Gerold Schneider, Janis Goldzycher and Martin Volk

Detecting and Mapping Hate in Religious Contexts

Abstract: Hate speech, a societal challenge with broad implications for democratic discourse, inclusivity and security of minorities, is propagated on online platforms. This chapter explores methods for detecting and analyzing hate speech, focusing on religious contexts. The chapter is structured into two main parts: The first part describes the key ideas underlying AI-based hate speech detection, highlights challenges of current approaches, and then introduces our approach for more modular, efficient hate speech detection systems. The second part dives into conceptual maps as a tool for visualizing associations in large datasets, applied to Twitter data. This analysis reveals a pattern of negative sentiment and the politicization of religion, in contrast to spirituality, which is associated with more positive connotations and perceptions. The chapter concludes by contextualizing these findings, emphasizing the potential of combining quantitative with qualitative analysis to obtain a more nuanced understanding of hateful discourses, and religion-related discourses on the internet.

Hassrede, eine gesellschaftliche Herausforderung mit weitreichenden Auswirkungen auf den demokratischen Diskurs, Inklusivität und die Sicherheit von Minderheiten, erfährt derzeit grosse Verbreitung auf Onlineplattformen. Dieses Kapitel untersucht Methoden zur Erkennung und Analyse von Hassrede mit einem Fokus auf religiöse Kontexte. Das Kapitel ist in zwei Hauptteile gegliedert: Der erste Teil beschreibt die zentralen Ideen, die der KI-basierten Erkennung von Hassrede zugrunde liegen, hebt die Herausforderungen für aktuelle Ansätze hervor und stellt anschließend unseren Ansatz für modularere und effizientere Systeme zur Erkennung von Hassrede vor. Der zweite Teil widmet sich konzeptionellen Karten als Werkzeug zur Visualisierung von Assoziationen in großen Datensätzen. Bei der Anwendung auf Twitter-Daten zeigt sich eine negative Assoziation und Politisierung von Religion, im Gegensatz zur Spiritualität, die positiver wahrgenommen wird. Das Kapitel schließt mit einer Kontextualisierung dieser Erkenntnisse ab und betont das Potenzial, quantitative und qualitative Analysen zu kombinieren, um ein differenzierteres Verständnis von Hassrede und religionsbezogenen Diskussionen im Internet zu gewinnen.

Open Access. © 2025 the author(s), published by De Gruyter. BY-NC-ND This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
https://doi.org/10.1515/9783111721729-009

1 Introduction

Hate speech is a serious challenge for our society. We follow the UN's definition of hate speech which goes as follows: Hate speech is "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."[1] Hate speech hurts people, and it damages a culture of mutual respect and understanding. Hate speech may be spurned by the anonymity of the internet. Social media seems to be particularly affected. The impression that it has increased in frequency can be confirmed. After Elon Musk has acquired Twitter and renamed it to X, several findings support this impression. Racist posts have become more frequent,[2] and the social media index GLAAD confirms that user safety scored for the LGBTQ+ community have decreased further, after they had already decreased a year ago.[3] Their study included all major social media platforms, including Facebook, Instagram, TikTok, YouTube, and X. The suspicion that hate speech is simply more frequent because Musk's policy may be to censor fewer posts is thus not an explaining factor. At the same time, Musk's policy changes had clear effects on X. On the one hand, Musk's posts and Republican posts generally are boosted by the algorithm (Graham and Andrejevic 2024). This may have had an effect of the US elections, which Trump won, and the boosting may become even stronger when he takes up office. On the other hand, X users are now more likely to be exposed to hateful content.[4] The amount of hate speech and thus exposure to it also depends on the topic. Politicized issues attract hate speech more than leisure topics. In this article we will assess the situation in view of topics related to religious debates, using data-driven approaches with few theoretical assumptions. We will see that many aspects of religion are affected very much by hate, politicization, and discrimination, while others, in particular spirituality, are often portrayed as offering innocent happiness.

In this chapter, we describe the two main contributions of our URPP project: First, how can we detect hate speech with state-of-the-art methods? For this we present the computational linguistic algorithms that we have applied and im-

[1] See https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech, last visited at 30.10.2024.

[2] https://www.ohchr.org/en/statements/2023/01/freedom-speech-not-freedom-spread-racial-hatred-social-media-un-experts.

[3] https://glaad.org/publications/social-media-safety-index-2023/.

[4] https://www.washingtonpost.com/technology/2023/03/30/elon-musk-twitter-hate-speech/.

proved, using extensive and targeted training material. Second, we address the question of how much hate and negative sentiment is present in the religious discourse. For this, we investigate frequently used words in the context of religion, spirituality, and faith from social media. Before we delve into these, we further motivate our research, by showing why the detection of hate speech is important.

Hate speech is unpleasant and hurts the attacked groups or individuals. Even the attackers are at least partly aware of this fact, or they may also intend to hurt people. They thus violate the human right(s) of liberty and security and dignity. To a certain extent, even the right of life is jeopardized, at least the attackers are willing to play with fire. Bereft of context, it is difficult for the attacked to distinguish an uncontrolled bout of anger from a well-planned attack.

But censoring hate speech is also problematic. Censorship violates the human right of free speech, freedom of opinion and expression. Therefore, hate speech can lead to a clash of human rights and a bias towards excluding some voices in society (see also Frei et al. this volume).

1.1 A Clash of Human Rights

While there is no unanimous definition of human rights, most sources testify a strong overlap. In order to assess which human rights are concerned, we adopt the definition given on the learning platform Vaia.[5] Their list comprises the following ten items:

1. Right to life: Every person has the right to live and not be deprived of life unlawfully.
2. Freedom from torture: No person should be subject to torture or cruel, inhuman, or degrading treatment.
3. Right to liberty and security: Everyone has the right to be free from arbitrary arrest or detention.
4. Freedom of thought, conscience, and religion: All individuals have the right to hold and practice their beliefs freely.
5. Freedom of opinion and expression: People have the right to hold and share opinions and ideas without interference or censorship.
6. Right to work and education: Everyone has the right to work in fair and safe conditions and to receive an education.
7. Right to privacy: All individuals have the right to privacy in their personal, family, home, and correspondence lives.

[5] https://www.hellovaia.com/explanations/law/human-rights-law/fundamental-human-rights/.

8. Right to participate in government: Every person has the right to take part in their country's political affairs and exercise their right to vote.
9. Freedom of movement: People have the right to move freely within their country and to leave and return to it.
10. Right to equality before the law: All individuals are entitled to equal protection of the law without discrimination.

Unchecked hate speech may violate right 4, and jeopardize rights 1, 2, and possibly 3. Right 4 is especially tricky, as it potentially contains a clash in itself: freedom of thought entails the right to strongly dislike or even hate something, while freedom of religion entails the right to practice any religion and belief without fear. There is a crucial difference between rights 4 and 5. Right 4 postulates that thoughts need to be free, while right 5 entails the danger that voicing one's opinions with check may lead to hate speech, which poses a danger to rights 1 to 3.

Defenders of unlimited free speech typically argue that verbal expressions do not lead to bodily harm, thus rights 1 to 3 are still guaranteed. However, there are two main reasons that challenge the validity of this claim: First, clearly expressed intentions and desires encourage the execution of violent acts in the real world. Second, attacked groups may also psychologically suffer from degrading treatment in verbal form.

Discussions on what is best for society will, of course, continue. If we compare this conflict to economical models, both the fully liberal free-market model and highly protected communist ideologies are seen as failed by most theorists. In a similar vein, it is likely that fully unchecked hate speech will lead to disaster as well as filtering all negative comments, and that the debate of what exactly hate speech is, and how one should deal with it is unlikely to find an easy settlement. What we need in either way are tools that detect hate speech allowing social media platforms, regulators and the government to take appropriate measures. Our contribution is to provide such tools.

1.2 Definitions of Hate Speech

Finding an appropriate balance between freedom of speech and protecting groups is tricky, and the debate may be heated. In hate speech detection research various definitions of hate speech have been adopted and proposed (Khurana et al. 2022). However, these definitions are typically guided by the UN's definition of hate speech, given at the beginning of this section.

This definition entails that a sentence like "I despise you" is not hate speech, because the attack does not reference a *protected characteristic* such as religion or ethnicity. However, the sentence “I despise these Muslims” or “I despise you, you Muslim” falls under the hate speech definition because the attack works via reference to such a characteristic. In this context the term *protected group* is often used in hate speech detection research to refer to groups that are defined via protected characteristics. For example, “black people” is a protected group because it is defined via the “race” or “colour” characteristic, but “teachers” is not a protected group because occupation is not a protected characteristic.

1.3 The Tenuous Relation between Thought and Action

While elucidating the clash between human rights in the previous subsection we have already hinted at the next clash: thoughts must be free and cannot be controlled, not even in the most oppressive dictatorship. However, the correlation of thought and action is disputed, in fact a classic in philosophy, ranging from common sense utterances that action speaks louder than words, that one should judge people (both in moral and legal terms) not on their thoughts but on their actions, up to the reminder that there is indeed a connection, that calls for violence often lead to real violence. Let the words of Nobel prize winner Maria Ressa speak, quoted in Hietanen and Eddebo (2023):

> “Online violence does not stay online. Online violence leads to real world violence.” Maria Ressa, Recipient of the Nobel Peace Prize (SVT, 2021, 1:04:03).

We could formulate this second clash as follows: While there is unanimity that the condemnation of violence is a cornerstone of civilized society, there is disagreement (1) on whether verbal expressions as such constitute harm and (2) on whether hate speech incites violence in the real world.

Point (1) involves several aspects: While it is clear that psychological harm can be as hurtful as physical harm, it is harder to measure. Theoretical discussions do not necessarily conceive of hate speech as harm per se (Barendt 2019); however, one could argue that it should be prohibited due to its potential effects on society.

Point (2) is easier to defend, and also Barendt (2019) points out: If hate speech leads to physical harm, it needs to be detected and censored, particularly if vulnerable minorities are affected. Also, defendants of direct democracy sometimes forget that the touchstone of democracy is not only the rule of majority but equally the protection of minorities.[6] The United Nations are unequivocal in

[6] https://www.principlesofdemocracy.org/majority.

stressing that hate speech is often a precursor to real violence. History provides too many proofs of this fact, ranging from the Holocaust to the Srebrenica genocide in Bosnia and Herzegovina. There is also mounting evidence that online hate can turn into real-life violence. Williams et al. (2019) investigate correlations between police crime and Twitter data to show that there is a positive correlation between social media hate and real-world crime. The authors conclude: "This research shows that online hate victimization is part of a wider process of harm that can begin on social media and then migrate to the physical world." (Williams et al. 2019, 114).

Further, psychological studies now also confirm that there is a cause-and-effect relation, as hate speech possibly causes mental differences: Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain, as Pluta et al. (2023) have found out. From a utilitarian perspective, it is also important to point out that calls for hatred and violence are pointless if they are not meant seriously in the majority of cases. If all authors of hate speech were convinced that there is a complete disjoint between actions and words there would be very little hate speech.

1.4 Why is Hate Speech Harmful?

After zooming in on the clashes of human rights and the freedom of speech, let us broaden the perspective again and remind ourselves of reasons why hate speech is harmful and may have a corrosive effect on society, and thus why tools such as ours are needed. This list is not encompassing.

- Hate speech affects inclusion and participation in society and its institutions.
- It creates fear and anxiety among the targeted groups.
- Children are at particular risk, as they are particularly defenseless and have less life experience to judge which threats are real.
- Hate speech divides and polarizes society.
- Hate speech is often not based on facts. In a post-fact society, any rumor can rise to a monster.
- Hate speech does not contribute to a solution to possibly real problems, it does not even intend to do so.
- Hate speech may violate human rights (as we have elaborated above).
- Hate speech is often a precursor to real violence (as we have elaborated above).
- Hate speech often meets no immediate resistance in the anonymous space of the internet.

- If unanswered, and not met with resistance such as counter speech, hate speech may lead to radicalization.
- Hate speech is often also harmful for the attackers: an uncontrolled fit of anger may lead to exclusion, to loss of reputation or even one's job, which can set off a vicious downward spiral.
- Hate speech is a real threat to democracy, because the attacked groups are often de-humanized, for instance their human rights are questioned, and the second pillar of democracy – protection of minorities – is jeopardized.

1.5 Why Can Hate Speech not be Filtered Manually?

The amount of posts on social media is simply too large to use manual filtering methods. Also, the fact that social platforms are under financial constraints makes it impossible in practice. Only filtering samples, for instance, to find perpetrators is not a viable answer either, as legal regulations (for example in the EU) demand reliable detection (see, for instance, the court case of EU vs. Musk[7]).

Even if enormous resources and a workforce were dedicated to the task, further problems remained. The task of manual filtering is tedious and very repetitive, and constant exposure to hate speech psychologically affects people, and an army of annotators would possibly be less consistent than a reproducible algorithm.

2 Theoretical Basis and Methodology

After showing why the detection of hate speech is important in the last section, we now turn to the methods needed to address our research objectives: First, we present how one can detect hate speech with state-of-the-art methods (Section 2.1). Second, we present a method that allows us to draw a map of hate and negative sentiment in the religious discourse (section 2.2).

[7] https://www.voanews.com/a/like-brazil-the-european-union-also-has-an-x-problem/7772200.html.

2.1 Hate Speech Detection

2.1.1 An Excursion into Classification Methods

On a computational level, hate speech detection is typically formulated as a text classification problem (Poletto et al. 2020). Each incoming text, typically a social media post or user comment under a news article is classified as either hate speech or not-hate speech. The classification of texts into classes is typically called text classification or document classification and is a long-established task. What all classification models have in common is that they do not only look at a few isolated single indicators but at a multitude, and their interaction. The fact that many indicators (we will use the term *features*) are used already reduces the risk of making errors.

Document Classification

Each document, whether a newspaper article, a web page, a book, a paragraph, a tweet, or a similar discourse unit, is assigned to a class. Classes can, for example, be broad topics divided into the binary classes of relevant or irrelevant documents for an Information Retrieval task (Jurafsky and Martin 2009, chapter 23.1, page 781 ff.; Manning and Schütze 2001, chapter 15.1, page 530 ff. for an introduction). The research questions can deal with e.g. a positive or negative assessment of a political issue in automated content analysis (Grimmer and Stewart 2013 for an introduction). In the majority of the implementations, the words in the documents are used as discriminators between the classes, typically without respecting their sequence or syntactic context.

It would be better to use word sequences, words in combination and their interaction. Document classification respects the combination of words in the sense that it includes all words in a given document. The interaction model is a radically simple one, though. It is only counted how often a word occurs in a given document. The sequence of words and their position in the document or in the sentence is not taken into consideration. That is why this method is often called a bag-of-words model. In order to include a minimal notion of word order, the model is often extended to include frequent sequences of two words (bigram model) or three words (trigram model). Longer sequences are typically not used, as most longer sequences have very low frequencies.

Distributional Semantics and Word Embeddings

One further problem for document classification is that a bag-of-words approach does not know which words are similar. One can extract this knowledge, however, from large texts by considering the typical contexts of each word. Similar words tend to occur in similar contexts, because human language is inherently redundant. The Firthian hypotheses, summarized by "You shall know a word by the company it keeps" (Firth 1957, 14) allows one to detect similar words from the sums of their contexts, and can therefore add semantic knowledge to language models.

Sahlgren (2006) shows that while a very narrow context such as word adjacency delivers linguistic collocations, i.e. relations at the syntagmatic level, larger context windows, such as 10 words before and after, deliver semantic relations and associations, i.e. relations at the paradigmatic level. This insight is also exploited by the hugely successful research paradigm of distributional semantics (Baroni and Lenci 2010), which aims to detect synonyms, antonyms, and hyponyms of words.

Models predicting word similarity are now often based on neural networks. This approach is called word embedding, it goes back to Baroni et al. (2014). The probabilities of words given their context are predicted in two alternative ways, we briefly describe now. The probably more frequently used one, which is known as CBOW (continuous bag-of-word), works as follows:

> If we use a context of 5 words, i.e. 2 words before and after the word to be predicted, the probabilities of the word to be predicted, in terms of conditional probabilities, is:
>
> $p(w_0 \mid w_{-2}, w_{-1}, w_{+1}, w_{+2})$ [assuming a context window of 5 words].
>
> All words would need to be predicted, which is too costly in practical terms. To reduce complexity, negative sampling and word hierarchies are used.

Supervised, Unsupervised, and Self-Supervised Learning

While document classification needs texts that are annotated for the classes that the algorithm should be able to detect for new documents, word embeddings are learned purely from the texts. Document classification is a typical instance of supervised learning, while word embedding is an instance of unsupervised learning.

More recently, an approach called self-supervised learning has become very influential, as it is the background of Large Language Models such as BERT and GPT, the latter is the base for the famous ChatGPT tool. Supervised learning approaches typically perform better than unsupervised approaches, but annotating data is very labor-intensive. It is usually not possible to annotate millions of documents. Unsupervised learning has the advantage that it can profit from the almost unrestricted amounts of data available today, such as complete web scrapes and Wikipedia dumps.

Self-supervised learning takes these enormous amounts of textual data and makes class predictions that are readily available, although they may seem to be very far away from the prediction that is required for a given annotation task. Self-supervised learning predicts the next word (this is why they are also called generative models, as they can directly generate text based on an initial sequence, for instance, a sentence), some models also predict missing words (gap filling, like in a cloze test) or the full sentence. Predicting missing words is an approach that we have just seen in word embeddings, so in a sense self-supervised approaches represent a further development of word embeddings, not only in terms of their historical development.

BERT models focus on predicting missing words in the following fashion: every 15th word is masked, and the training process learns to predict it as accurately as possible from the unmasked words. For this reason, these models are sometimes also called masked language models. Although self-supervised models are basically trained for the “wrong” task – unless you want to predict word sequences – their world knowledge is impressive. They have seen more text than an experienced human in his/her entire life. Due to this, they typically only need little adaptation to be tuned to a specific task, such as question answering, natural language inference, text summarization or hate speech detection.

Large Language Models (LLMs)

The models that emerge in state-of-the art self-supervised learning are several orders of magnitude larger than the largest supervised models. The number of features (often also called parameters) used for document classification or Distributional Semantics is roughly 10^4. BERT models and the first GPT model (GPT-1) have about 10^8 = 100 million parameters. BERT and BART models have between 110 million and 345 million (we use a large BART model in section 4.1). GPT-3 has about 175 billion (=10^{11} parameters). As the models are so complex, and often perform as well as document classification, but on a large variety of tasks, users no longer train then from scratch, which would also be unecological. Training a GPT-3 model from scratch uses as much energy as a thousand US households per year.

Neural Networks and Transformers

We mentioned each node in a neural can be thought of as a separate logistic regression. In addition to the logit function known from logistic regression, other activation functions can also be used to trigger a node or “neuron” to fire or not.[8]

Classical feed-forward networks arrange the nodes in a grid of several layers, each layer containing several nodes, and every node is connected to each node in the subsequent layer. The number of layers defines how “deep” the deep neural network is. These networks have been used successfully for many tasks, also in computational linguistics. A class of neural networks called transformers (Vaswani et al. 2017) is particularly successful, because they observe a very lage context window, by means of the so-called attention mechanism. The attention mechanism amplifies the weights of (the relatively few) important tokens in the context window and decreases the weight of all others. The aim of the attention mechanism is to simulate cognitive attention.

Pre-Training and Fine-Tuning

The large pre-trained models can be used directly for many tasks, without any adaptation, a so- called zero-shot approach. Alternatively, they may be adapted to a task with a small number of additional training instances. We use a fine-tuned model in section 3.1. These approaches are called few-shot. Usually, only the weights of the last few layers of the neural network are adapted based on the annotated, task-dependent material. The main advantage of fine-tuning is that far fewer training instances are needed than when training a model from scratch. Transformer-based LLMs have such detailed world knowledge that fine-tuning only needs to specify the particular task, for example, question answering, summarization, natural language inference, stance detection, hate speech detection, language level, etc.

Shortcomings and underlying reasons

A common criticism of deep neural networks is that they are “black boxes”, methods that are too complex to understand what happens in detail. Accordingly, it is hard to anticipate in which cases these models tend to fail. Neural models including transformers are typically very reliable, but sometimes they produce arbitrary, seemingly absurd results in sparse data situations, so- called hallucinations. Traditional model evaluation, i.e. computing the accuracy or F1-score over an en-

[8] See https://en.wikipedia.org/wiki/Activation_function#Comparison_of_activation_functions for an overview of activation functions.

tire test set, does not help in this situation, since it only measures the overall performance – not the specific strengths and weaknesses of a given model. As a solution to this lack of insight where models fail behavioral testing has been suggested by Ribeiro et al. (2020). This is the motivation for our project presented in section 3.1. In addition, adding difficult and rare cases systematically, as we are doing, also reduces the risk of hallucination.

In addition to the algorithmic shortcomings, to which we alluded above, and which we evaluate in detail in the results section, we are aware that our proposed methods (described in the remainder of this section) have many further shortcomings. For instance, it can be argued that detecting hate speech only finds the symptoms but does not address the underlying questions: why do some people feel so offended, marginalized, and threatened by society that they see no other way but to resort to uttering hate speech? Will people who feel patronized by the state ("Wutbürger") feel less patronized by AI?

A partial answer could be that recognizing hate speech is a first step in addressing it. Deleting offenders' posts or banning users if abuse persists at least protects the potential victims, the targeted groups. Ideally, counter speech and talking to the identified offenders about the situation of their victims would be the next step.

2.1.2 Why is Classifying Hate Speech Difficult?

Hate speech detection is a special case of the more general task of text classification. However, there are a number of attributes that make hate speech detection a unique and challenging classification task. In this section we will highlight five of these hate speech-specific challenges and in the following section, we will showcase our new approach addressing these challenges.

The first challenge is that there is no universally agreed on definition of hate speech (Fortuna et al. 2018). There exist many datasets and classification models for hate speech, but they are often based on similar but slightly disagreeing definitions. Further, even if datasets and definitions agree, studies have shown that annotators (humans labeling texts as hate speech or not-hate speech) interpret these definitions differently, leading to inconsistent annotations that lead to contradicting signals for models trained on these data. How to handle such disagreements in the training data is an ongoing field of research (Uma et al. 2021).

Secondly, hate speech is often not expressed in simple, overt ways, like "I really hate the Jews.". Instead, hate speech often relies on implicit language, metaphors, comparisons, and references to current events, making its detection challenging, even for state-of-the-art methods.

The third challenge is common to many text classification tasks: while there are many English hate speech datasets, there is little to no hate speech data for most of the world's languages. Current research addresses this on the one hand by creating datasets in more languages (Chhabra and Vishwakarma, 2022) and on the other hand by developing methods that require less training data (Röttger et al. 2022).

Fourth, when hate speech detection is used in content moderation systems to flag or block hate speech comments, then we expect to get an explanation for why a comment was flagged or blocked. However, current hate speech classifiers are black box systems that cannot provide such an explanation.

Finally, another line of research has shown that hate speech classifiers are often prone to keyword-based misclassifications. For example, Gröndal et al. (2019) demonstrate that simply adding the word "love" to the end of a hate speech comment will mislead many hate speech classifiers. Conversely, Röttger et al. (2021) have shown that hate speech classifiers tend to classify cursing or expressions of anger as hate speech, even if no protected group is attacked in the text.

Beyond these main challenges, there are more issues being researched such as (a) how to consider the context dependence of hate speech (b) how to deal with temporal shift since the styles and contexts of hate speech evolve constantly, and (c) how to handle multi-modality since hate speech is often not only expressed via text or speech but in combination with images or memes. Our goal for this section was to give an overview of current challenges in hate speech detection research. We will now turn to our own work, proposing a new approach for hate speech detection.

2.1.3 Why is Automated Hate Speech Detection Difficult?

Not only is it challenging to clearly define hate speech; detecting it may be even more difficult. Hate speech, or in general, all expressions with a similar meaning may be expressed in a multitude of ways. At the level of words, linguists speak of synonymity and ambiguity.

So-called synonyms express very similar meanings. For instance, *astronaut* and *cosmonaut* are synonyms, or *hate* and *despise*, or *kill* and *execute*. In order to cover all expressions potentially containing hate speech, one would need a very long list of words, and a large dictionary. These examples of synonyms also show that there are hardly any full synonyms: *astronaut* points to a U.S. or European setting, while *cosmonaut* refers to Russian space programs, with all the political and military implications. *Kill* refers to any form of taking life, while *execute* has more likely a legal setting. These subtle differences may have an effect on hate speech status: while *I think all Jews should be killed* is clearly hate speech, the statement *I think all terrorists should be executed* is clearly not – discussions on the death penalty need to be possible in democracies.

While the word *execute* on the one hand has a narrow meaning when it refers to taking life, it also has further meanings that are very different – it is highly ambiguous. Think of the contrast between *I think all police orders should be executed* and *I think all police staff should be executed.* Ambiguity is a main reason why using a large dictionary of words, so-called dictionary-based approaches, do not perform very well. In the example of *execute* many, probably the majority of, utterances do not contain hate speech. In *I think all police orders should be executed* the fact that police orders are not alive (linguists use the term animate) triggers the correct reading of *execute,* and also illustrates that the status of an utterance hate speech or not depends on whether the hate speech target is animate and member of a protected group.

These examples already show that we need to know more than individual used words as a filter. We minimally need to use words in combination and their interaction, and we need to know which words are similar. Text Technology, disciplines like Computational Linguistics and applications like media content analysis offer methods addressing these requirements. A relatively simple method to use words in combination is bag-of-words classification, and word similarities can be computed by the various methods that are called word embeddings.

2.1.4 Modular Zero-Shot Hate Speech Detection

We now describe a new approach for hate speech detection, developed by us and published in Goldzycher and Schneider (2022), and Goldzycher et al. (2023). In these papers, we propose a new framework that enables modular, and interpretable zero-shot hate speech detection. It is modular because the framework allows for extensions and modifications, e.g. to make the classifier more robust or tailor it to a specific hate speech definition. It is interpretable because the final hate speech prediction is composed of many “atomic” predictions that can be inspected.

Finally, the framework is “zero-shot” as it does not require hate speech examples to train on. In this section, we describe this approach in more detail.

Natural Language Inference for Text Classification: As explained in Section 2.1.1, text classifiers need to be trained (or fine-tuned) on thousands of examples that are marked with a specific class – in our case, the classes are *hate speech* or *not-hate speech.* Yin et al. (2019) propose a new zero-shot text classification approach (zero-shot means that no training for the specific task at hand is needed)

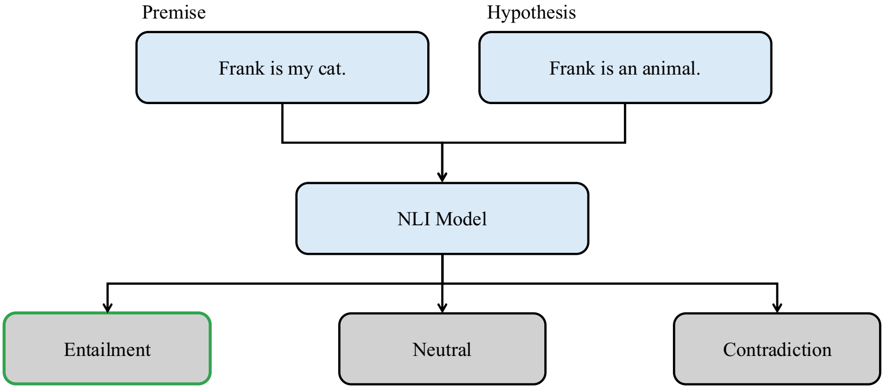

Figure 1: The inputs and outputs of a natural language inference model.

that works by repurposing natural language inference (NLI) models. An NLI model takes a premise and a hypothesis as input and produces a prediction of "entailment", "contradiction", or "neutral" as output. "Entailment", in this case, means that the hypothesis is entailed by the premise. "Contradiction" means that the hypothesis contradicts the premise. And "neutral" means that the hypothesis is neither entailed by the premise nor contradicts the premise. Figure 1 displays an example. Yin et al. (2019) propose to repurpose NLI models for text classification by inputting the text to classify as the premise and formulating a hypothesis of the form "This text is about X". We expect the NLI model to output entailment if the text is indeed about X and we expect it to output contradiction if the text is not about X. We remove the neutral option from the NLI model to force it into making a binary classification. Figure 2 displays this method of using NLI for text classification. Yin et al. (2019)'s experiments show that this approach can lead to high-performing topic classifiers.

Natural Language Inference for Hate Speech Detection: Our work (Goldzycher and Schneider, 2022) starts by applying this approach to hate speech classification. This means inputting the text to classify as the premise and formulating a hypothesis that says "This text is hate speech".

Analogously to the previous zero-shot topic classification examples, if the model predicts "entailment", we can interpret this as a prediction of *hate speech*. Conversely, we can interpret the model outputting "contradiction" as the prediction of *not-hate speech*. To evaluate how accurate this approach is we test it on a popular test suite for hate speech detection named HateCheck (Röttger et al. 2021). HateCheck contains over 3,700 sentences, either marked as *hate speech* or *not-hate speech*. These sentences are not collected from social media like most hate

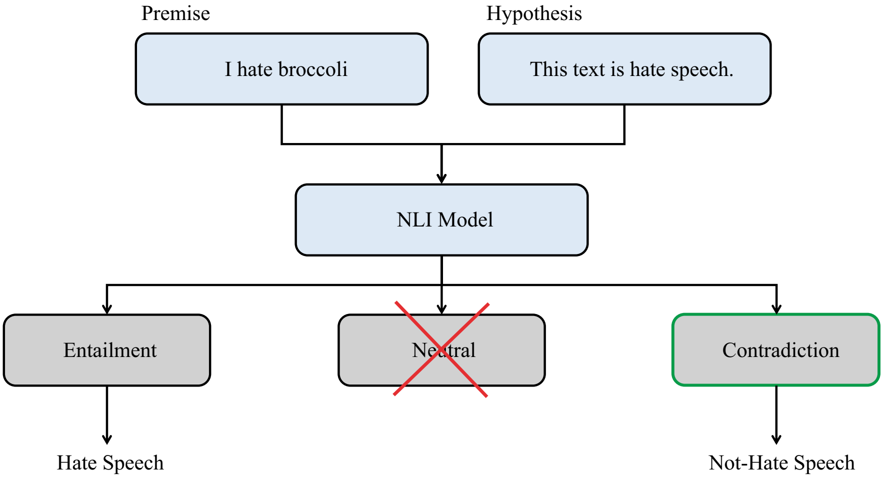

Figure 2: Using a natural language inference model for text classification.

speech datasets, but are based on manually created templates with the goal of covering common and challenging types of hate speech and especially difficult types of non-hate speech.

Using this dataset, we evaluate how well our NLI model can detect hate speech and observe an accuracy of 79.4%. This is a comparatively high accuracy that beats hate speech classifiers which have been trained on hate speech data. We now turn to the main contribution of our work, which is "Hypothesis Engineering" for hate speech detection.

Hypothesis Engineering: Text classification with an NLI model can be done in a zero-shot fashion, i.e. requires no task-specific training data, as we have just illustrated. This means that we can use an NLI model not only to predict if the input text contains hate speech, but also to predict any property that we are interested in, simply by adjusting the hypothesis. E.g. if we are interested in whether the input text contains anger, we can formulate a hypothesis "This text contains anger". This means that NLI models allow us to predict *any* properties of the input text that we might find helpful for a final decision with respect to the question of whether the input text contains hate speech. The flexibility of NLI models to predict any aspect represents a great opportunity. But how do we decide which aspects of the text we want to predict out of all the possible aspects one could predict? One response to that question, indeed the response that we explore in this paper, is that we aim to identify relevant aspects by analyzing where the current NLI hate speech detection model fails. These additional hypotheses are then aimed at mitigating the identified weaknesses.

Weaknesses: We use the HateCheck test suite to analyze the weaknesses of NLI models as zero-shot hate speech classifiers. The test suite is structured into 29 categories of hate speech and difficult-to-classify types of non-hate speech. Examples of such categories include “expression of strong negative emotions” (“I hate [PROTECTED GROUP]”), or “Denouncements of hate that make direct reference to it” (“You have to stop calling [PROTECTED GROUP] disgusting.”). When evaluating our NLI model on these categories, we identify four main clusters of weaknesses: (1) The model often predicts those texts as hate speech which contain abusive language or strong negative emotions, even if they do not contain a protected target group (a necessary condition for hate speech). (2) The model misclassifies denouncements of hate speech as actual hate speech. (3) Reclaimed slurs, e.g. “queers”, often lead to harmless texts being misclassified as hate speech. And (4), dehumanizing comparisons between groups of people and negatively associated animals, such as monkeys, pigs, rats, or “the plague” are often not correctly recognized as hate speech.

Strategies: For each weakness, we develop a strategy of how to mitigate that weakness using the NLI zero-shot text classification. All strategies are based on additional aspects being predicted by the NLI model. A strategy is a collection of hypotheses to predict these aspects and rules for how to combine these predictions with the prediction of the “main hypothesis” (“This text is hate speech.”). The strategies are called (1) Filtering by Target (FBT), (2) Filtering Counterspeech (FCS), (3) Filtering Reclaimed Slurs (FRS), and (4) Catching Dehumanizing Comparisons (CDC). We describe here the first strategy, FBT. For further strategies, we refer to the original paper (Goldzycher and Schneider, 2022).

Filtering by Target (FBT): As described, our strategy aims to mitigate false classifications as hate speech when the input text contains negative or abusive language but is not directed at a protected group. In this strategy, all protected groups and characteristics, according to a given hate speech definition are inserted into the text template “This text is about [PROTECTED GROUP]”. This leads to a list of hypotheses such as “This text is about black people”, “This text is about Jews”, “This text is about Muslims” etc. We collected the NLI model predictions for all hypotheses. Finally, we implement the rule that if the model predicts that none of these target group hypotheses is applicable, then a possible prediction of hate speech is corrected to not-hate speech. The process is sketched in Figure 1.

Extending the Zero-Shot Paradigm to More Languages: In Goldzycher et al. (2023), we extend this approach to a multilingual setting and evaluate, among other experiments, whether hypothesis engineering is beneficial for NLI-based hate speech detection in other languages when either no or few training examples are available.

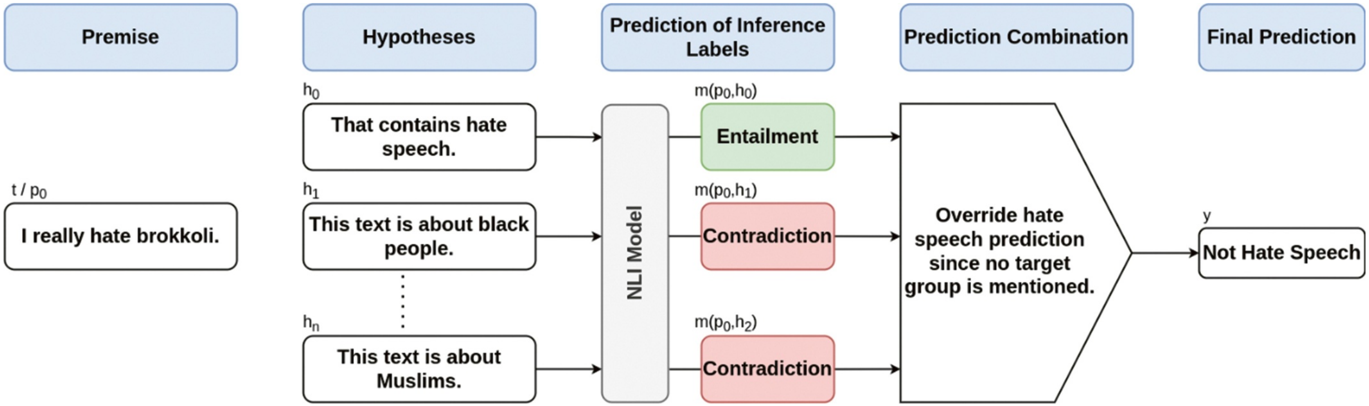

Figure 3: An overview of the “Filtering by Target” strategy. Image taken from Goldzycher and Schneider (2022).

We describe the evaluation of the hypothesis engineering and its extension to more languages, in Section 3.1.

2.2 Conceptual Maps

We now present the methods used to address the second of our research questions: how much hate and negative sentiment is present in religious discourses on social media? For this, we investigate frequently used words in the context of religion, spirituality and faith from social media, and create maps of frequently co-occurring words. We also introduce our motivation and data.

2.2.1 From Word Profiles to Conceptual Maps

Firth (1957) noticed that words which frequently co-occur, or which occur in similar contexts, are similar or related in meaning (section 2.1.1). Conceptual Maps use even more context than word embeddings. Words that occur together in contexts can be counted in word-word co-occurrence matrices, and the visualised in such a way that strongly connected words appear close together, and rarely connected words far away. Advantages of this method are that one can obtain an overview of a dataset at a glance while also allowing exploration of words that appear between clusters – and e.g. serving as bridges – or those that are important across the entire collection.

A disadvantage of the approach is that it is word-centred, it cannot profit from word embedding. Therefore, a statistical smoothing method is applied, namely Kernel Density Estimation (KDE, Zucchini 2003). The smoothing process makes sure that the method also works with relatively small datasets. We use KDE to create maps, which we call *conceptual maps*, but there is no broadly accepted established term, as the method is used only rarely and has significant potential. Some of the rare uses are Kaufmann (2020) and Eve (2022). The created maps are like *mind maps* (Buzan and Buzan 1993), although these are typically drawn manually. We argue that the method of creating term-term matrices with KDE and then plotting similar words together is an automated method to draw maps that are similar to mind maps.

Kernel Density Estimates are functions of these mutual co-occurrences which are learned from corpus data. Then, visualisations of similarities of words are done in the form of such networks, which we call conceptual maps. In order to obtain smoothed results, approximating functions are used which gloss over data fluctuations. We use the Python library *textplot* (McClure 2015) to calculate the term-term co-occurrence matrices and the spring attraction algorithm *ForceAtlas2* (Jacomy et al. 2014) in *gephi* to create two-dimensional maps.[9]

2.2.2 Theoretical Background

Religion is a frequent target of hate speech, discrimination, misunderstandings and intolerant ideology. Political oppression in the name of religion is a reality in many areas of the world. Accordingly, we need to expect an important amount of hateful utterances and very negative attitudes. We extend our study beyond hate speech to generally assess negative and positive sentiment and attitudes on religious topics, as they are mirrored in social media.

Neubert (2016) suggests a radically discursive, i.e. data-driven, definition of religion. The assumption that words are concepts and are defined by their use has already been suggested by Wittgenstein (1953), in this sense the Firthian hypothesis can be seen as a consistent operationalisation. Neubert (2016) argues that a definition of religion needs to include the social discourse.

> Die Suche nach einer wissenschaftlichen Definition geht jedoch weiter und stößt immer von Neuem auf die Frage, wie Religionsverständnisse in breitere soziale Diskurskontexte eingebunden sind. (Neubert 2016: 15)
>
> The search for a scientific definition, however, needs to go further and keeps encountering the question of how conceptions of religion should be embedded into larger discursive concepts. (Neubert 2016: 15, our translation)

Such a discursive understanding, Neubert (2016) argues, needs to be grounded in what religion means to people on an everyday basis, how they personally connect to it.

> Da empirisch also am Ausgangspunkt religionswissenschaftlicher Theoriebildung dieses Alltagsverständnis liegt, könnte man dieses bei konsequenter Theoretisierung auch explizit zum Ausgangspunkt machen.
>
> Because this everyday conception of religion is the starting point for scientific theory formation in religious science, it could also be explicitly made its initial point, following a consistent theoretisation. (Neubert 2016: 15, our translation)

We suggest to use social media as a proxy to such an everyday understanding, as a glimpse into the worries, discoveries, inspiration, awe, and prayers of a frag-

[9] https://www.gephi.org.

ment of society. We are aware that such a selection can hardly be representative of the population as a whole, but that we are rather shown an aggregation of the loudest voices on the market square.

An empirical bottom-up definition crucially depends on the data, accordingly we do not expect to see a comprehensive definition, we also assume that the results would be considerably different, if different texts, for example the scriptures and scholastic texts, were used.

Given large amounts of representative data, the method can be seen as an operationalisation not only of Neubert's (2016) idea, but more generally of de Saussure's linguistic concept, in which words are not defined by what they are, but in counter distinction to other similar words, until they form a system of pointers in the sense of Derrida's différance (Derrida 1968), where the ultimate grounding is constantly deferred.

The personal experience, which seems to be a central difference between religion and spirituality, is also often referred to by the term of *faith*. *Faith* is sometimes seen as a subset of spirituality and mainly a Christian concept, but other scholars give it the central position. In their discussion of the differences between faith, religion and spirituality, Paul Victor and Treschuk (2020) write that "Faith is more personal, subjective, and deeper than organized religion and relates to the relationship with God", and how individuals are touched by God and their religious experiences. It is also often described as psychological human universal, e.g. "Faith, from a more naturalistic, psychological perspective, is merely the innate drive to search for meaning, purpose and significance." (Popcak and Popcak 2014).

Newman (2004) sees *faith* as the underlying force behind both religion and spirituality.

> [I]n my model, spirituality and religion are a function of faith. Both religion and spirituality require faith as a foundation (. . .). In other words, faith is the guiding principle by which individuals are either religious or spiritual. Faith serves as both the source and the target of their religion or spirituality. Devotion to religion or perception of growth in spirituality may be seen as a measure of greater valence of understanding one's faith. (Newman 2004, p. 106)

To explore the question of the role of *faith*, we added it to the picture as a third term, allowing us to see the overlap in meaning between *faith* and *spirituality*, or to observe the differences.

For the conceptual map that we present in section 3.2, we use a collection of over 100,000 tweets that we collected in Spring 2021. We queried for the terms *religion*, *spirituality*, and *faith*. 51,277 tweets contain the term *religion*, 36,733 contain *spirituality*, and 31,145 contain *faith*. We compare spirituality and religion in order to verify the hypothesis that spirituality is seen as much more positive than religion, which is connected to negativity, hierarchy and scandals. This hypothesis is expressed by Neubert (2016) as follows:

> ‘Spiritualität’: in vielen Kontexten ‘Religion’ positiv gegenübergestellt wird – entweder als positive Überhöhung und ‘wahres Wesen’ von ‘Religion’ oder aber als in positiver Weise individualistisch, innerlich und heilsorientiert, wogegen ‘Religion’ negativ mit Organisation, Hierarchie, Dogmatismus und Weltlichkeit verknüpft wird. (Neubert 2016: 127)

> ‘Spirituality’ is often juxtaposed to religion as positive – be it as positive idealisation or the true essence of religion, or as positively individualistic, intimate and healing, whereas ‘religion’ is negatively connected with organisation, hierarchy, dogmatism and profanity. (Neubert 2016: 127, our translation)

In order to measure positive and negative sentiment, we have also included an automatic sentiment detection tool (Hartmann et al. 2022), which annotates every tweet with the pseudoword *sentinegative* and *sentipositive* respectively.

3 Main Findings

3.1 Hate Speech Detection

We now apply the methods given in section 2.1 to hate speech detection.

Hypothesis Engineering Evaluation: For the evaluation of the NLI approach for hate speech detection, i.e. hypothesis engineering, we use a BART model (Lewis et al. 2020) fine-tuned on the MNLI dataset (Williams et al. 2018). It is available in the huggingface transformers library (Wolf et al. 2020) under the name “bart-large-mnli”.

Table 1: Evaluation of hypothesis engineering strategies on HateCheck.

Strategy	Accuracy	Δ (percentage points
Without Strategy	79.4	0.0
Filtering by Target	82.7	+3.3
Filtering Counterspeech	84.0	+4.6
Filtering Reclaimed Slurs	80.1	+0.7
Catching Dehumanizing Comparisons	79.6	+0.2
Overall	87.3	+7.9

We evaluate the approach on two datasets: HateCheck and ETHOS (Mollas et al. 2022). The ETHOS consists of real social media posts collected from Youtube and Reddit. We use a part of the dataset that has binary annotations (hate speech vs.

Table 2: Evaluation of hypothesis engineering strategies on ETHOS.

Strategy	Accuracy	Δ (percentage points
Without Strategy	69.6	0.0
Filtering by Target	78.7	+9.1
Filtering Counterspeech	69.6	+0.0
Filtering Reclaimed Slurs	71.3	+1.7
Catching Dehumanizing Comparisons	69.5	−0.1
Overall	79.6	+10.0

non-hate speech), following a hate speech definition that is close to the definition in HateCheck, and contains 997 texts. In contrast to HateCheck, which informed our hypothesis engineering, i.e. the proposed strategies, ETHOS is new to this setup allowing us to measure the generalization capabilities of the proposed approach. Table 1 shows the evaluation results on HateCheck. First, we observe that the overall accuracy on HateCheck increases from 79.4% to 87.3%, an increase of 7.9 percentage points. Second, we see that each individual strategy increases the accuracy. And third, the table shows that the strategies Filtering by Target and Filtering Counterspeech lead to the largest increases in accuracy. Moving on to the results on the ETHOS dataset in Table X, we observe a similar improvement in accuracy from 69.6% to 79.6% – an increase of 10.0 percentage points. However, for ETHOS we see that two strategies ("Filtering Counterspeech" and "Catching Dehumanizing Comparisons") do not improve the accuracy at all, and that most of the improvement comes from the strategy "Filtering by Target". We highlight two findings from these results: (1) Hypothesis engineering can improve the accuracy of NLI-based zero-shot hate speech classifiers making them as accurate as fine-tuned classifiers, sometimes even surpassing them. (2) Hypothesis engineering for hate speech detection is not only useful on the domain that informed the hypothesis engineering, but it also generalizes to other domains, demonstrated by the fact that the results did not just improve on HateCheck but also on ETHOS. (3) The usefulness of individual hypothesis engineering strategies depends on the domain, platform, or genre of text to classify – as seen in the different results on HateCheck and ETHOS. Thus, having a range of strategies can make the classifier more robust to different types of input text.

Extending the Hypothesis Engineering to More Languages: To test the hypothesis engineering strategies in more languages, we need a multilingual NLI model and datasets to test on, in multiple languages. We focus on five languages for which there exist the necessary NLI training data and the appropriate hate speech datasets. These languages are Arabic, Hindi, Italian, Portuguese, and Span-

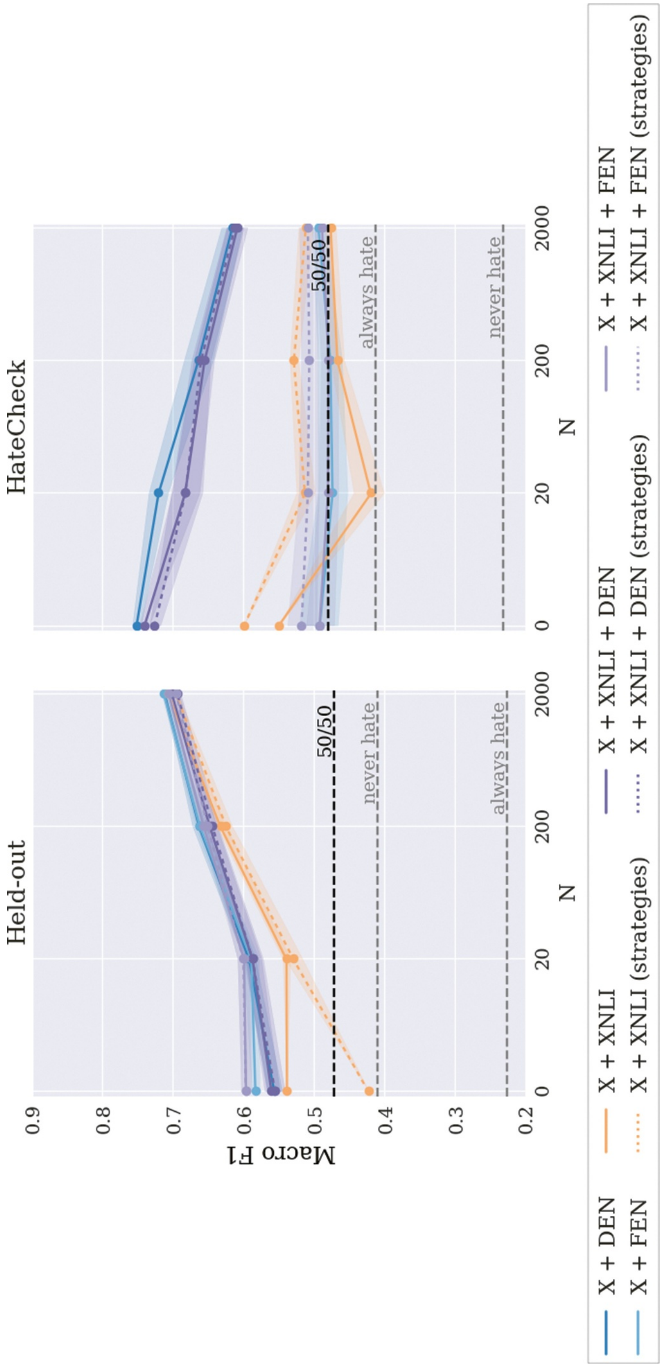

Figure 4: Results of hypothesis engineering (strategies) in five more languages.

ish. To get an NLI model for all those languages we train a RoBERTa-based model (Barbieri et al. 2022) on the multilingual natural language inference (MNLI) dataset (Williams et al. 2018). We then either directly perform hypothesis engineering in the target language or train further on 20, 200, or 2000 examples in the target language. The averaged results over all languages are displayed in Figure 4.

The left part of the plot named "Held-out" shows the average performance on Twitter datasets in the target languages and the plot on the right side shows the average performance on multilingual versions of the HateCheck dataset (Röttger et al. 2022). The plot uses a macro F_1 score (higher is better). The "N" in the X-axis corresponds to the number of training examples in the target language. The shaded areas denote bootstrapped 95% confidence intervals based on averaging ten models per setting. The different colors correspond to different training training datasets, where "X" refers to the multilingual base model, "XNLI" to multilingual NLI dataset (Williams et al. 2018). "DEN" and "FEN" refer to different hate speech datasets. Solid lines denote the results without hypothesis engineering and dotted lines in the same color denote results when the hypothesis engineering has been applied additionally. Thus, to analyze the effect of hypothesis engineering we need to compare the dotted with the solid lines in each color. We observe that on average, over the five languages, hypothesis engineering leads to mixed results – in contrast to the only positive results in English. While there are no positive effects on the Twitter datasets, we see improvements from hypothesis engineering in half of the settings on HateCheck. Overall, these results show that the impact of hypothesis engineering in multilingual setups depends on the language, application domain, and availability of training data in the target language.

3.2 A Map of Associations

We now turn to conceptual maps (see section 2.2 for the method) to explore associations and the presence of hate in the online religious debate. The data-driven, radically discursive conceptual map of the juxtaposition between religion, spirituality and faith is given in Figure 5. *Faith* and its co-occurring words are shown at the top, *spirituality* on the left and religion on the *right*. *Faith* appears close to *hope*, *god*, *Jesus* and the *word* of the *Bible*. The Bible connects *faith* to *religion*, which in turn is associated with abstract concepts like the institution of the *church*, *state*, but also calls for *freedom* and *respect*. In addition to mentioning some of the world's important religions including *Islam*, *Hinduism* and *Christianity*, *politics* highlights the conflicts involving religion, and these are often ignited with reference to different beliefs, such as *hate* and *race*, but also *gender* discrimination, which we can also spot here. The negations (*doesn*, *don*, *isn*) are due to complaints,

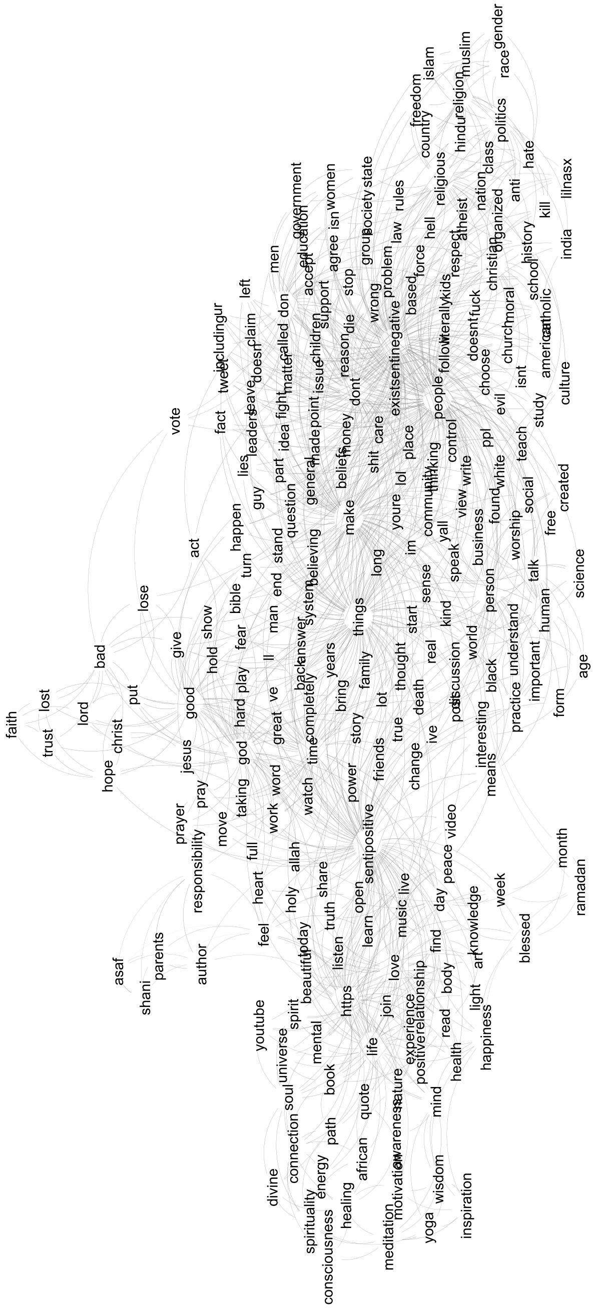

Figure 5: Conceptual map of spirituality, religion, and faith with the 250 most frequent words.

e.g. that (other) religions *don't respect* our values, or warnings, namely that *hate* doesn't solve any problem. Also, the positions of the pseudowords *sentinegative* and *sentipositive* confirm Neubert's (2016) hypothesis. Expressions of hate and strong negativity for religion is in fact stronger than we had expected.

4 Conclusion

Hate Speech is a harmful phenomenon and cannot be detected manually, due to its sheer volume on the internet. We have addressed our first research question of how we can detect hate speech with state-of-the-art methods by presenting computational linguistic algorithms that we have applied and improved. We summarized our research proposed in previous papers, in which we propose a new framework for hate speech detection that addresses four drawbacks of previous approaches, including adaptability to specific hate speech definitions, explainability of the final decision, and requiring less hate speech-specific training data than previous approaches. Our experiments show that these advantages can be achieved while upholding classifier accuracy. When applied to analyzing data on the internet, our framework enables a more fine-grained analysis of hate speech, such as detecting what types of attacks are prevalent on social media, which religions are targeted how often, and how the contexts, expressions and types of hate differ by religion.

In our second research question, we have addressed how much hate and negative sentiment is present in the religious discourse, in data-driven fashion. Conceptual maps are a data-driven method that learns associated words from the contexts, in the spirit of Firth (1957). We have used the method to obtain an overview map of the terms *religion*, *spirituality* and *faith*. We observed that *religion* is highly politicized, and frequently associated with *hate*, *discrimination*, polarization and the power of the church and other forms of institutionalization) as institution.

5 Future Research

5.1 Hate Speech Detection

In future work, we aim to enhance the accuracy of our approach, specifically in non-English languages. An obvious opportunity lies in using new, more competent large language models as the backbone model for hypothesis engineering. Further, since hate speech detection based on hypothesis engineering relies on multiple predictions, such as if a text is about a specific group, or if it contains a comparison to animals, we can use these intermediate predictions for a fine-grained analysis. For instance, we can examine who is targeted how frequently, what types of attacks are how prevalent and how specific groups (for example religious affiliations) are targeted with particular expressions, associations, or tropes. We envision this toolset being used as the foundation of a cycle of quantitative and qualitative analysis: For example, a qualitative deep dive into posts classified as using animal comparisons for Muslims can reveal underlying narratives, tropes, and contextual factors of these attacks. These insights, in turn inform improved hypothesis engineering strategies leading more accurate and granular quantitative analyses feeding a continuously improving cycle. The results of such an iterative approach can provide a much more nuanced understanding of hate online with broad potential implications for many fields including social sciences, religious studies and debates about free speech vs. content moderation.

5.2 Conceptual Maps

Our current selection was motivated by the question of which religious concepts attract hate and hate speech. We will do a more detailed analysis of which groups and issues are most contested and most affected by negative stance and hate. We will use automated versions of close reading harnessing the semantic detail of LLMs to detect these and their implications for democratic discourse, inclusivity and security of minorities, which have motivated the current study. We also aim to include further terms that are relevant in the discussion of religion and society, and also further datasets, for example scientific religious research papers. We will also use further methods to create semantic maps in addition to KDE, for example t-SNE (van der Maaten and Hinto 2008). In addition, we will investigate if Musks's and Trump's utterances further spurn hate, e.g. which features of right-wing populism and targeted hate (Wodak 2015) appear. Religion is often used as a pretext to exclude segments of society, we hope to shed further light on these questions by investigating associations and stance.

Bibliography

Barbieri, Francesco, Espinosa Anke, Luis, and Camacho-Collados, Jose. 2022. "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond". Proceedings of the Thirteenth Language Resources and Evaluation Conference, 258–266. https://aclanthology.org/2022.lrec-1.27.

Barendt, Eric. 2019. "What is the Harm of Hate Speech?" *Ethical Theory and Moral Practice*, 22(3), 539–553. http://www.jstor.org/stable/45217319.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 238–247, Baltimore, Maryland: Association for Computational Linguistics. http://www.aclweb.org/anthology/P14-1023.

Baroni, Marco and Alessandro Lenci. 2010. "Distributional Memory: A General Framework for Corpus-Based Semantics." *Computational Linguistics*, 36(4), 673–721.

Buzan, Tony and Barry Buzan. 1993. *The Mind Map Book: How to Use the Radiant Thinking to Maximize Your Brain's Untapped Potential*. London: Penguin.

Chhabra, Anusha and Dinesh Kumar Vishwakarma. 2023. "A Literature Survey on Multimodal and Multilingual Automatic Hate Speech Identification." *Multimedia Systems* 29, 1203–1230. https://doi.org/10.1007/s00530-023-01051-8.

Derrida, Jacques. 1968. *La "différance"*. Société Française de Philosophie, Bulletin 62(3), 73.

Eve, Martin Paul. 2022. *The Digital Humanities and Literary Studies*. Oxford: Oxford University Press.

Firth, John Rupert. 1957. "A Synopsis of Linguistic Theory 1930-1955." *Studies in Linguistic Analysis*, 1–32.

Fortuna, Paula and Sérgio Nunes. 2018. "A Survey on Automatic Detection of Hate Speech in Text." *ACM Computing Surveys*, 51(4), 1–30. https://doi.org/10.1145/3232676.

Goldzycher, Janis and Gerold Schneider. 2022. "Hypothesis Engineering for Zero-Shot Hate Speech Detection." In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying* (TRAC 2022), 75–90, Gyeongju: Association for Computational Linguistics.

Goldzycher, Janis, Moritz Preisig, Chantal Amrhein, and Gerold Schneider. 2023. „Evaluating the Effectiveness of Natural Language Inference for Hate Speech Detection in Languages with Limited Labeled Data." In *The 7th Workshop on Online Abuse and Harms (WOAH)*, 187–201. Toronto: Association for Computational Linguistics.

Graham, Timothy, and Mark Andrejevic. 2024. *A Computational Analysis of Potential Algorithmic Bias on Platform X during the 2024 US Election*. Technical report. https://eprints.qut.edu.au/253211/.

Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*, 21(3), 267–297.

Gröndahl, Tommi, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. "All You Need is 'Love': Evading Hate Speech Detection." In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (AISec'18). New York: Association for Computing Machinery, 2–12. https://doi.org/10.1145/3270101.3270103.

Hartmann, Jochen, Mark Heitmann, Christian Siebert, and Christina Schamp. 2022. *More Than a Feeling: Accuracy and Application of Sentiment Analysis*. Rochester: Social Science Research Network. https://doi.org/10.2139/ssrn.3489963.

Hietanen, Mika and Johan Eddebo. 2023. "Towards a Definition of Hate Speech – With a Focus on Online Contexts." *Journal of Communication Inquiry*, 47(4), 440–458. https://doi.org/10.1177/01968599221124309.

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. "ForceAtlas2, A Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software". *PLOS ONE* 9(6), e98679. https://doi.org/10.1371/journal.pone.0098679.

Jurafsky, Dan and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd Edition. Upper Saddle River: Pearson Prentice Hall.

Kaufman, Micki. 2020. "Everything on Paper Will Be Used Against Me." *Quantifying Kissinger*. http://blog.quantifyingkissinger.com.

Khurana, Urja, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. “Hate Speech Criteria: A Modular Approach to Task-Specific Hate Speech Definitions.” In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 176–191, Seattle: Association for Computational Linguistics.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Manning, Christopher D. and Hinrich Schütze. 2001. *Foundations of Statistical Natural Language Processing*. MIT Press.

McClure, David. 2015. *Textplot*. https://github.com/davidmcclure/textplot.

Mollas, Ioannis, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. “ETHOS: A Multi-Label Hate Speech Detection Dataset.” In *Complex & Intelligent Systems* 8, 4663–4678. https://doi.org/10.1007/s40747-021-00608-2.

Neubert, Frank. 2016. *Die diskursive Konstitution von Religion*. Berlin: Springer.

Newman, Leanne Lewis. 2004. “Faith, Spirituality, and Religion: A Model for Understanding the Differences.” *College Student Affairs Journal*, v23 n2, 102–110. https://eric.ed.gov/?id=EJ956981.

Paul Victor,Chitra and Judith V. Treschuk. 2020. “Critical Literature Review on the Definition Clarity of the Concept of Faith, Religion, and Spirituality.” *Journal of Holistic Nursing* 38(1), 107–113. https://journals.sagepub.com/doi/full/10.1177/0898010119895368.

Pluta, Agnieszka, Joanna Mazurek, Jakub Wojciechowski, Tomasz Wolak, Wiktor Soral, and Michal Bilewicz. 2023. “Exposure to Hate Speech Deteriorates Neurocognitive Mechanisms of the Ability to Understand Others’ Pain.” *Scientific Reports*, 13(1): art. no. 4127.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. “Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review.” *Language Resources and Evaluation* 55, 477–523.

Popcak, Rachel and Gregory Popcak. 2014. “Faith, Spirituality, Belief, Religion . . . What’s the Difference?” *Faith on the Couch blogs*. https://www.patheos.com/blogs/faithonthecouch/2014/05/faith-spirituality-belief-religion-whats-the-difference/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. “Beyond Accuracy: Behavioral Testing of NLP Models with CheckList.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.442.

Röttger, Paul, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. “HateCheck: Functional Tests for Hate Speech Detection Models.” In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 41–58. Association for Computational Linguistics.

Röttger, Paul, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. “Data-Efficient Strategies for Expanding Hate Speech Detection into Under-Resourced Languages.” In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5674–5691, Abu Dhabi: Association for Computational Linguistics.

Röttger, Paul, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. “Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models.” In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 154–169, Seattle: Association for Computational Linguistics.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Doctoral Thesis, University of Stockholm.

Uma, Alexandra N., Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. "Learning from Disagreement: A Survey." *Journal of Artificial Intelligence Research* 72(2021), 1385–1470.

Yin, Wenpeng, Jamaal Hay, and Dan Roth. 2019. "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), 3914–3923, Hong Kong: Association for Computational Linguistics.

Van der Maaten Laurens and Geoffrey Hinto. 2008. "Visualizing Data Using t-SNE". *Journal of Machine Learning Research*, 9, 2579–2605.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All You Need". In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 6000–6010. Curran Associates Inc., Red Hook, NY. https://doi.org/10.48550/arXiv.1706.03762.

Williams, Adina, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Williams, Matthew L., Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2019. "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime." *The British Journal of Criminology*, 60(1), 93–117. https://doi.org/10.1093/bjc/azz049.

Winiger, Fabian, Gerold Schneider, Janis Goldzycher, David Neuhold, and Simon Peng-Keller. Accepted for publication. "The 'Spiritual' and the 'Religious' in the Twittersphere: A Topic Model and Semantic Map." *Journal of Religion, Media and Digital Culture*.

Wittgenstein, Ludwig. 1953. Philosophical Investigations, ed. G. E. M. Anscombe and R. Rhees, trans. on facing pages by G. E. M. Anscombe. Oxford: Blackwell. 2nd edn 1958; revised edn 2001.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander Rush. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Association for Computational Linguistics.

Zucchini, Walter. 2003. *Applied Smoothing Techniques – Part 1: Kernel Density Estimation*. Unpublished Manuscript. http://staff.ustc.edu.cn/~zwp/teach/Math-Stat/kernel.pdf.