UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

NOISY SELF-TRAINING WITH DATA AUGMENTATIONS FOR OFFENSIVE AND HATE SPEECH DETECTION TASKS

JOÃO AUGUSTO LEITE

ORIENTADOR: DIEGO FURTADO SILVA

São Carlos - SP

June, 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

NOISY SELF-TRAINING WITH DATA AUGMENTATIONS FOR OFFENSIVE AND HATE SPEECH DETECTION TASKS

JOÃO AUGUSTO LEITE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Artificial Intelligence.

Orientador: Diego Furtado Silva

São Carlos - SP

June, 2024

João Augusto Leite

Noisy Self-Training with Data Augmentations for Offensive and Hate Speech Detection Tasks/ João Augusto Leite. – São Carlos - SP, June, 2024-

35 p. : il. (algumas color.) ; 30 cm.

Orientador: Diego Furtado Silva

– Universidade Federal de São Carlos, June, 2024.

1. Self-supervision 2. Offensive and hateful speech detection 3. Data augmentation I. Diego Furtado Silva. II. Universidade Federal de São Carlos. III. Departamento de Computação. IV. Noisy Self-Training with Data Augmentations for Offensive and Hate Speech Detection Tasks

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato João Augusto Leite, realizada em 16/07/2024.

Comissão Julgadora:

Prof. Dr. Diego Furtado Silva (UFSCar)

Prof. Dr. Alan Demétrius Baria Valejo (UFSCar)

Prof. Dr. Ricardo Marcondes Marcacini (USP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

RESUMO

As mídias sociais online estão repletas de comentários ofensivos e discursos de ódio, o que exige o desenvolvimento de sistemas automatizados de detecção para gerenciar o vasto volume de postagens geradas a cada segundo. Criar conjuntos de dados de alta qualidade rotulados por humanos para essa tarefa é desafiador e caro, principalmente porque as postagens não-ofensivas superam significativamente as ofensivas. Em contraste, dados não-rotulados são abundantes, mais acessíveis e mais baratos de obter. Esta tese explora a aplicação de métodos de *self-supervision*, que utilizam exemplos fracamente rotulados para aumentar os conjuntos de dados de treinamento. A contribuição central desta tese é o artigo "Noisy Self-Training with Data Augmentations for Offensive and Hate Speech Detection Tasks", que investiga a eficácia de abordagens de auto-treinamento com ruído utilizando técnicas *data augmentation* para melhorar a consistência das predições e a robustez contra dados ruidosos e ataques adversariais. Experimentos foram realizados com *self-training* padrão e com ruído, utilizando três diferentes técnicas de *data augmentation* textuais em cinco distintas arquiteturas BERT pré-treinadas de tamanhos variados. Os resultados indicam que o auto-treinamento com ruído e *data augmentations* textuais, apesar do sucesso em configurações semelhantes, prejudicam o desempenho dos modelos treinados para a tarefa de detecção de comentários ofensivos e discursos de ódio em comparação com o método padrão. Esse achado revela limitações dos métodos de *self-training* com ruído e *data augmentation* para domínios em que a modificação de certas palavras-chave gera alteração semântica.

Palavras-chave: self-supervision, comentários ofensivos e discursos de ódio, data augmentation

ABSTRACT

Online social media is rife with offensive and hateful comments, necessitating the development of automated detection systems to manage the vast volume of posts generated every second. Creating high-quality human-labeled datasets for this task is challenging and costly, primarily because non-offensive posts significantly outnumber offensive ones. In contrast, unlabeled data is abundant, more accessible, and cheaper to obtain. This thesis explores the application of self-training methods, which leverage weakly-labeled examples to augment training datasets, in the context of offensive and hate speech detection. The core of this thesis is the paper "Noisy Self-Training with Data Augmentations for Offensive and Hate Speech Detection Tasks", which investigates the efficacy of noisy self-training approaches incorporating data augmentation techniques to enhance prediction consistency and robustness against noisy data and adversarial attacks. Experiments are conducted with both default and noisy self-training using three different textual data augmentation techniques across five distinct pre-trained BERT architectures of varying sizes. The results indicated that noisy self-training with textual data augmentations, despite its success in similar settings, decreased performance in offensive and hate speech domains compared to the default method. This finding and reveals limitations of noisy self-training methods with data augmentations for domains such as offensive speech detection, where certain specific keywords cannot be modified without introducing semantic variations.

Keywords: self-supervision, offensive and hateful speech detection, data augmentation

LIST OF ABBREVIATIONS AND ACRONYMS

NLP	Natural Language Processing
SSL	Self-Supervised Learning
CNN	Convolutional Neural Network
ROC	Receiver Operating Characteristic curve
ROC-AUC	Area under the ROC curve
GRU	Gated Recurrent Unit
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GAN	Generative Adversarial Networks
BERT	Bidirectional Encoder Representations from Transformers
API	Application Programming Interface
DF	Default Fine-tuning
ST	Self-training
BT	Backtranslation
SS	Synonym Substitution
WS	Word Swap

Contents

Chapter 1
INTRODUCTION

The proliferation of social media platforms has transformed how individuals communicate and interact online. These platforms, while facilitating productive communication and information dissemination, have also become hotspots for offensive and hateful speech (MONDAL et al., 2017a). This phenomenon is exacerbated by factors such as user anonymity, which often emboldens individuals to express harmful sentiments without fear of repercussions. However, manual moderation of online content is not scalable due to the sheer volume of user-generated posts and the potential psychological impact on human moderators. This challenge underscores the importance of developing automated systems for hate speech detection. Such systems can leverage machine learning and natural language processing (NLP) techniques to identify and filter harmful content, thus protecting users and maintaining community standards (SCHMIDT; WIEGAND, 2017).

The development of robust detection models often requires large annotated datasets. However, obtaining labeled data is expensive and time-consuming, as it relies on human annotators to accurately identify and label instances of offensive and hate speech. On the contrary, vast amounts of unlabeled data are readily available across various domains, as the collection of such data is relatively inexpensive and scalable. This disparity between the availability of labeled and unlabeled data underscores the necessity for alternative learning paradigms. Self-supervised learning (SSL) emerges as a compelling approach, leveraging the abundance of unlabeled data to generate informative features without the exhaustive requirement for human annotation. This paradigm shift towards self-supervision not only mitigates the limitations imposed by the scarcity of labeled data but also enhances the model's ability to generalize from diverse and extensive datasets (BLUM; MITCHELL, 1998; CHEN et al., 2021; KARAMANOLAKIS et al., 2021).

Recently, novel self-supervised learning (SSL) methods incorporating noise-inducing techniques have demonstrated significant improvements across various domains. These approaches, such as the noisy student method (XIE et al., 2020b), integrate self-training with strategic noise introduction, both at the data level through augmentations and within model architectures using techniques like dropout. This combination enhances model robustness and generalization capabilities, leading to improved performance on diverse tasks such as computer vision (SCHIAPPA et al., 2022), audio and speech processing (LIU et al., 2022), and natural language processing (He et al., 2019). Despite these advances, the application of noisy SSL methods to the domain of offensive and hate speech detection remains relatively unexplored. While preliminary studies have employed default self-training approaches for this purpose (SANTOS et al., 2022), there is a notable gap in leveraging noise-induced strategies to enhance the detection accuracy of offensive and hate speech. The particularities of this domain present unique challenges that are not as prevalent in other applications. Specifically, the semantic meaning of offensive and hate speech is often highly dependent on specific keywords. These keywords are crucial indicators of the underlying class, and any alteration to them can significantly shift the intended meaning and, consequently, the classification outcome. For instance, replacing or modifying a single offensive term can transform a piece of text from hate speech to non-offensive content, thereby misleading the model during training and evaluation.

This thesis aims to address these challenges by exploring the application of noisy self-training combined with data augmentation techniques to enhance the detection of offensive and hate speech. It aims to provide comprehensive experimental results to verify whether such techniques lead to improvements in classification performance when applied to the domain of offensive and hate speech detection. This objective contributes to the development of more effective and scalable automated moderation systems. The structure of this thesis is as follows: Chapter 2 (Theoretical Foundations) - This chapter delves into the concepts of self-supervised learning, the teacher-student framework, and data augmentation methods for textual data, providing the theoretical underpinnings for the proposed approach. Chapter 3 (Noisy Self-Training with Data Augmentations for Offensive and Hate Speech Detection Tasks) - This chapter presents the research paper that is the core of this thesis, detailing its methodology, experiments, and results. Chapter 4 (Conclusion) - This chapter summarizes the findings, discusses the implications of the research, and outlines potential directions for future work.

Chapter 2

THEORETICAL FOUNDATIONS

2.1 Self-Supervision

Self-supervised learning (SSL) is a learning paradigm within the scope of semi-supervised methods (AMINI et al., 2022). In this category of methods, we frequently aim to learn a model using significant amounts of unlabeled data ($\propto 10^6$) combined with a commonly smaller set of labeled data ($\propto 10^4$). The idea to combine both types of data stems from a practical point of view: for many tasks, collecting unlabeled data is cheap, scalable, and automatable, while obtaining labeled data is often costly and time consuming, since it requires human labour. Self-supervised methods can be divided into two categories: self-predictive methods, and contrastive methods. The first comprises methods to train a model to predict an useen part of a data point, given the other parts of the same data point. For example, learning a model to predict a specific part of an image given the other parts of the same image. In this work, we are interested in the second type of self-supervised methods - the contrastive methods. More specifically, constrastive self-supervised methods make use of a supervised model to generate pseudo-labels from a different set of data points. Let us denote the human-annotated ground-truth labelled set as D_{gold} (i.e., the gold-standard data set), and the machine-inferred pseudo-labelled set as D_{silver} (i.e., the silver-labelled data set). A general framework for learning self-supervised models is shown in Algorithm 1.

Algorithm 1 General-SSL-Framework

```
Require: Ground-truth dataset $D_{gold}$, Unlabeled dataset $D_{unlabeled}$
1: $M_{t_0} \leftarrow$ Train($D_{gold}$)
2: $D_{silver} \leftarrow$ Inference($M_{t_0}$, $D_{unlabeled}$)
3: $M_t \leftarrow$ Train($D_{gold} \cup D_{silver}$)
```

Note that the word 'self' in self-supervision refers to the fact that we are learning a new model using the outputs generated by itself at an earlier stage. This paradigm has an important caveat: what if M_{t_0} is a poor predictor for the unlabelled data set? As a practical example, consider that M_{t_0} can predict the pseudo-labels for a given unlabelled data set with 30% accuracy. As a result, 70% of D_{silver} has incorrect labels, moreover, since $|D_{silver}| >> |D_{gold}|$, it is clear that the accuracy of M_t will be even lower than of M_{t_0}. Therefore, it is imperative for this framework that M_{t_0} is a decent predictor for the unlabeled data set. This can be difficult to measure, since there is no ground-truth labels for the unlabeled data set, therefore accuracy is not computable. Ultimately, self-supervision relies in making certain **assumptions** about the unlabeled dataset:

Manifold: The higher-dimensional input space contain lower-dimensional manifolds on which the data points lie, and data points on the same manifold belong to the same class.

Cluster: A set of data points that belong to the same cluster - i.e., are more similar to each other than to other sets of arbitrary data points - must belong to the same class.

Continuity: Given a labelled data point x_1, and two unlabelled data points x_2 and x_3. Let us assume x_2 is close to x_1 in the input space, and x_1 and x_2 belong to the same class y. Then, if x_3 is close to x_2, it should also belong to class y.

Low-density: The decision boundary between classes should not pass through high-density regions. In other words, regions in the input space where many data points are concentrated should not be a delimiter between two distinct classes.

2.2 Teacher-Student framework

A common approach in the self-supervised setting is to conceptually separate M_t from M_{t+1}, i.e., a model learned from a previous stage from the model learned in a subsequent stage. The model that generates D_{silver} is called the **teacher model**, and the model that learns with D_{silver} is called the **student model** (BLUM; MITCHELL, 1998; CHEN et al., 2021; KARAMANOLAKIS et al., 2021). Algorithm 2 displays the procedure to train a self-supervised model using the teacher-student framework.

Algorithm 2 Teacher-Student-SSL

```
Require: Ground-truth labelled dataset $D_{gold}$, Unlabeled dataset $D_{unlabeled}$, No. iterations $N$
$M_{teacher} \leftarrow$ Train($D_{gold}$)
for $i = 1$ to $N$ do
    $D_{silver} \leftarrow$ Inference($M_{teacher}$, $D_{unlabeled}$)
    $M_{student} \leftarrow$ Train($D_{gold} \cup D_{silver}$)
    $M_{teacher} \leftarrow M_{student}$
end for
```

This distinction allows (a) using different models to represent the teacher and the student, and (b) applying a different treatment specifically for the teacher model or for the student model. An example for (a) would be knowledge distillation. This task involves learning a student model $M_{student}$ that is smaller (i.e., has fewer parameters) than the teacher model $M_{teacher}$. Learning a smaller model that is (ideally) as accurate as a larger model has several practical advantages such as speeding up inference, reducing deployment costs, improving generalisation, etc. For the purpose of this thesis, we are more interested in (b) - learning $M_{teacher}$ in a slightly different way than $M_{student}$. This will be discussed in more detail in the next section.

2.3 Noisy Student

Developed by Xie et al. (2020b), the noisy student method integrates self-training with noising techniques to improve the robustness and generalization capabilities of the model. The framework builds upon the teacher-student self-training concept presented in Section 2.2 by introducing noise during the training of the student model. More specifically, noisy student introduces noise both in the model architecture, through the use of stochastic depth or dropout, and in the data, through the use of data augmentation methods. Inserting noise in the training of the student model is intended to increase robustness to variations in the input data, known as adversarial examples (see Table 1) (RASMUS et al., 2015; LAINE; AILA, 2017; MIYATO et al., 2018; He et al., 2019; XIE et al., 2020a). By training the student model on noisy variations of the data, the model is encouraged to learn more generalized representations, improving its performance on unseen data (CARMON et al., 2019a; ALAYRAC et al., 2019; NAJAFI et al., 2019). Another way to observe the advantages of training with noisy data is by observing the decision boundaries of the model. Adversarial examples lie on the frontiers of the decision boundary, thus, by fitting the model with the original data alongside the adversarial inputs, the model is capable of learning a smoother decision boundary that clearly separates distinct classes (see Figure 1).

Input	**Output**
All other religions should be extinct except mine	Offensive
All other religions should be extint except mine	Not offensive
All other religions should be 3xt1nct except mine	Not offensive
My religion should not be extinct, but all others should	Not offensive

Table 1 – Examples of adversarial inputs. The model is unable to correctly predict variations of the same input. Entries are as follows: (i) original input, (ii) word "extinct" is misspelled, (iii) word "extinct" written in leetspeak, (iv) original input is paraphrased.

A high-level description of the noisy student model can be summarized as follows:

1. **Train the Teacher Model**: A model is initially trained on the available gold-labeled data. There is no noise added to the teacher model, neither architectural noise (e.g. dropout) nor data noise (e.g. data augmentations).

2. **Generate silver labels**: The trained teacher model is used to infer the unlabeled dataset.

3. **Add noise to the silver-labeled dataset**: Use a method to introduce noise to the silver-labeled dataset (e.g. apply data augmentations).

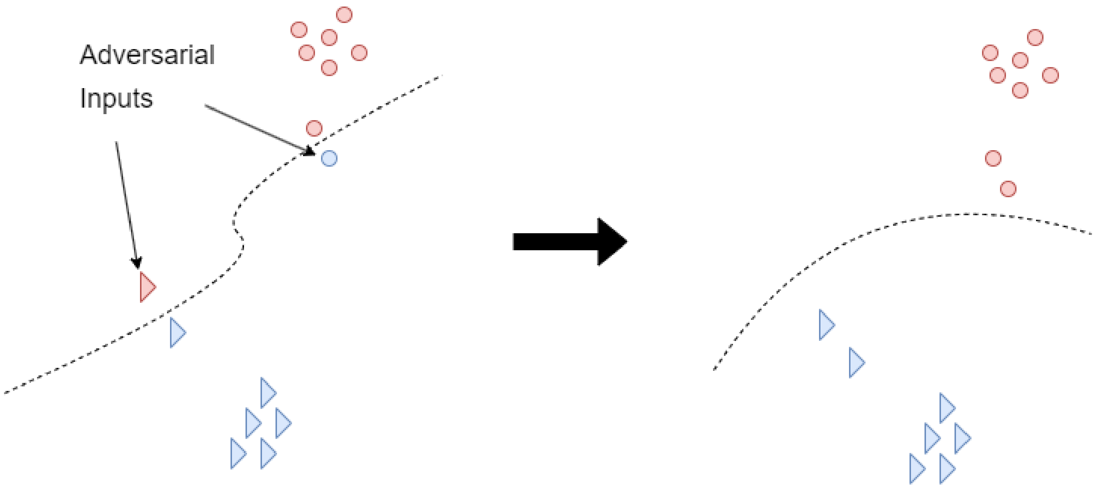

Figure 1 – Decision boundary smoothing through self-training with adversarial inputs. Triangles and circles represent distinct classes. Red and blue represent distinct model predictions.

4. **Train the Student Model**: A new model (the student) is trained on the combined gold and silver-labeled data, with noise introduced both in the architecture of the model (e.g. dropout) and in the input data (through augmenting the silver-labeled dataset).

5. **Iterate**: Repeat from 2. now using the current student model as the teacher.

More formally, the noisy student training procedure can be found in Algorithm 3.

Algorithm 3 Noisy Student Training Algorithm

```
Require: Ground-truth dataset $D_{gold}$, Unlabeled dataset $D_{unlabeled}$, No. iterations $N$
$M_{teacher} \leftarrow \text{Train}(D_{labeled})$
for $i = 1$ to $N$ do
    $D_{silver} \leftarrow \text{Infer}(M_{teacher}, D_{unlabeled})$
    $D_{silver\_noisy} \leftarrow \text{ApplyDataAugmentation}(D_{silver})$
    $M_{student} \leftarrow \text{Train}(D_{gold} \cup D_{silver\_noisy})$
    $M_{teacher} \leftarrow M_{student}$
end for
```

Finally, there are two implementation details worth noting. Firstly, it is common to compute the training *loss* L as a combination of the *loss* with respect to the gold (L_{gold}) and the silver (L_{silver}) labeled data (see Equation 3.1). This is done to ensure that L is normalised by the amount of training samples in each dataset. If this normalisation is not done, then $L \approx L_{silver}$, since typically the silver-labeled dataset is much larger than the gold-labeled dataset. Lastly, when inferring the unlabeled dataset, it is common to discard examples predicted with low confidence. This is done to reduce the likelihood that incorrect predictions are added to the silver-labeled dataset. A common practice is to determine a threshold (e.g. confidence $> 80\%$) and discard examples whose model confidence is lower than this amount.

$$L = \frac{1}{n}\sum_{i=1}^{n} L_{\text{gold}} + \frac{1}{m}\sum_{i=1}^{m} L_{\text{silver}} \tag{2.1}$$

2.4 Data Augmentation Methods for Textual Data

In the context of machine learning, data augmentation involves generating new data samples from the existing dataset to increase its size and variability. This process is especially beneficial when dealing with limited or imbalanced datasets, as it helps in preventing overfitting and improves the generalization capability of the models (MUMUNI; MUMUNI, 2022). Specifically for the domain of self-supervised methods, generating augmented examples is closely related to increasing the adversarial robustness of the model. As discussed previously in Section 2.3, adversarial robustness refers to a model's ability to maintain performance when exposed to similar (but not exactly equal) examples that cause the model to make incorrect predictions. In this scenario, data augmentation contributes significantly to enhancing adversarial robustness by simulating potential perturbations during training, thereby preparing the model to handle them during inference (CARMON et al., 2019b).

When it comes to textual data, data augmentation presents unique challenges and opportunities compared to other types of data, such as images. Textual data is inherently discrete, structured, and context-dependent, making it more susceptible to semantic shifts and the introduction of noise. Despite these challenges, several effective data augmentation methods have been developed and widely adopted for textual data (BAYER et al., 2022). This section will present the main methods used for data augmentation in natural language processing (NLP) tasks, specifically focusing on their application and efficacy in various scenarios. A non-exhaustive list of common textual data augmentation techniques is presented below:

Synonym substitution is a straightforward yet effective data augmentation technique where words in a text are replaced with their synonyms. This method relies on lexical resources such as WordNet or word embedding models to find suitable synonyms that do not significantly alter the meaning of the original text. Synonym substitution helps in increasing the lexical diversity of the dataset and is commonly used in sentiment analysis and offensive language detection tasks.

Word swap involves randomly shuffling or swapping words within a text to create new variations. This method aims to maintain the overall structure and meaning of the text while introducing syntactic variations. Word swap can be effective for tasks where syntactic diversity is important, such as part-of-speech tagging and syntactic parsing.

Random Insertion and Deletion involves randomly adding or removing words from a text. Similar in terms of applicability to the previous two methods above.

Backtranslation is a popular data augmentation technique for textual data, which involves translating a text from the source language to a target language and then translating it back to the source language. This method leverages the variability introduced by the translation process to generate paraphrases of the original text, thereby enriching the dataset without altering its semantic content. Backtranslation is particularly effective for tasks requiring robust semantic understanding, such as text classification, machine translation, and text summarization.

Chapter 3

Noisy Self-Training with Data Augmentations for Offensive and Hate Speech Detection Tasks

This chapter includes the research paper that is the core result of the Master's program. The paper "Noisy Self-Training with Data Augmentations for Offensive and Hate Speech Detection Tasks" has been published in the *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP)* (LEITE et al., 2023).

Abstract

Online social media is rife with offensive and hateful comments, prompting the need for their automatic detection given the sheer amount of posts created every second. Creating high-quality human-labelled datasets for this task is difficult and costly, especially because non-offensive posts are significantly more frequent than offensive ones. However, unlabelled data is abundant, easier, and cheaper to obtain. In this scenario, self-training methods, using weakly-labelled examples to increase the amount of training data, can be employed. Recent "noisy" self-training approaches incorporate data augmentation techniques to ensure prediction consistency and increase robustness against noisy data and adversarial attacks. In this paper, we experiment with default and noisy self-training using three different textual data augmentation techniques across five different pre-trained BERT architectures varying in size. We evaluate our experiments on two offensive/hate-speech datasets and demonstrate that (i) self-training consistently improves performance regardless of model size, resulting in up to +1.5% F1-macro on both datasets, and (ii) noisy self-training with textual data augmentations, despite being successfully applied in similar settings, decreases performance on offensive and hate-speech domains when compared to the default method, even with state-of-the-art augmentations such as backtranslation.

3.1 Introduction

Online social media platforms are widely used by modern society for many productive purposes. However, they are also known for intensifying offensive and hateful comments, attributed in part to factors such as user anonymity (MONDAL et al., 2017b). Manual identification of hate speech is impractical at scale due to the massive number of posts generated every second and the potential harm to the mental health of moderators. Therefore, there is a need for automatic approaches to detect offensive and hateful speech.

In recent years, research on this topic has increased, resulting in new models and datasets published in various languages and sources (FORTUNA; NUNES, 2018). A common characteristic among available datasets is label skewness towards the negative class (non-offensive/hateful), which is usually more frequent than the positive class (offensive/hateful). Apart from traditional ways of dealing with imbalanced classes (e.g. under or oversampling or applying class weighting), semi-supervised techniques such as self-training can be used to extend the training set with unseen examples that introduce new learning signals without the costly burden of manual data labeling.

Self-training is a technique that involves iteratively training models using both labelled and unlabelled data. The process begins by training a model using human-labelled data only, which is then used to infer labels for a set of unlabelled data, creating a weakly-labelled dataset. The weakly-labelled dataset and the human-labelled dataset are then aggregated and used to retrain the model. This iterative process is repeated for a fixed number of steps or until no performance improvement is observed. Self-training can be particularly useful when labelled data is scarce or expensive to obtain, and was successfully applied in a variety of domains such as computer vision (SCHIAPPA et al., 2022), audio and speech processing (LIU et al., 2022), and natural language processing (He et al., 2019).

Several variants of self-training have been proposed over the years (AMINI et al., 2022). One common approach is to use a teacher-student framework, in which the "student" model learns from the output generated by the "teacher" model (BLUM; MITCHELL, 1998; XIE et al., 2020b; CHEN et al., 2021; KARAMANOLAKIS et al., 2021). Additionally, a confidence threshold filter may be applied to remove examples that are too ambiguous or non-informative. This process is summarised in Figure 2.

Recent research on self-training has reported further improvements in performance by introducing perturbations directly into the raw input or to its latent representation, improving generalisation and convergence (RASMUS et al., 2015; LAINE; AILA, 2017; MIYATO et al., 2018; He et al., 2019; XIE et al., 2020a). These perturbations are often introduced in the form of data augmentations, which are widely applied in Computer Vision tasks but are less commonly explored in Natural Language Processing tasks, especially in the context of self-training. These "noisy self-training" methods can be particularly useful in settings where the input data is noisy or

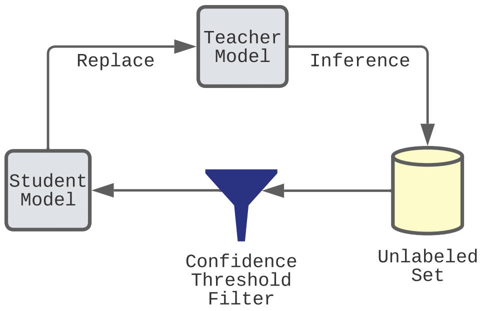

Figure 2 – Teacher-student self-training loop

subject to a high degree of variation, improving prediction consistency and adversarial robustness (CARMON et al., 2019a; ALAYRAC et al., 2019; NAJAFI et al., 2019).

Bayer et al. (2022) argue that data augmentation depends on the underlying classification task, thus it cannot be effectively applied in all circumstances. Previous work focusing solely on data augmentation methods, not coupled with self-training, has shown mixed results for the domain of offensive/hate speech classification (Section 3.2.1). This indicates that there may not be a best method, while some may even negatively impact performance.

An open question is whether noisy self-training with text data augmentations can contribute to text classification tasks using state-of-the-art transfer-learning BERT models that have been shown to be invariant to various data transformations (LONGPRE et al., 2020). The task of offensive/abusive speech detection poses a difficult challenge for generating high-quality semantic invariant augmented examples, since it is a domain that is intrinsically associated with specific keywords that, if modified, can completely change the semantics of the text. In this paper, we innovate by providing an extensive experimentation setup using three different data augmentation techniques - backtranslation, random word swap, and random synonym substitution - in a self-training framework, with five different pre-trained BERT architectures varying in size, on two different datasets.

We demonstrate that self-training, either with or without data noising, outperforms default fine-tuning regardless of model size, on both datasets. However, when comparing self-training without data noising vs ‘noisy’ self-training, we find that data augmentations decrease performance, despite the literature reporting the superiority of noisy self-training in other domains. We further investigate how the augmentation methods fail to create label-invariant examples for the offensive/hate speech domain. Finally, we discuss future research ideas to address the limitations found in this work.

3.2 Related Work

3.2.1 Data Augmentation

Bayer et al. (2022) present a survey on data augmentation methods for NLP applications, reporting performance gains on various tasks. In the domain of offensive/hate speech classification, Ibrahim et al. (2018) experiment with three different text augmentation techniques to expand and balance their Wikipedia dataset by augmenting negative (non-offensive) examples. From a binary view of the dataset, more than 85% of their examples are labelled as non-offensive, and from a multi-label view of the dataset, three of the six offensive classes are represented by less than 7% of the dataset. They report F1-score increases of +1.4% with unique words augmentation, +2.9% with unique words and random mask, and +3.6% with unique words, random mask, and synonym replacement.

Mosolova et al. (2018) use a custom synonym replacement augmentation method to experiment with a ‘toxic’ dataset with 6 classes from a Kaggle competition[1]. They experiment with character and word embeddings with a CNN architecture, and report a +3.7% and +5.1% ROC-AUC increase when applying their augmentation method with character embeddings on the public and private scores[2], respectively. However, when coupled with word embeddings, they find that their augmentations result in a decrease of -0.09% and -0.21% ROC-AUC scores on the public and private scores, respectively.

Rizos et al. (2019) propose three text-based data augmentation techniques to address the class imbalance in datasets, and apply them on three English hate speech datasets named HON (DAVIDSON et al., 2017), RSN-1 (WASEEM; HOVY, 2016) and RSN-2 (WASEEM, 2016). Their augmentation methods include (i) synonym replacement based on word embedding, (ii) warping of the token words along the padded sequence, and (iii) class-conditional RNN language generation. They compare the three methods on different architectures combining word embeddings, CNNs, GRUs, and LSTMs, and they report an average across four different architecture configurations of -6.3% F1-Macro using (i), +5% F1-Macro using (ii), and -4% F1-Macro using (iii).

Marivate e Sefara (2020) experiment with four different data augmentation techniques: WordNet synonym substitution, backtranslation between German and English, word embedding substitution according to cosine similarity, and mixup (ZHANG et al., 2018). Authors experiment with three datasets from different domains: Sentiment 140 (GO et al., 2009), AG News (ZHANG et al., 2015) and a Hate Speech dataset (DAVIDSON et al., 2017). They observe performance increases on both Sentiment 140 and AG News across different augmentation methods, up to +0.4% and +0.5% accuracy score on AG News and Sentiment 140, respectively. However, they

[1] <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>

[2] Public scores are computed over a smaller portion of the test set. At the end of the competition, private scores are computed with the remainder of the test set.

report performance decreases with all methods on the Hate Speech dataset, with decreases of 0.0% with mixup, -0.3% with embedding similarity, -0.8% with synonym substitution, and -2.3% with backtranslation.

3.2.2 Self-Training

Xie et al. (2020b) present a method called *noisy student*, which achieves state-of-the-art results on the ImageNet dataset (DENG et al., 2009) by performing self-training with a teacher-student approach, using student models that are equal or larger-sized than the teacher models, and adding noise both to the input data through random image augmentations and to the model via dropout.

He et al. (2019) apply a similar idea using textual data augmentation methods such as backtranslation (EDUNOV et al., 2018) and token modifications to a self-training LSTM architecture for the tasks of machine translation and text summarization. They find that both model noise, in the form of dropout, and data noise, in the form of data augmentations, are crucial to their observed increase in performance on both tasks.

Xie et al. (2020a) use six text classification and two image classification benchmark datasets to experiment with different types of noise-inducing techniques for self-training. They argue that state-of-the-art augmentations like backtranslation for text classification and RandAugment (CUBUK et al., 2020) for image classification, outperform simple noise inducing techniques, such as additive Gaussian noise.

The use of noisy self-training approaches in the domain of offensive/hate speech classification is still limited, but default ‘non-noisy’ self-training has been successfully applied in some recent works. Alsafari e Sadaoui (2021) collect unlabelled Arabic tweets and perform semi-supervised classification with self-training for the domain of Offensive and Hate Speech detection using multiple text representations such as N-grams, Word2Vec, AraBert and Distilbert, and multiple model architectures such as SVM, CNN and BiLSTM. They report up to 7% performance increase in low resource settings where only a few labelled examples are available.

Leonardelli et al. (2020) apply self-training in their submission to the HaSpeeDe shared task on Italian hate speech detection (task A). They fine-tune an AlBERTo model with the human-labelled dataset provided by the task organisers and extend it with a weakly-labelled dataset using self-training. Additionally, they oversample the human-labelled set in an attempt to make the model more robust to inconsistencies in the weakly-labelled set. Their submission achieve an F1-macro score of 75.3% on tweets, placing 11th out of 29 teams, and 70.2% on news headlines, placing 5th out of 29 teams.

Pham-Hong e Chokshi (2020) report experiments with the noisy student method from Xie et al. (2020b) in the OffensEval 2020 shared task, achieving 2nd place at subtask B (Automatic categorization of offense types). In their setup, although dropout is applied to a BERT-large model, no noise is injected into the data, which is a crucial component of the noisy student method. Because of this, we argue that this work is actually applying a default self-training method instead of a noisy self-training method. Also, OffensEval 2020's training data does not contain human-labelled data[3], thus both their weakly-labelled dataset and ground-truth dataset consist of inferred examples.

Richardson et al. (2022) detect hate speech on Twitter in the context of the Covid-19 pandemic. They employ a simple approach, utilizing a bag-of-words representation combined with an SVM classifier. Authors demonstrate that by employing self-training with only 20% of the training data, they manage to improve accuracy by +1.55% compared to default training using 80% of the training data.

To the best of our knowledge, Santos et al. (2022) is the only previous work in which a **noisy** self-training approach was attempted on an offensive/hate speech classification task. They propose an ensemble of two semi-supervised models to create FIGHT, a Portuguese hate speech corpus. Authors combine GANs, a BERT-based model, and a label propagation model, achieving 66.4% F1-score. They attempt to increase performance using backtranslation as data augmentation, but ultimately observe no performance gains, thus their best model is obtained with default self-training, not with noisy self-training.

3.3 Materials and Methods

This section presents the description of the datasets, data augmentation methods and self-training architectures used throughout our experiments. Our code is available at GitHub[4].

3.3.1 Data Description

We use two English binary offensive/hate speech detection datasets in our experiments. Table 2 presents their target class distributions.

	OLID		
	Train	Dev	Test
Not-Offensive	8,840	0	620
Offensive	4,400	0	240
	ConvAbuse		
	Train	Dev	Test
Not-Offensive	2,163	719	725
Offensive	338	112	128

Table 2 – Target class distribution for OLID and ConvAbuse.

[3] In OffensEval 2020, the labels in the training data are the average confidence score and confidence standard deviation aggregated from an ensemble of models.

[4] <https://github.com/JAugusto97/Offense-Self-Training>

Offensive Language Identification Dataset (OLID)

Zampieri et al. (2019) contains a collection of annotated tweets following three levels: Offensive Language Detection, Categorization of Offensive Language, and Offensive Language Target Identification. This work only uses the first level - Offensive Language Detection. The dataset was normalised by replacing URLs and user mentions with placeholders. The best model in Zampieri et al. (2019) achieves 80% macro-$F1$ using convolutional neural networks, with 70% and 90% of $F1$-Score for the positive and negative classes, respectively.

ConvAbuse

Curry et al. (2021) is a dataset on abusive language towards three conversational AI systems: an open-domain social bot, a rule-based chatbot, and a task-based system. Authors find that the distribution of abuse towards conversational systems differs from other commonly used datasets, with more than 50% of the instances containing sexism or sexual harassment. To normalise the data, web addresses were replaced with a placeholder. Authors provide standard train, development, and test sets and achieve up to 88.92% macro-$F1$ using a fine-tuned BERT model. In our experiments, we concatenate the interactions between the user and the chatbot into a single text document divided by new line separators, and we use majority voting between the annotations to consolidate the binary abusive vs. non-abusive label.

Unlabelled data

We collected 365,456 tweets in English with the Twitter API using an unbiased query rule: random tweets mentioning stop-words like “in”, “on”, “a”, “is”, “not”, “or” and so on. We also preprocess the data by removing user mentions, urls, punctuations, extra whitespace and accents.

3.3.2 Self-Training Architecture

Our noisy self-training system is similar to that introduced by Xie et al. (2020b) and Xie et al. (2020a), and works as follows:

1. A teacher model is trained to minimise the cross-entropy loss on the human-labelled training set exclusively.
2. The teacher model infers weak labels from the unlabelled dataset.
    - A confidence threshold filter is applied, and examples that fall below this threshold are removed.
    - Apply *downsampling* on the inferred examples, ending up with a perfectly balanced weakly-labelled dataset.

3. All the examples selected from the previous step are augmented once with one of the data augmentation methods, doubling the amount of weakly-labelled examples. The labels obtained with the 'clean/without noise' text in step 2 are replicated for the augmented texts.

4. An equal-sized student model minimises the combined cross-entropy loss on human-labelled and weakly-labelled datasets:

$$L = \frac{1}{n}\sum_{i=1}^{n} L_{\text{labelled}} + \frac{1}{m}\sum_{i=1}^{m} L_{\text{inferred}} \tag{3.1}$$

5. Repeat from step 2 using the current student model as the teacher model.

In our experiments, we compare this noisy self-training framework against the default 'non-noisy' self-training method, which simply skips step 3, meaning we do not apply any form of data augmentation.

3.3.3 Data Augmentation Methods

In each noisy self-training experiment we use `nlpaug`[5] to apply one of the three following data augmentation methods for textual data:

Random Synonym Substitution

Uses WordNet (MILLER, 1995) to randomly replace tokens by one of its synonyms. For each sentence, 30% of its tokens will be replaced.

Random Word Swap

Randomly swaps adjacent tokens in a sentence. For each sentence, 30% of its tokens are swapped.

Backtranslation

First translates the original texts into a second language, then translates them back from the second language to the original language. We use the backtranslation model from `nlpaug`, which uses the two different transformer models from Ng et al. (2019) to translate the data from English to German, then from German back to English.

[5] `<https://github.com/makcedward/nlpaug>`

3.4 Experimental Setup

Firstly, we experiment with each dataset to estimate the hyperparameters for the base models, which is the first teacher models in the self-training loop. We use a batch size of 128, maximum sequence length of 128, learning rate of 0.00001, 15% of the training set as warm-up batches, weight decay of 0.001 and 20 training epochs. We apply a dropout rate of 10% for both the attention and classification layers. The model with highest validation F1-macro score[6] obtained during training is loaded at the end of the last epoch. For the hyperparameters associated with the self-training method, we set the number of teacher-student iterations to 4 (including the first teacher model) and a confidence threshold filter of 80%, similarly to Xie et al. (2020a). Also, we experiment with five different pre-trained BERT models: DistilBERT, BERT-base-cased, BERT-large-cased, RoBERTa-base and RoBERTa-large, aiming to investigate the impact of model size in performance gains associated with self-training.

From the above-listed configurations, we designed two main classification scenarios. The first scenario accounts for a regular self-training loop without data noise injection through augmentations, while the second scenario uses the noisy self-training approach, introducing data noise with one of the three augmentation methods described in Section 3.3.3.

Finally, we conduct a deeper analysis of each augmentation method. We use the first teacher model, trained exclusively with the human-labelled data of each dataset, to infer both the ’clean/without augmentation’ and the ’noisy/augmented’ versions of the unlabelled dataset and verify the following: (i) Does the augmentation method create new tokens that are not present in the vocabulary of the ’clean/without augmentation’ unlabelled dataset? and (ii) Are the augmentations semantically invariant, meaning both the ’clean’ and ’noisy’ pairs of examples are assigned the same label?

3.5 Results

3.5.1 Default Fine-Tuning vs. Self-Training

Table 3 displays the mean and standard deviation $F1$-macro scores computed over three different random seed initializations for each experiment. Note that self-training, regardless of whether coupled with data augmentation methods or not, improves over default fine-tuning for every model architecture, increasing the F1-macro score from +0.7% up to +1.5% on OLID and +0.8% up to +1.5% on ConvAbuse depending on the pre-trained model architecture.

Also, we highlight how self-training can make smaller models, which require fewer resources to maintain in practical applications, achieving the same performance as larger and more costly models that are trained with default fine-tuning. Self-training on a DistilBERT (66M parameters) outperforms a BERT-large-cased (340M parameters) with default fine-tuning

[6] Lowest training loss in the case of OLID, since no development set is provided.

		OLID			
Architecture	DF	ST	ST + BT	ST + SS	ST + WS
DistilBERT	78.4 ± 0.1	**79.2 ± 0.2**	79.0 ± 0.3	79.0 ± 0.3	79.0 ± 0.3
BERT-base-cased	77.2 ± 0.3	**78.7 ± 0.1**	78.1 ± 0.1	78.3 ± 0.3	78.3 ± 0.3
BERT-large-cased	79.2 ± 0.2	**80.0 ± 0.3**	79.4 ± 0.1	79.3 ± 0.3	79.3 ± 0.3
RoBERTa-base	79.4 ± 0.7	**80.1 ± 0.3**	80.0 ± 0.4	80.0 ± 0.4	80.0 ± 0.4
RoBERTa-large	79.8 ± 0.3	80.4 ± 0.4	80.3 ± 0.4	**80.7 ± 0.7**	**80.7 ± 0.7**

		ConvAbuse			
Architecture	DF	ST	ST + BT	ST + SS	ST + WS
DistilBERT	85.7 ± 0.5	86.8 ± 0.3	87.1 ± 0.3	**87.2 ± 0.3**	**87.2 ± 0.3**
BERT-base-cased	86.8 ± 0.8	**87.6 ± 0.1**	87.2 ± 0.5	87.2 ± 0.5	87.2 ± 0.5
BERT-large-cased	87.1 ± 0.6	**87.9 ± 0.5**	87.4 ± 0.2	**87.9 ± 0.5**	**87.9 ± 0.5**
RoBERTa-base	84.5 ± 0.3	**85.5 ± 0.4**	85.3 ± 0.8	85.4 ± 0.5	85.4 ± 0.5
RoBERTa-large	86.0 ± 0.1	86.2 ± 0.3	86.6 ± 0.3	**86.9 ± 0.1**	86.8 ± 0.1

Table 3 – Mean ± 1 std F1-Macro scores obtained over three random seed initializations. DF=Default Fine-Tuning, ST=Self-Training, BT=Backtranslation, SS=Synonym Substitution, WS=Word Swap

on both OLID and ConvAbuse. On OLID, a RoBERTa-base architecture (125M parameters) with self-training outperforms a RoBERTa-large (354M parameters) architecture with default fine-tuning, although this does not hold true for ConvAbuse.

Furthermore, we point out that OLID and ConvAbuse's data come from different sources, the first being Twitter, and the second one representing conversations between humans and chatbots, thus their structure differs significantly. Since our unlabelled dataset is composed of Twitter data, it would be fair to assume that the benefits of self-training in our experiments would be more prominent for the OLID dataset, but our results do not show this, since models trained with ConvAbuse benefited from self-training with our Twitter-originated unlabelled dataset just as much as models trained with OLID.

3.5.2 Default Self-Training vs. Noisy Self-Training

After verifying that self-training is beneficial to both datasets on all model architectures, we compare default self-training with noisy self-training, and the impacts of adding data noise in the form of data augmentations. We find that introducing data augmentations to the self-training pipeline increases performance against default self-training only for RoBERTa-large on both OLID and ConvAbuse, with DistilBERT also showing improvements for ConvAbuse, but not for OLID. On all other architectures, for both datasets, default self-training without data augmentations achieves the highest scores.

In our results for offensive/hate speech classification, backtranslation does not achieve the highest score in any setup, while synonym substitution and word swap tie for highest score in three scenarios: ConvAbuse with DistilBERT, ConvAbuse with BERT-large-cased, and OLID with RoBERTa-large. Synonym substitution outperforms all the remaining methods on ConvAbuse with RoBERTa-large.

An important remark is that our results diverge from He et al. (2019), which finds that state-of-the-art data augmentation methods such as backtranslation outperform simpler methods on self-training for machine translation and text summarization. However, our results align with Marivate e Sefara (2020), although their work is not focused on self-training, but instead on how different data augmentation techniques impact their models on three datasets from different domains. They report backtranslation as their worst augmentation method on a hate speech dataset, decreasing accuracy by -2.3%. Our findings bridge this gap and reveal that backtranslation has significant limitations in the domain of offensive/hate speech detection, even when used in a noisy self-training approach.

3.5.3 Data Augmentation Analysis

Our first data augmentation analysis is to understand if the augmented text introduces new unseen tokens to the vocabulary of the 'clean' unlabelled set when both are combined. We find a vocabulary size increase of 39.5%, 9.0% and 4.7% averaging across all different pre-trained architectures for backtranslation, synonym substitution and word swap[7] respectively. This indicates that backtranslation is heavily superior in terms of introducing new unseen tokens, but this is not correlated with performance increase, as backtranslation appears as the worst augmentation method for noisy self-training in our classification experiments.

Next, in order to verify the performance of the data augmentation methods in generating semantically invariant examples, we use the base models trained exclusively with the human-labelled data from each dataset, on each pre-trained architecture, and use them to perform inference on both the 'clean' and the noisy/augmented unlabelled set. We then compare both predictions and analyse how augmentations may shift the underlying target class. We will refer to **positive shift** when a non-offensive example is classified as offensive after being augmented, and **negative shift** when an offensive example is classified as non-offensive after being augmented.

Table 4 presents the total class shift percentage for each augmentation method, averaging across both datasets and all model architectures, of which we further divide into positive and negative label shift percentages. Notice that backtranslation is the method that produces the highest amount of label shifting at 23.8%, of which 54.7% are negative shifts, which is a 6.6% increase over synonym substitution and a 4.8% increase word swap.

It is fair to assume that not all of the class shifting occurs from the augmentation changing the semantic that defines if an example is either offensive or not-offensive. In most cases, class shifting may occur because of small perturbations that are semantically invariant, meaning

[7] Word swap is unintuitively capable of creating new tokens depending on how a sentence is split into tokens and then merged back after swapping the tokens.

Augmentation	Total Shift	Positive Shift	Negative Shift
BT	23.8%	46.7%	54.7%
SS	23.5%	48.7%	51.3%
WS	23.3%	47.8%	52.2%

Table 4 – Average target class shift percentage on the weakly-labelled set. BT=Backtranslation, SS=Synonym Substitution, WS=Word Swap

Text	Augmented Text	Method
I HATE ALL OF YOU	ALL I HATE OF YOU	WS
Maybe I dont respect all women	Maybe I respect dont women all	WS
Bitches and sports	Females and Sport	BT
Wooooow what the fuck	Wooooow, what the hell?	BT
Bitch you better be joking	Gripe you good be joking	SS
The NYT has been showing its whole ass [...]	The NYT has follow showing its whole butt [...]	SS

Table 5 – Examples of Offensive to Not-Offensive semantic shift created by data augmentation. BT=Backtranslation, SS=Synonym Substitution, WS=Word Swap

Text	Augmented Text	Method
Is that Fat Albert	That Fat is Albert	WS
Man that is terrible	That man is terrible	WS
damn white people oppressing the blacks	fucking white people who oppress the blacks	BT
That damn staircase be beating my ass [...]	That fucking staircase will bang my ass [...]	BT
i will not get over this	i will not fuck off ended this	SS
Send me the link and Ill love you forever	Send pine tree state the link and Ill fuck you forever	SS

Table 6 – Examples of Not-Offensive to Offensive class shift created by data augmentation. BT=Backtranslation, SS=Synonym Substitution, WS=Word Swap

both the 'clean' and the augmented text's true underlying classes are still the same, even if the classifier predicted them as different classes. In these cases, when we set the label of the augmented text to be the same as the one obtained when inferring the 'clean' version of the text, as presented in section 3.3.2, we are reinforcing the model to be more robust against these small perturbations, which is one of the main benefits of noisy self-training. However, when augmentation methods create semantically different versions of the original texts, replicating the inferred label from the original text to the augmented text results in the addition of incorrect ground-truth labels to the train set, which may degrade performance.

Currently, to the best of our knowledge, there is no dataset annotated for offense/hate speech before and after applying data augmentation, which would enable a more accurate estimation of semantic variations produced by them. In tables 5 and 6 we show two examples for each augmentation method that suffered from positive shift (not-offensive to offensive) and negative shift (offensive to not-offensive), respectively.

An example of a recurrent theme among various target shifted examples is the substitution of the keywords 'fuck' with 'damn' or 'hell', indicating that despite these keywords being semantically similar, they are not always interchangeable with respect to the target class, and the mere replacement of one for another is enough to shift the target class. This could be expected, as offense detection is highly impacted by the mere presence or absence of offensive keywords.

3.6 Conclusion

In this work, we analysed the impact of self-training on offensive and hate speech classification tasks using five different pre-trained BERT models of varying sizes and two different datasets. We also experimented with noisy self-training using three different data augmentation techniques for textual data. We found that self-training improves classification performance for all model architectures on both datasets, with an increase in F1-Macro of up to +1.5%. However, our experiments comparing default self-training versus noisy self-training showed that noisy self-training does not improve performance, despite its success in other domains. Finally, we investigated the three data augmentation methods and showed that the domain of offensive/hate speech classification is highly sensitive to semantic variances produced by them, and we discussed future research ideas to mitigate these problems.

3.7 Future Work

We understand that some of the semantic variations discussed in this work could be mitigated by data augmentation methods that both preserve existing offensive keywords, and do not introduce new offensive keywords randomly, as these are often conditional to the underlying ground-truth class. For some languages, most of these keywords are extensively documented[8], thus they can be known a priori by these methods, and be treated differently, such as only substituting an offensive keyword by another offensive keyword, or not allowing a non-offensive keyword to be substituted by an offensive keyword. This custom approach can theoretically help mitigate semantic variations in this domain, but offensive/hateful comments can still be made without making use of a single offensive/hateful keyword. In these more subtle cases, a system would have to detect the offensive/hateful context without relying solely on keywords, and modify the example while still maintaining this context. We see potential benefits of using recent instruction-tuned large language models (OUYANG et al., 2022) as specialised data augmentation methods that are task-specific, and can be able to preserve the semantics associated with the task when modifying a given text. In this scenario, an instruction prompt can be designed to inform the system of the context of the task, and make it aware that this semantic must be preserved when modifying the given text. In the future, we aim towards extending this work with the above-mentioned research ideas.

[8] <https://hatebase.org/>

Chapter 4
Conclusion

This thesis explored the efficacy of noisy self-training with data augmentations for the task of offensive and hate speech detection. The research presented in the core paper "Noisy Self-Training with Data Augmentations for Offensive and Hate Speech Detection Tasks" provided a comprehensive experimental setup to study the supposed benefits of applying data noising in the form of data augmentations, specifically for the domain of offense and hate speech classification. Firstly, the study underscored the critical need for automated systems to detect offensive and hate speech on social media platforms, given the limitations of manual moderation and the psychological toll it can take on human moderators. This serves as motivation to incorporate automated systems that leverage training data without the need for human annotators. For the purpose of this work, we studied self-training methods that leverage an intermediate model (teacher) to infer labels for a second model (student) that is superior to its predecessor. Recent work using approach had been increasingly focusing on self-training methods that make use of noising techniques, due to it's success in several different domains ().

The core of the thesis, detailed in Chapter 3, presented a comprehensive methodology that incorporated several common data augmentation methods with noisy self-training, which were applied on different transformer-based models in varying sizes. Results indicated that self-training improves classification performance for all model sizes on both datasets considered, with an increase in F1-Macro of up to +1.5%. However, when comparing the default self-training method versus noisy self-training, the results indicate that there is no benefit in adding data noise through data augmentation. This finding is relevant especially since this is not observed in other similar tasks (cite). To further investigate the particular characteristics of the domain of offensive and hateful speech classification, an analysis was performed on the three data augmentation methods considered, highlighting that this domain is highly sensitive to semantic variances produced by these data augmentation methods. In conclusion, the advancements presented in this thesis contribute to the ongoing efforts to develop more effective and scalable automated moderation systems for online platforms.

Future work could benefit from these findings by exploring the integration of more sophisticated data augmentation techniques that take in consideration the task-specific keywords that cannot be modified without semantic shift. Several of the semantic shifts encountered could be mitigated by data augmentation methods that both preserve existing offensive keywords, and do not introduce new offensive keywords, as these are often conditional to the underlying ground-truth class. For some languages, most of these keywords are extensively documented[1], thus they can be known a priori by these methods, and be treated differently, such as only substituting an offensive keyword by another offensive keyword, or not allowing a non-offensive keyword to be substituted by an offensive keyword. This custom approach can theoretically help mitigate semantic variations in this domain, but offensive/hateful comments can still be made without making use of a single offensive/hateful keyword. In these more subtle cases, a system would have to detect the offensive/hateful context without relying solely on keywords, and modify the example while still maintaining this context. Also, there are potential benefits in using recent instruction- tuned large language models (OUYANG et al., 2022) as specialised data augmentation methods that are task-specific, and can be able to preserve the semantics associated with the task when modifying a given text. In this scenario, an instruction prompt can be designed to inform the system of the context of the task, and make it aware that this semantic must be preserved when modifying the given text.

[1] <https://hatebase.org/>

BIBLIOGRAPHY

ALAYRAC, J.-B.; UESATO, J.; HUANG, P.-S.; FAWZI, A.; STANFORTH, R.; KOHLI, P. Are labels required for improving adversarial robustness? In: WALLACH, H.; LAROCHELLE, H.; BEYGELZIMER, A.; ALCHé-BUC, F. d'; FOX, E.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. v. 32. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2019/file/bea6cfd50b4f5e3c735a972cf0eb8450-Paper.pdf>. Cited 2 times on pages 12 e 18.

ALSAFARI, S.; SADAOUI, S. Semi-supervised self-training of hate and offensive speech from social media. *Applied Artificial Intelligence*, Taylor Francis, v. 35, n. 15, p. 1621–1645, 2021. Disponível em: <https://doi.org/10.1080/08839514.2021.1988443>. Cited on page 20.

AMINI, M.-R.; FEOFANOV, V.; PAULETTO, L.; DEVIJVER, E.; MAXIMOV, Y. Self-training: A survey. *arXiv preprint arXiv:2202.12040*, 2022. Cited 2 times on pages 10 e 17.

BAYER, M.; KAUFHOLD, M.-A.; REUTER, C. A survey on data augmentation for text classification. *ACM Computing Surveys*, ACM New York, NY, v. 55, n. 7, p. 1–39, 2022. Cited 3 times on pages 14, 18 e 19.

BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. New York, NY, USA: Association for Computing Machinery, 1998. (COLT' 98), p. 92–100. ISBN 1581130570. Disponível em: <https://doi.org/10.1145/279943.279962>. Cited 3 times on pages 8, 11 e 17.

CARMON, Y.; RAGHUNATHAN, A.; SCHMIDT, L.; DUCHI, J. C.; LIANG, P. S. Unlabeled data improves adversarial robustness. In: WALLACH, H.; LAROCHELLE, H.; BEYGELZIMER, A.; ALCHé-BUC, F. d'; FOX, E.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. v. 32. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2019/file/32e0bd1497aa43e02a42f47d9d6515ad-Paper.pdf>. Cited 2 times on pages 12 e 18.

CARMON, Y.; RAGHUNATHAN, A.; SCHMIDT, L.; DUCHI, J. C.; LIANG, P. S. Unlabeled data improves adversarial robustness. In: WALLACH, H.; LAROCHELLE, H.; BEYGELZIMER, A.; ALCHé-BUC, F. d'; FOX, E.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. v. 32. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2019/file/32e0bd1497aa43e02a42f47d9d6515ad-Paper.pdf>. Cited on page 14.

CHEN, X.; YUAN, Y.; ZENG, G.; WANG, J. Semi-supervised semantic segmentation with cross pseudo supervision. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2021. p. 2613–2622. Disponível em:

<https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00264>. Cited 3 times on pages 8, 11 e 17.

CUBUK, E. D.; ZOPH, B.; SHLENS, J.; LE, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, 2020. p. 3008–3017. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPRW50498.2020.00359>. Cited on page 20.

CURRY, A. C.; ABERCROMBIE, G.; RIESER, V. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 7388–7403. Disponível em: <https://aclanthology.org/2021.emnlp-main.587>. Cited on page 22.

DAVIDSON, T.; WARMSLEY, D.; MACY, M.; WEBER, I. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 11, n. 1, p. 512–515, May 2017. Disponível em: <https://ojs.aaai.org/index.php/ICWSM/article/view/14955>. Cited on page 19.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*. Los Alamitos, CA, USA: IEEE Computer Society, 2009. p. 248–255. ISSN 1063-6919. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2009.5206848>. Cited on page 20.

EDUNOV, S.; OTT, M.; AULI, M.; GRANGIER, D. Understanding back-translation at scale. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 489–500. Disponível em: <https://aclanthology.org/D18-1045>. Cited on page 20.

FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 51, n. 4, jul 2018. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3232676>. Cited on page 17.

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, v. 1, n. 12, p. 2009, 2009. Cited on page 19.

He, J.; Gu, J.; Shen, J.; Ranzato, M. Revisiting Self-Training for Neural Sequence Generation. *arXiv e-prints*, p. arXiv:1909.13788, set. 2019. Cited 5 times on pages 9, 12, 17, 20 e 26.

IBRAHIM, M.; TORKI, M.; EL-MAKKY, N. Imbalanced toxic comments classification using data augmentation and deep learning. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.: s.n.], 2018. p. 875–878. Cited on page 19.

KARAMANOLAKIS, G.; MUKHERJEE, S. S.; ZHENG, G.; AWADALLAH, A. H. Self-training with weak supervision. In: *NAACL 2021*. NAACL 2021, 2021. Disponível em: <https://www.microsoft.com/en-us/research/publication/self-training-weak-supervision-astra/>. Cited 3 times on pages 8, 11 e 17.

LAINE, S.; AILA, T. Temporal ensembling for semi-supervised learning. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. Disponível em: <https://openreview.net/forum?id=BJ6oOfqge>. Cited 2 times on pages 12 e 17.

LEITE, J.; SCARTON, C.; SILVA, D. Noisy self-training with data augmentations for offensive and hate speech detection tasks. In: MITKOV, R.; ANGELOVA, G. (Ed.). *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, 2023. p. 631–640. Disponível em: <https://aclanthology.org/2023.ranlp-1.68>. Cited on page 16.

LEONARDELLI, E.; MENINI, S.; TONELLI, S. Dh-fbk@ haspeede2: Italian hate speech detection via self-training and oversampling. In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. [S.l.: s.n.], 2020. v. 2765. Cited on page 20.

LIU, S.; MALLOL-RAGOLTA, A.; PARADA-CABALEIRO, E.; QIAN, K.; JING, X.; KATHAN, A.; HU, B.; SCHULLER, B. W. Audio self-supervised learning: A survey. *Patterns*, v. 3, n. 12, p. 100616, 2022. ISSN 2666-3899. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666389922002410>. Cited 2 times on pages 9 e 17.

LONGPRE, S.; WANG, Y.; DUBOIS, C. How effective is task-agnostic data augmentation for pretrained transformers? In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020. p. 4401–4411. Disponível em: <https://aclanthology.org/2020.findings-emnlp.394>. Cited on page 18.

MARIVATE, V.; SEFARA, T. Improving short text classification through global augmentation methods. In: SPRINGER. *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*. [S.l.], 2020. p. 385–399. Cited 2 times on pages 19 e 26.

MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, ACM New York, NY, USA, v. 38, n. 11, p. 39–41, 1995. Cited on page 23.

MIYATO, T.; MAEDA, S.-i.; KOYAMA, M.; ISHII, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 41, n. 8, p. 1979–1993, 2018. Cited 2 times on pages 12 e 17.

MONDAL, M.; SILVA, L. A.; BENEVENUTO, F. A measurement study of hate speech in social media. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. New York, NY, USA: Association for Computing Machinery, 2017. (HT '17), p. 85–94. ISBN 9781450347082. Disponível em: <https://doi.org/10.1145/3078714.3078723>. Cited on page 8.

MONDAL, M.; SILVA, L. A.; BENEVENUTO, F. A measurement study of hate speech in social media. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. New York, NY, USA: Association for Computing Machinery, 2017. (HT '17), p. 85–94. ISBN 9781450347082. Disponível em: <https://doi.org/10.1145/3078714.3078723>. Cited on page 17.

MOSOLOVA, A.; FOMIN, V.; BONDARENKO, I. Text augmentation for neural networks. *AIST (Supplement)*, v. 2268, p. 104–109, 2018. Cited on page 19.

MUMUNI, A.; MUMUNI, F. Data augmentation: A comprehensive survey of modern approaches. *Array*, v. 16, p. 100258, 2022. ISSN 2590-0056. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2590005622000911>. Cited on page 14.

NAJAFI, A.; MAEDA, S.-i.; KOYAMA, M.; MIYATO, T. Robustness to adversarial perturbations in learning from incomplete data. In: WALLACH, H.; LAROCHELLE, H.; BEYGELZIMER, A.; ALCHé-BUC, F. d'; FOX, E.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. v. 32. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2019/file/60ad83801910ec976590f69f638e0d6d-Paper.pdf>. Cited 2 times on pages 12 e 18.

NG, N.; YEE, K.; BAEVSKI, A.; OTT, M.; AULI, M.; EDUNOV, S. Facebook FAIR's WMT19 news translation task submission. In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, 2019. p. 314–319. Disponível em: <https://aclanthology.org/W19-5333>. Cited on page 23.

OUYANG, L.; WU, J.; JIANG, X.; ALMEIDA, D.; WAINWRIGHT, C.; MISHKIN, P.; ZHANG, C.; AGARWAL, S.; SLAMA, K.; RAY, A. et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, v. 35, p. 27730–27744, 2022. Cited 2 times on pages 28 e 30.

PHAM-HONG, B.-T.; CHOKSHI, S. PGSG at SemEval-2020 task 12: BERT-LSTM with tweets' pretrained model and noisy student training method. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, 2020. p. 2111–2116. Disponível em: <https://aclanthology.org/2020.semeval-1.280>. Cited on page 20.

RASMUS, A.; VALPOLA, H.; HONKALA, M.; BERGLUND, M.; RAIKO, T. Semi-supervised learning with ladder networks. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. Cambridge, MA, USA: MIT Press, 2015. (NIPS'15), p. 3546–3554. Cited 2 times on pages 12 e 17.

RICHARDSON, C.; SHAH, S.; YUAN, X. Semi-supervised machine learning for analyzing covid-19 related twitter data for asian hate speech. In: IEEE. *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.], 2022. p. 1643–1648. Cited on page 21.

RIZOS, G.; HEMKER, K.; SCHULLER, B. 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*. New York, NY, USA: Association for Computing Machinery, 2019. (CIKM '19), p. 991–1000. ISBN 9781450369763. Disponível em: <https://doi.org/10.1145/3357384.3358040>. Cited on page 19.

SANTOS, R. B.; MATOS, B. C.; CARVALHO, P.; BATISTA, F.; RIBEIRO, R. Semi-Supervised Annotation of Portuguese Hate Speech Across Social Media Domains. In: CORDEIRO, J. a.; PEREIRA, M. J. a.; RODRIGUES, N. F.; PAIS, S. a. (Ed.). *11th Symposium on Languages,* *Applications and Technologies (SLATE 2022)*. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022. (Open Access Series in Informatics (OASIcs), v. 104), p. 11:1–11:14. ISBN 978-3-95977-245-7. ISSN 2190-6807. Disponível em: <https://drops.dagstuhl.de/opus/volltexte/2022/16757>. Cited 2 times on pages 9 e 21.

SCHIAPPA, M. C.; RAWAT, Y. S.; SHAH, M. Self-supervised learning for videos: A survey. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, dec 2022. ISSN 0360-0300. Just Accepted. Disponível em: <https://doi.org/10.1145/3577925>. Cited 2 times on pages 9 e 17.

SCHMIDT, A.; WIEGAND, M. A survey on hate speech detection using natural language processing. In: KU, L.-W.; LI, C.-T. (Ed.). *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: Association for Computational Linguistics, 2017. p. 1–10. Disponível em: <https://aclanthology.org/W17-1101>. Cited on page 8.

WASEEM, Z. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In: *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: Association for Computational Linguistics, 2016. p. 138–142. Disponível em: <https://aclanthology.org/W16-5618>. Cited on page 19.

WASEEM, Z.; HOVY, D. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, 2016. p. 88–93. Disponível em: <https://aclanthology.org/N16-2013>. Cited on page 19.

XIE, Q.; DAI, Z.; HOVY, E.; LUONG, M.-T.; LE, Q. V. Unsupervised data augmentation for consistency training. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020. (NIPS'20). ISBN 9781713829546. Cited 5 times on pages 12, 17, 20, 22 e 24.

XIE, Q.; LUONG, M.; HOVY, E.; LE, Q. V. Self-training with noisy student improves imagenet classification. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2020. p. 10684–10695. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01070>. Cited 5 times on pages 8, 12, 17, 20 e 22.

ZAMPIERI, M.; MALMASI, S.; NAKOV, P.; ROSENTHAL, S.; FARRA, N.; KUMAR, R. Predicting the type and target of offensive posts in social media. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 1415–1420. Disponível em: <https://aclanthology.org/N19-1144>. Cited on page 22.

ZHANG, H.; CISSE, M.; DAUPHIN, Y. N.; LOPEZ-PAZ, D. mixup: Beyond empirical risk minimization. In: *International Conference on Learning Representations*. [s.n.], 2018. Disponível em: <https://openreview.net/forum?id=r1Ddp1-Rb>. Cited on page 19.

ZHANG, X.; ZHAO, J.; LECUN, Y. Character-level convolutional networks for text classification. In: CORTES, C.; LAWRENCE, N.; LEE, D.; SUGIYAMA, M.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. v. 28. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf>. Cited on page 19.