C-REF: An Explainable Framework for Social Media Toxicity Classification

Daniel Yakubov
dyakubov@gradcenter.cuny.edu

Abstract

The rise of social media platforms has enabled many different people to be able to share their opinions in a public space with ease. Unfortunately, this has led to the rise of toxicity in digital spaces. This toxicity can be hateful and cause emotional distress to a reader or can directly incite violence towards individuals or groups. Therefore, it is out of necessity that systems that can aid in the detection of toxic discourse are developed. Such systems are well studied in the fields of Computational Linguistics, NLP, and Digital Humanities but there is still much ongoing research in the direction of creating explainable systems that provide users with rationales for the given classification label. This need is twofold as it comes from the lack of consensus on what is considered toxic and what isn't and the general conversation about free speech. The present work serves to propose a framework that not only classifies a given text, but provides linguistically informed rationales for the classification that mirror human rationales. The output provided by the framework provides justifications which can be used for educational purposes and for transparency in order for the system's labels to be trusted by individuals.

1 Introduction

Since its conception and widespread adoption, social media has undeniably played a vital role in digital interactions. It is an online medium by which many individuals communicate thoughts, news, and ideologies through their posts. Some of the posts shared on social media platforms are harmful to readers since they contain toxic language such as hate speech, and/or offensive speech. The issue of defining hate speech alone is a complex issue (Sellars 2016), this is evidenced by the fact that human annotator labels have low agreement on hate speech datasets (Kwok and Wang 2013; Vigna et al. 2017; Ousidhoum et al. 2019). A further complication is that the low agreement between annotators can also be attributed to individual traits of the annotators, such as their age (Sang and Stanton 2022).

Given the low agreement between annotators, any effort to effectively detect toxic discourse should be explainable. If one were to disagree with an automatic classification, the presentation of a rationale could serve to demonstrate the exact point or disagreement, or it can serve to educate one what makes a given text toxic. For the purposes of this work, an explainable system is one that provides a reliable rationale for its decision. The primary benefits of an explainable system for toxicity detection include education and transparency. Education is an effective approach to minimizing the presence of toxic discourse on social media platforms. An existing educational approach provides an explanation to an offender, and those who read the offender's post, on why the toxic statements presented in the discourse are factually incorrect and offensive (Tekiroğlu 2020). Transparency refers to the need for an individual interacting with the system to have access to the decision making process employed by the system. More broadly transparency is needed in intelligent systems in order for the decisions of the system to be trusted by users (Vorm 2018).

Hate speech Detection is a well researched task (Schmidt and Wiegand 2017) and more recently, the development of large language models such as BERT (Devlin et al. 2019) and GPT-3 (Brown et al. 2020) has led to the application of these models to the task with impressive results (Caselli et al. 2021; Chiu et al. 2021), but this class of models is not explainable in decision making processes. Further, the rationales for the given results cannot be understood from the internal matrix representations stored in the model. The latter issue is prevalent in all end-to-end deep learning approaches, and not just those employing the largest existing models. Due to this, there has been recent work in the direc-

tion of explainable hate speech classification (Jiyun et al. 2022; Karim et al. 2021), but it is still an active area of research, as the internal representation issue does not yet have a solution in end-to-end methodologies.

The present work contributes to the study of explainable hate speech classification systems by proposing the Constituent Rationale Explainability Framework (C-REF) and applying it to toxic social media posts. C-REF involves parsing a given text into a syntax tree and then using a classifier to detect the largest constituent that could be a rationale for toxic speech. A rationale is a span of text that a human annotator selects in a given post, p, as their reasoning for the label for post p. In doing so, C-REF performs slightly worse on F1 and accuracy than the baseline version of its classifier component trained to classify toxic speech end-to-end. However, the outputs of C-REF include linguistically meaningful rationales with the classification label, making the outputs transparent and allowing for educational research on toxic discourse.

2 Previous Research

2.1 Explainable Text Classification

The literature for explainable text classification can be divided into two primary categories: post-hoc systems that add an explanation for a decision made by a model after the classification label is outputted and self-explaining systems, which generate their explanations at inference time. A complete overview of such systems and the dichotomy between them is provided in the survey study done by Danilevsky et al. 2020. A representative approach in the category of post-hoc explanations is the well-known framework, LIME (Ribeiro et al. 2016). The LIME approach employs an explainer model that provides an interpretable explanation for a decision made by an existing model. While post-hoc approaches like LIME are useful for pre-trained models that cannot be easily modified, these approaches do not get the explanations from the decision making model itself, but only approximate the explanations using properties of the models outputs. Approaches in the category of self-explaining systems mitigate this effect. The MARTA framework developed by Arous et al. (2021) is a system that performs explainable text classification at the document level. MARTA achieves this by using incremental text classification at the sentence level, where each sentence is classified to be a possible label rationale or not. Then, using the summation of weighted labels for each sentence in the document, a classification label for the entire document is calculated. A limitation of the MARTA framework is with its current formulation, it can only work for entire documents containing multiple sentences. Furthermore, a rationale for a classification may not always be an entire sentence, but a sub-sentence unit, such as a single word or a constituent.

2.2 Hate Speech Detection

Early approaches for hate speech classification have involved creating lexical and linguistic heuristics and using them in statistical classifiers or rule based systems, such as the Smokey system proposed by Spertus (1997). These systems are explainable if the vocabulary is available to the user, but they are not robust to new environments, such as new hateful slang, and thus require expert maintenance. Further, these systems are not context sensitive and may filter the casual speech of communities which reclaim slurs. Two sensitive communities that have undertaken such reclamation efforts would be African Americans and LGBT individuals (Poppa-Wyatt 2020), as such, their online discourse would be inadvertently targeted by such systems.

More recent approaches to hate speech detection involve the application of deep learning (Vigna et al. 2017; Pitsilis et al. 2018; Malik et al. 2022). Malik et al. (2022) provides an exploration of which deep learning models have reported the best performance for the task of hate speech classification, showing that transformers provide the best performance, but are computationally more expensive than other approaches. Large language models have also been applied to the task. Caselli et al. (2021) explore the direction of using transfer learning to train a domain-specific BERT (Devlin et al. 2018) on a dataset of Reddit hate speech comments. The resulting model was named HateBERT. The authors aimed to evaluate HateBERT on its robustness and portability. This was achieved through the data chosen for evaluation, which consisted of three unbalanced datasets. Pretrained HateBERT showed an improvement over BERT in most metrics, but did not achieve state-of-the-art results. The large language model, GPT-3 (Brown et al. 2020), has also been applied to the task of hate speech detection, with accuracy up to 85% (Chiu et al. 2021). The paper explores different prompt-based approaches

including 0-shot, few-shot, and instruction-based few-shot in-context learning. Each approach shows better results than the last on hate speech classification. The few-shot in-context prompting led to open class classification for hate speech, but this direction was not explored further by the researchers.

Though these approaches using large language models end-to-end are promising, they are not explainable. These approaches do not provide justifications for classifications, and even if one of these models were modified to provide justifications, it is unclear if these justifications would be reliable, as the Chiu et al. 2021 discuss the ability GPT-3 has to produce counterfactual information. A major limitation to these recent works involving end-to-end deep learning systems is that, despite the impressive performance on evaluation metrics, the outputs of much of these systems are not explainable. Because of this, further research had to be done to show that these models have unintended biases and thus can harm the communities they are meant to benefit (Sap et al. 2021). In an explainable system, the bias would be more explicit and possibly be caught in an earlier stage of development.

2.3 Explainable Hate Speech Detection

Due to the issues highlighted in the study of hate speech detection, much research has begun in the field of explainable methods in the detection of hate speech. Shvets et al (2021) adapted a general concept extraction model (Shvets & Wanner 2020) to a hate speech domain by using a reference corpus and a series of algorithms in order to extract what the researchers call "targets" and "aspects". A target is defined as the concept in the text that is the recipient of the hate speech, it does not necessarily have to be a person, for example, "feminist books" can be a target (Shvets et al 2021). An aspect is defined as a concept that is attributed to the target. A target is extracted from a text by first identifying all the concepts in the text, and then weighting all the candidates on their likelihood to be a target, then sampling the target with the highest score. Aspect extraction is done in a similar way, but the relationship of a candidate to the target is considered. They report that the task of aspect extraction is difficult for their system, due to the syntactically complex relationship an aspect can have with a target. The goal of this work was to create an open-class hate speech classification system, and thus the system is not evaluated as a step in closed-class classification. However, the fine-grained labels outputted by this system make it more explainable than a closed-class binary classification system. This approach could be an explainable component of a hate speech classification system but by design it is not a closed class classification system.

Karim et al. (2021) develop an explainable system for hate speech detection in Bengali. The system developed in this work classifies a given text in Bengali using an end-to-end BERT model, and then applies a post-hoc explainability system consisting of probing and sensitivity analysis (Saltelli 2002) to explain the rationale of the classification model. Kim et al (2022) propose the task of Masked Rational Prediction for hate speech detection. The researchers pre-finetune BERT to predict human rationales as an intermediate task. Then, this model is further fine-tuned on the task of hate speech detection. This technique is self explaining and reports state-of-the-art results, but yet ultimately still employs an end-to-end classification system. The present work contributes to this body of research. In particular, this work explores the idea of providing linguistically coherent rationales and employing a deep learning classifier incrementally for the task of toxic speech classification, where previous work has focused on classifying the entire text at once.

3 Method

The Constituent Rationale Explainability Framework (C-REF) aims to provide human readable explanations for the classification of toxic social media posts. C-REF consists of two components, a syntactic constituency parser and a text classification model. In this work, the parser used is the Berkeley Neural Parser (Kitaev et al. 2018, 2019), which was chosen due to state-of-the-art performance. The other component of C-REF is an arbitrary classifier which classifies constituent spans of text as *toxic*, if the constituent span can be modeled as a human rationale for toxicity, *non-toxic* if the constituent span can not be modeled as a human rationale for toxicity, and *mixed*, if the constituent span contains some text that is a rationale and some text that is not a rationale for the toxicity of the text. A constituent span refers to a span of text that is a syntactic constituent. For this work, a DistilBERT model (Sanh et al. 2019) that was fine-tuned on constituent spans was used for the constituent span classification component.

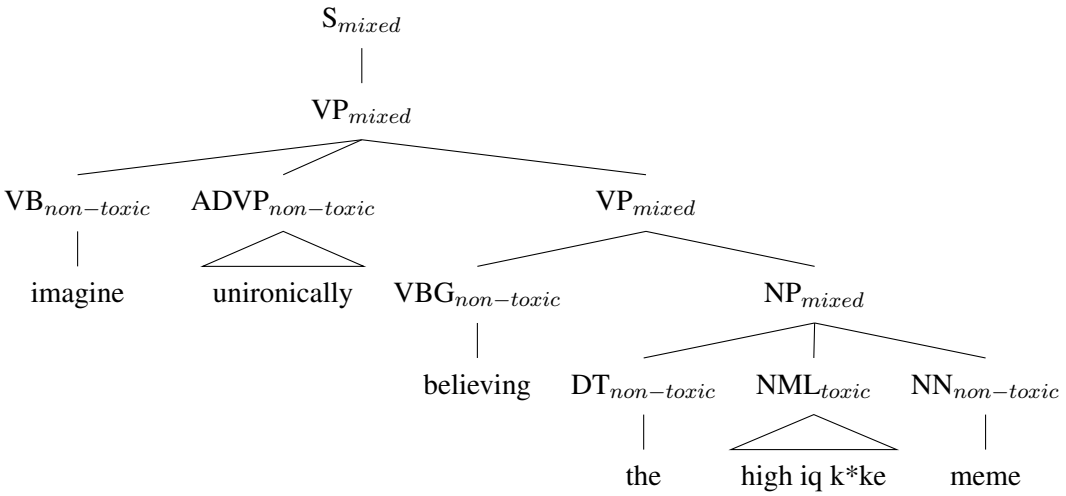

Figure 1: C-REF applied to a toxic post from the HateXplain dataset. Each node X represents a constituent span that it dominates and has a label in {*mixed*, *non-toxic*, *toxic*}. C-REF first parses a text into a syntax tree and then works Inorder classifying each X until hitting a node X that is not X_{mixed}

The fine-tuning parameters included three epochs over the data, a batch size of 64, and the AdamW optimizer (Loshchilov and Hutter 2017) initialized at 5e-5. The generation procedure for the training data is outlined in the data section.

Making a prediction using C-REF employs Algorithm 1

Algorithm 1 Constituent Span Classification

```
procedure S_CLASSIFICATION(S)
    for s_i in S do
        if is_tree(s_i) then
            t ← word_sequence(s_i)
            c ← to_span(t)
            l ← classify(c)
            if l is mixed then return c, l
            else return S_CLASSIFICATION(s_i)
        else
            c ← to_word(s_i)
            l ← classify(c)
            if l is mixed then
                label ← non-toxic
        return c, l
```

Given a text, t, where t is the word sequence $w_0, w_1, w_2...w_n$, a syntax tree S is generated from t using the parser component. The tree S consists of $s_0, s_1, s_2, ...s_n$, where each s_i is a sub-tree of S that dominates c_i, a subset of t. At the first time step, the classifier component outputs a label in {*toxic*, *non-toxic*, *mixed*} for S. If the label for S is not *mixed*, S is returned as the rationale. If the label is *mixed*, S is updated to be s_0. In subsequent time steps, S is reassigned to nodes that are more deeply embedded in the tree. When S is a bare node, or a node dominating only one item in t, further recursion is impossible. In this case, if the classification component incorrectly classifies c as *mixed*, l is corrected to be *non-toxic* as a heuristic. After the procedure is completed, the returned output is Z, series of constituent spans and respective corresponding labels. If any z_i in Z is labeled as toxic, the entire text is considered *toxic*, and all z_i in Z with the label of *toxic* are returned as the rationales.

4 Data

The binary classification version of the open-source dataset HateXplain (Mathew et al. 2021) is used to train and evaluate C-REF. This dataset contains 20148 entries that consist of a post_id, 3 annotators, and a list of tokens from the post. The post_ids correspond to a standard split for the dataset which is used in the present paper. For each annotator, there is a overall text label in the set of {*non-toxic*, *toxic*}. 8733 of the posts in the dataset are labeled as *non-toxic* and 11415 are labeled as *toxic*. For each post labeled as *toxic* the data is further annotated with a vector, r, that indicates which section of the text was identified by annotators as the rationale for their label. r contains n items, where n is the length of the list of tokens. In r, the tokens marked as a rationale by the annotators are stored as 1, all other tokens are marked as 0. This dataset

was selected as it is the largest dataset containing rationales for the task of hate speech detection.

Preprocessing the data involved taking the computing an average rationale vector using the r from all annotators per data point. This was done in order to have a condensed representation of the rationales. The tokens which had higher agreement between rationales have a value closer to one, and the tokens with lower agreement have a value closer to 0. Tokens marked with a 0 indicate that out of all three annotators, not one found that token to be a rationale for the label. The label for the text is decided by a majority vote. If all three annotators labeled a data point with a different label, the data point is discarded. Up to this point, the preprocessing procedure follows Mathew et al. 2021.

To make the data suitable for fine-tuning the classifier component of C-REF, the data is further subdivided into a file which contains all the constituent spans per post. This is done in a linguistically informed way. For each post a syntactic tree representation of the posts is computed using the parser component. Then, an algorithm traverses the tree Inorder in search of the largest constituents that correspond to either vectors of only non-zero rationales or vectors of only 0 rationales. The constituents that correspond to all zero vectors are labeled as *non-toxic* rationales, and the constituents that correspond to a rationale vector which contains all non-zero items are labeled as *toxic*. The intermediate constituents which contain both rationale and non-rationale tokens are also stored with the label of *mixed*. As a heuristic, function words are included in the rationale spans even if a human annotator did not specify it. This is required since for a given sentence such as "The X people are Y", the parser does not consider [$_{NP}$ X People] as a constituent, instead the tokens X and People are considered bare constituents, however, the inclusion of function words addresses the issue as [$_{DP}$ The X People] is considered as a constituent. The result of this preprocessing procedure is the constituent span dataset. The label balance is shown in Table 1.

The constituent label span dataset is skewed towards *non-toxic*. This is due to the fact that most toxic posts contain spans that are not rationales.

Constituent Span Label	Count
non-toxic	93773
mixed	53352
toxic	31546

Table 1: The label balance of the training set for the classifier component

5 Baseline

There are two baseline models used to evaluate C-REF: one that serves as a lower bound and one that serves as an upper bound for performance. The lower bound model is a dummy classifier with a most frequent configuration. This means the dummy classifier will classify everything as the most frequent label in its training set, which in the case of the post text classification, was *toxic*. The upper bound model is a DistilBERT model fine-tuned to classify posts. The fine-tuning procedure involved passing over the training data for three epochs at a batch size of 64 and using the AdamW optimizer initialized at 5e-5. In order to create comparable models, these hyper parameter values were not tuned to maximize performance for the baseline model, as was the case for the classifier component. This baseline model shows how well the classifier component of C-REF would work end-to-end as opposed to when trained on constituent spans.

6 Results/Evaluation

Model	F1	Accuracy
Dummy Classifier	0.372	0.593
C-REF	0.764	0.774
DistilBERT	0.778	0.788

Table 2: Results of evaluation, DistilBERT preformed best in both F1 and Accuracy

For evaluation, the metrics of Accuracy and macro F1 were used. Since the dataset is not perfectly balanced, F1 is a stronger indicator of performance than accuracy. Table 2 shows the performance of the C-REF model against the two baseline models.

C-REF preforms significantly better than the Dummy Classifier on both F1 and Accuracy, however, the DistilBERT model outperforms C-REF in both F1 and accuracy. The similarity between the scores of the DistilBERT model and C-REF could be attributed to the initial classification in C-REF. The first step of C-REF classifies the text

as *toxic*, *non-toxic*, *mixed*. This step provides a preliminary classification for the entire text using a similar DistilBERT model, leading to the similarity in scores.

6.1 Qualitative Analysis

Hate Speech Post
kim kardashian is a n*gr*ss by now anyway
rip nehlen he may be married to a b**ner and a christcuck but at least he aiming in the right direction
my gay ass can't take it anymore

Table 3: Examples of the C-REF applied to texts. Dark red means the span is a true positive, pink means the span is a false positive

The heuristic of including stop words in the classification window led to a difference in the predicted rationales and the human rationales. In Table 3, we can see that words like 'a', 'to', and 'and' were not considered by humans as part of the rationale but returned by the model as a rationale. The third example in the table shows an example of a false positive for the model. Unfortunately, some in-group usages of slurs were classified as hate speech rationales, this is also exemplified in this example as the word 'gay' is said to be the rationale for the false positive. There were several instances of this type of error in the results. The reason for this could be related to the lack of context that the model has at the time of classifying the span. However, the benefit of C-REF and similar systems is that the model's bias can be explicitly observed by a user thus providing transparency in these cases and allowing for researchers to observe exact causes of misclassifications.

Interestingly, C-REF catches newer slurs such as 'christcuck' which targets religious christian individuals. In general, false negatives were far rarer than false positive. This could be contributed to the fact that HateXplain was gathered using a lexicon approach where Mathew et al (2021) gathered tweets and gab posts that contained certain words, and with limited context, these selected words may be more likely to be classified as toxic.

7 Conclusion

Toxicity classification is a task that benefits from explainable models due to its controversial nature and low inter-annotator agreement (Kwok and Wang 2013; Vigna et al. 2017; Ousidhoum et al. 2019). While this is an active field of work (Kim et al 2022; Karim et al. 2021; Shvets & Wanner 2020), this work is unique as it incorporates syntactic knowledge into the task. This is done by the Constituent Rationale Explainability Framework (C-REF), a linguistically informed system that returns both a label for a text and the constituent spans that are the rationales for the label classification. C-REF consists of two components, a parser component to parse a text into its syntactic structure, and a classifier component that operates on the constituents returned from the parser component. Evaluation was performed on HateXplain (Mathew et al. 2021), a dataset for toxicity classification. C-REF outpreformed the lower bound baseline model in accuracy and F1 but was not equal to the upper bound baseline model in terms of performance. However, the primary benefit of C-REF is the rationales that it uses in order to make it's decision and the transparency of the decision making process. The rationales given for a label by the framework included tokens that human rationales did not include, such as function words, but this was largely due to a heuristic that was employed to make rationales constituents, and could be rectified easily if it were the goal to do so.

8 Future Work

Since the performance of C-REF relies on its two components, an exploration of the impact of different models would be interesting here. Particularly for the parser component, where a heuristic had to be employed to capture some constituents. Future work could also explore how this framework performs in other domains rather than toxicity detection. Some issues were encountered in the evaluation of the model that could be attributed to lack of context given, it would be interesting to explore how to incorporate some context into the framework as well. Finally, it would be interesting to see if an intelligent system like C-REF does help progress research in algorithmic bias and promote trust with end users.

References

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. Marta: Leveraging human rationales for explainable text classification. *Proceedings of the AAAI*

Conference on Artificial Intelligence, 35(7):5868–5876.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. Detecting hate speech with gpt-3.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. In *AACL*.

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, Minnesota. Association for Computational Linguistics.

Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md. Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. 2022. Why is it hate speech? masked rationale prediction for explainable hate speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6644–6655, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Annual Meeting of the Association for Computational Linguistics*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Jitendra Singh Malik, Guansong Pang, and Anton van den Hengel. 2022. Deep learning for hate speech detection: A comparative study.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *ArXiv*, abs/2012.10289.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

Mihaela Popa-Wyatt. 2020. Reclamation: Taking back control of words. *Grazer Philosophische Studien*, (1):159–176.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Andrea Saltelli. 2002. Sensitivity analysis for importance assessment. *Risk analysis*, 22(3):579–590.

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *Information for a Better World: Shaping the Global Future*, pages 425–444, Cham. Springer International Publishing.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Andrew Sellars. 2016. Defining hate speech. *Berkman Klein Center Research Publication*, (2016-20):16–48.

Alexander Shvets, Paula Fortuna, Juan Soler, and Leo Wanner. 2021. Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190, Online. Association for Computational Linguistics.

Alexander V. Shvets and L. Wanner. 2020. Concept extraction using pointer-generator networks and distant supervision for data augmentation. In *EKAW*.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of IAAI-97, the 9th Conference on Innovative Application of Artificial Intelligence*, pages 1058–1065.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Italian Conference on Cybersecurity*.

Eric S. Vorm. 2018. Assessing demand for transparency in intelligent systems using machine learning. *2018 Innovations in Intelligent Systems and Applications (INISTA)*, pages 1–7.

Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network.