2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Levels of Hate in Online Environments

Tor Berglind
Uppsala University
Uppsala, Sweden
Email: berglind.tor@gmail.com

Björn Pelzer, Lisa Kaati
Swedish Defence Research Agency
FOI
Stockholm, Sweden
Email: firstname.lastname@foi.se

***Abstract*—Hate speech in online environments is a severe problem for many reasons. The space for reasoning and argumentation shrinks, individuals refrain from expressing their opinions, and polarization of views increases. Hate speech contributes to a climate where threats and even violence are increasingly regarded as acceptable.**

The amount and the intensity of hate expressions vary greatly between different digital environments. To analyze the level of hate in a given online environment, to study the development over time and to compare the level of hate within online environments we have developed the notion of a *hate level*. The hate level encapsulates the level of hate in a given digital environment. We present methods to automatically determine the hate level, utilizing transfer learning on pre-trained language models with annotated data to create automated hate detectors. We evaluate our approaches on a set of websites and discussion forums.

I. Introduction

Freedom of expression is a fundamental human right, as the political discourse requires that all citizens must be able to voice their opinions and interests, and be allowed to discuss any issues pertinent to society. In open democratic societies where freedom of speech exists, misuse of it has traditionally been curtailed by social norms. Hate speech, slander and derogatory language have not been part of public conversation. However, in digital environments, such as Internet forums and social media, physical distance and the possibility of anonymity have in many cases weakened the traditional boundaries of civility [16].

Simultaneously, the views of fringe groups that used to reside in closed environments can now be made available for a worldwide audience. On the Internet, largely anyone can express their views in any way they like, including such ways intended to intimidate and silence the opposition. Hateful messages tend to catch people's attention and also feed more hate, thus drowning out more moderate messages. In the end this fosters a climate of aggression that may spill over into the physical world.

Thus online hate may severely hurt individuals and groups, increase polarization and stifle civil conversation. Moreover, hateful digital environments can form a breeding ground for radicalization, where some participants eventually choose to physically act against their targets. Reliable automatic detection of hate is for several reasons notoriously hard. The possibility of detecting every instance of hate is far away. However, for threat assessment of a digital environment, it would be sufficient to quantify the prevalence of hate expressions in a more general manner. We propose doing this by determining a *hate level* of a given digital environment. This hate level would numerically express the amount of hate found in that environment. If the hate level is computed using a standardized method, hate levels of different environments can be compared in a meaningful way. Moreover, by determining a baseline hate level of the "average internet" we can assess the degree of online hate in more extreme digital environments. Changes over time in the hate level of some environment can be studied by computing the hate level of forum postings in different intervals. An increasing hate level might indicate ongoing radicalization.

Our approach to hate detection is based on transfer learning on pre-trained language models. Transfer learning generally lessens the need for annotated data, which can be difficult to acquire. Basically, a language model representing the common relationships and orderings of words in a given language is created using unsupervised learning. As no annotated data is required for this, large amounts of readily available raw text can be used to produce a language model of good quality. In the transfer learning step this model is then fine-tuned with annotated data to detect hate. The idea behind this is that the language model with its general "understanding" of the language requires less annotated learning examples to recognize a particular aspect such as hate than a more traditional machine learning approach that relies entirely on annotated data.

This paper is outlined as follows. In Section II we describe our interpretation of hate speech and how measuring the usage of hate speech can be used to calculate a hate level for different digital environments. In Section III we present our approaches for measuring hate in digital environments. In Section IV we describe the training of these systems and evaluate them against each other, to select a best approach for hate level computation. In Section V we test our approach by computing

ASONAM '19, August 27-30, 2019, Vancouver, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6868-1/19/08...$15.00
http://dx.doi.org/10.1145/3341161.3343521

the hate level for a number of different digital environments. A discussion of the results is presented in Section VI and finally, some directions for future research are presented in Section VII.

II. Online Hate

Hate as a psychological construct is usually regarded as an emotional state involving a number of different emotions such as anger, fear, contempt, or disgust. The intensity of these emotions and the individual's degree of dedication determine the behavioral outcome, which ranges from avoiding, punishing or -in the most serious form- trying to destroy or annihilate the target [17]. As opposed to emotions such as anger, hate is considered to be relatively stable over time. Further, hate as an emotional state is maintained, even reinforced, by hateful behavior towards one or multiple target individuals or groups. This is another difference from emotions such as anger or sadness, where acting out that can often provide some relief from it.

One kind of hateful behavior is verbal abuse, including threats, libel, cruel and derogatory speech etc. When this kind of verbal hate is expressed in digital environments, it is what we refer to as digital hate. However, discerning expressions of hate from expressions of mere dislike, or just offensive language seems next to impossible in many cases. There are no clear boundaries, and also great individual disagreements as to when a particular expression constitutes hate. In training data sets annotated by multiple reviewers, such as the set established in [2], this is handled by providing detailed annotation accounts, leaving it to the user to decide on resolving disagreements before training, for example by majority vote.

Conceptual difficulties aside, the fact that there is a multitude of ever-evolving ways of expressing hate verbally makes automatic detection of digital hate a difficult task. Methods relying on a dictionary of hate terms fall short as users constantly invent new hate terms, express hate using individually non-hateful terms (e.g. *"I wish someone would shove X in front of the bus!"*), or use hateful terms in an ironic, joking or otherwise non-hateful manner.

As our machine-learning approaches rest on annotated data from different sources, we offer no concise definition of hate beyond the psychological one at the beginning of this section. Pragmatically speaking, our implementations effectively use a definition that is an amalgamation of the many individual definitions that are implicitly found in the annotations of the different data sets. One should note, though, that due to our selection of data sets this effective definition of hate and its expressions is intentionally wider than the common understanding of *hate speech*. The latter is usually reserved for denigrating statements against groups or individuals based on their race, religion, sex or other innate attributes. We on the other hand are interested in hate based on any reason.

A. Related work

Online hate and its detection have become a subject of considerable interest among law enforcement agencies, civil rights organizations and academia. A number of projects are investigating different ways of tackling the problem.

In [5] generalized hate and hate directed at individuals or entities are studied. Directed hate is defined as hate language towards a specific individual or entity while generalized hate is defined as hate language towards a general group of individuals who share a common protected characteristic, e.g., ethnicity or sexual orientation.

In [11] machine learning is used to separate between hate speech, profanity, and other texts. The data set that is used is an annotated data set of tweets. The three class problem was solved with an SVM and three different set of features: character n-grams, word n-grams and word skip-grams. The results showed that distinguishing profanity from hate speech is a very challenging task. The use of bag-of-words approaches tends to have high recall but leads to high rates of false positives since the presence of offensive words can lead to the misclassification of tweets as hate speech, something that was noticed in [10]. Kwok and Wang found that most of the time the reason that tweet was categorized as a racist tweet was because it contained offensive words.

Another example where machine learning is used is Google's Perspective API, which is built for detecting toxic comments online. However, research shows that detecting hate speech is a difficult task and that Perspective can easily be deceived [7].

The *Bag of Communities* approach [1] uses machine learning on posts from different communities to train a classifier for abusive online behavior. The training samples are labeled as abusive or non-abusive, but this is not done according to their specific content. Rather, the selected training communities are each assumed to be either wholly abusive or non-abusive, and this is then extended to all the respective posts. The method achieves a good accuracy of 75 percent, but given the training setup is unclear whether it actually detects abusive language or rather similarity to broader classes of communities. The same paper also presents a classifier trained on 100,000 comments evaluated by human moderators. This achieves an excellent accuracy of over 91 percent, but the source of the moderated training data is kept anonymous and the data itself is not made available. This highlights the importance and difficulty of finding suitable training data of the required sizes, and unfortunately it limits the practical use of the method.

The Online Hate Index[1] (OHI) is a tool developed by the ADL (Anti-Defamation League) and the UC Berkeley, intended to detect and quantify hate speech in online environments. The OHI is trained on 9,000 annotated comments from Reddit.[2] While intended for comparisons between different environments, the OHI project is in an early phase, and at the time of this writing it has not been tested on any other data beyond its training set.

[1] https://www.adl.org/resources/reports/the-online-hate-index

[2] https://www.reddit.com

Since most available data sets that can be used to study hate speech and abusive language are in English, there is a lack of studies of hate in other languages. However, some studies do focus on other languages. Techniques for detecting directed hate in Swedish is described in [14] and in [9]. A study on abusive language on Arabic social media is presented in [12] and a study on hate speech in German in [15]. Despite this there is still a need for more studies on hate speech and offensive language in other languages than English.

III. Models for Hate Detection

We implemented and trained three different models for the detection of hate, one SVM as a baseline and two fine-tuned language models. The purpose of this three-pronged approach was to gain a better understanding of the performance of each method, as we could then compare them on our annotated data with known hate content.

For the baseline we chose an SVM model. The SVM implementation in this paper is based on *scikit-learn*[3] [13]. It is implemented with a linear kernel and the classifier was trained with the square of the Hinge loss function $l = max(0, 1 - t * y)^2$. For feature representation we used traditional BoW (bag-of-words) features. This was simple to implement, and it has been shown in [18] that more complex BoW approaches (bigrams, trigrams etc.) only achieve minor improvements. Thus this most basic version should give us an impression of the general performance of SVM with BoW.

Our first language model of choice for fine-tuning was Google BERT[4] [4]. BERT represents the state of the art and has won a number of competitions in language classification. It is provided with an extensive pre-trained language representation of English, based on Wikipedia[5] and BookCorpus [20], a corpus of 11,038 books. This corpus selection ensures that the pre-trained BERT represents a wide range of English, including more casual and conversational language that is to be expected in forums and other social media. Thus we could use this English model as is for our fine-tuning.

Our second language model is based on the ULMFiT[6] method [8] as implemented in the *fast.ai*-framework.[7] ULMFiT is provided with a pre-trained model for English, based on Wikipedia. As the almost entirely encyclopedic language of this corpus may not be a good match when dealing with other types of texts, the authors recommend pre-training a language model from scratch when needed, and provide some tools for this. Therefore we pre-trained our own English language representation model, based on the Celebrity Profiling Corpus[8] [19], a collection of approximately 74 million tweets from celebrities. ULMFiT has also won competitions, but BERT scores higher in general and is built upon experiences with ULMFiT. Development continues on both methods.

The need to pre-train a model from scratch for ULMFiT may initially appear as a disadvantage, but it does showcase a flexibility that can be useful: Pre-training with the celebrity corpus consisting of about 1.6 billion tokens (fast.ai actually recommends 100 million to be sufficient) required approximately 72 hours on a GTX 1080 GPU, which shows that pre-training new models for other intended environments and languages is relatively easy to accomplish. Conversely, BERT pre-training is more expensive computationally, by orders of magnitude. While BERT comes with tools supporting pre-training from scratch, producing a language representation comparable to the included model would require months of computation on common hardware, and include challenging RAM bottlenecks to overcome. Thus in practice one is likely to be restricted to the included model, which has excellent all-round performance on English, but could face problems in specific environments. E.g. language models by their nature have a limited vocabulary, typically the 30,000 most frequent words in their corpus, and an environment with its own jargon (or even language) is likely to contain terms that are important to its community, but not found in a generalist vocabulary. Thus we were interested in the performance of the more flexible ULMFiT.

IV. Fine-tuning and Initial Evaluation

The three models of this paper have been fine-tuned (or trained in the case of the SVM) with annotated data from two different sources:

- *Twitter*: This social media data set from the SemEval 2019 competition task 5[9] consists of 9,000 samples of Twitter[10] posts (*tweets*).
- *Forum Posts*: The forum posts presented by [3] were extracted from the white nationalist forum *Stormfront*[11] and annotated at sentence level. The data consists of 10,568 samples.

These data sets are provided with annotations indicating whether each sample contains hate or not. The Twitter data is more specific in its conception of hate in that it focuses on hate speech against women and immigrants. The forum data is more general in its understanding of hate, but given the source the hateful comments tend to be directed against ethnicities perceived as non-white and groups regarded as opposing white supremacy.

We trained three variants of each of our approaches, one on the forum data, one on the tweets, and one on both, resulting in nine classifiers overall.

After the fine-tuning/training the nine models were evaluated on subsets of the data that had been excluded from the training. The results are summarized in the Tables I, II and III.

In each table the *training* column indicates which data set the respective classifier was trained on, i.e. the Forum, Twitter, or combined (F+T). The *test* column shows which dataset the

[3] https://scikit-learn.org

[4] *Bidirectional Encoder Representations from Transformers*, https://github.com/google-research/bert

[5] https://www.wikipedia.org

[6] *Universal Language Model Fine-tuning*

[7] https://www.fast.ai

[8] https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html

[9] *Shared Task on Multilingual Detection of Hate*, https://competitions.codalab.org/competitions/19935

[10] https://twitter.com

[11] https://www.stormfront.org

TABLE I
MODEL: BAG OF WORDS WITH SVM

training	test	accuracy	F1 score	precision	recall
Forum	Forum	0.722	0.708	0.745	0.674
Forum	Twitter	0.553	0.526	0.481	0.581
Twitter	Twitter	0.705	0.67	0.64	0.7
Twitter	Forum	0.615	0.505	0.707	0.393
F+T	F+T	0.705	0.676	0.685	0.667

TABLE II
MODEL: BERT

training	test	accuracy	F1 score	precision	recall
Forum	Forum	0.816	0.819	0.806	0.833
Forum	Twitter	0.639	0.601	0.571	0.635
Twitter	Twitter	0.798	0.769	0.763	0.775
Twitter	Forum	0.726	0.731	0.718	0.745
F+T	F+T	0.802	0.796	0.751	0.846

respective classifier was then tested upon. The Forum-based classifiers and the Twitter-based classifiers were each tested twice, once on Forum data and once on Twitter. The combined F+T classifiers were tested on combined test sets.

The results show that with few exceptions the fine-tuned language models outperform the baseline SVM, and BERT outperforms ULMFiT. We also see that hate detection in Twitter appears to be consistently more difficult than detecting hate in our chosen forum. This can in part be blamed on the different notions of hate in the training data: With the Twitter set being annotated primarily for hate speech against women and minorities, hate directed at other targets may lead to detection errors. An indication for this is the low recall of the Twitter-trained SVM and ULMFiT when tested on forum data; these systems apply the narrower definition of hate and thereby miss samples annotated with the broader definition. Interestingly, BERT suffers far less from this phenomenon. As we want to apply our method on a broad range of environments, the combination classifiers (F+T) give us the best idea of an expected general performance. These models can be said to have learned a merged conception of hate, based on the different definitions from the two data set annotations. The F+T variants all come close in performance to the best specialized classifier within their respective base model (in all three cases the Forum on Forum classifier). BERT shows the best performance also among the combination F+T classifiers. Given this we choose BERT F+T for the subsequent hate level computation. We also acknowledge that ULMFiT may be the more practical choice when expanding into certain environments and languages that are not covered by BERT's language model.

TABLE III
MODEL: ULMFIT

training	test	accuracy	F1 score	precision	recall
Forum	Forum	0.771	0.784	0.742	0.832
Forum	Twitter	0.655	0.636	0.579	0.705
Twitter	Twitter	0.722	0.698	0.651	0.752
Twitter	Forum	0.633	0.536	0.73	0.424
F+T	F+T	0.72	0.703	0.673	0.735

V. LEVELS OF ONLINE HATE

There are many places on the Internet where hate speech is a more or less common part of the discussions. By measuring the level of hate present in some digital environments we can get an idea on how much hate speech that is present in each environment and also study differences in the level of hate among different digital environments.

A. Digital Environments

To test our hate level computation we ran the BERT F+T classifier on three different digital environments: Stormfront, The Daily Stormer and Reddit.

1) Stormfront: Stormfront was launched in 1996 by the white supremacist Donald Black and was arguably the first "hate site". Stormfront is still one of the most visited and well-known sites among radical nationalists [6].

Selected texts from Stormfront had already been part of the initial training set as described in Section IV.

2) The Daily Stormer: The Daily Stormer[12] is an American neo-Nazi, white supremacist, and Holocaust denial website founded by Andrew Anglin in 2013. The site attracts a young audince and uses Internet memes and is similar to the imageboards 4chan and 8chan. In 2017 when a victim of a homicide was insulted on the site several domain registrars rejected to host the site and since then the site has changed domain several times.

3) Reddit: Reddit[13] is one of the largest discussion websites with currently 1.2 million different forums called "subreddits". Reddit is home to discussions on virtually any imaginable topic. For example, there is a subreddit devoted to the most disturbing content the internet has to offer, and there are subreddits for all kinds of games and sports.

Subreddits are managed by moderators who can remove posts or content that is seen as unwelcome. This includes content that encourages or incites violence as well as content that threatens, harasses, or bullies or encourages others to do so.

B. Measuring the Level of Hate

For each environment we selected a representative sample of posts, sized to achieve a margin of error below 1% and a confidence level above 99% (i.e. 17,000 posts for each). The

[12] https://dailystormer.name

[13] https://www.reddit.com

model classifies a text as hate or not hate. Table IV show the classification results. The hate level is the percentage of hateful posts in the total sample set.

TABLE IV
HATE LEVEL OF DIGITAL ENVIRONMENTS

	Stormfront	Daily Stormer	Reddit
Non-hate posts	14,606	9,952	15,810
Hate posts	2,394	7,048	1,190
Hate level	14%	41%	7%

It comes as no surprise that both nationalist sites have a higher proportion of hateful comments than Reddit. Perhaps more interestingly, the Daily Stormer scores considerably higher than Stormfront. This can in part be explained by the nature of each site. The Daily Stormer styles itself primarily as a news site with editorial articles from the staff, and the forum exists for discussion of the articles. Thus the forum retains a strong focus on white nationalism. Conversely, Stormfront regards itself as a community forum, and many of its subforums are dedicated to everyday life discussions among its members, featuring threads that would not be out of place in non-nationalist venues.

VI. Discussion

Computing the hate level for our three example environments yields surprising results. The Stormfront forum has gained some notoriety over the years as a foremost online place of racially motivated hate. While our hate level measurement shows that it is indeed twice as hateful as the internet in general as represented by Reddit, the proportion of hateful posts compared to non-hateful ones is still rather low, with approximately six non-hateful posts for every hateful one. The Daily Stormer with its much stronger focus on topics that are bound to attract or incite hate outclasses Stormfront by considerable margin - but even here the majority of posts is not hateful. Of course, our approach makes no attempt to measure the intensity of hate in a given posting. There is merely a binary decision between hate and non-hate. Thus it is possible that the average hateful post on Stormfront is much more threatening or hurtful than the average hateful post on Reddit, or vice versa. We intend to investigate this in the future.

From a more technical perspective, we can see that transfer learning is a viable solution that greatly ameliorates the problem of obtaining the large quantities of training data required by more traditional supervised learning approaches. Our method uses less than 20,000 annotated samples, yet exceeds the performance of the aforementioned *Bag of Communities* [1] model that was trained on several million labeled posts (although the differences between the systems and the respective data mean that one should be cautious about such direct comparisons). Good quality training data remains essential: The low ratio of hateful to non-hateful comments even in a highly toxic digital environment like the Daily Stormer implies that the tempting approach of simply using a very hateful forum as the “hate set” cannot be recommended as a shortcut alternative to actual annotations. Fortunately smaller training sets from varied sources are increasingly easier to obtain. Moreover, our combined classifier trained on samples from both the Stormfront forum and Twitter performs approximately as well on different targets as the respective specialized models. This indicates that fine-tuning a language model using several small training sample sets taken from a multitude of sources is likely to result in a classifier with good general performance on a variety of targets.

VII. Future Work

There are several interesting directions for future work. A natural extension is to measure the level of hate on a larger set of digital environments and also to verify the result manually.

It would also be interesting to study if individuals express more hate over time when they interact on discussion forums where the level of hate is high.

As mentioned in the previous section, another important objective is more fine-grained analysis of the nature and intensity of the hate in a given environment. For example, a cursory review of hateful posts indicates that the hate expressed in the nationalist forums is more hostile and threatening than the average, but it tends to be targeted at individuals and groups outside the respective community. Reddit on the other hand appears to contain less serious hate, but more of it happens directly between members. A deeper study could reveal interesting characteristics of online hate and possibly in the bigger picture some measure of the group cohesion in a digital community, and one approach would be to use techniques from aspect-based sentiment analysis to identify the targets of hate.

The advantages of using language models, as discussed in the previous section, are also beneficial when working with other languages that lack the linguistic resources available for English. Large annotated training sets are even more difficult to procure for such languages. Raw text for the unsupervised creation of a language model on the other hand is relatively easy to obtain, for example from the respective Wikipedia, forums and news sites. The ULMFiT-approach allows training such a language model in time frames ranging from hours to days, on hardware that is readily on hand in academia. If even the small amounts of annotated training data sufficient for the fine-tuning step are not available, manual annotation is feasible at this scale. While the performance of of our ULMFiT classifier lagged behind BERT, training classifiers on different languages to obtain a multilingual hate level would nevertheless be an interesting research direction.

Acknowledgment

This work was funded by a commission from the Swedish government.

References

[1] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3175–3187, New York, NY, USA, 2017. ACM.

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[3] O. de Gibert, N. Pérez, A. G. Pablos, and M. Cuadros. Hate speech dataset from a white supremacy forum. *CoRR*, abs/1809.04444, 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. *CoRR*, abs/1804.04257, 2018.

[6] L. Figea, L. Kaati, and R. Scrivens. Measuring online affects in a white supremacy forum. In *IEEE Conference on Intelligence and Security Informatics, ISI 2016, Tucson, AZ, USA, September 28-30, 2016*, pages 85–90, 2016.

[7] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective API built for detecting toxic comments. *CoRR*, abs/1702.08138, 2017.

[8] J. Howard and S. Ruder. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146, 2018.

[9] T. Isbister, M. Sahlgren, L. Kaati, M. Obaidi, and N. Akrami. Monitoring targeted hate in online environments. In E. Lefever, B. Desmet, and G. D. Pauw, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA).

[10] I. Kwok and Y. Wang. Locate the hate: Detecting tweets against blacks. In M. desJardins and M. L. Littman, editors, *AAAI*. AAAI Press, 2013.

[11] S. Malmasi and M. Zampieri. Detecting hate speech in social media. *CoRR*, abs/1712.06427, 2017.

[12] H. Mubarak, K. Darwish, and W. Magdy. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada, Aug. 2017. Association for Computational Linguistics.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

[15] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In M. Beißwenger, M. Wojatzki, and T. Zesch, editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9, 2016.

[16] C.-L. Sia, B. C. Y. Tan, and K.-K. Wei. Group polarization and computer-mediated communication: Effects of communication cues, social presence, and anonymity. *Info. Sys. Research*, 13(1):70–90, Mar. 2002.

[17] R. J. Sternberg and K. Sternberg. *The Nature of Hate*. Cambridge University Press, 2008.

[18] 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.

[19] M. Wiegmann, B. Stein, and M. Potthast. Overview of the Celebrity Profiling Task at PAN 2019. In L. Cappellato, N. Ferro, D. Losada, and H. Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, Sept. 2019.

[20] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society, 2015.