs has reached the implementation of systems that can flag online HS with remarkable accuracy (Schmidt & Wiegand, 2017), but the focus has been primarily on classifying explicit HS and overt abuse. The lack of research on the topic of implicit HS can be attributed not only to the challenge this task presents, but also to existing datasets being strongly skewed towards explicitly hateful text (Waseem & Hovy, 2016).

Many datasets have been created to aid the task of online HS detection, reflecting both the variation of this phenomenon and its relevance. Datasets can be classified in terms of their language(s), hate target, data source, and granularity of annotations (Chung et al., 2022). Most datasets are retrieved from social media platforms such as Facebook, Twitter, YouTube, or Instagram; English is the most common language, comprising over 50% of HS-related datasets, studies, and open-source projects (Jahan & Oussalah, 2023). The hate target is the most straightforward way to characterize hate datasets, but the boundary between hate identities can be vague and subjective (Poletto et al., 2021).

Most current datasets heavily rely on very overt lexical signals and, although some may contain implicit HS, there are no implicit hate labels to recognise that. ElSherief et al. (2021) build on the Gab Hate Corpus of Kennedy et al. (2018), which provides both explicit and implicit hate and target labels for a random sample of 27K Gab messages. They extend this work with a theoretically-grounded taxonomy and fine-grained labels for implicit hate speech on Twitter data. One of the first systems to classify implicit HS leveraging external knowledge is proposed by Lin et al. (2022) and uses the *Latent Hatred* dataset (ElSherief et al., 2021). Related work is that by Breitfeller at al. (2019) on microaggressions, Sue (2010) on the unconscious linguistic reflections of social bais, and the more recent research by Ghosh et al. (2022) on the importance of context for implicit HS detection.

HS detection can be approached as a classification problem, using Naive Bayes or Logistic Regression algorithms and features derived from lexical resources (Gitari et al., 2015), Support Vector Machines (Fortuna et al., 2021), n-grams, or knowledge bases (Dinakar et al., 2012); Araque and Iglesias (2022) suggest novel feature extraction methods that use resources from affective computing, showing that the combination of affective-aware features with studied textual representations can yield a performance improvement. The use of deep learning algorithms such as Convolutional Neural Networks (Zhang & Luo, 2019), Recurrent Neural Networks (Nobata et al., 2016) and Long Short-Term Memory networks (Arango et al., 2019) proves to be a valid alternative: Jahan and Oussalah's review of automatic HS detection (2023) revealed that deep-learning models outperformed popular classifiers in most studies. Moreover, in recent years the use of pre-trained language models such as BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019), or RoBERTa (Liu et al., 2019) has reached state-of-the-art results, with these becoming the top performing systems of offensive language identification at Semeval-2019, Semeval-2020 and the HASOC-2020 competitions (Jahan & Oussalah, 2023).

ElSherief et al. (2021) experiment with two classification tasks, distinguishing implicit hate speech from non-hate, and categorizing implicit hate speech with their fine-grained taxonomy. Their baselines, which we aim to replicate and improve on, are SVM and BERT models. Additionally, they experiment with data augmentation and knowledge-based features.

3. Data set and task

The dataset we are using (ElSherief et al., 2021) contains 22.056 tweets from the most prominent extremist groups in the US and is currently the most complete available benchmark for implicit HS to our knowledge. ElSherif et al. (2021) label their data into various steps called stages, each categorizing HS at one level of precision further and more fine-grained than the previous one.

- Stage 1: The data is labelled as No Hate, Explicit Hate, or Implicit Hate
- Stage 2: There are 6.346 tweets containing implicit hate, which are separated into the following categories:
  - Grievance
  - Incitement
  - Inferiority
  - Irony
  - Stereotypes
  - Threats
  - Other (Only 1.2% of implicit hate)
- Stage 3: For each implicit HS datapoint, the target group and an explanation of the implied harmful meaning is annotated.

We manipulate the data by taking a more gradual approach to classification: we add a 'Stage 0' to the data annotation, labelling Hate or No Hate; the Hate data is further annotated as Implicit or Explicit Hate and, following the HS taxonomy proposed by ElSherif et al. (2021), we use their Stage 2 labels for fine-grained distinctions on types of HS. For the purpose of this study, we don't make use of Stage 3 annotations. Figure 1 shows the different classification steps used in this study.

The labeled data is split between 80% training and 20% testing data. For validation, 10% of the training data is used, accounting for 8% of the entire corpus.

The data, especially for finer grained labels, tends to be quite imbalanced, and the lines between classes can be blurry and difficult to identify. Table 1 shows the initial counts per class. To tackle the impact of very imbalanced data on the models' performance, we will experiment with data augmentation techniques.

Stage	Label	Count
Hate vs. No Hate	No Hate	13291
Hate vs. No Hate	Hate	7349
Implicit HS vs Explicit	Explicit	1094
Implicit HS vs Explicit	Implicit	6255
Fine-grained Implicit HS	Grievance	1507
Fine-grained Implicit HS	Incitement	1240
Fine-grained Implicit HS	Stereotypes	1105
Fine-grained Implicit HS	Inferiority	867
Fine-grained Implicit HS	Irony	795
Fine-grained Implicit HS	Threats	666
Fine-grained Implicit HS	Other	79

Table 1. Counts of data points per class in original dataset

The tasks this project is concerned with are binary and multi-class classification, in particular for implicit HS detection. Throughout this project, we will explore how SVM, BERT, and CNN models perform in this task, eventually arguing for a hybrid approach that leverages BERT for encoding and the CNN architecture for classification. Our focus will be on how taking a multistep approach to classification, distinguishing between HS and not HS, explicit and implicit HS, and different types of implicit HS can lead to more insightful results. Moreover, to improve the performance of our models, we will employ data augmentation techniques while ensuring diversity in the data.

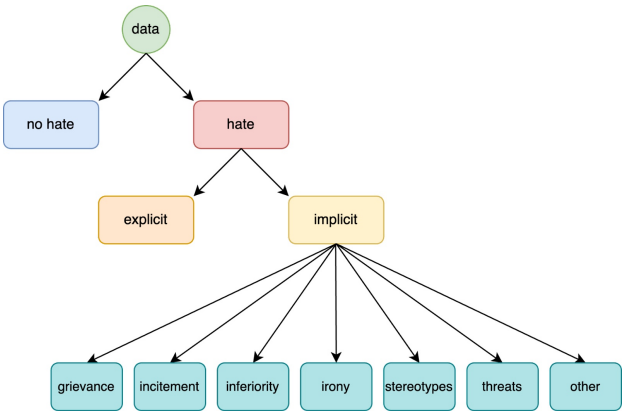

Figure 1. Breakdown of the different classification steps

4. Methodology

As can be seen from Figure 1, we plan to tackle the task of implicit and explicit hate speech detection by categorizing the stages individually: firstly we will perform binary classification between HS and Not-HS, then between implicit HS and explicit HS, until we divide the implicit HS data into 7 further categories. Our study will be articulated in three experimental stages:

1. Build a series of baseline models and evaluate their performance.
2. Explore any imbalances in the dataset, measure their effect on the performance of our model, and propose and implement ways to mitigate those effects to improve model performance.
3. Develop more complex deep models and evaluate their performance.

We believe our approach to HS detection would integrate well with modern systems used by social media platforms by providing more flexibility: instead of only focusing on whether the text is hateful or not, social media platforms can take different actions according to the type of hate speech detected by our model. An example of where this could be particularly beneficial would be in counter narrative generation for implicit hate speech (Chung et al., 2022).

Our approach also accounts for how HS detection models tend to bias towards certain keywords (De la Peña Sarracén & Rosso, 2023). Although imbalances in the original data may make our models susceptible to such biases, by implementing ways to resolve those imbalances (e.g., data augmentation, which we describe in section 4.3) we ensure that our hybrid model is less likely to bias towards specific words or phrases.

4.1. Evaluation Metrics

The four main performance metrics that we will use to evaluate our models are accuracy, precision, recall, and F1 score. We provide the formal definitions for each of these four metrics below (Sokolova et al., 2006) (note that we use the following abbreviations in the mathematical formulas: TP = true positives, FP = false positives, TN = true negatives, FN = false negatives):

1. **Accuracy:** the number of correct predictions divided by the total number of predictions

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

2. **Precision:** the proportion of true positive predictions in relation to all positive predictions made

$$\text{Precision} = \frac{TP}{TP+FP}$$

3. **Recall:** the proportion of actual positives that were correctly identified by the model

$$\text{Recall} = \frac{TP}{TP+FN}$$

4. **F1 score:** the harmonic mean of the precision and recall

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$$

4.2. Models Design

For the first stage of our study, we opted for two different baseline models: a smaller, shallow model, and a more complex, deep model.

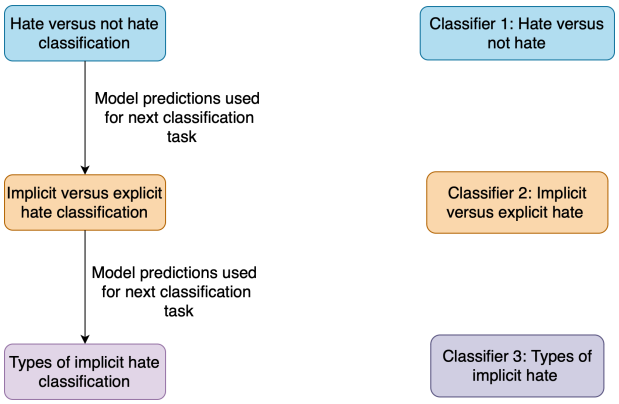

Figure 2. Classification procedure used for our baseline models. We show how the chained classification was implemented (left) and how it differs from having three individual classifiers (right).

4.2.1. Shallow Baseline Model

Our choice for the shallow model is a SVM classifier, due to their ability to handle non-linear data separation through the use of kernel tricks and their effectiveness in high-dimensional spaces. SVM classifiers are widely used in current literature as baseline models for tasks such as sentiment analysis, spam detection, and also hate speech classification (Vogel & Meghana, 2021).

We show the procedure we used to design our baseline experiments for the SVM classifier in Figure 2. As can be seen from the figure, we use the following two main approaches:

1. Three separate classifiers, one for each of the individual classification steps shown in Figure 1
2. One chained classifier that provides classification results at each individual step. This classifier is composed of three separate classifiers, and the predictions of each classifier are factored in as part of the input into the next classifier.

For the one classifier where previous predictions are used as input for the next classification task (e.g, predictions for hate versus not hate are factored in when classifying explicit versus implicit), we outline the procedure used through the algorithm below.

The procedure outlined in Algorithm 1 differs from that used for the individual classifiers, where the full dataset is used instead of filtering the data based on previous predictions. The reason for using previous predictions as a filtering mask for the chained classifier is due to SVM's infelxibility in altering the model's features space. By having both the three individual classifiers and the one chained classifier, we can measure the discrepancy in their performance at each classification step to determine the impact of cascading errors, the classification errors caused by a model's bias due to its previous predictions. These help us quantify to what extent factoring in the model's prior knowledge and bias impacts its ability to correctly and accurately perform a new different, related classification task.

Algorithm 1 SVM classification with previous predictions

procedure INCORPORATEPREVPREDICTIONS
  clf1 ← hate/no-hate classifier
  clf2 ← implicit/explicit classifier
  preds ← predictions by *clf1*
  dataIE ← dataset for implicit/explicit
  newData ← empty
  for *post* in *dataIE* **do**
    if *post*is hate in *preds* **then**
      add *post* **to** *newData*
    end if
  end for
  classify *newData* **with** *clf2*
end procedure

4.2.2. Deep Baseline Model

For our deep model, we have opted for BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). This choice can be justified due to BERT's reliable results on many NLP tasks using a pre-training/fine-tuning approach; an example of this is given by the results in (ElSherief et al., 2021), where BERT- base models consistently outperform the linear models. BERT is designed for pre-training deep bidirectional representations from unlabeled text, conditioning on left and right context in all layers. The open-source pre-trained model can be fine-tuned with one additional output layer for many tasks including HS detection and classification. For most tasks, no modifications to the architecture are necessary and, when using one of the smaller BERT models, the fine-tuning stage can be executed with limited computational resources. We decided to have a deep model as our second baseline in order to further showcase the motivation for our study, by showing that a single deep model alone also performs quite poorly on implicit hate speech classification tasks.

The procedure we used to design our baseline experiments for the BERT classifier is similar to that of the SVM classifier: we once again use the dual approach of individual classifiers and the one chained classifier to determine the impact of cascading errors for our deep baseline model. It is worth noting again that the 'chained classifier' is still technically composed of three different classifiers, with the addition being that, unlike the individual classifiers, we use the predictions from the previous classification step as input for the next step. We illustrate the procedure used for the BERT classifier using the algorithm below.

As can be seen from the procedure in Algorithm 2, unlike the SVM classifier, we use the full dataset for further classification steps when using a chained BERT classifier. This is because BERT provides us with more flexibility in terms of editing the model's feature space using previous predictions, allowing us to apply one-hot encoding to the model's previous predictions and include it as a feature for the next

Algorithm 2 Algorithm for factoring in previous predictions for the next classification task using BERT classifiers

procedure IncorporatePrevPredictions
  clf1 ← hate/no-hate classifier
  clf2 ← implicit/explicit classifier
  preds ← predictions by *clf1*
  dataIE ← dataset for implicit/explicit
  encode *preds* **using one-hot encoding**
  encodedpreds ← one-hot encoded*preds*
  append *encodedpreds* **to the set of existing features for** *clf2*
  classify *dataIE* **with** *clf2*
end procedure

classification step. This approach closer reflects what we described at the beginning of this section, where a model's predictions are factored into the model's knowledge instead of applied as a filter on the data.

4.2.3. Our Hybrid Approach

Recent results in literature appear to suggest that deep learning models provide better performance for hate speech classification tasks (Zimmerman et al., 2018). An example of such deep models would be convolutional neural networks (CNN), which are a regularized type of feed-forward neural network that is able to learn feature engineering by itself via filters optimization (Wu, 2017). This makes CNN particularly useful for tasks that involve finding patterns in the data and categorise them based on those patterns, such as the task of text classification. This makes CNNs a particular popular choice for hate speech detection and classification, with many examples in current literature leveraging this model architecture for more accurate results (Zhang & Luo, 2019; Jahan, 2020; Zampieri et al., 2019). Given the CNN's ability leverage patterns in the data for high accuracy, as well as BERT's ability to extract meaning from text data, we decided to use a hybrid approach. This approach would combine the representational power of BERT encoding with a deep learning algorithm for classification, such as a CNN. We show the structure of our CNN in Figure 3. The main aim of this combined approach is to leverage the strengths of each component to achieve more robust and accurate hate speech detection and classification, surpassing the results obtained by just using one of the two methods on its own.

4.3. Data Augmentation Techniques

As Table 1 shows, some of the classes in our original dataset had very few data points, impacting the performance of the models in their classification. The most noticeable example it 'other' in the fine-grained implicit HS labels, with only 79 occurrences. To overcome this issue, and use more balanced classes in the training and testing of our models, we decided to experiment with data augmentation techniques. In particular, we decided to use back-translation (Liu et al., 2023), practically achieving paraphrases of the original data by translating it into a different language and back into English.

Due to the amount of new data we needed to generate, we decided to use two different neural machine translation (NMT) models, namely FairSeq (Ott et al., 2019) and MarianMT (Junczys-Dowmunt et al., 2018), and try three different language combinations. We performed back-translation from Russian (en-ru-en) and French (en-fr-en) using the MarianMT model and from German (en-de-en) using the FairSeq model. For both models, we set the temperature parameter, which controls the creativity of the outputs, to 0.7, which, from empirical results, seemed to give us a good balance between favoring diversity in the output and controlling their reliability.

Original Sentence	Paraphrase
The white genocide continues: white race now minority in London too! Read more: via	White man's downfall continues: white race now minority in London. Read via

Table 2. Comparison of original sentence and paraphrase.

After iteratively generating new paraphrases through the three different model configurations, we processed the newly augmented data to remove potential duplicates and visually assess the quality of randomly picked samples (i.e 2). Table 3 shows the counts of the augmented dataset. To ensure that the data was still representative of a wide range of real-world scenarios, and that our paraphrases were introducing diversity to the dataset, we compared the distribution of the original data points with the distribution of newly created data. To do so, we encoded the data with TD-IDF vectorization and projected the encodings using UMAP (McInnes et al., 2018). As the plots in Figure 8 confirm, paraphrasing introduced a significant amount of diversity, and while some overlap between the original and newly generated data points' encodings is visible, this is a negligible amount in comparison to the overall data distribution.

Stage	Label	Count
Hate vs. No Hate	No Hate	13291
Hate vs. No Hate	Hate	**12959**
Implicit HS vs Explicit	Explicit	**3771**
Implicit HS vs Explicit	Implicit	**9188**
Fine-grained Implicit HS	Grievance	1507
Fine-grained Implicit HS	Incitement	**1242**
Fine-grained Implicit HS	Stereotypes	1105
Fine-grained Implicit HS	Inferiority	**1316**
Fine-grained Implicit HS	Irony	**1402**
Fine-grained Implicit HS	Threats	**1523**
Fine-grained Implicit HS	Other	**1100**

Table 3. Counts of data points per class in augmented dataset. The counts highlighted using bold text represent the classes for which we applied data augmentation.

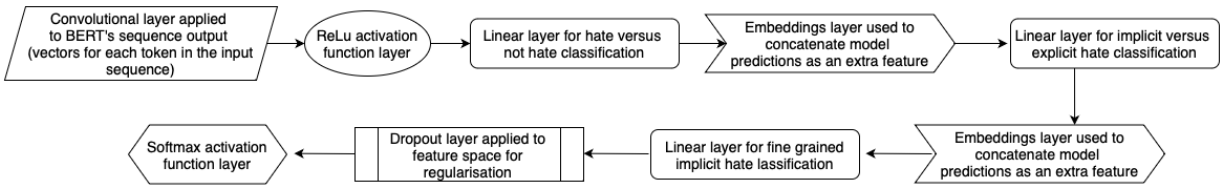

Figure 3. The sturcture of the convolutional neural network (CNN) we designed for the task of hate speech classification.

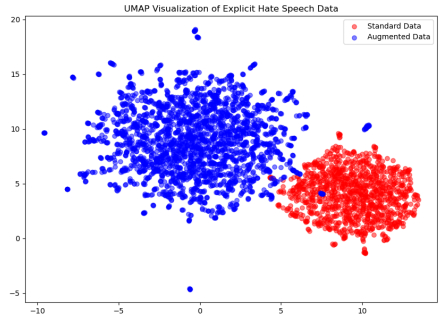

Figure 4. Explicit HS

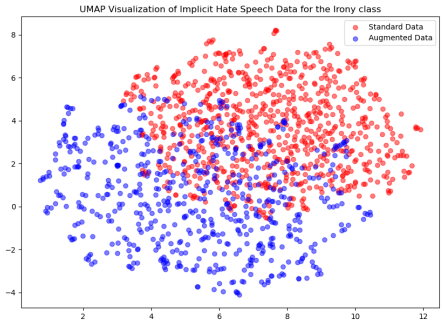

Figure 5. Irony Implicit HS

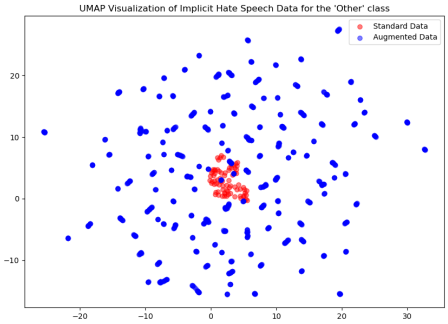

Figure 6. Other Implicit HS

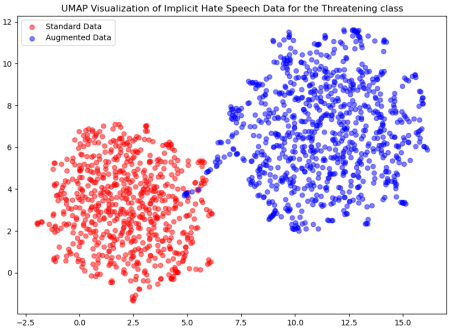

Figure 7. Threatening Implicit HS

Figure 8. UMAP plots comparing distribution of the original data and augmented data vectorised with TD-IDF

5. Experiments

We designed a series of experiments to test the performance of both the baseline models and our CNN model on the original and augmented dataset. We then designed our last experiment to test our final hybrid approach, where we used BERT instead of TD-IDF for extracting meaning and encoding the text data, before performing hate speech classification using the CNN. This setup enabled us to test the main hypothesis for our study, which states that *Since our hybrid approach combines two powerful deep learning models, leveraging the powerful meaning extraction of BERT and the high accuracy of a CNN, its performance will be significantly better than that of a single deep learning model.*

The experimental part of our project consisted in evaluating the performances of the following models:

- N-gram vs TF-IDF text vectorization for Chained SVMs and Single Task SVMs
- Chained Bert vs. single task Bert
- Bert with augmented data vs. CNN with augmented data
- BERT + CNN with augmented data vs. the rest

Tables 4 and 5 show a comparison between Chained SVM and Single-task SVM models across the three stages of classification: Hate vs. No Hate, Implicit vs. Explicit HS, and Fine-grained Implicit HS (categorizing the types of implicit HS).

In Table 4, with n-gram text vectorization, the Chained SVM model outperforms the Single-task SVM in terms of accuracy in the first stage (Hate vs No Hate) with an accuracy of 0.76 vs. 0.74. When distinguishing between Implicit and Explicit HS they achieve the same accuracy of 0.88, and finally, when diving down into the types of implicit HS, we can see a poor performance from both models, but Single-task SVM does outperform Chained SVM with 0.54 and 0.50 accuracy respectively.

For F1-score (F-1), precision (P), and recall (R), the Chained SVM generally shows better precision, except in the Fine-grained Implicit HS stage, where the Single-task SVM has a slightly higher precision (0.46 vs 0.42), F-1 (0.45 vs. 0.41) and recall (0.46 vs. 0.41). On the other hand, Single-task SVM seems to have a marginally better balance between precision and recall (as indicated by F-1

scores) in the Fine-grained Implicit HS stage compared to the Chained SVM.

In Table 5, using TF-IDF text vectorization, we see a similar trend where the Chained SVM tends to have higher accuracy. However, the differences in F-1, precision, and recall are less pronounced compared to Table 4, indicating that the choice of text vectorization technique may impact the relative performance of the two models. In terms of accuracy we see that the Chained SVM performs exactly the same regardless of the type of text vectorization method used, however, single task SVM performs the same.

In summary, while the Chained SVM model shows superior performance, the choice between these models also depend on the specific requirements for precision or recall. If precision is more critical (minimizing false positives), the Chained SVM might be preferable. Conversely, if recall is more important (minimizing false negatives), the Single-task SVM might be better, especially in the context of Fine-grained Implicit HS classification.

To understand the effect of "cascading" errors (what we have defined as chained models), where we carry our errors from previous classifications to the current one, we should look at our precision values. The reason for this is that given that stages 0 and 1 are binary classifications where we proceed by removing 1 class for the next classification, the only errors we are "cascading" are the false positives. When examining Tables 4 and 5, we see that precision values are very similar for one text vectorization technique for the equivalent steps. The only significant difference there is, is for stage 2 of TF-IDF, where precision is 10% higher when not cascading errors. Although this is a significant difference, since we decided not to do the stage 3 classification, we do not cascade those false positives to a future stage, making precision much less important than accuracy, which itself is much better.

Model	Stage	F-1	P	R	Acc.
Chained SVM	Hate vs. No Hate	0.70	0.75	0.69	0.76
Chained SVM	Implicit HS vs Explicit HS	0.56	0.84	0.55	0.88
Chained SVM	Fine-grained Implicit HS	0.41	0.42	0.41	0.50
Single-task SVM	Hate vs. No Hate	0.70	0.74	0.69	0.74
Single-task SVM	Implicit HS vs Explicit HS	0.53	0.88	0.53	0.88
Single-task SVM	Fine-grained Implicit HS	0.45	0.46	0.46	0.54

Table 4. Performance of SVM models with n-gram text vectorisation

Table 6 compares the performance of Chained BERT and Single-task BERT models on the three same levels of classification previously mentioned. Regarding F1-scores, precision, and recall, the Single-task BERT model slightly outperforms the Chained BERT model in the Hate vs. No Hate and Implicit vs. Explicit HS stages, with marginally higher values. This indicates that it is better at distinguishing between categories than the Chained BERT model. Once more, for Fine-grained Implicit HS, the Single-task BERT model shows a higher F1-score, precision, and recall compared to the Chained BERT model, indicating it has a better balance between precision and recall.

Model	Stage	F-1	P	R	Acc.
Chained SVM	Hate vs. No Hate	0.70	0.75	0.69	0.76
Chained SVM	Implicit HS vs Explicit HS	0.56	0.82	0.56	0.88
Chained SVM	Fine-grained Implicit HS	0.41	0.41	0.41	0.50
Single-task SVM	Hate vs. No Hate	0.68	0.71	0.67	0.72
Single-task SVM	Implicit HS vs Explicit HS	0.58	0.83	0.56	0.88
Single-task SVM	Fine-grained Implicit HS	0.45	0.51	0.45	0.53

Table 5. Performance of SVM models with TF-IDF text vectorisation

However, Chained BERT shows higher accuracy across all stages, with the largest difference when differentiating implicit and explicit HS, with an accuracy of 0.87 compared to Single-task BERT's accuracy of 0.75.

These results suggest that while the Single-task BERT may be more precise and recall more true positives in certain stages, the Chained BERT model may be more accurate overall in classifying the given stages.

For Table 7 we compare the performance of Chained BERT and CNN models with data augmentation. For the Chained BERT + Data Augmentation, we see a high performance across all stages, with the strongest performance across all KPIs seen in the Implicit HS vs Explicit HS stage (F1: 0.89, Precision: 0.88, Recall: 0.89, and Accuracy: 0.90). However, its performance drops significantly in the Fine-grained Implicit HS stage (accuracy = 0.67), indicating more complexity in distinguishing between finer categories.

The CNN + Data Augmentation excels particularly in the Implicit HS vs Explicit HS stage with remarkable F1, Precision, and Recall scores (0.91 each) and a high accuracy (0.87), suggesting this model's strength lies in this binary classification task. However, it falls behind the Chained BERT model in the Hate vs. No Hate classification stage.

Overall, it seems like each of these models have different strengths, with an overall stronger performance from BERT. However, BERT's strong lack in performance in categorizing types of implicit HS motivated us to combine the encoding by BERT with the CNN architecture, as we believe this will lead to our strongest model overall.

The results of this hybrid model are seen in Table 8 where we can see that this new approach achieves higher F1 scores,

precision, recall, and accuracy in the Hate vs. No Hate and Implicit vs. Explicit HS stages. Specifically, the hybrid model's performance in the Implicit vs. Explicit HS stage is notably higher (F1: 0.95) compared to the CNN model's performance in the same stage in Table 7 (F1: 0.91).

In the Fine-grained Implicit HS stage, the hybrid model's F1 score is also higher (0.72) compared to the CNN model in Table 7 (0.65). This suggests that combining BERT and CNN models with data augmentation may offer a more robust predictive capability across different classification challenges.

Overall the results of these experiments prove that our hypothesis were correct with the hybrid model performing significantly better than the others. Not only that, but as is made evident in Table 8, this model outperforms the the best performing model in ElSherief et al. 2021 by a very significant margin. Comparing the stage 2 classification from our final model, we see a very big increase in performance to 71% from (ElSherief et al., 2021) 's 64% accuracy.

Given that BERT is an external algorithm, we cannot attribute this improvement to it, however when looking at Tables 7 and 8 together we see that both our data augmentation and our CNN contributed to this improvement. The data augmentation seen in Table 7 leads to a 3% increase in accuracy, and the remaining 4% can therefore be attributed to our CNN.

Model	Stage	F-1	P	R	Acc.
Chained BERT	Hate vs. No Hate	0.81	0.81	0.80	0.82
Chained BERT	Implicit HS vs Explicit HS	0.71	0.72	0.70	0.87
Chained BERT	Fine-grained Implicit HS	0.6	0.63	0.59	0.64
Single-task BERT	Hate vs. No Hate	0.82	0.82	0.82	0.79
Single-task BERT	Implicit HS vs Explicit HS	0.83	0.82	0.84	0.75
Single-task BERT	Fine-grained Implicit HS	0.68	0.67	0.69	0.64

Table 6. Performance of BERT baseline models

6. Conclusions

In this project, we've explored how to further the classification of HS at 3 levels: hate vs. no hate, implicit vs. explicit, and finally, distinguish between 7 kinds of implicit HS. By experimenting with various machine learning models we have achieved results on par with SOTA and sometimes surpassing previous research done with the same dataset. We have evaluated both the performance of different models as well as the effect of "cascading" errors, and concluded that a hybrid model in which the text vectorization is done through BERT and classified by a CNN is optimal for performance. In the end we had the following accuracies of 91% for stage 0, 92% for stage 1, and 71% for stage 2, which clearly outperforms our baselines where we had 76%, 88%, and 50% respectively. Through data augmentation techniques, we created a new larger dataset with more balanced classes and, using UMAP visualisations, we ensured the quality and diversity of the data was not harmed. This contribution can offer new possible benchmarks to the field of Implicit HS detection.

Model	Stage	F-1	P	R	Acc.
Chained BERT + data augmentation	Hate vs. No Hate	0.83	0.83	0.83	0.83
Chained BERT + data augmentation	Implicit HS vs Explicit HS	0.89	0.88	0.89	0.90
Chained BERT + data augmentation	Fine-grained Implicit HS	0.62	0.64	0.61	0.67
CNN + data augmentation	Hate vs. No Hate	0.75	0.78	0.72	0.77
CNN + data augmentation	Implicit HS vs Explicit HS	0.91	0.91	0.90	0.87
CNN + data augmentation	Fine-grained Implicit HS	0.65	0.64	0.65	0.65

Table 7. Performance of BERT and CNN models after data augmentation

Model	Stage	F-1	P	R	Acc.
BERT + data augmentation + CNN	Hate vs. No Hate	0.90	0.88	0.92	0.91
BERT + data augmentation + CNN	Implicit HS vs Explicit HS	0.95	0.94	0.96	0.92
BERT + data augmentation + CNN	Fine-grained Implicit HS	**0.72**	**0.72**	**0.72**	**0.71**
BERT + Aug (ElSherief et al., 2021)	Fine-grained Implicit HS	0.59	0.59	0.59	0.64

Table 8. Performance of the hybrid BERT+CNN approach

The focus on Implicit HS detection, and in particular on the identification of fine-grained labels, is an under-researched field of study. Further work in this direction can have an impact that extends beyond academia, and be of interest to stakeholders such as SMPs, users from diverse and possibly marginalised groups, and institutional bodies concerned with digital citizenship, offering tangible contributions to the ongoing efforts to create safer online environments.

Further steps should include additional experimentation on our CNN's layers and hyperparameters to improve performance, collecting a multilingual dataset, and proposing next steps for SMPs depending on the type of hate speech.

References

Arango, Aymé, Pérez, Jorge, and Poblete, Barbara. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pp. 45–54, 2019.

Araque, Oscar and Iglesias, Carlos A. An ensemble method for radicalization and hate speech detection online empowered by sentic computing. *Cognitive Computation*, 14(1):48–61, 2022.

Breitfeller, Luke, Ahn, Emily, Jurgens, David, and Tsvetkov, Yulia. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 1664–1674, 2019.

Caplan, Robyn. Content or context moderation? 2018.

Chung, Yi-Ling, Tekiroglu, Serra Sinem, and Guerini, Marco. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*, 2021.

Chung, Yi-Ling et al. Counter narrative generation for fighting online hate speech. 2022.

Citron, Danielle Keats and Norton, Helen. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91:1435, 2011.

De la Peña Sarracén, Gretel Liz and Rosso, Paolo. Systematic keyword and bias analyses in hate speech detection. *Information Processing & Management*, 60(5):103433, 2023.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dinakar, Karthik, Jones, Birago, Havasi, Catherine, Lieberman, Henry, and Picard, Rosalind. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30, 2012.

ElSherief, Mai, Ziems, Caleb, Muchlinski, David, Anupindi, Vaishnavi, Seybolt, Jordyn, De Choudhury, Munmun, and Yang, Diyi. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*, 2021.

Fortuna, Paula, Soler-Company, Juan, and Wanner, Leo. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3): 102524, 2021.

Gerrard, Ysabel. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12):4492–4511, 2018.

Ghosh, Soumitra, Ekbal, Asif, Bhattacharyya, Pushpak, Saha, Tista, Kumar, Alka, and Srivastava, Shikha. Sehc: A benchmark setup to identify online hate speech in english. *IEEE Transactions on Computational Social Systems*, 10(2):760–770, 2022.

Gitari, Njagi Dennis, Zuping, Zhang, Damien, Hanyurwimfura, and Long, Jun. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

Jahan, Md Saroar. Team oulu at semeval-2020 task 12: Multilingual identification of offensive language, type and target of twitter post using translated datasets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 1628–1637, 2020.

Jahan, Md Saroar and Oussalah, Mourad. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, pp. 126232, 2023.

Junczys-Dowmunt, Marcin, Grundkiewicz, Roman, Dwojak, Tomasz, Hoang, Hieu, Heafield, Kenneth, Neckermann, Tom, Seide, Frank, Germann, Ulrich, Aji, Alham Fikri, Bogoychev, Nikolay, et al. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*, 2018.

Katsaros, Matthew, Yang, Kathy, and Fratamico, Lauren. Reconsidering tweets: Intervening during tweet creation decreases offensive content. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pp. 477–487, 2022.

Kennedy, Brendan, Atari, Mohammad, Davani, Aida Mostafazadeh, Yeh, Leigh, Omrani, Ali, Kim, Yehsong, Coombs, Kris, Havaldar, Shreya, Portillo-Wightman, Gwenyth, Gonzalez, Elaine, et al. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18, 2018.

Kwan, Irene, Dickson, Kelly, Richardson, Michelle, MacDowall, Wendy, Burchett, Helen, Stansfield, Claire, Brunton, Ginny, Sutcliffe, Katy, and Thomas, James. Cyberbullying and children and young people's mental health: a systematic map of systematic reviews. *Cyberpsychology, Behavior, and Social Networking*, 23(2): 72–82, 2020.

Lan, Zhenzhong, Chen, Mingda, Goodman, Sebastian, Gimpel, Kevin, Sharma, Piyush, and Soricut, Radu. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Lin, Jessica. Leveraging world knowledge in implicit hate speech detection. *arXiv preprint arXiv:2212.14100*, 2022.

Liu, Xiner, He, Jianshu, Liu, Mingzhe, Yin, Zhengtong, Yin, Lirong, and Zheng, Wenfeng. A scenario-generic neural machine translation data augmentation method. *Electronics*, 12(10):2320, 2023.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

McInnes, Leland, Healy, John, and Melville, James. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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

Ott, Myle, Edunov, Sergey, Baevski, Alexei, Fan, Angela, Gross, Sam, Ng, Nathan, Grangier, David, and Auli, Michael. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

Poletto, Fabio, Basile, Valerio, Sanguinetti, Manuela, Bosco, Cristina, and Patti, Viviana. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, 2021.

Schmidt, Anna and Wiegand, Michael. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pp. 1–10, 2017.

Sokolova, Marina, Japkowicz, Nathalie, and Szpakowicz, Stan. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pp. 1015–1021. Springer, 2006.

Soral, Wiktor, Świderska, Aleksandra, Puchała, Dominik, and Bilewicz, Michał. Desensitization to hate speech: Examination using heart rate measurement. *Aggressive Behavior*, 2023.

Sue, Derald Wing. *Microaggressions and marginality: Manifestation, dynamics, and impact*. John Wiley & Sons, 2010.

Vogel, Inna and Meghana, Meghana. Profiling hate speech spreaders on twitter: Svm vs. bi-lstm. In *CLEF (Working Notes)*, pp. 2193–2200, 2021.

Waseem, Zeerak and Hovy, Dirk. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

Wu, Jianxin. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23):495, 2017.

Zampieri, Marcos, Malmasi, Shervin, Nakov, Preslav, Rosenthal, Sara, Farra, Noura, and Kumar, Ritesh. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*, 2019.

Zhang, Ziqi and Luo, Lei. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945, 2019.

Zimmerman, Steven, Kruschwitz, Udo, and Fox, Chris. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.