Can Hate be Explained?
Investigating Generalisability of Hate Speech Detection using HateXplain

Student numbers: 21134487, 21169262, 21169259, 21111357

Group 20

Abstract

Building Automatic Hate Speech detection models is pivotal from the perspective of how social media platforms have been used to effectively reinforce violence and persecution of minorities in the recent past. However, predictions from such models have largely been found to be biased with poor interpretability and generalization. The aim of this report is to juxtapose the generalisation of two pre-trained models (BERT and DistilBERT) trained on the HateXplain dataset. After fine-tuning the model, we will test their generalisation performance on a second hate-speech dataset, ETHOS.

1 Introduction

Facebook alone removed 31.5 million hateful posts in Q2'2021, a 200 percent growth from Q1'2020, with automatic detection of 90 percent before users reported it(tra), but Hate Speech detection still has a long way to go, as before something can be reported and investigated, the consequences can transmute into real crimes, and culture distortion(Williams et al., 2020)(Appendix section B).

This is despite the fact, that the field has seen many advances, ranging from new datasets for training and detection(Ousidhoum et al., 2019),(Qian et al., 2019),(de Gibert et al., 2018),(Sanguinetti et al., 2018), (Mollas et al., 2020) shared tasks to effectuate the inductive transfer and learn better representations, to models that claim "state-of-the-art performance" on such datasets (Arango et al., 2019),(Gröndahl et al., 2018). However, pace to the reality of "successful detection", models trained on such datasets often fail to generalize on unseen data, erroneously learning a "trigger vocabulary" or associations to make biased predictions for commonly targeted identities such as non-binaries, black, and Muslims (Sap et al., 2019),(Davidson et al., 2017).

More lucidly phrased, the biased predictions can often result in non-toxic comments being labelled as hate speech, while unseen, hateful comments, about some targeted identities, can pass on without being labelled as "hateful"(Borkan et al., 2019).

The paper aims to test recent advancements to mitigate the lack of generalization, in line with the suggestion that "data-preparation" is more pivotal than "modelling"(Yang et al., 2019) and that augmentation alone is not sufficient if the "strong" features are hard to learn and can rather be counterproductive (Jha et al., 2020). We do so by using HateXpain, a dataset that takes into account robust data preparation to look at not only the commonly used 3-class-classification (hate, offensive, or normal), the target community (identity targeted for abuse), but also the rationales behind labelling of such data being labelled as the classifications by using a "human in the loop" (Amazon Mechanical Turk workers)(Mathew et al., 2020)

First, we use pre-trained BERT and Distilbert models, training them on HateXplain, and then testing them on another binary classification dataset ETHOS(Mollas et al., 2020), later we introduce two data-augmentation experiments, to test the hypotheses if "strong features" can be learnt to improve generalization on unseen data.

The idea is to investigate if data preparation with rationales makes hate more "explainable" to the model, resulting in learned representations that don't overfit on some "trigger vocabulary".

2 Literature Review

2.1 The road to Transformers: An overview of the architectural choice

Parting ways from a rather "ossified" architecture, locked into a stringent reliance on the order of input sequences to learn representations either from recursions (RNNs\LSTMs) (Hochreiter and Schmidhuber, 1997),(Sutskever et al., 2014),(Bahdanau et al., 2014) and convolutions (Conv2S) (Gehring et al., 2017) based on sequence-

to-sequence encoder-decoders, Transformer neural networks have provided appreciable gains in tasks related to machine translation (Lakew et al., 2018), large scale speech recognition (Karita et al., 2019) while providing faster training times and better computational efficiency achieved through simultaneous parallelization.

By “ossified” it is implied how the “stringent reliance” of sequence-to-sequence encoder-decoder in RNNs takes a token-at-a-time approach in both encoders (for generation of fixed-length state vectors) and decoders (token-at-a-time reading to predict the target and recursion on all previous positions). This “ossified” approach is a computational impediment for long input sequences and limits batching (Vaswani et al., 2017) albeit it was partly solved using Attention.

In contrast to a simple sequence-to-sequence encoder-decoder approach, attention can be lucidly described as a “human-inspired” tendency to look “attentively” at details and focus only on what is relevant rather than looking at the whole fixed-length state vector effectively training for a better log probability of the related prediction(Vaswani et al., 2017)

Transformers differ from RNNs and even other convolutional neural networks(CNNs) that have tried to reduce sequential computation by computing hidden representations parallelly such as ByteNet (Kalchbrenner et al., 2016), Conv2S, and Extended Neural GPU (Kaiser and Bengio, 2016) by reducing the number of operations to relate signals from two output or input positions to a constant number of operations. Transformers achieve this efficiency in sequential computation by using “multi-headed attention”, employed in three different ways: first in a fashion similar to the sequence-to-sequence encoder-decoder network; second by using self-attention layers in the encoder; third by using self-attention layers in the decoder as well. This allows for each position and both encoders and decoders to attend to all positions in the previous layer.

Cogently summed, it is analogous to being able to “attentively” find associations between two different positions while being “attentive” at each position to learn a specific representation.

All these advantages lead to significant improvements and inform our choice of models for the experiments, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and Distilbert (A distilled version of BERT) (Sanh et al., 2019).

Essentially, with our choice we wanted to ask, if a larger model with more parameters, with a similar pre-training did really perform better in "generalization" ? did this reduction in parameters affected the "trigger vocabulary" aspect without learning the context? And what role did augmentation play with the two models?

2.2 "State of the Art": But can it Generalize?

A succinct history of training the supervised classification task (Schmidt and Wiegand, 2019) reveals efforts from using classical methods with feature engineering (Coroiu, 2019)like SVM (Mehdad and Tetreault, 2016), logistic regression, Naïve Bayes, and random forests (Xiang et al., 2012), (Davidson et al., 2017) to using neural networks. Interestingly, even the classical methods can achieve high accuracies but do not generalize well, and essentially can be biased on some “target words” leading to inaccurate classifications.

Deep Learning methods(Goodfellow et al., 2016), on the other hand, can learn representations of the data automatically using a multi-layered structure that can learn abstract feature representations (Zhang et al., 2018).“State-of-the-art” architectures include RNNs (in particular, LSTMs)(Sutskever et al., 2014), CNNs (Gehring et al., 2017), Gated Recurrent Units (GRUs)(Zhang et al., 2018), and more recently Transformers(Vaswani et al., 2017). The usual CNN, RNN approach does comprise of a common sequence-to-sequence encoder-decoder architecture, the key difference often being CNNs use a bag-of-words approach or n-grams and RNNs employ LSTM cells and recursion to capture dependencies between the words in an instance of hate speech.

The claimed F-score using these methods is impressive or “state-of-the-art”, CNNs can achieve a near 80% score while RNNs can get a score close to 90% (Arango et al., 2019). But when it comes to generalization, that is, adapting to unseen data and being able to classify that into hate speech, offensive language, or normal expression, by finding words from unseen data that are similar, have semantic similarities, and understand connotations and sentiment related to the particular text that is being classified, these models falter.

For example, a "state-of-the-art" deep-learning

model (Badjatiya et al., 2017) drops from a claimed precision, recall, and F-score ∼90% to an average of ∼64% precision, ∼56% recall (with one class dipping to 12.5%, and ∼49% recall. Another model by (Agrawal and Awekar, 2018) saw similar drops too(Arango et al., 2019).

Machine learning literature is rife with comprehensive analysis that tries to dissect the lack of generalizability, with different approaches from the perspective of data augmentation(Jha et al., 2020) mostly concluding on the importance of annotation(Waseem, 2016), defining more metrics to check over-fitting (Arango et al., 2019).

Intuitively, a more linguistics oriented understanding can be taken from Dr. Chomsky's explanation of how humans converge linguistic generalizations(Chomsky, 1965) (Chomsky, 1980)that are similar even with largely different linguistic inputs they encountered in their childhood , while models do this in mercurial ways (McCoy et al., 2019).

Therefore, we test one such dataset prepared to counter the limitations arising from annotations using human rationales to classify target classes and related social groups by employing human annotators, HateXplain (Mathew et al., 2020). We test the trained data on another dataset albeit without rationale but a similar annotation standard, the ETHOS(Mollas et al., 2020) dataset. We also test the hypothesis if augmentation can improve generalization and if yes, with what sensitivity?

3 Methods

3.1 Explaining Hatred

Our primary question is to understand, does "explaining" hatred, through a dataset that has been annotated differently using "rationale" effectuates better hate speech detection? And our reason for this is to inquisitively look at an improvement in "generalizability", for humans "generalize" and converge such experiences very differently in a linguistic sense which we alluded to in the previous section.

A comprehensive linguistic analysis though is beyond the scope of our discussion but if we were to do a cursory analysis, we will find two broad categories in the hateful content online, directed or generalized. By "directed" we mean that it can specifically aim to target an individual and refer to them in first-person, with hateful content, expletives, and threats. However, when it is more generalized, it could refer to a more "collective" identity with generally hateful comments, scapegoating rhetoric, xenophobia, and more (ElSherief et al., 2018).

Thus, from the cursory analysis, it seems that hatred requires a more "nuanced" approach to "explaining" hatred and this informs the choice of our dataset, HateXplain. The dataset rather than looking at hateful/not-hateful in a dichotomous way and using just one annotator uses annotation in three types of annotation in what authors describe as "human attention" using data collected from two platforms where such studies have already been conducted, Twitter and Gab.

"Human attention" refers to how the human in the loop, that is Amazon Mechanical Turk workers have annotated the dataset by giving "rationales" to decide why they considered something hateful. The first annotation is a three-class label, deciding whether a speech is hateful, offensive, or normal. The second annotation looks at the target group of the associated text which could be related to race, religion, gender; sexual orientation, or be miscellaneous. The third annotation is to look at why a certain context was considered hateful, that is, what phrases or words were conclusive in suggesting if something is hateful or not.

Since the annotation is crowdsourced (Amazon Mturk), quality is ensured by not only ensuring a HIT approval threshold (task completion rate) of 95% on a minimum of 5,000 tasks. To select a good quality annotator, all annotators were first given a pilot annotation task, followed by the main annotation task only if the performance was satisfactory. In total the dataset contains 20,148 posts annotated by 253 workers selected from 621 who took part in the task. In all, each post was annotated by three workers and the label was decided using the majority vote count. The dataset overall has Krippendorff's alpha of 0.46 for inter-annotator agreement, much higher than other hate speech datasets(Ousidhoum et al., 2019),(Del Vigna12 et al., 2017).

A description of the Ethos dataset can be found in the appendix(section C).

3.2 The Hypothesis

To test our hypothesis, which stems from the conclusive literature around looking at data preparation, and emphasis on annotation for hate speech detection generalizability rather than "state-of-the-art" architectures, we choose pre-Trained BERT

and Distilbert as our models of choice to train on HateXplain because of their architectural advantages which allow them to capture contextual and syntactical relationships due to multi-headed attention and the success of these two particular models from previous related work and similar tasks. To test the training performance, we test the claimed performance of the dataset on a similar dataset although much smaller in size but largely similar in annotation approach. Our primary reason is to test if the claim of "generalizability" really stands and to what extent is it able to generalize on unseen data.

We further try to experiment and use data augmentation and wonder if it can really affect the performance if strong features are present as concluded in previous research.

After we proceed from one end of these experiments, we reverse the direction and train on Ethos and test the models on HateXplain, testing the if annotations, even without rationales, despite coming from a relatively much smaller dataset can help representations effectively that can generalize well on unseen data.

3.3 Overall Framework, Evaluation, and Metrics

We fine-tune our pre-trained models BERT and Distilbert according to the associated datasets and try to achieve similar training performances as cited by the authors in HateXplain and Ethos. A comprehensive account of the pre-training details of BERT and Distilbert along with a comparison can be found in the appendix (section A Figure 21 and Section C).

Our framework essentially comprised of training the model on a subset of the given dataset we used for training, then using the embeddings of the sentences, we train the models on the subset of the training dataset and then use the trained model to check the generalizability on a different dataset in its entirety.

In the process of training, we perform several experiments which are explained in the next section and check the sensitivity of evaluation metrics to them.

For our evaluation, we use four different metrics, Accuracy, F-Score, Precision, Recall.

4 Experiments

The overall idea was to seek a solution to the "trigger vocabulary" problem we have talked about in the previous section and if that could be achieved by some sort of augmentation in the existing data, or if generalizability depended on how the data was prepared. We tried different approaches for augmentation, sentence length, and random deletion to see if augmentation of this kind effectuated context-based learning rather than over-fitted response that springs up a conclusion the moment words belonging to a certain target group come up (Our model parameters can be found in Section F of the appendix).

4.1 Sentence Length

Our first experiment was to test the effect of the fine-tuning BERT and Distilbert on different sentence lengths and to see if this affected the generalization while testing on the same dataset. For fine-tuning, we picked a certain threshold of sentence length after visualizing the sentence length distribution and then tested both the models on sentences above this threshold on HateXplain and then later on the Ethos dataset. The largest sentence in HateXplain consisted of 165 words while the smallest consisted of 2 words. More granularly, 12-word sentences fell in the 25^{th} of percentile of sentence lengths, 21-word sentences at the 50^{th} percentile, and 34-word sentences at the 75^{th}percentile. The overall sentence-length distribution had a positive skew, as seen from the figure.

Both BERT and Distilbert were fine-tuned on sentences less than the 12 words, 21 words, and 34 words for 5 epochs. The trained models were tested on both HateXplain and Ethos dataset.

4.2 Random Deletion

Random deletion is a common data augmentation technique in NLP and involves deleting words from sentences in the training dataset according to a probability P. The probability P is considered an augmentation parameter α (Wei and Zou, 2019).

In this experiment, BERT and Distilbert were fine-tuned on HateXplain after randomly masking some of the words. Random masking was applied using α the range of $\alpha = 0.05, 0.10, 0.25, 0.50, 0.75$ of the sentences. The trained models were tested on the Ethos dataset to test for generalisability.

4.3 Testing on different types of Hate Speech

In section 2.2 we looked at how some of the "state-of-the-art" architectures achieved abysmal scores for some sub-categories or classes when tested for generalizability. To check if there were such drops in performance, we extracted sub-categories (gender, racism, violence, etc.) from the Ethos dataset and created a new dataset for each category to test for category-wise generalizability.

The models resulting from the different configurations in the random deletion and the sentence length experiments were tested on each sub-category of hate speech.

4.4 Training on Ethos and testing on HateXplain

HateXplain dataset claimed to be a benchmark hate-speech detection training dataset, particularly in the vein of implementing what has been suggested by various researchers from the perspective of improving generalizability through annotation. A larger crowd-sourced method of annotation which we discussed in the previous section is a major differentiating factor along with "rationales". For our final experiment, we wanted to test this claim by fine-tuning both of our models on ETHOS and then testing them on HateXplain, to see if a model, that followed a similar but less stringent approach to annotation while being much smaller and not using "rationales" was as effective for training as HateXplain.

Since ETHOS is a much smaller dataset and looking at our own results from the previous experiments, we do the training on Ethos without any data augmentation followed by testing on HateXplain.

5 Discussion of Results

5.1 Results

5.1.1 Sentence Length

The performance for Distilbert showed a rather decreasing trend after being trained on sentences of more than 12-words, that is the 25th percentile, as can be verified from the drops in F-Score and Recall from the plots (Figure 2). However, the performance, egregiously enough, was better than the Distilbert trained on the entire dataset and dropped below 0.62 for both recall and F-Score after training on sentences up to 34-words of length. The minute gain of F-Score and Recall from 0.62 to roughly around 0.63 (for both) is nothing significant but seems rather odd as to how the smaller, more compact Distilbert with much less parameters showed better performance when only trained on smaller sentences. The precision and accuracy peaked at the 50th percentile, that is, for sentences with up to 21 words, and dropped thereafter (Figure 19 and 20 appendix).

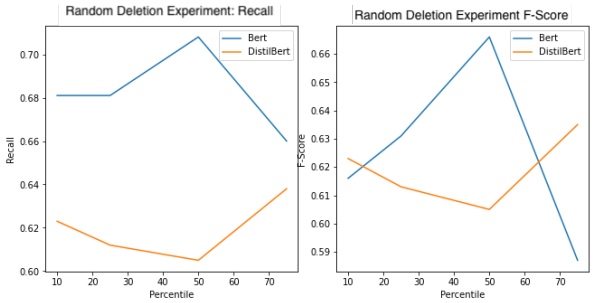


Figure 1: Random Deletion: Recall (Left) and F-Score (Right)

Bert model had around 20% better performance (F-Score) than Distilbert (without augmentation) when trained on the whole HateXplain dataset and tested on Ethos. BERT dropped to an F-Score of 0.63 but more than Distilbert when trained on sentences with 12 words or less and showed a minute gain when trained on sentences with up to 21 words (around 0.64) but increased sharply to around 0.73 when trained on larger sentences with up to 34 words. The performance however remained lower than the BERT trained on the whole dataset consistently. Recall, precision, and accuracy for BERT showed a similar trend (Base performance without augmentation for both BERT and Distilbert can be found in Figure 21, Appendix Section A).

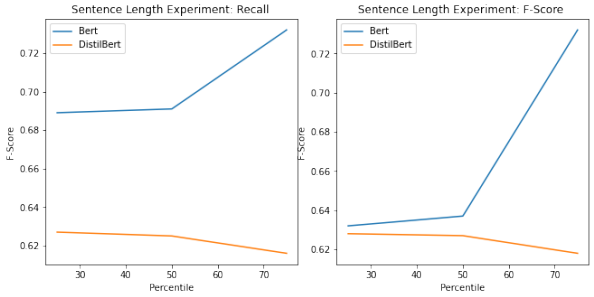


Figure 2: Sentence Length: Recall (Left) and F-Score (Right)

5.1.2 Random Deletion

The opposing trends continued with random masking as well, though this time, BERT and Distilbert came out to be absolute mirror opposites (Figure 1).

On removing 10% of words from a sentence randomly Distilbert maintained an almost similar performance (F-Score and Recall) to the one trained on sentences without any masking. On increasing the random masking percentage to 25%, Distilbert showed minimal drops but dropped to a minimum for all the metrics when the random masking was increased to 50%. But rather oddly, the performance saw an increase to around 0.64 for F-Score, Precision, and Recall when around 70% of the words were randomly removed outperforming the model trained without any deletion (For precision and accuracy refer Figure 18 and 19).

BERT with random masking didn't outperform the one without but reached a peak F-Score and Recall (around 0.66 and 0.71) for the experiment when 50% of the words were randomly deleted. The performance dropped sharply when 75% percent of words were randomly deleted, reaching a minimum for all the metrics.

5.1.3 Testing on different types of Hate Speech

For our test on different types of Hate Speech, we didn't notice a huge variation in how the performance generalized for different classes. Verily, it was better than what previous research has looked at in trying to test automatic hate speech detection with huge, aberrant drops in some classes as discussed in section 2.2.

However, as discussed in the first two experiments, BERT significantly outperformed Distilbert across the classes. The highest F-Score of 0.671 and corresponding Recall of 0.715 were achieved for the violence class and the lowest F-Score of 0.548 with a corresponding recall of 0.616 (for the best performing sentence length experiment with sentences up to 34 words). Overall, this was an 11% drop in performance with the best performing class (Figures 3 and 4).

With random deletion as well, the Violence class generalized the best although the drops were close to 15% when compared to the best performance in sentence length. More interestingly, what we noticed was that violence achieved an F-score of 0.52 with 25% words deleted and 0.57 with 75% words deleted with 0.57 being the highest F-Score for the masking experiment with a drop of over 26% in the worst performing class of disability (F-Score 0.479). Without any augmentation, the performance drop was 16% and the best performing and worst performing classes remained the same, Violence and disability for all the experiments in all scenarios, even without augmentation.

Sentence length also saw a much larger drop of around 24% in the best performing experiment, that is, in training with sentences of up to 34 words

Distilbert saw a performance lower than that of Bert but a 14% drop in performance (F-Score) with the random deletion experiment with violence being the best performing class and disability being the worst-performing class. Also, interestingly, since this wasn't a multi-class classification rather extracting each class and making a new data-set, Distilbert behaved rather oddly for sentence length and had the same scores for all the classes no matter what sentence length it was trained on.

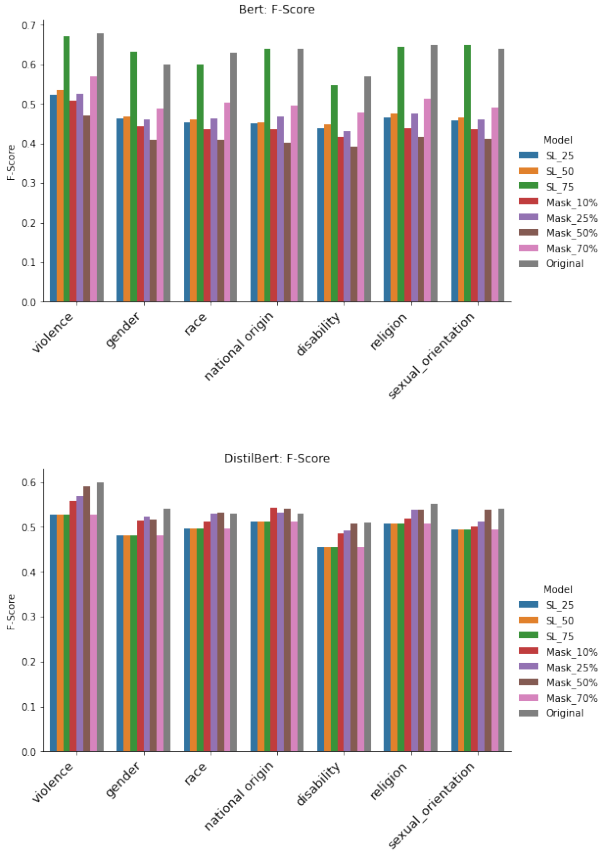


Figure 3: Bert F-Score (Up) and DistillBert F-Score (Bottom)

5.1.4 Training on Ethos and testing on HateXplain

Perhaps the most interesting finding of the entire experiment was how differently the two models behaved after being trained on a much smaller dataset. Bert was significantly outperformed by Distilbert, with Bert dropping to an abysmal precision of around 0.18, F-Score of 0.25, and a recall of 0.50. Distilbert fared much better as can be seen from Figure (Figure 5) with all metrics hovering

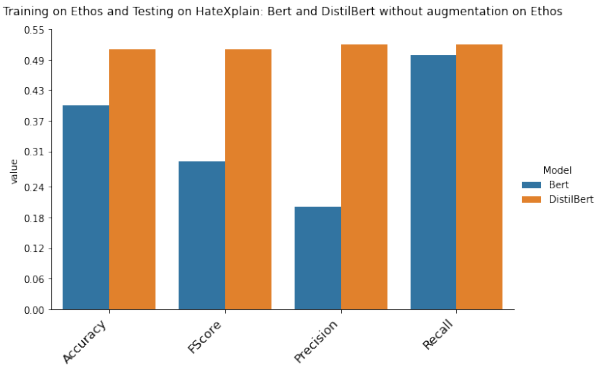


Figure 4: Training on Ethos and testing on HateXplain: Bert and DistillBert without augmentation on Ethos

around 0.50, an almost 100% better performance overall for the F-Score.

5.2 Discussion

Overall, the results were consistent with different arguments we tried to explore to construct our hypothesis in sections 1 and 2. The Bert Model, pre-trained on a much larger corpus, performed significantly better, owing to the number of representations and rationales it could learn on a much larger dataset. Though we fine-tuned both the models but Distilbert's lighter 40% fewer parameters didn't necessarily result in preserving the 97% performance(Sanh et al., 2019) which could have possibly been a limitation of not having the optimal hyper-parameter tuning. In generalization on a different dataset, it showed significantly lower performance as can be seen from Figure (Figure 5).

Curiously, with both the augmentation techniques, Distilbert responded better to the shorter end of the representations to lean the associations faster. It can be hypothesized that perhaps smaller sentences had "strong" features for the Distilbert to learn but the findings weren't consistent with what Bert showed. It could perhaps then be simply a scenario of learning "trigger" vocabulary and then learning classifications for such words from that "trigger" vocabulary(Sap et al., 2019),(Davidson et al., 2017),(Jha et al., 2020),(Yang et al., 2019).

The "trigger vocabulary" scenario is perhaps explainable from both sentence length and random deletion perspectives, especially from the random deletion experiment where a random removal of 70% percent words resulted in the best performance overall for Distilbert.

For BERT the performance rather followed the trend that we talked about in section 1, that "data-preparation" was more important than augmentation. For example, if there were no "strong features" or associations to learn, especially for a task like hate-speech detection that finds its representations, and rationales emanating from a complex sociological context, it would be hard for a model to "learn" and "explain" hatred.

For the multi-class classification test on Ethos, the results followed the same trend of scores for labels as they did with models tested without any augmentation. The violence class had the best performance and disability the lowest. This brings us to another question we had in mind while looking at the results, how is one kind of hatred better explained than the others?

Perhaps the answer lies in how annotation is done for one task and also the "strong" features available for it in terms of how the data was prepared for HateXplain (Section 3.1). A direct follow-up of this result can actually be seen from how Bert attained abysmal scores when trained on Ethos, a much smaller dataset with a similar annotation style but much less "explained" in the vein of how HateXplain is.

The much larger crowd-sourced voted annotation along with a quality check on the quality of annotators and inclusion of rationale for the model interpretability did really show a much better performance for our Bert model on Hatexplain than the one trained on ETHOS. So, does HateXplain explain hate better? Certainly, but it still doesn't reach a "state-of-the" art performance as claimed for previous models available in the literature (discussed in section 2.2).

Distilbert however, is able to perform much better than BERT but the performance is 20% less than that of what was achieved by training on HateXplain in terms of generalization. Whereas, HateXplain saw drops of around 200% here confirming the hypothesis around data preparation and annotation. Though Distilbert does perform better than BERT, a recall and F-Score of around 0.5 isn't really exceptional to be conclusive of anything. Rather, for a task like hate speech, perhaps having fewer parameters did make Distilbert more prone to a "trigger vocabulary" which definitely allows it to put up not-so-bad, hovering around 0.50 scores in evaluation metrics but doesn't really conclude that the fine-tuned, trained model is a "state-of-the-art" performer.

But would it be wise to say that BERT wasn't

prone to the "trigger vocabulary" problem? The confusion matrices (Figures 7-11, appendix section A) show that BERT consistently produced higher false positives throughout all but also got slightly lower false negatives, which can be very important even in small percentages given social media posts are in millions and wrongly classifying a post as non-hate speech when it is can have serious consequences.

BERT only performs better in comparison to Distilbert and though our transformer-based architecture did generalize better in comparison to previous research with different architectures it is still prone to the same problems though they don't seem as amplified because a better generalization is achieved.

5.3 Caveats

Our discussion can be seen as conclusive in some sense, but the performance of models trained on datasets like HateXplain needs to be tested on multiple other datasets to find out if the generalization achieved is consistent with what we observed on Ethos.

Different architectures, even the most popular choice for such tasks, SVM and even slightly older "state-of-the-art" neural architectures like LSTMs(despite the computational issues and efficiency problems) or Conv2s need to be compared with transformer based architectures to assess the role of architecture. This would be important in assessing to what extent is an architecture responsible for learning the "explained" hatred, for efficiency can be one argument but for the task at hand generalization on unseen data can prove useful for research in juxtaposing and coming up with better architectures to look at the interdisciplinary nature of understanding hatred, which we discuss in the next section.

6 Conclusion

6.1 So did we explain hate?

Throughout our discussion, we didn't merely meander through one development after the other but tried to look at automatic hate speech detection from a rather "human" lens especially focusing on the idea that data-preparation was important but again the very argument of "data-preparation" is girdled by a bit of ambiguity here.

Hate-speech detection is not merely learning associations or patterns and concluding if something is hateful or not hateful, it's something very interdisciplinary and while a usual approach can be to treat sociological context, linguistic aspects, and how polarization works in the context of global, national, or local events that shape them, especially for generalized hatred against some groups as a "black-box", hatred encompasses multiple domains and how it interacts with human memory, experience.

While HateXplain takes a much-needed approach to solve the cognitive bias in even identifying what's hateful and with what rationale which is consistent with cognitive research from the perspective of confirmation bias in large groups and how it can be corrected only through common rationales that minimize confirmation bias(Mercier, 2017), it still doesn't look at other interdisciplinary research available from a psychological/neurological (Sapolsky, 2018),(Zeki and Romaya, 2008) and representation perspective(Chomsky, 1965),(Chomsky, 1980) which in turn could bolster how these datasets are made.

6.2 And the architecture? What next?

The architectural choices for our project were only to test the generalizations with the best available solutions referenced from prior research in section 2.1. Certainly, there are caveats to our findings, and with more depth of research we can find how the current architectures really stack up as "state-of-the-art" when juxtaposed to a slightly more modified approach where we introduce multi-task learning and not look at hate-speech as an isolated task.

An approach to look at hate-speech detection using Emotion and Sentiment analysis as auxiliary tasks has already been explored with promising results(Plaza-del Arco et al., 2021) but a major drawback of how the resource requirement amplifies especially because related tasks require annotation and perhaps with stringent quality, yet are more "human-inspired" in a way even in context of these representations are recognized by humans in an inter-sectional way, something we discussed in the previous subsection.

References

2022. Meta platforms inc statistics 2022: Revenue, users, acquisitions amp; shares.

Samira Abnar and Willem H. Zuidema. 2020. Quantifying attention flow in transformers. *CoRR*, abs/2005.00928.

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer.

Aymé Arango, Jorge Pérez, and Bárbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

BBC. 2018. Facebook admits it was used to 'incite offline violence' in myanmar.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Jill Burstein, Christy Doran, and Thamar Solorio. 2019. 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). 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)*.

Noam Chomsky. 1965. Aspects of the theory of syntax cambridge. *Multilingual Matters: MIT Press*.

Noam Chomsky. 1980. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15.

Alexandra Coroiu. 2019. The generalization performance of hate speech detection using machine learning. B.S. thesis, University of Twente.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rohan Jha, Charles Lovering, and Ellie Pavlick. 2020. Does data augmentation improve generalization in nlp? *arXiv preprint arXiv:2004.15012*.

Łukasz Kaiser and Samy Bengio. 2016. Can active memory replace attention? *Advances in Neural Information Processing Systems*, 29.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Raksha Kumar. 2022. As religious riots grow in india, critics accuse facebook of fanning the flames.

Surafel M Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. *arXiv preprint arXiv:1806.06957*.

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 preprint arXiv:2012.10289*.

R Thomas McCoy, Junghyun Min, and Tal Linzen. 2019. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

H. Mercier. 2017. *The Enigma of Reason*. Harvard University Press.

Leo Mirani. 2015. Millions of facebook users have no idea they're using the internet.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328*.

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.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. *arXiv preprint arXiv:2109.10255*.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2019. Learning to decipher hate symbols. 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 3006–3015, Minneapolis, Minnesota. Association for Computational Linguistics.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and A Noah Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Robert M Sapolsky. 2018. Doubled-edged swords in the biology of conflict. *Frontiers in psychology*, page 2625.

Anna Schmidt and Michael Wiegand. 2019. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain*, pages 1–10. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.

Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. Data augmentation for bert fine-tuning in open-domain question answering. *arXiv preprint arXiv:1904.06652*.

Semir Zeki and John Paul Romaya. 2008. Neural correlates of hate. *PloS one*, 3(10):e3556.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

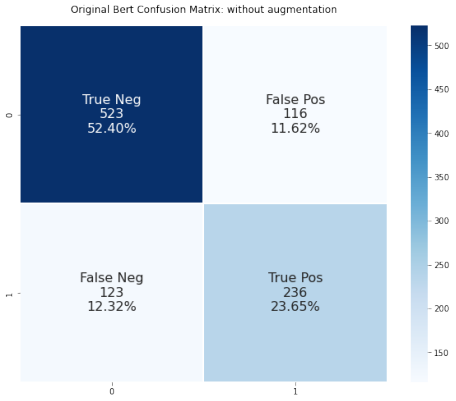


Figure 6: Original Bert Confusion Matrix without augmentation

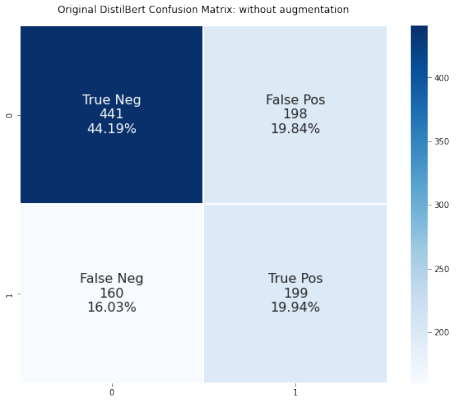


Figure 7: Original DistillBert Confusion Matrix without augmentation

A Appendix

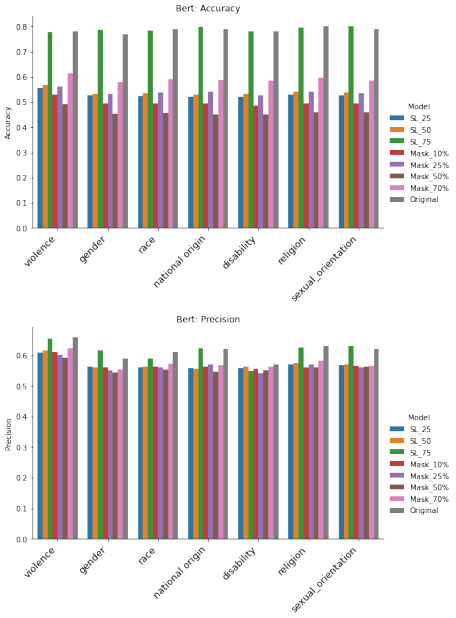


Figure 5: Bert Accuracy (Up) and Bert Precision (Bottom)

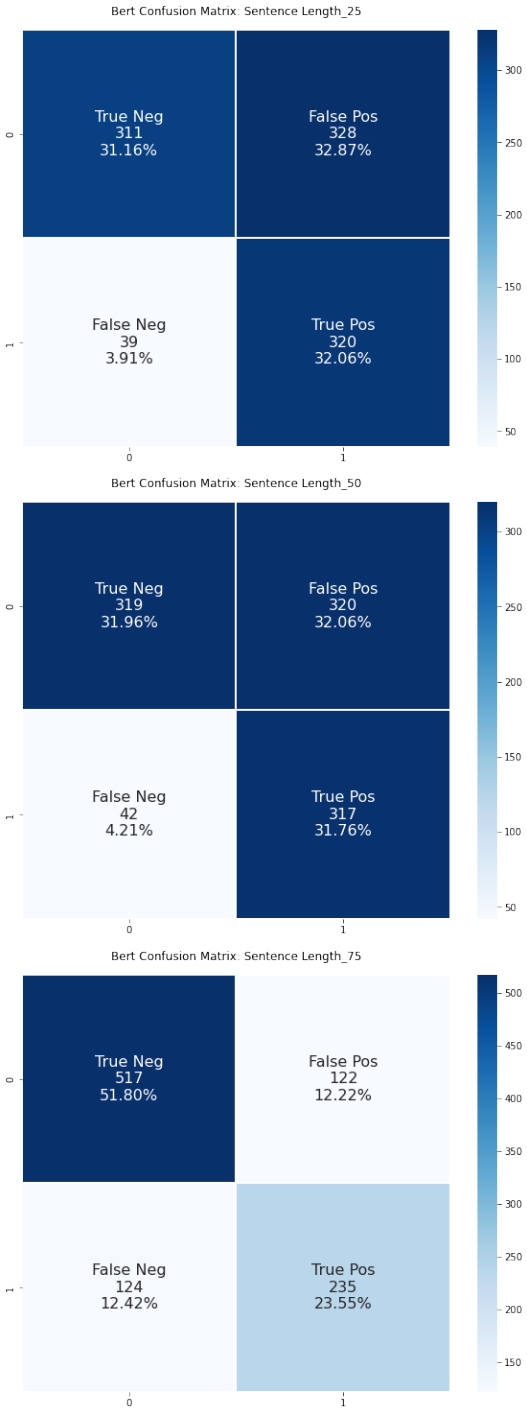
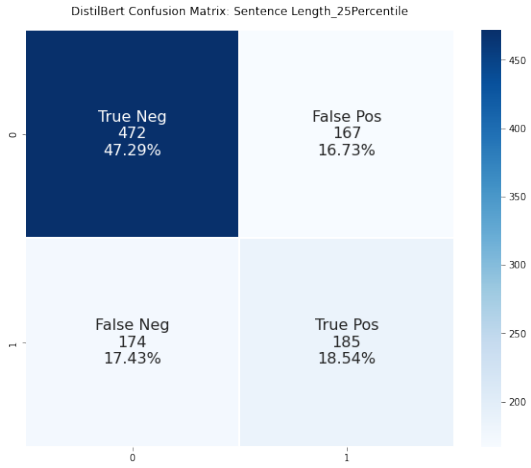
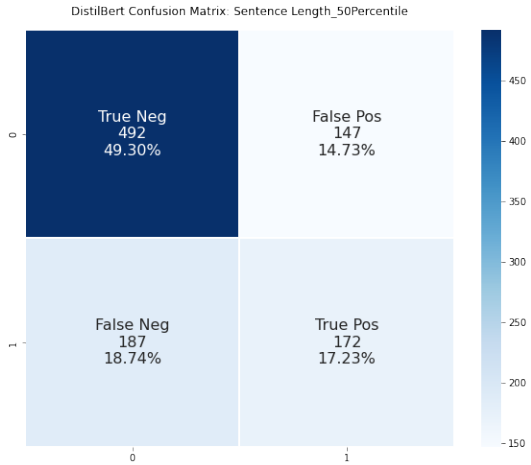


Figure 8: Bert Confusion Matrix: Sentence Length 25 % (Up), Sentence Length 50 (Center) and Sentence Length 75 % (Bottom)





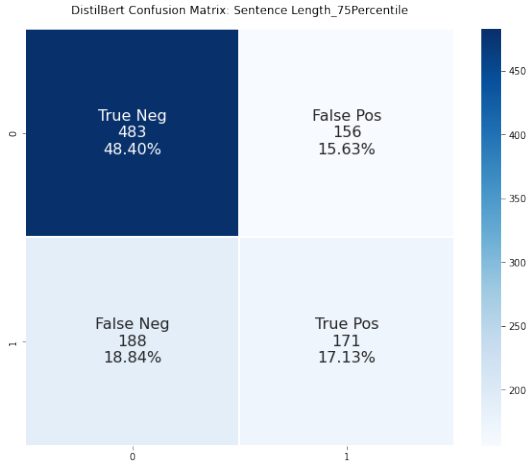
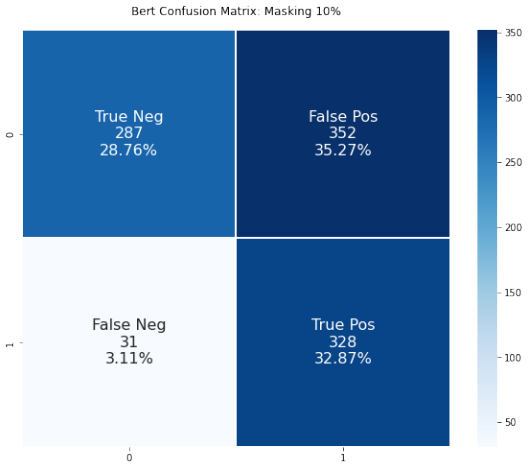


Figure 9: Distill Bert Confusion Matrix: Sentence Length 25 % (Up) Sentence Length 50 % (Center) and Sentence Length 75 % (Bottom)



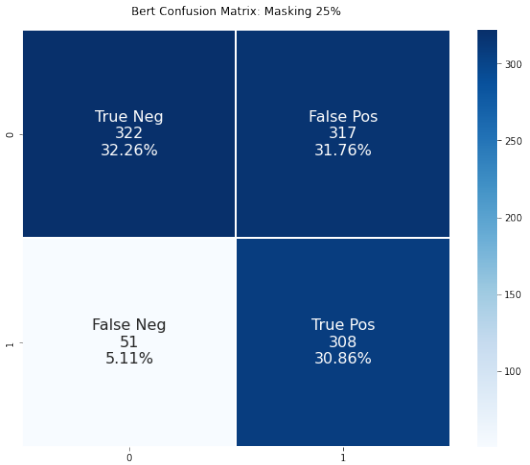


Figure 10: Bert Confusion Matrix: Masking 10 % (Up) and Masking 25 % (Bottom)

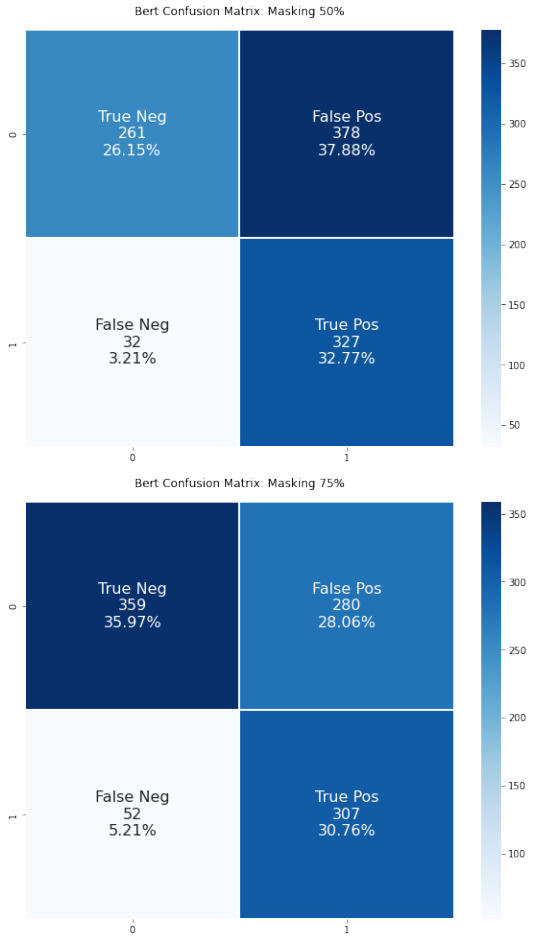


Figure 11: Bert Confusion Matrix: Masking 50 % (Up) and Masking 75 % (Bottom)

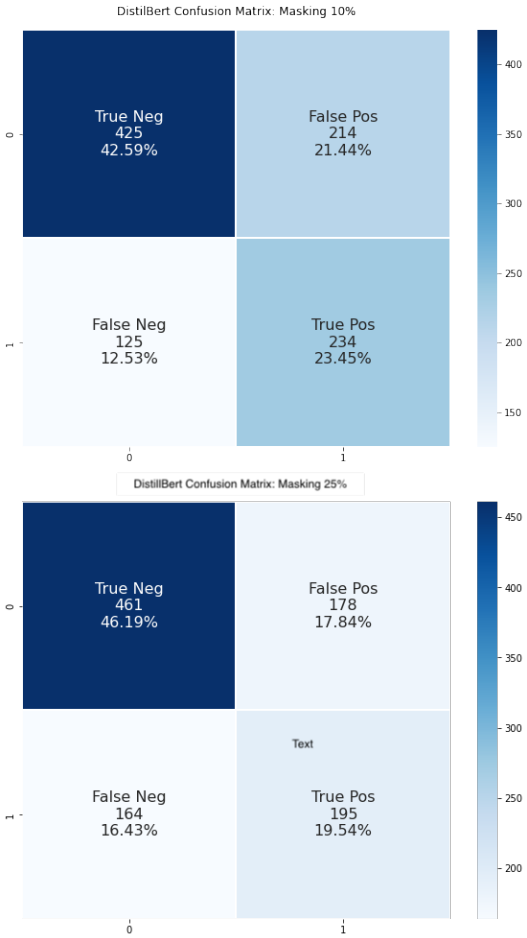


Figure 12: DistillBert Confusion Matrix: Masking 10 % (Up) Masking 25 % (Bottom)

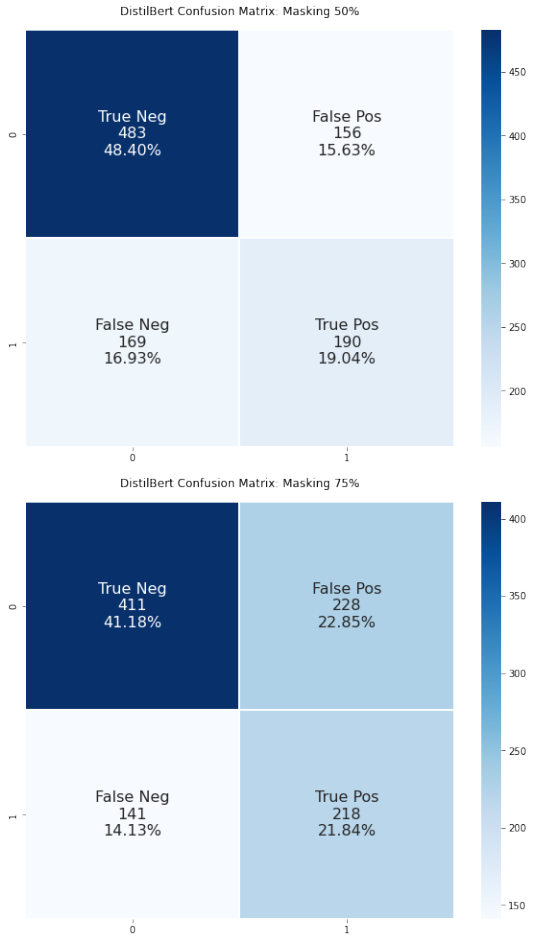


Figure 13: DistillBert Confusion Matrix: Masking 50 % (Up) Masking 75 % (Bottom)

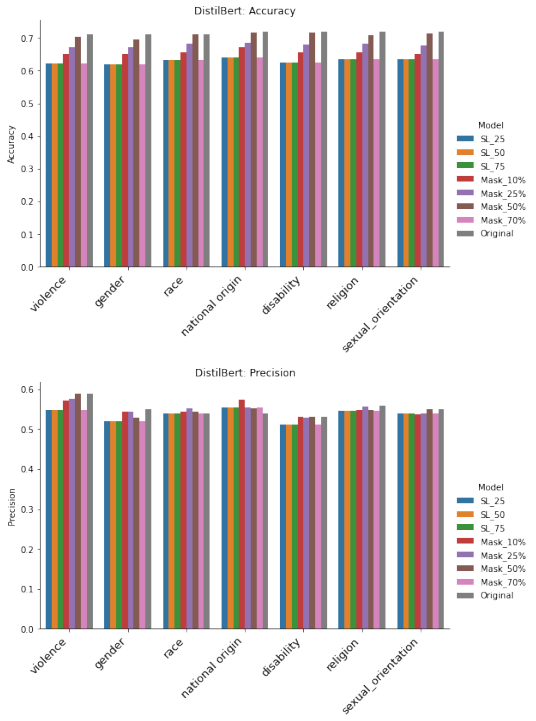


Figure 14: DistillBert Accuracy (Up) DistillBert Precision (Bottom)

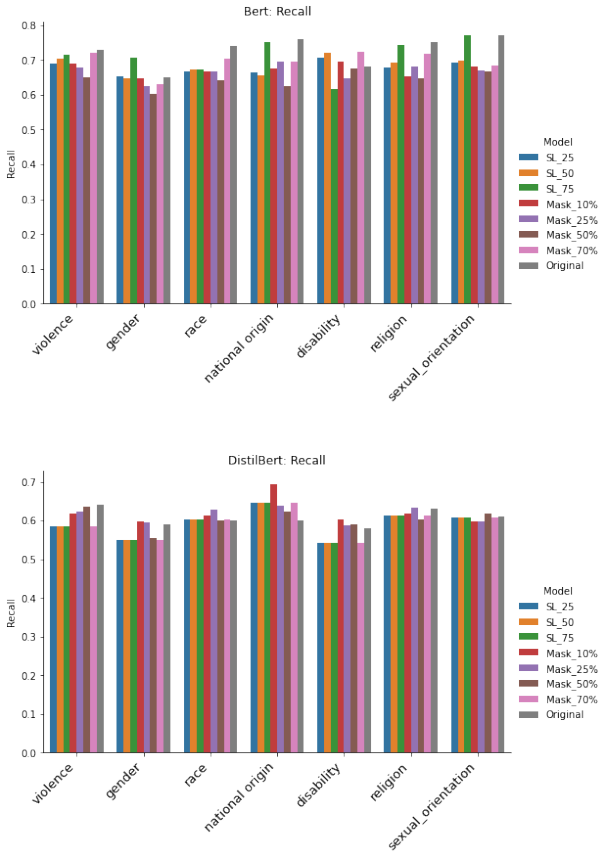


Figure 15: Bert Recall (Up) and DistillBert Recall (Bottom)

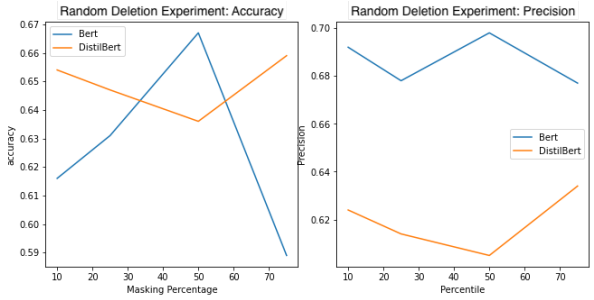


Figure 16: Random Deleition Experiment: Accuracy (Left) Random Deleition Experiment: Precision (Right)

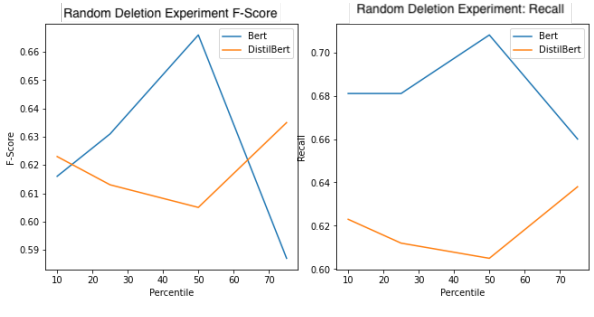


Figure 17: Random Deleition Experiment: F-Score (Left) Random Deleition Experiment: Recall (Right)

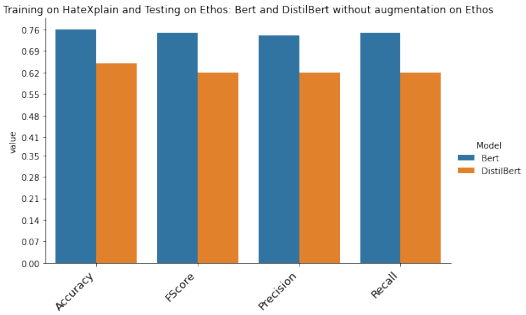


Figure 20: Training on HateXplain and testing on Ethos: Bert and DistillBert without augmentation on Ethos

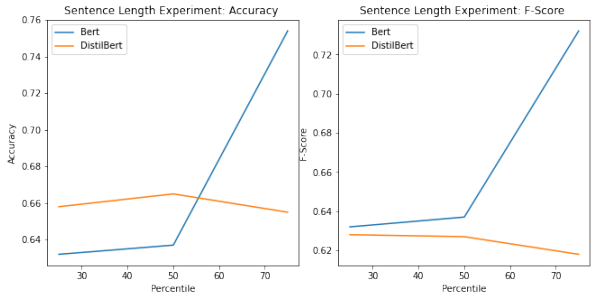


Figure 18: Sentence Length Experiment: Accuracy (Left) F-Score (Right)

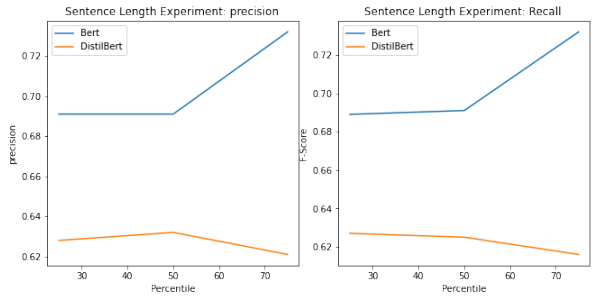


Figure 19: Sentence Length Experiment: Precision (Left) Recall (Right)

B Context

It wouldn't be hyperbolic to say that social media platforms have become a microcosm of the internet in the modern-day world. With a mammoth user base of around 3.6 billion active users (ear, 2022), Meta and its associated products like Facebook (2.9 billion (ear, 2022) active users have largely become the sole "idea of the internet" in some developing countries (Mirani, 2015)

With that amount of user base, sharing, re-sharing of messages, posts, videos, infographics, have seen a stratospheric rise in not only false information but also hate speech with very real-world consequences(hil), like inducing anxiety, self-worth issues, to social media reinforced genocides in Myanmar(BBC, 2018) and riots in India (Kumar, 2022)

	BERT	DistillBERT
Method	Bidirectional Transformer with Masked-Language Modeling and next sentence prediction	Ditilled BERT arquitecture
Number of parameters	Base: 110 million Large: 340 million	Base: 66 million
Performance	Best performance out of similar models	3% worse performance than BERT
Data		
Corpus	3.3 billion terms	3.3 billion terms

Figure 21: Bert model vs DistillBert model

C Models description

- Bert

Bidirectional Encoder Representations from Transformers (BERT) is a stack of transformer encoder layers of 12 encoders with 12 bidirectional self-attention heads (Burstein et al., 2019). BERT can pre-train deep bidirectional representations from unlabeled text by jointly conditioning on the context in every layer (Burstein et al., 2019).

- DistillBERT

DistillBERT is a distilled version of BERT. It has the same general architecture as BERT, with an optimized linear layer and layer normalisation (Sanh et al., 2019). As a result, DistilBERT makes it possible to reduce the size of a BERT model by 40 %, while retaining 97 % of its language understanding capabilities and being 60 % faster (Sanh et al., 2019).

D Data process description

- HateXplain

HateXplain is the first benchmark dataset for hate speech with word and sentence level span annotations that include human labeling (Mathew et al., 2020). The data (from around twenty thousand posts) was collected from previous studies to identify hate speech on Twitter and Gab.

The dataset was annotated by Amazon Mechanical Turk (MTurk) workers in three phases (Mathew et al., 2020). The first stage was to identify if the text could be considered hate speech, offensive speech, or normal speech. The second phase included the identification of the target communities. Third, to explain why the text was identified as hate/offensive speech. With this process involving human rationale, the dataset was labeled in three classes of speech (hateful, offensive, normal), ten target communities (African, Islam, Jewish, LGBTQ, Women, Refugee, Arab, Caucasian, Hispanic, Asian), and rationale explanations (Mathew et al., 2020).

- Ethos

Ethos is a hate speech detection dataset collected from social media platforms (Mollas et al., 2020). For the dataset creation, three stages were considered: platform selection and data collection, data prediction, and manual data annotation.

For the first stage (platform selection and data collection), the authors collected data from Hatebusters and Reddit through the Public Reddit Data Repository. After collecting the data, the Hatebuster Platform performs a classification score using a Support Vector Machine (SVM) model with a linear kernel embedded with TF- IDF to identify the "hate" component. For the second stage, data prediction, the authors used the comments extracted in the first stage to assign useful labels to the available unlabelled set and perform a grid search among some classification methods in the currently expanded dataset. For the Data Annotation stage, the authors used a combination of query strategies to pick informative comments for manual annotation (Mollas et al., 2020).

After the dataset was created, there was a Data Validation stage in which contributors were asked to identify if the comments contained hate speech or not. Afterward, three questions were considered: whether the comment incites violence, whether the comment includes directed hate speech or whether it contains a generalized hate speech. Finally, the contributors chose a category that better reflected the hate speech comment concerning the following categories: gender, race, national origin, disability, religion, and sexual orientation. For the final process, the dataset configuration, there was a manually checking of the results for any misclassification. Few errors were found, which reassured the quality of the annotators' work for classifying the dataset (Mollas et al., 2020).

From this process, two datasets were obtained. The first obtained dataset includes 998 comments and a binary label on the presence or absence of hate speech. The second obtained dataset includes 433 hate speech messages with eight labels: violence, directed vs. generalized, gender, race, national origin, disability, sexual orientation, and religion (Mollas et al., 2020).

E Attention weights

Figure 22 shows a heatmap for the raw attention weights for a Bert model trained on HateXplain dataset without augmentation. The heatmap shows the attention weights at the 7th layer and the 9th attention head. This heatmap shows that the word handsome has high weight for I. However, this may not be sufficient to interpret the model. "Attention rollout" and "attention flow" are two approaches that can be used in future analysis to interpret attention weights while considering the weighted flow of information from input embeddings to an output hidden layer (Abnar and Zuidema, 2020).

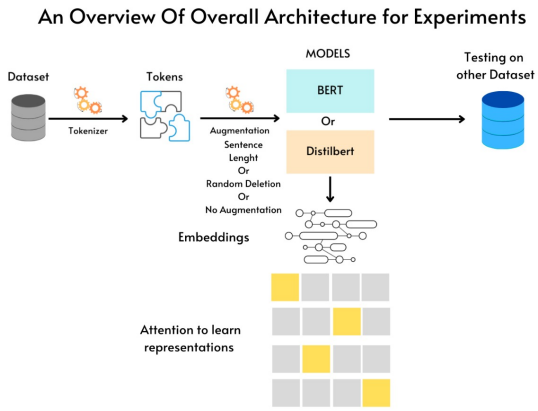


Figure 23: An Overview of Overall Architecture for Experiments

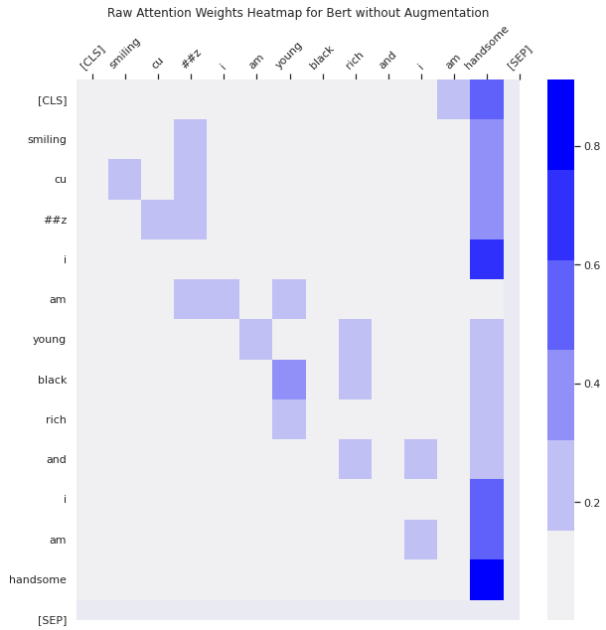


Figure 22: Raw attention weights heatmap for Bert model without augmentation. The figure represents the attention weights at the 7th layer and the 9th attention head

F Model parameters

Both Bert and DistilBert were fine tuned for 5 epochs. Adam optimiser was used for both models with learning rates of 2e-5 and 5e-5 for bert and

DistilBert respectively. The token length was set to 128 for both models to allow for fast processing. Both models had 12 attention heads but Bert had 12 hidden layers while DistilBert had 6 hidden layers. The dropout rate was set to 0.1 for Bert and 0.2 for DistilBert. The dataset was split into train, development and testing sets with a ratio of 8:1:1 respectively.