2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Harmful Communication

Detection of Toxic Language and Threats on Swedish

Lisa Kaati, Arvin Moshfegh, Kevin Lindén
Stockholm University
Stockholm, Sweden
Email: lisa.kaati@dsv.su.se

Amendra Shrestha
Mind Intelligence Lab
Uppsala, Sweden
Email: amendra@mindintelligencelab.com

Nazar Akrami
Uppsala University
Uppsala, Sweden
Email: nazar.akrami@psyk.uu.se

***Abstract*—Harmful communication, such as toxic language and threats directed toward individuals or groups, is a common problem on most social media platforms and online spaces. While several approaches exist for detecting toxic language and threats in English, few attempts have detected such communication in Swedish. Thus, we used transfer learning and BERT to train two machine learning models: one that detects toxic language and one that detects threats in Swedish. We also examined the intersection between toxicity and threat. The models are trained on data from several different sources, with authentic social media posts and data translated from English. Our models perform well on test data with an F1-score above 0.94 for detecting toxic language and 0.86 for detecting threats. However, the models' performance decreases significantly when they are applied to new unseen social media data. Examining the intersection between toxic language and threats, we found that 20% of the threats on social media are not toxic, which means that they would not be detected using only methods for detecting toxic language. Our finding highlights the difficulties with harmful language and the need to use different methods to detect different kinds of harmful language.**

I. Introduction

Social media and online environments have become an important part of our society and daily life functioning. In Sweden, 80% of the population uses social media daily [18]. Although the internet and social media offer great opportunities for everyone to, for example, get informed and participate in discussions anytime and anywhere, there are also downsides. For example, when individuals use the opportunities that online environments provide to threaten, incite violence, or spread hate [3], [16]. More than half of the Swedish population between the ages 16-45 expressed having been targets of both online hate speech and threats [6].

When the online conversation climate becomes contaminated by harmful communication, the impact on the mental well-being of targeted individuals can be severe. Individuals who are consistently subjected to toxic comments and threats may opt to disengage from public discourse, leading to the suppression of certain voices and the fading away of meaningful, nuanced discussions. Being the target of toxic language and threats can also result in psychological consequences for individuals, and when this targeting is persistent, it can be equated to bullying. Bullying has severe consequences and can lead to different types of exposure and vulnerability that reduce the quality of life and psychological well-being. It is also not uncommon for victims of bullying to feel both guilt and shame [14]. Moreover, spreading harmful communication undermines social cohesion and threatens democratic processes [25].

Moderating and removing all harmful communication is almost impossible [16], and there is a constant need for efficient tools that can assist social media platforms and other forms of communication services to automatically detect all kinds of harmful content, including toxic language and threats, without violating freedom of speech [1], [30], [38]. One of the difficulties with harmful communication is to decide what is harmful and what is not. Most social media platforms have their own community rules that determine what kind of communication that is allowed and what kind of communication that is considered harmful.

One type of harmful communication is what we in this work call *toxic language*. The term toxic language describes communication that poisons the climate of conversation on social media. Toxic language includes communication prohibited by law (such as hate speech) but also other forms of abuse, such as derogatory speech, invasion of privacy, or disrespect. However, not all harmful communication is toxic. Another form of harmful communication is different kinds of threats - both violent and non-violent. Threats do not necessarily have to be toxic and could therefore be difficult to detect for methods that are designed to detect toxic language. In this work, we explore the possibility of using two different methods for identifying harmful language online: one for detecting toxic language and one for detecting threats in Swedish. Thus, our aim is to explore the intersection between toxic communication, including hate speech, and communication that includes a direct or indirect threat.

Today, several methods are available for detecting hate speech and toxic language, but there have been fewer attempts to identify threats in online environments. One major concern that scholars have brought up is the lack of datasets, particularly for low-resource languages that have limited data and resources available [2]. Most existing tools and models are

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11.
https://doi.org/10.1145/3625007.3627597

based on datasets in English, which limits existing models' effectiveness in identifying and detecting threats in low-resource languages. The lack of datasets makes training models from scratch more resource-consuming and generally challenging.

A. Outline

This paper is outlined as follows. In Section II, we provide an introduction to harmful communication, more specifically to toxic language and threats. Section III describes some of the previous work on detecting toxic language and threats. Section IV describes how we trained the models and the datasets we used for training. In Section V, the performance of our models is presented, and in Section VI an analysis of the intersection between threats and toxic language is provided. Section VII contains a discussion of the results, and finally, some conclusions and directions for future research are presented in Section VIII.

II. Harmful communication

Harmful communication is an umbrella term for a wide range of communication that causes harm, distress, or negative consequences to individuals or groups. It can manifest in various ways, including but not limited to toxic speech, hate speech, trolling, cyberbullying, spreading misinformation or disinformation, shaming, and threatening individuals and groups. In the present work, we focus on toxic language and threats.

A. Toxic Language

Although many of us have an instinctive understanding of what toxic communication is, there is no unified definition. People generally hold diverse perspectives and varying levels of tolerance toward what qualifies as toxic communication. Furthermore, the context of language is important - a message may be perceived as toxic or not toxic at all by the same individual, depending on the context.

Toxic language is a broad term that captures various forms of offensive or harmful language. Pavlopous et al. [28] describe toxic language as an umbrella term that includes several different types of language, including offensive language, abusive language, and hateful language. Google's Counter Abuse Technology team, who developed the Perspective API, describes toxic language as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." When using toxic language as an umbrella term, it includes profanities, identity attacks, slights, insults, threats, sexually explicit content, demeaning language, language that incites violence, and hostile and malicious language targeted at a person or group.

B. Threats

The definition of what is considered to be a threat differs. In [7], a threat is defined as "a communication of intent to harm". While intent plays an essential role in a threat, it is also necessary to consider both the context and the interpretation of the target. In [7], threats are divided into three categories: direct, conditional, and indirect/veiled.

- **Direct threats** are characterized by being explicit in their nature and contain an intention to cause harm, often using pronouns combined with violent verbs and occasionally including specification of the time and way the threat will be carried out. Examples of such threats include *"I will kill you tomorrow"*.
- **Conditional threats** are characterized by an if-then-format, where the condition determines the execution of the threat. Expressions of such threats can, for example, be *"If you do not do X, I will [threat of harm/violence]"*. These types of threats often blur the line between a warning and an actual threat, which is often used to circumvent legal action and consequences. A conditional threat does not necessarily need to be written in an explicit if-then format. A threat can be conditional as long as there is a presence of conditional variables that determine whether the threat will be realized or not.
- **Indirect or veiled threats** are the most challenging kinds of threats due to the ambiguity and lack of explicitness or conditional variables. Indirect threats are often characterized by containing implications of violence, such as *"I could shoot you if I wanted to"*. Indirect or veiled threats sometimes completely lack indication of violence or harm and rather rely on the context to convey the potential threat. Examples of this can be *"I know where you live"* or *"Shame if something happened to you"*. In some contexts, these types of expression might be completely non-threatening, while in other contexts, they can be used as threats that potentially circumvent legal actions or consequences [7], [27].

III. Related work

There have been many attempts to detect hateful language in previous research. The detection of hateful language is commonly tackled in three main directions: building datasets and resources, detecting the binary distinction of hate versus non-hate content, and identifying various subcategories of hateful language [4].

There are different strategies regarding the techniques and methods that can be used to detect hate speech. The simplest approaches are lexicon or dictionary-based methods that use a list of words that are toxic and/or considered to be related to hate speech. If a word from the target list is present in a text – the text is considered hateful. Some lexicon-based approaches have been extended with automated reasoning to identify targets of hateful language [29].

Earlier work on detecting hateful language used different kinds of traditional machine learning with features such as in [10], [11]. However, with the recent advancement in text analysis and language processing, the use of deep learning-based technologies and transfer learning has become a leading approach for detecting various forms of hateful language. In [5], levels of hate are measured on some online platforms, an approach that was later used in [20] to analyze the levels of toxic language used in right-wing extremist communication online. In [23] RoBERTa was used to train a model to detect

toxic language in the English language. The model received an F1-score of 0.91 on the test data, but when tested in the wild on communication from a wide range of different social media platforms, the model received an F1-score ranging between 0.7-0.8 depending on the level of agreement between annotators when labeling the data.

Research on detecting hateful language in other than English is more limited. It is also often the case that the authors need to collect and annotate their own data, which makes it difficult to compare different approaches.

There are some recent examples of hate speech detection on Arabic (Tunisan) using an Arabic BERT (AraBERT) [31], and on Bengali using Bengali BERT [34]. In [33], automatically annotated tweets have been used to train a Dutch BERT model(BERTje). The results show that a monolingual language model fine-tuned with automatically annotated data is a competitive baseline against the zero-shot transfer of a multilingual model, but also that adding automatically annotated data to manually annotated data is a source of error that reduces the performance of models significantly. There have been some attempts to detect hateful language in Swedish. For example, a dictionary-based approach was presented in [19] and a machine learning-based approach in [13] using ULMFiT [17].

While the research on hate speech and toxic language is extensive, there have been fewer attempts to detect threats specifically. However, one example is the dictionary-based approach presented in [8], where a threat dictionary that indexes threat levels from texts across media platforms was developed. The threat dictionary shows convergent validity with objective threats in American history, including violent conflicts, natural disasters, and pathogen outbreaks.

Moreover, in [15], a method to automatically detect threats of violence using machine learning is presented. The dataset used to train the machine learning model is a previous version of the dataset described in [16] that we also use in our work. The dataset contains around 30 000 sentences from YouTube videos annotated manually as threats or non-threats. In [38], the same dataset was used to train a number of different classifiers to recognize threats. In [38], an F1-score of 0.61 was obtained on the YouTube data.

Research on detecting threats in languages other than English is limited. In [9], a work done in Urdu is described. Amjad et al. also highlight the challenges of using machine learning to detect threats on imbalanced datasets. Unless using techniques such as class weighting - imbalanced datasets impact the performance of machine learning models [39].

IV. Method

We have used Bidirectional Encoder Representations from Transformers (BERT) [12] to build our models for detecting hateful language and threats. We used the Swedish BERT (KB-BERT) developed by the KBLab for data-driven research at the National Library of Sweden [26]. Since the model is already pre-trained and has a general understanding of how specific languages work, both semantically and syntactically, it can be fine-tuned to perform a specific task without needing to train and create a model from scratch [12].

A. Datasets

The dataset used for creating the model for detecting toxic language consists of 6 101 toxic texts and 15 901 not toxic texts. The texts are from a wide range of Swedish social media and are annotated by at least three annotators. The dataset has been collected and annotated during several previous research projects [13], [21], [22]. In one of the data sources, names were removed from the text.

The dataset used for creating the model for detecting threats consists of the following sources:

- An English dataset from Hammer et al. [16] containing approximately 30 000 manually annotated sentences from YouTube videos.
- Swedish social media posts annotated as threats and non-threats.
- A synthetically generated dataset from Vidgen et al. [36] containing English texts labeled with different kinds of hate speech. Texts labeled as threatening were used as threats, and the remaining texts were used as non-threats.

All English data were translated into Swedish and partially re-annotated.

TABLE I
THE DATASETS

Category	Number of posts
Toxic	6 101
Not toxic	15 901
Threats	2 335
Not Threats	3 705

The threat dataset consists of 6 040 manually annotated posts where 2 335 (39%) were labeled as threats and 3 705 (61%) were labeled as non-threats. A total of 493 threats and 493 non-threats originate from the translated English synthetic data, 1 047 threats, and 1 047 non-threats from the translated English YouTube data. A total of 795 threats and 2 165 non-threats originate from several Swedish forums and social media. The dataset consists of 2 960 (49%) original Swedish texts and 3 080 (51%) translated texts. A summary of the datasets used for training and testing can be found in Table I.

B. Training

All experiments were done with 10 epochs, and the batch size was kept at 8. During the training process, we chose the best-performing model measured by accuracy on the validation set. While training the BERT model, Rectified Adam optimizer [24] was used with a learning rate of 3e-5. When training the threat model, the maximum sequence of tokens was fixed to 256 tokens. When training the toxic language model, the maximum sequence of tokens was fixed to 128 tokens.

The datasets were partitioned into training, validation, and testing sets with random splits. For the threat model, the data

was divided into splits of 80%, 10%, and 10%, and for the toxic language model, splits of 90%, 9%, and 1%, respectively.

A random seed was included to ensure the reproducibility of the splits, and stratified sampling was employed to maintain a balanced representation of the datasets.

The Python libraries Trainer and Training Arguments from HuggingFace were used throughout the development process. The model's performance during testing was based on each experiment's accuracy, precision, recall, and F1-score metrics on the test set. Each training and experiment used different combinations of hyperparameters.

V. Results

The performance of the model for classifying toxic language is presented in Table II. As can be seen in the table, the F1-score for classifying non-toxic language is 0.98, and for toxic language 0.94. These results are very promising and an improvement compared to previous work done on Swedish toxic language detection [13]. The performance of the model for classifying threats is presented in Table III. The F1-score for classifying threats was lower than that for toxic language: 0.92 for classifying non-threats and 0.86 for classifying threats. The results are encouraging, which allows us to analyze the models in the wild and examine the connection between toxic language and threat. We will explore this further in the following section.

TABLE II
Performance of the toxic language classification model

	Precision	Recall	F1-Score	Support
Non-toxic	0.99	0.97	0.98	161
Toxic	0.92	0.97	0.94	60
Accuracy			0.97	221
Macro Avg.	0.95	0.97	0.96	221
Weight Avg.	0.97	0.97	0.97	221

TABLE III
Performance of the threat classification model

	Precision	Recall	F1-Score	Support
Non-Threat	0.91	0.93	0.92	370
Threat	0.88	0.85	0.86	234
Accuracy			0.90	604
Macro Avg.	0.89	0.89	0.89	604
Weight Avg.	0.90	0.90	0.90	604

VI. The Relation Between Toxic Language and Threats

A. Testing in the Wild

One of our major aims is to explore the intersection between toxic language and threat. The basic assumption here is that toxic language does not necessarily need to contain threats, and threats do not necessarily need to contain toxicity. To explore the relationship between these two forms of harmful communication, we began by testing our models on a new dataset consisting of 40 000 social media posts from several Swedish forums and online environments. As a first step, the threat classifier was run on the dataset. A total of 1 076 (2.7%) posts were classified as threats and 38 924 posts (97.3%) as non-threats. From each label (threats and non-threats), we selected a representative sample with a 95% confidence level and a 5% margin of error. The samples were selected to reflect the characteristics of the two labels accurately and contain a smaller, more manageable representation of the two different labels. The representative samples consisted of 284 threats and 381 non-threats - a total of 665 posts.

TABLE IV
The models' performance on the social media dataset

Label	No. of Posts	Correctly classified
Threats	284	170 (60%)
Toxic	252	236 (94%)
Not threats	381	376 (99%)
Not toxic	413	369 (89%)

The representative samples were manually annotated. The results of the annotation showed that 60% (170 of 284) of the threats were true positives, while 40% (114 of 284) were false positives. These results mean that 40% of the threats were incorrectly classified. The threat classification model performed better on the non-threats: 99% (376 of 381) of the non-threats were true negatives, and 1% (5 of 381) were false negatives.

Next, we run the toxic language classification model on the sample of 665 posts (representative samples above combined). The results showed that a total of 252 posts were classified as toxic. The sample was manually annotated as to their toxicity, and the results showed that 94% (236 of 252) of the toxic posts were true positives, while 6% (16 of 252) were false positives. For the non-toxic comments, the model performance decreased, with 89% of the posts being correctly classified as non-toxic (true negatives), and 11% (46 of 413) incorrectly classified as non-toxic (false negatives).

Table IV shows the number of correct classifications when testing the two models on the 665 posts of new unseen data.

B. Toxic language and threats

To further explore the intersection between toxic language and threats, we analyzed the results specifically regarding the overlap between toxicity and threat. Here, we use the set of correctly identified threats and toxic language, i.e., 175 threats (170 + 5) threats and 280 (236 + 44) toxic comments.

We found that out of the 175 identified threats - 133 (76%) also contained toxic language. Out of the 280 posts that contained toxic language, 48% contained threats. Figure 1) shows the overlap between threats and toxic language in the dataset.

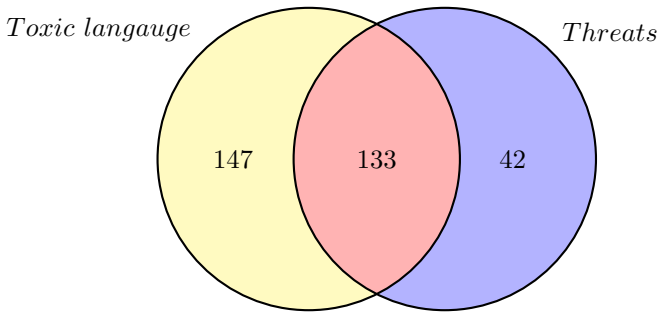

Fig. 1. The relationship between toxic language and threats. More than 2/3 of the threats also contain toxic language but around 1/3 of the threats did not contain any toxic language.

The posts that contain both toxic language and threats are often threats of violence or threats targeting a group of people, such as people of color, Jews, immigrants, or Muslims. Many of the threats are also vague and indirect.

The posts that contained only threats were often indirect or veiled threats. Often the threats were about non-violent acts and statements that can be interpreted differently depending on the reader and the context. The context can make it very difficult to determine if a comment is a threat or not.

VII. Discussion

While it is difficult to compare the performance of models when different datasets are used to train and test the models, our models outperform previous work done on other languages, for example, Urdu [9] both when it comes to detection of offensive language and detection of threats. Our toxic language classification model also performs better than previous work on Swedish [13], [22].

Exploring the performance of the models when they are applied to new unseen data, we conducted an experiment where both models were run on a new Swedish language social media dataset. A sample of the classified data was then manually annotated. The results showed that the model for classifying threats had difficulties in determining when a text contained a threat - 40% of the threats identified by the model were incorrectly classified, but the model performed well when determining if a text does not contain a threat. Only one percent was misclassified.

The threat classification model had a threshold of 0.5 to distinguish between threats and non-threats. By adjusting this threshold, it is possible to control the trade-off between reducing the number of false positives (missed non-threat classifications) and minimizing the number of false negatives (missed threat classifications). For example, increasing the threshold value, e.g., from 0.5 to 0.8, could reduce the number of false positives and thus increase the precision when detecting threats. However, increasing the threshold would also increase the number of false negatives, meaning that some threats could be missed. A better way to deal with this problem is probably more research and more and better training data.

Our model for toxic language detection performed better when it was tested on new unseen data: 94% of the comments that were classified as toxic were actually toxic. However, the model missed some of the toxic comments: 11% of the comments were classified as not toxic while they actually were toxic. The toxic language classification model was also set on a threshold of 0.5 to distinguish between toxic and non-toxic. This threshold can also be changed depending on what is considered to be important.

Further, we used the two classification models to explore the intersection between toxic language and threats. First, the results showed a significant overlap between threat and toxicity, with almost 80% of the threats being toxic and 50% of the toxic having the characteristics of threat. These findings are interesting in a context where it is more common to search for toxicity than threat. The implication of our findings would be that if approximately 20% of the threats found on social media are not toxic, then they would not be detected using methods for detecting toxic language only. This finding illustrates a challenge for online moderating efforts and highlights the difficulties with harmful language and the need to use different methods to detect harmful language. A more diverse range of models is needed to achieve optimal moderation, and these models need to target different aspects of harmful language.

Speaking of moderation and employing different models to moderate online communication, some scholars stress that there are several ethical and social risks of harm associated with using language models for classification [37]. For example, language models can introduce biases and stereotypes, and the performance of a model may differ depending on the target group. In [35] the concept of harmfulness in Scandinavian language models is examined. The results showed that Scandinavian models, including the BERT model we have used in the studies presented in this paper, generate disturbing, offensive, harmful, and gender-based stereotypes.

While we have fine-tuned the underlying language model with training data, the identified biases in the language model may still exist, and there is also a possibility that new biases have been introduced with the training data. The training data we have used consists of social media data, and it is likely that some targets, groups as well as individuals, of toxic language are overrepresented. The models for toxic language have not been tested for biases using functional tests such as HateCheck [30]. A natural direction for future work would be to use such

tests to identify weaknesses in the model.

VIII. Conclusions and Future Work

In this paper, we presented two different models: one for detecting toxic language and one for detecting threats in Swedish. Both models performed well on test data, obtaining an F1-Score of 0.94 and 0.86 when classifying toxic language and threats, respectively, and 0.98 and 0.92 when classifying posts as non-toxic and non-threats, respectively.

Determining what is a threat or not can be challenging since both the sender's intention and the recipient's interpretation of it need to be considered [7], [32]. The in-the-wild test showed that the model's performance for classifying threats decreased, and around 40% of the posts classified as threats were misclassified. The model was better at classifying non-threats and only misclassified 1% of the posts. More importantly, our analyses showed that while there is a large overlap between toxic communication and threats, more than 20% of the threats did not contain toxic language. Thus, models for toxic language detection will not be able to identify such threats.

For future work, it would be optimal to test the models on data from real scenarios and use the results to improve the models. Future research could also focus on exploring biases embedded within the models. A better understanding of biases would provide valuable insights into what kind of training data is needed to improve the overall fairness and effectiveness of the models.

Finally, the research on harmful language has advanced significantly in recent years. However, the focus of the research has mainly been on toxic language in general. Exploring the intersection between different forms of harmful languages, as we did in the present paper, would be a natural and highly pertinent direction for further investigation. We believe that insight from exploring the intersection between different forms of harmful language would improve not only general models of toxic language but also models that target specific forms of online harm, such as cyber harassment, cyberbullying, and the spread of self-harm content and terrorist-related content.

Acknowledgements

The research was supported by grants from Riksbankens Jubileumsfond to Nazar Akrami (P15-0603:1).
The computations/data handling were enabled by resources provided by Chalmers e-Commons at Chalmers.

References

[1] G. Al-Turaif and F. Fkih. A Review on Threat Detection Approaches in Social Networks. *International Journal of Computer Science and Network Security*, 21(10):353-361, Oct. 2021.

[2] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butta, H. I. Amjad, O. Vitman, and A. Gelbukh. Overview of Abusive and Threatening Language Detection in Urdu at FIRE 2021, July 2022.

[3] N. Ashraf, R. Mustafa, G. Sidorov, and A. Gelbukh. Individual vs. Group Violent Threats Classification in Online Discussions. In *Companion Proceedings of the Web Conference 2020*, WWW '20, pages 629-633, New York, NY, USA, Apr. 2020. Association for Computing Machinery.

[4] D. Battistelli, C. Bruneau, and V. Dragos. Building a formal model for hate detection in french corpora. *Procedia Computer Science*, 176:2358–2365, 2020. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020.

[5] T. Berglind, B. Pelzer, and L. Kaati. Levels of hate in online environments. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 842–847, 2019.

[6] Brottsoffermyndigheten. Näthat och självcensur det nya normala. 2021.

[7] P. Casula, A. Anupam, and N. Parvin. 'We found no violation!': Twitter's Violent Threats Policy and Toxicity in Online Discourse. *C&T '21: Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech*, pages 151-159, June 2021.

[8] V. K. Choi, S. Shrestha, X. Pan, and M. J. Gelfand. When danger strikes: A linguistic tool for tracking america's collective response to threats. *Proceedings of the National Academy of Sciences*, 119(4), 2022.

[9] M. Das, S. Banerjee, and P. Saha. Abusive and Threatening Language Detection in Urdu using Boosting based and BERT based models: A Comparative Approach. Nov. 2021.

[10] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, May 2017.

[11] G. A. De Souza and M. Da Costa-Abreu. Automatic offensive language detection from twitter data using machine learning and feature selection of metadata. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2020.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 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*, pages 4171–4186. Association for Computational Linguistics, 2019.

[13] J. Fernquist, O. Lindholm, L. Kaati, and N. Akrami. A study on the feasibility to detect hate speech in swedish. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4724–4729, 2019.

[14] M. S. Gibbs. Psychological impacts of toxic exposure in third world countries: An extrapolation. *Impact Assessment*, 8(4):7–18, 1990.

[15] H. L. Hammer. Detecting threats of violence in online discussions using bigrams of important words. In *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference*, JISIC '14, page 319, USA, 2014. IEEE Computer Society.

[16] H. L. Hammer, M. A. Riegler, L. Ovrelid, and E. Velldal. THREAT: A Large Annotated Corpus for Detection of Violent Threats. *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1-5, Sept. 2019.

[17] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[18] Internetstiftelsen. Sociala medier. https://svenskarnaochinternet.se/rapporter/svenskarna-och-internet-2022/sociala-medier/, 2022. Accessed: 2023-02-27.

[19] T. Isbister, M. Sahlgren, L. Kaati, M. Obaidi, and N. Akrami. Monitoring Targeted Hate in Online Environments. *Second workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*, 2018.

[20] L. Kaati, K. Cohen, and B. Pelzer. *Heroes and Scapegoats: Right-wing Extremism in Digital Environments*. European Commission and Directorate-General for Justice and Consumers., 2021.

[21] L. Kaati, K. Cohen, B. Pelzer, D. Wallgren, A. Akrami, and J. Yourstone. *Könsskillnader i utsatthet för toxiskt språk online*. FOI Memo 7741, Swedish Defence Research Agency, 2021.

[22] L. Kaati, K. Cohen, B. Pelzer, D. Wallgren, A. Akrami, and J. Yourstone. *Toxiskt språk i svenska digitala miljöer.* FOI Memo 7740, Swedish Defence Research Agency, 2021.

[23] L. Kaati, A. Shrestha, and N. Akrami. A machine learning approach to identify toxic language in the online space. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 396–402, 2022.

[24] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *CoRR*, abs/1908.03265, 2019.

[25] P. Lorenz-Spreen, L. Oswald, S. Lewandowsky, and R. Hertwig. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 7:1–28, 11 2022.

[26] M. Malmsten, L. Börjeson, and C. Haffenden. Playing with words at the national library of sweden – making a swedish BERT, 2020.

[27] M. E. O'Toole. *The School Shooter: A Threat Assessment Perspective*. Federal Bureau of Investigation, Jan. 1999.

[28] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos. Toxicity detection: Does context really matter? In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *ACL*, pages 4296–4305. Association for Computational Linguistics, 2020.

[29] B. Pelzer, L. Kaati, and N. Akrami. Directed digital hate. In *ISI*, pages 205–210. IEEE, 2018.

[30] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. B. Pierrehumbert. HateCheck: Functional Tests for Hate Speech Detection Models. 2020.

[31] P. O. Salomon, Z. Kechaou, and A. Wali. Arabic hate speech detection system based on arabert. In *2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing*, pages 208–213, 2022.

[32] K. Storey. The language of threats. *International Journal of Speech, Language and the Law*, 2(1):74-80, 2013.

[33] D. Theodoridis and T. Caselli. All that glitters is not gold: Transfer-learning for offensive language detection in dutch. *Computational Linguistics in the Netherlands Journal*, 12:141–164, Dec. 2022.

[34] S. R. Titli and S. Paul. Automated bengali abusive text classification: Using deep learning techniques. In *2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS)*, pages 1–6, 2023.

[35] S. Touileb and D. Nozza. Measuring harmful representations in scandinavian language models. *5th workshop on Natural Language Processing and Computational Social Science (NLP+CSS) at EMNLP 2022*, 2022.

[36] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. 2020.

[37] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359, 2021.

[38] A. Wester, L. Øvrelid, E. Velldal, and H. L. Hammer. Threat detection in online discussions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 66-71, San Diego, California, June 2016. Association for Computational Linguistics.

[39] Y. Yadav, P. Bajaj, R. K. Gupta, and R. Sinha. A Comparative Study of Deep Learning Methods for Hate Speech and Offensive Language Detection in Textual Data. In *2021 IEEE 18th India Council International Conference (INDICON)*, pages 1-6, Dec. 2021.