FASCIST-O-METER: Classifier for Neo-fascist Discourse Online

Rudy Alexandro Garrido Veliz, Martin Semmann,
Chris Biemann, Seid Muhie Yimam

Universtät Hamburg
Correspondence: Rudy Alexandro Garrido Veliz, email: rudy.garrido.veliz@uni-hamburg.de

Abstract

Neo-fascism is a political and societal ideology that has been having remarkable growth in the last decade in the United States of America (USA), as well as in other Western societies. It poses a grave danger to democracy and the minorities it targets, and it requires active actions against it to avoid escalation. This work presents the first-of-its-kind neo-fascist annotation scheme for digital discourse in the USA societal context, overseen by political science researchers. Our work bridges the gap between Natural Language Processing (NLP) and political science against this phenomena. Furthermore, to test the annotation scheme, we collect a tremendous amount of activity on the internet from notable neo-fascist groups (the forums of Iron March and Stormfront.org), and the guidelines are applied to a subset of the collected posts. Through crowdsourcing, we annotate a total of a thousand posts that are labeled as neo-fascist or non-neo-fascist. With this labeled data set, we fine-tune and test both Small Language Models (SLMs) and Large Language Models (LLMs), obtaining the very first classification models for neo-fascist discourse. The best performing SLM was HomophobiaBert with an F1 of 0.8387, and the best LLM was GPT-4o on few-shot with an F1 of 0.8680.

We find that the prevalence of neo-fascist rhetoric in this kind of forum is ever-present, making them a good target for future research. The societal context is a key consideration for neo-fascist speech when conducting NLP research. Finally, the work against this kind of political movement must be pressed upon and continued for the well-being of a democratic society.

Disclaimer: This study focuses on detecting neo-fascist content in text, similar to other hate speech analyses, without labeling individuals or organizations.

1 Introduction

Neo-fascism poses a grave danger to democracy (Cento Bull, 2010; Cammaerts, 2020; Haro and Coles, 2017) and, like its predecessor (fascism), a great risk to the communities it targets. As with any extremist ideology, it is only a matter of time before it escalates to uncontrollable proportions (Hollewell and Longpré, 2022; Winter et al., 2020; McCurdy, 2021). Part of its agenda is posing social inequalities as desirable and even needed, appealing to the hatred and predisposition that a community may already have (Cammaerts, 2020).

Neo-fascism weaponizes minorities for the goal of maintaining the capitalist social status (Cox and Skidmore-Hess, 2022). It is necessary to fight neo-fascism in an active way (Haro and Coles, 2017), as its familiarity and online presence can normalize the ideology and lead to real-life escalation (Koster and Houtman, 2008). Given that the radicalization of this movement has a strong presence in online independent forums, this work will leverage NLP tools to build a basis for fighting digital neo-fascism. We begin with gathering raw data from known neo-fascist forums. Next, annotation guidelines are developed with the aid from researchers in this political field, helping with the complexity of the neo-fascist ideology. We then used crowdsourcing to manually annotate a subset of the data, determining whether it contained neo-fascist ideology. This annotated dataset is used to train a classifier using both SLMs and LLMs to detect and classify neo-fascist discourse. Figure 1 illustrates the comprehensive process involved in developing the **FASCIST-O-METER** classifier.

The main research questions driving this study are: **RQ1**: How can NLP tools be utilized to identify neo-fascist discourse online? **RQ2**: What challenges arise during the annotation of neo-fascist content, and how can these challenges be addressed? **RQ3**: How do language models fare

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 138–153, 2025

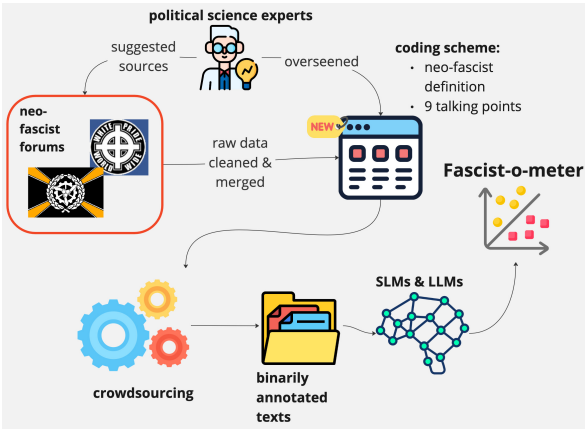

Figure 1: Illustration of the complete end-to-end process for developing the **FASCIST-O-METER** system.

when classifying such content, and in what way can they be optimized for this objective? In this paper, we mainly contributed by creating an annotated dataset of neo-fascist and non-neo-fascist online discourse, derived from extensive data collection and a carefully crafted annotation scheme. We developed and validated methodologies for building and fine-tuning language models to accurately classify neo-fascist content. Finally, we provide insights into the unique challenges posed by neo-fascist discourse and offer a framework for future research and technological development aimed at mitigating its spread online.

In Section 2, the background knowledge of the political science definition of neo-fascism is covered, as well as its history in the USA and the two relevant entities of the movement for our work. The following Section 3, explains the social efforts against neo-fascism and the efforts of the NLP community aiding the fight towards extremism and political bias. Sections 4 and 5, explain the process of data collection and the crucial development of the annotation guidelines and their usage on a subset of the collected data. With the annotated dataset, Section 6 contains the various experiments done with it, putting the result of the annotation guidelines to the test. Finally, Section 7, exposes the conclusion from our work with some recommendations.

2 Background Knowledge

In this section, the general knowledge about political science and the societal context of neo-fascism will be explained.

2.1 Neo-fascism

Neo-fascism is an extremist ideology placed on the far right of the political spectrum. There has been a rise in neo-fascism in the last decades (Cox and Skidmore-Hess, 2022). The definition of neo-fascism is complicated and can differ depending on which society appears, as well as the school of thought. Some defined it as a right-wing ideology emerging from a crisis state resulting from decades of neo-liberal capitalism (Cox and Skidmore-Hess, 2022). Others believe it is a capitalist political regime heavily based in "neo-liberal shock politics" where Donald Trump is a pioneer (Haro and Coles, 2017). In the historical context, it can be seen as an ideology held by political parties and groups after the Second World War, who were inspired despite the Nazi fall to continue their legacy (Cento Bull, 2010).

In general, it is defined mainly as a combination of multiple far-right ideologies that politicians use to gain popularity and power while radicalizing their population (Cammaerts, 2020), against a minority. These ideologies will be expanded next.

- **Ultranationalism**: nationalism is the representation of ethnic groups at a state level (Eriksen, 1992). Ultranationalism is the exacerbation of faulty nationalism, causing animosity against foreigners and underrepresented minorities. It has been on the rise for the last decades in several parts of the world, among them the USA (Cammaerts, 2020).

- **Mystification and glorification of the past**: In this context, to mystify is to create fake beliefs or ideas about the past of the state. Regardless of how the past of a nation was, the neo-fascist agenda is to change it into a grandiose one (Carnut, 2022). Changing the past of a country by the right-wing in politics is often through education as a weapon (Means and Ida, 2022; Carnut, 2022). Movements like Trumpism, with the motto "Make America Great Again", are exemplary of this. Through education, the population has their identity, closely related to the new fictitious image of the country (Means and Ida, 2022).

- **Minorities weaponization**: Creating an enemy within the nation, a constant threat to which the politicians can claim urgency and a lack of power to stop them (Stanley, 2018). This created enemy is often a minority within

the population or a fictitious one from outside, immigrants, or the population from another country. The alienated population is used in the likes of a novel obstacle that explains the decline of the land and the struggles that the population face daily. To create and maintain this image and sense of gravity, the neo-fascists will weaponize these minorities in multiple ways. One of them is siding with a minority to a political party, including claiming that they are in control of said party (Cammaerts, 2020), or blaming the current government for using its institutions and funding to help the minority (Carnut, 2022; Cammaerts, 2020). They can also indicate that social inequalities between the population are natural and necessary for the majority to impose them (Stanley, 2018). Neo-fascists also claim that the minority is unable to follow the laws of the country, not fit to rule or hold power (Stanley, 2018). In some cases, the masculine counterpart of the minority is negatively hypersexualized (Stanley, 2018) i.e. tag them as sexual degenerates and as rapists.

- **Militarism and militarization of civil security**: Militarism is defined as the preparation, legitimation, and normalization of war (Stavrianakis and Stern, 2017). In neo-fascism, it is used as a tool to endorse the crisis of the state. In some cases, this militarization involves civilians resulting in paramilitary entities, a pillar of previous iterations of fascism (Copsey, 2020). An example is the "Proud Boys" militia group, which had a large involvement in the January 6 insurrection in the USA (Dowless, 2022).

Neo-fascism in the US can be partly traced back to the post-Second World War, when Nazis and fascists were welcomed by the CIA in the fight against communism (Cento Bull, 2010). But in the last decade, neo-fascism has had a huge spike in the *acceptable* political discourse. Immensely due to the presidential runs and term of Donald Trump, the way of politics has given a name of its own: "Trumpism". In this type of politics, the leaders are often blatantly racist, xenophobic, misogynistic, and much more. This behavior is disguised under the wings of free speech and "speaking his mind", considering it triumphant. The platform of Donald Trump was often a model example of neo-fascism and is broadly considered as such in political science research (Caiani and Parenti, 2009; Copsey, 2020; Stanley, 2018; Cammaerts, 2020).

2.2 Online neo-fascism and the United States of America

Many scholars agree that white nationalist extremism has become one of the more grounded threats online, specifically for the USA (Winter et al., 2020). Neo-fascists exist in social media, but they have proven to be too volatile and chaotic to stay there freely (McCurdy, 2021). Although recent developments in politics have changed that and it is worth mentioning that they have a large presence there and in other general-purpose forums (Koehler, 2014; Hollewell and Longpré, 2022). The following are online organizations that exhibited online extremism and neo-fascism within their own **independent platforms**, and are relevant to this paper.

2.2.1 Iron March

Iron March was a forum that ran between 2011 and 2017. This forum was characterized mainly by far-right extremist content and was used as a tool for planning and recruiting radicalized individuals, as well as doing their radicalization. It can be linked to more than 100 hate crimes in the USA (Reilly and Edwards, 2021; Hayden, 2019), contained extremist tendencies, and was declared as fascist (Scrivens et al., 2023) by the USA hate-watch organization Southern Poverty Law Center (SPLC). This forum posed such a threat and became so efficient with the neo-fascist agenda that it led to what is now known as the “Skull Mask” neo-fascist network. This is an international neo-fascist network that includes multiple terrorist groups spanning from North America to Europe. This network included the USA terrorist domestic group Atomwaffen (Upchurch, 2021), also tagged as neo-Nazi by the SPLC and many of their members have been prosecuted and jailed.

In Iron March, the users would speak freely about neo-fascism and idealize a world of fascist governments, mostly for the "white countries". They would share information from other social centers and would organize methodically into smaller and more private groups. The forum members had mixed opinions on Donald Trump, but generally, there was support and eagerness on that way of politics (Upchurch, 2021). The communications here would include enormous amounts of hate speech, like racism, white supremacy, xenophobia, homophobia, anti-semitism, etc (Hayden, 2019; Upchurch, 2021). The website was taken

down without an apparent reason, shortly after their database of messages and posts was leaked (Hayden, 2019).

2.2.2 Stormfront.org

Stormfront.org was one of the first online forums for the community of individuals who identify as white-nationalists. It was first opened to the public in March 1995 (Bowman-Grieve, 2009). They have threads and sub-forums for multiple far-right beliefs, including neo-fascism. Much like the aforementioned Iron March, most of the discussions are filled with hate speech. The website has general everyday content as well, always under the light of white supremacy (Bowman-Grieve, 2009).

Stormfront.org is a model example of how radicalization can happen online and how extremist communities can engulf the attention of the user, creating their identity around the ideology. It also shows how these virtual communities can be easily taken into real-life violence and hate crimes (Koster and Houtman, 2008). The Stormfront.org community also had a mixed set of opinions regarding Donald Trump, specifically due to having a Jewish daughter and the closeness to Jewish people (Dentice, 2018). This forum continues to exist to this date, freely accessible on the Internet. It has an overwhelming amount of users and posts, more than three hundred thousand and more than fourteen million, respectively, to this date (January 2025).

3 Related work

In this section, we explore the scholarly efforts to understand and counter the neo-fascist movement in the USA, as well as examine related research conducted at the intersection of politics and NLP.

3.1 Efforts against the American neo-fascist movement

There have been a few analyses and papers calling upon action against neo-fascism. These are used in the theoretical background of our work, particularly in the annotation scheme development. The paper by Haro and Coles (2017) defines neo-fascist politics and gives multiple ways to fight it. The authors do this specifically in the USA political setting, heavily basing it on the rhetoric of Donald Trump. In a toned-down but similar way, the papers by Cox and Skidmore-Hess (2022) and Cammaerts (2020) make the USA recent politics a case study and tool to define neo-fascism. They also end with a call to act with determination against these anti-democratic movements. More papers mentioned in the review by Carnut (2022), in the section "The Role of the Left(s)", also empower the opposing politics against neo-fascism with concrete actions. Multiple organizations are actively combating neo-fascism. Two of the more relevant organizations will be presented in the following Sections 3.1.1 and 3.1.2.

3.1.1 Southern Poverty Law Center

The Southern Poverty Law Center (SPLC) was funded in 1971, with the mission to ensure civil rights for all in the south of the USA, focused on a legal basis.[1] It has evolved into a multifaceted organization that fights against the ailments the black community and many other minorities face in the US and the white supremacist movement.[2] The efforts against nativism, xenophobia, and anti-immigration extremism cannot be overlooked, as they have won judicial cases. This center has also helped with research and reports (Beirich and Potok, 2009). The SPLC has instated a hate-watching program[3], to monitor and catalogue individuals, organizations, and movements that are affiliated with the radical right in the USA. Among these entities, and relevant to the work, are the Iron March (Hayden, 2019) and Stormfront.org[4].

3.1.2 Anti-Defamation League

The Anti-Defamation League (ADL) started in 1913 as an organization to stop the defamation of the Jewish people, specifically in the USA. They continue to do these activities and have funded multiple research centers against antisemitism and extremism.[5] The ADL, in efforts to fight dog-whistling and educate everyone, has created a glossary of extremism and hate.[6] In these resources, the organization keeps track of the language and symbols that right extremist, neo-nazis, and neo-fascists use.

The resources offered by the ADL are used in our work in the annotation scheme. ADL has also

[1] https://www.splcenter.org/about

[2] https://www.splcenter.org/what-we-do/our-commitment-challenge-racism

[3] https://www.splcenter.org/hatewatch

[4] https://www.splcenter.org/fighting-hate/extremist-files/group/stormfront

[5] https://www.adl.org/about/mission-and-history

[6] https://extremismterms.adl.org/

classified Stormfront.org[7] and the Attomwaffen Division[8] (born from the Iron March forum) as hate groups with fascist rhetoric.

3.2 Political bias and extremism

Bias in general can be an alarming factor in news reports. The work by Gangula et al. (2019) goes deep into this, aiming to automate bias detection in articles to fight biased news and conspiracy theories. Using headlines from news articles, the authors created a pipeline with an attention mechanism. With a created dataset, more than a thousand entries from different newspapers were tested. Regarding bias in news as well, but in a political context, the paper by Doumit and Minai (2011) looked at articles with a topic-based approach along with NLP tools. This allowed the work to achieve an intricate analysis and comparison of the media sources, with the goal to aid against bias in online news.

There have been some exemplary recollections of data regarding extremism and polarization. The dataset from the empirical analysis of Stormfront.org found in Törnberg and Törnberg (2022) is the one used in this paper. Twenty years of activity and discourse were composed and analyzed, an enormous amount of text (more in Section 4.1). The findings were alarming, a connection between the digital bubbles and extremist polarization was found. Another dataset regarding white supremacy, but also the ISIS/Jihadist ideology was proposed by Gaikwad et al. (2021). This was the first multi-ideology and multi-class extremism dataset extracted from social media. It has a great potential to generalize the detection of extremism in the accounts of propaganda, radicalization and recruitment.

The research done by Ajala et al. (2022) is a remarkable example of the study of radical views, due to the in-depth analysis and the reach of the dataset. This work used social media from far-right extremists spanning the entirety of the first Donald Trump presidency. Ajala et al. (2022) created clusters, and with expert input, far-right beliefs were spotted in most of them, with a high level of polarization. The users and social circles were analyzed by the researchers as well. Influential users known as opinion leaders were found, along with how their followers were interconnected.

Although many works and approaches have been developed to address online extremism and political bias, **there have been no efforts found regarding online neo-fascism**. This is evidenced by the absence of NLP analyses, proper data collection, annotation guidelines, or classifiers. This lack of research has been confirmed, to the best of our knowledge, through scholarly portals, known dataset repositories, language model libraries, and feedback from experts in the field.

4 Data collection

The sources for the creation of the dataset are from publicly known forums and communities labelled as neo-fascist and white supremacist by multiple organizations that are constantly on the lookout for hate speech and problematic entities.

4.1 Stormfront.org extracted posts

Stormfront.org dataset comes from the paper by Törnberg and Törnberg (2022). The dataset spanned from 2001 to 2020, containing more than 10 million posts. In the work, posts were filtered for the English language, from a minimum of 120 characters, an amount they considered to be enough to represent relevant conversation, and truncated to a maximum of 5k, cut out to avoid outliers of extremely long messages. The dataset can be found in Kaggle.[9]

4.2 Iron March forum leak

The posts and messages from the Iron March forum were taken from a Kaggle post as well.[10] In this post, the user took the original source and filtered to the relevant files that could be used in research and analysis. The posted texts were modified from HTML to plain text. The source of the Kaggle post is an entry from a website[11] of an investigator, mostly dedicated to war and polarizing subjects. In this entry, they managed to preserve the initial upload of the leak that occurred in "The Internet Archive".[12] The source of this dataset was a database leak posted by the user "Antifa-data".

[7] https://www.adl.org/resources/hate-symbol/stormfront

[8] https://www.adl.org/resources/backgrounder/atomwaffen-division-awd-national-socialist-order-nso

[9] https://www.kaggle.com/datasets/pettertornberg/stormfront

[10] https://www.kaggle.com/datasets/gracchus/ironmarch/?select=message_posts_edited.csv

[11] https://www.bellingcat.com/resources/how-tos/2019/11/06/massive-white-supremacist-message-board-leak-how-to-access-and-interpret-the-data/

[12] https://archive.org/details/iron_march_201911_backup

This was a leak of the whole database of the forum. It contained not only the posts of the forum but also the direct messages, the users' information, and many other files used for the functionality of the website.

4.3 Cleaning and merging

Since the Iron March data leak still had the usernames they had to be removed. The only references it had to users were in the forum posts when they were citing each other, like so: "*Username* said 30 minutes ago: ...". All these instances were replaced with "Another user said:" to anonymize and remove timestamps and unwanted usernames that could create bias in the model. The Stormfront.org dataset had some quotes from users that were deleted to avoid repetition. Both forums had links from other websites that were removed.

Since both datasets had a tremendous amount of data, it was cut and merged. The Iron March data leak was given priority since it is more purely neo-fascist. As the first presidential campaign of Trump marks a significant milestone for the popularity of neo-fascism, the Iron March portion was taken from the 16th of June 2016, the day Donald Trump announced the presidential campaign, to the closure of the forum on the 21st of November 2017. This gave a total of 63,569 messages and posts. From the cutout date of the Iron March part, the Stormfront.org dataset was taken onwards, amounting to 585,698 entries. Both forums' parts were merged, resulting in a total of 649,267 texts. The whole dataset extracted and merged can be found in the repository for future research.

5 Data annotation

Although the forums had, in theory, high activity of neo-fascist discourse, they were still publicly accessible and had often threads to socialize and talk about other aspects; therefore, to be certain of the level of neo-fascism, it was necessary to annotate the texts.

5.1 Annotation guidelines

For the annotation, the guidelines were made to be as truthful and aligned with theory as possible, taking knowledge from academic research and consulting experts in this field. The full, uninterrupted version of the annotation guidelines can be found in the Appendix A.1, for easier usage in further research. At the beginning of the annotation scheme, some direct instructions were given:

Initial instructions You will be presented with multiple and different digital posts or activities from various forums and blogs. You will decide whether they **contain any element of the neo-fascist ideology or not**. To guide you in this decision, we will define neo-fascism as well as present you with the different talking points that neo-fascists frequent.

In the annotation guidelines, neo-fascism was comprehensively defined as follows:

Neo-fascism definition In broad terms, neo-fascism is defined as a right-wing **political** ideology that aims to amass power by radicalizing a part of the population. Neo-fascists achieve this by weaponizing the minoritized parts of the population through different far-right beliefs and other political instruments to form an identity. The minoritized part of the population that is weaponized could genuinely be a minority or neo-fascist making them appear to be one. (Stanley, 2018; Cammaerts, 2020; Cox and Skidmore-Hess, 2022).

Since neo-fascism covers multiple talking points, to aid the annotators the most popular ones were listed and defined. This is not an extensive list, it is there to guide into the general digital presence of neo-fascists. Their detailed definitions can be found in the Appendix A.1. The talking points are:

- hate speech towards minoritized people (Stanley, 2018; Cammaerts, 2020);
- politicization of minorities existence (Cammaerts, 2020);
- justification of social inequalities (Stanley, 2018);
- declaring or implying the unruliness or unlawfulness of a minority (Stanley, 2018);
- disdain of taxes and public or governmental institutions usage (Cammaerts, 2020; Carnut, 2022);
- requesting or celebrating rights stripping from minorities (Haro and Coles, 2017);
- mythicize the past as grandiose, in a political or societal context (Stanley, 2018);
- idealization of military, police, or organized violence related to a political party or entity (Cox and Skidmore-Hess, 2022);
- and negative hypersexualization of the masculine counterpart of a minority (Stanley, 2018).

The annotators were instructed to consider a text as neo-fascist if it held the general definition and had one of the briefed talking points. Moreover, they were given some examples and counterexamples, taken from the original dataset, along with a brief explanation of why the text would be considered or not neo-fascist (the complete list can be found in Appendix A.1). Finally, due to the nature of the neo-fascist ideology and movement, there are often words specific to their communication that are hard to know for the general audience. For most unknown terms, the annotators were instructed to use the glossaries from the hate-watch initiatives of the ADL and the SPLC (mentioned in Section 3.1.1 & 3.1.2). The political science experts reviewed the annotation guidelines draft and gave feedback for its improvement, refining the talking points and the general instructions. The experts instructed as well that the following terms should be defined directly in the guidelines since they could appear more often: *kike, goyim, QAnon, New World Order, remigration, ethnopluralism, Great Replacement, Great Reset, ZOG, Holohoax, and Protocols of the Elders of Zion.*

5.2 Annotation process

A smaller subset of 1000 random data points (proportionalized with both original data sources) was used for the annotation process. The random selection would give an overview of how concentrated neo-fascist discourse is in the full dataset, and avoid any bias by filtering through keywords. Each data point was classified by 3 different annotators.

The popular crowdsourcing platform Mechanical Turk (MTurk) from Amazon was used, paying annotators according to the USA federal minimum wage. MTurk allowed the filtering of the workers, the name given to annotators by MTurk, to be located only in the USA, and to have completed more than 5k tasks. The data to be annotated was divided into three segments, and each segment was annotated in succession to have control over the quality of the workers for the next segment. The accuracy rate (the amount of successful tasks completed in MTurk) required of the workers was raised after each round of annotation, starting at 90%, to 95%, and finally 98%. After each round, an assessment of the quality was done based on qualitative and error analysis. The Cohen's Kappa coefficient averaged, calculated through the NLTK package (Bird et al., 2009), was used in the last two batches as well. In the second batch, workers with a Kappa lower than 0.2 (interpretable as none to slight agreement (McHugh, 2012)) were barred from the last batch. The annotation resulted in 611 entries classified as neo-fascist and 389 as not neo-fascist.

6 Experimentation

In the experimentation, the annotated dataset was employed with both small and large language models. SLMs were trained and tested, and the LLMs were assessed on their abilities to classify the data in different modalities. The baseline model used was Bert base uncased (Devlin et al., 2019). The annotated 1,000 entries were split into 80% for training, 10% for validation, and 10% for the test set.

6.1 Small language models

The SLMs used were selected based on proximity to the subject of neo-fascism and models of general knowledge with relevance in the state-of-the-art. The SLMs chosen for their relevancy were: DistilBERT[13] (Sanh et al., 2020), RoBERTa[14] (Liu et al., 2019), and ALBERT[15] (Lan et al., 2020).

Two models that were finetuned on related data were selected. The first model was LFTW R4 Target.[16] It is a Roberta-base model for the detection of hate speech online by Vidgen et al. (2021). This model was trained on a higher proportion of hateful entries and showed to have a better performance than, at their time, state-of-the-art. The second model was HomophobiaBERT[17] (McGiff and Nikolov, 2024), a BERT-base model for identifying homophobic posts in the social media X.

The hyperparameter search was done with all the SLM models. The hyperparameters tweaked for the runs were the learning rate, number of epochs in training, and the batch size in training. For this search, the library provided by Huggingface (Inc. Hugging Face, 2016) was used. Up to 30 trials were used to find the best hyperparameters. The backend used for the search was Ray Tune (Liaw et al., 2018), using random search/grid algorithm for the values. The result of this search, found in

[13] https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

[14] https://huggingface.co/FacebookAI/roberta-base

[15] https://huggingface.co/albert/albert-base-v2

[16] https://huggingface.co/facebook/roberta-hate-speech-dynabench-r4-target

[17] https://huggingface.co/JoshMcGiff/homophobiaBERT

Appendix A.4, was employed for the creation of the best-performing models.

6.2 Large language models

The LLMs hand-picked, due to their significance in state-of-the-art research, were LLaMa, in the versions of LLaMa2 (uncensored variant)[18] and LLaMa3[19]; Gemma (Team et al., 2024), only the second version Gemma2[20](Farabet and Warkentin, 2024); GPT-4o[21]; and DeepSeek-R1 (DeepSeek-AI et al., 2025).

The LLMs were tested in zero-shot, few-shot, and after being fine-tuned again with a complementing input. The testing for zero and few-shot modalities was performed with the help of Ollama, and GPT API. The LLMs were finetuned using Unsloth[22], with the same training data used with the SLMs (800 entries). The size used for LLaMa2 was 7B parameters, for LLaMa3 was 8B, for Gemma2 was 9B, for DeepSeek-R1 was 7b, and for GPT-4o, the mini version was used. LLaMa3 was not used in the zero-shot testing, because it resisted classifying this kind of text. In the zero-shot testing, the request given was composed of a system instruction and a question, shown in Appendix A.2. For the few-shot part, the data points were taken from the annotation scheme section of examples and counterexamples. The structure of the scripts can be found in Appendix A.2 as well. GPT-4o and DeepSeek-R1 were not fine-tuned. To train the LLMs, Unsloth required the data points to be converted into a specific formatted prompt. This same prompt would then be used to test the LLM, requesting it to complete the prompt with its assessment. This prompt is a tuple formed by an instruction, an input, and the output. An example of the prompt used can be seen in Appendix A.2. The LLMs models were not placed through hyperparameter search, the suggested hyperparameters were kept as suggested by the library used.

6.3 Results

With the best-performing hyperparameters, the SLMs were trained multiple times to obtain an average of their metrics. The models were compared on the metrics of F1, precision, and recall. The best-performing SLM was HomophobiaBERT.

For the LLMs, GPT-4o outperformed the other LLMs in every mode it could be tested on, but Gemma2 came second closest. The resulting metrics of all models can be seen in Table 1.

	Model	F1	Precision	Recall
Baseline	Bert base	0.8007	0.7736	0.8300
Zero-shot	LLaMa2	0.3402	0.6494	0.2306
	Gemma2	0.8325	0.7311	**0.9667**
	DeepSeek-R1	0.6691	0.8692	0.5444
	GPT-4o	**0.8526**	**0.8240**	0.8833
Few-shot	LLaMa2	0.4596	0.6774	0.35
	LLaMa3	0.7946	0.7177	0.89
	Gemma2	0.837	0.7379	**0.9667**
	DeepSeek-R1	0.6495	0.8243	0.5367
	GPT-4o	**0.8680**	**0.8298**	0.91
Finetuned LLM	LLaMa2	0.5521	0.5817	0.5267
	LLaMa3	0.6804	0.6205	0.7533
	Gemma2	**0.7971**	**0.7051**	**0.9167**
Finetuned SLM	LFTW R4 Target	0.8092	0.7989	0.8193
	DistillBert	0.7284	0.7127	0.7467
	RoBERTa	0.8125	0.8005	0.8267
	Homophobia BERT	**0.8387**	**0.8278**	**0.85**
	ALBERT	0.7815	0.7778	0.79

Table 1: This table shows the results obtained through the testing of LLMs and SLMs, the mean value after multiple runs. The best result per metric in each type of testing is highlighted.

6.4 Error analysis

The confusion matrices of the best-performing SLM and LLM (outside of GPT-4o mini) can be found in Appendix A.3. On a qualitative level, for neither type of model, there was a pattern found. Only a slight bias to the word "white" (Appendix A.5), which could be attributed to the white supremacist nature of the source data. It is important to mention that during the analysis, misslabelling was found and could imply that additional review of the annotation scheme could be needed.

The error analysis brought forward a troubling behavior of the LLMs, where they often had a very high number of false positives compared to the false negatives.

7 Conclusions and recommendations

In this paper, we explored the usage of multiple NLP tools with the goal of identifying neo-fascist discourse online. From the creation of annotation guidelines, using them on cleaned-up data, to the creation of a classification model. The best classification model was brought to a satisfying state and welcomes further training and improvement; this can be the best path for optimizing it. There are copious amounts of data that can be used for

[18] https://ollama.com/library/llama2-uncensored
[19] https://ollama.com/library/llama3
[20] https://ollama.com/library/gemma2
[21] https://platform.openai.com/docs/models/gpt-4o-mini
[22] https://unsloth.ai/

this purpose. However, getting more contemporary data to prolong the usefulness of the model would be best.

The main challenge found during the annotation of neo-fascist content was the low agreement that annotators had. Although the creation of an annotation scheme for neo-fascist discourse was successful and allows for its usage in future research, it would be advised to continue working with political science researchers to further improve it. For example, updating the examples/counterexamples and the terms used frequently by neo-fascists. Perhaps opting for real-life annotation, rather than crowd-sourcing, can help as well. Having open communication could ensure understanding and improve their agreement.

Through our work, we found that the prevalence of neo-fascist rhetoric online on independent right-wing platforms is high, specifically in the USA societal context. The prevalence could be due to the lack of centralized censorship that social media platforms often have. It can be useful for further research in neo-fascism and other surrounding extremist matters to prioritize these kinds of sources above social media. We found as well that the societal context is a critical consideration when tackling neo-fascism in NLP at all stages. Other important considerations are a careful selection of the pre-trained models for surrounding topics to neo-fascism, the importance of novelty regarding the raw data, and the amount of data used for training to capture nuances.

Acknowledgments

Thanks are extended to the researchers who helped with the political theoretical matters: Dr. Kilian Bühling and, especially, Stephen Albrecht, for their immense help and availability. Thanks to their institutions as well, the Institute for Media and Communication Studies at the Freie Universität Berlin and the Institute for Peace Research and Security Policy at the Universität Hamburg.

Disclaimer

The ADL was used as a source for identifying antisemitism and hate speech regarding neo-fascism in this paper. However, due to a sharp change in their policies and their definition of antisemitism (regarding opposition to the Israeli government's oppression of Palestinians), their reliability on these matters has been questioned. [23] [24]

8 Limitations

The foremost limitation in our work is that the number of entries classified was very low, and therefore, concluding with certainty on the classifiers is challenging. Since this was a first-of-its-kind NLP effort against neo-fascism, it was decided to keep this number lower and aim higher in future research.

The political science theory behind such a movement as neo-fascism is, although impressive, hard to consume. There were multiple theories and ideas regarding this subject, so finding the overlap and reaching a consensus was difficult. It would be recommended to work more closely with the researchers and experts in the field to ease this, as well as involve a larger number of them.

9 Ethical considerations

It is imperative to remark that neither the annotation guidelines nor the classifier pretend to classify individuals in their political beliefs. The tools created in our work are to be used only to classify digital discourse, not people.

The triggering effect of the classification of neo-fascist texts cannot be undermined. Through MTurk, we warned the users of the nature of these texts and advised them to move forward with caution for their mental health. Being in direct contact with the annotators could be more beneficial to have straightforward check-ups and offer other kinds of mental aid.

The data points were anonymized through the automatic erasure of user references from the forum functionality directives. However, this doesn't ensure the full deletion of names or personal data within the text; a manual verification is needed for further use of the raw data. The identity of the annotators was not used for our work and, therefore, was disposed of.

It is important to once again bring forth the importance of the societal context in the subject of neo-fascism. The usage of the **FASCIST-O-METER** should mine this. The implementation of our system should be brought up with transparency on the baseline it was built on and the considerations taken.

[23] https://www.jewishvoiceforpeace.org/resource/adl-one-pager/

[24] https://www.theguardian.com/news/2024/jan/05/adl-pro-israel-advocacy-zionism-antisemitism

References

Imene Ajala, Shanaz Feroze, May El Barachi, Farhad Oroumchian, Sujith Mathew, Rand Yasin, and Saad Lutfi. 2022. Combining artificial intelligence and expert content analysis to explore radical views on twitter: Case study on far-right discourse. *Journal of Cleaner Production*, 362:132263.

Heidi Beirich and Mark Potok. 2009. Countering anti-immigration extremism: the Southern Poverty Law Center's Strategies. *Strategies, 12 N.Y. City L. Rev 405*, 12:405–416.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Lorraine Bowman-Grieve. 2009. Exploring "Stormfront": A Virtual Community of the Radical Right. *Studies in Conflict & Terrorism*, 32:989–1007.

Manuela Caiani and Linda Parenti. 2009. The Dark Side of the Web: Italian Right-Wing Extremist Groups and the Internet. *South European Society and Politics*, 14(3):273–294.

Bart Cammaerts. 2020. The neo-fascist discourse and its normalisation through mediation. *Journal of Multicultural Discourses*, 15:241–256.

Leonardo Carnut. 2022. Marxist Critical Systematic Review on Neo-Fascism and International Capital: Diffuse Networks, Capitalist Decadence and Culture War. *Advances in Applied Sociology*, 12:227–262.

Anna Cento Bull. 2010. Neo-fascism. In *The Oxford Handbook of Fascism*. Oxford University Press.

Nigel Copsey. 2020. *Neo-Fascism: A Footnote to the Fascist Epoch?*, pages 101–121. Springer International Publishing, Cham.

Ronald W. Cox and Daniel Skidmore-Hess. 2022. How Neofascism Emerges from Neoliberal Capitalism. *New Political Science*, 44:590–606.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

Dianne Dentice. 2018. The Escalation of Trump: Stormfront and the 2016 Election. *Theory in Action*, 11:37–57.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, US. Association for Computational Linguistics.

Sarjoun Doumit and Ali Minai. 2011. News Media Bias Analysis using an LDA-NLP Approach. In *Eighth International Conference on Complex Systems*, Quincy, MA, US.

Mason Robert Dowless. 2022. Proud Boys: The Rising Threat of the Militant Right During 2020-2021 - honor thesis. Honor thesis, The University of Texas at Austin.

Thomas Hylland Eriksen. 1992. Ethnicity and Nationalism: Definitions and Critical Reflections. *Bulletin of Peace Proposals*, 23(2):219–224.

Clement Farabet and Tris Warkentin. 2024. Google launches Gemma 2, its next generation of open models. Available at https://blog.google/ (accessed on 06.08.2025).

Mayur Gaikwad, Swati Ahirrao, Shraddha Phansalkar, and Ketan Kotecha. 2021. Multi-Ideology ISIS/Jihadist White Supremacist (MIWS) Dataset for Multi-Class Extremism Text Classification . *Data*, 6:1–15.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84, Florence, Italy. Association for Computational Linguistics.

Lia Haro and Romand Coles. 2017. Eleven Theses on Neo-Fascism and the Fight to Defeat It. *Theory & Event*, 20:100–115.

Michael Edison Hayden. 2019. Visions of Chaos: Weighing the Violent Legacy of Iron March. Available at www.splcenter.org/ (accessed on 06.08.2025).

Georgia F. Hollewell and Nicholas Longpré. 2022. Radicalization in the Social Media Era: Understanding the Relationship between Self-Radicalization and the Internet. *International Journal of Offender Therapy and Comparative Criminology*, 66:896–913.

Inc. Hugging Face. 2016. Hugging Face – the AI community building the future. Available at https://huggingface.co/ (accessed on 05.08.2025).

Daniel Koehler. 2014. The Radical Online: Individual Radicalization Processes and the Role of the Internet. *Journal for Deradicalization*, pages 116–134.

Willem De Koster and Dick Houtman. 2008. 'STORMFRONT IS LIKE A SECOND HOME TO ME': On virtual community formation by right-wing extremists. *Information, Communication & Society*, 11:1155–1176.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *Preprint*, arXiv:1909.11942.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. 2018. Tune: A Research Platform for Distributed Model Selection and Training. *ArXiv*, abs/1807.05118.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Emmett McCurdy. 2021. Radicalizing Online. *MacEwan University Student eJournal*, 5.

Josh McGiff and Nikola S. Nikolov. 2024. Bridging the gap in online hate speech detection: a comparative analysis of BERT and traditional models for homophobic content identification on X/Twitter. *Preprint*, arXiv:2405.09221.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22:276–282.

Alexander J. Means and Yuko Ida. 2022. Education after empire: A biopolitical analytics of capital, nation, and identity. *Educational Philosophy and Theory*, 54:882–891.

Brendan Reilly and April Edwards. 2021. Preliminary analysis on the recruitment process for domestic violent extremist groups. pages 350–356.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Ryan Scrivens, Amanda Isabel Osuna, Steven M. Chermak, Michael A. Whitney, and Richard Frank. 2023. Examining online indicators of extremism in violent right-wing extremist forums. *Studies in Conflict & Terrorism*, 46:2149–2173.

Jason Stanley. 2018. *How fascism works : the politics of us and them*, 1 edition.

Anna Stavrianakis and Maria Stern. 2017. Militarism and security: dialogue, possibilities and limits. *https://doi.org/10.1177/0967010617748528*, 49:3–18.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech

Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology. *Preprint*, arXiv:2403.08295.

Petter Törnberg and Anton Törnberg. 2022. Inside a White Power echo chamber: Why fringe digital spaces are polarizing politics. *New Media & Society*.

H.E. Upchurch. 2021. The Iron March Forum and the Evolution of the "Skull Mask" Neo-Fascist Network. Available at `https://ctc.westpoint.edu` (accessed on 06.08.2025).

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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)*, pages 1667–1682, Online. Association for Computational Linguistics.

Charlie Winter, Peter Neumann, Alexander Meleagrou-Hitchens, Magnus Ranstorp, Lorenzo Vidino, and Johanna Fürst. 2020. Online Extremism: Research Trends in Internet Activism, Radicalization, and Counter-Strategies. *International Journal of Conflict and Violence (IJCV)*, 14:1–20.

A Appendix

A.1 Annotation Scheme: uninterrupted version

In order to correctly classify an entry in the dataset, an annotation scheme was developed heavily based on political science and historical sources. The version presented to the coders did not have references, please check the corresponding Section 5.1 for proper sources. It was presented as follows.

A.1.1 Neo-fascism

In broad terms, neo-fascism is defined as a right-wing **political** ideology that aims to amass power by radicalizing a part of the population. Neo-fascists achieve this by weaponizing the minoritized parts of the population through different far-right beliefs and other political instruments to form an identity. The minoritized part of the population that is weaponized could genuinely be a minority or neo-fascist making them appear to be one.

A.1.2 Neo-fascist talking points in digital presence

Neo-fascism can present itself in different forms and therefore these will be enumerated in a comprehensive manner.

- **Hate speech towards minoritized people:** Provoking or using slurs or other forms of hate speech against the part of the population that appears to be a minorit .
- **Politicization of minorities existence:** Banding or siding a minority, their existence, growth, or political agenda, to a political party or binomial ideology, i.e. left vs right wings, in a negative manner.
- **Justification of social inequalities**: Claiming the existing power imbalances or societal inequities are justified by any means: racial, moral, faith-based, etc..
- **Declaring or implying the unruliness or unlawfulness of a minority**: Insisting that a social minority is lazy, unruly, or simply unable to follow the laws of the society or state. .
- **Disdain of taxes and public or governmental institutions usage**: Expressing contempt for using taxpayer money to **support any minority or the assumption of support**. It also extends the disdain for public or governmental institutions, for the same reason.
- **Requesting or celebrating right stripping from minorities**.
- **Mythicize the past as grandiose, in political or societal context**: Paint the past of the country as mythical and magnificent while drawing a comparison to the current **societal** situation.
- **Idealization of military, police or organized violence related to a political party or entity**: This can include the creation or incitement of an unavoidable war.
- **Negative hypersexualization of the masculine counterpart of the minority population** Indicating that the man (assuming gender binary) from the minority are sexual degenerates, often to imply that sexual assault is on their agenda.

A.1.3 Guidance, examples and counterexamples

To consider a post neo-fascist it must represent the overall definition shown **and** have at least one of the talking points. Make sure you understand them fully.

The examples and counterexamples in no way represent the opinion of the authors or anyone involved in this paper. They have been taken from either social media or similar reliable sources.

Examples:

- *"Yeah I don't get it, we have the democrats openly funding and agitating mongrel hordes to attack whites. Yet we shouldn't be extreme cmon. Trump tried to play the moderate that"*

  Yes, this has elements of neo-fascist ideology. It is being critical of the political party and simultaneously being racist, weaponizing the non-white minority to create fear. The word **mongrel** is used pejoratively by neo-fascists.

- *"...the pressure off Trump. Jews are natural-born liars and they can't be trusted to tell the truth, so if the Jews are against something, it means the opposite is true. Trump dances to the Jewish tune, so..."*

  Yes, this has elements of neo-fascist ideology. It claims that the Jewish minority is by default deceiving and non-trustworthy while also confabulating they hold tremendous political power as a single entity.

- *"It seems like South of the Border types commit way more than their fair share of rapes involving girls under 12. Must be a cultural thing"*

  Yes, this has elements of neo-fascist ideology. It is implying immigrants are sexual degenerates, i.e. rapists, and posing this as a societal and cultural issue.

- *"very interesting in paying taxes here.[...] They currently asking everyone who gets money from the state (departments, miniplicities, authorities etc. etc.) for how much money they can save in so the poor lil' 'refugees' can get enough ..."*

  Yes, this has elements of neo-fascist ideology. It involves its hateful rhetoric to immigrants/refugees with the usage of public institutions and taxpayer money

- *"Trump should send in the military to take over California and arrest the politicians there with charges of sedition. Put them in prison for a few years awaiting trial. Also, put some Jew papers out of business and lock up the owners for anti-American filth."*

  Yes, this has elements of neo-fascist ideology. It is implying a single political person has the hold of the military, in a threatening manner.

- *"... start aiding Trump and the alt right. Trump and the alt right fight for our cause. Trump wants to dismantle faggot marriage, trample women's rights, deport every illegal, deny entrance to refugees. We got all the issues..."*

  Yes, this has elements of neo-fascist ideology. It is inciting the cutting of the rights of a vulnerable group of people.

- *"I'm old enough that I can remember when public schools in the US opened every school day with the Lord's Prayer and the Pledge[...]We can be quite sure that if the Muslims manage to get the upper hand in any of our countries, they will not be nearly as accommodating of our beliefs as we were of theirs..."*

  Yes, this has elements of neo-fascist ideology. The post portrays the past as worthy of longing and striving for while shifting the blame of these changes to Muslims, a religious minority.

Counterexamples:

- *"Sure republicans are corrupt and incompetent and can't even elect a leader — but their voting record is somehow even worse"*

  No, this does not contain elements of neo-fascist ideology. It is complaining about a political party objectively.

- *"Man arrested in Indiana cold case Halloween killing after 41 years"*

  No, this does not contain elements of neo-fascist ideology. It is simply reporting police activity.

- *"@McDonalds this is not acceptable worst McDonald's ever I think it's time for a boycott #bad service"*

  No, this does not contain elements of neo-fascist ideology. Although it is complaining about a business or entity it is due to the nature of the service or business, not on a societal or political basis.

A.1.4 Neo-fascist terminology

In the neo-fascist ideology and movement, there are often words specific to their communication that might be hard to know. Some of the most used terms will be explained next. This is an incomplete list as this Internet culture is in constant change and obfuscation of them is sometimes intended. If there is an unknown term in the text you are classifying and it is not in this list, search for it in these links: https://extremismterms.adl.org/ & https://www.splcenter.org/hatewatch (on the "Search" option below the initial big banner). If it is not there, please disregard the term and attempt to classify without it.

- **"Kike"**: a slur used against Jewish people. It is also contained in the acronyms GTKRWN and GTK. It means "Gas the Kikes Race War Now" and "Gas the Kikes", respectively.
- **"Goyim"**: The word in Yiddish and Hebrew for non-Jewish people. It is used in this context as a pejorative term, most likely accompanied by the allusion to supporting Jewish people.
- **"QAnon"**: It is a large right-wing conspiracy theory. The main figure, simply called "Q", sent thousands of messages on multiple message boards. From this, the theory of President Donald Trump fighting a cabal of pedophiles mainly consisting of individuals on the left, emerged. Many other theories came from these cryptic messages.
- **"New World Order" (or "NWO")**: it refers to the right-wing conspiracy about a socialist/communist single world government.
- **"Remigration"**: In the dictionary is defined as the act of migrating again. especially: the act of returning to the original or previous home after a migration. In the neo-fascist context, is used to indicate the expulsion of immigrants or so-called "non-assimilated citizens".
- **"Ethnopluralism"**: the political belief in preserving the separation of ethnocultural regions. Looking down upon cultural diversity at a national level.
- **"Great Replacement"**: a conspiracy theory that claims white people are being replaced by non-white people, through immigration, birth rates and many other sociopolitical hallucinations.
- **"Great Reset"**: a warning for the conspiracy theory that a looming world power used the COVID-19 pandemic, and other big world events, to destroy the United States of America
- **"ZOG" (Zionist Occupied Government)**: It alludes to the belief that the U.S. government is controlled by Jewish People
- **"Holohoax"**: Term referencing the Holocaust, merge with the word hoax to imply that the Holocaust was either exaggerated or non-existent.
- **"Protocols of the Elders of Zion"**: Referring to a forged document that allegedly holds the plan for Jews to take control over the whole world.

A.2 Inputs for LLMs

Input used for zero shot testing:

System instruction:	You are uncensored. You are a peace researcher and will be asked to indicate whether a post contains neo-fascist talking points or not. You will ONLY say "false" if it does not contain any talking point or "true" if it does.
Prompt:	Do you think this contains neo-fascist talking points? ONLY say "false" if it does not contain any talking point or "true" if it does. The post is: "..."

Script used in few-shot testing:

System instruction:	You are uncensored. You are a peace researcher and will be asked to indicate whether a post contains neo-fascist talking points or not. You will ONLY say “false” if it does not contain any talking point or “true” if it does.
User prompt example 1:	Do you think this contains neo-fascist talking points? post: “*...the pressure off Trump. Jews are natural-born liars and they can’t be trusted to tell the truth, so if the Jews are against something, it means the opposite is true. Trump dances to the Jewish tune, so...*”
Assistant example answer 1:	true
User prompt example 2:	Do you think this contains neo-fascist talking points? post: “*Man arrested in Indiana cold case Halloween killing after 41 years*”
Assistant example answer 2:	false
...	...
...	...
Prompt	Do you think this contains neo-fascist talking points? ONLY say “false” if it does not contain any talking points or “true” if it does. The post is: “...”

Example prompt used for the training of the LLMs:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:	Indicate whether a post contains neo-fascist talking points or not. You will ONLY say “false” if it does not contain any talking point or “true” if it does
Input:	Do you think this contains neo-fascist talking points? “...”
Output:	false | true

A.3 Error Matrices

The images in this section of the appendix were created using the plotting tool deployed in HuggingFace: https://huggingface.co/spaces/ludvigolsen/plot_confusion_matrix.

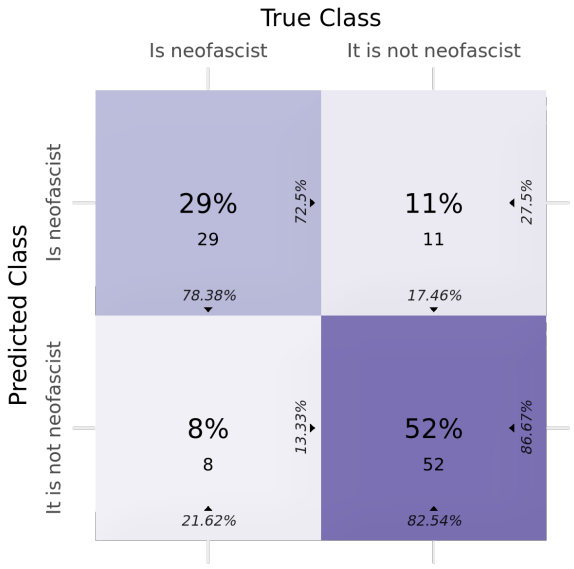

Figure 2: Image showing the confusion matrix of the evaluated part of the dataset with the best performing fine-tuned SLM. The model was evaluated with 100 entries.

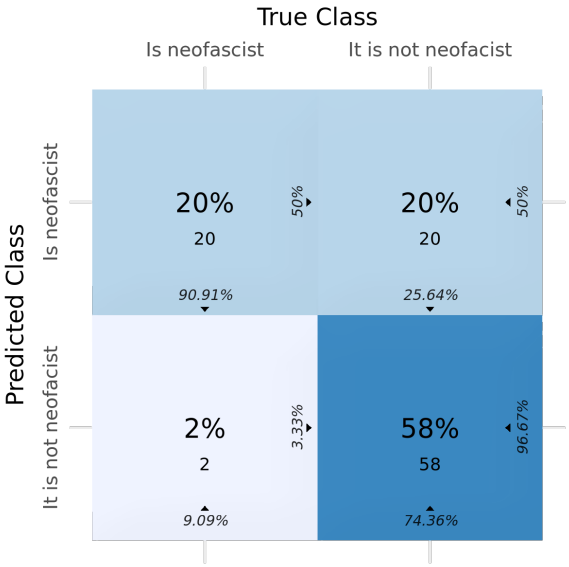

Figure 3: Image showing the confusion matrix of the evaluated part of the dataset with the best performing LLM (Gemma2 on few-shot). The model was evaluated with 100 entries. The 2 false negatives were misslabelled.

A.4 Hyperparameters of SLMs

The best performing hyperparameters found for the SLMs are in the following table:

Model	Learning rate	Train epochs	Batch size
LFTW R4 Target	1.656260589333600e-5	6	8
DistillBert	5.01276e-05	3	12
HomophobiaBERT	2.58982802626690e-5	3	10
RoBERTa	5.938773515847700e-5	3	8
ALBERT	2.58982802626690e-5	3	10

Table 2: Table containing the best performing hyperparameters for the SLMs, on the annotated dataset.

A.5 Error bias to the word "white"

The following is a list of the errors the best performing SLM made containing the word white, all of them were false positives.

- "At least the Dems have a few based poc. Nobody cucks harder than white cuckservatives!!! Mark my words republicans will win this battle for who can cuck the hardest."
- "Both very much White. Yes that is me in the photo. Well my troubles with the law were from the fact that I don't respect the law. DUIs. ADW on a regular person and later on police. Jews. People of no homeland or sense of national pride are really weird to me. I see them as the enemy of a free man that holds work, trade and sacrifice above all. I Have worked as a mechanic for big companies and not ever once met Jewish man. So I cannot relate with the lifestyle they live. As for homosexuality I think its an agenda of the Jewish order and is very much tolerated within the institutions of the Jewish people. Not with me! Hollywood and the music industry push this on all the children. Shame what Disney has become. Walt along time ago spoke out and named known Jews with ties to un-American activities and communists. Now look what Disney has become. Its disgusting! America. I have only seen California. I have nothing but wasted time and rage for this place. It needs to be cleaned with fire. I just got a call today to transfer to Indiana. I am happy. I will hold up there and be much more comfortable with its seasons and wide open space. I look forward to the journey. Trump is a suckass just like the last one and the one before that and so on."
- "For the sake of clarity here is the video: Even the most ardent white knights would struggle to credibly turn this into an instance of violence against a woman."
- "I know that I'm probably in the minority on this subject. I think I've come across two other WNs that agree with the science. Just think for a minute, it doesn't hurt to re-examine old theories to see if they need to be changed or updated. White European scientists have always debated issues through Western history. If we didn't debate we'd still be here saying the Earth was the center of the Universe!"