

# A Tale of Two Domains: Automatic Identification of Hate Speech in Cross-Domain Scenarios

Gustaf Gren

Department of Linguistics  
Bachelor's Thesis 15 ECTS credits  
Linguistics – Bachelor's Course, LIN622  
Bachelor's Programme in Linguistics 180 ECTS credits  
Spring semester 2023  
Supervisor: Robert Östling  
Swedish title: Automatisk identifikation  
av näthat i domänöverföringsscenarion



# A Tale of Two Domains: Automatic Identification of Hate Speech in Cross-Domain Scenarios

Gustaf Gren

## Abstract

As our lives become more and more digital, our exposure to certain phenomena increases, one of which is hate speech. Thus, automatic hate speech identification is needed. This thesis explores three strategies for hate speech detection for cross-domain scenarios: using a model trained on annotated data for a previous domain, a model trained on data from a novel methodology of automatic data derivation (with cross-domain scenarios in mind), and using ChatGPT as a domain-agnostic classifier. Results showed that cross-domain scenarios remain a challenge for hate speech detection, results which are discussed out of both technical and ethical considerations.

## Keywords

NLP, hate speech detection, transformers, BERT, ChatGPT

## Sammanfattning

I takt med att våra liv blir allt mer digitala ökar vår exponering för vissa fenomen, varav ett är näthat. Därför behövs automatisk identifikation av näthat. Denna uppsats utforskar tre strategier för att upptäcka hatretorik för korsdomänscenarion: att använda inferenserna av en modell tränad på annoterad data för en tidigare domän, att använda inferenserna av en modell tränad på data från en ny metodologi för automatisk dataderivatisering som föreslås (för denna avhandling), samt att använda ChatGPT som klassifierare. Resultaten visade att korsdomänscenarion fortfarande utgör en utmaning för upptäckt av näthat, resultat som diskuteras utifrån både tekniska och etiska överväganden.

## Nyckelord

Språkteknologi, näthat, hatretorik, transformers, BERT, ChatGPT

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background</b>	<b>2</b>
2.1	The problem of definition . . . . .	2
2.1.1	Definition . . . . .	2
2.1.2	Hate speech as speech acts . . . . .	3
2.2	The societal effects of hate speech . . . . .	4
2.3	Machine learning to the rescue . . . . .	4
2.3.1	Transformers . . . . .	4
2.3.2	BERT . . . . .	5
2.3.3	ChatGPT and prompt engineering . . . . .	5
2.4	Related research . . . . .	6
2.4.1	Deep learning approaches . . . . .	6
2.4.2	Traditional machine learning approaches . . . . .	6
2.5	Gaps . . . . .	7
<b>3</b>	<b>Data &amp; Proposed Methodology</b>	<b>8</b>
3.1	Data . . . . .	8
3.1.1	Flashback . . . . .	8
3.1.2	Familjeliv . . . . .	8
3.2	Pipelines . . . . .	10
3.2.1	Pipeline for baseline model ⟨Q1⟩ . . . . .	10
3.2.2	Semi-automatic derivation of training data ⟨Q2⟩ . . . . .	11
3.2.3	Pipeline for model based on semi-automatically derived data in regards to the first domain ⟨Q2⟩ . . . . .	12
3.3	Cross-domain scenarios ⟨Q3⟩ . . . . .	12
3.3.1	ChatGPT as a domain agnostic classifier . . . . .	12
3.3.2	Summary of cross-domain scenario methodologies . . . . .	13
<b>4</b>	<b>Results</b>	<b>15</b>
4.1	Data derivation using SVM classifier ⟨Q2⟩ . . . . .	15
4.2	Model performance in relation to method of data derivation . . . . .	16
4.3	Patterns in model inference confusion . . . . .	17
4.3.1	Analysis of the discrepancies in inference . . . . .	18
<b>5</b>	<b>Discussion</b>	<b>20</b>
5.1	Research questions . . . . .	20
5.2	Data discussion . . . . .	20
5.3	Methodology discussion . . . . .	21
5.4	Discussion of results . . . . .	21
5.5	Ethical Concerns . . . . .	22
<b>6</b>	<b>Conclusion</b>	<b>24</b>
6.1	Answers to research questions . . . . .	24
6.2	Summary . . . . .	25
6.3	Future Research . . . . .	25

<b>References</b>	<b>26</b>
<b>A Annotation Guidelines</b>	<b>30</b>
<b>B Hyper Parameters</b>	<b>32</b>

## List of Figures

1	Data breakdown concerning proportion test/dev-set . . . . .	9
2	Pseudocode for the distillation of multi-category annotations to binary labels . .	11
3	Prompt for ChatGPT . . . . .	13
4	Precision/recall curve for Familjeliv SVM classifier . . . . .	15
5	MLM hyper-parameters . . . . .	32
6	BERT hyper-parameters for annotated data . . . . .	32
7	BERT hyper-parameters for automatic model . . . . .	32

## List of Tables

1	Shortened annotation guidelines . . . . .	10
2	Summary of methodologies for domain transfer . . . . .	14
3	SVM classifier results . . . . .	15
4	Flashback results . . . . .	16
5	Familjeliv results . . . . .	16
6	Three-way confusion matrix for Annotated model/ChatGPT . . . . .	17
7	Three-way confusion matrix for Automatic model/ChatGPT . . . . .	17
8	BERT prediction discrepancies . . . . .	18
9	ChatGPT prediction discrepancies . . . . .	19
10	Annotation guidelines . . . . .	30
11	Annotation guidelines (continued) . . . . .	31

# 1 Introduction

Most don't equate freedom of expression to the freedom to say any thing to anyone for any reason whatsoever. Just like we introduce laws to disallow individuals to do harm to others — physically or psychologically — we have laws and norms to regulate the extent of that freedom of expression so that it isn't used to hurt others or impend the freedom of expression of others (similar to accepting the conclusion of Popper (1945, p. 109) concerning the paradox of intolerance: any free society tolerating intolerance will eventually have its liberties destroyed by the intolerant). Sometimes this is the definition used for what is known as *hate speech*, but, despite the common sense intuition that we *know* what hate speech is, namely speech which expresses *hate* towards either specific people or groups, in trying to close in on a definition it grows ever more elusive — both in the common sense intuition of people (as evident from poor annotator agreement (Roß et al. 2016)), as well as more linguistic definitions (Brown 2017; Brown 2018). Despite the difficulty in definition, identifying hate speech is necessary if we want to avoid the real consequences exposure to hate speech can have to individuals, which results from studies like Bilewicz and Soral (2020) suggest can increase polarization in society. Because of the immense amount of data on the internet, automatic parsing and identification of unfiltered posts is needed.

Hate speech detection is a challenging but in-demand task within natural language processing (NLP). This thesis aims to investigate methodologies to alleviate challenges during cross-domain scenarios, i.e. when we have data for one domain (e.g. one discussion forum) and then transfer it to another. The reasoning for this stems from other studies in computational socio-linguistics, who found numerous lexical and semantic differences between communities on the internet (Bamman et al. 2014; Hemphill and Otterbacher 2012; Lucy and Mendelsohn 2018; Yoder 2021), which theoretically also should apply to hate speech constructions in accordance to Brown (2017). Specifically, methods for automatic extraction of training data from unannotated data (using a small amount of annotated data from another domain), as well as the feasibility of using ChatGPT for annotation of hate speech, is researched.

## 2 Background

In this section the following will be presented: the surrounding literature regarding hate speech itself, the methodologies commonly used for hate speech detection, and the underlying architecture of those methodologies themselves necessary to make sense of the proposed methodology for this thesis.

### 2.1 The problem of definition

To investigate how we can identify hate speech we must first try to define what it is. This turns out to be a greater challenge than one might think, especially when we take into consideration that any definition of hate speech must tread the line between what can be constituted as being under the protection of free speech, or what can't, very carefully (Brown 2017). After all, as Shakespeare put it so many years ago: *"All's not offence that indiscretion finds."*<sup>1</sup> One common strategy is to lean on an already established legal definition. For example, according to the EU's Framework Decision 2008/913/JHA, hate speech is punishable by means of criminal law, stated as follows:

*"Public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, colour, descent, religion or belief, or national or ethnic origin;"*

However, leaning only on legal definitions is risky if our primary goal is to understand what hate speech actually *is*, since legal definitions draw heavily from people's intuition to make final judgments, which isn't ideal considering the large problems with annotator agreement in studies based on those types of subjective judgments (e.g. is this hate speech, or not)(Roß et al. 2016). As an example of how this subjectivity issue is instantiated in less academic circles, take the infected debate climate surrounding the state of Israel. One side argues that any criticism against Israel is effectively hate speech (e.g. Wolkoff (2020)), while the other claims that the critique is merely an attempt to shut down debate (e.g. Forman (2018)). And while these examples may both be slightly tongue-in-cheek, the consequence is that the line itself between what can be considered hate speech, and what can't, paradoxically becomes impossibly strict and impossibly subjective at the same time.

This problem of definition is pervasive in the realm of automatic identification of hate speech across datasets, and available datasets often make use of different definitions, making a comparison between them difficult (MacAvaney et al. 2019).

#### 2.1.1 Definition

However, despite the difficulty in definition, a working definition must be used since this is a classification task wherein there must be a hard set boundary of what can be constituted as hate speech, and what can't. The definition of hate speech that will be formally used in this thesis is derived from Fortuna and Nunes (2018), who gathered several definitions of hate speech from different sources and identified common dimensions between them, as well as the definition used in Essen and Jansson (2020), seeing that the data used in this thesis is partly from their research (for more details on data choice see Section 3.1 on page 8). The working definition is as follows:

---

<sup>1</sup>King Lear, Act II Scene IV

1. Incites violence and/or hate against groups/individuals,
2. Attacks or diminishes groups/individuals,
3. Targets specific groups/individuals.

This working definition is chosen not only because Fortuna and Nunes’ (ibid.) definition has been used as inspiration for previous hate speech identification research (e.g. MacAvaney et al. (2019)), but primarily because it is close to the definition used in Essen and Jansson (2020) to annotate posts containing hate speech. However, worth noting is that this definition is not universally accepted, and it has its fair share of issues. Additionally, it does not explain the reasoning behind cross-domain scenarios, and why it is needed for hate speech identification on the internet. This will be further discussed in the section that follows.

### 2.1.2 Hate speech as speech acts

Trying to amend this gap of the lack of a universally accepted definition, Brown (2017) attempts to define hate speech not out of a definite set of properties, but as he puts it on page 427, instead as being the “*expressive dimensions of identity-based envy, hostility, conflict, mistrust and oppression*”. In other words, that hate speech — despite what the term’s constituents imply — needn’t necessarily stem from emotion or malicious intent but can be explained in virtue of a special kind of speech act; i.e. that hate speech can have different *functions* for the user depending on the context in which it is uttered. Brown (2017) doesn’t deny that it can be used stemming from hate, but that it also can have functions for that individual’s identity. One can imagine, for example, that a member of the Ku Klux Klan is expected to express hate speech despite their own opinion on the matter if that person wants to fit in with that particular group. In a follow-up article, Brown (2018) argues that what makes *online* hate speech special are the elements of anonymity and instantaneousness (i.e. that one doesn’t have to carefully consider what one says before posting), which are defining features of social media platforms in comparison to real life. These aspects, he argues, have led to a radical shift in how hate speech is used and encountered in modern society compared to when the term referred to what we might consider the prototypical sense of the term, in that we encounter it quite a lot more frequently than we used to, and that online hate speech can propagate in tight-knit communities where hate speech serves as one aspect of that group’s identity.

This is in line with other studies from computational sociolinguistics, where numerous studies have investigated the diversity in language use and adoption between distinct communities, for example: Bäck et al. (2018) investigated a Swedish discussion forum and found that as users get more involved in that particular community, they use the first-person singular pronoun (*I*) less and the first-person plural pronoun (*we*) more. Hemphill and Otterbacher (2012) found that female users of a movie review website tend to adapt their language use in movie reviews to their male counterparts, which they explain in virtue of men receiving more prestige in that particular community. Numerous other studies in computational sociolinguistics have additionally shown that groups with different identity on the internet differ in their language use (Yoder 2021), both in lexical choice (Bamman et al. 2014) as well as how users use the same word to convey different meanings (Lucy and Mendelsohn 2018). In other words, we expect linguistic variation between communities on the internet just like in regular linguistic communities, variation that — if Brown (2017) is correct — also applies to hate speech constructions.

## 2.2 The societal effects of hate speech

A cross-sectional study by Keipi et al. (2016) found that 40/44% (female/male) of subjects had reported being exposed to hate speech online. This is not ideal, since as Bilewicz and Soral (2020) argues, exposure to hate speech leads to political radicalization and it deteriorates intergroup relations. This is supported by research like He et al. (2021) who found a correlation between exposure to hate speech and anti-Asian hate on Twitter after COVID-19. Psychologically, Bilewicz and Soral (2020) explain this as being in virtue of empathy being replaced by a ‘us vs. them’-response since the increasing presence of hate speech creates a sense of the existence of a descriptive norm which allows (and encourages) outgroup scorn. In other words, the increase in hate speech is a psychological process powered by our desire to adhere to social norms. These effects can be observed in real-life observations as well, Winiewski et al. (2017) for example, saw that Polish youths who are more exposed to hate speech not only tend to avoid being around ethnic minorities in their physical environment but also show greater support for harsher treatment of immigrants.

## 2.3 Machine learning to the rescue

Many areas concerning the analysis of linguistic behavior on the internet must use machine learning simply due to the enormous amount of data that needs to be processed. Sifting through that data manually simply isn’t feasible, no matter the manpower. This often includes some classification task, i.e. a task wherein we are interested in an algorithm that can ‘classify’ pieces of text according to certain labels. One such example is spam detection, where we are interested in labeling emails as either spam or not-spam. Originally, this kind of task was approached using a simple algorithm, where our features were — for example — based on certain words. This would mean that a spam detector would only be able to classify something as spam if it contained a certain amount of words that are predictive as spam according to that algorithm. However, text classification tasks are often much more difficult than that; hate speech, case and point. The task of hate speech identification is one which not only lexical choice but also a syntactic and — perhaps most importantly — a semantic dimension play differently important roles. To capture these multifaceted dimensions these algorithms and models need to not only possess a large number of dimensions themselves, but they also need *a lot* of data to fill those dimensions. Many of these difficult aspects require context-sensitive systems, something which new machine-learning models have attempted to incorporate.

### 2.3.1 Transformers

The transformer model is a deep-learning model introduced by Vaswani et al. (2017), and it is built on the idea of *self-attention*, which gives certain parts of the input more importance while diminishing other parts. Specifically, transformers were designed to solve specific problems for machine translation (where context is essential for many languages to assign meaning and correct grammar for a sentence) on longer documents. Consider translating the two sentences [taken from the original blog post by Uszkoreit (2017)] “*The animal didn’t cross the street because it was tired*” and “*The animal didn’t cross the street because it was too wide*”. The translation of the anaphoric pronoun ‘it’ depends on whether we mean ‘the street’ or ‘the animal’ from English to French, since French has gendered nouns. The introduction of self-attention meant that in contrast to previous approaches, transformers could make these types of inferences using a constant amount of calculations instead of having to take the entire input in as a sequential list (which makes dealing with longer documents exceptionally demanding

in regards to computation), and since it processes all tokens of the input at the same time also makes parallelization easier, which decreases training time. The original transformer is made up of two neural networks: the *encoder* and the *decoder*. The encoder is a type of neural network which takes data as input and produces a latent representation as its output, where a latent representation can be thought of as a compressed representation of the most important features of the data. The decoder then takes that compressed output and generates an output sequence. Some models make explicit use of these components and only use one or the other. GPT, which stands for **Generative Pretrained Transformer**, introduced by Radford et al. (2018) only uses the decoder to generate text by taking certain tokens as input and then using those guesses the next token, which are then appended to the input and used as output for the next token prediction. This architecture was then augmented in a supervised manner to better act as dialogue agents in ChatGPT, released in late 2022 (OpenAI 2022).

### 2.3.2 BERT

BERT, which stands for **Bidirectional Encoder Representations from Transformers**, is an encoder-only transformer model used for various language inference tasks, introduced by Devlin et al. (2018). BERT was introduced to mitigate the limitation they observed in the way that the then-current transformers represented text since they did so unidirectionally. BERT can instead utilize both directions for its representations. For BERT there are two specific steps in which training occurs: *pre-training* and *fine-tuning*. For pre-training the two main components of BERT are trained, namely i.) the *masked language model* (MLM) and ii.) the *next sentence predictor* (NSP), using two tasks: i.) mask some percentage<sup>2</sup> of the input tokens at random and then let the model predict those tokens using all the tokens in that particular input, and ii.) give the model pairs of sentences, half of which are true pairs and half random, and let it predict whether they truly follow or not. These pre-trained representations are task agnostic, which reduces the need for specifically engineered task-specific architectures. Instead, task-specific engineering is performed with the second step, the fine-tuning stage, in which we use a smaller amount of data to train the model for some specific downstream task (e.g. hate speech detection). This is what is also called *transfer learning*, where we take the knowledge already present in a model and leverage that to some more specific downstream task, i.e. we *transfer* it.

There is an additional distinction here regarding text classification tasks during transfer learning and if the model in question has seen examples of what it is designed to accomplish: *few-shot learning*, meaning if the model has only seen a few examples of the task at hand, the goal being that the model generalizes the prior knowledge from that data to unseen data; and *zero-shot learning* where the model hasn't seen *any* labeled examples of the downstream task, and simply infers the label from previously trained concepts.

### 2.3.3 ChatGPT and prompt engineering

ChatGPT is a conversational model based on the GPT architecture. Users write what is called a *prompt*, a set of instructions which the model creates a latent representation of and then outputs the most probable response to that according to the GPT model. This also means that *prompt engineering* becomes an important part of the process of interacting with these conversational language models (White et al. 2023), where prompt engineering is how we best align our prompts to more reliably produce a certain kind of output.

---

<sup>2</sup> $p = 0.15$  in Devlin et al. (2018)

The extent of what downstream tasks these models can robustly solve, however, is still unexplored. Gilardi et al. (2023) looked into using ChatGPT as an annotator, where ChatGPT outperformed crowd-workers for several annotation tasks, including relevance, stance, topics, and frames detection.

## 2.4 Related research

Automatic hate speech detection has been broadly studied for many languages, although most efforts have been focused on English since the resources for other languages are scarce (Jahan and Oussalah 2021). For Swedish, there are a few studies: Fernquist et al. (2019) looked into the feasibility of automatic hate speech detection on three Swedish discussion forums using ULM-FIT (Howard and Ruder 2018), where they managed to achieve an  $F_1$  score of  $\sim 0.8$ ; Essen and Jansson (2020) instead focuses on investigating how hate speech changed after a group of journalists managed to get a hold of the real identity of some prominent users on the discussion forum ‘Flashback’ in 2014, where they find that hate speech against immigrants decreased while hate against women increased. This was done by training a logistical regression machine-learning model in a supervised manner, which then allowed them to quantify hate speech trends.

Generally speaking, the following methodologies are used for the task of automatic hate speech identification:

### 2.4.1 Deep learning approaches

Deep learning approaches are arguably the most common method currently used for automatic hate speech identification tasks (Jahan and Oussalah 2021), especially models based on the transformer architecture like BERT. For example, Dowlagar and Mamidi (2021) used the conventional pipeline of fine-tuning a pre-trained BERT with a smaller labeled dataset for automatic hate speech detection in English, which outperformed their other baselines. Sai and Sharma (2020) found the same results for multilingual hate speech detection for English, German, and Hindi respectively. BERT outperformed their other baselines which were constituted by SVMs (with various methods of representing text). SVM, or *support vector machine*, is a supervised algorithm used for classification and regression analysis. The algorithm is, simply put, based on the idea of finding the best possible boundary (i.e. hyperplane) that separates data into two classes. Sometimes, an additional step of pretraining for the masked language model is used (which adapts the model to the domain in question by altering the latent representation produced by e.g. BERT) since it has been shown to improve performance for other downstream tasks (Gururangan et al. 2020), which Pham et al. (2020) for example used for automatic hate speech detection in Vietnamese.

### 2.4.2 Traditional machine learning approaches

While deep-learning models have shown state-of-the-art performance on hate speech identification tasks, simpler machine learning methods are also used. MacAvaney et al. (2019) used multi-view SVMs to achieve near state-of-the-art performance, while simultaneously avoiding the issue models like BERT face in regards to features that are difficult to interpret. Aluru et al. (2020) used LASER embeddings (Schwenk and Douze 2017) in concordance with logistic regression, which outperformed their BERT implementation for low-resource languages. Logistic regression is a statistical method used to model the probability of a binary outcome based on predictor values, and LASER embeddings another type of sentence embedding created specifically

for use in code-switching scenarios. These results suggest that traditional machine learning approaches might still be useful in hate speech detection, seeing that they are much faster to train, and in some cases, still comparable in performance.

## 2.5 Gaps

Most previous research has been focused on optimizing automatic hate speech identification for metrics set by the shared tasks where datasets already exist (Jahan and Oussalah 2021). As far as I can tell, no work has been done regarding developing system pipelines for scenarios of cross-domain transfer (i.e. if we develop the system for one discussion forum, and then transfer it to another), and the feasibility of doing so. Since it is unreasonable to think that owners of these discussion forums will have the resources available to hire annotators to train a supervised model for hate speech detection [not to mention the technical knowledge required], I will look into the feasibility of doing this in a manner which could enable cross-domain transfer (as well as updating that model when linguistic expressions change). If hate speech constructions differ according to context, as e.g. the theory of Brown (2017) predicts, then cross-domain transfer (rather than relying on inferences based on another domain) is needed for different communities on the internet. This requires the automatic identification of hate speech since the amount of data is too large for a human to sift through. What would such a system look like, and what is the feasibility of it?

### Research Questions

- ⟨Q1⟩ What is the baseline performance of hate speech detection in Swedish using BERT for one domain trained using manually annotated data?
- ⟨Q2⟩ How does the performance of hate speech detection on a single domain using BERT trained on semi-automatically derived training data (derived from an SVM-classifier trained on a small amount of annotated data) compare to the baseline set in ⟨Q1⟩?
- ⟨Q3⟩ Does a model trained on semi-automatically derived data (derived from the inferences of the model trained on the previous domain) improve performance for hate speech identification in Swedish in cross-domain scenarios compared to using baseline method in ⟨Q1⟩ which has just seen unannotated data from the new domain? And, how does the performance of both of those methodologies compare to the domain-agnostic methodology of using ChatGPT to perform zero-shot classification of hate speech?

## 3 Data & Proposed Methodology

In this section, I will first discuss data choice, then move on to the annotation process, and then I will finally go through the proposed methodology. The code written for the various methodologies is available at <https://github.com/skogsgren/a-tale-of-two-domains>.

### 3.1 Data

The data used in this thesis is taken from two Swedish discussion forums: Flashback<sup>3</sup> and Familjeliv<sup>4</sup>. These two domains were chosen because they simultaneously show a necessary similarity in containing subforums about the same topics, as well as an important difference in the fact that they differ in *group identity*. Group identity here is a term intentionally used vaguely to encompass the intuition of the sense that the forums differ in what is considered acceptable linguistic behavior for their respective group. Studies from sociolinguistics have shown for a long time that linguistic variation correlates to certain demographic features, like socioeconomic status (Labov 2006), age and gender (Koch et al. 2022) or political beliefs (Hall-Lew et al. 2010), among others. From the basis of this, the two discussion forums ought to have linguistic variation, as one is focused on the freedom of expression (which attracts individuals ascribing to certain political ideologies), while the other is focused on discussions about parenting, which interests different demographics.

#### 3.1.1 Flashback

Flashback is the largest Swedish discussion forum, with more than a million registered users (Essen and Jansson 2020). It consists of numerous subforums, covering topics such as drugs, sex, immigration, politics, and many others. As the topics suggest, their motto is “true freedom of expression” (Essen and Jansson 2020). Within each subforum, there are threads created by users, within which lie posts written by users.

Data is obtained from Essen and Jansson (2020), data which consists of posts from three subforums: *immigration*, *politics*, and *feminism*. In total, there are ~3.9 million unannotated posts and 4040 annotated posts (labeled either as hate speech or not-hate speech). The annotated posts are first split into an 80/20 split to form an intermediate dataset which is then used to derive the final test set and the dev set. This is visualized in Figure 1 on the next page. Proportion was kept in mind for each of these splits, and each subset maintained the same proportion of HS/non-HS as the full dataset, to maintain consistency and avoid sources of error. One could, for example, consider the scenario where - by chance - just randomly choosing the testset skews the distribution, either so that it consists of *more* or *less* hate speech than the original dataset, which would give an inaccurate metric of how performant the particular model is. The unannotated data is used to train the masked language model of BERT for  $\langle Q1 \rangle$  and  $\langle Q2 \rangle$ .

#### 3.1.2 Familjeliv

Familjeliv is a Swedish discussion forum focusing on discussions about various topics relating to parenting. A study of Swedish internet users from 2021 found that 1% use the website at least on a weekly basis (Internetstiftelsen 2021). Similarly to Flashback, there are numerous subforums for a range of various topics (although not as diverse as Flashback), from which

---

<sup>3</sup><https://flashback.org>

<sup>4</sup><https://familjeliv.se>

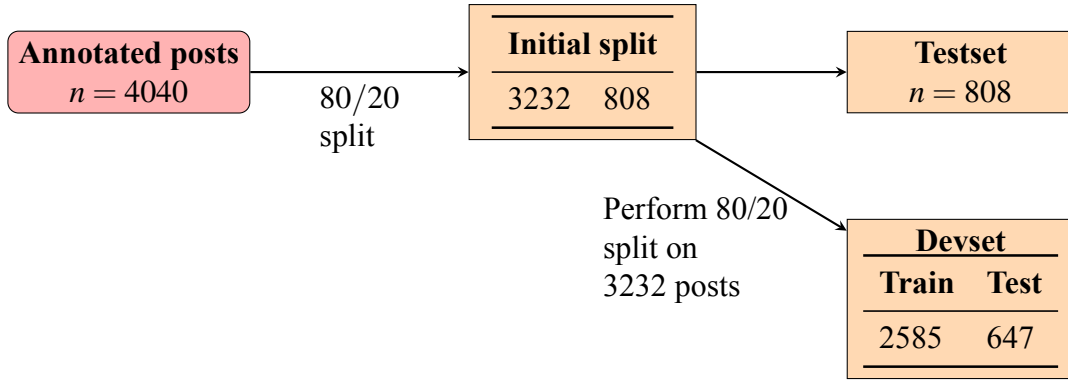


Figure 1: Data breakdown concerning proportion test/dev-set

the forums *gender* and *politics* were chosen. Intuitively the reason for this is that hate speech is much less likely to occur in contexts where one — for example — discusses the various methods of breastfeeding. This was manually pruned, because there are several subforums on Familjeliv which are somewhat similar in topics to the Flashback equivalent (one example being *society* on Familjeliv vs. *politics* on Flashback), but which on manual inspection were too different to include. The structure of threads and posts is practically identical to Flashback in that users start threads and then other users respond.

Since there are currently no publicly available datasets, data is gathered employing crawling, which is done using a Python script written using BeautifulSoup<sup>5</sup>. In total  $\sim 760,000$  posts are crawled. From those 988 posts are randomly extracted using the same technique used in Essen and Jansson (2020): a thread is randomly selected, then a number of posts from the beginning of the thread are chosen, then a number from the end. In my case, I chose 10 from the beginning and 3 from the end. No filtering is done regarding the quality of these posts (i.e. the language isn’t identified, token count, etc), instead being left to the annotation stage. This process is repeated until 988 posts are extracted (i.e. 76 threads). The number 988 is arbitrarily chosen, and was chosen because it: i.) adheres to the methodology just described by being a multiple of the number of having posts from the beginning and end of the post, ii.) is comparable both to the testset for Flashback (808 posts), and iii.) is reasonable for a single person to annotate given the timeframe.

This data is then manually annotated by me using a custom-built annotation UI<sup>6</sup>. For the annotation process, the guidelines are based on the guidelines in Essen and Jansson (2020) in an attempt to maintain similarity to their annotation process. A simplified version is available below in Table 1 on the following page, with the full version available in Section A on page 30. During annotation posts that are deemed inapplicable (e.g. only contain a URL, emojis, or HTML code) are discarded. This lessened the testset to 978 in total. To extract a binary classification from that multi-feature annotation process a simple list of conditionals is used based on the presence of certain annotated features, and is presented in pseudocode in Figure 2 on page 11. What features were utilized to form the conditionals was based on the definition laid out in Section 2 on page 2 in that a specific group had to be picked out, and some kind of threat/aggression. Certain features were not used due to the combinations not being present (hence the exclusion of e.g. female preference in combination with aggressiveness, while male preference in combination with aggressiveness is included). Wholly ironic/sarcastic posts were also excluded to potentially

<sup>5</sup><https://www.crummy.com/software/BeautifulSoup/>

<sup>6</sup>Code for both the crawler and annotation UI is available at <https://github.com/skogsgren/a-tale-of-two-domains>

avoid sources of error. One thing worth mentioning is that while the `us_vs_them == 1 AND aggressiveness > 1` *could* theoretically apply to, for example, somebody angry at soccer teams, because of the choice in subforums, this was not an issue for this dataset (since it focused on political discussion).

<b>Party politics sentiment</b>	If the post expressed sentiment (positive/neutral/negative) towards specific parties or coalitions
<b>Aggressiveness</b>	If the post expressed aggressiveness in tone
<b>Hate</b>	If the post expressed explicit hate towards specific groups or individuals
<b>Threat</b>	If the post expressed explicit threats towards specific groups or individuals
<b>Gender preference</b>	If the post expressed gender superiority for each gender
<b>Us vs. Them</b>	If the post expressed explicit us vs. them sentiment
<b>Gender/ethnicity disadvantage</b>	If the post expressed whether a specific gender/ethnicity is disadvantaged in society.
<b>Attitude towards foreigners/religion</b>	If the post either expressed positive/neutral/negative sentiment toward foreigners/religion.
<b>Irony/Sarcasm</b>	If the post expresses irony/sarcasm.

Table 1: Shortened annotation guidelines, properties concatenated for the sake of brevity. For full guidelines see Section A on page 30.

## 3.2 Pipelines

The general methodology for the pipelines answering  $\langle Q1 \rangle$  and  $\langle Q2 \rangle$  is very similar and differs only in the derivation of the data that the model is trained on, as well as the additional steps in cross-domain scenarios. Therefore the pipeline for the baseline model (based on annotated data) will be presented first, followed by the way that the automatic model differs. For both pipelines, the Swedish BERT model trained by The National Library of Sweden (Malmsten et al. 2020) is used as a base model <sup>7</sup>. All the hyperparameters used are available in Section B on page 32.

### 3.2.1 Pipeline for baseline model $\langle Q1 \rangle$

First, we adapt our base model to our specific domain from the recommendation of Gururangan et al. (2020), which in short suggests that we train one specific component of BERT’s two,

<sup>7</sup>Available here: <https://huggingface.co/KB/bert-base-swedish-cased>

```

IF (sarcasm_irony != 2)
{
    IF (
        hate == 1,
        OR threat == 1,
        OR us_vs_them == 1 AND aggressiveness > 1,
        OR foreigner_attitude == 2 AND aggressiveness > 1,
        OR religion_attitude == 2 AND aggressiveness > 1,
        OR male_preference == 1 AND aggressiveness > 1
    )
    {
        set label to hate speech
    }
}

```

Figure 2: Conditionals presented in pseudo code for how annotated posts were reduced to binary labels. A value of 2 for `foreigner_attitude` and `religion_attitude` respectively indicate a *negative* attitude. A value of 2 for `sarcasm_irony` indicates that the post is wholly sarcasm, so in this case I’m asking *as long as it is not wholly sarcasm/irony*. For full details see guidelines in Section A on page 30.

namely the *masked language model*. The masked language model layer predicts masked words. Finetuning the masked language model adapts the model to a domain by making certain domain-specific constructions *more likely* to be predicted by that masked language model. This is performed by semi-supervised means, wherein certain words are masked based on a probability for the text from the target domain, and then the model is asked to predict that word, adjusting its weights depending on those predictions. After that the second component of BERT is added, namely the *text classification head*. This is a layer that is trained for some specific downstream task, in this case with the task of classifying the labels (hate speech / not hate speech) of the annotated data as provided from Essen and Jansson (2020). By suggestion of Srivastava et al. (2014), dropout was increased, in this case from the default 0.1 to 0.2, to curb overfitting on the relatively small amount of training data. Dropout is when a given percentage of sub-layers are randomly set to zero to encourage more robust generalization.

### 3.2.2 Semi-automatic derivation of training data (Q2)

Deriving training data from unannotated data in an automatic fashion should optimally be performed through a system that is scalable and transferable. The best solution would be something that is semi-supervised since this wouldn’t require any annotated data at any part of the process. Zhu et al. (2021) had success with this in regards to euphemism detection, where they used the contrast in *log*-probability between a masked language model trained on domain-specific data and generalized data to find euphemisms using only seed-terms as the user-specified data; however, due to the complex nature of hate speech constructions (i.e. them not necessarily simply being based on lexical co-occurrence or a set amount of seed terms), a more supervised approach is deemed to be required.

The system proposed is based on research from MacAvaney et al. (2019), in that an SVM-classifier is trained on a small amount of domain-specific annotated data (in this case the dev-set

in Figure 1 on page 9) using all the tokens in that amount of data as features from unigram to 6-gram combinations. This results in a list of features where the weights indicate the predictive value of each token(s) for identifying hate speech in a given piece of text. This is a novel methodology introduced in this thesis, specifically meant to alleviate challenges in cross-domain scenarios.

### 3.2.3 Pipeline for model based on semi-automatically derived data in regards to the first domain [Q2](#)

This pipeline is very similar to the baseline pipeline. First, we finetune the masked language model beneath the base model using domain-specific data to adapt the model to the domain. Then we train an SVM classifier as detailed in Section 3.2.2 on the preceding page which is used to classify a much larger amount of unannotated data (25,000 vs 2500) than for our baseline. This semi-automatically derived dataset is then used as training data for BERT in the subsequent step. The idea is that BERT will overcome the noise present in the data by inferring the underlying linguistic (semantic and syntactic) properties of hate speech, and thus achieve comparable performance to the baseline of using annotated data.

## 3.3 Cross-domain scenarios [Q3](#)

For cross-domain scenarios, the masked language model will be additionally finetuned to data from the new domain for both methodologies using unannotated data from the new domain. For the baseline, the text classification head will remain the same as the one trained using the process described in Section 3.2.1 on page 10 (meaning it will see no annotated data from the new domain). If we have no other annotated data we can only hope that the two domains are similar enough for the already-trained model to infer the differences. The method described in Section 3.2.3 gives us a different possible path:

- (a.) Take the [Q2](#) model trained on the first domain and predict the labels for posts for the new domain.
- (b.) Using those inferences train a new SVM-classifier for the new domain, which hypothetically should give us a new list of features for tokens most predictive of hate speech in that new domain.
- (c.) Train the masked LM for the base model using unannotated data from new domain.
- (d.) Use the SVM from [\(b.\)](#) to automatically label data from the new domain.
- (e.) Train the model using that data in the same manner as described in Section 3.2.3.

Reasoning being that this results in a model which gives better performance by virtue of being more adapted to the new domain, and if performance is good, is a pipeline that could be easily adapted to new domains without the need for expensive annotation efforts.

### 3.3.1 ChatGPT as a domain agnostic classifier

The two other previously discussed methodologies are complemented using an additional classifier inspired by the results of Gilardi et al. (2023), who found ChatGPT as a promising alternative for manual annotation. The model GPT-3.5 Turbo from OpenAI (the same used for ChatGPT <sup>8</sup>)

---

<sup>8</sup>The base model is the same, however, due to the proprietary nature of ChatGPT it is impossible to know whether or not the version of ChatGPT currently running on their website is different than the model available via API; i.e. if the website version has been additionally augmented using additional data.

is prompted to classify the same posts from that new domain, but without any prior training (effectively zero-shot learning). This methodology is domain agnostic since it has not seen *any* annotated data from *any* of the two domains. The full prompt used is available in Figure 3. Important to note here is that ChatGPT was primarily trained on English, but the posts provided were in Swedish, straight from the crawled data. This is important because ChatGPT has seen very little Swedish data, not to mention Swedish data concerning definitions of hate speech, or examples of hate speech. This is also the reason why the prompt was written in English instead of Swedish, in the hopes that the model better understands its task rather than focusing on “translating” the prompt from Swedish to a task description.

This is a novel contribution of this thesis since ChatGPT has — to my knowledge — never been used before for hate speech detection, in English or Swedish.

```
You are an annotator in a linguistic study, where your task is to
identify whether the post provided is hate speech or not. You will
only respond with the corresponding label, 1 for hate speech, and 0
if it is not hate speech.
POST: {DOCUMENT}
```

Figure 3: Full prompt used for querying ChatGPT for hate speech annotation, where {DOCUMENT} is replaced by the post in question.

### 3.3.2 Summary of cross-domain scenario methodologies

A summary of the various methodologies is presented henceforth in Table 2 on the next page.

<b>Annotated</b>	<b>Automatic</b>	<b>ChatGPT</b>
1.) Train MLM of model trained on previous domain to unannotated data from the new domain .	Use the inferences from previous domain model to train an SVM classifier.	Using the OpenAI API, predict the testset using GPT3.5 Turbo.
2.) Predict testset for new domain using that model.	Derive larger training dataset using that SVM classifier.	
3.)	Train MLM of model trained on previous domain to unannotated data from new domain.	
4.)	Train model (with fine-tuned MLM) on data derived from SVM classifier.	
5.)	Predict testset for new domain using that model.	

Table 2: Summary of the various proposed methodologies for cross-domain scenarios.

## 4 Results

*Precision* measures how many of the positive predictions were correct, while *recall* measures how many of the positive labels the model managed to predict (i.e. how many of the instances of hate speech it managed to detect).  $F_1$ -score is the combined metric of the two, or more precisely, the harmonic mean of precision and recall, providing a single metric that balances both metrics.

### 4.1 Data derivation using SVM classifier [Q2](#)

Domain	Confusion Matrix		$F_1$	Precision	Recall
<i>Flashback</i>	525	94	0.47	0.48	0.46
	102	87			
<i>Familjeliv</i> <sup>9</sup>	643	288	0.17	0.09	0.65
	16	31			

Table 3: Results of the SVM classifier for the two domains. Confusion matrices read thusly (left-right, top-bottom): true negatives, false positives, false negatives, true positives.

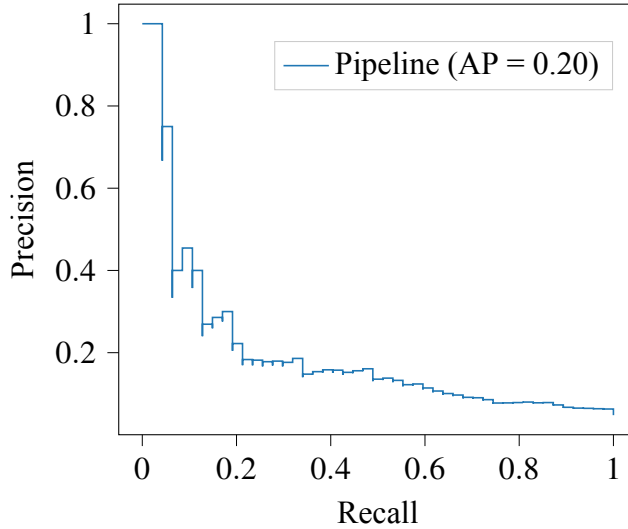


Figure 4: Precision/recall curve for Familjeliv SVM classifier; maximum  $F_1$  score of 0.24 achieved with precision 0.49 and recall 0.16

The performance of the SVM-classifier scored higher on the initial domain with an  $F_1$ -score of 0.47, and much lower during domain transfer with an  $F_1$ -score of 0.17. Adjusting the probability thresholds for the Familjeliv SVM-classifier shows an increase in  $F_1$ -score to 0.24, but this comes at the tradeoff of a much lower recall value (0.16 in comparison to 0.65).

<sup>9</sup>N.B. the classifier for *Familjeliv* didn't see the testset until all experimentation was complete, here presented simply as a metric to gauge performance of the domain-transfer process.

## 4.2 Model performance in relation to method of data derivation

Model	Confusion Matrix	$F_1$	Precision	Recall
Flashback Baseline <a href="#">⟨Q1⟩</a>	$\begin{matrix} 589 & 30 \\ 75 & 114 \end{matrix}$	0.68	0.79	0.60
Flashback Automatic <a href="#">⟨Q2⟩</a>	$\begin{matrix} 507 & 112 \\ 76 & 113 \end{matrix}$	0.55	0.50	0.60

Table 4: Results of the different models for the domain *Flashback*. Confusion matrices read thusly (left-right, top-bottom): true negatives, false positives, false negatives, true positives. Training took 10 minutes for the baseline model trained on annotated data, and 40 minutes for the automatic model on an A4000.

For the initial domain, as shown in Table 4, the model trained on annotated data had better performance due to the less frequent false positives. The recall for both methodologies was nearly identical (the automatic model made one additional mistake).

Model	Confusion Matrix	$F_1$	Precision	Recall
Familjeliv Baseline <a href="#">⟨Q1⟩</a>	$\begin{matrix} 883 & 48 \\ 29 & 18 \end{matrix}$	0.32	0.27	0.38
Familjeliv Automatic <a href="#">⟨Q3⟩</a>	$\begin{matrix} 806 & 125 \\ 20 & 27 \end{matrix}$	0.27	0.18	0.57
Familjeliv ChatGPT <a href="#">⟨Q3⟩</a>	$\begin{matrix} 866 & 65 \\ 27 & 20 \end{matrix}$	0.30	0.24	0.43

Table 5: Results of the different models for the domain *Familjeliv* [⟨Q3⟩](#). Confusion matrices read thusly (left-right, top-bottom): true negatives, false positives, false negatives, true positives. Training took 40 minutes for the automatic model on an A4000.

During cross-domain scenarios, presented above in Table 5, the automatic model also fared worse in  $F_1$ -score than both the baseline model, as well as that of ChatGPT, which received a similar  $F_1$ -score of 0.32 and 0.30 respectively. The automatic model landing a little lower, at 0.27. Again, similar to the results of the initial domain, the precision for the automatic model is worse than for other methodologies.

### 4.3 Patterns in model inference confusion

The three-way confusion matrices below track the difference in inference between the two different methodologies and the respective classification performed by ChatGPT on the testset for the cross-domain scenario. Because the most interesting differences are where they *disagreed*, they are colored differently below; the blue column denominating where the methodology predicted one (while ChatGPT predicted zero), the yellow column denominating where ChatGPT predicted one (while the methodology predicted zero).

		PREDICTED LABEL			
		annt=0	annt=1	annt=0	annt=1
		cgpt=0	cgpt=0	cgpt=1	cgpt=1
ACTUAL LABEL	0	840	<b>26</b>	<b>43</b>	22
	1	18	<b>9</b>	<i>11</i>	9

Table 6: Three-way confusion matrix tracking differences in inference between the model trained on annotated data (the Baseline model in Table 5 on the previous page), in comparison to the domain-agnostic classification from ChatGPT for the testset in the cross-domain scenario. annt is shorthand for the annotated model, while cgpt is shorthand for ChatGPT. **Bold-face** font refers to the cases where the annotated model was correct, while ChatGPT was wrong, while an *italic* font refers to the cases where the opposite was true; namely, where the annotated model was wrong, and ChatGPT was correct.

		PREDICTED LABEL			
		auto=0	auto=1	auto=0	auto=1
		cgpt=0	cgpt=0	cgpt=1	cgpt=1
ACTUAL LABEL	0	772	<b>94</b>	<b>34</b>	31
	1	17	<b>10</b>	<i>3</i>	17

Table 7: Three-way confusion matrix tracking differences in inference between the model trained on semi-automatically derived data (the Automatic model in Table 5 on the previous page), in comparison to the domain-agnostic classification from ChatGPT for the testset in the cross-domain scenario. auto is shorthand for the annotated model, while cgpt is shorthand for ChatGPT. **Bold-face** font refers to the cases where the automatic model was correct, while ChatGPT was wrong, while an *italic* font refers to the cases where the opposite was true; namely, where the automatic model was wrong, and ChatGPT was correct.

### 4.3.1 Analysis of the discrepancies in inference

Manually going through the posts for each case where the models disagreed with ChatGPT, there were some observed patterns of behavior. The BERT models, both in the cases of the Annotated/Automatic model, were keener to label posts as hate speech based on the occurrence of “trigger” words, whether that be false negatives because it fails to recognize that certain words are more predictive of hate speech (e.g. if they aren’t present in the training data, like in the second post in Table 8), or false positives (e.g. just mentioning ‘holocaust’). This sensitivity to occurrences of certain tokens was especially evident for the Automatic model (see the confusion matrix in Figure 7 on the previous page), where there are — in comparison to ChatGPT — a large number of false positives. This did not hold for the annotated model (see Figure 6 on the preceding page). ChatGPT, meanwhile, seemingly used more semantic properties in predictions; meaning that ChatGPT tended to not pick up on things like paraphrasing other posters, and classified such occurrences as HS even when the poster was only quoting somebody else, or when ChatGPT missed certain occurrences of actual HS because it was simply stating a fact (which gives the appearance of neutrality).

For examples supporting these conclusions see Table 8 and Table 9 on the following page<sup>10</sup>. In the discrepancies between the Annotated and the Automatic model, there were no observable patterns seen in the discrepancies in inference except for the larger amount of false positives for the Automatic model, hence its exclusion from its own third-way confusion matrix in Section 4.3 on the previous page.

Label explanation	Paraphrased post
Post is not HS but labelled by BERT as HS	<i>I believe in the holocaust and I do not hate jews. Just stop with assuming that we’re all like that, it’s just distasteful.</i>
Post is HS but labelled by BERT as not HS	<i>The Quran is much worse than Mein Kampf, so no, Muslims aren’t as dangerous as Hitler. They’re much more dangerous.</i>

Table 8: Discrepancies showing the tendency of BERT to label based on lexical occurrences in comparison to ChatGPT, where ChatGPT classified these examples correctly.

<sup>10</sup>Due to privacy concerns I cannot release the full results. These examples were manually picked from a list, translated, and then paraphrased slightly to obfuscate any identifying details.

Label explanation	Paraphrased post
Post is not HS but labelled by ChatGPT as HS	<i>“If my wife had done that, then she would never get in the house again. A woman that acts so irrational is neither a good partner or a fit mother.” Is that what you are saying to OP?</i>
Post is HS but labelled by ChatGPT as not HS	<i>If you have an IQ of over 50 then it is impossible to be brainwashed by a feminist.</i>

Table 9: Discrepancies showing the tendency of ChatGPT to label based on semantic properties, where BERT classified these examples correctly.

## 5 Discussion

### 5.1 Research questions

**⟨Q1⟩:** What is the baseline performance of hate speech detection in Swedish using BERT for one domain trained using manually annotated data?

The performance is acceptable, albeit still not what could be considered *good*, with an  $F_1$ -score of 0.68 for the first domain (in comparison to for example Fernquist et al. (2019) who displayed an  $F_1$ -score of  $\sim 0.8$ ). This is a novel baseline for hate speech detection in Swedish using BERT.

**⟨Q2⟩:** How does the performance of hate speech detection on a single domain using BERT trained on semi-automatically derived training data (derived from an SVM-classifier trained on a small amount of annotated data) compare to the baseline set in **⟨Q1⟩**?

The methodology employed was an SVM-classifier using a small amount of annotated data, which for cross-domain scenarios can then be retrained using the inferences using the model of the previous domain. While a semi-supervised approach is preferred, it was not explored due to the complex nature of hate speech. This is discussed further in Section 5.2.

The model trained on automatically derived data achieves consistently worse performance, due to it being more sensitive to lexical trigger words (i.e. the *precision* is lower than for the initial domain, while *recall* remains about the same). This is discussed further in Section 5.3 on the next page.

**⟨Q3⟩:** Does a model trained on semi-automatically derived data (derived from the inferences of the model trained on the previous domain) improve performance for hate speech identification in Swedish in cross-domain scenarios compared to using baseline method in **⟨Q1⟩** which has just seen unannotated data from the new domain? And, how does the performance of both of those methodologies compare to the domain-agnostic methodology of using ChatGPT to perform zero-shot classification of hate speech?

For cross-domain scenarios performance was poor across the board, hovering around  $F_1 \sim 0.3$ , with slight variations depending on the method. The model trained on the automatic pipeline (outlined in Section 3.2.3 on page 12), similar to the performance on the previous domain, made predictions that resulted in a large number of false positives, yet recall remained high even during cross-domain scenarios. For the model trained on annotated data (from the previous domain), as well as for ChatGPT, there are no distinct patterns in the predictions. This is discussed further in Section 5.4 on the next page.

### 5.2 Data discussion

The choice of data and the way the data was annotated could also have skewed the results towards better, or worse metrics for the models. It could be the case that, say in comparison to Fernquist et al. (2019), that their dataset contained “easier” to identify hate speech (i.e. in this case based on lexical co-occurrence), while the datasets in this thesis had the more difficult to identify instances of hate speech constructions (such as those requiring factual knowledge), in turn leading to poor

performance in terms of metrics. However, as mentioned in the background, this is one of the problems of automatic hate speech identification, and there is no easy fix. But, one could argue along the lines of Thomas and Uminsky (2022), in that hate speech is one of those areas where a single metric (like  $F_1$ ) perhaps isn't the absolute judge of performance, but instead, if the model does what it is supposed to do, to the best of the model's ability, in relation to the perceived harm if it performs poorly (i.e. a weighing the benefits of increased recall to lower precision, and vice versa). In that case, additional research is needed in qualitative terms of what the most effective trade-off is in terms of hate speech identification. This is further discussed in Section 5.5 on the following page.

### 5.3 Methodology discussion

One potential drawback with the methodology for the semi-automatic derivation of training data is that it uses a small amount of annotated training data to get the SVM classifier going. This would in the best case not be needed, and instead be automatically derived in either a semi-supervised fashion or by using the capabilities of a larger language model, or even better, be derived from user-reported posts (i.e. posts that have been manually reported by users of that particular website, and then subsequently removed by the administrators of the website); however, due to the scope of this thesis that couldn't be performed within the given time-frame. This is not unproblematic, however, seeing that those posts would not adhere to any strict definition of hate speech, but instead often come down to the purely subjective judgments from moderators/users, which in and of itself is a potential source of error.

When it came to cross-domain scenarios I chose to use the Automatic model from the previous domain to train the SVM-classifier for the new domain. While training the SVM-classifier on inferences from a model trained on annotated data from the previous domain could have improved performance, it was not explored, as the novel proposed methodology was meant to alleviate the need for annotated data in such scenarios, and was therefore not explored.

For the additional domain-agnostic ChatGPT methodology there are a lot of variables that can affect its performance. Not only can the choice of model have an impact on performance, where one can consider using language-specific models (like the recently released GPT-SW3 (Ekgren et al. 2022), a LLM trained on Swedish) or newer models (like GPT-4), but also the prompt itself. The prompt used in this thesis was not evaluated in any way except for some manual tweaking, and spending more time on the prompt in a more systematic way could increase performance. For example, what if the prompt had been written in Swedish, or what if the prompt had also included the definition of hate speech used for this thesis? Additionally, what if this is combined with few-shot learning, seeing that what is used in this thesis is effectively zero-shot learning?

The point regarding model choice holds for the BERT-based methodologies as well. A change in the model to a larger multi-lingual model (e.g. XLM-R (Conneau et al. 2019)) could have a positive impact on the model's performance by being more adapted in code-switching contexts.

### 5.4 Discussion of results

The issues with detecting hate speech depended on the choice of model. Many of the mistakes that BERT made were related to lexical co-occurrence, whether the model was trained on automatically derived data or not. Meanwhile, ChatGPT tended to make mistakes when it came to things like discourse-level paraphrasing. This could be in virtue of some kind of overfitting

issue in regards to BERT, in that it overfits to the lexical occurrences in the training data, which in the case of ChatGPT is not an issue because it is parametrically huge and trained for a different downstream task entirely (text generation), and therefore much more contextually sensitive. Therefore, there is reason to believe here that the future forward in regards to hate speech detection is a massive language model which is sensitive enough to pick up on the most minute of details. The argument for this stems from the likes of Sutton (2019), who argues that there is a [bitter] lesson to be learned through how the performance of downstream tasks in machine learning has improved, namely that we achieve the best results if we focus on methods which allows us to simply throw more computation at the problem, rather than specifically engineered architectures. But, there is still a long way to go, a view supported by the results from e.g. Liu et al. (2023), where they showed the difficulty LLMs have in accounting for ambiguity, even in the most recent versions like GPT-4. Since ambiguity is one of the key factors in much of the more nuanced hate speech, this is one (among many) of the challenges needed to be overcome before a robust system could be put forth. The difficulty ChatGPT had in terms of (discourse-level) paraphrasing is also in line with previous research, where Kurfali and Östling (2021), showed that several large language models had high variance on various discourse-level tasks in transfer scenarios.

The tendency for the automatic model to overpredict hate speech is an issue that could stem from the inherent difference in how humans annotate compared to algorithms like SVM used during cross-domain scenarios for this thesis, especially in the cases where we do the final judgment purely based on normative principles. Just because someone talks *about* a topic usually coinciding with hate speech, doesn't necessarily mean that they are exhibiting hateful linguistic behavior. In these cases we make normative judgments, based in virtue of factual and contextual clues, to make the final decision, or even *despite* factual and contextual clues for which e.g. hate speech is predicated upon. This paradoxical claim is supported by a recent study from Balagopalan et al. (2023), who found that human subjects made different judgments based on normative claims (e.g. does an image break against this specific list of norms, and if so why?) rather than factual claims (e.g. does this image contain the presence of these factual features), and that models trained on normative annotated data performed better.

## 5.5 Ethical Concerns

As Bender et al. (2021) argues, we must take incredible care before putting systems such as these into production, as many of these large language models risk exacerbating already present biases in the data, which in this case could potentially lead to an *increase* in hate speech toward certain demographics, especially if the systems put in place simply remove posts it sees as hate speech, as it drives people away from mainstream platforms to more esoteric platforms where hate speech perhaps is propagated in a more accepted fashion. It could also, seeing that the BERT-based methodologies in this thesis had an observed pattern of motivating its choices on lexical occurrence, make it harder to have productive discussions about controversial subjects which use similar terms to hate speech (like for example discussions about sexual assault). Instead, one could consider a system of two components, one concerned with identification, and one concerned with generating *counter-speech*, where counter-speech — as the name implies — is speech that counters the point made in the post containing hate speech, either by asking for specification or pointing out factual flaws and so on. This is in line with findings from He et al. (2021) where he found that counter-speech was successful in reducing the amount of hate speech produced by others. But, it is also worth noting here that these types of systems could also potentially be used by bad-faith actors trying to exacerbate radicalization and civil unrest. We

could consider a discussion forum that works by users up-voting or down-voting posts, which weighs the importance of that post. In that forum, a bad faith actor could build a bot that only picks out those posts which promote hate speech to up-vote, and then down-votes the rest for destabilization purposes.

Additionally, we have to keep the carbon footprint of these big models in mind. As Lacoste et al. (2019) argues, we must take care not only to run the training (or even inference in the case of massive models) in areas of the world where electricity is green, but also use other smaller machine learning models when possible, and (as used in this thesis), fine-tuning pre-trained model for downstream tasks instead of training a new one from scratch. In regards to hate speech, especially in regards to the discussion about increasing performance by just increasing computation, one could debate whether or not a less performant algorithm could be useful in production, seeing that it cuts the cost of computation, especially if combined with acceptable performance.

## 6 Conclusion

### 6.1 Answers to research questions

⟨Q1⟩: What is the baseline performance of hate speech detection in Swedish using BERT for one domain trained using manually annotated data?

The baseline performance using BERT trained on annotated data for one domain is  $F_1 = 0.68$ . While the baseline performance isn't state-of-the-art compared to previous work in hate speech detection, the result could be in virtue of a multitude of factors: the choice in hate speech definition (which affects the way the data is annotated), the choice in model, the nature of the data (i.e. less/more prevalence of more ambiguous hate speech), just to name a few. And, since no previous work has been done regarding the performance of hate speech detection using BERT in Swedish, the results present a new baseline to the literature.

⟨Q2⟩: How does the performance of hate speech detection on a single domain using BERT trained on semi-automatically derived training data (derived from an SVM-classifier trained on a small amount of annotated data) compare to the baseline set in ⟨Q1⟩?

The methodology presented in this thesis used a small amount of annotated data for the initial domain to train an SVM-classifier. The performance of the model trained on that data performed worse, with an  $F_1$ -score of 0.55 despite a larger training pool. One explanation for these results is that the BERT model overfits to the lexical features which the SVM classifier makes its predictions upon, rather than learning the underlying features of hate speech itself. This methodology used *factual* properties in the text (i.e. the occurrence of certain tokens) as the lowest predictive nominator, which could have negatively affected performance (as compared to having *normative* judgments as its base).

⟨Q3⟩: Does a model trained on semi-automatically derived data (derived from the inferences of the model trained on the previous domain) improve performance for hate speech identification in Swedish in cross-domain scenarios compared to using baseline method in ⟨Q1⟩ which has just seen unannotated data from the new domain? And, how does the performance of both of those methodologies compare to the domain-agnostic methodology of using ChatGPT to perform zero-shot classification of hate speech?

A model trained on annotated data performs best ( $F_1 = 0.32$ ), with ChatGPT closely behind ( $F_1 = 0.30$ ) and the model trained on semi-automatically derived data last ( $F_1 = 0.27$ ). One explanation for these results is that the BERT-based methodologies over attribute lexical cues as predictive of hate speech, while ChatGPT instead made mistakes when it came to things such as discourse level paraphrasing and other fine-grained semantic cues. The results suggest that the way forward in the area of automatic hate speech detection is not set in stone: smaller models could remain useful as it is less expensive both in terms of economy and the environment, and the extent to how good current LLMs can be in terms of hate speech detection remains open-ended.

## 6.2 Summary

This thesis attempted to find a general methodology for hate speech detection in cross-domain scenarios by evaluating three methodologies: using the inferences from a model trained on annotated for another domain, a model trained on semi-automatically derived training data, and using ChatGPT for annotation.

Results showed that performance was generally poor, with the model trained on semi-automatically derived data achieving consistently worse  $F_1$ -score for both the initial domain as well as for cross-domain scenarios (higher recall, much lower precision). An explanation for the results is that both methods using BERT overfit to their respective domains to some capacity, while ChatGPT is confused by semantic properties (like paraphrasing). These results reflect the difficulty in automatic hate speech detection, with the automatic methodology introduced in this thesis shedding light on the relationship between how much BERT can overcome noisy data for the downstream task of hate speech detection in virtue of an increase in training pool size.

Novel contributions from this thesis are as follows: using BERT to perform hate speech detection in Swedish (and with it a novel methodology, as well as a novel baseline for hate speech detection in Swedish), using ChatGPT for hate speech identification in Swedish.

Moving forward, much care has to be put into the ethical concerns regarding hate speech detection before it can be put effectively into production, regardless of its need. Not only could these systems have the opposite effect to curbing hate speech (by making users move to other more esoteric forums where hate speech is encouraged), but they could also be misused by bad faith actors for destabilization purposes.

## 6.3 Future Research

The potential of counter-speech generation for hate speech in Swedish is an unexplored area, which has greater harm-reducing potential than merely identifying hate speech in combination with the removal of said posts as noted previously, as well as exploring the capabilities of LLMs in regards to hate speech detection (prompt engineering, combining with few-shot learning, etc.) — which is not only unexplored for Swedish but as far as I can tell, even for English. Another interesting area is if one were able to achieve cooperation with some of the owners of these discussion forums, one could see how the performance would change if one got access to user-reported posts. Additionally, there is room to explore the qualitative sides of successful hate speech identification (in addition to metrics like  $F_1$ -score), and future research could look into exactly what those possible qualitative metrics could be so that the systems we produce are aligned to human goals. Lastly, additional research looking into comparing data annotated purely based on factual properties (like those used in this thesis, see Section A on page 30) to data annotated on normative principles, and how that compares in terms of automatic hate speech identification performance.

## References

- Aluru, Sai Saketh, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee (2020). *Deep Learning Models for Multilingual Hate Speech Detection*. DOI: [10.48550/ARXIV.2004.06465](https://doi.org/10.48550/ARXIV.2004.06465). URL: <https://arxiv.org/abs/2004.06465>.
- Bäck, Emma A., Hanna Bäck, Marie Gustafsson Sendén, and Sverker Sikström (2018). From I to We: Group Formation and Linguistic Adaption in an Online Xenophobic Forum. In: *Journal of Social and Political Psychology* 6.1, pp. 76–91. DOI: [10.5964/jspp.v6i1.741](https://doi.org/10.5964/jspp.v6i1.741). URL: <https://jspp.psychopen.eu/index.php/jspp/article/view/5043>.
- Balagopalan, Aparna, David Madras, David H Yang, Dylan Hadfield-Menell, Gillian K Hadfield, and Marzyeh Ghassemi (May 2023). Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. In: *Science Advances* 9.19, eabq0701.
- Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen (2014). Gender identity and lexical variation in social media. In: *Journal of Sociolinguistics* 18.2, pp. 135–160.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. ISBN: 9781450383097. DOI: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922). URL: <https://doi.org/10.1145/3442188.3445922>.
- Bilewicz, Michał and Wiktor Soral (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. In: *Political Psychology* 41, pp. 3–33.
- Brown, Alexander (Aug. 2017). What is hate speech? Part 1: The Myth of Hate. In: *Law and Philosophy* 36.4, pp. 419–468. ISSN: 1573-0522. DOI: [10.1007/s10982-017-9297-1](https://doi.org/10.1007/s10982-017-9297-1). URL: <https://doi.org/10.1007/s10982-017-9297-1>.
- (2018). What is so special about online (as compared to offline) hate speech? In: *Ethnicities* 18.3, pp. 297–326. DOI: [10.1177/1468796817709846](https://doi.org/10.1177/1468796817709846). eprint: <https://doi.org/10.1177/1468796817709846>. URL: <https://doi.org/10.1177/1468796817709846>.
- Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2019). Unsupervised Cross-lingual Representation Learning at Scale. In: *CoRR* abs/1911.02116. arXiv: [1911.02116](https://arxiv.org/abs/1911.02116). URL: <http://arxiv.org/abs/1911.02116>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. DOI: [10.48550/ARXIV.1810.04805](https://doi.org/10.48550/ARXIV.1810.04805). URL: <https://arxiv.org/abs/1810.04805>.
- Dowlagar, Suman and Radhika Mamidi (2021). *HASOCOne@FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection*. DOI: [10.48550/ARXIV.2101.09007](https://doi.org/10.48550/ARXIV.2101.09007). URL: <https://arxiv.org/abs/2101.09007>.
- Ekgren, Ariel, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlin-den, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren (June 2022). Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3509–3518. URL: <https://aclanthology.org/2022.lrec-1.376>.
- Essen, Emma von and Joakim Jansson (2020). Misogynistic and xenophobic hate language online: a matter of anonymity. In: *IFN Working Paper* 1350. DOI: [10.2139/ssrn.3682069](https://doi.org/10.2139/ssrn.3682069).

- Fernquist, Johan, Oskar Lindholm, Lisa Kaati, and Nazar Akrami (2019). A Study on the Feasibility to Detect Hate Speech in Swedish. In: *2019 IEEE International Conference on Big Data (Big Data)*, pp. 4724–4729. DOI: [10.1109/BigData47090.2019.9005534](https://doi.org/10.1109/BigData47090.2019.9005534).
- Forman, Ira (2018). *Is Anti-Israel Anti-Semitic?* Accessed 2023/03/08. URL: <https://berkleycenter.georgetown.edu/responses/is-anti-israel-anti-semitic>.
- Fortuna, Paula and Sérgio Nunes (2018). A survey on automatic detection of hate speech in text. In: *ACM Computing Surveys (CSUR)* 51.4, pp. 1–30.
- Framework Decision 2008/913/JHA (2008). *Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law*. European Parliament and Council. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:133178>.
- Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*. arXiv: [2303.15056](https://arxiv.org/abs/2303.15056) [cs.CL].
- Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith (2020). *Don't Stop Pretraining: Adapt Language Models to Domains and Tasks*. DOI: [10.48550/ARXIV.2004.10964](https://doi.org/10.48550/ARXIV.2004.10964). URL: <https://arxiv.org/abs/2004.10964>.
- Hall-Lew, Lauren, Elizabeth Coppock, and Rebecca L Starr (2010). Indexing political persuasion: Variation in the Iraq vowels. In: *American Speech* 85.1, pp. 91–102.
- He, Bing, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar (2021). Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 90–94.
- Hemphill, Libby and Jahna Otterbacher (2012). Learning the lingo? Gender, prestige and linguistic adaptation in review communities. In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pp. 305–314.
- Howard, Jeremy and Sebastian Ruder (2018). *Universal Language Model Fine-tuning for Text Classification*. DOI: [10.48550/ARXIV.1801.06146](https://doi.org/10.48550/ARXIV.1801.06146). URL: <https://arxiv.org/abs/1801.06146>.
- Internetstiftelsen (2021). *Svenskarna och internet*. Accessed 2023/03/08. URL: <https://svenskarnaochinternet.se/rapporter/svenskarna-och-internet-2021/sociala-medier/youtube-facebook-och-instagram-ar-de-tre-sociala-medier-som-anvants-mest-det-senaste-aret/>.
- Jahan, Md Saroar and Mourad Oussalah (2021). *A systematic review of Hate Speech automatic detection using Natural Language Processing*. DOI: [10.48550/ARXIV.2106.00742](https://doi.org/10.48550/ARXIV.2106.00742). URL: <https://arxiv.org/abs/2106.00742>.
- Keipi, Teo, Matti Näsi, Atte Oksanen, and Pekka Räsänen (2016). The rise of online hate. In: *Online Hate and Harmful Content*. Routledge. Chap. 4, pp. 53–74. ISBN: 9781315628370.
- Koch, Timo K, Peter Romero, and Clemens Stachl (2022). Age and gender in language, emoji, and emoticon usage in instant messages. In: *Computers in Human Behavior* 126, p. 106990.
- Kurfali, Murathan and Robert Östling (2021). *Probing Multilingual Language Models for Discourse*. arXiv: [2106.04832](https://arxiv.org/abs/2106.04832) [cs.CL].
- Labov, William (2006). *The social stratification of English in New York city*. Cambridge University Press. DOI: [10.1017/CB09780511618208](https://doi.org/10.1017/CB09780511618208).
- Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres (2019). *Quantifying the Carbon Emissions of Machine Learning*. arXiv: [1910.09700](https://arxiv.org/abs/1910.09700) [cs.CY]. URL: <https://arxiv.org/abs/1910.09700>.

- Liu, Alisa, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi (Apr. 2023). *We're Afraid Language Models Aren't Modeling Ambiguity*. URL: <https://arxiv.org/abs/2304.14399>.
- Lucy, Li and Julia Mendelsohn (2018). *Using Sentiment Induction to Understand Variation in Gendered Online Communities*. DOI: [10.48550/ARXIV.1811.07061](https://doi.org/10.48550/ARXIV.1811.07061). URL: <https://arxiv.org/abs/1811.07061>.
- MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder (Aug. 2019). Hate speech detection: Challenges and solutions. In: *PLOS ONE* 14.8, pp. 1–16. DOI: [10.1371/journal.pone.0221152](https://doi.org/10.1371/journal.pone.0221152). URL: <https://doi.org/10.1371/journal.pone.0221152>.
- Malmsten, Martin, Love Börjeson, and Chris Haffenden (2020). *Playing with Words at the National Library of Sweden – Making a Swedish BERT*. arXiv: [2007.01658](https://arxiv.org/abs/2007.01658) [cs.CL].
- OpenAI (2022). *Introducing ChatGPT*. Accessed 2023-04-28. URL: <https://openai.com/blog/chatgpt>.
- Pham, Quang Huu, Viet Anh Nguyen, Linh Bao Doan, Ngoc N Tran, and Ta Minh Thanh (2020). From universal language model to downstream task: improving RoBERTa-based Vietnamese hate speech detection. In: *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, pp. 37–42.
- Popper, Karl (1945). *The open society and its enemies*. Routledge Classics. Reprint 2012. London, England: Routledge.
- Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. (2018). *Improving language understanding by generative pre-training*. OpenAI. URL: <https://openai.com/research/language-unsupervised>.
- Roß, Björn, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. en. In: DOI: [10.17185/DUEPUBLICO/42132](https://doi.org/10.17185/DUEPUBLICO/42132). URL: <http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=42132>.
- Sai, Siva and Yashvardhan Sharma (2020). Siva@ HASOC-Dravidian-CodeMix-FIRE-2020: Multilingual Offensive Speech Detection in Code-mixed and Romanized Text. In: *FIRE (Working Notes)*, pp. 336–343.
- Schwenk, Holger and Matthijs Douze (2017). *Learning Joint Multilingual Sentence Representations with Neural Machine Translation*. arXiv: [1704.04154](https://arxiv.org/abs/1704.04154) [cs.CL].
- Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). Dropout: a simple way to prevent neural networks from overfitting. In: *The journal of machine learning research* 15.1, pp. 1929–1958.
- Sutton, Rich (2019). *The Bitter Lesson*. Accessed 2023-01-08. URL: <https://web.archive.org/web/20230108040226/http://incompleteideas.net/IncIdeas/BitterLesson.html>.
- Thomas, Rachel L. and David Uminsky (2022). Reliance on metrics is a fundamental challenge for AI. In: *Patterns* 3.5, p. 100476. ISSN: 2666-3899. DOI: <https://doi.org/10.1016/j.patter.2022.100476>. URL: <https://www.sciencedirect.com/science/article/pii/S2666389922000563>.
- Uszkoreit, Jakob (2017). *Transformer: A Novel Neural Network Architecture for Language Understanding*. Accessed 2020/03/28. Google. URL: <https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). *Attention Is All You Need*. DOI: [10.48550/ARXIV.1706.03762](https://doi.org/10.48550/ARXIV.1706.03762). URL: <https://arxiv.org/abs/1706.03762>.

- White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. arXiv: 2302.11382 [cs.SE]. URL: <https://arxiv.org/abs/2302.11382>.
- Winiewski, Mikolaj, Wiktor Soral, and Michal Bilewicz (2017). *Contempt speech, hate speech. Report from research on verbal violence against minority groups*. URL: [www.ngofund.org.pl/wp-content/uploads/2017/02/Contempt\\_Speech\\_Hate\\_Speech\\_Full\\_Report.pdf](http://www.ngofund.org.pl/wp-content/uploads/2017/02/Contempt_Speech_Hate_Speech_Full_Report.pdf).
- Wolkoff, Robert L. (2020). *Criticism of Israel is Anti-Semitism. Really*. Accessed 2023/03/08. URL: <https://blogs.timesofisrael.com/criticism-of-israel-is-anti-semitism-really/>.
- Yoder, Michael Miller (2021). *Computational Models of Identity Presentation in Language*. PhD thesis. Carnegie Mellon University. URL: [https://raw.githubusercontent.com/michaelmilleryoder/michaelmilleryoder.github.io/master/files/yoder\\_thesis.pdf](https://raw.githubusercontent.com/michaelmilleryoder/michaelmilleryoder.github.io/master/files/yoder_thesis.pdf).
- Zhu, Wanzheng, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat (2021). Self-Supervised Euphemism Detection and Identification for Content Moderation. In: *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 229–246. DOI: 10.1109/SP40001.2021.00075.

## Appendix A Annotation Guidelines

NB. the party politics sentiment categories were annotated, but not used for the creation of the datasets in this thesis.

1	<b>Party politics positive</b>	
	0 = No, not positive to any coalition of parties 1 = Yes, the red-green coalition 2 = Yes, the liberal-conservative coalition 3 = Yes, the Sweden Democrats 4 = Feminist Initiative 5 = Feminist initiative and the Left Party 6 = The seven traditional parties 7 = Sweden Democrats and the Right	Only to be coded "yes" if it is obvious, e.g. when the parties or their representatives are mentioned, either explicitly or through paraphrases
2	<b>Party politics negative</b>	
	0 = No, not negative to any coalition of parties 1 = Yes, the red-green coalition 2 = Yes, the liberal-conservative coalition 3 = Yes, the Sweden Democrats 4 = Feminist Initiative 5 = Feminist initiative and the Left Party 6 = The seven traditional parties 7 = Sweden Democrats and the Right.	Only to be coded "yes" if it is obvious, e.g. when the parties or their representatives are mentioned, either explicitly or through paraphrases
3	<b>Aggressiveness</b>	
	0 = Not at all aggressive 1 = Partly aggressive 2 = Predominately aggressive	If some parts are aggressive annotate as 1, if it mostly contains aggressive text, annotate as 2
4	<b>Hatred</b>	
	0 = No 1 = Yes Another user Public persona Sex/gender Born abroad/ parents born abroad Ethnicity Sexual preference	If the post contains hatred against a specific person/group (column to the left);  Possible examples: threat, expressions of disrespect, insults.
5	<b>Threat</b>	
	0 = No 1 = Yes Another user Public persona Sex/gender Born abroad/ parents born abroad Ethnicity Sexual preference	If the post contains threat / assault against a specific person/group (column to the left). Assault means that someone threatens to harm an individual or his/her property. The assault can be targeted against other people/animals/properties which are important to that person.
6	<b>Male preference</b>	
	0 = No 1 = Yes	Annotate as yes if the post includes words and/or expressions which state the superiority of men over women, otherwise annotate as no.
7	<b>Female preference</b>	
	0 = No 1 = Yes	Annotate as yes if the post includes words and/or expressions which state the superiority of women over men, otherwise annotate as no.
8	<b>Gender equality preference</b>	
	0 = No 1 = Yes	Annotate as yes if the post includes word and/or expression which state a preference for gender equality

Table 10: Annotation guidelines

9	<b>Attitude towards foreigners</b>	
	0 = No opinion 1 = Positive Attitude 2 = Negative Attitude 3 = Neutral Attitude	Foreigners specifically referring to simply people who were born abroad.
10	<b>Attitude towards religion</b>	
	0 = No opinion 1 = Positive Attitude 2 = Negative Attitude 3 = Neutral Attitude	Annotate as negative if the post expresses negative opinion against explicitly stated religion without specific criticism or misguided criticism. For example, one could imagine a post criticizing a religion on its view on women in bad faith (i.e. not based on any examples) as being labelled as negative. A positive example would be the opposite. A neutral example would simply be commenting on aspects of the religion.
11	<b>Gender disadvantage</b>	
	0 = No opinion 1 = Men are disadvantaged 2 = Women are disadvantaged	
12	<b>Ethnicity disadvantage</b>	
	0 = No opinion 1 = Swedes are disadvantaged 2 = Immigrants are disadvantaged	
13	<b>Us vs. Them</b>	
	0 = No 1 = Yes	If the post explicitly contains a language of "us and them" or clearly expresses an in-group out-group view the variable should be coded yes.
14	<b>Sarcasm / Irony</b>	
	0 = No 1 = Yes, partly 2 = Yes, fully	

Table 11: Annotation guidelines (continued from last page)

## Appendix B Hyper Parameters

Hyper-parameters are also present in the code itself on

<https://github.com/skogsgren/a-tale-of-two-domains>.

The specific commit SHA for the code in use as of the writing of this thesis is 925a3e7376c25db90edab7a3d7a9b4db503e55d3.

```
DATASET_SIZE : {train: 100000, test: 10000}  
EPOCHS : 15  
OPTIMIZER : AdamW  
LR : 5e-5  
BATCH_SIZE : 12  
MASKING_PROBABILITY : 0.15
```

Figure 5: Hyper parameters for masked language model finetuning

```
DATASET : {train: 2585, test: 647}  
EPOCHS : 3  
OPTIMIZER : AdamW  
LR : 1e-5  
BATCH_SIZE : 16  
DROPOUT : 0.2  
WEIGHT_DECAY : 0.0005
```

Figure 6: Hyper parameters for text classification head using annotated data (i.e. [Q1](#))

```
DATASET : {train: 25000, test: 647}  
EPOCHS : 3  
OPTIMIZER : AdamW  
LR : 5e-5  
BATCH_SIZE : 16  
DROPOUT : 0.1  
WEIGHT_DECAY : 0.001
```

Figure 7: Hyper parameters for text classification head using automatically derived training data (i.e. [Q2](#))

Stockholm University  
SE-106 91 Stockholm, Sweden  
Telephone +46 (0)8 16 20 00  
<https://www.su.se/>



**Stockholm**  
**University**